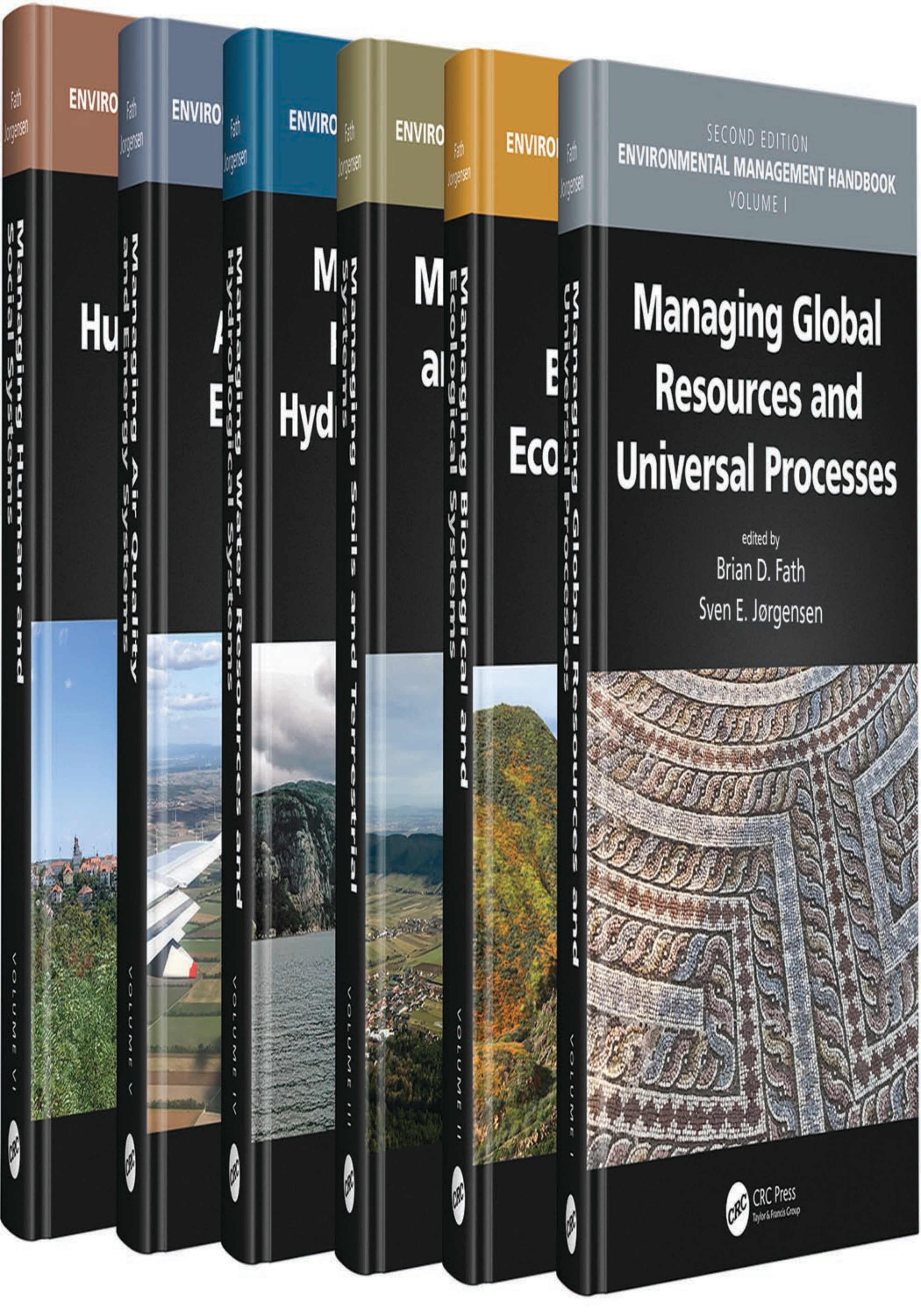


SECOND EDITION

ENVIRONMENTAL MANAGEMENT HANDBOOK

VOLUME I

Managing Global Resources and Universal Processes

edited by

Brian D. Fath

Sven E. Jørgensen

CRC Press
Taylor & Francis Group

SECOND EDITION

ENVIRONMENTAL MANAGEMENT HANDBOOK

VOLUME I

Managing Global Resources and Universal Processes

edited by

Brian D. Fath

Sven E. Jørgensen

CRC Press
Taylor & Francis Group

Managing Global Resources and Universal Processes

Environmental Management Handbook, Second Edition

Edited by
Brian D. Fath and Sven E. Jørgensen

Volume 1
Managing Global Resources and Universal Processes

Volume 2
Managing Biological and Ecological Systems

Volume 3
Managing Soils and Terrestrial Systems

Volume 4
Managing Water Resources and Hydrological Systems

Volume 5
Managing Air Quality and Energy Systems

Volume 6
Managing Human and Social Systems

Managing Global Resources and Universal Processes

Second Edition

Edited by
Brian D. Fath and Sven E. Jørgensen

Assistant to Editor
Megan Cole

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

Cover photo: Paphos Archeological Park, Cyprus, N. Fath

Second edition published 2021
by CRC Press
6000 Broken Sound Parkway NW, Suite 300, Boca Raton, FL 33487-2742

and by CRC Press
2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

First edition published by CRC Press 2013
CRC Press is an imprint of Taylor & Francis Group, LLC

ISBN: 978-1-138-34263-7 (hbk)
ISBN: 978-0-429-34613-2 (ebk)

Typeset in Minion
by codeMantra

Contents

SECTION I Anthropogenic Chemicals: Human Manufactured and Activities

SECTION II Natural Elements and Chemicals

SECTION III Basic Environmental Processes

Preface

Given the current state of the world as compiled in the massive Millennium Ecosystem Assessment Report, humans have changed ecosystems more rapidly and extensively during the past 50 years than in any other time in human history. These are unprecedented changes that need certain action. As a result, it is imperative that we have a good scientific understanding of how these systems function and good strategies on how to manage them.

In a very practical way, this multivolume *Environmental Management Handbook* provides a comprehensive reference to demonstrate the key processes and provisions for enhancing environmental management. The experience, evidence, methods, and models relevant for studying environmental management are presented here in six stand-alone thematic volumes, as follows:

VOLUME 1 – Managing Global Resources and Universal Processes
VOLUME 2 – Managing Biological and Ecological Systems
VOLUME 3 – Managing Soils and Terrestrial Systems
VOLUME 4 – Managing Water Resources and Hydrological Systems
VOLUME 5 – Managing Air Quality and Energy Systems
VOLUME 6 – Managing Human and Social Systems

In this manner, the handbook introduces in the first volume the general concepts and processes used in environmental management. The next four volumes deal with each of the four spheres of nature (biosphere, geosphere, hydrosphere, and atmosphere). The last volume ties the material together in its application to human and social systems. These are very important chapters for a wide spectrum of students and professionals to understand and implement environmental management. In particular, features include the following:

- The first handbook that demonstrates the key processes and provisions for enhancing environmental management.
- Addresses new and cutting-edge topics on ecosystem services, resilience, sustainability, food–energy–water nexus, socio-ecological systems, etc.
- Provides an excellent basic knowledge on environmental systems, explains how these systems function, and gives strategies on how to manage them.
- Written by an outstanding group of environmental experts.

Since the handbook covers such a wide range of materials from basic processes, to tools, technologies, case studies, and legislative actions, each handbook entry is further classified into the following categories:

APC: Anthropogenic chemicals: The chapters cover human-manufactured chemicals and their activities

COV: Indicates that the chapters give comparative overviews of important topics for environmental management

CSS: The chapters give a case study of a particular environmental management example
DIA: Means that the chapters are about diagnostic tools: monitoring, ecological modeling, ecological indicators, and ecological services
ELE: Focuses on the use of legislation or policy to address environmental problems
ENT: Addresses environmental management using environmental technologies
NEC: Natural elements and chemicals: The chapters cover basic elements and chemicals found in nature
PRO: The chapters cover basic environmental processes.

Overall, these volumes will be a valuable resource for all libraries supporting programs in environmental science and studies, earth science, geography, and policy.

In this volume, #1, the collection of over 50 entries provides an overview of global resources and universal processes. This serves as a good introduction to the key aspects of environmental management and includes descriptions of elements of the periodic table as well as organic and inorganic processes leading to pollution and alteration of natural conditions. A new chapter on telecoupling shows the long distance relations and interactions that mark most environmental systems.

Brian D. Fath
Brno, Czech Republic
December 2019

Editors

Brian D. Fath is Professor in the Department of Biological Sciences at Towson University (Maryland, USA) and Senior Research Scholar at the International Institute for Applied Systems Analysis (Laxenburg, Austria). He has published over 180 research papers, reports, and book chapters on environmental systems modeling, specifically in the areas of network analysis, urban metabolism, and sustainability. He has co-authored the books *A New Ecology: Systems Perspective* (2020), *Foundations for Sustainability: A Coherent Framework of Life–Environment Relations* (2019), and *Flourishing within Limits to Growth: Following Nature's Way* (2015). He is also Editor-in-Chief for the journal *Ecological Modelling* and Co-Editor-in-Chief for *Current Research in Environmental Sustainability*. Dr. Fath was the 2016 recipient of the Prigogine Medal for outstanding work in systems ecology and twice a Fulbright Distinguished Chair (Parthenope University, Naples, Italy in 2012 and Masaryk University, Czech Republic in 2019). In addition, he has served as Secretary General of the International Society for Ecological Modelling, Co-Chair of the Ecosystem Dynamics Focus Research Group in the Community Surface Modeling Dynamics System, and member and past Chair of Baltimore County Commission on Environmental Quality.

Sven E. Jørgensen (1934–2016) was Professor of environmental chemistry at Copenhagen University. He received a doctorate of engineering in environmental technology and a doctorate of science in ecological modeling. He was an honorable doctor of science at Coimbra University (Portugal) and at Dar es Salaam (Tanzania). He was Editor-in-Chief of *Ecological Modelling* from the journal inception in 1975 until 2009. He was Editor-in-Chief for the *Encyclopedia of Environmental Management* (2013) and *Encyclopedia of Ecology* (2008). In 2004, Dr. Jørgensen was awarded the Stockholm Water Prize and the Prigogine Medal. He was awarded the Einstein Professorship by the Chinese Academy of Sciences in 2005. In 2007, he received the Pascal Medal and was elected a member of the European Academy of Sciences. He had published over 350 papers, and has edited or written over 70 books. Dr. Jørgensen gave popular and well-received lectures and courses in ecological modeling, ecosystem theory, and ecological engineering worldwide.

Contributors

Imad A.M. Ahmed
Lancaster Environment Center
Lancaster University
Lancaster, United Kingdom

Ronald G. Amundson
College of Natural Resources
University of California—Berkeley
Berkeley, California

Jirapat Ananpattarachai
Faculty of Engineering
Center of Excellence for Environmental Research and Innovation
Naresuan University
Phitsanulok, Thailand

Lars Bergström
Department of Soil Science
Swedish University of Agricultural Sciences (SLU)
Uppsala, Sweden

Angelika Beyer
Department of Analytical Chemistry
Chemical Faculty
Gdansk University of Technology
Gdansk, Poland

Marek Biziuk
Department of Analytical Chemistry, Chemical Faculty
Gdansk University of Technology
Gdansk, Poland

Frederick Paxton Cardell Blamey
School of Agriculture and Food Sciences
The University of Queensland
St. Lucia, Queensland, Australia

Elke Bloem
Institute for Crop and Soil Science
Julius Kuhn Institute (JKI)
Braunschweig, Germany

Pascal Boeckx
Faculty of Agricultural and Applied Biological Sciences
University of Ghent
Ghent, Belgium

John Borden
Department of Biological Sciences
Simon Fraser University
Burnaby, British Columbia, Canada

Céline Boutin
Science and Technology Branch
Environment Canada
Carleton University
Ottawa, Ontario, Canada

Dominic A. Brose
University of Maryland
College Park, Maryland

Ray Correll
Commonwealth Scientific and Industrial Research Organization (CSIRO)
Adelaide, South Australia, Australia

Marianna Czaplicka
Institute of Non-Ferrous Metals
and
Department of Analytical Chemistry
Silesian University of Technology
Gliwice, Poland

Vera Lucia S.S. de Castro
Ecotoxicology and Biosafety Laboratory
Brazilian Agricultural Research Corporation
(Embrapa Environment)
São Paulo, Brazil

Malcolm Devine
Aventis CropScience Canada Co.
Saskatoon, Saskatchewan, Canada

Peter Dillon
Commonwealth Scientific and Industrial
Research Organization (CSIRO)
Adelaide, South Australia, Australia

J.K. Dubey
Department of Entomology
Dr. Y.S. Parmar University of Horticulture
and Forestry
Solan, India

Renata Gaj
Institute of Soil Science
Agricultural University
Poznan, Poland

Ardell D. Halvorson
U.S. Department of Agriculture (USDA)
Fort Collins, Colorado

Silvia Haneklaus
Institute for Crop and Soil Science
Julius Kuhn Institute (JKI)
Braunschweig, Germany

Philippe Hinsinger
Sun and Environment Unit
National Institute for Agricultural Research
(INRA)
Montpellier, France

Zhengyi Hu
Institute of Soil Science
Chinese Academy of Sciences
Nanjing, China

Hayriye Ibrikci
Soil Science and Plant Nutrition Department
Cukurova University
Adana, Turkey

Alexandra Izosimova
St. Petersburg Agricultural Physical
Research Institute
St. Petersburg, Russia

Bruce R. James
University of Maryland
College Park, Maryland

Philip M. Jardine
Oak Ridge National Laboratory
Oak Ridge, Tennessee

Sven Erik Jørgensen
Institute A, Section of Environmental
Chemistry
Copenhagen University
Copenhagen, Denmark

Puangrat Kajitvichyanukul
Faculty of Engineering
Center of Excellence for Environmental Research
and Innovation
Naresuan University
Phitsanulok, Thailand

Gabriella Kakonyi
Kroto Research Institute
Sheffield University
Sheffield, United Kingdom

Keith A. Kelling
Professor Emeritus, Department of
Soil Science
University of Wisconsin—Extension
Madison, Wisconsin

Rami Keren
Agricultural Research Organization of Israel
Bet-Dagan, Israel

Peter Kleinman
Pasture Systems and Watershed Management Research Unit
U.S. Department of Agriculture (USDA)
University Park, Pennsylvania

Rai Kookana
Commonwealth Scientific and Industrial Research Organization (CSIRO)
Adelaide, South Australia, Australia

Peter Martin Kopittke
School of Agriculture and Food Sciences
The University of Queensland
St. Lucia, Queensland, Australia

David P. Kreutzweiser
Canadian Forest Service
Natural Resources Canada
Sault Sainte Marie, Ontario, Canada

Tore Krogstad
Department of Plant and Environmental Sciences
Norwegian University of Life Science
Aas, Norway

Witold Kurylak
Institute of Non-Ferrous Metals
Gliwice, Poland

Gamini Manuweera
Chemicals and Health Branch
United Nations Environmental Program
Geneva, Switzerland

Richard McDowell
AgResearch Ltd., Invermay Agricultural Center
Mosgiel, New Zealand

Leslie D. McFadden
Department of Earth and Planetary Sciences
University of New Mexico
Albuquerque, New Mexico

Ronald G. McLaren
Soil, Plant, and Ecological Sciences Division
Lincoln University
Canterbury, New Zealand

Mike J. McLaughlin
Land and Water, Commonwealth Scientific and Industrial Research Organization (CSIRO)
Glen Osmond, South Australia, Australia

Mallavarapu Megharaj
Commonwealth Scientific and Industrial Research Organization (CSIRO)
Adelaide, South Australia, Australia

Neal William Menzies
School of Agriculture and Food Sciences
The University of Queensland
St. Lucia, Queensland, Australia

J. David Miller
Department of Chemistry
Carleton University
Ottawa, Ontario, Canada

H. Curtis Monger
Department of Agronomy and Horticulture
New Mexico State University
Las Cruces, New Mexico

Ravendra Naidu
Commonwealth Scientific and Industrial Research Organization (CSIRO)
Adelaide, South Australia, Australia

Niels Erik Nielsen
Plant Nutrition and Soil Fertility Laboratory
Department of Agricultural Sciences
Royal Veterinary and Agricultural University
Frederiksberg, Denmark

David R. Parker
Department of Soil and Environmental Sciences
University of California—Riverside
Riverside, California

Judith F. Pedler
Department of Environmental Sciences
University of California—Riverside
Riverside, California

Mark B. Peoples
Agriculture and Food, Commonwealth Scientific and Industrial Research Organization (CSIRO)
Canberra, Australian Capital Territory, Australia

Robert Pietrzak
Department of Chemistry
Adam Mickiewicz University
Poznan, Poland

Paola Poli
Department of Genetics, Biology of Microorganisms, Anthropology, and Evolution
University of Parma
Parma, Italy

Abdul Rashid
Pakistan Atomic Energy Commission
Islamabad, Pakistan

Lisa A. Robinson
Independent Consultant
Newton, Massachusetts

Alexandru V. Roman
School of Public Administration
Florida Atlantic University
Boca Raton, Florida

John Ryan
International Center for Agricultural Research in the Dry Areas (ICARDA)
Aleppo, Syria

Vilma Sandström
Sustainability Science
LUT-University
Lahti, Finland

William H. Schlesinger
Department of Geology and Botany
Duke University
Durham, North Carolina

Ewald Schnug
Institute for Crop and Soil Science
Julius Kuhn Institute (JKI)
Braunschweig, Germany

A. Paul Schwab
Department of Agronomy
Purdue University
West Lafayette, Indiana

Andrew N. Sharpley
University of Arkansas
Fayetteville, Arkansas

Paul K. Sibley
School of Environmental Sciences
University of Guelph
Guelph, Ontario, Canada

Bogdan Skwarzec
Faculty of Chemistry
University of Gdansk
Gdansk, Poland

Rolf Sommer
International Center for Agricultural Research in the Dry Areas (ICARDA)
Aleppo, Syria

Gerd Sparovek
College of Agriculture
Graduate School of Agriculture Luiz de Queiroz (ESALQ)
University of São Paulo
São Paulo, Brazil

Scott J. Sturgul
Nutrient and Pest Management Program
University of Wisconsin
Madison, Wisconsin

Donald L. Suarez
U.S. Salinity Laboratory
Agricultural Research Service (USDA-ARS)
U.S. Department of Agriculture
Riverside, California

Meena Thakur
Department of Entomology
Dr. Y.S. Parmar University of Horticulture and Forestry
Solan, India

Oswald Van Cleemput
Faculty of Agricultural and Applied Biological Sciences
University of Ghent
Ghent, Belgium

Doug Walsh
Washington State University
Prosser, Washington

Leszek Wachowski
Department of Chemistry
Adam Mickiewicz University
Poznan, Poland

Johannes Bernhard Wehr
School of Agriculture and Food Sciences
The University of Queensland
St. Lucia, Queensland, Australia

W.W. Wenzel
Institute of Soil Research
University of Natural Resources and Life Sciences
Vienna, Austria

Larry P. Wilding
Department of Soil and Crop Sciences
Texas A&M University
College Station, Texas

He Zhong
Pesticide Environment Impact Section
Public Health Entomology Research and Education Center
Florida A&M University
Panama City, Florida

I

Anthropogenic Chemicals: Human Manufactured and Activities

1

Acaricides

Doug Walsh

Introduction

Approximately 45,000 species of mites and hundreds of species of ticks are described worldwide. Many thousands of species still remain unidentified. About half are plant-feeding species, and among these, about half are in the superfamily Eriophyoidea (gall, bud, and rust mites). Most of the other plant-feeding mites are classified into the superfamilies Tetranychoidea and Tarsonemidae. The superfamily Tetranychoidea includes the economically important spider, flat, and fowl mites, and the superfamily Tarsonemidae includes the economically important broad, cyclamen mites and Varroa mites. Over another 3000 mite species are loosely classified in the order Astigmata. Economically important species include feather and scabies mites. Ticks are placed in the superfamily Ixodoidea and all are ectoparasites (blood feeders) of vertebrate animals.[1]

Most mites are small to minute and mites are universally cryptic, making them difficult to detect. Often infestations are overlooked. Mites are often colonizers of new or disturbed habitats, and once established on a new host, mites possess biological characteristics that permit rapid increases in population abundance. Factors in most mites' lifestyle that lead to rapid population buildup include high egg production, various modes of reproduction (parthenogenesis, pedogenesis, and sexual), short life cycles, a myriad of dispersal techniques, and adaptability to diverse ecological conditions.[1] These traits combined with an exponential increase in worldwide transport of humans and plant and animal products will likely contribute to increased concerns over mite pests in the future.

In plant-based agriculture, Van de Vrie et al.[2] observed that outbreaks of mite populations were uncommon historically in systems where productivity languished far below the levels achieved in modern production agriculture. Spider mite populations stayed below observable levels due to natural regulation by predators, disease, and poor nutrition from low-quality host plants. Van de Vrie et al.[2] went on to observe that mite populations often experienced outbreaks in agroecosytems where production levels were bolstered by the use of synthetic inputs including fertilizers and pesticides. When crop production is optimized (i.e., not limited by water, nutrients, competition from weeds, or predatory mites and insects), the plants in production become an excellent food source for mite pests. Under these conditions, the developmental rate, fecundity, and life span of mites are increased and contribute to population outbreaks.

Spider Mite Pests

A number of mite species can achieve pest status at high population abundance. Spider mites develop through several stages: egg, six-legged larva, eight-legged protonymph, deutonymph, and adult. Males typically reach maturity before females and will position themselves near developing quiescent females. When an adult female emerges, copulation will often occur immediately. Under optimal conditions, most mite species can develop from egg to adult in 6 to 10 days. Egg laying can begin as soon as one or two days after maturing to adults. Most spider mite species overwinter as mated adult females. A notable exception is the European red mite that overwinters as eggs.[3]

A Big Drain from the Feeding of Such Small Pests

At the microscopic level, significant quantities (relative to mite size) of plant juices pass through the digestive tract of spider mites as they feed on leaf tissues. McEnroe[4] estimated this volume at 1.2×10^{-2} microliters per mite per hour. This quantity represents roughly 50% of the mass of an adult female spider mite. Leisering and Beitrag[5] calculated that the number of photosynthetically active leaf cells that are punctured and emptied per mite is 100 per minute. In gut content studies of two-spotted mites, Mothes and Seitz[6] observed only thylakoid granules inside their digestive tract following feeding. The thylakoid grana on which spider mites focus their feeding are key photosynthetic engines in plant cells. The grana consist of 45%–50% protein, 50%–55% lipid, and minute amounts of RNA and DNA.[7] Water and other low-density plant cell contents are directly excreted.[4]

At the macroscopic level, damage from mite feeding can cause leaf bronzing, stippling, or scorching. For most horticultural crops, economic loss is caused by a drop in yield or quality due to reduction in photosynthesis.

Spider Mite Outbreaks Are Promoted by Hot, Dry Weather

Water stress, wind, and dust all contribute to the potential for mite outbreaks. When mite outbreaks occur, acaricide treatments are often used for suppression.

Varroa mites *Varroa jacobsoni* provide an ideal example of how rapidly a mite species can spread and exploit a new habitat. First recorded in honeybee colonies in Southeast Asia in 1904, Varroa mites are now pandemic. Varroa mites feed parasitically on an individual bee's hemolymph fluid, weakening the bee and often causing premature death. Mites attach themselves to foraging workers in order to spread themselves from one hive to another. This mite can severely weaken bees, and an unchecked mite population will almost certainly lead to the premature death of a honeybee colony. Apiculturists speculate that Varroa mite has contributed substantially to the collapse of feral honeybee populations worldwide.[8]

The northern fowl mite *Ornithonyssus sylviarum* is a common pest of domestic fowl and other wild birds commonly associated with human settlements. The nymphs and adults have piercing mouthparts and seek blood meals. Mite populations build up rapidly and a generation can be completed in 5 to 12 days. Several generations occur each year. Northern fowl mite spends virtually its entire life on the host bird.[9]

Deer and dog ticks *Ixodes scapularis* and *Dermacentor variabilis* are two common ticks to which acaricides are applied for on a consistent basis, especially since both are parasitic feeders on mammals. Deer ticks are a significant concern since they are the primary vector for Lyme disease.[10]

Mange or scabies in livestock is a skin condition caused by microscopic mites in or on the skin. The mites cause intense itching and discomfort that is associated with decreased feed intake and production. Scratching and rubbing result in extensive damage to hides and fleece. Mange mites are able to cause mange on different species of livestock but are somewhat host specific, thus infecting some species more severely than others. The three most important types of mange are as follows: sarcoptic mange, caused by *Sarcoptes scabiei* feeding; psoroptic mange, caused by *Psoroptes ovis* feeding; and chorioptic mange, caused by *Chorioptes bovis* feeding.[11] Infestations of these mites on their respective livestock, domestic pet, or human host will cause skin irritation and itching and leave entry points for secondary infections. Weight gain can be reduced in livestock, pets can lose hair and itch persistently, and disfigurement can occur in humans. Acaracides are often applied to suppress mite populations parasitizing humans, pets, and livestock.

Smothering Agents

Solutions containing petroleum-based horticultural oils, vegetable oils, or agricultural soaps are applied to many crops and, occasionally, livestock. Application of these types of products kills spider mites through suffocation. Unfortunately, oils and soaps can prove phytotoxic to crop plants and are typically not effective on mites or ticks infesting livestock, pets, or humans. Mites on animal hosts are typically cryptic or subcutaneous, so acaricide coverage is an impediment to effective control.

Organochlorines

Endosulfan and dicofol are organochlorine miticides registered for use on many crops. Unlike many other organochlorine pesticides, endosulfan and dicofol are relatively non-persistent in the environment. Organochlorine acaricides interfere with the transmission of nerve impulses and disrupt the nervous system of pest mites. Organochlorine acaricides are more effective at killing mites at warmer temperatures. Overuse of organochlorine acaricides in commercial situations has resulted in the development of tolerance in many pest mite populations. Organochlorines were used substantially in the mid-20th century, but regulatory actions and public health and environmental concern have eliminated their use in most developed countries (though some continue to use them). Lindane was commonly used for mange mite in pets, livestock, and humans. Only in limited circumstances is lindane still permitted as a pharmaceutical second-line treatment. However, use of lindane continues in developing countries due to its low cost, effectiveness, and persistence.

Organophosphates and Carbamates

Many organophosphate and carbamate pesticides have acaricidal activity. Studies have demonstrated significant mite control with applications of parathion, TEPP (tetraethyl pyrophosphate), and aldicarb. Spider mites are listed as target pests on many organophosphate and carbamate products. However, many mite populations following long-term exposure have developed resistance to the toxic effects of organophosphates.[12] Carbaryl, a common carbamate, continues to be a mainstay for mite control on livestock and poultry, but its use on domestic pets and households is no longer permitted in most developed countries. The use of carbaryl continues extensively in many developing countries in domestic settings.

Organotins

Miticides in this category were synthesized in the 1960s and 1970s and registered for commercial use in the United States in the 1970s. They have been used extensively for their ability to quickly knock down spider mite populations through contact activity. Fenbutatin-oxide has been used extensively since the 1970s. Cyhexatin was used extensively in the 1970s and 1980s, but regulatory actions have now limited its use. Efficacy of the organotin acaricides is improved with warmer weather. Overuse of cyhexatin during the mid-1980s led to the development of resistance in several cropping and livestock production systems. However, populations of pest mites can regain susceptibility to organotins following a period of non-exposure.[13]

Propargite

This acaricide has been used since the 1960s. It provides effective suppression of pest mites on many crops. Regulatory constraints have resulted in the cancellation of a number of uses. Identification of propargite as a dermal irritant has led to substantial increases in time required following application before re-entry is permitted into the treated site.

Amidines

Amitraz is a miticide that once had significant use in plant and animal agriculture. At present, its use is restricted to only a small subset of the domestic pet care market.

Ovicides

Clofentazine and hexythiazox are selective carboxamide ovicidal acaricides. Spider mite eggs exposed to either compound fail to hatch. These acaricides are selective and aid in the conservation of populations of beneficial arthropods. These acaricides are typically used relatively early in the production season before mite populations reach outbreak conditions.

Antimetabolites

A number of miticidal compounds have been developed within the past 30 years. These include avermectins, pyridazinones, carbazates, and pyrroles. Pest mortality results from disruption of metabolic pathways typically within the mitochondria of nerve cells of spider mites.[14] Avermectins, ivermectins, and related compounds are fermentation products derived from mycelial extracts of *Streptomyces* species (reviewed by Burg and Stapley).[17] Avermectins are locally systemic (translaminar) in plant tissues, [15] and ivermectins can be applied dermally, by injection, suppository, or in a bolus to livestock and domestic pets. The ivermectins are the predominant parasiticide used in livestock production today. A number of products have been commercialized in recent years. Pyridaben is a pyridazinone recently registered for use on ornamentals and some tree crops. Bifenazate is a carbazate acaricide. It has a new mode of action that is not clearly understood, but it has proven toxicologically safe in mammalian studies. Bifenazate is registered on ornamentals and food products. Chlorfenapyr is a synthetic pyrrole that has been commercially available on cotton. Other uses are pending.

Synthetic Pyrethroids

Fenpropathrin and bifenthrin are two synthetic pyrethroid insecticides registered for control of spider mites in plant agriculture. Permethrin is registered for mite control on livestock and poultry. Mites have a well-documented history of rapidly developing resistance to pyrethroid insecticides in both plant and animal production systems, and resurgence of spider mite populations following pyrethroid application is typical.[18]

Tetronic Acids

Spiromesifen and spirotetramat are acaricides in a recently introduced class of selective chemistry tetronic acids that exhibit a broad-spectrum insecticidal acaricidal activity against mites. Their mode of action is by inhibition of lipid biosynthesis that affects the egg and immature stages of mites. Foliar sprays of spiromesifen are translaminar in plants and effective against mites in many cropping systems. Spirotetramat has a relatively unique property among currently registered acaricides in that it is phloem systemic within the plant it is applied to. These two acaricides have recently entered the acaricide market and are quickly gaining in use in production agriculture for mite control.

Application Technology

Mite pests can prove difficult to control with acaricides due to their potential for high population abundance, small size, and propensity to live on the bottom surfaces of leaves or within the folds of plant tissues. Good acaricide spray coverage is essential for mite control, particularly for acaricides that kill on contact with the pest mite.

Combating Miticide Resistance

Following repeated exposure, spider mite populations have a history of rapidly developing resistance to acaricides.[16] Alternating acaricides that have different modes of action reduces the potential for development of resistance to acaricides within specific modes of activity. Other techniques to discourage resistance development include spraying only when necessary and treating only infested portions of the crop. Organophosphate, carbamate, and pyrethroid insecticide applications can induce spider mite outbreaks. If possible, avoid early-season insecticide application or apply insecticides that are less disruptive to beneficial arthropods. Careful selection and use of insecticides can potentially reduce the number of miticide applications required later in the season.

References

1. Krantz, G.W.; Walter, D.E., Eds. *A Manual of Acarology,* 3rd Ed.; Texas Tech University Press: Lubbock, Texas, 2009; 807 pp.
2. Van de Vrie, M.; McMurtry, J.A.; Huffaker, C.B. Ecology of mites and their natural enemies. A review. III Biology, ecology, pest status, and host plant relations of tetranychids. Hilgardia **1972**, *41,* 345–432.
3. Bostanian, N.J. The relationship between winter egg counts of the European red mite *Panonychus ulmi* (Acari: Tetranychidae) and its summer abundance in a reduced spray orchard. Exp. Appl. Acarol. **2007**, *42,* 185–195.
4. McEnroe, W.D. The role of the digestive system in the water balance of the two-spotted spider mite. Adv. Acarol. **1963**, *1*, 225–231.
5. Leisering, R.; Beitrag, O. Beitrag zum phytopatologischen Wirkungsmeechanismus von Tetranychus urticae. Pflanzenschutz **1960**, *67,* 525–542.
6. Mothes, U.; Seitz, K.A. Functional microscopic anatomy of the digestive system of *Tetranychus urticae* (Acari: Tetranychidae). Acarologia **1981**, *22*, 257–270.
7. Noggle, G.R. The organization of plants. In *Introductory Plant Physiology;* Noggle, G.R., Fritz, G.J., Eds.; Prentice Hall: Englewood Cliffs, New Jersey, 1983; 9–38.
8. Mangum, W.A. Honey bee biology: The third annual report on the coexistence of my North Carolina bees with varroa mites. Am. Bee J. **2009**, *149,* 63–65.
9. Mullens, B.A. Temporal changes in distribution, prevalence and intensity of northern fowl mite (*Ornithonyssus sylviarum*) parasitism in commercial caged laying hens, with a comprehensive economic analysis of parasite impact. Vet. Parasitol. **2009**, *160,* 116–133.

10. Diuk-Wasser, M.A. Field and climate-based model for predicting the density of host-seeking nymphal *Ixodes scapularis,* an important vector of tick-borne disease agents in the eastern. U. S. Global Ecol. Biogeogr. **2010**, *19,* 504–514.
11. Vercruysse, J. World Association for the Advancement of Veterinary Parasitology (W.A.A.V.P.) guidelines for evaluating the efficacy of acaricides against (mange and itch) mites on ruminants [electronic resource]. Vet. Parasitol. **2006**, *136,* 55–66.
12. Smissaeret, H.R.; Voerman, S.; Oostenbrugge, L.; Reenooy, N. Acetylcholinesterases of organophosphate-susceptible and resistant spider mites *Tetranychus urticae*. J. Agric. Food Chem. **1970**, *18,* 66–75.
13. Hoy, M.A.; Conley, J.; Robinson, W. Cyhexatin and fenbutatin-oxide resistance in Pacific spider mite (Acari: Tetranychidae) stability and mode of inheritance. J. Econ. Entomol. **1988**, *81,* 57–64.
14. Hollingsworth, R.M.; Ahammadsahib, K.I.; Gadelhak, G.; McLaughlin, J.L. New inhibitors of Complex I of the mitochodrial electron transport chain with activity as pesticides. Biochem. Soc. Trans. **1994**, 22, 230–233.
15. Walsh, D.B.; Zalom, F.G.; Shaw, D.V.; Welch, N. C. Effect of strawberry plant physiological status on the translaminar activity of avermectin B1 and its efficacy on the two-spotted spider mite *Tetranychus urticae* Koch (Acari: Tetranychidae). J. Econ. Entomol. **1996**, *89* (5), 1250–1253.
16. Leeuwen, T.V.; Dermauw, W.; Tirry, L.; Vontas, J.; Tsagkarakou, A. Acaricide resistance mechanisms in the two- spotted spider mite *Tetranychus urticae* and other important Acari. Insect Biochem. Mol. Biol. **2010**, *40,* 563–572.
17. Burg, R.W., and E.O. Stapley. Isolation and characterization of the producing organism. In W.C. Campbell (ed) Ivermectin and Abamectin. Springer-Verlag, New York, N.Y. 1989. pp. 24–32.
18. Leigh, T.F. Cotton. In W. Helle and M.W. Sabelis (eds.) World Crop Pests: Spider Mites. Elsevier Press, Amsterdam, the Netherlands. 1990. pp. 349–358.

2

Endocrine Disruptors

Vera Lucia S.S.
de Castro

Introduction

Endocrine-disrupting chemicals (EDCs) refer to anthropogenic compounds that are able to mimic, antagonize, alter, or modify normal hormonal activity. Dichlorodiphenyl-trichloroethane (DDT), an insecticide first produced on a wide scale in 1945, was used extensively during the 1960s and 1970s and was the first chemical found to be estrogenic. Subsequently, other organochlorine insecticides such as dieldrin, endosulfan, and methoxychlor were found to be estrogenic. Endocrine-disrupting chemicals include environmental estrogens such as *o,p*-DDT, endosulfan, non-planar polychlorinated biphenyl (PCB), octyl-and nonylphenols, the antiandrogens such as vinclozolin and DDE, and the thyroid hormone disrupters such as fenvalerate and benzene hexachloride.[1]

Endocrine-disrupting chemicals are a significant public health concern since these compounds interfere with normal function of pathways responsible for both reproduction and development and can affect the endocrine system, interfering in the production or action of hormones or compromising sexual identity, fertility, or behavior.[2,3] Besides, many of them are persistent in the environment, can be found in waters and sediments, and are easily transported long distances in the atmosphere.[4]

In recent years, numerous studies have suggested that many environmentally persistent chemicals have a potential to disrupt normal functions of the endocrine system. The field of endocrine disrupters, such as the special susceptibility of the developing organism and early induction of latent effects, has come a long way since its initial impetus in 1991.[5] Specially, the possible effects of EDCs on early events of proper gonadal development—which is dependent on intercellular signaling mechanisms—deserve attention since the early steps in mammalian sexual development are vulnerable to genetic and environmental perturbation.[6]

Exposure to EDCs is associated with dysfunctions of metabolism, energy balance, thyroid function, and reproduction, and an increased risk of endocrine cancers. These multifactorial disorders can occur through molecular epigenetic changes induced by exposure to EDCs early in life, the expression of which may not manifest until adulthood.[7] Effects attributed to the EDCs include developmental demasculinization and feminization in reptiles, mammals, amphibians, fish, and birds; reduced fecundity in reptiles, birds, and fish; and possibly increased breast cancer rates and reduced sperm counts in humans.[1]

Since hormones, in synergy with genes, are responsible for sex-related differences in anatomical, physiological, and behavioral traits, even if EDCs are present in minute amounts in environment, their effects in male and female physiology could be greater than before expected. They might also prejudice the sex steroid hormone–induced integrated physiological responses in women and men. In addition, differences in male and female susceptibility to EDCs could be present even if there is still scarce information available on this aspect.[8]

Mechanisms of Action

Several EDCs may work by multiple mechanisms, including uncharacterized mechanisms of action. Because of cross talk between different components of the endocrine systems, effects may occur unpredictably in endocrine target tissues other than the system predicted to be affected. A few modes of action could contribute to the same outcome, including aromatase inhibition, antiestrogenicity, testosterone biosynthesis disruption, and antiandrogens that alter upregulation of aromatase in the target regions within the brain. More complex biological responses to EDCs will generally represent combinations of several physiological processes integrated through multiple biological pathways.[9,10]

Endocrine disrupters may interfere with the functioning of hormonal systems in at least three possible ways: 1) by mimicking the action of a naturally produced hormone, producing similar but exaggerated chemical reactions in the body; 2) by blocking hormone receptors, preventing or diminishing the action of normal hormones; and 3) by affecting the synthesis, transport, metabolism, and/or excretion of hormones, thus altering the concentrations of natural hormones.[11]

The first characterized mechanism of action of EDCs is to act directly as ligands to steroid hormone nuclear receptors (NRs), in particular, estrogen, androgen, and thyroid NRs. Nuclear receptors are a class of proteins found within cells. In response to the presence of hormones, these receptors work in concert with other proteins to regulate the expression of specific genes by a conformation change. Schematically, NRs may be classified into four classes according to their dimerization and DNA-binding proprieties.[10]

Cross talk between NR-mediated and other signal-transduction pathways is an important aspect of NR-based regulation. This so-called genomic or genotropic signaling is normally slow and sustained, taking hours before biological outcomes become manifest. For example, in the classic view of estrogen action, the effects of 17β-estradiol (E2) were thought of as mediated by the NRs estrogen receptor α and β, acting as ligand-dependent transcription factors, thereby regulating gene expression by binding estrogen response elements in the DNA.[12]

Another type of NR cross talk that has recently been recognized is the non-genomic action of several NRs. Some non-genomic actions of NR ligands are apparently mediated through membrane receptors that are not part of the NR superfamily.[10] For example, it has become clear that E2 can also rapidly and transiently trigger a variety of second messenger signaling events, including the induction of cyclic adenosine monophosphate (cAMP) and adenylate cyclase; the mobilization of intracellular calcium; and the stimulation of PI3K, PKB, and Src with consequent activation of the extracellular-regulated kinases Erk1 and Erk2 in the Src/Ras/Erk cascade. All these effects are believed to be mediated through a membrane-associated or cytosolic estrogen receptor (ER) and have, therefore, been termed non-genomic or extranuclear actions of E2.[12]

The cellular activities of estrogens and xenoestrogens are the result of a combination of extranuclear (non-genomic) and nuclear (genomic) events and highlight the need to take non-genomic effects and signaling cross talk into consideration when screening for environmental estrogen.[12]

Disruption of the endocrine system by xenobiotic compounds is consistently reported in humans and wildlife and is a matter of concern worldwide. A great variety of natural or synthetic chemicals such as

EDCs are thought to exert an acute effect at different levels of the thyroid cascade. It is consensual that EDCs probably act by interfering with thyroid hormone (TH) synthesis, cellular uptake, and metabolism, at the level of TH receptors and also TH transport, by binding to thyroid hormone distributor proteins (THDPs). The TH transport system in particular may be quite susceptible to EDCs as many chemicals are structurally related to THs and may bind THDPs and disturb homeostasis of extracellular TH levels or even cellular uptake.[13]

Steroid hormone synthesis is controlled by the activity of several highly substrate-selective cytochrome P450 enzymes and a number of steroid dehydrogenases and reductases. Cytochrome P450 monooxygenases (CYPs) form a large group of enzymes found in most organisms from bacteria to mammals and can be grouped into 281 families. According to their function CYPs can be classified into enzymes metabolizing xenobiotics and enzymes that are part of key biosynthetic pathways, with narrow substrate specificity. Particularly, aromatase (CYP19), the enzyme that converts androgens to estrogens, has been the subject of studies into the mechanisms by which chemicals interfere with sex steroid hormone homeostasis and function, often related to (de)feminization and (de)masculinization processes.[14,15]

After all, several findings suggest that responses to EDCs cannot be assumed to be monotonic across a wide dose range and that change can occur in response to extremely low concentrations. In particular, low-dose effects may be mediated by endocrine-signaling pathways, evolved to act as powerful amplifiers, with the result that large changes can occur in response to extremely low concentrations. Dose–response relationship, however, is perhaps one of the most controversial issues in EDC studies. Reports on non-linearities in dose–response functions are highly controversial and the subject of intense research: non-monotonic, linear, and even threshold responses are all possible outcomes of low-dose exposure.[16] A non-monotonic response decreases testing efficiency and multiplies the time and other resources necessary to understand the potential hazard posed by a chemical. Because the issue of low-dose effects of EDCs was based on unknown and unexpected mechanisms, the actual features of these effects were not readily resolved.[17]

Low-dose effects of EDCs are based on unknown and unexpected mechanisms. Recent developments in the biological sciences, including homeostatic regulatory disturbance and epigenetic response, have aided in clarifying the mechanisms underlying the low-dose issue. Elucidating the xenobiotic effects of EDCs requires development of systems toxicology, i.e., deciphering the toxicity mechanisms underlying homeostatic regulatory disturbances.[17,18]

Hormones, Reproductive Aspects, and EDCs

In vertebrates, the ability to attain reproductive competence in adulthood involves the organization of a complex, steroid sensitive network in hypothalamic–preoptic–limbic brain regions during critical developmental windows. This process includes the establishment of the hypothalamic neural network of gonadotropin-releasing hormone (GnRH) cells, together with their regulatory inputs from other neuronal and glial cells in the brain, which enable feedback effects of steroid hormones on pulsatile GnRH release and the preovulatory GnRH/LH–luteinizing hormone surge in females. The anatomical development of this steroid-sensitive hypothalamic network occurs early in life, typically the late embryonic and early postnatal period in mammals, and its organization is important to the attainment and activation of appropriate reproductive functions in adulthood. Importantly, this same early developmental period is also a critical period for sexual differentiation of hypothalamic–limbic neural networks that must be organized perinatally to enable proper behavioral activation in adulthood. During mammalian development, the fetal organism is exposed to its own gonadal hormones, placental steroids, and maternal hormones that may cross the placental barrier. There are sex differences in exposures to androgens and estrogens that appear to underlie normal reproductive neuroendocrine development. Aberrations in these developmental patterns in females can cause masculinization (acquisition of a male-typical trait) or defeminization (loss of a female-typical trait) and in males may cause feminization or demasculinization (comparably defined). Perinatal hormones have permanent imprinting effects on the hypothalamus, manifested early on as morphological sex differences in the brain and manifested much later on as physiological and behavioral differences between the sexes.[19–21] Besides, androgens and estrogens can play a special role in the development of sexually dimorphic behaviors.[22]

Some populations exposed to chemicals from industrial accidents or chemical misadventures are of particular interest. Data from these select populations with higher levels of exposure than the common population seem to suggest that some of these chemicals have a role in genitourinary development as endocrine disrupters. Furthermore, animal studies of EDC effects on genitourinary development have confirmed that changes occur with exogenous manipulation of steroid levels or hormone receptors. These findings in animals have led to observational and epidemiological studies in humans to document a link between environmental exposure and human disease.[23]

Global declines in semen quality were suggested to be associated with enhanced exposure to environmental chemicals that act as endocrine disrupters as a result of increased use of pesticides, plastics, and other anthropogenic materials. A significant body of toxicology data based upon laboratory and wildlife animals studies suggests that exposure to certain endocrine disrupters is associated with reproductive toxicity, including the following: 1) abnormalities of the male reproductive tract (cryptorchidism, hypospadias); 2) reduced semen quality; and 3) impaired fertility in the adult.[24]

Recently, there has been increasing concern about the potential for environmental EDCs as fungicides to alter sexual differentiation in mammals. In this direction, observations demonstrate that vinclozolin (a systemic dicarboximide fungicide) can affect embryonic testicular cord formation in vitro. This transient in utero exposure to the fungicide increases apoptotic germ cell numbers in the testis of pubertal and adult animals. This effect is correlated with reduced sperm motility in the adult and putative effects on spermatogenic capacity later in adult life. In conclusion, transient exposure to this fungicide during the time of testis differentiation alters testis development and function.[25]

A higher prevalence of cryptorchidism and hypospadias was found in areas with extensive farming and pesticide use and in sons of women working as gardeners. Recently, a relation has been reported between cryptorchidism and persistent pesticide concentration in maternal breast milk.[26]

Other commonly used fungicides, such as the azoles, may also act as endocrine disrupters in vivo. They showed endocrine-disrupting potential when tested for endocrine-disruptive effects using a panel of in vitro assays. Overall, the imidazoles (econazole, ketoconazole, miconazole, prochloraz) were more potent than the triazoles (epoxiconazole, propiconazole, tebuconazole). The critical mechanism in vitro seems to be disturbance of steroid biosynthesis.[27]

Regarding in vivo effects, many of the commonly used azole fungicides act as endocrine disrupters, although the profile of action varies. Common features for azole fungicides are that they increase gestational length, virilize female pups, and affect steroid hormone levels in rat fetuses and/or dams.[28] For example, prochloraz causes reproductive malformations in androgen-dependent tissues of male offspring of exposed rats.[29] Also, tebuconazole has been found to demasculinize the male offspring and to possess some of the same endocrine effects as prochloraz. These effects strongly indicate that one major underlying mechanism for the endocrine-disrupting effects of azole fungicides is disturbance of key enzymes like CYP17 involved in the synthesis of steroid hormones.[28]

Also, triazole-induced male reproductive toxicity includes disruption of testosterone homeostasis. Elevated serum testosterone, increased testis, weights and anogenital distance, and hepatomegaly indicative of altered liver metabolism of steroids are the key events consistent with this mode of action.[30] Developmental exposure to triazole fungicides such as propiconazole, myclobutanil, and triadimefon can adversely impact reproduction in the female rat.[31] In this way, epoxiconazole and ketoconazole may be fetotoxic, increasing postimplantation loss and late resorptions.[32]

Aside from triazoles, the organic fungicide fenarimol possesses estrogenic properties[33] and acts both as an estrogen agonist and as an androgen antagonist.[34] In addition, fenarimol affects rat aromatase activity in vivo, inhibiting estrogen biosynthesis in rat microsomes[35] and in human tissues.[3]

This compound also affects other enzymes of the cytochrome P450 gene family that are involved in the metabolism of steroids.[36]

Induction of reactive oxygen species (ROS) by environmental contaminants and associated oxidative stress also have a role in defective sperm function and male infertility, although there are some controversial data. This is evidence for the existence of a link between endocrine-mediated and ROS-mediated adverse effects of environmental contaminants on male reproduction. Another link is the antioxidant enzyme superoxide dismutase, which has been shown to have a superoxide scavenging effect as well as act as an alternate regulatory switch in testicular steroidogenesis.[37]

Endocrine-disrupting chemicals can also impact female fertility by altering ovarian development and function, purportedly through estrogenic, antiestrogenic, and/or antiandrogenic effects. These compounds may also cause transgenerational effects by targeting oocyte maturation and maternal sex chromosomes.[38]

In girls, earlier age at menarche was reported after exposure to PCBs, polybrominated biphenyls, persistent pesticides (DDT), and phthalate esters. However, several other studies found no effect of these compounds on age at menarche. In boys, exposure to PCBs, PCDFs, or the pesticide endosulfan was associated with delayed puberty or decreased penile length. Much of the results found in population studies are in accordance with experimental studies in animals. However, the mixture of different components with antagonistic effects (estrogenic, antiestrogenic, antiandrogenic) and the limited knowledge about the most critical window for exposure (prenatal, perinatal, and pubertal) may hamper the interpretation of results.[39]

In human and rodent models, EDCs also interfere with the development of cognition and behaviors. In this way, fenvalerate is a potential EDC and is a candidate environmental risk factor for cognitive and behavioral development, especially in the critical period of development.[40] Also, prenatal phthalate exposure was associated with childhood social impairment in a multiethnic urban population.[41]

Recently, the interference of EDCs with receptors regulating metabolism has been proposed especially in relation with the etiology of metabolic diseases such as obesity and diabetes. In particular, the harmful action of EDCs on normal adipocyte development, homeostatic control of adipogenesis, early energy balance, and, in turn, body weight has been demonstrated. Much remains to be studied about the endocrine pathways responding to EDC exposure, especially those controlling feeding behavior, as their impairment represents a real risk factor for metabolic and feeding disorders.[42]

Screening Assays and Biomarkers for EDCs

Environmental stresses as presence of EDCs due to human activities are increasingly likely to pose habitat disturbances that could have potential deleterious effects on physiological function in vertebrates. These effects could result in major impacts on the life cycles of organisms, affecting morphology, physiology, and behavior. However, because animals live in diverse habitats, there is variation in susceptibility to disruption of response systems to environmental cues. While some populations of vertebrates, from fish to mammals, temporarily resist environmental stresses and breed successfully, many others show varying degrees of failure, sometimes resulting in population decline.[43] The development of targeted bioassays in combination with adequate chemical analyses is important for EDC risk assessment.

It is acknowledged that EDCs can affect humans and animals at low exposure levels and that responses to EDCs are in many cases complex, activating a range of different molecular events, e.g., receptor agonism/antagonism and enzyme induction, in multiple hormone systems. As a result, regulatory testing for these effects and evaluating the results is complicated.[44,45] In the typical case of assessing human risk, a scientifically justified validation could only mean an experimentally validated mechanistic link between EDC assay results and human susceptibilities at environmental exposures, sustained by reliable sensitivity and specificity benchmarks. Analytical methods have long been used to determine concentrations of chemical residues that persist in environment and accumulated in biota. Although they are useful, the development of EDC screening and monitoring procedures may help in the establishment of EDC exposure and biological responses. In this way, several in vitro and in vivo procedures have been proposed to screen and monitor individual EDCs or their mixtures.[45]

As an in vitro model, the use of the bovine ovarian follicle has already been recommended as a valuable instrument to unravel reproductive events in women due to the similarities in ovarian follicular dynamics and endocrine control.[46]

As an in vivo test, some external biomarkers of prenatal androgen disruption may be used, including the anogenital distance and the juvenile nipple/areola number. The anogenital distance is defined as the distance between the genital papilla and the anus; male rodents have an anogenital distance that is approximately twice the length as that of females. Areolae are dark areas surrounding the nipple bud and. their presence as measured at postnatal day 2–3 is indicative of adult nipples. Adult female rats typically have 12 nipples, whereas males have none. Both of these biomarkers vary with prenatal exposure to androgens or antiandrogens in females and males, respectively. Reduction of anogenital distance and/or retention of nipples in male rats is indicative of prenatal exposure to antiandrogens.[47]

In this context, a wide spectrum of potential biomarkers also could be applied to the study of endocrine disruption in the aquatic environment. In fish, they include changes in hormone titers (steroid hormones, thyroid hormones), abnormal gonad development, low gamete viability, and alterations in some enzyme activities (i.e., aromatases) and protein levels (i.e., vitellogenin, zona radiata proteins, spiggin). Likewise, evidence is slowly growing that indicates that gamete development and vitellogenesis of marine bivalve mollusks are targets of EDCs.[48]

On the other hand, although it is known that aquatic invertebrates contain different classes of steroids,[49] a clear cause–effect relationship between exposure and specific responses for most EDCs is far from being established.[48]

In crustacean populations, the attribution of endocrine toxicity to observed disturbances requires the identification of definitive biomarkers of such toxicity. Mortality, reduced fecundity, lowered recruitment, and impaired growth all might serve as indicators of endocrine disruption in crustaceans; however, such end points are indicative of adversity involving a variety of mechanisms. An exception to this premise is excess males in parthenogenic branchiopod populations that normally exist predominantly as females.[50]

Other organisms, such as amphibians, may be used to study the endocrine system and can serve as sentinels for detection of the modes of action of EDCs. Recently, amphibians are being reviewed as suitable models to assess (anti)estrogenic and (anti)androgenic modes of action influencing reproductive biology as well as (anti)thyroidal modes of action interfering with the thyroid system.[51]

Biochemical end points can also be useful biomarkers since environmental toxicants can trigger biological effects at the organism level only after initiating biochemical and cellular events. The cellular response to stress is characterized by the activation of genes involved in cell survival to counteract the physiological disturbance induced by physical or chemical agents. As an example, Hsps are suitable as an early-warning bioindicator of environmental hazard by various pollutants such as EDCs, because of their sensitivity to even minor changes in cellular homeostasis and their conservation along the evolutionary scale.[52] In addition, a combined testing strategy, considering both markers of endocrine/hormonal maturation and behavioral end points under hormonal control, may evidence even subtle perturbations of the neuroendocrine homeostasis, which often go undetected.[53]

Guidelines for Regulatory Purposes

In recent years, under the current European Union chemical regulation REACH (Registration, Evaluation, Authorization and Restriction of Chemicals), which revised plant protection product and biocide directives,[54] evaluation of endocrine-disrupting properties of chemicals has become a regulatory effort.

The initial framework for regulatory purposes has been revised by the Endocrine Disrupters Testing and Assessment (EDTA) Task Force at its meetings to reflect the Organization for Economic Cooperation and Development (OECD) member countries' views. The conceptual framework agreed upon by the EDTA6 in 2002 is not a testing scheme but rather a toolbox in which the various tests that can contribute information for the detection of the hazards of endocrine disruption are placed. The toolbox is organized into five compartments or levels, each corresponding to a different level of biological complexity (for both toxicological and ecotoxicological areas). Even though the conceptual framework may be full of testing tools, this does not imply that they all will be needed for assessment purposes. Tools will be added as they are validated in future. The conceptual framework is subject to further elaboration and discussion as the work on endocrine disrupters proceeds.[55] The OECD adopted in 2007 the uterotrophic bioassay as a standardized screening test with international regulatory acceptance. This assay may be used to screen for estrogenic properties of chemicals. However, generally, EDCs are handled as such only if their endocrine-disruption potential has been previously identified via, for example, academic research or is indicated by effects observed in required toxicity tests.[44]

The Endocrine Disruptor Screening Program (EDSP) of the United States Environmental Protection Agency (EPA) has been working to reach a consensus validation on a battery of screens and long-term tests for endocrine disrupters.[45] The Endocrine Disrupter Screening and Testing Advisory Committee (EDSTAC) was established by the EPA in 1996 as a federal advisory committee to provide advice in developing and implementing new screening and testing procedures for endocrine effects as mandated by the U.S. Congress (through the Food Quality Protection Act of 1996) in response to public concern.[56] The ED-STAC assesses the current state of the science and assists the agency in developing an endocrine screening program. The EDSTAC consists of scientists and others representing various interests, including advocates of the endocrine-disruption theory and the regulated community.

The EDSTAC concluded that the assays necessary to determine the potential endocrine activity of chemical substances varied significantly in their degree of development and validation. Several screens had an extensive history, e.g., the uterotrophic and the Hershberger screens, but others were only partially developed or were only hypothetically useful as screens, e.g., the amphibian developmental screen and the fish gonadal recrudescence screen. The fundamental validation principles are to clearly state the purpose and biological basis for the assay and to verify the performance of the assay against validation criteria using a common set of test chemicals across multiple laboratories.[57]

At the same time, EDSTAC recommended that EPA develop an extensive program that would subject all chemicals to screening and testing for estrogenic, antiestrogenic, androgenic, antiandrogenic, and thyroid effects in both humans and wildlife. Specifically, EDSTAC recommended, among other things, that the EPA do the following: 1) adopt a two-tiered, hierarchical testing and evaluation framework; and 2) initiate a research program, composed of both basic and applied research, to develop, standardize, and validate the necessary endocrine test methods. The EPA's EDSP was implemented in 2009–2010 with the issuance of test orders requiring manufacturers and registrants of 58 pesticide active ingredients and 9 pesticide inert/high-production-volume chemicals to evaluate the potential of these chemicals to interact with the estrogen, androgen, and thyroid hormone systems. Despite this great effort, numerous questions and uncertainties remain as to the usefulness and limitations of the specific assays. Understanding the tests' strengths and limitations is critical for interpretation of the screening results and for decision making based on those results.[57,58]

During the time EDSTAC was meeting, OECD began collaborating with its member countries, including the U.S. EPA, to develop internationally harmonized test guidelines.[48] Although the EPA and OECD endocrine screening and testing methods have been substantially harmonized, the framework of OECD's endocrine screening and testing program differs significantly from EPA's two-tiered EDSP. The EPA screening will entail evaluation of responses in the Tier 1 Endocrine Screening Battery, consisting of 11 distinct in vitro and in vivo assays. The OECD framework provides the flexibility to enter and exit at any level depending on information needs and encourages the maximal use of all existing relevant information that may be equally predictive and reduce vertebrate testing. The screening results are collectively intended to identify chemicals for which subsequent Tier 2 testing is necessary. Tier 2 testing uses test methods that encompass reproduction and developmental life stages in several species to provide data on adverse effects and dose response for risk assessment.[57]

In the years that the EPA worked on developing, standardizing, and validating the EDSTAC-recommended assays and implementing the EDSP, significant advances have been made in both computational and molecular technologies for discerning potential endocrine activity.[57] Accordingly, there are efforts to model EDC effects using computational approaches by the development and validation of mechanistically based computational models of hypothalamic–pituitary–thyroid (HPT); hypothalamic–pituitary–gonadal (HPG); hypothalamic–pituitary adrenal (HPA) axes in ecologically relevant species to better predict accommodation and recovery of endocrine systems.[59]

Overview of EDC Exposure

Human Exposure Effects

Models for estimating human exposure to endocrine disruptor (ED) pesticides are an important risk management tool. Many of them are harmful at very low doses, especially if exposure occurs during sensitive stages of development, producing effects that may not manifest for many years or that affect descendants via epigenetic changes. The main requirement for the use of such models is more quantitative data on the sources and pathways of human ED pesticide exposure. Quantifying the risks posed by the different routes of exposure will play an important part in designing and implementing effective risk mitigation for ED pesticides. In fact, it is difficult to assess the relative importance of some routes of exposure because no data sets that would allow these to be calculated are available. Pesticide exposure from the use of pesticides for medicinal purposes and exposure from cut flowers and ornamental plants both need to be quantified, and better data sets are required for pesticide exposure from spray drift, home use, municipal use, and travel.[60]

Food and water are both chronic exposure routes affecting the entire population. Food residues are currently thought to be the most important exposure pathway, for although residue levels present in food tend to be below the maximum residue levels permitted by law, they do result in constant measurable low-level exposure.[51] Food as a major xenobiotic and heavy metal exposure route to humans is studied intensively. More than 100 chemicals have been identified as antiandrogens, including certain phthalates, widely used as plasticizers, pesticides, and various other chemicals found in food and consumer products.[61]

Indeed, typical food contaminants, like pesticides, dioxins, PCBs, methylmercury, lead, etc., are well characterized in food. In contrast, the role of food and beverage packaging as an additional source of contaminants has received much less attention, even though food packaging contributes significantly to human xenobiotic exposure. Especially, EDCs in food packaging are of concern since even at low concentrations, chronic exposure is toxicologically relevant. Thus, non-intentionally added substances migrating from food contact materials need toxicological characterization.[62]

Some chemicals used in food processing have an environmental endocrine-disrupting effect that affects reproduction in wildlife. For example, bisphenol-A is a monomer of polycarbonate plastics and a constituent of epoxy and polystyrene resins, which are used in the food cans and found as a contaminant not only in the liquid of the preserved foods but also in the water autoclaved in the cans. This chemical is also released from polycarbonate flasks during autoclaving. Moreover, it has been reported that significant amounts of bisphenol-A are detected in the saliva of dental patients treated with fissure sealants. The exposure to low doses of this chemical was reported to affect the rate of growth and sexual maturation, hormone levels in blood, reproductive organ function, fertility, immune function, enzyme activity, brain structure, brain chemistry, and behavior.[63]

Another important route of EDC exposure is occupational. Relatively high levels of exposure to environmental endocrine disrupters in the form of pesticides occur among people working in agriculture. Some pesticides are able to influence the synthesis, storage, release, recognition, or binding of hormones, which may lead to alterations in reproductive hormone levels. The issue of male infertility caused by occupational exposure is pertinent worldwide. A significant increase in the incidence of male infertility has been described in the international literature. Part of this effect may result from synthetic toxic substances acting on the endocrine system, many of which are routinely used in work processes. However, progress is needed in the knowledge of possible effects of exposure on male fertility since monitoring these effects requires sufficient time for the manifestations to occur. Such progress will allow the development of preventive measure within the field of workers' health.[64]

Apart from EDC effects on males, several studies on occupational exposure to pesticides and adverse effects on human reproduction have been performed, including end points such as prolonged time to pregnancy, spontaneous abortion or stillbirth, low birth weight, and developmental disorders.[65]

Complex EDC Mixtures

Concerns increase when humans are exposed to mixtures of similar-acting EDCs and/or during sensitive windows of development. It is difficult to predict biological effects directly from the composition of pollutant mixtures. In addition to simple additive effects, interactions between different chemicals in a mixture may result in either a weaker (antagonistic) or a stronger (synergistic) combined effect than would be expected from knowledge about the toxicity and mode of action of each individual compound. Such interactions may take place in the toxicokinetic phase (i.e., processes of uptake, distribution, metabolism, and excretion) or in the toxicodynamic phase (i.e., effects of chemicals on the receptor, cellular target, or organ). A chemical mixture may contain a number of xenoestrogens enhancing the response of endogenous estrogens, or it may contain xenoantiestrogens that inhibit the normal action of endogenous estrogens.[66] Substances of concern include certain phthalates, pesticides and chemicals used in cosmetics and personal care products. A lack of knowledge about relevant exposure scenarios presents serious obstacles for better human risk assessment. A disregard for combination-effect studies may lead to underestimations of risks. In this way, the study of EDC mixture effects by developing biomarkers that capture cumulative exposure to endocrine disrupters is needed.[67]

Doses of endocrine-disrupting pesticides that appear to induce no effects on gestation length, parturition, and pup mortality when alone induced marked adverse effects on these end points together with other pesticides. They can also affect the sexual differentiation of offspring.[68] Chemicals that act on different fetal tissues via diverse cellular mechanism of action may produce additive effects. This fact indicates that the current framework for conducting cumulative risk assessments should not only consider including chemicals from different classes with the same mechanism of toxicity but also include chemicals that disrupt differentiation of the same fetal tissue at different sites in the androgen signaling pathway.[5] Compounds that act by disparate mechanisms of toxicity to disrupt the dynamic interactions among the interconnected signaling pathways in differentiating tissues produce cumulative dose-additive effects, regardless of the mechanism or mode of action of the individual mixture component.[69] Predictive approaches are generally based on the mathematical concepts of concentration addition and independent action, both predicting the toxicity of a mixture based on the individual toxicities of the mixture components.[26]

In this sense, a combination of five pesticides with dissimilar mechanisms of action produced greater androgen sensitive end-point responses than would be expected using response-addition modeling.[70] Deltamethrin, methiocarb, prochloraz, simazine, and tribenuron-methyl are all commonly used for agricultural and horticultural purposes. In vivo, the levator ani/bulbocavernosus muscle and adrenal gland weight changes indicated that the pesticides had an accumulating effect that was not observed for the individual pesticides. Several pesticide-induced gene expression changes were observed, indicating that these may be very sensitive antiandrogenic end points. In another study,[71] dexamethasone appeared to exacerbate the reproductive anomalies induced by in utero exposure of male rats to dibutyl phthalate.

In a recent study, male Sprague Dawley rats were sub-chronically exposed to single doses of dibutyl phthalate, single doses of benzo(a)pyrene, and combined doses of both EDCs. Significant adverse effects were observed on the reproductive system, including decreased sperm count, increased production of abnormal sperm, changes in serum testosterone levels, and irregular arrangements of the seminiferous epithelium. It is also observed that biochemical analyses showed that the activities of superoxide dismutase and glutathione peroxidase decreased after exposure to these EDCs. Therefore, the data suggest that exposure to them, in either separate or combined doses, can affect the reproductive system of male rats adversely via oxidative stress-related mechanisms.[72]

Thus, assessment of risks posed by chemicals causing reproductive effects and protection of future generations are important public health tasks. To determine the levels of significant human exposure to a given chemical and associated health effects, the Agency for Toxic Substances and Disease Registry's (ATSDR's) toxicological profiles examine and interpret available toxicological and epidemiological data. The ATSDR categorizes the health effects according to their seriousness as serious (effects that prevent the organism from functioning normally or can cause death), less serious (changes that will prevent an organ or organ system from functioning in a normal manner but will not necessarily lead to the inability of the whole organism to function normally), or minimal (effects that reduce the capacity of an organ or organ system to absorb additional toxic stress but will not necessarily lead to the inability of the organ or organ system to function normally). The ATSDR uses the highest no-observed-adverse-effect level or the lowest-observed-adverse-effect level (LOAEL) in the available literature to derive a health-based guidance value called a minimal risk level (MRL). An MRL is defined as "an estimate of the daily human exposure to a substance that is likely to be without an appreciable risk of adverse, non-cancer effects over a specified duration of exposure." Minimal risk levels based on reproductive and endocrine effects were described in a review by Pohl et al.[73]

Some Examples of Animal Exposure Effects

There is widespread exposure to EDCs, which can disrupt the reproduction and development of various non-target organisms. Effects of EDCs have been shown by observed adverse reproductive and developmental effects. Indeed, most studies of potential EDC effects are based on indirect evidence of endocrine disruption rather than defined endocrine pathways.

Some domestic mammals may come into contact with EDCs by sewage exposure. As an example, sewage sludge is sometimes recycled to arable land or pasture and contains large amounts of a variety of pollutants, including EDCs and heavy metals, derived from industrial, agricultural, and domestic sources. A demasculinizing effect of exposure to higher pollutant concentrations with respect to exploratory sheep behavior was observed.[74] These observations demonstrate the need to take into account the effects of pollutant combinations, even at very low, environmental concentrations, and further highlight the usefulness of ethotoxicology for the study of biological effects of environmental pollutants.

Endocrine-disrupting chemicals have been found in sewage effluent in low concentrations (ng/L). Some of these estrogens bind with estrogen receptors in exposed organisms and have the potential to exert effects at extremely low concentrations. Data from laboratory experiments support the hypothesis that EDCs in the aquatic environment can impact the reproductive health of various fish species, but evidence in the aquatic environment is still weak and needs a dependable method or indicator to assess reproduction of fish in situ. The link between endocrine disruption and reproductive impairment that cause an ecologically relevant impact on the sustainability of fish populations remains to be better understood.[75]

Surface waters are the main sink of EDCs, which are mainly of anthropogenic origin. Thus, aquatic organisms, especially lower vertebrates such as fish and amphibians, are the main potential targets for EDCs, being at direct or indirect risk via ingestion and accumulation of EDCs via exposure or the food chain. These compounds may play an important role in the decline of the amphibian population.[51] Several incidents in the wildlife population strongly correlated decreased reproductive capacity with exposure to specific industrial chemicals, and the organisms may be viewed as sentinels of human health effects. Reported reproductive disorders in wildlife have included morphologic abnormalities, eggshell thinning, population declines, sex reversal, impaired viability of offspring, altered hormone concentration, and changes in sociosexual behavior.[76]

The ED are prevalent over a wide range of chemicals in the aquatic ecosystems, most of them being resistant to environmental degradation and considered ubiquitous contaminants.[48,77] Some imidazole (prochloraz, imazalil) and triazole (epoxiconazole) agricultural fungicides induced oocyte maturation in rainbow trout. Prochloraz, epoxiconazole, and imazalil strongly potentiated the induction of oocyte maturation by gonadotropin in a dose-dependent manner.[78] Above all, prochloraz caused responses consistent with aromatase inhibition, although there were indications that the fungicide may also be disturbing the balance between estrogens and androgens via effects elsewhere in the steroidogenic pathway.[77]

In U.K. rivers, a widespread feminization of wild fish was observed involving contributions from both steroidal estrogens and xenoestrogens and from other yet-unknown contaminants with antiandrogenic properties. The widespread occurrence of feminized male fish downstream of some wastewater treatment works has led to substantial interest from ecologists and public health professionals. This concern stems from the view that the effects observed have a parallel in humans and that both phenomena are caused by exposure to mixtures of contaminants that interfere with reproductive development.[79] Some authors reported the occurrence of fish feminization as well as reproduction and development interference with other aquatic organisms,[80] although there is no universally accepted bioassay or chemical technique to quantify EDCs in the aquatic environment.[81] Endocrine-disrupting chemicals can also promote disrupting effects in vitro on ovarian follicular cells exposed to environmentally relevant doses of mixtures of persistent organic pollutants extracted from marine and freshwater ecosystems.[66]

Population studies have revealed alterations in crustacean growth, molting, sexual development, and recruitment that are indicative of environmental endocrine disruption. However, environmental factors other than pollution (i.e., temperature, parasitism) also can elicit these effects and definitive causal relationships between endocrine disruption in crustacean field populations, and chemical pollution is generally lacking.[50] Also, temperature and photoperiod are the two most important environmental cues in the regulation of the annual cycles of circulating sex steroid hormones and reproduction in fish. Thus, these variables may alter the endocrine-disruption effects induced by EDCs.[82]

In contrast to mammals and birds, the mechanisms underlying sex determination and differentiation in fishes vary widely and are changeable or labile in response to environmental parameters. These environmental parameters include temperature, behavioral cues and demographic structure of the local population, and EDCs. Understanding the gender similarities and differences in how organisms respond following exposure to environmental chemicals is important to determine the relative risk of these agents to wildlife and human populations. Given the central role of sex steroid hormones in the sex determination and sexual differentiation of fishes, amphibians, and reptiles, future research that includes sex as a factor is recommended. Thus, the risk assessment can address the probable gender differences in effects from exposure to chemicals in the environment.[83]

Dealing with Environment EDC Emission

Municipal wastewater contains a complex mixture of EDCs originating from different sources. A number of organic pollutants, such as polycyclic aromatic hydrocarbons, PCBs, and pesticides, are resistant to degradation and represent an ongoing toxicological threat to both wildlife and human beings. Furthermore, recently, wastewater sludge has been subjected to reuse for production of value-added products. These facts have heightened the need for novel and advanced bioremediation techniques to effectively remove EDCs from a variety of contaminated environmental media including water, wastewater sludge, sediments, and soils. One possibility to solve this problem is the use of microbial potential to degrade or detoxify EDCs and other toxic intermediates.[80,84]

Also, there are physical methods such as absorption by activated carbon and rejection by membranes to remove EDCs. However, pollutant removal from wastewater is a process with high energy consumption, where cost and efficiency are the key considerations for their application. Biodegradation processes have proven to be the most cost-effective.[85]

Water companies became aware of the endocrine-disruption problem when a survey confirmed the observation by anglers of hermaphrodite fish in wastewater treatment plant lagoons after being exposed to significant levels of persistent man-made chemicals. The evolving regulatory context related to micropollutants in the environment may have a decisive impact on wastewater management and requires an increased knowledge of the fate of micropollutants during wastewater treatment. Advanced treatments such as oxidation (ozone) are known to be able to enhance the removal of micropollutants, but technical, economic, and environmental risk/benefit evaluations must be performed before implementing such additional processes. In any case, the reduction of the pollution at the source, i.e., upstream of the wastewater treatment plant, represents the most sensible option, which should be promoted.[86]

Although numerous studies have investigated degradation of individual EDCs in laboratory or natural waters, chemical-based analytical methods cannot represent the combined or synergistic activities between water quality parameters and/or the EDC mixtures at environmentally relevant concentrations since natural variations in water matrices and mixtures of EDC in the environment may confound analysis of the treatment efficiency. In conjunction with standard analytical approaches, bioanalytical assessments of residual estrogenic activity in treated water will enable estimates of the interactions and/or combined estrogenic activity among mixtures of EDCs and the water matrix in natural water.[87]

By contrast, the agricultural sector, a significant user of veterinary pharmaceuticals, has no such treatment—compounds are deposited straight to the ground in dung and urine or washed from hides in the case of topical applications. There has been little research as to whether any of these compounds leach into and persist in local soil and aquatic ecosystems. The extent to which the active ingredients of any of these chemicals (and their metabolites) leach into pastures, soil, runoff, and groundwater is a matter for field research. Also, much spraying of pesticides as herbicides and insecticides is done by ground crews. In such circumstances, it is not known whether they react with each other as well as pesticides and herbicides, forming further compounds which, either acting individually or in combination, could adversely affect bacteria, fungi, and higher organisms.[88]

Above all, a ranking system that could be customized for specific geographical locations will aid public policies in prioritizing EDCs that need monitoring and removal of aquatic sources as drinking water.[89] The establishment of simple but integrative screening assays for regulatory purposes is allowed by a strong correlation between xenoestrogen exposure and reproductive impairment. In fact, molecular screening assays could contain a battery of molecular targets allowing a more comprehensive approach in the identification of endocrine-disrupting compounds in fish and vertebrates in general.[90]

Different assays can be successfully employed as a battery of assays to screen environmental water samples for estrogenicity. The results obtained from this battery of assays should be interpreted as a first-tier screen for estrogenicity. Samples that test positive should be further investigated using second-and third-tier screens with routine sampling in order to monitor rivers for estrogenicity.[91] Complementarily, a fugacity-based model may be applied to simulate the distribution of EDCs in reservoirs of recycled waste-water,[92] or a fugacity-hydrodynamic model may be used for predicting the concentrations of the organic pollutants in surface water.[93]

Conclusion

Endocrine-disrupting chemicals can cause a wide range of reproductive damage and developmental, growth, immune, and behavior effects even in low doses and by different mechanisms of action. They encompass a variety of chemical classes, including hormones, plant constituents, pesticides, compounds used in the plastics industry and in consumer products, and other industrial by-products and pollutants. Some of them are widely dispersed in the environment. Exposure to EDCs can occur through direct contact with these chemicals or through ingestion of contaminated water, sediment, air, soil, and food and consumer products.

In humans, it is difficult to definitively link a particular EDC with a specific effect because the studies have inconclusive results. However, fetuses and embryos, whose growth and development are highly controlled by the endocrine system, are more vulnerable to exposure and may suffer reproductive abnormalities. The timing of exposure is also presumed to be critical, since different hormone pathways are active during different stages of development. Perinatal exposure, in some cases, can lead to permanent alterations that may be overt in adulthood.

Compared with humans, the evidence that wildlife has been affected adversely by exposures to EDCs is extensive. Available evidence seems to indicate that endocrine disruption caused by xenobiotics is primarily an ecotoxicologic problem. These chemicals may be extremely challenging for aquatic organisms and mammals that have a large habitat and that consume fish from many different areas throughout their lives. Low concentrations of endocrine disrupters can have synergistic effects in various organisms as amphibians. For removal of these compounds from aquatic sources, the most cost-effective process is biodegradation.

In spite of the need to manage the environmental, human health, and economic impacts of EDCs, most attention is focused on pharmaceutically active chemicals instead of those for agricultural use. The impact of these latter compounds is understudied.

The legal approach has been improved by new test protocols. Progress has been made in the identification and quantification of a wide array of chemicals with endocrine-active properties, especially those that persist and bioaccumulate in organisms and their environment. Studies with mammals have shown that exposure to endocrine-active compounds during early development may result in adverse health impacts that are not realized until adulthood. However, from a regulatory perspective, the ability of animals to recover from chemical insults is problematic because it complicates efforts to establish acceptable levels of exposure. Consequently, research to define the limits and biological cost recovery using standardized test designs is needed.[94]

However, exposure complexities, including transient and low-concentration exposure to EDCs, maternal metabolism of bioaccumulated EDCs, varying vulnerability and response by developmental stage, poorly understood exposure sources, mixtures and synergies, and cultural, social, and economic patterns, make it difficult for science to make solid exposure determinations. While there has been a great deal of research and effort in context with the hazard assessment and regulation of EDCs, there are also remaining uncertainties and issues. These include animal rights concerns due to significant increases in the use of animals to fulfill testing requirements; associated needs for alternative testing concepts such as *in vitro, in silico,* and modeling approaches; and lack of understanding of the relevance of exposure of humans and wildlife to EDCs.[95] Given the dynamic nature of the endocrine system, future efforts in the study of EDCs need more focus on the timing, frequency, and duration of exposure to these chemicals.

References

1. Kretschmer, X.C.; Baldwin, W.S. CAR and PXR: Xenosensors of endocrine disrupters? Chem.-Biol. Interact. **2005**, *155*, 111–128.
2. Castro, V.; Mello, M.A.; Diniz, C.; Morita, L.; Zucchi, T.; Poli, P. Neurodevelopmental effects of perinatal fenarimol exposure on rats. Reprod. Toxicol. **2007**, *23*, 98–105.
3. Vinggaard, A.; Jacobsen, H.; Metzdorff, S.; Andersen, H.; Nellemann, C. Antiandrogenic effects in short-term in vivo studies of the fungicide fenarimol. Toxicology **2005**, *207*, 21–34.
4. Waissmann, W. Health surveillance and endocrine disruptors. Cadernos de Saude Publica **2002**, *18* (2), 511–517.

5. Hotchkiss, A.K.; Rider, C.V.; Blystone, C.R.; Wilson, V.S.; Hartig, P.C.; Ankley, G.T.; Foster, P.M.; Gray, C.L.; Gray, L.E. Fifteen years after "wingspread"—Environmental endocrine disrupters and human and wildlife health: Where we are today and where we need to go. Toxicol. Sci. **2008**, *105* (2), 235–259.
6. Koopman, P. The delicate balance between male and female sex determining pathways: Potential for disruption of early steps in sexual development. Int. J. Androl. **2010**, *33*, 252–258.
7. Walker, D.M.; Gore, A.C. Transgenerational neuroendocrine disruption of reproduction. Nat. Rev. Endocrinol. **2011,** *7* (4), 197–207.
8. Bulzomi, P.; Marino, M. Environmental endocrine disruptors: Does a sex-related susceptibility exist? Front. Biosci. **2011,** *16* (7), 2478–2498.
9. Andersen, M.; Dennison, J. Mechanistic approaches for mixture risk assessments—Present capabilities with simple mixtures and future directions. Environ. Toxicol. Pharmacol. **2004**, *16,* 1–11.
10. Porcher, J.; Devillers, J.; Marchand-Geneste, N. Mechanism of endocrine disruptions—A tentative overview. In *Endocrine Disruption Modeling;* Devillers, J., Ed.; CRC Press: Boca Raton, 2009; 11–46.
11. Flynn, K. Dietary exposure to endocrine-active pesticides: Conflicting opinions in a European workshop. Environ. Int. **2011,** *37* (5), 980–990.
12. Silva, E.; Kabil, A.; Kortenkamp, A. Cross-talk between non-genomic and genomic signaling pathways—Distinct effect profiles of environmental estrogens. Toxicol. Appl. Pharmacol. **2010**, *245,* 160–170.
13. Morgado, I.; Campinho, M.A.; Costa, R.; Jacinto, R.; Power, D.M. Disruption of the thyroid system by diethylstilbestrol and ioxynil in the sea bream (*Sparus aurata*). Aquat. Toxicol. **2009**, *92,* 271–280.
14. Trosken, E.R.; Adamskab, M.; Arand, M.; Zarn, J.A.; Patten, C.; Volkel, W.; Lutz, W.K. Comparison of lanosterol-14α-demethylase (CYP51) of human and *Candida albicans* for inhibition by different antifungal azoles. Toxicology **2006**, *228,* 24–32.
15. Sanderson, J.T. The steroid hormone biosynthesis pathway as a target for endocrine-disrupting chemicals. Toxicol. Sci. **2006**, *94* (1), 3–21.
16. Belloni, V.; Dessì-Fulgheri, F.; Zaccaroni, M;. Di Consiglio, E.; De Angelis, G.; Testai, E.; Santochirico, M.; All-eva, E.; Santucci, D. Early exposure to low doses of atrazine affects behavior in juvenile and adult CD1 mice. Toxicology **2011,** *279,* 19–26.
17. Hirabayashi, Y.; Inoue, T. The low-dose issue and stochastic responses to endocrine disruptors. J. Appl. Toxicol. **2011**, *31,* 84–88.
18. Calabrese, E.J. Hormesis: A revolution in toxicology, risk assessment and medicine—Re-framing the dose–response relationship. EMBO Rep. **2004**, *5*, S37–S40.
19. Gore, A.C. Developmental programming and endocrine disruptor effects on reproductive neuroendocrine systems. Front. Neuroendocrinol. **2008,** *29,* 358–374.
20. Anway, M.D.; Skinner, M.K. Epigenetic transgenerational actions of endocrine disruptors. Endocrinology **2006**, *147* (6), S43–S49.
21. Crews, D.; McLachlan, J.A. Epigenetics, evolution, endocrine disruption, health, and disease. Endocrinology **2006**, *147* (6), S4–S10.
22. Li, A.A.; Baum, M.J.; McIntosh, L.J.; Day, M.; Liu, F.; Gray L.E., Jr. Building a scientific framework for studying hormonal effects on behavior and on the development of the sexually dimorphic nervous system. Neurotoxicology **2008,** *29,* 504–519.
23. Yiee, J.H.; Baskin, L.S. Environmental factors in genitourinary development. J. Urol. **2010**, *184,* 34–41.
24. Phillips, K.P.; Tanphaichitr, N. Human exposure to endocrine disrupters and semen quality. J. Toxicol. Environ. Health, Part B **2008**, *11* (3–4), 188–220.

25. Uzumcu, M.; Suzuki, H.; Skinner, M.K. Effect of the anti-androgenic endocrine disruptor vinclozolin on embryonic testis cord formation and postnatal testis development and function. Reprod. Toxicol. **2004**, *18* (6), 765–774.
26. Mnif, W.; Hassine, A.I.H.; Bouaziz, A.; Bartegi, A.; Thomas, O.; Roig, B. Effect of endocrine disruptor pesticides: A review. Int. J. Environ. Res. Pub. Health **2011,** *8* (6), 2265–2303.
27. Kjærstad, M.B.; Taxvig, C.; Nellemann, C.; Vinggaard, A.M.; Andersen, H.R. Endocrine disrupting effects in vitro of conazole antifungals used as pesticides and pharmaceuticals. Reprod. Toxicol. **2010,** *30,* 573–582.
28. Taxvig, C.; Hass, U.; Axelstad, M.; Dalgaard, M.; Boberg, J.; Andersen, H.R.; Vinggaard, A.M. Endocrine-disrupting activities in vivo of the fungicides tebuconazole and epoxiconazole. Toxicol. Sci. **2007**, *100* (2), 464–473.
29. Noriega, N.C.; Ostby, J.; Lambright, C.; Wilson, V.S.; Gray, L.E., Jr. Late gestational exposure to the fungicide prochloraz delays the onset of parturition and causes reproductive malformations in male but not female rat offspring. Biol. Reprod. **2005**, *72*, 1324–1335.
30. Goetz, A.K.; Ren, H.; Schmid, J.E.; Blystone, C.R.; Thillai-nadarajah, I.; Best, D.S.; Nichols, H.P.; Strader, L.F.; Wolf, D.C.; Narotsky, M.G.; Rockett, J.C.; Dix, D.J. Disruption of testosterone homeostasis as a mode of action for the reproductive toxicity of triazole fungicides in the male rat. Toxicol. Sci. **2007**, *95* (1), 227–239.
31. Rockett, J.C.; Narotsky, M.G.; Thompsona, K.E.; Thillai-nadarajah, I.; Blystone, C.R.; Goetz, A.K.; Rena, H.; Best, D.S.; Murrell, R.N.; Nichols, H.P.; Schmid, J.E.; Wolf, D.C.; Dix, D.J. Effect of conazole fungicides on reproductive development in the female rat. Reprod. Toxicol. **2006**, *22*, 647–658.
32. Taxvig, C.; Vinggaard, A.M.; Hass, U.; Axelstad, M.; Metzdorff S.; Nellemann, C. Endocrine-disrupting properties in vivo of widely used azole fungicides. Int. J. Androl. **2008**, *31,* 170–177.
33. Grünfeld, H.; Bonefeld-Jørgensen, E. Effect of in vitro estrogenic pesticides on human oestrogen receptor α and β mRNA levels. Toxicol. Lett. **2004**, *151* (3), 467–480.
34. Andersen, H.; Vinggaard, A.; Rasmussen, T.; Gjermandsen, I.; Bonefeld-Jørgensen, E. Effects of currently used pesticides in assays for estrogenicity, androgenicity, and aromatase activity in vitro. Toxicol. Appl. Pharmacol. **2002**, *179* (1), 1–12.
35. Hirsch, K.; Weaver, D.; Black, L.; Falcone, J.; Maclusky, N. Inhibition of central nervous system aromatase activity, a mechanism for fenarimol-induced infertility in the male rats. Toxicol. Appl. Pharmacol. **1987**, *91,* 235–245.
36. Paolini, M.; Mesirca, R.; Pozzetti, L.; Sapone, A.; Can-telli-Forti, G. Molecular non-genetic biomarkers related to fenarimol cocarcinogenesis, organ-and sex-specific CYP induction in rat. Cancer Lett. **1986**, *101,* 171–178.
37. Saradha, B.; Mathur P.P. Effect of environmental contaminants on male reproduction. Environ. Toxicol. Pharmacol. **2006**, *21,* 34–41.
38. Uzumcu, M.; Zachow, R. Developmental exposure to environmental endocrine disruptors: Consequences within the ovary and on female reproductive function. Reprod. Toxicol. **2007**, *23,* 337–352.
39. Hond, E.; Schoeters, G. Endocrine disrupters and human puberty. Int. J. Androl. **2006**, *29,* 264–271.
40. Meng, X.; Liu, P.; Wang, H.; Zhao, X.; Xu, Z.; Chen G.; Xu, D. Gender-specific impairments on cognitive and behavioral development in mice exposed to fenvalerate during puberty. Toxicol. Lett. **2011**, *203* (3), 245–251.
41. Miodovnik, A.; Engel, S.M.; Zhu, C.; Ye, X.; Soorya, L.V.; Silva, M.J.; Calafat, A.M.; Wolff, M.S. Endocrine disruptors and childhood social impairment. Neurotoxicology **2011**, *32* (2), 261–267.
42. Migliarini, B.; Piccinetti, C.C.; Martella, A.; Maradonna, F,; Gioacchini, G.; Carnevali, O. Perspectives on endocrine disruptor effects on metabolic sensors. Gen. Comp. Endocrinol. **2011**, *170,* 416–423.

43. Wingfield, J.C.; Mukai, M. Endocrine disruption in the context of life cycles: Perception and transduction of environmental cues. Gen. Comp. Endocrinol. **2009**, *163,* 92–96.
44. Beronius, A.; Rudén, C.; Hanberg, A.; Håkansson, H. Health risk assessment procedures for endocrine disrupting compounds within different regulatory frameworks in the European Union. Regul. Toxicol. Pharmacol. **2009**, *55*, 111–122.
45. Gori, G. Regulating endocrine disruptors. Regul. Toxicol. Pharmacol. **2007**, *48,* 1–3.
46. Campbell, B.K.; Souza, C., Gong, J.; Webb, R., Kendall, N.; Marsters, P., Robinson, G.; Mitchell, A.; Telfer, E.E.; Baird, D.T. Domestic ruminants as models for the elucidation of the mechanisms controlling ovarian follicle development in humans. Reprod. Suppl. **2003**, *61,* 429–43.
47. Wilson, V.S.; Blystone, C.R.; Hotchkiss, A.K.; Rider, C.V.; Gray, L.E., Jr. Diverse mechanisms of anti-androgen action: impact on male rat reproductive tract development. Int. J. Androl. **2008**, *31,* 178–187.
48. Porte, C.; Janer, G.; Lorusso, L.C.; Ortiz-Zarragoitia, M.; Cajaraville, M.P.; Fossi, M.C.; Canesi, L. Endocrine disruptors in marine organisms: Approaches and perspectives. Comp. Biochem. Physiol., Part C **2006**, *143,* 303–315.
49. Lafont, R.; Mathieu, M. Steroids in aquatic invertebrates. Ecotoxicology **2007**, *16,* 109–130.
50. LeBlanc, G.A. Crustacean endocrine toxicology: A review. Ecotoxicology **2007**, *16,* 61–81.
51. Kloas, W.; Lutz, I. Amphibians as model to study endocrine disrupters. J. Chromatogr. A **2006**, *1130,* 16–27.
52. Morales, M.; Planelló, R.; Martfnez-Paz, P.; Herrero, O.; Cortés, E.; Martfnez-Guitarte, J.L.; Morcillo, G. Characterization of Hsp70 gene in *Chironomus riparius:* Expression in response to endocrine disrupting pollutants as a marker of ecotoxicological stress. Comp. Biochem. Physiol. **2011**, *153,* 150–158.
53. Calamandrei, G.; Maranghi, F.; Venerosi, A.; Alleva, E.; Mantovani, A Efficient testing strategies for evaluation of xenobiotics with neuroendocrine activity. Reprod. Toxicol. **2006**, *22*, 164–174.
54. Danga, Z.C.; Ru, S.; Wang, W.; Rorije, E.; Hakkert, B.; Vermeire, T. Comparison of chemical-induced transcriptional activation of fish and human estrogen receptors: Regulatory implications. Toxicol. Lett. **2011**, doi:10.1016/j.toxlet.2010.12.020.
55. OECO Conceptual Framework for the testing and Assessment of Endocrine Disrupting Chemicals, available at http://www.oecd.org/document/580.3343,en_2649_34377_2348794_1_1_1_1.00.html (accessed May 2011).
56. Juberg, D.R. An evaluation of endocrine modulators: Implications for human health. Ecotoxicol. Environ. Safety **2000**, *45,* 93–105.
57. Borgert, C.J.; Mihaich, E.M.; Quill, T.F.; Marty, M.S.; Levine, S.L.; Becker, R.A. Evaluation of EPA's Tier 1 Endocrine Screening Battery and recommendations for improving the interpretation of screening results. Regul. Toxicol. Pharmacol. **2011**, *59,* 397–411.
58. Vogel, J.M. Perils of paradigm: Complexity, policy design, and the Endocrine Disruptor Screening Program. *Environ. Health: Global Access Sci. Source* **2005**, *4*, 2, doi:10.1186/1476-069X-4-2.
59. Nichols, J.W.; Breen, M.; Denver, R.J.; Distefano, J.J.; Edwards, J.S.; Hoke, R.A.; Volz, D.C.; Zhang, X. Predicting chemical impacts on vertebrate endocrine systems. Environ. Toxicol. Chem. **2011**, *30* (1), 39–51.
60. McKinlay, R.; Plant, J.A.; Bell, J.N.B.; Voulvoulis, N. Calculating human exposure to endocrine disrupting pesticides via agricultural and non-agricultural exposure routes. Sci. Total Environ. **2008**, *398*, 1–12.
61. Kortenkamp A.; Faust M. Combined exposures to anti-androgenic chemicals: Steps towards cumulative risk assessment. Int. J. Androl. **2010**, *33*, 463–474.
62. Muncke, J. Exposure to endocrine disrupting compounds via the food chain: Is packaging a relevant source? Sci. Total Environ. **2009**, *407,* 4549–4559.
63. Narita, M.; Miyagawa, K.; Mizuo, K.; Yoshida, T.; Suzuki, T. Prenatal and neonatal exposure to low-dose of bisphenol-A enhance the morphine-induced hyperlocomotion and rewarding effect. Neurosci. Lett. **2006**, *402*, 249–252.

64. Queiroz, E.K.R.; Waissmann, W. Occupational exposure and effects on the male reproductive system. Cadernos de Saúde Pública **2006**, *22* (3), 485–493.
65. Bretveld, R.W.; Hooiveld, M.; Zielhuis, G.A.; Pellegrino, A.; van Rooij, I.A.L.M.; Roeleveld, N. Reproductive disorders among male and female greenhouse workers. Reprod. Toxicol. **2008**, *25*, 107–114.
66. Gregoraszczuk, E.L.; Milczarek, K.; Wojtowicz, A.K.; Berg, V.; Skaare, J.U.; Ropstad, E. Steroid secretion following exposure of ovarian follicular cells to three different natural mixtures of persistent organic pollutants (POPs). Reprod. Toxicol. **2008**, *25,* 58–66.
67. Kortenkamp A. Low dose mixture effects of endocrine disrupters: implications for risk assessment and epidemiology. Int. J. Androl. **2008**, *31*, 233–240.
68. Jacobsen, P.R.; Christiansen, S.; Boberg, J.; Nellemann C.; Hass, U. Combined exposure to endocrine disrupting pesticides impairs parturition, causes pup mortality and affects sexual differentiation in rats. Int. J. Androl. **2010**, *33*, 434–442.
69. Rider, C.V.; Furr, J.R.; Wilson, V.S.; Gray, L.E., Jr. Cumulative effects of in utero administration of mixtures of reproductive toxicants that disrupt common target tissues via diverse mechanisms of toxicity. Int. J. Androl. **2010**, *33*, 443–462.
70. Birkhoj, M.; Nellemann, C.; Jarfelt, K.; Jacobsen, H.; Andersen, H.R.; Dalgaard, M; Vinggaard, A.M. The combined antiandrogenic effects of five commonly used pesticides. Toxicol. Appl. Pharmacol. **2004**, *201,* 10–20.
71. Drake, A.J.; van den Driesche, S.; Scott, H.M., Hutchison, G.R.; Seckl, J.R.; Sharpe, R.M. Glucocorticoids amplify dibutyl phthalate-induced disruption of testosterone production and male reproductive development. Endocrinology **2009**, *150,* 5055–5064.
72. Chen, X.; An, H.; Ao, L.; Sun, L.; Liu, W.; Zhou, Z.; Wanga, Y.; Cao, J. The combined toxicity of dibutyl phthalate and benzo(a)pyrene on the reproductive system of male Sprague Dawley rats in vivo. J. *Hazard.* Mater. **2011**, *186* (1), 835–841.
73. Pohl, H.R.; Luukinen, B.; Holler, J.S. Health effects classification and its role in the derivation of minimal risk levels: Reproductive and endocrine effects. Regul. Toxicol. Pharmacol. **2005**, *42,* 209–217.
74. Erhard, H.W.; Rhind, S.M. Prenatal and postnatal exposure to environmental pollutants in sewage sludge alters emotional reactivity and exploratory behaviour in sheep. Sci. Total Environ. **2004**, *332,* 101– 108.
75. Mills, L.J.; Chichester, C. Review of evidence: Are endocrine-disrupting chemicals in the aquatic environment impacting fish populations? Sci. Total Environ. **2005**, *343,* 1– 34.
76. Fox, G.A. Wildlife as sentinels of human health effects in the Great Lakes-St. Lawrence Basin. Environ. Health Perspect. **2001**, *109* (Suppl. 6), 853–861.
77. Kinnberg, K.; Holbech, H.; Petersen, G.I.; Bjerregaard, P. Effects of the fungicide prochloraz on the sexual development of zebrafish *(Danio rerio).* Comp. Biochem. Physiol., Part C **2007**, *145,* 165–170.
78. Monod, G.; Rime, H.; Bobe, J.; Jalabert, B. Agonistic effect of imidazole and triazole fungicides on in vitro oocyte maturation in rainbow trout *(Oncorhynchus mykiss).* Mar. Environ. Res. **2004**, *58,* 143–146.
79. Jobling, S.; Burn, R.W.; Thorpe, K.; Williams, R.; Tyler, C. Statistical modeling suggests that antiandrogens in effluents from wastewater treatment works contribute to widespread sexual disruption in fish living in English rivers. Environ. Health Perspect. **2009**, *117* (5), 797–802.
80. Barnabé, S.; Brar, S.K.; Tyagi, R.D.; Beauchesne, I.; Surampalli, R.Y. Pre-treatment and bioconversion of wastewater sludge to value-added products—Fate of endocrine disrupting compounds. Sci. Total Environ. **2009**, *407,* 1471–1488.
81. Nelson, J.; Bishay, F.; van Roodselaar, A.; Ikonomou, M.; Law, F.C.P. The use of in vitro bioassays to quantify endocrine disrupting chemicals in municipal wastewater treatment plant effluents. Sci. Total Environ. **2007**, *374,* 80–90.

82. Jin, Y.; Shu, L.; Huang, F.; Cao, L.; Sun, L.; Fu, Z. Environmental cues influence EDC-mediated endocrine disruption effects in different developmental stages of Japanese medaka (*Oryzias latipes*). Aquat. Toxicol. **2011**, *101*, 254–260.
83. Orlando, E.F.; Guillette, L.J., Jr. Sexual dimorphic responses in wildlife exposed to endocrine disrupting chemicals. Environ. Res. **2007**, *104,* 163–173.
84. Robinson, B.J.; Hellou, J. Biodegradation of endocrine disrupting compounds in harbour seawater and sediments. Sci. Total Environ. **2009**, *407,* 5713–5718.
85. Liu, Z.; Kanjo, Y.; Mizutani, S. Removal mechanisms for endocrine disrupting compounds (EDCs) in wastewater treatment—Physical means, biodegradation, and chemical advanced oxidation: A review. Sci. Total Environ. **2009**, *407,* 731–748.
86. Janex-Habibi, M.; Huyard, A.; Esperanza, M.; Bruchet, A. Reduction of endocrine disruptor emissions in the environment: The benefit of wastewater treatment. Water Res. **2009,** *43,* 1565–1576.
87. Chen, P.; Rosenfeldt, E.J.; Kullman, S.W.; Hinton, D.E.; Linden, K.G. Biological assessments of a mixture of endocrine disruptors at environmentally relevant concentrations in water following UV/H_2O_2 oxidation. Sci. Total Environ. **2007**, *376,* 18–26.
88. Fisher, P.M.J.; Scott, R. Evaluating and controlling pharmaceutical emissions from dairy farms: A critical first step in developing a preventative management approach. J. Cleaner Prod. **2008**, *16,* 1437–1446.
89. Kumar, A.; Xagoraraki, I. Pharmaceuticals, personal care products and endocrine-disrupting chemicals in U.S. surface and finished drinking waters: A proposed ranking system. Sci. Total Environ. 2010, *in press*
90. Scholz, S.; Mayer, I. Molecular biomarkers of endocrine disruption in small model fish. Mol. Cell. Endocrinol. **2008**, *293*, 57–70.
91. Swart, J.C.; Pool, E.J.; van Wykb, J.H. The implementation of a battery of in vivo and in vitro bioassays to assess river water for estrogenic endocrine disrupting chemicals. Eco-toxicol. Environ. Saf. **2011**, *74,* 138–143.
92. Cao, Q.; Yu, Q.; Connell, D.W. Fate simulation and risk assessment of endocrine disrupting chemicals in a reservoir receiving recycled wastewater. Sci. Total Environ. **2010**, *408*, 6243–6250.
93. Zhang, Y.; Song, X.; Kondoh, A.; Xia, J.; Tang, C. Behavior, mass inventories and modeling evaluation of xenobiotic endocrine-disrupting chemicals along an urban receiving wastewater river in Henan Province, China. Water Res. **2011**, *45,* 292–302.
94. Nichols, J.W.; Breen, M.; Denver, R.J.; Distefano, J.J., III; Edwards, J.S.; Hoke, R.A.; Volz, D.C.; Zhang X. Predicting impacts on vertebrate endocrine systems. Environ. Toxicol. Chem. **2011**, *30* (1), 39–51.
95. Hecker, M.; Hollert, H. Endocrine disruptor screening: Regulatory perspectives and needs. Environ. Sci. Eur. **2011**, *23*, 15, doi:10.1186/2190-4715-23-15.

3

Herbicides

Malcolm Devine

Herbicide Discovery

Traditionally, herbicides have been discovered through large chemical synthesis and screening programs. Given the high cost of discovery and development of new compounds, efforts have focused on herbicides that target major weeds in major world crops (e.g., corn, rice, wheat, soybean, cotton). However, these compounds often find uses in minor crops, also. The process normally involves a step wise progression from greenhouse screening on a few crop and weed species, to more extensive indoor testing, and eventually to field trials in many locations examining the interactions between different weed species, soil types, weather conditions, etc. Toxicological testing and formulation improvement proceed concurrently, to ensure that regulatory and efficacy requirements are met.

A recent innovation is the use of combinatorial chemistry to identify lead compounds. Rather than being synthesized and evaluated in isolation, compounds are produced and screened as mixtures. When combined with in vitro screens (activity testing at the biochemical or cellular level) rather than whole-plant assays, this can allow the testing of 20- to 50-fold more compounds per year than a traditional herbicide discovery program.

"Rational" discovery of herbicides involves identification of a candidate herbicide target site in the plant, followed by design of inhibitors that specifically block that target. While this has led to the discovery of some novel enzyme inhibitors, to date no commercial herbicides have been discovered through this approach. One difficulty is that compounds predicted to have high activity might not penetrate the tissue satisfactorily or may be rapidly degraded inside the tissue.

Finally, herbicides can be developed from bacterial, fungal, or plant toxins. One commercial herbicide, glufosinate (= phosphinothricin), was developed from the bacterial toxin bialaphos from *Streptomyces hygroscopicus.*

Herbicide Classification

Method or Timing of Application

Preplant incorporated herbicides are applied to the soil surface and mechanically incorporated into the upper 5–10 cm of the soil, in order to minimize photodecomposition and volatilization losses. Preemergence herbicides are applied to the soil surface and often rely on precipitation or soil moisture to transport them to the plant root or shoot for uptake. Postemergence herbicides are applied to exposed foliage after the plants have emerged.

Chemical Structure or Mode of Action

Herbicides within the same structural family usually have the same mode of action, with varying degrees of activity or selectivity depending on the structural variations between compounds. However, herbicides from different chemical families can have the same mode of action (see later).

Components of Herbicide Action

To be effective, herbicides must penetrate the tissue and reach the target site in sufficiently high concentrations to block its activity. The overall process of herbicide action can be separated into the following components:[1]

Absorption: Herbicides can be absorbed by the roots or directly into the leaves. Root uptake occurs through mass flow of herbicide in soil moisture and is a function of root distribution in the soil, soil moisture status, the physical properties of the soil, and the behavior of the herbicide in the soil. Foliar absorption occurs following application to the leaves. To facilitate entry through the cuticle (the waxy layer on most leaf surfaces), herbicides are usually formulated with an array of inert ingredients including surfactants, emulsifiers, etc. Once inside the tissue, further penetration through cell walls and membranes usually occurs by simple diffusion.

Translocation: Long-distance transport of herbicides in the plant can occur in the xylem and/or phloem. In some instances, distribution through the plant is a critical component of overall activity. The amount of translocation depends on the plant stage of development, the physicochemical properties of the herbicide, and the rate at which herbicide injury slows down the movement of endogenous compounds.

Metabolic degradation: Plants have evolved various enzyme systems to detoxify potentially harmful compounds. Fortuitously, some of these enzymes (e.g., cytochrome P450 monooxygenases, glutathione S-transferases) can degrade herbicides to inactive compounds.

Interaction at the target site: Finally, all herbicides must interfere with some critical process in the plant. In most cases this involves binding to a protein (usually a functional enzyme, or a transport or structural protein) so that it cannot carry out its normal function. Over time, through the combined effect of this direct action and other indirect actions, the plant dies.

Herbicide Selectivity

Most herbicides are selective, that is, they kill some plant species but not others. A few herbicides, on the other hand, are nonselective, and kill essentially all species. Selectivity is normally based on one of the following:[1]

- Failure to absorb the herbicide, due to either selective placement (e.g., directed spray on weeds growing between the crop rows) or failure of the herbicide to reach the roots of deep rooted crops, while killing shallow rooted weeds.

- Enhanced rate of metabolic degradation of the herbicide in the crop. This is the most common basis of selectivity of agricultural herbicides. After entry into the plants, the crop metabolizes the herbicide to inactive compounds, whereas degradation does not occur or is slower in the susceptible weeds.
- Differential sensitivity of the target site. This is particularly common in the case of herbicide resistant weeds (see later).

Herbicide Mode of Action

Herbicides kill plants by interfering with an essential process in the plant. The major modes of action of herbicides, the biochemical target sites, and some examples of chemical groups that interfere with those targets, are shown in Table 1.

In addition, many herbicides have been identified that interact with other, unique target sites in plants. However, most of the herbicides that have been developed over the past 50 years target about 15 distinct molecular targets.

Herbicide Resistance in Weeds

The repeated use of herbicides can lead to the development of herbicide-resistant weed populations. The use of herbicides per se does not create herbicide resistance. Rather, the continuous selection pressure through herbicide use creates a niche, allowing resistant individuals to survive and increase in

TABLE 1 Modes of Action of Major Herbicide Groups[a]

Mode of Action	Target Site	Representative Chemical Groups
	Inhibition of amino acid biosynthesis	
Branched chain amino acids	Acetolactate synthase	Sulfonylureas, imidazolinones, triazolopyrimidines
Glutamine synthesis	Glutamine synthetase	Glufosinate
Aromatic amino acid biosynthesis	Enolpyruvylshikimate phosphate synthase	Glyphosate
	Photosynthesis	
Photosynthetic electron transport (PS II)	Q_b or D1 protein	S-Triazines, phenylureas, benzonitriles
Photosynthetic electron transport (PS I)	PS I electron acceptor	Bipyridiliums
	Pigment biosynthesis	
Chlorophyll synthesis	Protoporphyrinogen oxidase	(Nitro) diphenylethers
Carotenoid synthesis	Phytoene desaturase and others	Aminotriazole, clomazone
	Lipid biosynthesis	
Fatty acid synthesis	Acetyl-coA carboxylase	Aryloxyphenoxypropionates, cyclohexanediones
Fatty acid elongation	"Elongase" complex	Thiocarbamates
	Cell division	
Spindle formation	β-Tubulin	Dinitroanilines, carbamates
	Other	
Auxin disruption	Auxin binding proteins (?)	Phenoxyacetic acids, benzoic acids
Homogentisate biosynthesis	4 hydroxyphenylpyruvate dioxygenase	Isoxazoles

[a] From Devine.[1]

the population as susceptible weeds are killed. Thus, starting with an initial population of perhaps one resistant weed in a population of 10^6–10^9, resistance builds up over time until the resistant weeds become predominant in the field.

Herbicide resistance in weeds was first observed in the late 1950s, but was a minor problem until the mid-1970s, when resistance to triazine herbicides became a widespread concern. Since then the occurrence of resistance has increased dramatically, with >200 cases now reported.[2]

In most cases, resistance is due to a point mutation in the gene coding for the herbicide target site.[3] This alters the structure of the target protein in such a way that it can still perform its natural function, but herbicide binding is reduced. This type of resistance often confers cross-resistance to herbicides in the same chemical family or mode-of-action group, although exceptions exist and each case must be analyzed separately. However, target-site mutations do not alter the sensitivity of the weed to herbicides with other mechanisms of action.

Resistance can also be conferred by elevated activity of the enzyme(s) responsible for herbicide degradation. These weeds can be cross-resistant to other herbicides with different mechanisms of action. Again, the possibilities for cross-resistance have to be analyzed on a case-by case basis.

Although herbicide resistance has become widespread, in almost all cases alternative control methods are available, through the use of other herbicides, changes in cropping or tillage practices, or some combination of these. Avoidance and management of herbicide resistance has become an integral part of good farming practices in modern agriculture.

Herbicide-Resistant Crops

A recent development in selective weed control has been to create resistance to certain herbicides in crops where it did not exist previously. While this has been done primarily to extend the market share of certain products, it offers farmers the advantage of broad-spectrum weed control in crops with a single herbicide application.[4] In some cases this has substantially reduced the total amount of herbicide required in a single season.

Herbicide-resistant crops can be produced by three methods:[4]

1. Making crosses between the crop (sensitive to the herbicide) and a related, resistant species. This method was used to develop triazine-tolerant canola (*Brassica napus*).
2. Selecting resistant cells in tissue culture, through random mutation or by selecting somaclonal variants, and regenerating resistant plants from these cells. Corn lines resistant to the herbicide sethoxydim were generated in this way.
3. Transfer of an herbicide-resistance gene through genetic engineering. The gene is identified in an unrelated species (often a bacterium), cloned, and transferred into the crop species of interest. These procedures were used to develop canola varieties resistant to the nonselective herbicides glufosinate and glyphosate.

Herbicide-resistant crops have greatly facilitated weed control, but present some new research questions that have had to be addressed. These include the likelihood and long term ecological consequences of gene flow to related species, and the need to control volunteer plants in the following season(s). These issues do not present insurmountable obstacles but, again, need to be dealt with on a case-by-case basis.

Safety and Environmental Fate of Herbicides

Environmental safety is of prime concern in the development of new herbicides. This includes an understanding of herbicide toxicology, safe handling and application procedures, and environmental behavior and fate. In most countries, approval of herbicides by the relevant decision making bodies is dependent on the registrant satisfying the regulatory requirements imposed by those countries.

Toxicological requirements vary from country to country, but usually include data on oral and dermal toxicity in a range of species, in tests of varying duration. Based on the data collected, maximum residue levels are established in food or food products. Data from these tests and field experiments are used to establish maximum application doses and safe intervals between product application and harvest or grazing. Data may also be required on effects on nontarget organisms and ecosystems that may be exposed to low herbicide doses.

Appropriate handling procedures are an important aspect of herbicide safety. The use of appropriate safety clothing (gloves, coveralls, masks, etc.), more benign formulations (e.g., dispersible granules rather than wettable powders), etc., contribute to reduced applicator exposure. Recently, novel formulations have been developed that further reduce applicator exposure when adding products to the spray tank.

Herbicide drift immediately after spraying can be a source of off-site contamination, resulting in injury to adjacent sensitive crops and other species. Various measures, including spraying only under calm conditions, use of wind deflectors, and avoiding very small droplets, can substantially reduce the risk of spray drift.

Herbicides in soil are lost by a combination of microbial and chemical degradation, plant uptake, and, in some instances, leaching or surface run-off. Stringent environmental regulations have been introduced in many countries to minimize the possibility of groundwater contamination. Herbicide residues in soil can provide extended weed control, but also may limit crop rotation options in future seasons. Field research is conducted to establish the risk of such carryover and the effects on future crops.

References

1. Devine, M.D.; Duke, S.O.; Fedtke, C. *Physiology of Herbicide Action*; Prentice-Hall: Englewood Cliffs, NJ, 1993; 441.
2. Heap, I.M. International survey of herbicide-resistant weeds. http://www.weedscience.com (accessed April 2000).
3. Devine, M.D.; Eberlein, C.V. Physiological, Biochemical and Molecular Aspects of Herbicide Resistance Based on Altered Target Sites. In *Herbicide Activity: Toxicology, Biochemistry and Molecular Biology*; Roe, R.M., Burton, J.D., Kuhr, R.J., Eds.; IOS Press: Amsterdam, 1997; 159–185.
4. *Herbicide-Resistant Crops. Agricultural, Environmental, Economic, Regulatory and Technical Aspects*; Duke, S.O., Ed.; CRC Press: Boca Raton, FL, 1996; 420.

4

Herbicides: Non-Target Species Effects

Céline Boutin

Introduction

Pesticides such as inorganic chemicals (e.g., sulfur, arsenic, or other metal compounds) have been used in agriculture for centuries to protect crops. In more recent times, especially after World War II, organic pesticides have been discovered and increasingly used to suppress unwanted plants, insects, and other organisms that interfere with crop production. The main pesticides used are categorized by their target organisms: insecticides used to kill or suppress insects, fungicides used for pathogens, and herbicides for weeds. In this entry, we are mostly concerned with herbicides because of their considerable use, especially in North America, and their phytotoxic effects. The definition accepted by the Weed Science Society of America is that an herbicide is “a chemical substance or cultured organism used to kill or suppress the growth of plants.”[1]

Phytotoxicity refers to the capability of herbicides or other pesticides to exert toxic effect on plant growth, reproduction, and survival. By extension and in addition to terrestrial and aquatic vascular plants, it usually includes other primary producers such as algae and cyanobacteria, which will not be considered in this entry.

This entry will first consider the history, types, and main uses of herbicides. There are advantages and disadvantages of using herbicides in agriculture and forest management. Benefits have long been established; however, the limitations and undesirable effects attributed to herbicide use are still debated. The environmental impact will be examined, including environmental exposure, phytotoxicity, and toxicity to different trophic levels, as well as the multitudinous factors to take into account in risk assessment. Lastly, the techniques and limitations of the phytotoxicity assessments currently used to determine environmental risk evaluations will be discussed. Mitigation measures and alternatives to herbicide use will be discussed in conclusion.

History, Types of Herbicides, and Their Use

DNOC or 4,6-dinitro-*o*-cresol, developed in 1935, was the first organic herbicide used to control weeds. This herbicide was also used as an insecticide, fungicide, and a defoliant and was shown to be toxic to animals.[2] It is no longer used in many countries. In the 1940s, several phenoxy herbicides were discovered and many are still used today, including 2,4-D and MCPA [(4-chloro-2-methylphenoxy)acetic acid]. In the 1950s, other herbicides appeared on the market, including diuron, diquat, paraquat, and triallate. In the 1960s, the triazine herbicides were developed and are still considerably used, especially atrazine. Glyphosate, which is still the most widely used herbicide worldwide, was developed in the 1970s. In the 1980s, low-dose high-efficacy herbicides such as sulfonylureas and imidazolinones were found. Since the late 1990s, a new method for controlling weeds has emerged with the development of herbicide tolerant crops, i.e., genetically engineered crops resistant to glyphosate and, to a lesser extent, glufosinate ammonium. More genetically modified herbicide tolerant crops will no doubt be engineered or bred in the future.

Herbicides exhibit different mechanisms of action.[1,3] They include disruption of photosynthesis (uracils, substituted ureas, and triazines), inhibition of lipid biosynthesis (carbamothioates such as EPTC [s-ethyl dipropylthiocarbamate], triallate, clethodim, fluazifop, and metolachlor), inhibition of cell division (dinitroanilines such as trifluralin and pendimethalin), plant hormone mimics (the phenoxy herbicide 2,4-D, MCPA, the benzoic acid dicamba, and the picolinic acid picloram), inhibition of amino acid biosynthesis (glyphosate, sulfonyl ureas including metsulfuron methyl and chlorsulfuron, imidazolinones including imazethapyr), blockage of carotenoid biosynthesis (clomazone), and disruption of cell membranes (acifluorfen and bipyridylium compounds such as diquat and paraquat). Other herbicides such as glufosinate ammonium act by inhibiting glutamine synthetase, thus leading to a complete breakdown of ammonia metabolism in affected plants. Uncouplers of oxidation phosphorylation, such as the widely used bromoxynil, interfere with plant respiration.

Many herbicides act primarily on systems unique to plants, e.g., photosynthesis, but some herbicides act at more than one site of action. Undoubtedly, the secondary mode of action of some herbicides could explain their relatively high toxicity to animals (see below). In the case of some herbicides, the precise mode of action is unknown and exact molecular sites of action remain to be determined.[4]

The number of herbicides listed in the Weed Science Society of America reached 374 in 2010.[5] In Canada, there are 500 pesticide active ingredients (the ingredient to which the pesticide is attributed) and 7000 pesticide formulated products (mixture containing one or several active ingredients and formulants) available since many formulated products contain a mixture of active ingredients.[6] The number of herbicide active ingredients registered in Canada amounts to approximately 125.[6] The majority of herbicide use occurs in agriculture where in modern practices they dominate weed control practice. Approximately 90% of areas planted with corn, cotton, potato, wheat, and soybean were sprayed with herbicides in the United States in 2004.[1]

While the benefits of using herbicides from an agronomic perspective are well known, the undesirable effects of herbicides on the environment have not always been considered carefully. Regardless of the method of application, it is generally accepted that misplacement will take place through drift at the time of application or through runoff, leaching, and volatilization from soil or plants or from particles moving with contaminated soil after application has occurred.

Benefits to Agriculture and Forest Management

Benefits of herbicide to agriculture, forest management, and other agronomic applications leave no doubt as to their utility, although some applications for cosmetic reasons (e.g., domestic use, horticulture, golf courses) are more questionable. The negative relationship between crop yield and weed density is well established.[1] Weeds are well-adapted species that compete with crops in disturbed environments and will reduce crop yield depending on their germination timing and densities, growth patterns, and growth rates. However, there may be cases where weeds can be beneficial to crops. As an example, field experiments were conducted to evaluate the effects of nicosulfuron and imazethapyr, a sulfonyl urea and an imidazolinone herbicide, respectively, for the control of johnsongrass (*Sorghum halepense* L.), a weed difficult to control in corn fields.[7] It was noted that corn vigor was greatly reduced in plots where these two herbicides were applied and where johnsongrass was reduced. In order to verify if there was an unwanted effect of the two herbicides on corn, small plots were sprayed with the two herbicides separately. Two other treatments were also included in the experimental design: control of johnsongrass via mechanical means and no treatment. It was noticed that in plots where johnsongrass was removed, corn was more prone to being attacked by the maize dwarf mosaic and maize chlorotic dwarf viruses, which are transmitted by aphids or leafhoppers. Results revealed that in treated plots, virus disease was increased because the preferred host johnsongrass was suppressed. Furthermore, there was no change in yield between treatments. Crop yield was reduced in treated plots (whether through mechanical or chemical treatments) due to increased virus severity, while in non- treated plots, reduced corn yield was due to more competition from johnsongrass.

Herbicide Use and Exposure to Primary Producers

The amount of herbicide active ingredients used worldwide exceeded 950 million kg in 2007, of which 39% was herbicides.[8] Herbicides are used most extensively in North America. For instance, in 2003, herbicides accounted for 77% of total pesticide sales in Canada, while fungicides represented 9%, insecticides 8%, and other products 6%.[6] In the United States, 47% of pesticides used were herbicides, while insecticides and fungicides amounted to 8% and 6%, respectively.[8] In Europe, fungicides are used extensively mainly due to the use of sulfur fungicides on vineyards by France, Italy, Spain, and Greece.[9] Herbicides constitute less than 35% of pesticides used in Europe, but nevertheless, this is approximately 75,000 tons of applied active ingredients.[10] Of note, it appears that these European data underestimated the actual amount of pesticides used.[9] France is the biggest herbicide user followed by Germany and the United Kingdom. In other parts of the world, insecticides constitute the bulk of the pesticides used.[11]

Large amounts of herbicides are annually used in terms of hectares sprayed and in terms of quantity per hectares. In Canada, in excess of 25 million ha are sprayed with atleast one herbicide on an annual basis.[12] In the United States, 110 million ha were treated with herbicides in the late 1980s,[13] and in 1997, 209 million kg of herbicide active ingredients were used (Olszyk et al.[14] and references therein). Total land area devoted to agriculture worldwide is considerable,[15] especially in Europe where around 70% of the land is cropland or pastureland.[16] As a result, total areas treated with herbicides reach large figures. Herbicide application is highly dependent on the types of crops, the prevailing climate conditions, the land use, and other factors such as topography. In the United Kingdom where 77% of the land is used for agriculture, pesticide use was also intensive at 5.8 kg/ha in 1988[17] but seemed to have declined in 1999 at 2.7 kg/ha (1.4 kg-ai of herbicides), not taking into account the underestimation mentioned previously.[9] By contrast, in France, the intensity of pesticide used increased from 4.4 kg-ai/ha in 1988 to 6.0 kg-ai/ha in 1999, including 2.0 kg-ai/ha of herbicides. This is well above the average for European member states, which used approximately 4.5 kg-ai/ha of pesticides and 1.3 kg-ai/ha of herbicides. In the Netherlands, 11.8 kg-ai/ha of pesticides are used on an annual basis (mostly non-herbicides).[18] On the other hand, in Canada, pesticide use is low at 0.9 kg/ha (largely herbicides).[6] The United States uses 23% of all the pesticides used in the world, and 28% of all herbicides.

These figures, however, dissimulate important facts. In Canada, where less than 10% of the land is devoted to agriculture, two ecoregions are particularly important in this regard, the Prairies of western Canada and the Mixed- wood Plains of the Great Lakes–Saint Lawrence corridor. The Prairies, which constitute the largest agricultural areas in Canada (5.1% of the land),[19] has been almost totally modified to satisfy increasing needs for more cropping areas. Consequently, there are almost no pristine grassland prairies left where native plant and animal communities can survive. The Mixedwood Plains are a smaller area in southern central Canada (1.5% of the land cover) where 50% of the Canadian population lives and where agriculture is also very intense. In this region, a large portion of the Carolinian forest and the mixed-wood forest has disappeared to satisfy human needs. In the United States, agriculture covers 48% of the land and is concentrated in the mid-west regions[20] where original ecosystems have largely vanished. The same pattern is repeated in Europe where mostly seminatural habitats remain in most countries. What remains of these ecosystems interspersed in a sea of intensively cultivated land can be greatly impacted by herbicide use.

Herbicide exposure to non-target environments can occur when application is performed with aircraft, mist-blower, and ground applications through overspray or via spray drift, vapor drift, revolatilization from soil and plants, runoff, or dust particles moved by wind or water.[14] Depending on the equipment and prevailing weather conditions during application, the amount of sprayed herbicide that will deposit in hedgerows and other field edges from multiple consecutive spray tracks can reach 1% to 10% of the application rate within 10 m of a single swath with ground equipment (Boutin and Jobin[21] and references therein), and much more with mist-blower sprayers and aerial applications.[22] Herbicides can travel a considerably longer distance with aerial equipment applications, for example, 500 m downwind from the source.[23,24]

With ground application, it was found that herbicides could cover long distances. A study was undertaken to assess the protection afforded by buffer zones from herbicide drift (Boutin and Baril, unpublished data). Surveys of nontarget plants situated in two small woodlots adjacent to crop fields were conducted prior to (May) and after (May, June, and July) herbicide application. Vegetation was surveyed for community composition and symptoms of herbicidal impact. The experimental work was conducted in southwestern Ontario, Canada, under normal field operation conditions, for soybean sprayed with imazethapyr in 1993, corn with dicamba in 1994, and wheat with MCPA in 1996, following the usual rotation in southwestern Ontario. The buffer zone was defined as a 12 m wide seeded strip of crop, upwind of a woodlot, where herbicides were not applied. Each treatment consisted of four transects, divided into sampling points, at 1 m, 2 m, 4 m, 8 m, 16 m, and 32 m distances into the woodlots. Herbicide application occurred in the early morning or evening, when no precipitation was forecasted, when wind speed was at or less than 8 km/hr, and when the direction of the wind was across the soybean field into the woodlot.

Results revealed that herbicides could move up to 32 m into woodlots (Figure 1). Up to 43% of the vegetation of one quadrat (of the 23 species) in the woodlots showed visual effects characteristics of herbicidal injury: discoloration, bleaching, epinasty, yellow or brown spots, etc. Effects were less pronounced in transects abutted to buffer zones. The plants most affected [e.g., raspberry (*Rubus idaeus* L.), goldenrod (*Solidago canadensis* L.), and ash tree (*Fraxinus* spp.)] were species of open areas growing in the first few meters of the woodlots. In some cases, effects lasted for more than 2 mo.

Vapor drift can also migrate a long way (Franzaring et al.[25] and references therein) causing recurrent sublethal effects on native plants not only in bordering seminatural habitats but also in more remote habitats. Presence of airborne herbicides in the atmosphere has been reported in Europe[26] and North America.[27] In the Netherlands, it was found that non-target vegetation was repeatedly exposed to small amounts of herbicides (Franzaring et al.[25] and references therein). Pesticides used in agriculture in southern Canada, including some herbicides, have been reported in the arctic environments.[28] Of the 10 chemicals surveyed by Hoferkamp et al.,[29] 9 were detected in the arctic. Traveling distances for these chemicals ranged from 55 km to 12,100 km. Persistence in the environment of these herbicides can explain their presence in the arctic where they are transported via the air or by dust.

Undoubtedly, the large number of herbicides available for use, the geographical extent of their use on different crops, and the quantity applied together with the method of application suggest a high probability of exposure to primary producers and other wildlife.

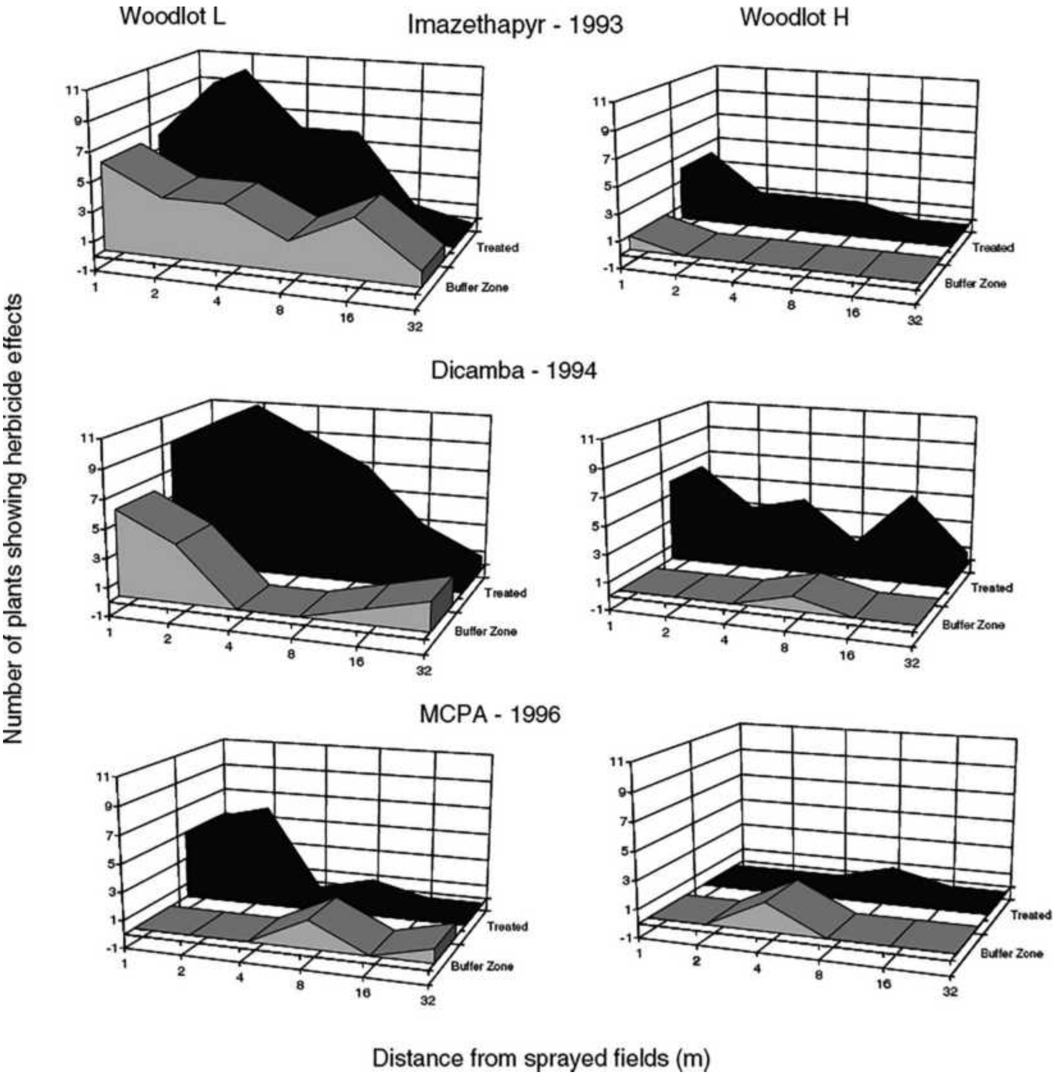


FIGURE 1 Number of plants showing herbicide effects in two woodlots (L and H) surveyed during 3 years prior to and after herbicide spray under normal field operations. Fields were sprayed with the herbicides imazethapyr in 1993, dicamba in 1994, and MCPA in 1996. Plants were surveyed in quadrats situated between 1 and 32 m from crop fields. A 12 m buffer zone within fields was used in half of the fields while the other half was sprayed right to the field edge. Four transects were placed for each treatment. In total, 116 species were inventoried and 35 species were found to be sensitive to herbicides during the course of the 3 years the study lasted.

Herbicide Toxicity to Humans and Animals

Herbicides are generally not very toxic to humans or to animals, including mammals, birds, amphibians, and invertebrates, as well as to microbial organisms. There are some exceptions, however, notably paraquat, which is very toxic to humans. It is still used extensively in the United States[30] and many developing countries. A large epidemiological study suggested that the herbicides pendimethalin and EPTC may be linked with human pancreatic cancer.[31] Sulfonyl ureas, which comprise a large number of low-dose herbicides widely used in agriculture, are known to affect humans as they are used in medicine to treat Type 2 diabetes and other illnesses.[32] They act by increasing insulin release from the beta cells in the pancreas, and side effects such as incidence of hypoglycemia have been reported.

As early as 1976, the herbicide diquat was shown to have an adverse effect on body size and pigmentation of *Xenopus laevis,* the South African clawed frog (family: Pipidae; subfamily: Xenopinae). Moreover, when treated with diquat and the fungicide nabam, the deleterious effect was more pronounced.[33] More recently, the herbicide atrazine has been implicated in a number of reports demonstrating effects on animals as an endocrine disruptor.[34] Similarly, the formulated product containing glyphosate is widely known to affect amphibians.[34] Glyphosate formulation has been implicated in a synergistic effect with a trematode species on fish.[35] Survival of juvenile fish was unaffected by exposure to glyphosate alone or by an infectious trematode parasite alone. However, simultaneous exposure to infection and glyphosate significantly reduced fish survival. Spinal malformations of juvenile fish were also enhanced when both stressors were present. A species of snail acts as the vector between the fish and the trematode. Glyphosate at high concentration killed all the snails, but at moderate concentrations, the snail produced more trematodes than in control and low-concentration groups.[35] This elegant experiment demonstrated the intricate interactions between different components of ecosystems.

Herbicides are categorized for their effect on unwanted plants, yet a number of them demonstrate toxicity to microbial organisms. A recent study revealed the control activities of glyphosate against rust diseases (*Puccinia striiformis* and *Puccinia triticina*) on glyphosate-resistant wheat and soybeans.[36] Control was equivalent to that of registered fungicides. Similarly, diquat was investigated for its potential to control the bacterial infection Columnaris disease (*Flavobacterium columnare*) affecting several fish species and was found to reduce Columnaris infection.[37]

These findings demonstrate that herbicides can be directly toxic not only to plants but also to different types of life form at various trophic levels, including humans.

Toxicity to Plants and Effects to Terrestrial Habitats

Herbicides are especially designed to control weeds and therefore are of particular concern for unwanted effects on desirable plants. However, several types of pesticides can injure plants since pesticides are mostly classified by target organisms. For example, the widely used insecticides chlorpyrifos, diazinon, carbaryl, malathion and others can cause injury to several crops and ornamental species.[38] Several fungicides belonging to the benzimidazole chemical class were shown to be toxic to several plants.[39] Nevertheless, this entry will concentrate on herbicides.

Effects of herbicides on crops and weeds are well documented for obvious agronomic and economic reasons. Unwanted effects on native species is usually assessed through regulatory processes using crops as surrogate species (see below) and via studies performed sporadically by research scientists. By and large, herbicides affect native species at different levels. Herbicides can alter biochemical and developmental processes in plants as well as plant morphology. Habitats within agroecosystems can experience modifications in their species abundance, composition, and diversity when subjected to herbicides. Plants as terrestrial primary producers constitute the basis of terrestrial ecosystems. Herbicide effects on them can have cascading effects at other trophic levels, on overall biodiversity, and on ecosystem functions. These will be addressed in turn.

Effects at the Species Level

Herbicides can modify biochemical processes in plants and as a result can increase plant susceptibility to pests and diseases. Most of the work demonstrating these effects has been performed with crop plants, but there is no reason to believe that native plants would not be affected in the same manner. The incidence of mildew on spring wheat was enhanced by three different herbicides as a result of stress induced by metabolic interference.[40] Wheat treated with 2,4-D was higher in protein content resulting in a proliferation of aphids.[41] The concentration of nitrogenous compounds in crops was enhanced by sublethal doses of several phenoxy, triazine, and uracil herbicides,[42] sometimes leading to an increase in pathogens and pests on corn[43] and rice.[44]

Symbiotic processes with plants such as nodulation and mycorrhizae, which are vital for the biochemical activities of most terrestrial plants, can be greatly modified by herbicides. Nodulation and nitrogen fixation was disrupted by herbicides resulting in deleterious effects on growth and reproduction in crops such as dry bean (*Phaseolus vulgaris*), soybean (*Glycine max*), broad bean (*Vicia faba*), and peanut (*Arachis hypogea*) (Schnelle and Hensley[45] and references therein). Mycorrhizal activities were affected by MCPA sprayed on (*Pisum sativum*) with ensuing decreases in growth observed.[46]

Herbicides can exert anomalous effects on plant developmental processes. They have been shown to affect seed production and seed germination. It was demonstrated as early as 1948 that applications of 2,4-D caused a delay in seed germination and growth of wheat plants sufficient to favor an increase of wireworm damage.[47] It has been known for some time that when some of these herbicides are applied to cereal crops late in their growth stage, just before seed formation, the plants produced far fewer seeds. Greenhouse experiments recently revealed that at typically used rates, dicamba and picloram reduced all or nearly all seed production while 2,4-D was much less effective.[48] Further field test experiments supported these greenhouse results.[49]

Research with glyphosate showed that depending upon the plant species, application rate, and the timing of application, effects on seed production, seed germination, and seedling development have been observed on a large number of plant species from various families.[50] Glyphosate produced an inhibitory effect on pollen germination and seed formation when applied at the flower bud stage of goldenrod (*S. canadensis*).[51] Germination, emergence, and plant establishment of native Australian plant species were impeded by the herbicide fluazifop-*p*-butyl.[52] Plants of *Stellaria media* treated with the herbicide glufosinate ammonium produced seeds with reduced germination and emergence.[53]

Herbicides can instigate unexpected effects on plants that have not been studied as part of the registration package because they are not relevant from an agronomical viewpoint. Sulfonylurea herbicides are selective herbicides that act by inhibiting the enzyme acetolactate synthase, which catalyzes the synthesis of the three branched-chain amino acids valine, leucine, and isoleucine. They are very potent herbicides applied at exceedingly low doses, in the order of a few grams per hectare. When applied at the fraction of the recommended label rate at the onset of flower formation, however, they were found to reduce the reproductive outputs of several species with significant reproductive damage occurring with only scant visible symptoms on leaves. Cherry trees sprayed at doses as low as 0.2% of the field application rate of the sulfonylurea chlorsulfuron showed a significant reduction in the production of fruits, with almost no observable damage to vegetative parts.[54] Low doses of metsulfuron methyl produced important injury on the vegetative biomass when crop and native species were sprayed at the seedling stage, but plants sprayed at later stages showed considerable reduction in the reproduction.[55] Potato bulking was greatly affected by low rates of sulfometuron methyl whereas few visible symptoms were observed on the vegetative parts.[56] Under field situations, it was found that berry production in hawthorn (*Crataegus monogyna* Jacq.) was severely affected by average spray drift concentrations higher than 2.5% of the label rate of the sulfonyl urea metsulfuron methyl (0.1 g-ai/ha) and that the effect was still observed 1 year after the spray event.[57,58]

Imidazolinones are another chemical family classified as low-dose herbicides. They also inhibit the enzyme acetolactase synthase and as such have been implicated in unforeseen effects. Potato (*Solanum tuberosum* L. from the Solanaceae family) tuber size and quality were most detrimentally affected by exposure to imidazolinone herbicides, as low as 0.1 times the recommended field rate, when exposure occurred during tuber bulking, as compared to exposure at seedling emergence or at tuber initiation.[59] While reductions in the weight and the overall yield of sensitive species were problematic, another major concern was the potential lowering of fruit quality, which could lead to significant economic losses in the case of crops.[56,59] Effects on native species have not been investigated, but it is easy to stipulate that drift of imidazolinones to non-target habitats could elicit a reduction in weight of sensitive species, including effects on underground and reproductive parts. This may lead to species of the Solanaceae family and other families to produce fruits and storage organs that are less attractive to wildlife.

Damage to or modification of plant morphology, including epinasty, inhibition of leaf expansion, and stem and root distortion, has been reported in weeds and crops, particularly for phenoxy herbicides.[60] Herbicides can cause morphological deformation with unforeseen repercussions. For instance, it was possible to demonstrate that morphological deformations of the flowering parts in *Arabidopsis thaliana* (L.) Heynh. induced by some herbicides prevented pollination and seed production.[61] Bud and flower abortion has been reported for sulfonylureas (see above). Such effects, usually undetected, can have long term adverse impacts, particularly on monocarpic species.

Population/Community Level

Herbicide use in modern agriculture alters species abundance, composition, and diversity in non-target habitats. Effects of low doses of herbicides on plants (grown in pots and placed at different distances from the spray swath) have been reported.[62,63] Marshall and Bernie[64] showed that several of the broad-leaved species found in field margins were susceptible to the six different herbicides tested separately in pots. In the field, Marrs et al.[65] assessed effects of spray drift in relation to plant damage and yield for a range of plant species of conservation interest in Britain after applying each of six herbicides with a standard agricultural hydraulic ground sprayer. They observed lethal effects at 2–6 m from field edges and damage at greater distances, although damaged plants were able to recover within the growing season. Effects on seedlings were observed up to 20 m.[66] The long-term effect of such damage to plants remains unknown. In North America, work performed by Jobin et al.[67] showed that recurrent applications of herbicides had a long-term effect on plant populations inhabiting hedgerows and woodlot edges adjacent to crop fields. In Britain, several native arable weeds are considered endangered due to destruction of their habitats and extensive use of agrochemicals.[68]

In the Netherlands, it was calculated that 9.5% of all pesticides applied was dispersed outside croplands.[18] Drift scenarios together with herbicide toxic effects investigated using bioassays and taking into account distances from spray events were used to estimate impact on biodiversity. In 2005, 41% of the linear landscape features near cropland were affected. This was an improvement since, in 1998, 59% of the area was affected. Natural areas located within farming regions were also affected by herbicide displacement in 31% and 11% of the area in 1998 and 2005, respectively. Measures in place such as unsprayed buffer zones and better equipment as well as reduced reliance on herbicides were largely responsible for this decline in unwanted effects on plant diversity. Also in the Netherlands, small plot experiments used to investigate the effects of the herbicide fluroxypyr on plants monitored for 3 years showed a decline in diversity and biomass.[69] Change in plant populations was also noticed.

Ecosystem/Trophic Level

Alterations by herbicides on primary producers through effects on morphology, physiology, phenology, species composition, diversity, and abundance can resonate considerably to other trophic levels. The best documented study implicating herbicide repercussions was conducted in Britain over several decades. The grey partridge (*Perdix perdix*) has been surveyed since 1933 in the margins of crop fields, and it was found that numbers declined by 80% between 1952 and the mid-1980s.[70] Studies conducted from the 1960s led to the conclusion that the use of herbicides and, to a lesser extent, insecticides precipitated the decline of grey partridge populations. Although partridges are largely herbivorous, newly hatched chicks feed largely on arthropods during the first 2–3 weeks of their lives. The falling number of grey partridges in agricultural land was attributed to declining chick survival early in the season due to weed removal by herbicides. A reduction in weed diversity and density on which insects feed and inhabit caused a food shortage at this very crucial period of the year.[71] Removal of field margins was also a contributing factor.

Many subsequent studies have shown that as a result of herbicidal effects, cover and diversity of flowering plant species are reduced in crop fields and field margins, thus subsequently reducing the resources available to flower- visiting insects and other arthropods.[72–74] Likewise, abundant floral diversity was found to be the prevailing factor related to high Lepidopteran diversity in farmland habitats (Boutin et al.[75] and references therein).

Factors Interacting with Herbicide Effects

Toxicity to plants and other organisms may be exacerbated by the chemical and physical properties of a compound, namely, volatility, mobility, and persistence in the environment. For example, atrazine is a widely used product in North America applied for the pre- or postemergence control of broadleaf weeds and grasses, predominately on corn crops. It is a selective compound with systemic activity resulting in chlorosis and eventual plant death. The primary mode of degradation of atrazine is through hydrolysis. It has been reported to persist in soil, water, and aquifers for years.[76] In a study conducted in Germany, it was found that the herbicide atrazine was still present 22 years after its application in soil.[77] It is regularly found in groundwater in the United States and Canada where it is widely used. Unquestionably, the persistence of atrazine explains its presence and accumulation in the environment and long-term effects.[78]

Most herbicides are highly water soluble and therefore can move away from crop fields with rainwater. Volatile compounds can elicit damage to non-target plants; clomazone applied in crop fields produced striking bleaching effects that caused the U.S. regulatory agencies to consider wild plants in their risk assessments.[79]

Persistent herbicides may cause long-lasting and unexpected contamination. Clopyralid is a growth-regulator-type herbicide used for the control of broad-leaved species. It was used on sugar beet that was later fed to cattle with the resulting manure containing sufficient residues to contaminate crops on which it was subsequently spread.[80] The persistence of clopyralid in compost and mulches was also demonstrated in the United States.[81] The same problem occurred with the related compound aminopyralid, an auxinic herbicide registered for the control of broad-leaved weeds on grassland and rangeland. Grass treated with aminopyralid persisted in the silage for more than 1 year. Cattle or horses fed with the hay produced contaminated manure toxic to receiving crops.[82]

Herbicide toxic effects may vary depending on the characteristics of the receiving environment. Residues of the sulfonylurea chlorsulfuron have been detected and caused crop damage as long as 7 years after application in Alberta, Canada.[83] Sulfonylureas are known for their increased activity at soil pH above neutrality, which is often observed in soils in western Canada.[84] Temperature may also exert effects on persistence and thus on long-term toxicity. In Canada and elsewhere where winters are long and cold, many pesticides, including chlorsulfuron, will take longer to degrade.[85]

Direct toxicity of herbicides to primary producers and other organisms is typically assessed without taking into account confounding stressors present under normal agriculture practices or occurring in natural conditions. For instance, high fertilizer application in intensive agriculture is the norm. Few studies on the misplacement of fertilizer in non-target adjacent areas have been published, but Rew et al.[86] in Britain demonstrated that off-target deposition could vary from 2% to 133.3% of the application rate in field margins, depending on the type of machinery used. Some species will thrive under high nutrient conditions, including grasses.[69] Field margins where intensive agriculture prevails tend to become simplified and dominated by grasses because both herbicides and fertilizers tend to suppress broad-leaved species.[21] Grasses are wind pollinated and, as a consequence, pollinators are reduced or eliminated.

A plethora of unforeseen events and interactions can occur when herbicides are released in the environment. Measuring phytotoxicity is a first step for assessing unwanted effects in the environment.

Measuring Phytotoxicity

Herbicides (and other pesticides) are special contaminants because they are intentionally released in the environment. Therefore, before they are registered for use, risk assessment of herbicide phytotoxicity is conducted using data submitted by registrants. In current guidelines, phytotoxicity tests designed to assess the impact to terrestrial plants are conducted under controlled greenhouse conditions, using crop plants growing individually in pots, and effects are assessed at the juvenile stage, usually before reproduction occurs.[87,88] These tests are used for risk assessment of native plants growing within communities exposed to variable outdoor conditions at various phenological stages (Figure 2). Therefore, there is a large gap in regulatory testing between actual phytotoxicity tests performed with selected non-native species under artificial conditions and risk assessment conducted for protection of native plants in natural habitats. Ecological risk assessment for effects on native plants usually takes into account both herbicide use patterns and herbicide characteristics. Elements that are typically not considered are biotic and abiotic factors, which will affect plant sensitivity to herbicides. Current phytotoxicity testing assumes similar effects at different phenological stages and does not take into account recurrent exposure of sublethal doses.

In the past, phytotoxicity testing was designed for the protection of agronomically relevant species from accidental herbicide drift; therefore, crop species were appropriately used in testing. However,

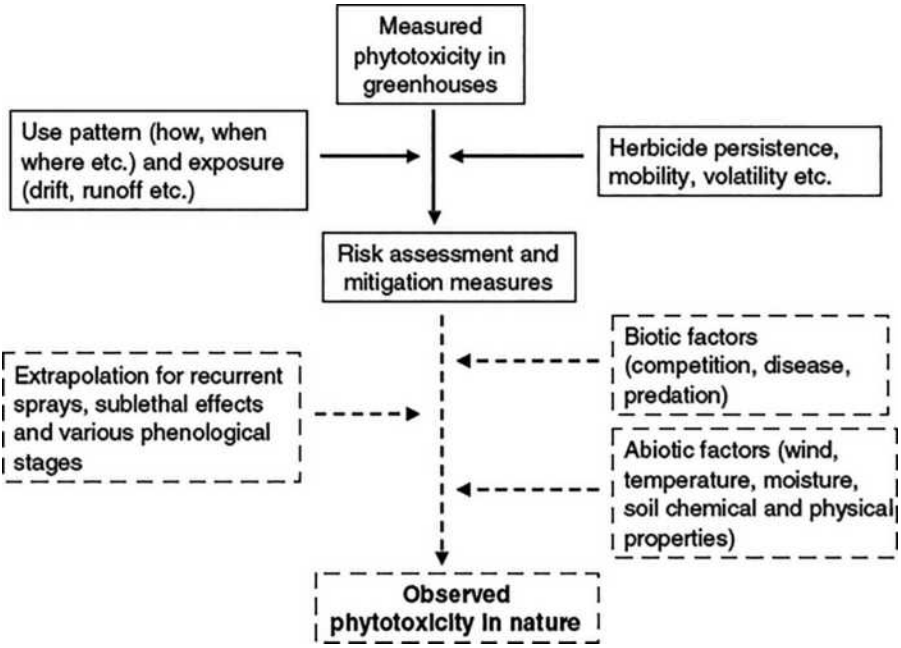


FIGURE 2 Schematic representation of measured and observed herbicide phytotoxicity testing. Phytotoxicity is measured following routine test guidelines. Subsequent ecological risk assessment for effects on native plants usually takes into account herbicide use pattern such as how (aerial or ground applications), when (time of the year and day), and where (type of terrains) herbicides are applied, which will determine to some extent exposure through drift, runoff, or overspray. Herbicide persistence, mobility, and volatility will increase exposure. Elements represented in dashed boxes are usually not considered and determined the observed phytotoxicity under natural conditions. Biotic factors include competition or interactions with other plant species, effects of diseases, and predation on plant health. Wind, low or high temperature, drought or flooding, and soil properties may affect plants and modify sensitivity to herbicides. Phytotoxicity testing assumes similar effects at different phenological stages and does not take into account recurrent exposure of sublethal doses.

it has been increasingly recognized that the vegetation bordering or in the vicinity of crop fields was important for the native plant species and for the other trophic levels relying on primary producers for food, habitat, and shelter. In studies aimed at comparing crop and noncrop species, results indicated that plant sensitivity is both herbicide and species dependent and that no obvious pattern emerged with numerous native and crop plant species tested with a number of herbicides.[89,90] For some herbicides, results with crops would mean underprotection of native species, while for other herbicides, testing with crop species would entail overprotection. However, these studies were limited to a narrow taxonomic range in line with the crop species usually tested, e.g., short-lived species. It was also found that the numerous native species selected for the studies were, for the most part, easy to grow and maintain in the greenhouse. These results suggest that testing should cover a broader range of native species in toxicity testing for regulatory purposes.

Further results indicated that a large variability in herbicide response existed among crop varieties and native plant ecotypes.[89,91,92] The number of species needed to be tested to encompass the range of sensitivity of a given herbicide has not been resolved. Boutin and Rogers[93] showed that the range of species sensitivity increases with an augmentation of numbers of species tested up to 40, suggesting that the number of species ($n = 6$ or 10) tested in current guidelines is insufficient. These studies and others indicated that the large emphasis placed on collecting very precise EC_{25} or EC_{50} values (effective concentration or dose causing a 25% or 50% effect on test plants) may be problematic and erroneous and that extrapolation factors should be determined in order to alleviate some of the uncertainty in testing that is currently not considered.

Tests in current guidelines are required for plants at the two- to four-leaf stage, the assumption being that the juvenile stage is more sensitive than the adult period. However, natural vegetation bordering crop fields consists of plants at various phenological stages. It was found in several studies that indeed the young stage is very sensitive to herbicides.[66,94] Conversely, plants at later stages can show higher sensitivity to some herbicides.[54,94]

For practical reasons, phytotoxicity tests are conducted under ideal greenhouse conditions using plants growing singly in pots and devoid of intra- or interspecific competition. This contrasts with plants growing within natural communities subjected to the occasional drought or flood, continuous wind, attacks from herbivores or pathogens, and other abiotic or biotic elements that interfere with herbicide effects. It is of course impossible and impractical to study the interactions of all these factors. In the past, studies comparing greenhouse and field tests elicited contradictory results with species being more or less sensitive when tested in greenhouses as compared to field tests.[53,69,95–97] A few studies have examined the direct effect of herbicide use of communities of native plants[21,64–67] (Figure 1), and although responses were extremely variable, most species showed some effects. There is considerable uncertainty in current phytotoxicity testing, and further studies are required to monitor herbicide effects on terrestrial native plants.[98]

Conclusion

Herbicides are used extensively, mostly in agriculture, with sometimes unforeseen environmental effects in non-target areas. Although measuring phytotoxicity can be straightforward when conducted on known crop plants under simple laboratory conditions (in greenhouses), determining phytotoxicity is a multifaceted affair as demonstrated by the numerous factors interacting under natural conditions (Figure 2). The linkage between measured phytotoxicity in greenhouses and observed phytotoxicity in nature implies the inclusion of environmental biotic and abiotic factors as well as considerations of test conditions, herbicide characteristics, level of exposure, and use pattern (Figure 2). It was also shown that phytotoxicity can have consequences at all trophic levels, and this is seldom taken into account.

Ecological and agronomic needs are often conflicting. Weed plants want to be suppressed in favor of crops, yet they are necessary to sustain some wildlife in otherwise desolate agrarian systems. This wildlife can in turn provide vital ecological services (e.g., pollinators for crops or biological control).

Mitigation measures to reduce phytotoxicity on desirable plants, including the elimination of aerial application, implementation of unsprayed buffer zones, more advanced equipment, and avoidance of adverse spray conditions (e.g., high wind, temperature inversion, or forecast rain), can substantially reduce herbicide impacts on the environment and are increasingly considered by the farming communities and regulators. Nevertheless, there appears to be an excessive dependence on herbicides for the control of weeds. New uses for herbicides are still being developed, including the desiccation of crops at the end of the growing season with foliar contact herbicides to facilitate harvest or the use of herbicide-resistant, genetically modified crops. Alternatives to herbicides exist and are widely used in organic farming. They comprise mechanical weed control using tillage or mowing, non-mechanical weed control through crop rotation or the planting of companion crops, and biological weed control using parasites, predators, or pathogens. Another complementary measure that can be promoted in conventional farming involves Integrated Pest Management (IPM), a management system that uses all available tactics to manage pest and weeds including crop rotation, host-resistant varieties, mechanical and physical controls, and chemical control. Weed control in agriculture should and can be made compatible with wildlife values.

Acknowledgments

The author wishes to thank D. Carpenter for useful comments on the manuscript and B. Daigneault for assistance in formatting. Funding was provided by Environment Canada (including Pesticide Science Fund) for the experimental work presented by the author and collaborators.

References

1. Zimdahl, R.L. *Fundamentals of Weed Science,* 3rd Ed.; Academic Press: San Diego, CA, 2008.
2. Matsumoto, M.; Hirose, A.; Ema, M. Review of testicular toxicity of dinitrophenolic compounds, 2-sec-butyl-4, 6-dinitrophenol, 4,6-dinitro-*o*-cresol and 2,4-dinitrophenol. Reprod. Toxicol. **2008**, *26* (3–4),185–190.
3. Duke, S.O. Overview of herbicide mechanisms of action. Environ. Health Perspect. **1990**, *87,* 263–271.
4. Walsh, M.J.; Maguire, N.; Powles, S.B. Combined effects of wheat competition and 2,4-D amine on phenoxy herbicide resistant *Raphanus raphanistrum* populations. Weed Res. **2009**, *49,* 316–325.
5. Anonymous. Common and chemical names of herbicides approved by the Weed Science Society of America. Weed Sci. **2010**, *58* (4), 511–518.
6. Brimble, S.; Bacchus, P.; Caux, P.-Y. Pesticide utilization in Canada: A compilation of current sales and use data, GoC Cat. No. En4–56/2005E; Environment Canada: Gatineau, QC, 2005.
7. Eberwine, J.W.; Hagood, E.S. Effect of johnsongrass (*Sorghum halepense*) control on the severity of virus diseases of corn (*Zea mays*). Weed Technol. **1995**, *9*, 73–79.
8. United States Environmental Protection Agency, 2001, available at http://www.epa.gov/pesticides/pestsales/ (accessed May 2011).
9. IRENA (Indicators Reporting on the Integration of Environmental Concerns into Agricultural Policy). Fact Sheet no. 9. Consumption of Pesticides, 1999.
10. Eurostat. Statistical books. The use of plant protection products in the European Union. Data 1992–2003. Statistical books. 2007 edition; 222 pp.
11. Wilson, C.; Tisdell, C. Why farmers continue to use pesticides despite environmental, health and sustainability costs. Economics, Ecology and the Environment, Working Paper no. 53; School of Economics, University of Queensland: Brisbane, Australia, 2000.
12. Statistics Canada. Census of Agriculture. Ministry of Supply and Services Canada, 1987.
13. National Research Council. Alternative Agriculture National Research Council: Washington, DC, 1989; 448 pp.

14. Olszyk, D.; Burdick, C.; Pfleeger, T.; Lee, E.H.; Watrud, L.S. Assessing the risks to non-target plants from herbicides. Jpn. J. Agric. Meteorol. **2004**, *60,* 221–242.
15. Foley, J.A.; DeFries, R.; Asner, G.P.; Barford, C.; Bonan, G.; Carpenter, S.R.; Chapin, F.S.; Coe, M.T.; Daily, G.C.; Gibbs, H.K.; Helkowski, J.H.; Holloway, T.; Howard, E.A.; Kucharik, C.J.; Monfreda, C.; Patz, J.A.; Prentice, I.C.; Ramankutty, N.; Snyder, P.K. Global consequences of land use. Science **2005**, *309* (5734), 570–574.
16. Cobham, R.; Rowe, J. Evaluating the wildlife of agricultural environments: An aid to conservation. In *Wildlife Conservation Evaluation;* Usher, M.B., Ed.; Chapman and Hall Ltd.: London, 1986; 223–246.
17. Available at http://www.fraseramerica.org/commerce.web/product_files/Environmental_Indicators_2002.pdf (accessed February 2011).
18. de Jong, F.M.W.; de Snoo, G.R.; van de Zande, J.C. Estimated nationwide effects of pesticide spray drift on terrestrial habitats in the Netherlands. J. Environ. Manage. **2008**, *86* (4), 721–730.
19. Available at http://en.wikipedia.org/wiki/EcozonesofCanada (accessed February 2011).
20. Pfleeger, T.G.; Olszyk, D.; Burdick, C.; King, G.; Kern, J.; Fletcher, J. Using a geographic information system to identify areas with potential for off target pesticide exposure. Environ. Toxicol. Chem. **2006**, *25* (8), 2250–2259.
21. Boutin, C.; Jobin, B. Intensity of agricultural practices and effects on adjacent habitats. Ecol. Appl. **1998**, *8* (2), 544–557.
22. Ware, G.W.; Cahill, W.P.; Gerhardt, P.D.; Witt, J.M. Pesticide drift, IV. On target deposits from aerial application of insecticides. J. Econ. Entomol. **1970**, *63* (6), 1982–1983.
23. Davis, B.N.K.; Williams, C.T. Buffer zone widths for honeybees from ground and aerial spraying of insecticides. Environ. Pollut. **1990**, *63* (3), 247–259.
24. Conacher, J.; Conacher, A. Herbicides in Agriculture: Minimum Tillage, Science and Society, Geowest No. 22; Department of Geography, University of Western Australia: Nedlands, WA, 1986.
25. Franzaring, J.; Kempenaar, C.; van der Eerden, L.J.M. Effects of vapours of chlorpropham and ethofumesate on wild plant species. Environ. Pollut. **2001**, *114* (1), 21–28.
26. Follak, S.; Hurle, K. Effect of airborne bromoxyniloctanoate and metribuzin on non-target plants. Environ. Pollut. **2003**, *126* (2), 139–146.
27. Yao, Y.; Harner, T.; Blanchard, P.; Tuduri, L.; Waite, D.; Poissant, L.; Murphy, C.; Belzer, W.; Aulagnier, F.; Sverko, E. Pesticides in the atmosphere across Canadian agricultural regions. Environ. Sci. Technol. **2008**, *42* (16), 5931–5937.
28. Welch, H.E.; Muir, D.C.G.; Billeck, B.N.; Lockhart, W.L.; Brunskill, G.J.; Kling, H.J.; Olson, M.P.; Lemoine, R.M. Brown snow: A long-range transport event in the Canadian Arctic. Environ. Sci. Technol. **1991**, *25* (2), 280–286.
29. Hoferkamp, L.; Hermanson, M.H.; Muir, D.C.G. Current use pesticides in Arctic media; 2000–2007. Sci. Total Environ. **2010**, *408* (15), 2985–2994.
30. Eilsler, R. *Paraquat Hazards to Fish, Wildlife, and Invertebrates: A Synoptic Review,* Biological Report 85(1.22); U.S. Fish and Wildlife Service, 1990; 1–38.
31. Andreotti, G.; Freeman, L.E.B.; Hou, L.; Coble, J.; Rusiecki, J.; Hoppin, J.A.; Silverman, D.T.; Alavanja, M.C.R. Agricultural pesticide use and pancreatic cancer risk in the agricultural health study cohort. Int. J. Cancer **2009**, *124* (10), 2495–2500.
32. Shorr, R.I.; Ray, W.A.; Daugherty, J.R.; Griffin, M.R. Incidence and risk factors for serious hypoglycemia in older persons using insulin or sulfonylureas. Arch. Intern. Med. **1997**, *157* (15), 1681–1686.
33. Anderson, R.J.; Prahlad, K.V. The deleterious effects of fungicides and herbicides on *Xenopus laevis* embryos. Arch. Environ. Contam. Toxicol. **1976**, *4* (1), 312–323.
34. Mann, R.M.; Hyne, R.V.; Choung, C.B.; Wilson, S.P. Amphibians and agricultural chemicals: Review of the risks in a complex environment. Environ. Pollut. **2009**, *157* (11), 2903–2927.

35. Kelly, D.W.; Poulin, R.; Tompkins, D.M.; Townsend, C.R. Synergistic effects of glyphosate formulation and parasite infection on fish malformations and survival. J. Appl. Ecol. **2010**, *47* (2), 498–504.
36. Feng, P.C.C.; Clark, C.; Andrade, G.C.; Balbi, M.C.; Caldwell, P. The control of Asian rust by glyphosate in glyphosate-resistant soybeans. Pest Manage. Sci. **2008**, *64* (4), 353–359.
37. Darwish, A.M.; Mitchell, A.J. Evaluation of diquat against an acute experimental infection of *Flavobacterium columnare* in channel catfish, *Ictaluruspunctatus* (Rafinesque). J. Fish Dis. **2009**, *32* (5), 401–408.
38. Thomson, W.T. *Agricultural Chemicals. Book 1 Insecticides*; Thomson Publications: Fresno, CA, 1985–1986.
39. van Lersel, M.W.; Bugbee, B. Phytotoxic effects of benzimidazole fungicides on bedding plants. J. Am. Soc. Hortic. Sci. **1996**, *121* (6), 1095–1102.
40. Heitefuss, P. Assessment of herbicide effects on interactions of weeds, crop plants, pathogens and pests. In *Field Methods for the Study of Environmental Effects of Pesticides, BCPC Monograph No. 40;* Greaves, M.P., Greig-Smith, P.W., Smith, B.D., Eds.; British Crop Protection Council: Thornton Heath, U.K., 1988; 265–274.
41. Adams, J.B.; Drew, M.E. Grain aphids in New Brunswick. III. Aphid populations in herbicide-treated oat fields. Can. J. Zool. **1965**, *43,* 789–794.
42. Ries, S.K. Subtoxic effects on plants. In *Herbicides;* Audus, L.J., Ed.; Academic Press: London, 1976; Vol. 2, 313–344.
43. Oka, I.N.; Pimentel, D. Herbicide (2,4-D) increases insect and pathogen pests on corn. Science **1976**, *193* (4249), 239–240.
44. Ishii, S.; Hirano, C. Growth responses of larvae of the rice stem borer to rice plants treated with 2,4-D. Entomol. Exp. Appl. **1963**, *6,* 257–262.
45. Schnelle, M.A.; Hensley, D.L. Effects of pesticides upon nitrogen fixation and nodulation by dry bean. Pestic. Sci. **1990**, *28* (1), 83–88.
46. Garcia-Romera, I.; Ocampo, J.A. Effect of the herbicide MCPA on VA mycorrhizal infection and growth of *Pisum sativum.* J. Plant Nutr. Soil Sci. **1988**, *151* (4), 225–278.
47. Fox, W.B. 2,4-D as a factor in increasing wireworm damage to wheat. Sci. Agric. **1948**, *28* (9), 423–424.
48. Rinella, M.J.; Haferkamp, M.R.; Masters, R.A.; Muscha, J.M.; Bellows, S.E.; Vermeire, L.T. Growth regulator herbicides prevent invasive annual grass seed production. Invasive Plant Sci. Manage. **2010**, *3* (1), 12–16.
49. Rinella, M.J.; Masters, R.A.; Bellows, S.E. Growth regulator herbicides prevent invasive annual grass seed production under field conditions. Rangeland Ecol. Manage. **2010**, *63* (4), 487–490.
50. Blackburn, L.G.; Boutin, C. Subtle effects of herbicide use in the context of genetically modified crops: A case study with glyphosate (Roundup®). Ecotoxicology **2003**, *12,* 271–285.
51. Guo, S.L.; Jiang, H.W.; Fang, F.; Chen, G.Q. Influences of herbicides, uprooting and use as cut flowers on sexual reproduction of *Solidago canadensis*. Weed Res. **2009**, *49* (3), 291–299.
52. Rokich, D.P.; Harma, J.; Turner, S.R.; Sadler, R.J.; Tan, B.H.; Fluazifop-*p*-butyl herbicide: Implications for germination, emergence and growth of Australian plant species. Biol. Conserv. **2009**, *142* (4), 850–869.
53. Riemens, M.M.; Dueck, T.; Kempenaar, C. Predicting sublethal effects of herbicides on terrestrial non-crop plant species in the field from greenhouse data. Environ. Pollut. **2008**, *155* (1), 141–149.
54. Fletcher, S.J.; Pfleeger, T.G.; Hillman, C.R. Potential environmental risks associated with the new sulfonylurea herbicides. Environ. Sci. Technol. **1993**, *27* (10), 2250–2252.
55. Boutin, C.; Elmegaard, N.; Kjaer, C. Toxicity testing of fifteen non-crop plant species with six herbicides in a greenhouse experiment: Implications for risk assessment. Ecotoxicology **2004**, *13* (4), 349–369.

56. Pfleeger, T.; Olszyk, D.; Plocher, M.; Yilma, S. Effects of low concentrations of herbicides on full-season, field-grown potatoes. J. Environ. Qual. **2008**, *37*, 2070–2082.
57. Kjær, C.; Strandberg, M.; Erlandsen, M. Metsulfuron spray drift reduces fruit yield of hawthorn (*Crataegus monogyna* L.). Sci. Total Environ. **2006**, *356*, 228–234.
58. Kjær, C.; Strandberg, M.; Erlandsen, M. Effects on hawthorn the year after simulated spray drift. Chemosphere **2006**, *63* (5), 853–859.
59. Eberlein, C.V.; Guttieri, M.J. Potato (*Solanum tuberosum*) response to simulated drift of imidazolinone herbicides. Weed Sci. **1994**, *42* (1), 70–75.
60. Tottman, D.R.; Davies, E.L.P. The effect of herbicides on the root system of wheat plants. Ann. Appl. Biol. **1978**, *90* (1), 93–99.
61. Ratsch, H.C.; Johndro, D.J.; McFarlane, J.C. Growth inhibition and morphological effects of several chemicals in *Arabidopsis thaliana* (L.) Heynh. Environ. Toxicol. Chem. **1986**, *5* (1), 55–60.
62. Breeze, V.G.; van Rensburg, E. Vapour of the free acid of the herbicide 2,4-D is toxic to tomato and lettuce plants. Environ. Pollut. **1991**, *72* (4), 259–267.
63. Marrs, R.H.; Frost, A.J. A microcosm approach to the detection of the effects of herbicide spray drift in plant communities. J. Environ. Manage. **1997**, *50* (4), 369–388.
64. Marshall, E.J.P.; Bernie, J.E. Herbicide effects on field margin flora. Proceedings of the BCPC International Congress—Weeds; British Crop Protection Council: Farnham, Surrey, U.K., 1985; 1021–1028.
65. Marrs, R.H.; Williams, C.T.; Frost, A.J.; Plant, R.A. Assessment of the effects of herbicide spray drift on a range of plant species of conservation interest. Environ. Pollut. **1989**, *59* (1), 71–86.
66. Marrs, R.H.; Frost, A.J.; Plant, R.A. Effects of herbicide spray drift on selected species of nature conservation interest: The effects of plant size and surrounding vegetation structure. Environ. Pollut. **1991**, *69* (2–3), 223–235.
67. Jobin, B.; Boutin, C.; DesGranges, J.-L. Effects of agricultural practices on the flora of hedgerows and woodland edges in Southern Québec. Can. J. Plant Sci. **1997**, *77*, 293–299.
68. Marshall, E.J.P. The impact of landscape structure and sown grass margin strips on weed assemblages in arable crops and their boundaries. Weed Res. **2009**, *49* (1), 107–115.
69. Kleijn, D.; Verbeek, M. Factors affecting the species composition of arable field boundary vegetation. J. Appl. Ecol. **1997**, *37* (2), 256–266.
70. Sotherton, N.W.; Dower, J.W.; Rands, N.R.W. The effects of pesticide exclusion strips on faunal populations in Great Britain. Ecol. Bull. **1988**, *39*, 197–199.
71. Potts, G.R. The effects of modern agriculture, nest predation and game management on the population ecology of partridges (*Perdix perdix* and *Alectoris rufa*). Adv. Ecol. Res. **1980**, *11*, 2–79.
72. Lagerlöf, J.; Stark, J.; Svensson, B. Margins of agricultural fields as habitats for pollinating insects. Agric., Ecosyst. Environ. **1992**, *40*, 117–124.
73. Longley, M.; Sotherton, N.W. Factors determining the effects of pesticides upon butterflies inhabiting arable farmland. Agric., Ecosyst. Environ. **1997**, *61* (1), 1–12.
74. Holzschuh, A.; Steffan-Dewenter, I.; Kleijn, D.; Tscharntke, T. Diversity of flower-visiting bees in cereal fields: effects of farming system, landscape composition and regional context. J. Appl. Ecol. **2007**, *44* (1), 41–49.
75. Boutin, C.; Baril, A.; McCabe, S.K.; Martin, P.A.; Guy, M. The value of woody hedgerows for moth diversity on organic and conventional farms. Environ. Entomol. **2011**, *40* (3): 560–569.
76. Lazorko-Connon, S.; Achari, G. Atrazine: Its occurrence and treatment in water. Environ. Rev. **2009**, *17*, 199–214.
77. Jablonowski, N.D.; Köppchen, S.; Hofmann, D.; Schäffer, A.; Burauel, P. Persistence of ^{14}C-labeled atrazine and its residues in a field lysimeter soil after 22 years. Environ. Pollut. **2009**, *157* (7), 2126–2131.
78. Graymore, M.; Stagnitti, F.; Allinson, G. Impacts of atrazine in aquatic ecosystems. Environ. Int. **2001**, *26*, 483–495.

79. Poster, J. Command herbicide. The rookie battles controversy. Crops Soils Mag. **1986**, *39* (1), 9–11.
80. Eagle, D.J. Agrochemical damage to U.K. crops. Pestic. Outlook **1990**, *1* (2), 14–16.
81. Nilsson, H.; Aamisepp, A. Persistence in plants and transfer of clopyralid (3,6-dichloropicolinic acid) through plant remains. In *Weeds and Weed Control,* 25th Swedish Weed Conference, Vol. 1 Reports. Uppsala, Sweden, 1984.
82. http://www.pesticides.gov.uk/garden.asp?id=2480 and http://www.guardian.co.uk/environment/2008/jun/29/food.agriculture (accessed May 2011).
83. Moyer, J.R.; Esau, R.; Kozub, G.C. Chlorsulfuron persistence and response of nine rotational crops in alkaline soils of southern Alberta. Weed Technol. **1990**, *4* (3), 543–548.
84. Blair, A.M.; Martin, T.D. A review of the activity, fate and mode of action of sulfonylurea herbicides. Pestic. Sci. **1988**, *22* (3), 195–219.
85. Walker, A. Simulation of herbicide persistence in soil. II. Simazine and linuron in long-term experiments. Pestic. Sci. **1976**, *7* (1), 50–58.
86. Rew, L.J.; Theaker, A.J.; Froud-Williams, R.J.; Boatman, N.D. Nitrogen fertilizer misplacement in field boundaries. Aspects Appl. Biol. **1992**, *30,* 203–206.
87. Organisation for Economic Co-operation and Development (OECD). Terrestrial Plant Test: Vegetative Vigour Test. New TG 227. Seedling Emergence and Seedling Growth Test. Updated TG 208, 2006.
88. United States Environmental Protection Agency (USEPA). Ecological Effects Test Guidelines: Terrestrial Plant Toxicity—Vegetative Vigor, OPPTS 850.4250, EPA 712-C-96- 163 and Early Seedling Toxicity Test, OPPTS 850.4130, EPA 712–C–96–347. Washington, DC, 1996.
89. White, A.L.; Boutin, C. Herbicidal effects on non-target vegetation: Investigating the limitations of current pesticide registration guidelines. Environ. Toxicol. Chem. **2007**, *26* (12), 2634–2643.
90. Carpenter, D.; Boutin, C. Sublethal effects of the herbicide glufosinate ammonium on crops and wild plants: Short-term effects compared to vegetative recovery and plant reproduction. Ecotoxicology **2010**, *19* (7), 1322–1336.
91. Boutin, C.; White, A.L.; Carpenter, D. Measuring variability in phytotoxicity testing using crop and wild plant species. Environ. Toxicol. Chem. **2010**, *29* (2), 237–242.
92. Bhatti, M.A.; Felsot, A.S.; Al-Khatib, K.; Kadir, S.; Parker, R. Effects of simulated chlorsulfuron drift on fruit yield and quality of sweet cherries (*Prunus avium* L.). Environ. Toxicol. Chem. **1995**, *14* (3), 537–544.
93. Boutin, C.; Rogers, C.A. Pattern of sensitivity of plant species to various herbicides—An analysis with two databases. Ecotoxicology **2000**, *9* (4), 255–272.
94. Boutin, C.; Lee, H.B.; Peart, T.E.; Batchelor, S.P.; Maguire, R.J. Effects of the sulfonylurea herbicide metsulfuron methyl on growth and reproduction of five wetland and terrestrial plant species. Environ. Toxicol. Chem. **2000**, *19* (10), 2532–2541.
95. Dalton, R.L.; Boutin, C. Comparison of the effects of glyphosate and atrazine herbicides on nontarget plants grown singly and in microcosms. Environ. Toxicol. Chem. **2010**, *29* (10), 2304–2315.
96. Damgaard, C.; Mathiassen, S.K.; Kudsk, P. Modeling effects of herbicide drift on the competitive interactions between weeds. Environ. Toxicol. Chem. **2008**, *27* (6), 1302–1308.
97. Clark, J.; Ortego, L.S.; Fairbrother, A. Sources of variability in plant toxicity testing. Chemosphere **2004**, *57* (11), 1599–1612.
98. Boutin, C.; Aya, K.L.; Carpenter, D.; Thomas, P.J.; Rowland, O. Phytotoxicity testing for herbicide regulation: shortcomings in relation to biodiversity and ecosystem services in agrarian systems. Sci. Total Environ. **2011**, *40* (3): 560–569.

5

Insecticides: Aerial Ultra-Low-Volume Application

He Zhong

Introduction

Mosquito control is necessary in order to protect public health from mosquito-borne diseases such as West Nile virus, eastern equine encephalitis, St. Louis encephalitis, malaria, and dengue. Aerial ultra-low-volume (ULV) application of mosquito insecticides is one of the most effective techniques for controlling adult mosquitoes and preventing mosquito-borne diseases.[1] During ULV application, large insecticide droplets may sometimes result in unwanted mortality to nontarget organisms.[2–8] For many years, nontarget mortality caused by mosquito adulticiding was usually accepted as a "casualty of war." With the advancement of mosquitocide residue monitoring and new spray technologies, we now recognize that "casualties" can be reduced to a minimum. The important factors that contribute to mosquito control efficacy and associated nontarget mortality are insecticide deposition,[7–9] droplet size,[1,10] spray time,[10] application dose,[1,10] topography,[10] and weather conditions (such as wind velocity, direction, and atmospheric stability).[10] Due to the complexity of aerial ULV applications, it may be difficult to achieve optimal mosquito reduction without nontarget mortality during an aerial spray mission. However, control efficacy can be increased, and nontarget mortality can be minimized, if aerial application is conducted at the right place (by increasing retention time of mosquitocide droplets in the air in order to enhance their contact with flying mosquitoes),[11,12] at the right time (dusk, or night when adult mosquitoes are actively flying),[10] and at the right dose (proper insecticide concentration in the air to kill mosquitoes but not nontarget organisms).[1,9]

Right Place

Current aerial spray technology is capable of targeting a specific zone where adult mosquitoes are actively flying. The droplet size of aerially applied insecticides governs downwind dispersal and subsequent impingement on targets.[1] Smaller droplets [5–25 pm volume median diameter (VMD)] remain aloft longer and offer a better probability for impingement on flying adult mosquitoes.[1] Larger insecticide droplets (more than 100 pm VMD) will deposit on the ground sooner after application, thereby reducing the likelihood of contact with flying mosquitoes. Insecticide that is deposited on the ground not only is wasted but also may adversely affect nontarget organisms.[7,8] Environmental contamination can be reduced by adopting application techniques that maintain droplets in the air and promote controlled downwind movement of the insecticide cloud while minimizing ground deposition (particularly in environmentally sensitive areas).[7,9] This concept is different from agricultural applications, where insecticide deposition is needed to coat the surface of crops. Moreover, agricultural applications try to reduce insecticide drift away from the target zone, such as a crop or field, rather than maximize droplet suspension in the air column.

Right Time

The best time of the day for mosquito adulticide applications (also called the "spray window") is at dusk, dawn, or nighttime when mosquitoes are actively flying, and this is when the optimal control efficacy will be achieved.[1,10] Applying an insecticide in that period will reduce impact on daytime active nontarget organisms like honeybees, dragonflies, and butterflies. These spray windows are the time periods when daytime nontargets are resting and, therefore, protected from insecticide exposure. Mosquito control operation in such spray windows will protect many daytime active nontargets from being adversely impacted by insecticide dispersal clouds.

Spraying at the right time also means spraying under optimal meteorological conditions, currently recognized as 3–10 miles/hr wind velocity without the presence of a temperature inversion. Understanding, as well as achieving, "optimal" meteorological conditions for mosquito spraying is often difficult.[1] Although aerial mosquito application technology can accurately calibrate the amount of insecticide delivered using the nozzle systems. Once released, the aerosol is in the hands of Mother Nature. The spray cloud, as it is carried by wind and influenced by gravity, starts its journey to the ground from an altitude of 100300 ft. Wind velocity, direction, and atmospheric stability greatly affect the distribution of the spray cloud.[1] Also, downwind movement and deposition of insecticide residue can vary greatly from one spray mission to another.[11,12] As wind speed increases, the impingement force of spray droplets will increase on targets and nontargets. This situation creates considerable variation in control efficacy and can often influence whether the effects on nontargets are minimal or substantial.[13]

Right Dose

Increasing the application rate may not always improve the level of control but will generally increase the risk of nontarget mortality due to escalating exposure levels. In reality, it is sometimes very difficult to apply the proper application rate to achieve adequate mosquito control without causing nontarget mortality. Nontargets' differential tolerance to insecticides may be the result of physiological as well as geographical differences within and among those organisms. The larger body size of some nontarget species may increase their tolerance levels to the insecticide. Natural topographic barriers, such as trees, bushes, and grasses, may provide refuge for them to escape lethal exposure from the insecticide aerosol.[8] Conversely, adult mosquitoes in vegetated areas may also be protected by the physical barriers,[1] and if the application rate is increased to compensate for this, adverse effects to nontargets may occur.

To determine the proper application rate or "right dose," studies are needed to determine the correlation of the insecticide concentration in the air column with adult mosquito mortality[11,12] and that of ground deposition concentrations with nontarget mortality.[7–9] The insecticide concentration at its final destination, whether at an airborne target or on the ground, is referred to as the terminal insecticide concentration (TIC), which is different from and can be influenced by the initial application dose. TIC is influenced by many environmental variables and therefore needs to be frequently monitored. If the TIC is adequate to kill the majority of adult mosquitoes and low enough that it spares nontargets of concern, the application dose will be considered appropriate. In this way, TIC critically affects control efficacy and nontarget impact. Determination of TIC during routine application of mosquitocides provides a mechanism to assess or cross-compare control efficacy and impact on nontargets during aerial mosquito control missions. This process will ensure the proper application dose to achieve the delicate balance between effective mosquito control and minimal nontarget impact.

Novel Application Technology

Mosquito control programs worldwide continue to develop novel application technologies to increase control efficacy and lessen damage to nontargets. In the late 1990s, James Robinson's group at Florida's Pasco County Mosquito Control District led an effort to develop a high-pressure nozzle system to deliver small insecticide droplets (<30 μm VMD).[11,12] The high-pressure system, with insecticide applied at two-thirds the label rate, achieved better adult mosquito control compared with a conventional flat-fan nozzle system (>80 μm VMD)[11] and dramatically reduced the mortality of caged fiddler crabs, *Uca pugilator* (Bosc), from 80% to 0 in Collier County, Florida.[7] Another field trial with honeybees (*Apis mellifera* L.) conducted during routine nighttime aerial adult mosquito control missions in Manatee County, Florida, further demonstrated the advantage of the new high-pressure system. Honeybees that clustered outside of beehive entrances were exposed to naled sprays for mosquito control. The larger insecticide droplets produced by flat-fan nozzles killed more than 90% of the bees clustered outside of the hives and resulted in an average of 35% reduction in honey yield at the end of the season.[8] In contrast, bee mortality and average honey yield were not significantly different compared with hives similarly exposed to the smaller droplets of the high-pressure nozzle system.[9] Several spray nozzles such as air-assisted nozzles, high-pressure nozzles, and Micronair nozzles— capable of delivering smaller spray droplets (<30 μm VMD) for adult mosquito control—are now available.

Insecticide Residue Monitoring

At present, bioassay techniques are widely used to measure mosquito control efficacy and nontarget organism impact. Generally, bioassays only answer "yes or no" (i.e., dead or alive) and are primarily acute toxicity tests. However, bioassays do not address critical issues such as where or how much insecticide is present following the ULV application. Insecticide residue monitoring is encouraged because it can identify TICs following mosquito control application. Research shows that the level of TIC is critically reflective of mosquito control efficacy and nontarget impacts based on the dose–response relationship. When TIC is used in conjunction with bioassays, mosquito control programs will have a powerful quantitative tool to combat mosquitoes while protecting the environment.[7–9,11,12] Insecticide residues cannot be observed by the human eye. However, they can be detected and quantified by modern analytical techniques such as gas chromatography or high-performance liquid chromatography. By monitoring insecticide residues, we can 1) determine the insecticide concentrations in the air and use these data to facilitate adjustments to the spray system to attain the right application dose; 2) identify any excessive insecticide deposited on the ground or into the water and use these data to optimize the spray system; 3) monitor the distance of aerosol dispersed downwind for an effective spray offset; 4) establish appropriate impact thresholds for nontarget mortality and effective thresholds for mosquito control efficacy; and 5) compare different application equipment or operational scenarios.

Conclusion

In summary, controlling adult mosquitoes through ULV aerial application of insecticides while minimizing nontarget impacts can be achieved by operational mosquito control programs. Instituting a residue monitoring program can be critical in protecting the environment. TIC levels identified from insecticide residue monitoring provide a powerful tool to optimize the application so that proper application of insecticides can be achieved at the right time, at the right place, and at the right dose.

Acknowledgments

I thank James Cilek and Cate Brock for their critical reviews, helpful comments, and suggestions.

References

1. Mount, G.A.; Biery, T.L.; Haile, D.G. A review of ultralow-volume aerial sprays of insecticide for mosquito control. J. Am. Mosq. Control. Assoc. **1996**, *12* (4), 601–618.
2. Hill, E.F.; Eliason, D.A.; Kilpatrick, J.W. Effects of ultralow volume application of malathion in Hale County, Texas. J. Med. Entomol. **1971**, *8* (2), 173–179.
3. McKenney, C.L.; Shirley, M.A.; Pierce, R.H. *A Two-Year Research Project in the Rookery Bay National Estuarine Research Reserve to Evaluate Populational and Physiological Responses of Crabs from Various Habitats in Relation to the Fate of a Mosquito Control Pesticide*, Contract No. NC-ND2100–3–00042; 1997; 9.
4. Pankiw, T.; Jay, S.C. Aerially applied ultra-low-volume malathion effects on caged honey bees (Hymenoptera: Apidae), caged mosquitoes (Diptera: Culicidae), and malathion residues. J. Econ. Entomol. **1992**, *85,* 687–691.
5. Pankiw, T.; Jay, S.C. Aerially applied ultra-low-volume malathion effects on colonies of honey bees (Hymenoptera: Apidae). J. Econ. Entomol. **1992**, *85,* 692–699.
6. Schoor, W.P.; Middaugh, D.P.; Marcovich, D.T. Effects of aerially applied fenthion on survival and reproduction of the panacea sand fiddler, *Uca panacea,* in laboratory habitats. Arch. Environ. Contam. Toxicol. **2000**, *38,* 327–333.
7. Zhong, H.; Dukes, J.; Greer, M.; Hester, P.; Shirley, M.; Anderson, B. Ground deposition impact of aerially applied fenthion on the fiddler crabs, *Uca pugilator*. J. Am. Mosq. Control. Assoc. **2003**, *19,* 47–52.
8. Zhong, H.; Latham, M.; Hester, P.G.; Frommer, R.L.; Brock, C. Impact of aerial applied ultra-low-volume insecticide naled with flat-fan nozzle system on honeybee survival and productivity. Arch. Environ. Contam. Toxicol. **2003**, *45,* 216–220.
9. Zhong, H.; Latham, M.; Payne, S.; Brock, C. Minimizing the impact of mosquito adulticide naled on honeybees *Apis mellifera* (Hymenoptera: Apidae): Aerial ultra-low volume application using a high-pressure nozzle system. J. Econ. Entomol. **2004**, *97,* 1–7.
10. Mount, G.A. A critical review of ultralow-volume aerosols of insecticide applied with vehicle-mounted generators for adult mosquito control. J. Am. Mosq. Control. Assoc. **1998**, *14* (3), 305–334.
11. Dukes, J.; Zhong, H.; Greer, M.; Hester, P.; Hogan, D.; Barber, J.A.S. A comparison of two spray nozzle systems used to aerially apply the ultra-low-volume adulticide fenthion. J. Am. Mosq. Control Assoc. **2004**, *20* (1), 27–35.
12. Dukes, J.; Zhong, H.; Greer, M.; Hester, P.; Hogan, D.; Barber, J.A.S. A comparison of two ultra-low-volume spray nozzle systems by using a multiple swath scenario for the aerial application of fenthion against adult mosquitoes. J. Am. Mosq. Control Assoc. **2004**, *20* (1), 36–44.
13. Zhong, H.; Hribar, L.J.; Daniels, J.C.; Feken, M.A.; Brock, C.; Trager, M.D. Aerial ultra-low-volume application of naled: Impact on nontarget imperiled butterfly larvae (*Cyclargus thomasi bethune-bakeri*) and efficacy against adult mosquitoes (*Aedes taeniorhynchus*). Environ. Entomol. **2010**, *39* (6), 1961–1972.

6

Neurotoxicants: Developmental Experimental Testing

Vera Lucia S.S.
de Castro

Introduction

Neurotoxic substances may play a role in a number of neurodevelopmental disorders. They can be released by industrial facilities and by agricultural practices, much of which ends up in the air or groundwater. Since most neuroteratogens affect multiple regions and processes, they can result in various behavioral defects.

Given the importance placed on fostering optimal cognitive development and the fact that chemical exposures can perturb the exquisite spatial and temporal choreography of brain development, it is not surprising that neurodevelopmental deficit frequently serves as the critical adverse health effect in risk assessments.[1]

Detection and characterization of chemical-induced toxic effects in the central and peripheral nervous system represent a big challenge. Prediction of neurotoxic effects is a key feature in the toxicological profile of compounds and is therefore required by many regulatory testing schemes.[2]

Despite the increasing recognition of the need to evaluate developmental neurotoxicity (DNT) in safety assessment, only very few of the commercial chemicals in current use have been examined with respect to neurodevelopmental effects. Validated rodent models exist, but they are considered expensive and are only infrequently used. The neurodevelopmental disorders include learning disabilities, attention deficit hyperactivity disorder, autism spectrum disorders, developmental delays, and emotional and behavioral problems. The causes of these disorders are unclear, and interacting genetic, environmental, and social factors are likely determinants of abnormal brain development.[3,4]

In calculations of environmental burdens of disease in children, lead neurotoxicity to the developing brain is a major contributor. Pesticide effects could well be of the same magnitude, or larger, depending on the exposure levels.[4] For example, an emerging literature provides evidence of neurobehavioral consequences resulting from exposure to relatively low levels of organochlorine and organophosphate pesticides in infants and children.

Organophosphate pesticides continue to be applied widely in agriculture and in residences throughout the world, representing about half the total annual amount of insecticides used. One of the major concerns with these agents is their propensity to elicit DNT at exposures below the threshold for any systemic symptoms, so that potentially damaging fetal or childhood exposures may go undetected until persistent functional impairments become expressed.[5]

Organophosphate-poisoned populations have shown a consistent pattern of deficits when compared to a nonexposed or non-poisoned population on measures of motor speed and coordination, sustained attention, and information processing speed.[6] In experimental models, developing animals have been shown to be more susceptible than adult animals to the acute toxicity of the organophosphate pesticide chlorpyrifos, which can cause neurobehavioral abnormalities.[7] Neonatal diazinon exposure below the threshold for appreciable cholinesterase inhibition in a non-monotonic dose–effect caused persisting neurocognitive deficits in adulthood. The organophosphorous insecticide can affect transmitter systems supporting memory function, differently, implying participation of mechanisms other than their common inhibition of cholinesterase.[5]

Future studies should examine the neurodevelopment effects in human beings associated with pesticide mixtures and other classes of pesticides (e.g., carbamates, pyrethroids), and with pesticide mixtures, because there is increasing use of these pesticides in certain communities that are replacing the organophosphate and organochlorine pesticides.[8] In this direction, the U.S. Agency for Toxic Substances and Disease Registry (ATSDR) developed a program for chemical mixtures of which an integral part is a mixtures health risk assessment. ATSDR has completed evaluations for several simple mixtures of child-specific exposure concern.[9]

Behavioral Aspects of DNT

Behavior represents an integrated response of the nervous system that can reveal functional changes important to the overall fitness and survival of the organism exposed to single pesticides or mixtures. Although some developmental neurotoxicants are structural teratogens as well, behavioral dysfunctions may be more serious than structural defects under certain circumstances.[10] The major developmental sensory systems of concern in toxicology include visual, auditory, olfactory, nociceptive (pain and other noxious stimuli), somatosensory, and vestibular. However, neurobehavioral functions are influenced by subject variables such as age, sex, education, and social and (especially in humans) cultural background.

Brain Development and Maturation

The development and maturation of the mammalian brain is an extremely complex process. Brain development involves cell division, migration and differentiation, programmed cell death (apoptosis), cell-to-cell interactions (e.g., for migration and synaptic communication), and multiple other processes under different timetables for the various brain regions. Genetic, epigenetic, and environmental factors (e.g., exposure to toxic chemicals, including certain heavy metals, industrial chemicals, and pesticides), particularly during the susceptible periods of development and aging, can result in many possible adverse central nervous system (CNS) consequences, ranging from mild to severe and involving various functions (e.g., cognition, motor, or sensory dysfunction). DNT refers to any adverse effect of perinatal exposure to a toxic substance on the normal development of nervous system structure and/or function.[11]

The mammalian brain undergoes a period of rapid brain growth, which in humans occurs perinatally, spanning from the third trimester of pregnancy throughout the first 2 years of life. In rats and mice, the brain growth spurt occurs in the neonate, spanning the first 3–4 weeks of life and reaching its peak around postnatal day 10. This period is characterized by axonal and dendritic outgrowth and the establishment of neuronal connections, and during this period, animals acquire many new motor and sensory abilities. Neurotypic and gliotypic proteins can serve as sensitive indicators of time- and region-specific effects of chemicals on the developing nervous system. The presence of xenobiotics in the brain during this defined period of maturational processes is a critical factor for induction of persistent changes in behavior and transmitter systems.[12] For example, it has been reported that multiple neurotransmitter systems are altered following exposure to organophosphorus insecticides. Developmental chlorpyrifos exposure produces persistent deficiencies in cholinergic synaptic neurochemistry.[13] Also, there is increasing evidence that polychlorinated biphenyls (PCBs) and methyl mercury also have neurotoxic effects. An enhanced effect of these toxicants, due to either synergistic or additive effects, would be considered as a risk for fetal development. It is postulated that these neurotoxicants might interact.[14]

Epidemiological studies have demonstrated a relationship between perinatal exposure to persistent organic pollutants among others and neurological and behavioral disturbances in infants and children. Studies in animals have confirmed that contaminants like PCBs, metals, and pesticides can disrupt behavioral functioning.[15]

Evidence indicates that exposure to environmental chemicals could have an impact on children's health and development. The developing CNS of fetus and children is particularly susceptible to chemically induced damage compared with the brain of adults due to the different pharmacokinetic factors, diminished defense mechanisms, or the fact that the developing nervous system undergoes a highly complex series of ontogenetic processes that are vulnerable to chemical perturbation.[16]

Periods of Vulnerability

The developing nervous system is particularly sensitive to environmental insults during critical periods that are dependent on the temporal and regional emergence of specific and sequential developmental processes (i.e., proliferation, migration, differentiation, synaptogenesis, myelination, and apoptosis). Evidence from numerous sources demonstrates that neural development extends from the embryonic period through adolescence. In general, the sequence of events is comparable among species, although the time scales are considerably different. Developmental exposure of animals or humans to numerous agents (e.g., x-ray irradiation, ethanol, lead, methyl mercury, or chlorpyrifos) demonstrates that interference with one or more of these developmental processes can lead to DNT.[17]

For many behaviors, a critical period exists during which the animal is sensitive to these organizational effects. Functional and structural life-lasting modifications can be induced by alterations of natural conditions during these adaptive developmental stages of maturation. The critical periods have been described for some cortical circuits involved in many different sensory systems such as the auditory, somatosensory, and olfactory systems. These critical periods occur also during postnatal life. Indeed, experience-dependent plasticity during critical periods of postnatal development shapes the adult brain anatomy and function.[18,19]

Protocols for Experimental Studies

In order to reduce the risk regarding the exposure to pollutants, addressing the behavioral aspects by appropriate investigation to sustain the safe use of the compounds is suggested. Evaluation of pre- and postnatal developmental parameters can be improved by including different tests on and safety assessment of chemicals to indicate the proper functioning of the sensory, motor, emotional, and cognitive domains.

Laboratory experimental studies suggest that many currently used pesticides such as organophosphates, carbamates, pyrethroids, ethylenebisdithiocarbamates, and chlorophenoxy herbicides can cause neurodevelopmental toxicity. Adverse effects on brain development can be severe and irreversible.[4]

Emotional processes can be viewed as adaptive events or states that are likely to occur across the animal kingdom, but that may or may not have subjective components, comprising physiological, behavioral, and subjective components, depending on the species and circumstances involved. Although these different components usually act in concert, they are potentially dissociable, not always operating as a functional whole.[20]

The utilization of an experimental protocol containing indices related to reproduction and animal development can identify initial damages due to exposure to environmental pollutants. Behavioral experimental methods are used to detect and characterize developmental neurotoxic effects on sensory, cognitive, and motor system functions. Neurobehavioral evaluations are widely used to examine the potential neurotoxicity of pesticides and other chemicals, [21] since neurobehavioral performance can be a sensitive biomarker of the neurodevelopmental consequences of exposure to environmental agents.

Prevention of possible damages due to pesticides during the development of young organisms, like newborns, requires an integrated strategy capable of monitoring the standard use of these products as well as the integration of the potential effects to improve evaluation. If available, biomarkers of exposure are useful for assessing the bioavailability of toxicants to the dam and offspring in utero and after birth. The evaluation of these biomarkers needs to differentiate normal variability from changes that are adverse in response.[22]

Animal models are used to understand neurophysiological processes on the basis of human exposure to xenobiotics. They represent a basis for understanding their pathophysiological traits. There is a variety of methodologies that can be utilized to assess these processes. Cross-species comparability between human and experimental animals supports the assumption that DNT effects in animals indicate a potential to affect development in humans.[10]

The first guideline specifically designed to evaluate DNT was developed and implemented by the U.S. Environmental Protection Agency (EPA) in 1991 and has later on been updated. The Organization for Economic Cooperation and Development (OECD) initiated the development of a DNT guideline (TG 426) following the recommendations of the OECD Working Group on Reproduction and Developmental Toxicity in Copenhagen in 1995. The first draft based on the U.S. EPA DNT guideline was prepared following a 1996 Expert Consultation Meeting that addressed a number of significant issues and incorporated improvements. The draft TG 426 was distributed to National Coordinators for comments in 1998, and significant technical issues in the comments were further discussed and revised.[10,23]

Developmental toxicity may result from either prenatal or postnatal exposure, may manifest at any life stage, and may be expressed as functional deficits. The DNT study is a specialized type of developmental toxicity study designed to screen for adverse effects of pre- and postnatal exposure on the development and function of the nervous system and to provide dose–response characterizations of those outcomes. The U.S. EPA and OECD DNT guidelines recommend administration of the test substance during gestation and lactation. Cohorts of offspring (typically rat) are randomly selected from control and treated litters for evaluations of gross neurologic and behavioral abnormalities during postnatal development and adulthood. These include assessments of physical development, behavioral ontogeny, motor activity, motor and sensory function, learning and memory, and postmortem evaluation of brain weights and neuropathology.[23]

There are a number of stimulus properties shared by all sensory systems, including intensity, frequency, duration, and location in space. In this way, behavioral tests of motor dysfunction in animals include those used to detect spontaneous movement disorders such as changes in gait, tremors, and myoclonus, and those used to detect changes in induced movement such as reflexes, reactions, and movements under operant control. Tests of motor function include observation of locomotion, measurement of locomotor activity, and tests of reflexes and reactions. Also, assessment of cognitive function is a critical component of a DNT assessment to address concerns over potential long term consequences of exposures to toxicants during brain development. Cognitive function is thought to encompass learning, memory, and attention processes.[24–27]

In this way, more effort is needed to adequately evaluate the neurotoxic effects. More elaborated experimental protocols are continuously proposed. They will focus on the interpretation of data obtained in studies that link xenobiotics exposure and functional (behavioral) deficits due to specific neurotransmitter and synaptic mechanisms,[28] identifying possible chemical class-specific targets and biomarkers of effect. Above all, gene expression could be also used as a sensitive tool for the initial identification of DNT effects induced by different mechanisms of toxicity in both cell types (neuronal and glial) and at various stages of cell development and maturation.[16]

Recent literature have examined specific end points across multiple guideline DNT studies to demonstrate the value of current methods in hazard characterization and explore further opportunities for methodologic refinement, examining the interpretation of neurodevelopmental end points for human health risk assessment, data interpretation and variability, positive control data, and statistical analysis.[11,29,30]

Test method reliability, reproducibility, and relevance are attributable in part to the high level of standardization of the test methods;[23] in some cases, the variability of some end points (e.g., motor activity) is very large. Methods have been suggested to decrease such variability.[29–31] Sources of variability include factors related to environmental conditions, personnel, experimental procedures, and equipment.[21,31] The detection, measurement, and interpretation of DNT effects depend on appropriate study design and execution, using established methods with appropriate controls.

Furthermore, the nature and extent of developmental neurotoxic effects often are dependent on the timing of exposure to a toxic agent or combinations of agents and environmental conditions; i.e., organisms exhibit distinct temporal windows of susceptibility. Variations in neurotoxic outcomes across species are expected because stages of nervous system development can vary significantly between species in relation to the time of birth. Thus, the time and duration of exposure in animal models must also be selected carefully to match the window of exposure in the human situation and allow cross-species extrapolation.[32]

Detection and characterization of chemical-induced toxic effects in the central and peripheral nervous system represent a major challenge for employing newly developed technologies in the field of neurotoxicology. For example, those using specific brain cell types can produce results of general mechanism of action but not specific to the chemical tested. In addition, toxicokinetic models are to be developed in order to properly evaluate absorption, distribution, metabolism, and excretion, as well as the blood–brain barrier. Behavioral toxicologists will be needed to contribute for the experimental tests and computational models to anchor molecular initiating events to adverse outcomes. Therefore, an intensive search for the development of alternative methods using in silico models for neurotoxic hazard assessment is appropriate.[33] The following are some of the challenges that need to be overcome: predicting behavior using models of complex neurobiological pathways, standardizing study designs and dependent variables to facilitate creation of databases, and managing the cost and efficiency of behavioral assessments.[34]

Conclusion

There is growing evidence of the adverse impact of exposures to ambient and indoor air pollutants on fetal growth and both early childhood and animal neurodevelopment. The normal structure and function of the nervous system may be altered as a result of exposure to some pollutants before or after birth. Its analysis is particularly relevant in assessing the interference of a chemical pollutant with neuroendocrine maturation by behavioral methods, as it is a sensitive and broad marker of perturbation of both nervous and endocrine functions.

A number of methods can evaluate alterations in sensory, motor, and cognitive functions in laboratory animals exposed to toxicants during nervous system development. Assessment methods are being developed to examine other nervous system functions, including social behavior, autonomic processes, and biologic rhythms.[35] Fundamental issues underlying proper use and interpretation of these methods include 1) consideration of the scientific goal in experimental design; 2) selection of an appropriate animal model; 3) expertise of the investigator; 4) adequate statistical analysis; and 5) proper data interpretation.

Acknowledgments

I wish to thank Dr. Stachetti for his precious contribution to this entry, consisting of the initial text, over which the present one has been constructed. Also, I would like to thank him for the opportunity offered to contribute to this prestigious publication, by referring my name as author. Without his help and his confidence on my work, this achievement would not have been possible. I also would like to thank Dr. Stachetti for the final text comments and suggestions.

References

1. Bellinger, D.C. Interpreting epidemiologic studies of developmental neurotoxicity: Conceptual and analytic issues. Neurotoxicol. Teratol. **2009**, *31,* 267–274.
2. Crofton, K.M.; Makris, S.L.; Sette, W.F.; Mendez, E.; Raffaele, K.C. A qualitative retrospective analysis of positive control data in developmental neurotoxicity studies. Neuro-toxicol. Teratol. **2004**, *26,* 345–352.
3. Dietrich, K.N.; Eskenazi, B.; Schantz, S.; Yolton, K.; Rauh, V.A.; Johnson, C.B.; Alkon, A.; Canfield, R.L.; Pessah, I.N.; Berman, R.F. Principles and practices of neurodevelopmental assessment in children: Lessons learned from the centers for children's environmental health and disease prevention. Res. Environ. Health Perspect. **2005**, *113* (10), 1437–1446.
4. Bjørling-Poulsen, M.; Andersen, H.R.; Grandjean, P. Potential developmental neurotoxicity of pesticides used in Europe. Environ. Health **2008**, *7,* 50.
5. Timofeeva, O.A.; Roegge, C.S.; Seidler, F.J.; Slotkin, T.A.; Levin, E.D. Persistent cognitive alterations in rats after early postnatal exposure to low doses of the organophosphate pesticide, diazinon. Neurotoxicol. Teratol. **2008**, *30,* 38–45.
6. Rohlman, D.S.; Lasarev, M.; Anger, W.K.; Scherer, J.; Stupfel, J.; McCauley, L. Neurobehavioral performance of adult and adolescent agricultural workers. NeuroToxicology, **2007**, *28,* 374–380.
7. Richardson, J.; Chambers, J. Effects of repeated oral postnatal exposure to chlorpyrifos on cholinergic neurochemistry in developing rats. Toxicol. Sci. **2005**, *84,* 352–359.
8. Eskenazi, B.; Rosas, L.G.; Marks, A.R.; Bradman, A.; Harley, K.; Holland, N.; Johnson, C.; Fenster, L.; Barr, D.B. Pesticide toxicity and the developing brain. Basic Clin. Pharmacol. Toxicol. **2008**, *102,* 228–236.
9. Pohl, H.R.; Abadin, H.G. Chemical mixtures: Evaluation of risk for child-specific exposures in a multi-stressor environment. Toxicol. Appl. Pharmacol. **2008**, *233,* 116–125.
10. Hass, U. The need for developmental neurotoxicity studies in risk assessment for developmental toxicity. Reprod. Toxicol. **2006**, *22*, 148–156.
11. Tyl, R.W.; Crofton, K.; Moretto, A.; Moser, V.; Sheets, L.P.; Sobotka, T.J. Identification and interpretation of developmental neurotoxicity effects: A report from the ILSI Research Foundation/Risk Science Institute expert panel on neurodevelopmental endpoints. Neurotoxicol. Teratol. **2008,** *30* (4), 349–381.
12. Viberg, H. Exposure to polybrominated diphenyl ethers 203 and 206 during the neonatal brain growth spurt affects proteins important for normal neurodevelopment in mice. Toxicol. Sci. **2009**, *109* (2), 306–311.
13. Johnson, F.O.; Chambers, J.E.; Nail, C.A.; Givaruang-sawat, S.; Carr R.L. Developmental chlorpyrifos and methyl parathion exposure alters radial-arm maze performance in juvenile and adult rats. Toxicol. Sci. **2009**, *109* (1), 132–142.
14. Andersen, I.S.; Voie, O.A.; Fonnum, F.; Mariussen, E. Effects of methyl mercury in combination with polychlorinated biphenyls and brominated flame retardants on the uptake of glutamate in rat brain synaptosomes: A mathematical approach for the study of mixtures. Toxicol. Sci. **2009,** *112* (1), 175–184.

15. Bowers, W.J.; Nakai, J.S.; Chu, I.; Wade, M.G.; Moir, D.; Yagminas, A.; Gill, S.; Pulido, O.; Meuller, R. Early developmental neurotoxicity of a PCB/organochlorine mixture in rodents after gestational and lactational exposure. Toxicol. Sci. **2004**, *77*, 51–62.
16. Hogberg, H.T.; Kinsner-Ovaskainen, A.; Coecke, S.; Hartung, T.; Bal-Price, A.K. mRNA expression is a relevant tool to identify developmental neurotoxicants using an in vitro approach. Toxicol. Sci. **2010**, *113* (1), 95–115.
17. Rice, D.; Barone, S., Jr. Critical periods of vulnerability for the developing nervous system: Evidence from humans and animal models. Environ. Health Perspect. **2000**, *108* (3), 511–533.
18. Soiza-Reilly, M.; Azcurra, J.M. Developmental striatal critical period of activity-dependent plasticity is also a window of susceptibility for haloperidol induced adult motor alterations. Neurotoxicol. Teratol. **2009**, *31,* 191–197.
19. Crews, F.; He, J.; Hodge, C. Adolescent cortical development: A critical period of vulnerability for addiction. Pharmacol. Biochem. Behav. **2007**, *86,* 189–199.
20. Paul, E.S.; Harding, E.J.; Mendl, M. Measuring emotional processes in animals: The utility of a cognitive approach. Neurosci. Biobehav. Rev. **2005**, *29,* 469–491.
21. Slikker, W., Jr.; Acuff, K.; Boyes, W.; Chelonis, J.; Crofton, K.; Dearlove, G.; Li, A.; Moser, V.; Newland, C.; Rossi, J.; Schantz, S.; Sette, W.; Sheets, L.; Stanton, M.; Tyl, S.; So-botka, T. Behavioral test methods workshop. Neurotoxicol. Teratol. **2005**, *27,* 417–427.
22. Raffaele, K.C.; Rowland, J.; May, B.; Makris, S.L.; Schumacher, K.; Scarano, L.J. The use of developmental neurotoxicity data in pesticide risk assessments, Neurotoxicol. Teratol., **2010**, *32*(5):563–572. doi:10.1016/j.ntt.2010.04.053.
23. Makris, S.L.; Raffaele, K.; Allen, S.; Bowers, W.J.; Hass, U.; Alleva, E.; Calamandrei, G.; Sheets, L.; Amcoff, P.; Delrue, N.; Crofton, K.M. A retrospective performance assessment of the developmental neurotoxicity study in support of OECD test guideline 426. Environ. Health Perspect. **2009,** *117* (1), 17–25.
24. Després, C.; Richer, F.; Roberge, M.C.; Lamoureux, D.; Beuter, A. Standardization of quantitative tests for preclinical detection of neuromotor dysfunctions in pediatric neurotoxicology. NeuroToxicology **2005**, *26*, 385–395.
25. Luft, J.; Bode, G. Integration of safety pharmacology endpoints into toxicology studies. Fundam. Clin. Pharmacol. **2002**, *16,* 91–103.
26. Moser, V.; Phillips, P.; Levine, A.; McDaniel, K.; Sills, R.; Jortner, B.; Butt, M. Neurotoxicity produced by dibromo-acetic acid in drinking water of rats. Toxicol. Sci. **2004,** *79,* 112–122.
27. Sarter, M. Animal cognition, defining the issues. Neurosci. Biobehav. Rev. **2004**, *28*, 645–650.
28. Yanai, J.; Brick-Turin, Y.; Dotan, S.; Langford, R.; Pinkas, A.; Slotkin, T.A. A mechanism-based complementary screening approach for the amelioration and reversal of neurobehavioral teratogenicity. Neurotoxicol. Teratol., **2010**, *32* (1), 109–113.
29. Crofton, K.M.; Foss, J.A.; Hass, U.; Jensen, K.; Levin, E.D.; Parker, S.L. Undertaking positive control studies as part of developmental neurotoxicity testing. Neurotoxicol Teratol. **2008**, *30* (4), 266–287.
30. Raffaele, K.C.; Fisher, J.E.; Hancock, S.; Hazelden, K.P.; Sobrian, S.K. Determining normal variability in a developmental neurotoxicity test. Neurotoxicol. Teratol. **2008**, *30* (4), 288–325.
31. Castro, V.L.; Silva, P.A. Validation of neurobehavioral studies for evaluating the perinatal effects of single and mixture exposure to pesticides. In *Progress in Pesticides Research;* Kanzantzakis, C.M., Ed.; Nova Science Publishers: New York, 2009; 371–395.
32. Hines, R.N.; Sargent, D.; Autrup, H.; Birnbaum, L.S.; Brent, R. L.; Doerrer, N.G.; Cohen Hubal, E.A.; Juberg, D.R.; Laurent, C.; Luebke, R.; Olejniczak, K.; Portier, C.J.; Slikker, W. Approaches for assessing risks to sensitive populations: Lessons learned from evaluating risks in the pediatric population. Toxicol. Sci. **2010**, *113* (1), 4–26.

33. Coecke, S.; Eskes, C.; Gartlon, J.; Kinsner, A.; Price, A.; van Vliet, E.; Prieto, P.; Boveri, M.; Bremer, S.; Adler, S.; Pellizzer, C.; Wendel, A.; Hartung, T. The value of alternative testing for neurotoxicity in the context of regulatory needs. Environ. Toxicol. Pharmacol. **2006**, *21,* 153–167.
34. Bushnell, P.J.; Kavlock, R.J.; Crofton, K.M.; Weiss, B.; Rice, D.C. Behavioral toxicology in the 21st century: Challenges and opportunities for behavioral scientists—Summary of a symposium presented at the annual meeting of the Neurobehavioral Teratology Society, 2009. Neurotoxicol. Teratol. **2010**, *32,* 313–328.
35. Cory-Slechta, D.A.; Crofton, K.M.; Foran, J.A.; Ross, J.F.; Sheets, L.P.; Weiss, B.; Mileson, B. Methods to identify and characterize developmental neurotoxicity for human health risk assessment. I: Behavioral effects. Environ. Health Perspect. **2001**, *109* (1), 79–91.

7

Persistent Organic Pesticides

Gamini Manuweera

Introduction

Persistent organic pesticides refer to a set of synthetic organic chemical substances meant to control pests in the human environment including those concerning agriculture, veterinary health, and public health. Persistent organic pesticides are part of a larger group of chemicals known as persistent organic pollutants or POPs.[1] In addition to pesticides, POPs include industrial chemicals and unintentionally produced chemical substances or by-products of anthropogenic origin. Dioxins, for example, are unintentionally produced POPs formed during incomplete combustion processes involving organic matter and chlorine. Chemically, POPs include linear and cyclic halogenated hydrocarbons. In some POPs, functional moieties may also exist in the hydrocarbon molecule as in the case of perfluorooctane sulfonic acid and its salts.

POPs are highly toxic to living organisms. This group of chemicals does not easily undergo common environmental degradation processes including chemical, microbial, or photolytic reactions. Once released, POPs stay for a long period of time in the environment, posing higher risk of long-term exposure to human populations and ecosystems. POPs are lipophilic (has affinity to fat, lipids, etc.) due to the nonpolar organic nature of the substances. The lipophilicity allows chemical substances to readily accumulate in fatty tissues of living organisms (the process referred to as bioaccumulation). Once accumulated, the concentration of POPs in the living organisms builds up through the food chain via biomagnification processes, increasing the risk of adverse effects at the higher tropic levels.

Some of the physicochemical properties of the POPs facilitate long-range transport in the environment. POPs are found in the alpine and mountainous regions, the Arctic, Antarctica, and remote Pacific islands far away from where activities associated with POPs are taking place. In the environment, POPs undergo sorption into organic matter, intra-media, and inter-media dispersion (diffusion), as well as advection (transport mechanism of substances due to the bulk motion of the medium). Most of the POPs are semi-volatile in nature. With long residence time in transport media, POPs can travel very long distances across regions through environmental transportation processes, including atmospheric transport, making them available for human and environmental exposure on a global scale.

The toxic effects of POPs are mainly linked to long-term, low-level exposure scenarios mostly resulting in chronic health problems, while some POPs could also exert acute effects. The toxic endpoints of POPs include cancer, birth defects, reproductive problems, damages to specific organs such as the liver and kidneys, among others.

Much attention was drawn to this group of substances at the international level after it became apparent that they travel long distances across borders. As a consequence, several countries started banning POPs in the 1970s. However, actions by a limited number of countries alone were unable to control continued environmental pollution and adverse health effects from such border-crossing substances. A regional legal agreement that specifically addresses POPs was adopted in 1998 with the Aarhus Protocol on Persistent Organic Pollutants under the regional Convention on Long-Range Transboundary Air Pollution (LRTAP) of the UN Economic Commission for Europe (UNECE).[2] As an agreement at the regional level was not sufficient to ensure satisfactory protection of human health and the environment from adverse effects of POPs, negotiations of an international legally binding instrument to reduce or eliminate, where possible, releases of POPs were initiated under the auspices of the United Nations Environmental Programme (UNEP) in 1998. In May 2001, more than 100 countries agreed and adopted a global treaty, now known as the Stockholm Convention on Persistent Organic Pollutants.[3,4] Some aspects of the life cycle of POPs are considered in other international legally binding instruments. POP wastes are included in the Basel Convention on the Control of Transboundary Movements of Hazardous Wastes and Their Disposal,[5] while the international trade of most POP pesticides is addressed by the Rotterdam Convention on the Prior Informed Consent Procedure for Certain Hazardous Chemicals and Pesticides in International Trade.[6] Several other international initiatives also address POPs, notably the Global Programme of Action for the Protection of the Marine Environment from Land-Based Activities (GPA)[7] and a number of regional sea agreements.

Persistent Organic Pesticides under the Stockholm Convention

As of 2019, there were 30 POPs included in the Stockholm Convention, of which, 17 were pesticides. At the time of its entry into force in 2001, the Stockholm Convention had 9 pesticides among the total of 12 POPs listed therein.

The Convention provides provisions for Parties to make proposals to add new POPs. These proposals must contain information on the chemical relating to the screening criteria established under the Convention that include persistence, bioaccumulation, potential for long-range environmental transport, and adverse effects. The POPs Review Committee established by the Convention reviews submissions on candidate POPs, including related information from other sources, prepares a risk profile, and undertakes a risk management evaluation. This step is undertaken once the committee evaluate the risk profile and agrees that the candidate chemical is likely, as a result of its long-range environmental transport, to lead to significant adverse human health and/or environmental effects, such that global action is warranted.

Upon that it undertakes a risk management evaluation to identify specific details of control actions relevant to the chemical to accompany with its recommendation to the COPs. If the candidate chemical satisfactorily meets the POP screening criteria of the Convention, the POPs Review Committee makes recommendations to the Conference of the Parties for considering the new chemical under the Convention. The Conference of the Parties then evaluates the recommendations made by the POPs Review Committee for the inclusion of the candidate chemical.

At its fourth meeting held in 2009, the Conference of the Parties to the Convention considered and included nine new chemical substances consisting five pesticides as POPs in the Convention. During the fifth, seventh, and ninth Conferences of the Parties held in 2011, 2017 and 2019, endosulfan and pentachlorophenol and its salts and esters and dicofol were listed under the Convention, respectively, as pesticides (see Table 1).

TABLE 1 Pesticides Listed as POPs under the Stockholm Convention

Aldrin
Alpha hexachlorocyclohexane
Beta hexachlorocyclohexane
Chlordane
Chlordecone
DDT
Dieldrin
Dicofol
Endosulfan
Endrin
Heptachlor
Hexachlorobenzene
Lindane
Mirex
Pentachlorobenzene
Pentachlorophenol and its salts and esters
Toxaphene

The Stockholm Convention requires Parties to take measures to reduce or eliminate releases from intentional production and use of POPs through two different approaches:

- Prohibit and/or take the legal and administrative measures necessary to eliminate the chemical substances listed in Annex A of the Convention.
- Restrict production and use of the chemicals listed in Annex B of the Convention in accordance with specific measures provided in that Annex.

The chemical substances for elimination are listed in Annex A of the Convention. When chemical substances with existing commercial uses are included in the Convention for elimination, some chemicals may still have certain specific use or uses for which Parties to the Convention may require a transition period to eliminate completely the reliance on the chemical substance. For such chemicals, exemptions are provided for those specific uses and related production for a limited period of time. These specific exemptions are initially available for a 5-year period. The Parties opt for the use of specific exemptions that require conformation to a set of precautionary measures relevant to its available uses established by the Convention to ensure reduced releases of the substance and effective elimination. When there are no longer any Parties registered for a particular type of specific exemption, no new registrations may be made with respect to it. Aldrin, for example, was listed in Annex A of the Convention for elimination with a specific exemption of use only as local ectoparasiticide and insecticide with no further production. At the end of the initial period, the exemptions for production and use are not available, if Parties have effectively eliminated the reliance on those POPs.

During the fourth meeting of the Conference of the Parties in 2009, it was noted that the specific exemptions provided for six POP pesticides initially listed in Annex A of the Convention were no longer needed. Accordingly, the POP pesticides aldrin, chlordane, dieldrin, heptachlor, hexachlorobenzene (also found as a by-product and industrial chemical), and mirex listed in Annex A have no specific exemptions available for Parties anymore. In addition to those six POP pesticides, Annex A also contains alpha hexachlorocyclohexane (also found as a by-product), beta hexachlorocyclohexane (also found as a by-product), chlordecone, endrin, and toxaphene, for which no specific exemptions were provided at the time of listing them in the Convention. Dicofol was listed under the Convention without exemptions. The three remaining POP pesticides listed in Annex A with specific exemptions are lindane, technical endosulfan and its related isomers and pentachlorophenol and its salts and esters.

Lindane was listed under the Convention with prohibition of its production. All uses have also been prohibited except as human health pharmaceutical for control of head lice and scabies as second-line treatment. For Parties that may wish to use lindane for the control of head lice and scabies, the specific exemptions were initially available for that purpose until 2015. When listing the chemical, the Conference of the Parties requested the Secretariat to develop a work plan in collaboration with the World Health Organization (WHO) for reporting and reviewing requirements on the use of lindane for the specific exemptions allowed by the Convention. Lindane has been listed from 1977 and 1990 in the WHO Model List with benzyl benzoate and permethrin as a cost-effective alternative to lindane for head lice and scabies. It was removed from the WHO Model List in 1992 on the basis that "it is toxic to the environment and humans, and safer alternatives are available." The specific exemptions for lindane provided under the Stockholm Convention expired for most Parties in 2015. It was noted at the ninth Conference of the Parties held in 2019 that no Party has registered for specific exemption for lindane. Hence, registration for the exemptions of lindane will no longer be available for Parties.

The specific exemptions for technical endosulfan expired for most Parties in 2017. As of 2018, two Parties, which had made declarations pursuant to the Convention, had registered exemptions that will be applicable for them until 2019.

The chemical substances listed in Annex B are those identified with specific uses for which there are no alternatives available at present or that the alternatives are not accessible or effectively available under certain settings. Such uses are recognized under the Convention as acceptable purposes with no set timeline for elimination. Parties are allowed to produce and use the chemicals for those purposes according to the recommended practices provided with respective uses. In such cases, a process is established to periodically review the continued need of the chemical in question for the acceptable purposes. The Convention has listed dichlorodiphenyltrichloroethane (DDT), for example, under Annex B with an acceptable purpose for disease vector control. It requires activities associated with DDT to be in accordance with the WHO's recommendations and guidelines on the use of DDT. Further, the use of DDT is allowed when locally safe, effective, and affordable alternatives are not available to the Party in question.

Following Article 4 of the Convention, the Parties that use POP pesticides according to the specific exemptions and acceptable purposes provided in the Convention require notifying the Secretariat to register for that purpose. The register of specific exemptions is made publicly available on the Convention's website (www.pops.int).

Major Issues concerning Persistent Organic Pesticides

Over the years, there has been an increase in both the general understanding and concern about adverse effects of POP pesticides. The latest concerns include interference of POPs with hormonal activities, acting as "endocrine disruptors," and possible interlinkages between climate change and POPs. Release, distribution, and degradation of POPs are highly dependent on environmental conditions. Climate change and increasing climate variability have the potential to affect POPs' contamination via higher releases from primary sources and environmental reservoirs, changes in transport processes and pathways, and routes of degradation. Exposure to POPs and related impacts on environmental and human health can be further exacerbated by higher atmospheric temperatures.

Most of the POP pesticides included in the Stockholm Convention are first-generation pesticides discovered in the World War II era. The subsequent advent of pesticides from the chemical families of organophosphates, carbamates, and synthetic pyrethroids challenged the continuity of favorable market for the first-generation pesticides. However, the remarkable successes achieved in the malaria eradication programs in the 1950s and 1960s using DDT for indoor residual spraying (IRS)[8] and the continued need to rely on DDT for disease vector control compelled the global community to place DDT under Annex B of the Convention, thereby allowing its continued production and use for disease vector control. The WHO recommends DDT only for IRS. Countries can use DDT as long as necessary, in the quantity needed, provided that the guidelines and recommendations of the WHO and the Stockholm Convention are all met and until locally appropriate and cost-effective alternatives are available for a sustainable transition from DDT. The continued need for DDT for disease vector control is evaluated at regular Conferences of the Parties, held every 2 years. The evaluation is undertaken in consultation with the WHO on the basis of available scientific, technical, environmental, and economic information. As a separate process, the WHO also reviews new information on adverse health effects of DDT periodically to facilitate the evaluation of DDT by the Convention. In 2010, the WHO expert consultation report on "DDT in Indoor Residual Spraying: Human Health Aspects" identified several potential hazards of DDT and its toxic metabolites. These include acute poisoning hazards for children with accidental ingestion, carcinogenicity, developmental toxicity, male reproductive effects, and concerns for women of childbearing age who live in DDT IRS-treated dwellings. However, in terms of relevant exposure scenarios for the general population in countries using IRS, the expert panel concluded that available evidence does not point to concern about levels of exposure for any of the endpoints that were assessed. The report demands further research to better evaluate risks that were suggested in the studies reviewed.[9]

The chemical substances currently added to the Convention as new POPs often have many active uses. For some of those uses, either alternatives are not currently found or cost-effective alternative products and options are not readily accessible under a certain setting, often under the conditions prevailing in developing countries. Therefore, it is not uncommon that new POPs added to the Convention consist of a relatively longer list of use exemptions. The later addition to POP pesticides, endosulfan, which is listed in Annex A, has a number of uses provided under its specific exemptions (see Table 2). Eliminating exposure to POPs is more challenging when the chemicals are included in Annex B of the Convention with many acceptable purposes, where there is no time-bound phaseout requirement.

Indiscriminate use of pesticides could result in the development of pest resistance and resurgence of new pests requiring increased dependence on pest control actions. Limitations in effective deployment

TABLE 2 Specific Exemptions of Crop–Pest Complexes of Endosulfan Available for Parties under the Stockholm Convention

Crop	Pest
Apple	Aphids
Arhar, gram	Aphids, caterpillars, pea semilooper, pod borer
Bean, cowpea	Aphids, leaf miner, whiteflies
Chilli, onion, potato	Aphids, jassids
Coffee	Berry borer, stem borers
Cotton	Aphids, cotton bollworm, jassids, leaf rollers, pink bollworm, thrips, whiteflies
Eggplant, okra	Aphids, diamondback moth, jassids, shoot, and fruit borer
Groundnut	Aphids
Jute	Bihar hairy caterpillar, yellow mite
Maize	Aphids, pink borer, stem borers
Mango	Fruit flies, hoppers
Mustard	Aphids, gall midges
Rice	Gall midges, rice hispa, stem borers, white jassid
Tea	Aphids, caterpillars, flushworm, mealybugs, scale insects, smaller green leafhopper, tea geometrid, tea mosquito bug, thrips
Tobacco	Aphids, oriental tobacco budworm
Tomato	Aphids, diamondback moth, jassids, leaf miner, shoot and fruit borer, whiteflies
Wheat	Aphids, pink borer, termites

of nonchemical pest control interventions, especially in developing countries, lead to further reliance on chemical control options. Even if the alternative pesticides are not highly toxic or do not possess POPs' characteristics, the increase of chemical load on the environment is inevitable, leading to undesirable consequences. Under certain settings where locally appropriate, cost-effective, and safer alternatives are not accessible when managing pest resistance has become a serious challenge would demand possible reintroduction or continued use of POP pesticides.

Alternative Approaches

Alternatives to POP pesticides include chemical and nonchemical products as well as control interventions that focus on avoiding pest interference or creating conditions not favorable for the prevalence of the targeted pest. Often, any one of those potential alternatives alone would not fulfill all desirable features of the POP pesticides as a successful replacement. It requires formulating approaches with combination of viable options appropriate for the given situation. The availability of a wider range of choices from different pest control options is vital for the development of such alternative approaches. New developments in organic chemistry such as increased flexibility to modify known functional insecticide chemical backbones to produce new toxophores have expanded the prospects for new chemical alternatives. Current research on nonchemical options ranges from conventional biopesticide products to gene technology and interventions on physical ecosystem management. These initiatives should present promising opportunities for a wider selection of vector control options for efficient integration.

Major issues related to the elimination of pesticides in the present list of POPs particularly concern the use of DDT for disease vector control. There are only three classes of pesticides currently available for public health vector control as alternatives to DDT: synthetic pyrethroids, organophosphates, and carbamates. They represent two different modes of actions limiting the choice of pesticides available for the management of vector resistance. The situation demands an urgent need for bringing new cost-effective public health pesticides to disease-endemic countries. All current public health pesticide active ingredients were initially developed for the agrochemical market. In the 1980s, the shift of the agrochemical target product profile, from broad-spectrum contact insecticides to target site specific selective toxicity insecticides, delivered a number of new agrochemicals that could be repurposed for public health. In efforts to eliminate DDT, governments should seek alternative approaches that are sustainable in situations prevailing in the country. It will be important to ensure that these POP pesticides are not simply replaced by other pesticides, but that the principles of integrated pest and vector management are adopted with due consideration on resistance management. Intergovernmental agencies such as the United Nations Environment Programme are promoting initiatives to demonstrate sustainable replacement of DDT in disease vector control using integrated multidisciplinary approach. It is also important to ensure that the strategies used will not be compromised by measures in other sectors. Efforts for effective management of resistance in disease vectors are hindered where the same insecticides are used in agriculture. Similarly, unplanned environmental modifications for developmental purposes could create more breeding grounds for malaria mosquitoes.

Assurance of close collaboration between sectors and key stakeholders is vital in endeavors by countries to find more sustainable solutions to POPs. These collaborations also help strengthen the base of the civil society in communities to increase other social benefits. Community participation, multisectoral initiatives on public awareness campaigns, and local social surveillance have been successfully integrated in a program implemented in Mexico and Central America on demonstrating the effectiveness of alternative methods to DDT for malaria control.[10] Such solutions must be based on the local conditions and can be best sustained through active community participation. Structures established under one sector such as Farmer Field Schools in agriculture may, for example, serve the purposes of public health and the environment. The interrelationship between the environment, agriculture, and health is, hence, a key for identifying sustainable strategies that will effectively and efficiently protect agriculture from pests, communities from diseases like malaria, and ecosystems from persistent pesticides.

Global Initiatives to Promote Alternatives to DDT

The fourth Conference of the Parties of the Stockholm Convention concluded that countries currently using DDT for disease vector control may need to continue such use until locally appropriate and cost-effective alternatives are available for a sustainable transition away from DDT. To support the countries still using DDT to reduce their reliance, it also endorsed the establishment of the Global Alliance for the development and deployment of products, methods, and strategies as alternatives to DDT for disease vector control.

The Global Alliance for alternatives to DDT provides an instrument for partnership and collaboration among all stakeholders at the global and national level to increase momentum on achieving the common goals and to catalyze new initiatives for the development and deployment of alternatives towards the elimination of reliance on DDT. The work of the Global Alliance is organized in a manner that respects its noninvolvement in funding and executing programs on the ground, yet addresses expectations that it will trigger significant actions in support of the development and deployment of alternatives to DDT.

The instrument is expected to stimulate the research community and chemical industry to accelerate the release of safer chemical alternatives to DDT. It draws global development initiatives to strengthen in-country capacity and knowledge base for efficient integration of vector control options into cost-effective and sustainable programs in disease-endemic countries. It involves global authorities and experts to review and develop support tools for countries to implement related activities and support efficient and effective networks and communication to promote indigenous knowledge and innovative concepts on nonchemical approaches.

The platform has been established with partners from parties to the Convention with due consideration on malaria disease-endemic countries and WHO, including research and academic institutions, the donor community, civil society organizations, and the pesticide industry.[11]

The UNEP in response to an invitation by the sixth Conference of the Parties to the Stockholm Convention held in 2013 prepared a road map for the development of locally safe, effective, affordable, and environmentally sound alternatives to DDT in consultation with WHO, the DDT expert group, established by the Conference of the Parties to assess continued need for DDT for disease vector control, and the Secretariat of the Convention. The basis of the Conference of the Parties to develop a road map was the conclusion that countries relying on DDT for disease vector control may need to continue such use until locally safe, effective, affordable, and environmentally sound alternatives are available for a sustainable transition away from DDT. The purpose of the road map is to provide a thematic guide and logical steps that are needed to make locally safe, effective, affordable, and environmentally sound alternatives to DDT[12].

Conclusion

The purpose of pesticides as a tool to control and, including in many cases, to kill living organisms presents the greatest drawback for its own existence and use. While the adverse effects of pesticides are many and diverse, some pesticides pose unique risks. Irrespective of the point of release, POP pesticides exert serious adverse health and environmental effects on the global scale, leaving little or no options except avoiding the reliance on them towards total elimination.

In spite of constant efforts, challenges for the development and deployment of sustainable solutions to avert continued reliance on some POP pesticides remain. The socioeconomic and ecological dimensions in specific settings may impede bringing in straightforward global solutions without compromising the benefits associated with certain uses of POP pesticides. The conventional approach of finding a chemical replacement with same pesticidal properties of the POP pesticide is not always viable. When long residual effects on the control of targeted pest, a property link to POPs' characteristics, are among the reasons for continued need for a POP pesticide, finding a chemical replacement with similar properties becomes even more challenging. The solution should, therefore, respect a multidisciplinary approach targeting interventions on a broader scope encompassing the life cycle of the pest consisting of a series of control options. Any individual element of such multidisciplinary control approach shall not produce a complete control over the targeted pest. The implementation of properly formulated complementary control interventions offers a successful and sustainable outcome. Such an approach requires the collaboration of different sectors of the society in implementing respective control actions with proper coordination in a strategic framework. It also includes action by stakeholders at global, regional, and local levels to ensure enhanced sustainability of the initiatives. Benefits to the global community on such integrated approach are not limited to the protection of human health and the environment from adverse effects of POP pesticides. It also helps to achieve enhanced coordination and collaboration within the civil society for sustainable development especially under resource-limited settings.

References

1. United Nations Environment Programme. Stockholm Convention on Persistent Organic Pollutants (POPs), 2009.
2. United Nations Economic Commission for Europe: Environment, available at www.unece.org/leginstr/env.html (accessed November 2011).
3. Persistent Organic Pollutant; United Nations Environmental Programme, available at www.chem.unep.ch/pops (accessed August 2011).
4. About the Convention. Stockholm Convention—Protecting Human Health and the Environment from Persistent Organic Pollutants, available at http://chm.pops.int/Conven-tion/tabid/54/Default.aspx (accessed October 2011).
5. Basel Convention on the Control of Transboundary Movements of Hazardous Wastes and their Disposal, available at http://archive.basel.int/ (accessed March 2011).
6. Rotterdam Convention: Shared Responsibility, available at www.pic.int/ (accessed November 2011).
7. Global Programme of Action for the Protection of the Marine Environment for Land-based Activities (GPA); United Nations Environment Programme, available at www.gpa.unep.org/ (accessed November 2011).
8. Mörner, J.; Bos, R.; Fredrix, M. *Reducing and/or Eliminating Persistent Organic Pesticides—Guidance on Strategies for Sustainable Pest and Vector Management*; United Nations Environment Programme: Geneva, 2002.
9. World Health Organization. *DDT in Indoor Residual Spraying: Human Health Aspects, Environmental Health Criteria*; WHO Press: Geneva, 2011, p. 241.
10. Regional Program of Action and Demonstration of Sustainable Alternative to DDT for Malaria Vector Control in Mexico and Central America (project DDT/UNEP/GEF/PAHP), available at www.paho.org/english/ad/sde/ddt-regionals.htm (accessed October 2011).
11. Global Alliance for Alternatives to DDT. Stockholm Convention: Protecting Human Health and the Environment from Persistent Organic Pollutants, available at www.pops.int/ga (accessed March 2011).
12. Road map for the development of alternatives to DDT, UNEP/DTIE, Chemicals Branch, February 2015, available at www.unenvironment.org/explore-topics/chemicals-waste/what-we-do/persistent-organic-pollutants/roadmap-alternatives-ddt (accessed July 2019).

8

Pollutants: Organic and Inorganic

A. Paul Schwab

Introduction

A soil pollutant can be broadly defined as any chemical or other substance which either is not normally found in soil or is present at high enough concentrations to be harmful to any living organisms.[1] This very general definition could be applied to human-derived (anthropogenic) as well as naturally occurring constituents. This entry will focus primarily on those pollutants that have anthropogenic origins and will address carbon-based (organic) and inorganic chemicals. Organic contaminants can include pesticides, solvents, preservatives, petroleum products, hormones, and antibiotics. Inorganic pollutants are comprised of heavy metals (e.g., lead and mercury), nonmetals (selenium), metalloids (arsenic and antimony), radionuclides, and simple soluble salts (sodium chloride). Whether the source of the contaminants is natural or anthropogenic, understanding their chemistry, toxicity, and bioavailability is crucial to responding to soil pollution.

Organic Pollutants

Thousands of organic chemicals exist in nature, and many are acutely toxic as well as carcinogenic, mutagenic, or teratogenic. When discussing toxic organic chemicals in soils, the focus normally is on those pollutants resulting from human activities because their release has the potential to be controlled. Typical synthetic organic pollutants include fuels, lubricants, herbicides, fungicides, insecticides, solvents, and propellants.

Soil Contamination by Organic Chemicals

Organic chemicals can find their way into the soil accidentally through spills, leaking storage tanks, and unintentional discharges. However, not all soil pollution is accidental. More than one-half million metric tons of pesticides are used annually, the majority of which are used on agricultural fields.[1]

Approximately 20,000 metric tons of pesticides are used for non-agricultural applications, including railroad right of way weed control, turf, and horticulture. Although many pesticides do not persist in soils, others are highly persistent and have been studied extensively because of their negative impacts. Chlordane (termite control), dichlorodiphenyltrichloroethane (DDT) (mosquitoes), and atrazine (weeds) are excellent examples of organic pesticides that have been determined to have human health effects and severe ecological impacts; chlordane and DDT have been banned in the United States, and the banning of atrazine has been debated.

Thousands of organic chemicals are in use today, and listing all the specific compounds that contaminate soils is beyond the scope of this entry. However, contaminants can be sorted into categories of chemicals that are used frequently and are commonly found in soils (Table 1). Insecticides, herbicides, fungicides, and nematicides may be the most frequently encountered, and some pesticides are quite toxic. Atrazine has been found to be somewhat persistent and is mobile such that atrazine applied to soil can migrate to groundwater and surface water. The insecticide DDT and its metabolites continue to be found in ecosystems despite being banned for decades. Methyl bromide (1,2-dibromomethane) is used as a nematicide, but it is a gas that is known to be ozone depleting. Trichloroethylene (TCE) is a useful solvent that, when disposed on soil, moves rapidly downward and contaminates groundwater. Petroleum-based fuels such as diesel and gasoline are problematic when spilled or originating from leaking underground storage tanks. Many of the other classes of compounds listed are of ecological concern when found in soil, again due to their potential toxicity to a wide range of organisms.

On December 11, 1980, the United States Congress enacted the Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA), commonly known as Superfund, to provide wide-ranging federal authority to respond to contamination or threats of contamination by hazardous substances. Among other actions, CERCLA established a list of priority pollutants with known or suspected

TABLE 1 Some Classes of Organic Chemicals Found in Soils

Contaminant Class	Example(s)	Sources, Impacts
Insecticides	DDT, chlordane, diazinon	Used for insect control; DDT and chlordane have serious health effects; diazinon has been found in water supplies.
Herbicides	Atrazine, 2,4-D, 2,4,5-T	Atrazine used on corn; 2,4-D widely used in lawns and agriculture; defoliant 2,4,5-T implicated in health effects in Agent Orange.
Fungicides	Benomyl, propiconazole, chlorothalonil	Agricultural fungicides that are directly applied to soil.
Nematicides	Methyl bromide (1,2-dibromomethane)	Used in fumigation of soil to remove nematodes; neurotoxin and ozone depleting.
Solvents	Trichloroethylene, trichloroethane	Widely used solvents and degreasers; chronic health effects not clear.
Fuels	Diesel, kerosene	From spills, leaking tanks; can lead to groundwater contamination. Some constituents of fuels are carcinogenic.
Polyaromatic hydrocarbons	Chrysene, benzo[a]pyrene	Components of petroleum, particularly after combustion; many of these are carcinogenic.
Polychlorinated biphenyls	Aroclors (1260, 1016, 1242); coplanar congeners (3,4,3’, 4’- tetrachlorobiphenyl, 3,4,5,3’, 4’- pentachlorobiphenyl)	Dielectric fluids in transformers, capacitors, and coolants. Can lead to skin conditions, ocular lesions, teratogenic effects in animals, endocrine disruption.
Explosives, propellants	TNT, RDX	Residuals from manufacture are the largest source of contamination; acute and chronic toxicities documented.

Source: Adapted from Schwarzenbach, Gschwend, and Imboden[2] and Evangelou.[3]

ecological or health impacts. The list is periodically updated, and the pollutants are prioritized. Thirty-two of these compounds are given in Table 2 along with the CERCLA priority ranking[4] and soil screening levels (SSLs).[5] This list includes compounds that have been banned in the United States (e.g., DDT) as well as those that are part of our everyday lives (e.g., BTEX—benzene, toluene, ethylbenzene, xylenes—found in gasoline). Most of the SSLs are in the range of 1 to 50 mg/kg. Some compounds have significantly higher SSL values (5000 mg/kg for toluene), indicating that these compounds are far less toxic. Others have very low SSL values (5.0×10^{-4} mg/kg for benzidine), suggesting that these compounds are a threat at very low concentrations. Soil concentrations of these compounds required to prevent threats to groundwater are approximately 100 times lower than the residential SSLs.

TABLE 2 Organic Compounds on the CERCLA Priorities List and Their Ranking, Health Impacts, and SSLs

		Soil Screening Level	
		Residential Soil	Protect Groundwater
Constituent	CERCLA Rank	mg/kg	
Vinyl chloride	4	6.0×10^{-2}	5.6×10^{-6}
PCBs[a]	5	1.4×10^{-1}	5.2×10^{-3}
Benzene	6	1.1×10^{-0}	2.1×10^{-4}
Polycyclic aromatic hydrocarbons	8	1.5×10^{-2}	2.7×10^{-4}
Chloroform	11	2.9×10^{-1}	5.3×10^{-5}
DDT, *p*, *p*'	12	1.7×10^{0}	6.7×10^{-2}
Trichloroethylene	16	2.8×10^{0}	7.2×10^{-4}
Dieldrin	17	3.0×10^{-2}	1.7×10^{-4}
Chlordane	20	1.6×10^{0}	1.3×10^{-2}
DDE, *p*, *p*'	21	1.4×10^{0}	4.7×10^{-2}
Hexachlorobutadiene	22	6.2×10^{0}	1.7×10^{-3}
Aldrin	24	2.9×10^{-2}	6.5×10^{-4}
DDD, *p*, *p*'	25	1.4×10^{0}	6.6×10^{-2}
Benzidine	26	5.0×10^{-4}	2.4×10^{-7}
Toxaphene	31	4.4×10^{-1}	9.4×10^{-3}
Hexachlorocyclohexane, γ(lindane)	32	5.2×10^{-1}	3.6×10^{-4}
Tetrachlorethylene	33	5.5×10^{-1}	4.9×10^{-5}
Heptachlor	34	1.1×10^{-1}	1.2×10^{-3}
1,2-Dibromomethane	35	2.5×10^{1}	2.5×10^{-3}
Hexachlorocyclohexane, ß	36	2.7×10^{-1}	2.2×10^{-4}
Acrolein	37	1.5×10^{-1}	8.4×10^{-6}
Disulfoton	38	2.4×10^{0}	2.7×10^{-3}
3,3'-Dichlorobenzidine	40	1.1×10^{0}	9.4×10^{-4}
Endrin	41	1.8×10^{1}	4.4×10^{-1}
Pentachlorophenol	45	3.0×10^{0}	5.7×10^{-3}
Heptachlor epoxide	46	5.3×10^{-2}	1.5×10^{-4}
Carbon tetrachloride	47	6.1×10^{-1}	1.7×10^{-4}
Diazinon	56	4.3×10^{1}	1.6×10^{-1}
Xylenes	58	6.3×10^{2}	2.0×10^{-1}
Toluene	71	5.0×10^{3}	1.6×10^{0}
Ethylbenzene	99	5.4×10^{0}	1.7×10^{-3}

All screening levels are for residential soils unless stated otherwise. Radionuclides were excluded from this list.

[a] This is a general class of contaminants, and individual members have unique SSLs. The most restrictive value was chosen for this table.

The legacy of high use of persistent, potentially toxic compounds became clear in a study published in 2010 in which the presence of organochlorine pesticides, polychlorinated biphenyls, and perfluorinated compounds were determined in food samples purchased in supermarkets in the United States.[6] The tracked compounds were detected in nearly all the food samples: DDT metabolite *p*, *p*′-dichlorodiphenyldichloroethylene was found in milk products; polychlorinated biphenyls (PCBs) were found in fish; and perfluorinated compounds were found in over half of all samples. Results such as these add to already enhanced sensitivities concerning organic contaminants in soils.

Potential Impacts of Organic Contaminants

After an organism is exposed to an organic pollutant, a number of antagonistic effects are possible if concentrations are high enough. The most dramatic impact is acute toxicity, in which symptoms are quickly apparent and readily identified. Consuming large quantities of the pure contaminant is not necessarily a requirement for acute toxicity. A case of acute parathion poisoning was reported when a child consumed contaminated soil.[7] Although reports of such cases are rare, the possibility for acute poisoning through soil consumption by children with pica is realistic for other compounds, such as phenol.[8] The estimated lethal dose for phenol is estimated to range from 10 to 50 mg/kg body mass, but the ingestion of only 5 g of soil contaminated with an SSL of 47,000 mg/kg would result in a dose of 18 mg/kg.

At lower concentrations, the impacts of organic contaminants become less obvious and take longer to be expressed. For many carcinogens, for example, decades are required to develop cancerous tumors. Pathway of exposure, concentrations, and duration of the exposure all dictate the resulting health effects and are essential components of risk analysis. Exposure to contaminated soil can result in chronic toxicity if the soil repeatedly comes in contact with the skin, if particulates are frequently inhaled, or if volatile compounds migrate into closed living spaces. Most SSLs have been developed based on risk associated with chronic exposure.

Bioremediation of Organic Contaminants

Although contamination of soil by organic compounds is an important environmental problem, many of these pollutants can be removed from the soil through bioremediation. Soil microorganisms have a remarkable capacity to degrade organic contaminants. Degradation can be direct, using the organics as a source of energy and carbon, or indirect, in which the compounds are cometabolized by organisms seeking similar compounds. End products can be CO_2(g) after total mineralization; alteration and humification; or incorporation into the microbial biomass. Polyaromatic hydrocarbons (PAHs) are readily degraded in the soil, as are many pesticides and components of BTEX. However, highly chlorinated compounds such as PCBs, some solvents, and the explosive cyclotrimethylenetrinitramine (RDX) are far more difficult to degrade, and some of the initial degradation products are as toxic as or more toxic than the parent compound. Reviews have been published for the bioremediation of pesticides,[9] a wide range of organic contaminants by fungi,[10] bacterial degradation of aromatic compounds,[11] PAHs,[12] aliphatic hydrocarbons,[13] explosives,[14] petroleum hydrocarbons by mycorrhizae,[15] earthworm-assisted bioremediation,[16] and composting as a general bioremediation approach.[17] Bioremediation of specific compounds also has been reviewed: pyridine, indole, and quinoline[18]; dieldrin and endrin[19]; and catechols.[20] This subject is treated in depth in another section of this entry.[21]

Phytoremediation is another approach taken to enhance the dissipation of organic contaminants in soil. The mechanism of removal of the contaminant from the soil depends upon the properties of the organic compound, the soil, and the chemistry of the roots. In some instances, organic contaminants are assimilated by the plants and either degraded or volatilized as part of the transpiration stream. Trichloroethylene has been observed to be effectively remediated in the root zone of poplar trees, but the mechanism has been the subject of debate. In some instances, uptake of TCE and eventual volatilization have been observed,[22] but other studies detected no volatilization of TCE and complete degradation in the soil.[23] In nearly all cases for PAHs and PCBs, plant uptake is negligible, and phytoremediation of these compounds is accomplished by microbial degradation in the rhizosphere.

Uptake of organic contaminants during phytoremediation is an important consideration for many reasons. From an ecological standpoint, accumulation of contaminants in the aboveground portions of plants is undesirable because of the potential for introduction into the food chain or dispersal of the contaminants. From the remediation perspective, uptake may be desirable to help remove the compounds from the soil and allow degradation within plant tissues. Uptake of volatile compounds followed by release to the atmosphere would be prohibited in many regulatory environments. Therefore, efforts have been taken to predict the transfer from soil to roots to the transpiration stream of higher plants. The most useful parameter in this analysis is the transpiration stream concentration factor (TSCF):

$$\mathrm{TSCF} = \frac{\textit{Concentration in xylem sap}}{\textit{Concentration in external solution}}$$

Concentrations in the xylem are difficult to quantify and are estimated as the amount of a compound assimilated over a given period of time that has been corrected for degradation with the plant. If the compound of interest is neither actively accumulated nor excluded from the plant (i.e., passive uptake), then TSCF = 1.0. Non-ionized organic compounds are of particular interest in phytoremediation because target contaminants are in this class, including many pesticides, PAHs, PCBs, etc. Trends in TSCF as a function of the octanol-water coefficient (K_{OW} or log K_{OW}) have been investigated.[24,25] Compounds that are soluble in water have small or even negative values of log K_{OW}; hydrophobic compounds have high values of log K_{OW}. Briggs et al.[24] investigated substituted phenylureas and *o*-methylcarbamoyloximes, and Burken and Schnoor[25] investigated a suite of compounds including RDX, phenol, benzene, atrazine, TCE, pentachlorophenol (PCP), and others. The combined data from the two studies are shown in Figure 1 and follow the relationship

$$\mathrm{TSCF} = 0.76\ \exp\left(-\left(\log K_{OW} - 2.45\right)^2 / 4.38\right)$$

This relationship predicts that compounds with a log K_{OW} of 2.45 will have the maximum movement into the transpiration stream. The shape of the curve is viewed as reflecting the balance of the various tendencies for the compounds to desorb from soil surfaces and pass through the hydrophobic

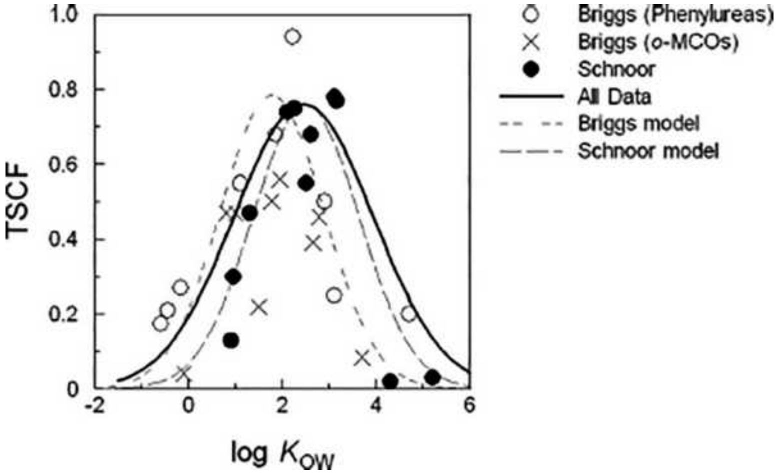


FIGURE 1 Transpiration stream concentration factor (TSCF) as a function of log K_{OW} (octanol-water partition coefficient). Data come from Briggs et al.[24] and Burken and Schnoor.[25] The models take the following form: TSCF = *a* exp(−(log K_{OW} − *b*)^2 *c*), in which *a, b,* and *c* are variables.

cellular membranes. Compounds with a low log K_{OW} will desorb readily from the soil but will not pass through the hydrophobic plasmalemma. Organic contaminants with high log K_{OW} values are predicted to have the capability to penetrate the root membranes, but they either will not desorb from soil components or will irreversibly adsorb to the lipophilic root membranes.[26]

Inorganic Pollutants

Inorganic pollutants are those contaminant compounds that do not contain carbon (with a few exceptions, including cyanide and carbonate). These contaminants include acutely toxic heavy metals, metalloids, oxyanions, and radioactive elements and the far-less-toxic soluble salts. Although all elements are naturally occurring, thousands of inorganic compounds are formed only through human activities. Some are highly toxic in trace quantities, but all have a threshold concentration above which they can be harmful.

Classes of Inorganic Pollutants and Concentration Ranges

As with organic pollutants, thousands of inorganic contaminants exist and are found over a wide range of concentrations. It is convenient to group them into classes (Table 3) consistent with the periodic table. Metals include Pb, Zn, Ni, and Cd, and their typical soil concentrations vary over a wide range, as do the concentrations considered anomalous. Most metals in soils occur as the cations in solution and in the solid phase. Nearly all metals can have toxic effects on humans, although some metals (such as Zn) require high concentrations for toxicity. Metalloids, including As and Sb, generally occur in soil solution and solid phases as oxyanions (such as AsO_4^{3-}), resulting in significantly different chemical behavior than the metals. Metalloids can be highly toxic. Soluble salts are typical inorganic species that, in normal concentrations, generally do not constitute a health or environmental threat. Typical salts are sulfates, chlorides, and nitrates of calcium, magnesium, potassium, and sodium. These components are always present in soil and do not pose an environmental problem until concentrations become excessive (Table 3).

Similar to organic contaminants, inorganic compounds have been listed and prioritized by CERCLA,[4] and most have SSLs established by the U.S. Environmental Protection Agency.[5] Radionuclides are not included in these listings because radioactive elements are considered under different legislative initiatives. Table 4 lists 27 inorganic species, their CERCLA priority ranking, and SSLs. The inorganic constituents have a prominent position in the top 10 on the CERCLA priority list, with arsenic being number one and lead, mercury, and cadmium also in the top ten. However, the priority ranking for the remaining inorganic contaminants is generally lower than those for the

TABLE 3 Inorganic Contaminants in Soil, Their Typical Ranges Found in Uncontaminated Soils, and Those Concentrations Considered to Be Unusual

Contaminant	Sources, Impacts	Typical Range in Soil	Extreme Concentrations
Metals Pb, Zn, Cd, Cu, Ni, Ba, Sr, Mn, Cr	Mining, smelting, electroplating	0.1 to 1000 mg/kg	25 to >10,000 mg/kg (metal dependent)
Metalloids As, B, Ge, Sb	Industrial activities, smelting	<0.1 to 100 mg/kg	>50 mg/kg
Nonmetals Se, I, P, S	Manufacturing, processing	<0.1 to 1000 mg/kg	up to 2000 mg/kg (element dependent)
Soluble salts Halides, alkalis, alkaline earths, sulfates, nitrates	Mining, manufacturing, petroleum extraction, natural sources	<500 mg/kg	>1000 mg/kg

TABLE 4 Inorganic Compounds on the CERCLA Priorities List and Their Ranking, Health Impacts, and SSLs

		Soil Screening Level	
		Residential Soil	Protect Groundwater
Constituent	CERCLA Rank	mg/kg	
Antimony	219	4.0×10^{1}	6.6×10^{-1}
Arsenic	1	3.9×10^{-1}	1.3×10^{-3}
Barium	109	1.5×10^{4}	3.0×10^{2}
Beryllium	42	1.6×10^{2}	2.5×10^{1}
Cadmium	8	7.0×10^{1}	1.4×10^{0}
Chlorine	91	7.5×10^{2}	1.6×10^{0}
Chromium(VI)	18	2.9×10^{-1}	8.3×10^{-4}
Cobalt	49	2.3×10^{1}	4.9×10^{-1}
Copper	128	3.1×10^{3}	5.1×10^{1}
Cyanide	28	1.6×10^{3}	7.4×10^{0}
Fluoride	211	3.1×10^{3}	NL[a]
Hydrazine	84	2.1×10^{-1}	NL[a]
Iodine	NL[a]	7.8×10^{2}	NL[a]
Lead	2	4.0×10^{2}	NL[a]
Manganese	117	1.8×10^{3}	5.7×10^{1}
Mercury	3	2.3×10^{1}	5.7×10^{-1}
Molybdenum	NL[b]	3.9×10^{2}	3.7×10^{0}
Nickel	53	1.5×10^{3}	4.8×10^{1}
Nitrate	216	1.3×10^{5}	NL[a]
Nitrite	212	7.8×10^{3}	NL[a]
Perchlorates	NL[b]	5.5×10^{1}	NL[a]
Phosphorus (white)	19	1.6×10^{0}	2.7×10^{-3}
Selenium	147	2.9×10^{2}	9.5×10^{-1}
Silver	214	3.9×10^{2}	1.6×10^{0}
Uranium	98	2.3×10^{2}	4.9×10^{1}
Vanadium	198	5.5×10^{1}	2.6×10^{1}
Zinc	74	2.3×10^{4}	6.8×10^{2}

All screening levels are for residential soils unless stated otherwise. Radionuclides were excluded from this list.

[a] Not listed.

[b] Not listed because radionuclides are treated separately.

organics (Table 2). The residential SSLs for the inorganics tend to be far higher than for the organics and typically have a range of 50 to 3000 mg/kg. Arsenic (0.39), Cr(VI) (0.29), and hydrazine (0.21) are the only inorganics with SSL values less than 1 mg/kg. SSLs for protection of groundwater are roughly 100 times lower than residential SSLs.

Potential Health Impacts of Inorganic Contaminants in Soils

Many of the metals, metalloids, and oxyanions have negative effects on human health, and these impacts have been studied for decades. The most widely distributed and best- known toxic metal in soils is Pb. If Pb-contaminated soil is consumed, the risk increases for lead-associated nervous system, brain, and blood disorders. The impacts of Pb in soil from car exhaust[27] and smelter sites[28] are among the many studies examining this well-known metal. The SSL for Pb is 400 mg/kg.

Cadmium exposure for humans occurs mostly through the consumption of food grown in Cd-contaminated soils. Health impacts of Cd include kidney function, bone strength, and central nervous system disorder. Cadmium contamination in soils can be the result of mining, refining, smelting, battery production, battery disposal, and other industrial operations. Generally, cadmium uptake by plants increases with decreasing pH and increasing total Cd in the soil.[29] The mechanism of cadmium uptake by plants is similar to that of zinc and calcium.

Chromium in soils is viewed to be problematic only if the Cr is in the Cr(VI) (chromate) oxidation state. Reduced chromium, Cr(III), is less mobile and far less toxic. High chromate can result in renal failure, DNA damage, and cancer. Because Cr(VI) is present in the soil as the anion CrO_4^{2-}, it is not strongly held by the soil, tends to be mobile, and is readily assimilated by plants. Fugitive dust from contaminated soil also can lead to increased incidence of cancer.[30] If Cr(VI) can be reduced to Cr(III) in the soil, the health and environmental threats are greatly diminished.[31,32] Labile organic matter, microorganisms, and reduced inorganic species such as Fe^{2+} can readily convert Cr(III) to Cr(VI). Although surface soils in equilibrium with atmospheric O2(g) are predicted to maintain chromium as Cr(VI), even small amounts of organic matter in moist, shallow subsoils can reduce Cr(VI) to Cr(III).

Mercury contamination in humans can result in skin rashes, hypertension, rapid pulse, kidney dysfunction, and memory failure. The most common pathway of exposure to Hg poisoning is through fish and meat, but high concentrations in soils can result in increased exposure through plant uptake and inhalation. The SSL for Hg is 23 mg/kg for residential soils and 0.57 mg/kg to protect groundwater (Table 4). Mercury contamination in soil is fairly uncommon, but in some mining areas, concentrations in the soil as high as 2700 mg/kg have been observed and are associated with plant concentrations as high as 1100 mg/kg. A threshold concentration in the soil is required before plant concentrations increase significantly above background.[33]

In most natural systems, arsenic occurs as one of two redox species: arsenate, AsO_3^{3-}, or arsenite, AsO_4^{3-}, Although both forms appear to induce toxicity in humans, the arsenite is more readily absorbed by human tissues.[34] The residential soil SSL for arsenic is 0.39 mg/kg and 0.013 mg/kg to prevent impacts on groundwater. Most of the attention to arsenic poisoning in recent years has been paid to As in drinking water, particularly as a result of extreme concentrations of total As in wells in Bangladesh. However, plant uptake of As and the exposure of foods to high-As irrigation water also can result in arsenic intake in excess of the World Health Organization's recommended limits.[35]

Remediation of Inorganic Contaminants

The best approach to keeping soil free from contamination is prevention. When this approach fails, removing the contaminants may be necessary, and many options are available.[36] With few exceptions, biological approaches to decontamination of soils with high concentrations of pollutant metals are limited to phytoremediation. Some inorganic contaminants, such as nitrate, chromate, perchlorate, and other redox-active species in which one redox state is less of a threat than another, can be bioremediated through processes in which the chemical state of the contaminant is altered. Nitrate can be denitrified through the production of gaseous nitrogen compounds.[37] Perchlorate can be reduced to benign chloride,[38] and Cr(VI) can be reduced to Cr(III).[39,40] For metals or metalloids in which only one chemical state is generally found in soils, a biological approach not involving phytoremediation is rare.

During the past two decades, phytoremediation of metals in soils has been vigorously explored. For some metals, phytoremediation has been highly successful and has found commercial applications: hyperaccumulation of Zn, Cd, and Ni by *Thlaspi caerulescens*;[41] brake ferns for removal of As from soil;[42] and phytovolatilization of mercury.[43,44] One of the most common metal contaminants in soils in Pb, but effective means of phytoremediation of Pb-contaminated soils have been elusive.[45]

Phytoremediation is not the only remediation alternative for metal-contaminated soils. Traditional and engineering approaches have been used with reasonable success. Although basic excavation methods are effective at cleaning up sites, the soil remains contaminated and is merely transported to a containment facility. Excavation is the least expensive means of remediation but damages the site and does not improve the contaminated soil resource. Soils may be incinerated to burn off organic pollutants and to volatilize metals. Other means of containment include capping the site, encasing the soil in cement, or in situ vitrification in which the soil is melted followed by solidification of the soil into a single mass of glass. These methods are effective in isolating the soil, but considerable waste is generated, the contaminants remain in place, and the soil is usually lost as a viable resource.[46]

Less destructive approaches often are preferred because the soil is not seriously altered or destroyed, and the techniques are often less expensive. Such methods are typified by soil washing,[47] electrokinetics and electromigration processes,[39] and chemical stabilization.[39]

Pathways of Exposure

Once in the soil, organic and inorganic pollutants can follow many pathways to potentially impacted organisms.[48] For all contaminants, direct consumption of contaminated soil can be an important mechanism of exposure for soil-borne organisms, grazing mammals that consume the soil, and humans. For some pollutants, direct soil consumption is the only means of exposure if the contaminants have low mobility and low bioavailability; Pb and five-or six-ring PAHs are examples.

Eight pathways exist by which contaminants in soil can reach humans.[48] The pathways include direct soil consumption; uptake of contaminants from soil into food crops and subsequent consumption by humans; consumption of animals that have been contaminated by tainted soil; inhalation of metals in air or airborne dust; and drinking contaminated groundwater or surface water. Half of the mechanisms involve plant uptake and subsequent consumption of the plants. Most anionic inorganic species are readily assimilated by plants, as are many cationic species. However, some inorganic species form highly insoluble solid phases or are otherwise non-bioavailable, and they are not translocated within the plant. For nonionic organic species, plant uptake is a function of water solubility and hydrophobicity/lipophilicity. Compounds with intermediate lipophilicity are most suitable for plant uptake. Once the pollutants are in the plant, organisms that consume the vegetation may be exposed to the pollutant, and some of the contaminants are passed along the food chain.

Other pathways of exposure include the contaminant moving from the soil to the water or air or becoming airborne on particulates. All of these mechanisms have been shown to contribute to negative health effects. Drinking contaminated water is one of the better-known pathways. If the drinking water originates from groundwater, then the pollutants must be soluble in water and percolate readily through the soil and into the groundwater. When surface water is used for drinking, the transport mechanism can be movement of dissolved contaminant in the water, or contaminants may be adsorbed onto the suspended sediment in the soil runoff. Virtually every soil pollutant can be transported by this mechanism, and only careful control of runoff water and sediments can avoid this problem.

Direct volatilization from the soil into the air is important for a limited number of contaminants, but when volatilization of a potentially toxic compound or element occurs, exposure is directly through the lungs. Mercury (Hg) is a well-known volatile metal, and direct inhalation of high concentrations of gaseous Hg or dimethyl mercury can be poisonous. Inhalation of airborne, contaminated particles is a more frequent occurrence but less insidious. After inhalation, contaminants must be released from the dust before they can be assimilated. Inhalation of contaminated particulates can be reduced through standard soil erosion control measures, such as establishing windbreaks and adequate plant cover on contaminated soils.

Conclusions

Hundreds of contaminants are found in soils, and many reach concentrations that require action. Regulated organic contaminants are usually anthropogenic in their origins and, as a result, their impact can be minimized by controlling their manufacture and use. Even after strict regulations, including outright banning of manufacture or use, detectable or even dangerous concentrations can be found for decades. Residual DDT and its degradation products are examples of highly persistent organic compounds that remain dangerous long after the termination of their use. Metals and other inorganic contaminants can be as problematic as persistent organics because they are not easily transformed to benign forms and they will persist in soils for extended periods. Although most metals do not have a pathway of exposure that could result in human toxicity, some metals (e.g., Se, Mo, Cd, Pb, As) do have such a pathway and must be monitored carefully.

References

1. Pierzynski, G.M.; Sims, J.T.; Vance, G.F. Classification of pollutants. In *Soils and Environmental Quality,* 2nd Ed.; Lewis Publishers: Boca Raton, Florida, 1993; 47–50, 167183, 185–210.
2. Schwarzenbach, R.P.; Gschwend, P.M.; Imboden, D.M. *Environmental Organic Chemistry;* Wiley-Interscience: New York, 1992; 8–41, 255–344.
3. Evangelou, V.P. Water quality. In *Environmental Soil and Water Chemistry*; Wiley Interscience: New York, 1998; 476–499.
4. Agency for Toxic Substances Registry. 2007 CERCLA Priority List of Hazardous Substances, available at http://www.agriculturedefensecoalition.org/sites/default/files/pdfs/4H_2007_CDC_ATSDR_CERCLA_Priority_List_of_Hazard ous_Substances_2007.pdf, (accessed May 2012).
5. United States Environmental Protection Agency. Regional Screening Levels (RSL) Summary Table May 2010, available at http://www.epa.gov/reg3hwmd/risk/human/rb-concentration_table/Generic_Tables/pdf/master_sl_table_run_MAY2012.pdf, (accessed May 2012).
6. Schecter, A.; Colacino, J.; Haffner, D.; Patel, K.; Opel, M. Perfluorinated compounds, polychlorinated biphenyls, and organochlorine pesticide contamination in composite food samples from Dallas, Texas, USA. Environ. Health Per-spect. **2010**, *118* (6), doi:10.1289/ehp.0901347.
7. Quinby, G.E.; Clappison, G.B. Parathion poisoning—A nearfatal pediatric case treated with 2-pyridine aldoxime methio- dide (2-PAM). Arch. Environ. Health **1961**, *3,* 438–447.
8. Calabrese, E.J.; Stanek, E.J.; James, R.C.; Roberts, S.M. Soil ingestion—A concern for acute toxicity in children. J. Environ. Health **1999**, *61,* 18–23.
9. Castelo-Grande, T.; Augusto, P.A.; Monteiro, P.; Estevez, A.M.; Barbosa, D. Remediation of soils contaminated with pesticides: A review. Internat. J. Environ. Anal. Chem. **2010**, *90,* 438–467.
10. Pinedo-Rivello, C.; Aleu, J.; Collado, I.G. Pollutants biodegradation by fungi. Curr. Org. Chem. **2009**, *13,* 11941214.
11. Seo, J.-S.; Keum, Y.S.; Li, Q.X. Bacterial degradation of aromatic compounds. Int. J. Environ. Res. Public Health **2009,** 6, 278–309.
12. Bamforth, S.M.; Singleton, I. Bioremediation of polycyclic aromatic hydrocarbons: Current knowledge and future directions. J. Chem. Technol. Biotechnol. **2005,** *80,* 723736.
13. Stroud, J.L.; Paton, G.I.; Semple, K.T. Microbe-aliphatic hydrocarbon interactions in soil: Implications for biodegradation and bioremediation. J. Appl. Microbiol. **2007,** *102,* 1239–1253.
14. Lewis, T.A.; Newcombe, D.A.; Crawford, R.L. Bioremediation of soils contaminated with explosives. J. Environ. Manag. **2004,** *70,* 291–307.
15. Roberston, S.J.; McGill, W.B.; Massicotte, H.B.; Rutherford, P.M. Petroleum hydrocarbon contamination in boreal forest soils: A mycorrhizal ecosystems perspective. Biol. Rev. **2007,** *82,* 213–240.
16. Hickman, Z.A.; Reid, B.J. Earthworm assisted bioremediation of organic contaminants. Environ. Internat. **2008,** 34, 1072–1081.
17. Semple, K.T.; Reid, B.J.; Fermor, T.R. Impact of composting strategies on the treatment of soils contaminated with organic pollutants. Environ. Pollut. **2001,** *112,* 269–283.
18. Fetzner, S. Bacterial degradation of pyridine, indole, quinoline, and their derivatives under different redox conditions. Appl. Microbiol. Biotechnol. **1998,** *49,* 237–250.

19. Bhatt, P.; Kumar, M.S.; Chakrabarti, T. Fate and degradation of POP-hexachlorocyclohexane. Crit. Rev. Environ. Sci. Technol. **2009,** *39,* 655–695.
20. Zeyaullah, M.; Abdelkafe, A.S.; Ben Zabye, W.; Ali, A. Biodegradation of catechols by micro-organisms—A short review. African J. Biotechnol. **2009,** *13,* 2916–2922.
21. Radosevich, M.; Rhine, E.D. Pollutants (Ch. 281). In *Encyclopedia of Soil Science,* 2005; 1338–1343.
22. Ma, X.; Burken, J.G. TCE diffusion to the atmosphere in phytoremediation applications. Environ. Sci. Technol. **2003,** *37,* 2534–2539.
23. Strand, S.E.; Dossett, M.; Harris, C.; Wang, X. Doty, S.L. Mass balance studies of volatile chlorinated hydrocarbon phytoremediation. Z. Naturforsch. **2005,** *60,* 325–330.
24. Briggs, G.G.; Bromilow, R.H.; Evans, A.A.. Relationship between lipophilicity and root uptake and translocation of nonionized chemicals by barley. Pestic. Sci. **1982,** *13,* 495–504.
25. Burken, J.G.; Schnoor, J.L. Predictive relationships for uptake of organic contaminants by hybrid poplar trees. Environ. Sci. Technol. **1998,** *32,* 3379–3385.
26. Bromilow, R.H.; Chamberlain, K. Principles governing uptake and transport of chemicals. In *Plant Contamination: Modeling and Simulation of Organic Chemical Processes*; Trapp, S, McFarlane, C., Eds.; Lewis Publishers: Boca Raton, Florida, 1995; 37–68.
27. Rhue, R.D.; Mansell, R.S.; Ou, L.T.; Tang, S.R.; Ouyang, Y. The fate and behavior of lead alkyls in the environment—A review. Crit. Rev. Environ. Control **1992,** *22,* 169–193.
28. Aelion, C.M.; Davis, H.T.; McDermott, S.; Lawson, A.B. Soil metal concentrations and toxicity: Associations with distances to industrial facilities and implications for human health. Sci. Total Environ. **2009,** *407,* 2216–2223.
29. Nawrot, T.S.; Staessen, J.A.; Roels, J.A.; Roels, H.A.; Munters, E.; Cuyupers, A.; Richart, T.; Ruttens, A.; Smeets, K.; Clijsters, H.; Vangronsveld, J. Cadmium exposure in the population: From health risks to strategies of prevention. Biometals. **2010,** *23,* 769–782.
30. Le Bot, B.; Gilles, E.; Durand, S.; Glorrenc, P. Bioaccessible and quasi-total metals in soil and indoor dust. Euro. J. Mineral **2010,** *22,* 651–657.
31. Stern, A.H. A quantitative assessment of the carcinogenicity of hexavalent chromium by the oral route and its relevance to human exposure. Environ. Res. **2010,** *110,* 798–807.
32. Dhal, B.; Thatoi, H; Das, N; Pandey, B.D. Reduction of hexavalent chromium by *Bacillus* sp. isolated from chromite mine spoils and characterization of reduced product. J. Chem. Technol. Biotechnol. **2010,** *85,* 1471–1479.
33. Molina, J.A.; Oyarzun, R.; Esbri, J.M., Higuereas, P. Mercury accumulation in soils and plants in the Almaden mining district, Spain: One of the most contaminated sites on Earth. Environ. Geochem. Health **2006,** *28,* 487–498.
34. Saha, J.C.; Dikshit, A.K.; Bandyopadhyay, M.; Saha, K.C. A review of arsenic poisoning and its effects on human health. Crit. Rev. Environ. Sci. Technol. **1999,** *29,* 281–313.
35. Roychowdhry, T. The impact of sedimentary arsenic through irrigated groundwater on soil, plant, crops and human continuum from Bengal delta: Special reference to raw and cooked rice. Environ. Eng. Sci. **2003,** *20,* 405–422.
36. LaGrega, M.D.; Buckingham, P.L.; Evans, J.C. In *Hazardous Waste Management;* McGraw-Hill: New York, 1994; 447–831.
37. Nie, S.W.; Gao, W.S.; Chen, Y.Q.; Sui, P.; Eneji, A.E. Review of current status and research approaches to nitrogen pollution in farmlands. Agric. Sci. China **2009,** *7,* 843–849.
38. Xu, J.L.; Song, Y.U.; Min, B.K.; Steinberg, L.; Logan, B.E. Microbial degradation of perchlorate: Principles and application. Environ. Eng. Sci. **2003,** *20,* 405–422.
39. Jeyasingh, J.; Somasundaram, V.; Philip, L.; Bhallamudi, S.M. Bioremediation of Cr(VI) contaminated soil/sludge: Experimental studies and development of a management model. Chem. Eng. J. **2010,** *2,* 556–564.

40. Mulligan, C.N.; Yong, R.N.; Gibbs, B.F. Remediation technologies for metal-contaminated soils and groundwater: An evaluation. Eng. Geol. **2001,** *1–4,* 197–203.
41. Milner, M.J.; Kochian, L.V. Investigating heavy-metal hyperaccumulation using *Thlaspi caerulescens* as a model system. Ann. Bot. **2008,** *102,* 3–13.
42. Gonzaga, M.I.S.; Santos, J.A.G.; Ma, L.Q. Phytoextraction by arsenic hyperaccumulator *Pteris vittata* L from six arsenic-contaminated soils: Repeated harvests and arsenic redistribution. Environ. Pollut. **2008,** *2*, 212–218.
43. Nagata, T.; Morita, H.; Akizawa, T.; Pan-Hou, H. Development of a transgenic tobacco plant for phytoremediation of methylmercury pollution. Appl. Microbiol. Biotechnol. **2010,** *87,* 781–786.
44. Lomonte, C.; Doronila, A.I.; Gregory, D.; Baker, A.J.M.; Spas, D.K. Phytotoxicity of biosolids and screening of selected plants with potential for mercury phytoextraction. J. Haz. Mater. **2010,** *173,* 494–501.
45. Hadi, F.; Bano, A.; Fuller, M.P. The improved phytoextraction of lead (Pb) and the growth of maize (*Zea mays* L.); the role of plant growth regulators (GA(3) and IAA) and EDTA alone and in combination. Chemosphere **2010,** *80,* 457–462.
46. Fleri, M.A.; Whetstone, G.T. In situ stabilisation/solidification: Project lifecycle. J. Haz. Mater. **2007,** *Special Issue S1,* 141, 441–456.
47. Lestan, D.; Luo, C.L.; Li, X.D. The use of chelating agents in the remediation of metal-contaminated soils: A review. Environ. Pollut. **2008,** *153,* 3–13.
48. Chaney, R.L; Ryan, J.A.; Brown, S. Environmentally acceptable endpoints for soil metals. In *Environmental Availability of Chlorinated Organics, Explosives, and Metals in Soils;* Anderson, W.C., Loehr, R.C., Reible, D., Eds.; American Academy of Environmental Engineers: Annapolis, Maryland, 1999; 111–155.

9

Pollution: Genotoxicity of Agrotoxic Compounds

Vera Lucia S.S.
de Castro and
Paola Poli

Introduction

Thousands of chemicals are in common use, but only a portion of them has undergone significant toxicological evaluation, leading to the need to prioritize the remainder for targeted testing. Various assays were set up and performed to identify the potential hazard (genotoxicity, carcinogenicity, reproductive toxicity, etc.) of pollutants for human health. Increasing need for effective tools to assess the risks derived by the large number of both natural and anthropic-origin organic and inorganic noxious substances, both of natural origin and released in the environment by human activities, leads to the development of very sensitive detectors of harmful substances. It is an accepted assumption that the simple measurement of chemical concentration, with reference to established regulatory rules, will not give an accurate account of the environmental hazard. The measurement of environmental physical–chemical parameters is the first step in monitoring environmental quality; however, attention and alarm thresholds of these parameters only concern the toxic effects of the studied polluting substance but do not take into consideration the issue of chronic exposure at low doses of physical agents and chemicals, frequently present in complex mixtures. On the other hand, the monitoring by biological assays can effectively define the risks for the environment and humans. These analyses may be able to assess the complexity of natural environment, in terms of both different organisms and differences in physiological status, to establish cause/effect relationships between presence and concentration of pollutants and consequent environmental damages and to detect the possible synergistic effect of complex mixtures of chemicals. Therefore, much attention was paid to biological sensors, markers, or detectors able to provide information on the effects of exposure to and/or susceptibility against a variety of environmental contaminants through the knowledge of action mechanisms and the identification of possible endpoints.

Due to their role as bioactive chemicals, pesticides tend to form electrophilic metabolites capable of reacting and combining with biological macromolecules. Preferred sites of action include nucleophilic oxygen as well as nitrogen atoms, both of which are abundant in DNA, predisposing the genetic material to mutagenic covalent binding. One main deleterious effect resulting from exposure to environmental mutagens is the possible initiation of cancer. Other manifestations of genotoxic environmental pollutants such as pesticides include heritable genetic diseases, reproductive dysfunction, and birth defects.[1] Epidemiological studies provide evidence that several types of tumors and other carcinogenic manifestations are in excess among farmers and other occupational groups associated with pesticide handling.[2–4] In addition to epidemiological evidence, laboratory and field monitoring data indicate that increases in chromosome aberration, recombination, sister chromatid exchange (SCE), and other genotoxic events are in excess for pesticide-exposed groups, pointing out a generic genetic activity of pesticides in humans.[5] Recently, a number of laboratory investigations and field studies have documented a correlation between genotoxic pollutants and heritable reproduction effects on individuals as well as a potential link to the declines of fish populations.[6]

Reproductive Toxicity

It is estimated that close to 30% of all pregnancies end in spontaneous abortion. Although about 60% of spontaneous abortions are thought to be due to genetic, infectious, hormonal, and immunological factors, the role of the environment remains poorly understood. Pregnancy involves a delicate balance of hormonal and immunological functions, which may be affected by environmental substances. Many toxic substances that are persistent in the environment and accumulate in the fatty tissues may disrupt this equilibrium.[7]

Several evidence link paternal exposure to genotoxic agents with an increased risk of pregnancy loss, developmental and morphological defects, infant mortality, infertility, and genetic diseases in the offspring, including cancer.[8] Toxicogenomic analysis of environmental chemicals may be performed to investigate the ability of genomics to predict toxicity, categorize chemicals, and elucidate mechanisms of toxicity. The concordance of in vivo observations and gene expression findings demonstrated the ability of genomics to accurately categorize chemicals, identify toxic mechanisms of action, and predict subsequent pathological responses.[9]

The central nervous system appears to be especially susceptible to toxic insults during development and there is evidence that functional changes can be induced at a lower exposure level than those resulting in toxicity in adults. As an example, fetal exposure to some environmental contaminants such as organophosphorus pesticides at apparently non-toxic doses may alter some important behaviors at adulthood in mice.[10] Neurobehavioral performance can be a sensitive biomarker of the neurodevelopmental consequences of exposure of environmental agents.[11,12]

Environmental contaminants as pesticides can come from a variety of sources, including diet, drinking water, and both indoor and outdoor residential use. These chemicals may alter gene expression profiles.[13] Also, some metals can disturb the reproduction process. A number of mechanisms of cadmium toxicity have been suggested, including ionic and molecular mimicry, interference with cell adhesion and signaling, oxidative stress, apoptosis, genotoxicity, and cell cycle disturbance.[14] Some mercury compounds are known as teratogenic agents, especially affecting the normal development of the central nervous system. Since the 1990s, a genotoxic effect has been demonstrated in human populations exposed to mercury through diet,[15] occupation,[16] or carrying dental fillings.[17]

In fact, concentrations of methylmercury causing significant genotoxic alterations in vitro below both safety limit and concentration were associated with delayed psychomotor development with minimal signs of methylmercury poisoning. Based on mercury's known ability to bind sulfhydryl groups, several hypotheses were raised about potential molecular mechanisms for the metal genotoxicity. Mercury may be involved in four main processes that lead to genotoxicity: generation of free radicals and oxidative stress, action on microtubules, influence on DNA repair mechanisms, and direct interaction with

DNA molecules.[18] DNA damage can also be a sensitive bioindicator of mercury contamination in other organisms as in fish.[19]

Some pesticides may act as endocrine disruptors. As an example, the fungicide fenarimol acts both as an estrogen agonist and as an androgen antagonist.[11] In addition, fenarimol affects rat aromatase activity in human tissues.[20] This compound also affects other enzymes of the cytochrome P450 gene family that are involved in the metabolism of steroids.[21] Studies of reproductive toxicity in rats have shown that, as a result of fenarimol maternal exposure, some neuromuscular and behavioral deficits in nursing pups may occur principally during the last gestational period and lactation.[12]

Pesticides may also operate through hormonal or genotoxic pathways to affect male reproduction. They may penetrate the blood–testis barrier to potentially affect spermatogenesis, by affecting either genetic integrity or hormone production. Effects may be at different stages of the cell cycle such as during meiotic disjunction, and such abnormalities can have deleterious effects on reproduction and offspring.[22]

Human and Environmental Genotoxicity

Mutation is a manifestation of change of the structure of DNA. Agents that cause DNA damage and mutation, as well as being potentially capable of causing hereditary disorders in the offspring and succeeding generations of exposed populations, are also likely to be carcinogenic. For these reasons, testing for the induction of DNA damage and for mutagenicity, using a variety of short-term tests, has become an accepted part of the toxicological evaluation of drugs, industrial intermediates, cosmetics, food and feed additives, pesticides, biocides, etc. Regulatory agencies and international authorities recommend a test scheme consisting of *in vitro* and *in vivo* methods to identify genotoxic/mutagenic substances as those described on Organization for Economic Cooperation and Development (OECD) and United States Environmental Protection Agency (USEPA) guidelines, as recent examples.[23,24]

The current EPA mutagenicity testing battery is required for pesticides and toxic substances that are regulated under the Federal Insecticide, Fungicide and Rodenticide Act (FIFRA) and the Toxic Substances Control Act (TSCA), respectively. The battery is a three-tiered system of various mutagenicity tests (including the *Salmonella,* mouse lymphoma, and mouse micronucleus assays in the first tier). Guidelines for the conduct of the tests employed in this battery were issued and were ultimately harmonized within EPA and with the OECD.[25]

In addition to the risks to human health posed by mutagenic pollutants such as pesticides, there are important ecological risks as well, such as the threat posed by pesticides to the stability of the ecosystems through the cumulative introduction of deleterious mutations into the genetic pool of populations. Indeed, it has been demonstrated that pesticides and other environmental genotoxicants are capable of altering the genetic makeup of some natural populations.[26,27]

The occurrence of genotoxic pollutants in the aquatic environment is of increasing concern. Pollution with anthropogenic toxicants may create pronounced environmental gradients that impose strong local selection pressures. Toxic contaminants may also directly impact genetic structure in natural populations by exhibiting genotoxicity. The genetic variation within natural populations and, hence, the potential for local adaptation can itself be impacted by anthropogenic pollution.[26]

Reservoirs are complex aquatic systems mediating between rivers and lakes; they usually reflect multiple impacts generated by a variety of anthropogenic activities. The sediment compartment is the intermediate or final receptor of insoluble (or slightly water soluble) pollutants and can act as a sink for various substances. Sediments accumulate chemicals up to concentrations many times higher than in the free water column. As pollutants may be made available under certain environmental conditions (such as dredging or flood events), sediments can also become a source of diffuse contamination to the free water space. Sediment pollutants are linked not only to organisms in aquatic ecosystem dynamics but also to human health via water and fish consumption. Due to their ability to metabolize xenobiotics and accumulate pollutants, fish represent important monitoring systems within aquatic genotoxicity assessment.[6]

In general, chemical mixtures may influence local ecosystems in a site-specific way defined by all aspects along the source and availability of the mixture and the nontarget receptor. Also, humans are concurrently exposed to a number of chemicals via food and environment. These chemicals may have a combined action that causes a lower or higher toxic effect than would be expected from knowledge about the single compounds. Consequently, combined actions need to be addressed in the risk assessment process. In addition, developments in the area of toxicogenomics have also been suggested as a way of increasing our knowledge of mechanism of toxicity in order to better understand and improve the approaches for risk assessment of combined actions of chemicals.[28]

Both chronic and acute contamination of watershed by pesticides dissolved or adsorbed to soil particles can affect aquatic organisms. Pesticide toxicity may occur in a broad range of non-target aquatic organisms, both in plants from microalgae to macrophytes and in animals from microinvertebrates to fish predators. The possible effects of pesticide exposure on fish are of interest because of the position of fish in the food chain and because early life stages of fish have been shown to be highly sensitive to pollutants. Genotoxicity assessment in fish has been highlighted since the implications of the genotoxic effects are impacting on fitness traits such as reproductive success, genetic patterns, and subsequent population dynamics.[29]

Pesticide Genotoxicity Assessment

Traditionally, the impact of pollutants discharged into the environment by human activities has been assessed using chemical assays or by evaluating physical parameters. High-performance liquid chromatography or gasmass spectrometry techniques have been widely used. However, the information is limited to the concentration of the pollutants, not on their toxicity. To overcome these limitations, biological analyses have been introduced. The use of effect-based screening tools has the advantage of indicating the real impact of all chemicals present in a given sample or ecosystem. Rapid biological tests are playing their major role in hazard and risk assessment, especially at the screening level. Bioassays measure changes in physiology or behavior of living organisms resulting from stresses induced by biological or chemical toxic compounds, which can cause disruption of the metabolism.[30]

Toxicological animal-based assays on organisms from different trophic levels (algae and plants, worms and crustaceans, fishes, etc.), employed in the identification of hazardous chemicals are sometimes expensive and time consuming, require large sample volumes, and raise ethical problems. In vitro methods, also commonly used for screening and ranking chemicals, must be included in battery tests for risk assessment purposes. Their major promise is to supply mechanism-derived information, considered pivotal for adequate risk assessment. Mutagenesis short-term assays can directly detect the genetic effect of chemical and physical agents on the tested cells/organisms, able to assess the DNA damage resulting from exposure to both single chemicals and heterogeneous mixtures.

Fundamental research in the mechanisms of induced mutation and carcinogenesis induction has benefited from the development of highly refined short-term tests for genotoxicity. The knowledge of toxicity pathways that is derived from genomic data would highlight the potential mechanisms. The challenge is to comprehensively integrate the disparate chemical, biological, toxicological, and toxicogenomic data in order to elucidate the mechanisms and networks involved in toxicity and to develop quantitative models capable of accurately predicting thresholds.[31]

However, despite the progress on newer methodologies, quantitative data on toxicological effects of some widely used pesticides in agricultural practices are not well established even at the single-organism level. Among the pesticides in common use, attention is drawn towards the use of triazole fungicides that are used in the control of several fungal and plant diseases. These fungicides have demonstrated some adverse effects in mammalian species derived from genomic data.[32] Also, genotoxicity studies conducted in microalgae showed that the exposure to epoxiconazole can result in an increase in the extent of DNA strand breaks depending on the concentration.[33]

Nevertheless, results from this kind of assays are prone to the criticisms concerning the differences between the real conditions in the ecosystems and those of the in vitro assays. In fact, since various factors (chemical, physical, and biological) affect environmental conditions, the transfer of results obtained by in vitro techniques to the field is a complex task, and the establishment of extrapolation parameters is a crucial issue. In this context, a combination of in vitro and in situ (in vivo) bioassays represents a promising approach to better understand both the real exposure situation in the environment and the action mechanisms studied by the in vitro bioassays.[34]

Also, although the genotoxicity testing strategies employed prior to product registration are designed to identify potential in vivo genotoxins, concerns that exposure to pesticides may result in long-term adverse effects still exist.[35] A positive association between occupational exposure to complex pesticide mixtures and the presence of chromosomal aberrations (CAs), SCEs, and micronuclei has been detected in a number of studies, although some of these failed to detect cytogenetic damage. Chromosomal damage induced by pesticides appears to have been transient in acute or discontinuous exposure, but cumulative in continuous exposure to complex agrochemical mixtures.[36]

Assessment Methodology

Human and wildlife risks associated with pesticide mutagenic capability can be assessed by using screening systems that are sensitive and able to detect the whole mutagenic spectrum. There are more than 100 short-term bioassays for evaluating the potential genotoxic effects of pesticides, but since no single system encompasses the whole spectrum of possible genetic toxic effects, a combination of evaluation procedures is recommended for the assessment of pesticide mutagenesis. Genotoxicity testing batteries were specifically established for hazard identification, the first step in risk assessment. Studies of sensitivity and correlation among test systems are fundamental for a more accurate evaluation of the environmental risks, as well as extrapolation of data to other target organisms.

As a consequence of the very large number of genetic toxicity evaluation techniques described in the specialized literature, most assays can be considered ancillary and will be employed only for specific ends.[37] The assays accepted for routine evaluation of pesticides fall in one of six testing categories: 1) microbial assays: a) prokaryotic (bacterial, such as the *Salmonella typhimurium, Escherichia coli, Bacillus subtilis*) or b) eukaryotic assays (fungi, such as *Saccharomyces cerevisiae, Neurospora crassa*); 2) in vitro isolated eukaryotic cell lines; 3) host-mediated assays; 4) in vivo animal; and 5) in vivo plant bioassays.

In order to approximate the studies to higher-organism subjects, several assays involve in vitro exposure of mammalian cell lines. Mouse bone marrow cells, erythrocytes, and white blood cells; hamster ovary cells; and human lung fibroblast are some examples, and several genetic endpoints can be evaluated, such as DNA damage, unscheduled DNA synthesis, chromatid exchanges, and micronucleus frequency. Aside from the practical advantages offered by the microbial and isolated cell line bioassays (ease of manipulation, asepticism, small space and large population assayed, low cost), the tests often depend upon an external metabolic activation complement, since the microbial and isolated cell lines may be incapable of responding to pro-mutagenic compounds if these are not partially metabolized.[38] The host-mediated assays are devised to circumvent this deficiency. Mammalian subjects are exposed to the pesticide; the microbial cell line is injected into this treated subject and later recovered and evaluated for mutation induction. Alternatively, for plant metabolic activation, whole plants are treated with the pesticide and their extracts are applied directly into the microbial assay.

Finally, plant bioassays involve a variety of endpoints, from DNA damage in leaf cells of *Impatiens balsamina* to micronuclei in pollen mother cells (*Tradescantia*) and root tip meristematic cells (*Allium*, *Vicia*), reversion and crossing over in chlorophyll-deficient lines (*Hordeum*, *Glycine*), sugar-specific starch production in pollen (*Zea*), flower pigmentation alteration (*Tradescantia*), and many other endpoints in many different species.[39–41]

Some recent studies[42–49] mainly use bioassays such as Comet, micronucleus, CAs, and SCE to detect pesticide genotoxic potential both in vitro and in vivo.

Comet Assay

The single-cell gel electrophoresis assay (or Comet assay) is a simple, rapid, and sensitive technique for analyzing and measuring DNA damage in individual mammalian (and to some extent prokaryotic) cells. The method published by Sing et al.[50] makes the Comet assay versatile and forms the basis for all developments that have taken place in this field. In the Comet assay, the cells are embedded in a thin agarose gel on a microscope slide. The cells are lysed to remove all cellular proteins and the DNA subsequently allowed unwinding under alkaline/neutral conditions. Following unwinding, the DNA is electrophoresed and stained with a fluorescent dye. During electrophoresis, broken DNA fragments (damaged DNA) or relaxed chromatin migrates away from the nucleus. The Comet assay essentially measures the sizes of DNA fragments within the cell. It is therefore necessary to convert DNA damage to DNA fragments by introducing breaks at the sites of DNA damage before they can be detected on Comet assay. The simplest types of DNA damage detected by Comet assay are the double-strand breaks (DSBs). DSBs within the DNA result in DNA fragments and can be detected by merely subjecting them to electrophoretic mobility at the neutral pH. Single-strand breaks (SSBs) do not produce DNA fragments unless the two strands of the DNA are separated/denatured. This is accomplished by unwinding the DNA at pH 12.1. Other types of DNA damage broadly termed as alkali labile sites are expressed when the DNA is treated with alkali at pH greater than 13.

The in vivo Comet assay is a well-established genotoxicity test.[51–55] It is currently mainly performed with somatic cells from different organs to detect a genotoxic activity of potential carcinogens. It is regarded as a useful test for follow-up testing of positive or equivocal in vitro test results and for the evaluation of local genotoxicity. In current test strategies, the in vivo Comet assay is mainly performed with somatic cells from different organs to detect the genotoxic activity of potential carcinogens and is regarded as a useful test for follow-up testing of positive in vitro findings. Furthermore, the comet assay also has the potential to be a useful tool for investigating genotoxicity in germ cells.[56]

Micronucleus Assay

The purpose of the micronucleus assay is to detect chromosome structure modifying agents: segregation is the way to lead the induction of micronuclei in interphase cells. These micronuclei may originate from acentric fragments (chromosome fragments lacking a centromere) or whole chromosomes unable to migrate with the remainder of the chromosomes during the anaphase of cell division.[57–59] The in vitro micronucleus assay is a mutagenic test system for the detection of chemicals that induce the formation of small membrane-bound DNA fragments, i.e., micronuclei in the cytoplasm of interphase cells.

After exposure to a test substance and addition of cytochalasin B for blocking cytokinesis, cell cultures are grown for a period sufficient to allow chromosomal damage to lead to the formation of micronuclei in bi- or multinucleated interphase cells. Harvested and stained interphase cells are then analyzed microscopically for the presence of micronuclei. Micronuclei are scored in those cells that complete nuclear division following exposure to the test item. Additionally, the cells are classified as mononucleates, binucleates, or multinucleates to estimate the proliferation index as a measure of toxicity.

The micronucleus test is also an efficient biological assay for monitoring population exposure to mixtures of agrochemicals as shown in a study that was performed in farm workers directly exposed to large amounts of agrochemicals (fungicides, insecticides, and herbicides) in an area of grain farming (wheat and soybeans).[36,60,61] High MN frequency was detected in people at higher cancer risk due to occupational/environmental exposure to a wide variety of carcinogens.[62–66]

CA Test

Chromosome mutations and related events are the cause of many human genetic diseases, and there is substantial evidence that chromosome mutations and related events causing alterations in oncogenes and tumor suppressor genes of somatic cells are involved in cancer induction in humans and experimental animals.

The purpose of the in vitro CA test is to identify agents that cause structural chromosome aberrations in cultured cells.[67,68] With the majority of chemical mutagens, induced aberrations are of the chromatid type, but chromosome-type aberrations also occur.

Structural CAs may be induced by direct DNA breakage, by replication on a damaged DNA template, by inhibition of DNA synthesis, and by other mechanisms (e.g., topoisomerase II inhibitors).[69] Based on morphological criteria, structural CAs can be divided into two main classes: chromosome-type aberrations (CSAs), involving both chromatids of one or multiple chromosomes, and chromatid-type aberrations (CTAs) involving only one of the two chromatids of a chromosome or several chromosomes.[69,70] An increase in polyploidy may indicate that a chemical has the potential to induce numerical aberrations.[70] CTAs (e.g., chromatid type breaks and exchanges) arise predominantly in vitro during S-phase of cultured lymphocytes, in response to base modifications and SSBs induced in vivo by S-phase-dependent clastogens.[69,70]

Proliferating cells are treated with the test substance in the presence and absence of a metabolic activation system for 3–6 hr and sampled at a time equivalent to about 1.5 normal cell cycle length after the beginning of treatment.[71] If this protocol gives negative results both with and without activation, an additional experiment without activation should be done, with continuous treatment until sampling at a time equivalent to about 1.5 normal cell cycle lengths. Certain chemicals may be more readily detected by treatment/sampling times longer than 1.5 cycle lengths. At predetermined intervals after exposure of cell cultures to the test substance, they are treated with a metaphase arresting substance (e.g., colchicine), harvested, and stained, and metaphase cells are analyzed microscopically for the presence of chromosome aberrations. Though the purpose of the test is to detect structural chromosome aberrations, it is important to record polyploidy and endoreduplication when these events are seen.

Structural CAs in peripheral blood lymphocytes as assessed by the chromosome aberration assay in vivo have been used for more than 30 years in occupational and environmental settings as a biomarker of early effects of genotoxic carcinogens. A high frequency of structural CAs in lymphocytes (reporter tissue) is predictive of increased cancer risk.[36,60,72]

Sister Chromatid Exchange

SCE is the exchange of homologous stretches of DNA sequence between sister chromatids and occurs normally in cells during mitosis. In the presence of genotoxic agents that provoke DNA damage, the rate of SCE increases. Equal SCE has been thought to be an important mechanism of DSB repair in eukaryotes, but this has never been proven due to the difficulty of distinguishing SCE products from parental molecules.[73]

The SCE analysis was adopted as an indicator of genotoxicity. SCEs represent the interchange of DNA replication products at apparently homologous loci. The exchange process presumably involves DNA breakage and reunion, although little is known about its molecular basis.[73–80]

Some studies revealed that the nucleotide pool imbalance can have severe consequences on DNA metabolism and it is critical in SCE formation. The modulation of SCE by DNA precursors raises the possibility that DNA changes are responsible for the induction of SCE in mammalian cells.[81,82]

Detection of SCEs in in vitro test requires some means of differentially labeling sister chromatids, and this can be achieved by incorporation of bromodeoxyuridine (BrdU) into chromosomal DNA for two cell cycles.[83,84] Cells in an exponential stage of growth are exposed to the test substance for a suitable period of time; in most cases, 1–2 hr may be effective, but the treatment time may be extended up to two complete cell cycles in certain cases. Cells without sufficient intrinsic metabolic activity should be exposed to the test chemical in the presence and absence of an appropriate metabolic activation system. At the end of the exposure period, cells are washed free of test substance and cultured for two rounds of replication in the presence of BrdU. As an alternative procedure, cells may be exposed simultaneously to the test chemical and BrdU for the complete culture time of two replication cycles.[85] Cells are analyzed in their second posttreatment division, ensuring that the most sensitive cell cycle stages have been exposed to the chemical. Cell cultures are treated with a spindle inhibitor (e.g., colchicine) 1–4 hours prior to harvesting. Chromosome preparations are made by standard cytogenetic techniques. Staining of slides to show SCEs can be performed by several techniques (e.g., the fluorescence plus Giemsa method).[85,86]

While increased levels of CA have been associated with increased cancer risk,[87,88] a similar conclusion has not been reached for SCE. However, high levels of SCE frequency have been observed in persons at higher cancer risk due to occupational or environmental exposure to a wide variety of carcinogens.[62–66]

Biomonitoring

Plants

Higher plants are recognized as excellent genetic models to detect environmental mutagens and are frequently used in monitoring studies.[89–94] Plant systems represent a complex multicellular environment where the efficiency of different protection or repair mechanisms can be modulated by cellular homeostasis. Higher plants, even showing low concentrations of oxidase enzymes and a limitation in the substrate specification in relation to other organism groups,[40] present consistent results that may serve as a warning to other biological systems, since the target is DNA, common to all organisms.[95]

They represent a stable sensor in an ecosystem and hence allow following the evolution of the genotoxic impact. Well-defined higher plants represent an excellent basis for cytogenetic evaluations after exposure to genotoxic pollutants, especially since the maturation of their gametes (meiosis) follows the same patterns as in animals and humans,[96] although sometimes plant and animal assays are differentially responsive to some pesticides as pendimethalin.[97]

Among the plant species, *Allium cepa* has been widely used to evaluate DNA damage, such as chromosome aberrations and disturbances in the mitotic cycle. Employing *A. cepa* as a test system to detect mutagens dates back to the 1940s.[40,98,99] It has been used to this day to assess a great number of chemical agents, which contributes to its increasing application in environmental monitoring. *A. cepa* is characterized as a low-cost test. It is easily handled and has advantages over other short-term tests that require previous preparations of the samples, as well as the addition of exogenous metabolic system. *A. cepa* test also enables the evaluation of different endpoints (mitotic index, chromosome aberrations, nuclear abnormalities, and micronucleus).[40,99,100] Among the endpoints, chromosome aberrations have been the most used one to detect genotoxicity along the years. The mitotic index and some nuclear abnormalities are used to evaluate both cytotoxicity and mutagenicity of different chemicals. Moreover, *A. cepa* test system provides important information to evaluate action mechanisms of an agent about its effects on the genetic material (clastogenic and/or aneugenic effects). Transgenic/transformed plants could be used as a model to better check whether, in response to environmental stress, DNA damage might be regulated by alterations in single genes and to understand the mode and mechanisms of DNA damage.[101]

Animals and Human

The response to a genotoxic agent can range considerably in the exposed populations. It is currently accepted that susceptibility to genotoxic exposure varies interindividually and that this could be the result of hereditary or acquired characteristics. For this reason, attention has been focused on genetic polymorphisms, which are able to modulate human response to genotoxic environmental agents. This individual variation suggests the importance of individual susceptibility factors. Many studies[102–108] showed the influence of metabolic and DNA repair polymorphisms in the individual response to exposure to carcinogens. Genetic polymorphisms of metabolizing enzymes widely influence xenobiotic effective dose.[109] On the other hand, the role of polymorphisms linked to the DNA repair genes appears to be essential in the modulation of genotoxic risk linked to carcinogen exposure.[105,107,110]

Tests for Some Novel Materials: The Case of Nanomaterials

Nanomaterials are generally defined as having one or more external dimensions or an internal or surface structure on the nanoscale (about 1–100 nm). Nanomaterials display novel properties (small size, particular shape, large surface area, and large surface activity) that make them attractive in many applications. However, these properties may contribute to their toxicological profile and may also affect nanomaterials' possible direct or indirect interaction with the DNA.[111]

Although it has been shown that cationic functionalized carbon nanotubes can condense with DNA and that gold nanoparticles binding to the major groove of DNA is associated with killing of cancer cells, for most nanomaterials, it is even unknown whether they directly interact with DNA or whether they promote indirect effects. The identification of the different ways by which various nanomaterials interact with DNA will improve the extrapolations of genotoxicity test results to human risk evaluation. Application of standard methods to nanomaterials demands, however, several adaptations, and the interpretation of results from the genotoxicity tests may need additional considerations. The use of a battery of standard genotoxicity testing methods covering a wide range of mechanisms towards establishing methodological adaptations and better test conditions is a practical and pragmatic approach for this genotoxic evaluation.[111]

Recent reviews have concluded that information on the genotoxicity of nanomaterials is still inadequate for general conclusions, e.g., on its characteristics critical for genotoxicity. It is presently unclear how well standard genotoxicity tests, designed for soluble chemicals, can be used to assess the genotoxicity of nanomaterials. It appears that more genotoxicity studies on nanomaterials are urgently required, and the need for novel methods to assess nanomaterials–genome interactions is imminent. Current evidence indicates that various nanomaterials may carry genotoxic potential although the mechanisms of which remain to be identified.[112]

Above all, some nanomaterials can generate great concern about their possible adverse effects. As an example, although titanium dioxide is frequently used for industrial and cosmetics purposes because of its low toxicity, nanosized or ultrafine TiO_2 (UF-TiO_2) (<100 nm in diameter) can generate pulmonary fibrosis and lung tumor in rats and cytotoxicity in rat lung alveolar macrophages.[113] These authors also observed that it can cause genotoxicity and cytotoxicity in cultured human cells in experimental models as well.

TiO_2 nanoparticles in the absence of photoactivation are potentially genotoxic to fish cells under in vitro conditions. This effect becomes more pronounced in the presence of UVA, along with cytotoxic effects, which only occurred during combined exposure to TiO_2 and UVA.[114]

Although the biological effects of some nanomaterials have already been assessed, information on toxicity and possible mechanisms of various particle types are insufficient. The comparative analysis demonstrated that particle composition probably played a primary role in the cytotoxic effects of different nanoparticles, whereas the genotoxicity potential might be mostly attributed to particle shape.[115] Furthermore, the genotoxic effectiveness could be mediated through lipid peroxidation and oxidative stress as in the case of ZnO nanoparticles on human epidermal cells, even at low concentrations.[116]

Future Concerns and New Genomic Techniques

The evidence of metabolic activation of pesticides into direct-acting mutagens and the complex interactions that may occur as a result of the synergistic action of the multitude of chemicals presently under common usage warrant devoted attention to the evaluation of the potential genetic consequences of pesticide mixtures and derived metabolites accumulated through food chains. This complexity is further deepened by the unpredictability of mutational effects, as a result of pleiotropism and other unforeseeable events. Studies of this nature should attempt to 1) point out the impossibility of predicting all genetic effects of pesticides and 2) recognize that small, improbable effects may have important consequences when imposed onto the very large populations exposed to pesticides in the environment.

With the advent of new technologies (e.g., genomics, automated analyses, and in vivo monitoring), new regulations (e.g., the reduction of animal tests by the European Registration, Evaluation, and Authorization of Chemicals), and new approaches to toxicology (e.g., Toxicity Testing in the 21st Century, National Research Council), the field of regulatory genetic toxicology is undergoing a serious re-examination. However, it is appropriate to apply a prudent approach to risk assessment, maintaining current testing standards that are working properly until others have proven superior by rigorous scientific evidence and widespread agreement.[117]

Human cytogenetics has proven to be effective over a 50 years life span. It is in reality a collection of techniques that, while common, are cheap, fast, and wide ranging. Therefore, in genotoxicology, they continue to be useful to identify mutagenic agents as well as to evaluate and analyze exposed populations.[118]

The introduction of advanced molecular techniques leads to improved risk assessment and also provides an alternative to the massive use of animal testing. Transcriptional profiling and DNA chips are highly informative and are among the most promising novel techniques for environmental risk assessment. Moreover, information discerned from these chips enables the identification of new discriminative biomarker genes. Based on these biomarker genes, cellular reporters can be constructed. These can be used in a high-throughput setup and can significantly facilitate ecotoxicological risk assessment.[119] However, some important technical and interpretative hurdles still need to be overcome before a full implementation of ecotoxicogenomics in regulatory settings can occur. Toxicogenomics uses molecular biology's entire arsenal to analyze the changes in DNA induced by genotoxic agents.[120,121] These changes include those induced in certain genes not related to the mechanisms of genotoxicity, like early response genes. These genes can be used as indicators for which there is a stress response to the genotoxic agent. Quantitative real-time polymerase chain reaction (RT-PCR) and microarrays are the most important procedures to analyze these genes, which can be numerous since current technology allows it.

RT-PCR uses fluorescent probes to measure the exact amount of a nucleic acid. In genotoxicology, it can be used to quantify gene expression to detect genetic polymorphisms and to quantify chromosome deletions. These probes are used to verify the expressed gene or to selectively analyze its expression over time or dose parameters. It will also become more important to analyze expression in specific cell populations in order to profile the global alterations in gene expression involved in chronic chemical exposure that may lead to tumor development.[122]

The results of the microarray studies in toxicogenomics show that genotoxic agents are able to induce changes in gene expression profiles. These studies demonstrate that the gene expression pattern is able to generate the information necessary to determine the agent's type of action and even to predict it. Chemically induced changes in gene expression can result in the identification of simple, sensitive, and relevant biomarkers of effect that can be used in dose–response studies to more readily identify precursor effects in the low-dose range.[123] Therefore, the ability to analyze the effect of mutagenic agents on a large number of genes in a single experiment using gene expression profiling analysis has been used to demonstrate that certain genes may or may not become activated when exposed to a toxic agent, depending on the type of cells and the type of genotoxic agent they have been exposed to.[124]

Furthermore, the DNA damage itself can activate genes codifying proteins involved in DNA repair and/or induce the apoptosis process, when the genes associated with this process are activated. Although broad consensus is still lacking concerning the guidelines for the reproducibility of experiments, toxicogenomic studies could lead to development of early biomarkers of toxic injury and may also help to resolve issues related to interspecies extrapolation and susceptibility variation among individuals.[122] As an application of this toxicogenomic approach, a study[32] reports data that linked genomic data to specific toxicological endpoints of the antifungal triazole, which causes varying degrees of hepatic toxicity and disrupts steroid hormone homeostasis in rodent in vivo models. Overall, these analyses revealed functional categories of chemical response genes that indicate mechanisms and provide direction for further research on triazole mechanisms of action.

Conclusion

The occurrence of genotoxic pollutants in the environment is of increasing concern. Genotoxicity studies can be applied to elucidate potential mechanisms of physiological or molecular alterations to minimize exposure of contaminants to levels that maintain sustainability of the environment. There is a large number of genetic toxicity evaluation assays described in the specialized literature that are able to identify action modes and biological targets both in the environment and in the organisms. Microorganisms as well as cell lines are widely used as in vitro models for risk assessment. Animal and plant models may be valuable for a better comprehension of the metabolic pathways of genotoxic chemicals. Various methods can be applied both in vitro and in epidemiological studies. New approaches through molecular techniques are also proposed for testing the genotoxic load of environmental pollutants. Moreover, in order to better establish biological hazards, it is also necessary to study the effects induced by chemicals on organism development, physiology, and behavior to evaluate their consequences. In relation to this, few attempts to link the genotoxic effects of contaminants with effects at the physiological or behavioral level have been made, which would lead to a detrimental impact on vital life processes including the reproductive potential of the organisms.

Acknowledgments

We wish to thank Dr. Stachetti for his precious contribution to this entry, consisting of the initial text, over which the present one has been constructed. Also, we would like to thank him for the opportunity offered to contribute to this prestigious publication, by referring our names as authors. Without his help and his confidence on our work, this achievement would not have been possible. We also would like to thank Dr. Stachetti for the final text comments and suggestions.

References

1. Waters, M.; Sandhu, S.S.; Simmon, V.F.; Mortelmans, K.E.; Mitchell, A.D.; Jorgenson, T.A.; Jones, D.C.L.; Valencia, R.; Garrett, N.E. Study of pesticide genotoxicity. In Genetic Toxicology: An Agricultural Perspective; Fleck, R.A., Hollaender, A., Eds.; Basic Life Sciences, Plenum Press: New York, 1982; Vol. 21, 275–326.
2. Maele-Fabry, G.; Duhayon, S.; Mertens, C.; Lison, D. Risk of leukaemia among pesticide manufacturing workers: A review and meta-analysis of cohort studies. Environ. Res. **2008,** *106*, 121–137.
3. Rull, R.P.; Gunier, R.; Behren, J.; Hertz, A.; Crouse, V.; Buffler, P.A.; Reynolds, P. Residential proximity to agricultural pesticide applications and childhood acute lymphoblastic leukemia. Environ. Res. **2009,** *109*, 891–899.
4. Alavanja, M.C.R.; Hoppin, J.A.; Kamel, F. Health effects of chronic pesticide exposure: Cancer and neurotoxicity. Annu. Rev. Public Health **2004,** *25*, 155–197.
5. Maroni, M.; Fait, A. Health effects in man from long-term exposure to pesticides—A review of the 1975–1991 literature. Toxicology **1993,** *78* (1–3), 1–180.
6. Rocha, P.S.; Luvizotto, G.L.; Kosmehl, T.; Bottcher, M.; Storch, V.; Braunbeck, T.; Hollert, H. Sediment genotoxicity in the Tiete River (São Paulo, Brazil): In vitro comet assay versus in situ micronucleus assay studies. Ecotoxicol. Environ. Saf. **2009,** *72*, 1842–1848.
7. Weselak, M.; Arbuckle, T.; Walker, M.; Krewski, D. The influence of the environment and other exogenous agents on spontaneous abortion risk. J. Toxicol. Environ. Health, Part B 2008, *11* (3–4), 221–241.

8. Aitken, R.J.; Koopman, P.; Lewis, S.E. Seeds of concern. Nature **2004,** *432*, 48–52.
9. Martin, M.T.; Brennan, R.J.; Hu, W.; Ayanoglu, E.; Lau, C.; Ren, H.; Wood, C.R.; Corton, J.C.; Kavlock, R.J.; Dix, D.J. Toxicogenomic study of triazole fungicides and perflu-oroalkyl acids in rat livers predicts toxicity and categorizes chemicals based on mechanisms of toxicity. Toxicol. Sci. **2007,** *97* (2), 595–613.
10. Ricceri, L.; Venerosi, A.; Capone, F.; Cometa, M.; Lorenzini, P.; Fortuna, S.; Calamandrei, G. Developmental neurotoxicity of organophosphorous pesticides: Fetal and neonatal exposure to chlorpyrifos alters sex-specific behaviors at adulthood in mice. Toxicol. Sci. **2006,** *93* (1), 105–113.
11. Andersen, H.; Vinggaard, A.; Rasmussen, T.; Gjermandsen, I.; Bonefeld-Jørgensen, E. Effects of currently used pesticides in assays for estrogenicity, androgenicity, and aromatase activity in vitro. Toxicol. Appl. Pharmacol. **2002,** *179* (1), 1–12.
12. Castro, V.L.; Mello, M.A.; Diniz, C.; Morita, L.; Zucchi, T.; Poli, P. Neurodevelopmental effects of perinatal fenarimol exposure on rats. Reprod. Toxicol. **2007,** *23*, 98–105.
13. Tully, D.B.; Bao, W.; Goetz, A.K.; Blystone, C.R.; Ren, H.; Schmid, J.E.; Strader, L.F.; Wood, C.R.; Best, D.S.; Narotsky, M.G.; Wolf, D.C.; Rockett, J.C.; Dix, D.J. Gene expression profiling in liver and testis of rats to characterize the toxicity of triazole fungicides. Toxicol. Appl. Pharmacol. **2006,** *215*, 260–273.
14. Thompson, J.; Bannigan, J. Cadmium: Toxic effects on the reproductive system and the embryo. Reprod. Toxicol. **2008,** *25*, 304–315.
15. Pinheiro, M.C.N.; Crespo-Lopez, M.E.; Vieira, J.L.F.; Oikawa, T.; Guimarães, G.A.; Araujo, C.C.; Amoras, W.W.; Ribeiro, D.R.; Herculano, A.M.; do Nascimento, J.L.M.; Silveira, L.C.L. Mercury pollution and childhood in Amazon riverside villages. Environ. Int. **2007,** *33*, 56–61.
16. Zachi, E.C.; Ventura, D.F.; Faria, M.A.; Taub, A. Neuropsychological dysfunction related to earlier occupational exposure to mercury vapor. Braz. J. Med. Biol. Res. **2007,** *40* (3), 425–433.
17. Di Pietro, A.; Visalli, G.; La Maestra, S.; Micale, R.; Baluce, B.; Matarese, G.; Cingano, L.; Scoglio, M.E. Biomonitoring of DNA damage in peripheral blood lymphocytes of subjects with dental restorative fillings. Mutat. Res **2008,** *650* (2), 115–22.
18. Crespo-López, M.E.; Macêdo, G.L.; Pereira, S.I.D.; Arrifano, G.P.F.; Picanço-Diniz, D.L.W.; do Nascimento, J.L.M.; Herculano, A.M. Mercury and human genotoxicity: Critical considerations and possible molecular mechanisms. Pharmacol. Res. **2009,** *60* (4), 212–220.
19. Della Torre, C.; Petochi, T.; Corsi, I.; Dinardo, M.M.; Baroni, D.; Alcaro, L.; Focardi, S.; Tursi, A.; Marino, G.; Frigeri, A.; Amato, E. DNA damage, severe organ lesions and high muscle levels of As and Hg in two benthic fish species from a chemical warfare agent dumping site in the Mediterranean Sea. Sci. Total Environ. **2010,** *408* (9), 2136–2145.
20. Vinggaard, A.; Jacobsen, H.; Metzdorff, S.; Andersen, H.; Nellemann, C. Antiandrogenic effects in short-term in vivo studies of the fungicide fenarimol. Toxicology **2005,** *207*, 21–34.
21. Paolini, M.; Mesirca, R.; Pozzetti, L.; Sapone, A.; Cantelli-Forti, G. Molecular non-genetic biomarkers related to fenarimol cocarcinogenesis, organ- and sex-specific CYP induction in rat. Cancer Lett. **1986,** *101*, 171–178.
22. Perry, M.J. Effects of environmental and occupational pesticide exposure on human sperm: A systematic review. Hum. Reprod. Update **2008,** *14* (3), 233–242.
23. OECD. *OECD Guideline for the Testing of Chemicals. 414 Prenatal Developmental Toxicity Study;* Organisation for Economic Co-operation and Development: Paris, 2009.
24. Rudén, C.; Hansson, S.O. Registration, Evaluation, and Authorization of Chemicals (REACH) is but the first step— How far will it take us? Six further steps to improve the European chemicals legislation. Environ. Health Perspect. **2010,** *118* (1), 6–10.
25. Dearfield, K.L.; Cimino, M.C.; McCarroll, N.E.; Mauer, I.; Valcovic, L.R. Genotoxicity risk assessment: A proposed classification strategy. Mutat. Res. **2002,** *521*, 121–135.
26. Coors, A.; Vanoverbeke, J.; De Bie, T.; De Meester, L. Land use, genetic diversity and toxicant tolerance in natural populations of Daphnia magna. Aquat. Toxicol. **2009,** *95*, 71–79.

27. Bickham, J.W.; Sandhu, S.; Hebert, P.D.; Chikhi, L.; Athwal, R. Effects of chemical contaminants on genetic diversity in natural populations: Implications for biomonitoring and ecotoxicology. Mutat. Res. **2000,** *463*, 33–51.
28. Andersen, M.E.; Krewski, D. Toxicity testing in the 21st century: Bringing the vision to life. Toxicol. Sci. **2009,** *107*, 324–330.
29. Bony, S.; Gillet, C.; Bouchez, A.; Margoum, C.; Devaux, A. Genotoxic pressure of vineyard pesticides in fish: Field and mesocosm surveys. Aquat. Toxicol. **2008,** *89*, 197–203.
30. Girotti, S.; Ferri, E.N.; Fumo, M.G.; Maiolini, E. Monitoring of environmental pollutants by bioluminescent bacteria. Anal. Chim. Acta **2008,** *608*, 2–29.
31. Boverhof, D.R.; Zacharewski, T.R. Toxicogenomics in risk assessment: Applications and needs. Toxicol. Sci. **2006,** *89*, 352–360.
32. Goetz, A.K.; Dix, D.J. Toxicogenomic effects common to triazole antifungals and conserved between rats and humans. Toxicol. Appl. Pharmacol. **2009,** *238*, 80–89.
33. Akcha, F.; Arzul, G.; Rousseau, S.; Bardouil, M. Comet assay in phytoplankton as biomarker of genotoxic effects of environmental pollution. Mar. Environ. Res. **2008,** *66*, 59–61.
34. Pellacani, C.; Buschini, A.; Furlini, M.; Poli, P.; Rossi, C. A battery of in vivo and in vitro tests useful for genotoxic pollutant detection in surface waters. Aquat. Toxicol. **2006,** *77*, 1–10.
35. Bull, S.; Fletcher, K.; Boobis, A.R.; Battershill, J.M. Evidence for genotoxicity of pesticides in pesticide applicators: A review. Mutagenesis **2006**, *21* (2), 93–103.
36. Bolognesi, C. Genotoxicity of pesticides: A review of human biomonitoring studies. Mutat. Res., Rev. Mutat. Res. **2003,** *543* (3), 251–272.
37. Epstein, S.S.; Legator, M.S. The Mutagenicity of Pesticides: Concepts and Evaluation; The MIT Press: Cambridge, 1971; 220 pp.
38. Hrelia, P.; Vigagni, F.; Maffei, F.; Morotti, M.; Colacci, A.; Perocco, P.; Grilli, S.; Cantelli-Forti, G. Genetic safety evaluation of pesticides in different short-term tests. Mutat. Res. **1994,** *321* (4), 219–228.
39. Poli, P.; de Mello, M.A.; Buschini, A.; de Castro, V.L.S.S.; Restivo, F.M.; Rossi, C.; Zucchi, T.M.A.D. Evaluation of the genotoxicity induced by the fungicide fenarimol in mammalian and plant cells by the use of single-cell gel electrophoresis assay. Mutat. Res. **2003,** *540*, 57–66.
40. Leme, D.M.; Marin-Morales, M.A. Allium cepa test in environmental monitoring: A review on its application. Mutat. Res. **2009,** *682*, 71–81.
41. Rodrigues, G.S.; Ma, T.H.; Pimentel, D.; Weinstein, L.H. Tradescantia bioassays as monitoring systems for environmental mutagenesis—A review. Crit. Rev. Plant Sci. **1997,** *16* (4), 325–359.
42. Li, X.; Li, S.; Liu, S.; Zhu, G. Lethal effect and in vivo genotoxicity of profenofos to Chinese native amphibian (Rana spinosa) tadpoles. Arch. Environ. Contam. Toxicol. 2010, *59*, 478–483. DOI 10.1007/s00244-010-9495-4.
43. Soloneski, S.; Larramendy, M.L. Sister chromatid exchanges and chromosomal aberrations in Chinese hamster ovary (CHO-K1) cells treated with the insecticide pirimicarb. J. Hazard. Mater. **2010,** *174*, 410–415.
44. Kumar, R.; Nagpure, N.S.; Kushwaha, B.; Srivastava, S.K.; Lakra, W.S. Investigation of the genotoxicity of malathion to freshwater teleost fish Channa punctatus (Bloch) using the micronucleus test and comet assay. Arch. Environ. Contam. Toxicol. **2010,** *58*, 123–130.
45. Candioti, J.V.; Natale, G.S.; Soloneski, S.; Ronco, A.E.; Larramendy, M.L. Sublethal and lethal effects on Rhinella arenarum (Anura, Bufonidae) tadpoles exerted by the pirimicarb-containing technical formulation insecticide Aficida. Chemosphere **2010,** *78*, 249–255.
46. Yin, X.; Zhu, G.; Li, X.B.; Liu, S. Genotoxicity evaluation of chlorpyrifos to amphibian Chinese toad (Amphibian: Anura) by comet assay and micronucleus test. Mutat. Res. **2009,** *680*, 2–6.
47. Vega, L.; Valverde, M.; Elizondo, G.; Leyva, J.F.; Rojas, E. Diethylthiophosphate and diethyldithiophosphate induce genotoxicity in hepatic cell lines when activated by further biotransformation via cytochrome P450. Mutat. Res. **2009,** *679*, 39–43.

48. Mladinic, M.; Berend, S.; Vrdoljak, A.L.; Kopjar, N.; Radic, B.; Zeljezic, D. Evaluation of genome damage and its relation to oxidative stress induced by glyphosate in human lymphocytes in vitro. Environ. Mol. Mutagen. **2009,** *50*, 800–807.
49. Moore, P.D.; Yedjou, C.G.; Tchounwou, P.B. Malathion-induced oxidative stress, cytotoxicity, and genotoxicity in human liver carcinoma (HepG2) cells. Environ. Toxicol. **2009,** *25*, 221–226. DOI 10.1002/tox.20492.
50. Sing, N.P.; McCoy, M.T.; Tice, R.R.; Schneider, E.L. A simple technique for quantitation of low levels of DNA damage in individual cells. Exp. Cell Res. **1988,** *175*, 184–191.
51. Tice, R.R.; Agurell, E.; Anderson, D.; Burlinson, B.; Hartmann, A.; Kobayashi, H.; Miyamae, Y.; Rojas, E.; Ryu, J-C.; Sasaki, Y.F. Single cell gel/comet assay: Guideline for in vitro and in vivo genetic toxicology testing. Environ. Mol. Mutagen. **2000,** *35*, 206–221.
52. Tsuda, S.; Matsusaka, N.; Madarame, H.; Miyamae, Y.; Ishida, K.; Satoh, M.; Sekihashi, K.; Sasaki, Y.F. The alkaline single cell electrophoresis assay with eight mouse-organs: Results with 22 monofunctional alkylating agents (including 9 dialkyl N-nitrosoamines) and 10 DNA crosslinkers. Mutat. Res. **2000,** *467*, 83–98.
53. Sekihashi, K.; Yamamoto, A.; Matsumura, Y.; Ueno, S.; Watanabe-Akanuma, M.; Kassie, F.; Knasmüller, S.; Tsuda, S.; Sasaki, Y.F. Comparative investigation of multiple organs of mice and rats in the comet assay. Mutat. Res. **2002,** *517*, 53–74.
54. Castro, V.L.; Mello, M.A.; Poli, P.; Zucchi, T.M. Prenatal and perinatal fenarimol-induced genotoxicity in leukocytes of in vivo treated rats. Mutat. Res. **2005,** *583*, 95–104.
55. Tewari, A.; Dhawan, A.; Gupta, S.K. DNA damage in bone marrow and blood cells of mice exposed to municipal sludge leachates. Environ. Mol. Mutagen. **2006,** *47*, 271–276.
56. Speit, G.; Vasquez, M.; Hartmann, A. The comet assay as an indicator test for germ cell genotoxicity. Mutat. Res. **2009,** *681*, 3–12.
57. Fenech, M. The cytokinesis-block micronucleus technique. In Technologies for Detection of DNA Damage and Mutations; Pfeifer, G.P., Ed.; Plenum Press: New York, 1996; 25–36.
58. Fenech, M. The in vitro micronucleus technique. Mutat. Res. **2000,** *455*, 81–95.
59. Parry, J.M.; Parry, E.M. The use of the in vitro micronucleus assay to detect and assess the aneugenic activity of chemicals. Mutat. Res. **2006,** *607* (1), 5–8.
60. Joksic, G.; Vidakovic, A.; Spasojevic-Tisma, V. Cytogenetic monitoring of pesticide sprayers. Environ. Res. **1997,** *75*, 113–118.
61. Pacheco, A.O.; Hackel, C. Chromosome instability induced by agrochemicals among farm workers in Passo Fundo, Rio Grande do Sul, Brazil. Cad. Saude Publica **2002,** *18* (6), 1675–1683.
62. Fucic, A.; Markucic, D.; Mijic, A.; Jazbec, A.M. Estimation of genome damage after exposure to ionising radiation and ultrasound used in industry. Environ. Mol. Mutagen. **2000,** *36*, 47–51.
63. Vaglenov, A.; Nosko, M.; Georgieva, R.; Carbonell, E.; Creus, A.; Marcos, R. Genotoxicity and radioresistance in electroplating workers exposed to chromium. Mutat. Res. **1999,** *446*, 23–34.
64. Somorovska, M.; Szabova, E.; Vodicka, P.; Tulinska, J.; Barancokova, M.; Fabri, R.; Liskova, A.; Riegerova, Z.; Petrovska, H.; Kubova, J.; Rausova, K.; Dusinska, M.; Collins, A. Biomonitoring of genotoxic risk in workers in a rubber factory, comparison of the Comet assay with cytogenetic methods and immunology. Mutat. Res. **1999,** *445*, 181–192.
65. Fenech, M.; Perepetskaya, G.; Mikhalevic, L. A more comprehensive application of the micronucleus technique for biomonitoring of genetic damage rates in human populations. Experiences from the Chernobyl catastrophe. Environ. Mol. Mutagen. **1997,** *30*, 112–118.
66. Sinues, B.; Sanz, A.; Bernal, M.L.; Tres, A.; Alcala, A.; Lanuza, J.; Ceballos, C.; Saenz, M.A. Sister chromatid exchanges, proliferating rate index, and micronuclei biomonitoring of internal exposure to vinyl chloride monomer in plastic industry workers. Toxicol. Appl. Pharmacol. **1991,** 108, 37–45.
67. Evans, H.J. Cytological methods for detecting chemical mutagens. In *Chemical Mutagens, Principles and Methods for their Detection*; Hollaender, A., Ed.; Plenum Press: New York and London, 1976; Vol. 4, 1–29.

68. Ishidate, M., Jr.; Sofuni, T. The in vitro chromosomal aberration test using Chinese hamster lung (chl) fibroblast cells in culture. In *Progress in Mutation* Research; Ashby, J., et al., Eds.; Elsevier Science Publishers: Amsterdam–New York–Oxford, 1985; Vol. 5, 427–432.
69. Albertini, R.J.; Anderson, D.; Douglas, G.R.; Hagmar, L.; Hemminki, K.; Merlo, F.; Natarajan, A.T.; Norppa, H.; Shuker, D.E.; Tice, R.; Waters, M.D.; Aitio, A. IPCS guidelines for the monitoring of genotoxic effects of carcinogens in humans, International Programme on Chemical Safety. Mutat. Res. **2000,** *463*, 111–172.
70. Hagmar, L.; Stromberg, U.; Bonassi, S.; Hansteen, I.L.; Knudsen, L.E.; Lindholm, C.; Norppa, H. Impact of types of lymphocyte chromosomal aberrations on human cancer risk: Results from Nordic and Italian cohorts. Cancer Res. **2004,** *64*, 2258–2263.
71. Galloway, S.M.; Aardema, M.J.; Ishidate, M., Jr.; Ivett, J.L.; Kirkland, D.J.; Morita, T.; Mosesso, P.; Sofuni, T. Report from working group on in in vitro tests for chromosomal aberrations. Mutat. Res. **1994,** 312, 241–261.
72. Mateuca, R.; Lombaert, N.; Aka, P.V.; Decordier, I.; Kirsch-Volders, M. Chromosomal changes: Induction, detection methods and applicability in human biomonitoring. Biochimie **2006,** *88*, 1515–1531.
73. González-Barrera, S.; Cortés-Ledesma, F.; Wellinger, R.E.; Aguilera, A. Equal sister chromatid exchange is a major mechanism of double-strand break repair in yeast. Mol. Cell Biol. **2003,** *11* (6), 1661–1671.
74. Domínguez, I.; Pastor, N.; Mateos, S.; Cortés, F. Testing the SCE mechanism with non-poisoning topoisomerase II inhibitors. Mutat. Res. **2001,** *497*, 71–79.
75. Speit, G.; Hochsattel, R.; Vogel, W. The contribution of DNA single-strand breaks to the formation of chromosome aberrations and SCEs. Basic Life Sci. **1984,** *29*, 229–244.
76. Das, B.C. Factors that influence formation of sister chromatid exchanges in human blood lymphocytes. Crit. Rev. Toxicol. **1988,** *19* (1), 43–86.
77. Shiraishi, Y. Nature and role of high sister chromatid exchanges in Bloom syndrome cells. Some cytogenetic and immunological aspects. Cancer Genet. Cytogenet. **1990,** *50* (2), 175–187.
78. Tucker, J.D.; Auletta, A.; Cimino, M.C.; Dearfield, K.L.; Jacobson-Kram, D.; Tice, R.R.; Carrano, A.V. Sister-chromatid exchange: Second report of the Gene-Tox Program. Mutat. Res. **1993,** *297* (2), 101–180.
79. Tilman, G.; Loriot, A.; Van Beneden, A.; Arnoult, N.; Londoño-Vallejo, J.A.; De Smet, C.; Decottignies, A. Subtelomeric DNA hypomethylation is not required for telomeric sister chromatid exchanges in ALT cells. Oncogene **2009,** *28* (14), 1682–1693.
80. White, J.S.; Choi, S.; Bakkenist, C.J. Transient ATM kinase inhibition disrupts DNA damage-induced sister chromatid exchange. Sci. Signal. **2010,** *3*, ra44. DOI: 10.1126/scisignal.2000758.
81. Popescu, N.C. Sister chromatid exchange formation in mammalian cells is modulated by deoxyribonucleotide pool imbalance. Somatic Cell Mol. Genet. **1999,** *25*, 101–108.
82. Ashman, C.R.; Davidson, R.L. Bromodeoxyuridine mutagenesis in mammalian cells is related to deoxyribonucleotide pool imbalance. Mol. Cell. Biol. **1981,** *1*, 254–260.
83. Latt, S.A. Sister chromatid exchanges, indices of human chromosome damage and repair: Detection by fluorescence and induction by mitomycin-C. Proc. Natl. Acad. Sci. U. S. A. **1974,** *71*, 3162–3166.
84. Latt, S.A.; Schreck, R.R. Sister chromatid exchange analysis. Am. J. Hum. Genet. **1980,** *32*, 297–313.
85. EPA. In vitro sister chromatid exchange assay. Test methods and guidelines/OPPTS Harmonized test guidelines OPPTS 870.5900. U.S. Environmental Protection Agency, 1998, available at http://www.epa.gov/epahome/research.htm (accessed May 2010).
86. Morgan, W.F.; Schwartz, J.L.; Murnane, J.P.; Wolff, S. Effect of 3-aminobenzamide on sister chromatid exchange frequency in X-irradiated cells. Radiat. Res. **1983,** *93*, 567–571.

87. Hagmar, L.; Brogger, A.; Hansteen, I.L.; Heim, S.; Hogstedt, B.; Knudsen, L.; Lambert, B.; Linnainmaa, K.; Mitelman, F.; Nordenson, I.; Reuterwall, C.; Salomaa, S.I.; Skerfving, S.; Sorsa, M. Cancer risk in human predicted by increased levels of chromosomal aberrations in lymphocytes: Nordic Study Group on the Health Risk of Chromosome Damage. Cancer Res. **1994,** *54*, 2919–2922.
88. Hagmar, L.; Bonassi, S.; Stromberg, U.; Brogger, A.; Knudsen, L.; Norppa, H.; Reuterwall, C. The European Study Group on Cytogenetic Biomarkers and Health. Chromosomal aberrations in lymphocytes predict human cancer. A report from the European Study Group on Cytogenetic Biomarkers and Health (ESCH). Cancer Res. **1998,** *58*, 4117–4121.
89. Gichner, T.; Menke, M.; Stavreva, D.A.; Schubert, I. Maleic hydrazide induces genotoxic effects but no DNA damage detectable by the comet assay in tobacco and field beans. Mutagenesis **2000,** *15*, 385–389.
90. Citterio, S.; Aina, R.; Labra, M.; Ghiani, A.; Fumagalli, P.; Sgorbati, S.; Santagostino A. Soil genotoxicity assessment: A new strategy based on biomolecular tools and plant bioindicators. Environ. Sci. Technol. **2002,** *36* (12), 2748–2753.
91. Restivo, F.M.; Laccone, M.C.; Buschini, A.; Rossi, C.; Poli, P. Indoor and outdoor genotoxic load detected by the Comet assay in leaves of *Nicotiana tabacum* cultivars Bel B and Bel W3. Mutagenesis **2002,** *17* (2), 127–134.
92. Klumpp, A.; Ansel, W.; Klumpp, G.; Calatayud, V.; Garrec, J.P.; He, S.; Penuelas, J.; Ribas, A.; Ro-Poulsen, H.; Rasmussen, S.; Sanz, M.J.; Vergne, P. Tradescantia micronucleus test indicates genotoxic potential of traffic emissions in European cities. Environ. Pollut. **2006,** *139* (3), 515–522.
93. Liu, W.; Zhu, L.S.; Wang, J.; Wang, J.H.; Xie, H.; Song, Y. Assessment of the genotoxicity of endosulfan in earthworm and white clover plants using the comet assay. Arch. Environ. Contam. Toxicol. **2009,** *56* (4), 742–746.
94. Villarini, M.; Fatigoni, C.; Dominici, L.; Maestri, S.; Ederli, L.; Pasqualini, S.; Monarca, S.; Moretti, M. Assessing the genotoxicity of urban air pollutants using two in situ plant bioassays. Environ. Pollut. **2009,** *157* (12), 3354–3356.
95. Rodrigues, F.P.; Angeli, J.P.; Mantovani, M.S.; Guedes, C.L.; Jordão, B.Q. Genotoxic evaluation of an industrial effluent from an oil refinery using plant and animal bioassays. Genet. Mol. Biol. **2010,** *33* (1), 169–175.
96. Sadowska, A.; Pluygers, E.; Niklinska, W.; Maria, M.R.; Obidoska, G. Use of higher plants in the biomonitoring of environmental genotoxic pollution. Folia Histochem. Cytobiol. **2001,** *39*, 52–53.
97. Dimitrov, B.D.; Gadeva, P.G.; Benova, D.K.; Bineva, M.V. Comparative genotoxicity of the herbicides Roundup, Stomp and Reglone in plant and mammalian test systems. Mutagenesis **2006,** *21* (6), 375–382.
98. Grant, W.F. The present status of higher plant bioassays for the detection of environmental mutagens. Mutat. Res. **1994,** *310* (2), 175–185.
99. Ma, T.H.; Cabrera, G.L.; Owens, E. Genotoxic agents detected by plant bioassays. Rev. Environ. Health **2005,** *20* (1), 1–13.
100. Feretti, D.; Zerbini, I.; Zani, C.; Ceretti, E.; Moretti, M.; Monarca, S. Allium cepa chromosome aberration and micronucleus tests applied to study genotoxicity of extracts from pesticide-treated vegetables and grapes. Food Addit. Contam. **2007,** *24* (6), 561–572.
101. Mancini, A.; Buschini, A.; Restivo, F.M.; Rossi, C.; Poli, P. Oxidative stress as DNA damage in different transgenic tobacco plants. Plant Sci. **2006,** *170*, 845–852.
102. Norppa, H. Cytogenetic biomarkers and genetic polymorphisms. Toxicol. Lett. **2004,** *149* (1–3), 309–334.
103. Vodicka, P.; Koskinen, M.; Naccarati, A.; Oesch-Bartlomowicz, B.; Vodickova, L.; Hemminki, K.; Oesch, F. Styrene metabolism, genotoxicity, and potential carcinogenicity. Drug Metab. Rev. **2006,** *38* (4), 805–853.
104. Lampe, J.W. Diet, genetic polymorphisms, detoxification, and health risks. Altern. Ther. Health Med. **2007,** *13* (2), S108–S111.

105. Lin, J.; Swan, G.E.; Shields, P.G.; Benowitz, N.L.; Gu, J.; Amos, C.I.; de Andrade, M.; Spitz, M.R.; Wu, X. Mutagen sensitivity and genetic variants in nucleotide excision repair pathway: Genotype–phenotype correlation. Cancer Epidemiol. Biomarkers Prev. **2007,** *16* (10), 2065–2071.
106. Rueff, J.; Teixeira, J.P.; Santos, L.S.; Gaspar, J.F. Genetic effects and biotoxicity monitoring of occupational styrene exposure. Clin. Chim. Acta **2009,** *399* (1–2), 8–23.
107. Rohr, P.; da Silva, J.; Erdtmann, B.; Saffi, J.; Guecheva, T.N.; Henriques, J.A.; Kvitko, K. BER gene polymorphisms (OGG1 Ser326Cys and XRCC1 Arg194Trp) and modulation of DNA damage due to pesticides exposure. Environ. Mol. Mutagen. **2010** [Epub ahead of print].
108. Goode, E.L.; Ulrich, C.M.; Potter J.D. Polymorphisms in DNA repair genes and associations with cancer risk. Cancer Epidemiol. Biomarkers Prev. **2002,** *11*, 1513–1530.
109. Buschini, A.; De Palma, G.; Poli, P.; Martino, A.; Rossi, C.; Mozzoni, P.; Scotti, E.; Buzio, L.; Bergamaschi, E.; Mutti, A. Genetic polymorphism of drug-metabolizing enzymes and styrene-induced DNA damage. Environ. Mol. Mutagen. **2003,** *41* (4), 243–252.
110. Vodicka, P.; Kumar, R.; Stetina, R.; Sanyal, S.; Soucek, P.; Haufroid, V.; Dusinska, M.; Kuricova, M.; Zamecnikova, M.; Musak, L.; Buchancova, J.; Norppa, H.; Hirvonen, A.; Vodickova, L.; Naccarati, A.; Matousu, Z.; Hemminki, K. Genetic polymorphisms in DNA repair genes and possible links with DNA repair rates, chromosomal aberrations and single-strand breaks in DNA. Carcinogenesis **2004,** *25*, 757–763.
111. Landsiedel, R.; Kapp, M.D.; Schulz, M.; Wiench, K.; Oesch, F. Genotoxicity investigations on nanomaterials: Methods, preparation and characterization of test material, potential artefacts and limitations—Many questions, some answers. Mutat. Res. **2009,** *681*, 241–258.
112. Savolainen, K.; Alenius, H.; Norppa, H.; Pylkkänen, L.; Tuomi, T.; Kasper, G. Risk assessment of engineered nanomaterials and nanotechnologies—A review. Toxicology **2010,** *269*, 92–104.
113. Wang, J.J.; Sanderson, B.J.S.; Wang, H. Cyto- and genotoxicity of ultrafine TiO_2 particles in cultured human lymphoblastoid cells. Mutat. Res. **2007,** *628*, 99–106.
114. Reeves, J.F.; Davies, S.J.; Dodd, N.J.F.; Jha, A.N. Hydroxyl radicals (•OH) are associated with titanium dioxide (TiO_2) nanoparticle-induced cytotoxicity and oxidative DNA damage in fish cells. Mutat. Res. **2008,** *640*, 113–122.
115. Yang, H.; Liu, C.; Yang, D.; Zhanga, H.; Xi, Z. Comparative study of cytotoxicity, oxidative stress and genotoxicity induced by four typical nanomaterials: The role of particle size, shape and composition. J. Appl. Toxicol. **2009,** *29*, 69–78.
116. Sharma, V.; Shukla, R.K.; Saxena, N.; Parmar, D.; Das, M.; Dhawan, A. DNA damaging potential of zinc oxide nanoparticles in human epidermal cells. Toxicol. Lett. **2009,** *185*, 211–218.
117. Elespuru, R.K.; Agarwal, R.; Atrakchi, A.H.; Bigger, C.A.H.; Heflich, R.H.; Jagannath, D.R.; Levy, D.D.; Moore, M.M.; Ouyang, Y.; Robison, T.W.; Sotomayor, R.E.; Cimino, M.C.; Dearfield, K.L. Current and future application of genetic toxicity assays: The role and value of in vitro mammalian assays. Toxicol. Sci. **2009,** *109* (2), 172–179.
118. Garcia-Sagredo, J.M. Fifty years of cytogenetics: A parallel view of the evolution of cytogenetics and genotoxicology. Biochim. Biophys. Acta **2008,** *1779*, 363–375.
119. Robbens, J.; van der Ven, K.; Maras, M.; Blus, R.; De Coen, W. Ecotoxicological risk assessment using DNA chips and cellular reporters. Trends Biotechnol. **2007,** *25* (10), 460–466.
120. Gant, T.W.; Zhang, S.D. In pursuit of effective toxicogenomics. Mutat. Res. **2005,** *575*, 4–16.
121. Ellinger-Ziegelbauerb, H.; Aubrechta, J.; Kleinjans, J.C.; Ahr, H.-J. Application of toxicogenomics to study mechanisms of genotoxicity and carcinogenicity. Toxicol. Lett. **2009,** *186*, 36–44.
122. Waters, M.D.; Olden, K.; Tennant, R.W. Toxicogenomic approach for assessing toxicant-related disease. Mutat. Res. **2003,** *544*, 415–424.
123. Sen, B.; Mahadevan, B.; DeMarini, D.M. Transcriptional responses to complex mixtures: A review. Mutat. Res. **2007,** *636*, 144–177.
124. Lettieri, T. Recent applications of DNA microarray technology to toxicology and ecotoxicology. Environ. Health Perspect. **2006,** *114*, 4–9.

10

Pollution: Pesticides in Agro-Horticultural Ecosystems

J.K. Dubey and
Meena Thakur

Introduction

Agro-horticultural ecosystems are man-made ecosystems that are greatly influenced by human activities. In the last few decades, agricultural intensification and diversification in most of the developing and developed countries have resulted in high-input farming, i.e., excessive use of pesticides (herbicides, insecticides, fungicides) and fertilizers, which has helped to meet the rising food demand. Agricultural and horticultural ecosystems are now dominated by monoculture and hybrid varieties to enhance crop yields. Hybrid varieties are high yielding and more demanding. The excessive use of inorganic fertilizers results in rapid multiplication and subsequent outbreaks of many pests, simultaneously resulting in enhanced pesticide use. Rice and cotton are important crops grown worldwide, and because of monoculture, they are attacked by hundreds of pests and receive a disproportionately high share of pesticides (17% and 24%) worldwide.[1] It has been observed that chemicals that have been banned for most of the food crops are still being used on cotton as it is not consumed directly. For example, nearly one-quarter of all pesticides used in the United States are applied to cotton, and the overall amount and intensity per acre are increasing every year. Worldwide, more and extremely toxic pesticides are sprayed on cotton than on any other crop,[2] which find an entry in the human body in the form of salad dressings, baked goods, and snacks like Fritos and Goldfish.

Based on the trend of herbicide use in the United States, agriculture accounts for three quarters of total chemical pesticides used. According to the data published by the US Environmental Protection Agency (EPA), in 2001, 675 million lb of chemical pesticides were used in agriculture.[3] Thus, horticultural and agricultural ecosystems are highly dependent on pesticides. It is estimated that if pesticides are not used, food supplies would fall to 30%–40% due to the ravages of pests.[4] Worldwide, research data show that without effective pest management, preharvest losses in crops would average about 40%

and the world's food and fiber production as well as environmental and human health would be seriously threatened.[5]

According to market research report, pesticide consumption has been found to be highest in China followed by the United States and Brazil and India ranked 11th after Spain, and as per Tata Strategic Management Group[6] (2016), the world agrochemical market is expected to be worth $60 billion in 2016 of which Indian market will account for $6.8 billion of the total revenues and is expected to hike at $8 billion by 2025 in comparison to world market share of $8.0 billion.[6]

No doubt, pesticides play an important role in enhancing agricultural productivity, but intensive use of pesticides in agricultural and horticultural ecosystems has resulted in degradation of environment, whether it is in the form of development of resistant pests, secondary pest outbreak, or pest resurgence, which leads to changes in ecosystem biodiversity or adverse effects on pollinators, natural enemies, and many incurable human diseases. Several pesticides are known to persist for a longer period in the environment or on the substrate to which they are applied and have long-term side effects on human health and the natural environment. Persistent organic pesticides, especially DDT (dichloro diphenyl trichloroethane) and HCH (hexachloro cyclohexane), have been detected in various systems, even in human blood, fat, and milk samples. Even the soft drinks that are water-based flavored drinks are known to contain pesticides. The hazards of pesticide pollution have been much realized presently, and the demands for their reduction are desired globally in various pesticide monitoring programs.

Pesticide Consumption Trend

By using nonchemical techniques such as sanitation, cultivation, crop rotation, resistant cultivars, and biological control (including introduction of transgenic) for pest control, many pests cannot be controlled adequately. Therefore, there is a continuous need for application of substantial quantities of chemical pesticides, as a result of which crop protection in many developing countries is still dominated by an increasing use of pesticides.

The world pesticide market now amounts $60 billion and is expected to increase to $80 billion in 2025, whereas, Indian market of pesticides is recently reported to be $6.8 billion out of which $2.9 billion is by domestic use and $3.9 billion is earned by export of pesticides to various countries. Both the domestic and the export markets are expected to enhance to $4.0 billion each in 2025, thereby increasing the expected Indian pesticide market to $8.0 billion.[6]

Globally, pesticide consumption has been found to be highest in China 18,07,000 and 17,72,449 metric tonnes in 2014 and 2015, respectively. In the United States, 4,07,779 metric tonnes of pesticide consumption in 2012 was observed. After China and the United States, next in the list are Brazil (3,52,336 and 3,95,646 metric tonnes in 2014 and 2015), Argentina (2,07,706 metric tonnes in 2014), Mexico (98,814 metric tonnes in 2014), Ukraine (78,201 metric tonnes in 2014), Canada (76,314 metric tonnes in 2014), and France (75,339 and 63,818 metric tonnes in 2014 and 2015).[7]

According to the data given by FAOSTAT 2018, the pesticide consumption reported in Europe is 4,36,361, 4,28,506, 4,49,392, and 3,56,279 metric tonnes in 2012, 2013, 2014, and 2015, respectively.[7] Out of all European countries, the top five consumers have been France, Italy, Spain, Germany, and Poland with the highest consumption in France, i.e., 63,844, 66,676, 75,339, and 63,818 metric tonnes in 2012, 2013, 2014, and 2015, respectively.

The per hectare consumption of pesticides in India is among the lowest in the world and currently stands at 0.6 kg/ha against 5 kg/ha in the UK and at almost 20 times approximately 13 kg/ha in China. In order to increase yield and ensure food security for its enormous population, agrochemical penetration in India is bound to go up.[6]

Further, government policies on subsidies, establishment of market for agrochemicals, availability of technology packet, etc., encourage the farmers to use pesticides excessively. Today, more than 80% of worldwide pesticide sales fall to the share of only six companies. In 2004, three agrochemical companies, each with sales of more than $4 billion, together controlled the global market for pesticides,[8] by controlling such a large stake of the market, these companies have a considerable influence on the way in which plant protection is practiced in farming ecosystems.

Pesticides Use Pattern in India

In many cases, farmers go for prophylactic applications, whether they are required or not. India was one of the first countries in the third world to start large-scale use of pesticides for the control of insect pests of public health as well as of agricultural importance, initially DDT and HCH were imported soon after independence. A moderate beginning in indigenous manufacture of pesticides was made with the setting of the plant to produce HCH in 1952 at Rishra (West Bengal). This was followed by the indigenous production of DDT.[9]

The export of various pesticides from India has been higher than their import. In 2010, pesticides of $1.14 billion were exported which gradually increased to $1.38, $1.66, and $2.13 billion in 2011, 2012, and 2013, respectively, whereas the export of pesticides from India was decreased after 2013 and was nearly of $1.94 and $1.93 billion in 2014 and 2015, respectively. The amount of pesticides imported into India increased from $0.63 (in 2010) to $0.99 billion (in 2014) and somewhat decreased to $0.89 billion in 2015.[7]

Considering the pesticide consumption and area under cultivation in 2016–2017 in different states, the consumption of pesticides (kg per hectare) has been found to be highest in Jammu and Kashmir (1.89) followed by Tripura (0.78), Punjab (0.74), Haryana (0.62), Telangana (0.61), and Maharashtra (0.58). The pesticide consumption in Kerala, Uttar Pradesh, Himachal Pradesh, Tamil Nadu, Jharkhand, and West Bengal was found to be 0.41, 0.39, 0.36, 0.34, 0.32, and 0.27 kg/ha, respectively.[10]

In India, Kerala is one of the leading agricultural states of India, and currently, India is the leading manufacturer of basic pesticides in Asia and ranks 12th globally. Compiling the data on the consumption level of pesticides in agriculture in Kerala (1995–1996 to 2007–2008), the total quantity is estimated at 462.05 metric tonnes (2007–2008). Pesticide application in the state is prophylactic and is one of the most important risk management strategies; e.g., pesticide application on bitter gourd starts from the time of transplanting. The prophylactic application of the pesticides is resorted to at an interval of 2 weeks initially, which gets reduced to 2 days as the crop nears flowering and fruit set. There is a tendency among farmers to change the chemicals in each spray. Thus, on an average, acetamiprid is sprayed 6 times, phorate and dimethoate 5 times each, quinalphos and indoxacarb 4 times each, and the rest 3–4 times each. During a crop cycle of 90 days in bitter gourd, farmers apply pesticides as many as 50 times.[11]

The pesticides that are used in the state include chemicals that are banned for sale in Kerala (endosulfan), banned for use in fruits/vegetables (monocrotophos), and those permitted for restricted use only (methyl parathion, lindane, and methoxy ethyl mercury chloride). The farmers are investing a large portion of their income on pesticides, and pesticide consumption is reported to be the primary method of suicide in Kerala. Of the 900–1000 suicides per year, 60% are by consuming poisons. The commonly used poisons are furadan, malathion, and rat poison.[12] Moreover, farmers also go for suicide because of indebtedness due to purchase of chemicals and sometimes due to complete failure of crops in spite of heavy investment.[12]

Impact of Pesticide Use on the Pollinators

Wild bees, bumblebees, honeybees, and solitary bees are well known and valued as important pollinators of crops/plants and are in commercial use for pollination. About 33% of all crops require pollination. Intensive cultivation and excessive use of pesticides lead to a sharp decline in the population of these pollinators, which is one of the major causes of low productivity of agricultural and horticultural crops. Pesticides, such as DDT, benzene hexachloride (BHC), cyclodiene, and most of the organophosphorus and carbamate compounds are highly toxic to bees. Although endosulfan is listed as a persistent organic pollutant, it is the only available pesticide known to be safe for honeybees and other beneficial insects and is still extensively used in many countries. Health and environmental causalities related to excessive use of this pesticide are reported from many parts of the world and European countries withdrew its registration in 2005. Still, there are many instances where most of the pesticides fail to be effective and endosulfan is recommended. For example, in 2008, there was heavy weevil infestation on hazelnut crops in Italy, and the Italian government had to prescribe the use of endosulfan for 120 days although it endangered the health of its citizens. Pesticides like neonicotinoids used as a substitute for endosulfan in agricultural ecosystems in countries such as Germany, France, UK, and the United States have resulted in mass bee killing and colony collapse disorders. In 1959, carbaryl was used against certain orchard pests and later registered for many other crops. In 1967, it caused the destruction of an estimated 70,000 colonies of honeybees in California from use in cotton and an estimated 33,000 colonies in Washington from use in corn. The estimated national loss from all pesticide poisoning from the same year was 500,000 colonies. Carbaryl is still one of the most destructive bee-killing chemical.[13]

A horticultural ecosystem, particularly the apple ecosystem, is heavily polluted with pesticides. Apple is one of the important commercial horticultural crops grown in temperate regions of the world. Honeybees are important pollinators of this crop, and quality beehives are placed in apple orchards for enhancing apple productivity. Since the crop is important from an economic point of view, much attention is paid to it. In India, the temperate northern regions of the country (Himachal Pradesh, Jammu, and Kashmir) are known for quality apple production. To obtain quality fruits in Himachal Pradesh, a number of pesticides, which affect the pollinators, are applied on the crop right from fruit set to harvest. The situation has reached an alarming level in the state, and if required measures are not initiated to conserve and rear the population of these pollinators, it could impinge on the total agricultural and fruit production in the years to come. Currently, about 943,000 ha of land is under fruit cultivation in the state, and it requires at least 5 lakh colonies of honeybees alone for pollination to enhance the production. There is a great need to encourage organic farming to enhance the population of natural pollinators and avoid pesticide applications when crops, cover crops, weeds, and wildflowers are in bloom in the treatment area or nearby.

Impact of Pesticide Use on Natural Enemies

Pesticides including insecticides and miticides are primarily used to regulate insect and mite populations in agricultural and horticultural crop production systems. However, continuous dependence on pesticides may eventually result in a number of potential ecological problems including resistance, secondary pest outbreaks, and/or target pest resurgence.[14,15] Therefore, implementation of alternative management strategies is justified in order to conserve existing pesticides and produce crops with minimal damage from insect pests. One option that has gained interest by producers is integrating pesticides with biological control agents or natural enemies including parasitoids and predators.[16] This is often referred to as "compatibility," which is the ability to integrate or combine natural enemies with pesticides so as to regulate arthropod pest populations without directly or indirectly affecting the life history parameters or population. This may also refer to pesticides being effective against targeted insect pests but relatively non-harmful to natural enemies.[17,18] Pesticides vary in their activity, which not only impacts how they kill arthropod pests but also how they may indirectly influence natural enemy populations. Pesticides may be classified as contact, stomach poison, systemic, and/or translaminar.[19,20] In addition, the application method—foliar versus drench or granular—may determine the extent of any indirect effects on natural enemies[21] as well as the pesticide mode of action. The type of natural enemy-parasitoid or predator may be influenced differently based on the factors mentioned above. Furthermore, the type of pesticide may substantially contribute to any indirect effects on natural enemies. For example, broad-spectrum, nerve toxin pesticides such as most of the older pesticides in the chemical classes, organophosphate (acephate and chlorpyrifos), carbamate (carbaryl and methiocarb), and pyrethroid (bifenthrin and cyfluthrin) may be both directly and indirectly more harmful to natural enemies than non-nerve toxin-type pesticides (often referred to a "selective pesticides") including insect growth regulators (kinoprene and pyriproxyfen), insecticidal soaps (potassium salts of fatty acids), horticultural oils (petroleum or neem-based), selective feeding blockers (flonicamid and pymetrozine), and microbials (entomopathogenic fungi and bacteria and other microorganisms).[22] The non-nerve toxin pesticides are generally more specific or selective in regards to arthropod pest activity with broader modes of action than nerve toxin pesticides.[16] The effects of pesticides on natural enemies are typically associated with determining direct effects such as mortality or survival over a given time period (24–96 hours).[23] While evaluations associated with the direct effects of pesticides on natural enemies are important, what are actually more relevant are the indirect or delayed effects of pesticides because this provides information on the long-term stability and overall success of a biological control program when attempting to integrate the use of pesticides with natural enemies.[24–27] Any indirect effects, which are sometimes referred to as sublethal, latent, or cumulative adverse effects, may be associated with interfering with the physiology and behavior of natural enemies by inhibiting longevity, fecundity, reproduction (based on the number of progeny produced or eggs laid by females), development time, mobility, searching (foraging) and feeding behavior, predation and/or parasitism, prey consumption, emergence rates, and/or sex ratio.[15,25,28,29–30]

In apple ecosystems, many host species of phytophagous arthropods among which red spider mite and two-spotted spider mite are substantial worldwide. Predatory mites play an important role in checking the population of these mites. In India, the red spider mite *Panonychus ulmi* was a minor pest up to 1990, but the commercialization of apple led to excessive and repeated use of pesticides for quality apple production, as a result of which the natural mite predators were destroyed and it emerged as a serious pest of apples. Most of the spray schedules are now focused to this pest, further deteriorating the condition. Pyrethroids and carbamates, i.e., carbaryl (Sevin), are highly toxic to predatory mites, e.g., *Typhlodromus occidentalis* and use of Sevin for thinning causes mite flare-ups. Another well-known example is the resurgence of brown plant hopper (BPH) in rice ecosystem. If no pesticides are used, BPH is kept under control by its natural enemies (mirid bugs, ladybird beetles, spiders, and various pathogens). Since rice is a heavily sprayed crop, pesticides kill the natural enemies and create a situation where BPH can multiply rapidly. Thus, similar to *P. ulmi*, it has also become a serious man-made pest. Synthetic pyrethroids result in spider mite resurgence.[31] In a study conducted by Beers,[32] pyrethroids, carbamates, organophosphates, Assail, Calypso, and Actara are toxic to *T. occidentalis* and *Zetzellia mali*, which are mostly found associated with *P. ulmi* and other phytophagous mites. Organophosphates and carbamates are reported to cause high levels of mortality to coccinellids and lacewings.

Pesticide Use and Bt Transgenics

A common pest management technology used in agroecosystems is the use of *Bacillus thuringiensis* (Bt) transgenic. Bt crops, particularly cotton, are grown all over the world. Bt crops were mainly introduced with an aim to reduce pesticide use, but growing secondary pest populations and efforts to control them have further increased the use of pesticides The major cotton-growing countries, i.e., the United States, China, India, and Argentina have quickly adopted this technique for cotton seeds. For example, before the commercialization of Bt cotton, the Chinese farmers applied an average of 20 pesticide treatments in a season to control bollworm infestations. With the adoption of Bt, the average number of treatments has fallen to only 6.6 in the early stages of Bt adoption.[33] As a result, the pesticide use decreased by 43.3 kg/ha in 1999, i.e., a 71% decrease in pesticide use. For the years 2000 and 2001, Bt cotton was associated with an average reduction of 35.7 kg/ha of pesticide or a percentage deduction of 55%.[34] Similar results have been found in other major cotton-growing countries, and Indian farmers save 39% of expenditures by planting Bt crops.[35] Argentine farmers save 47% of expenditures,[36] Mexican farmers can save 77%,[37] and South African farmers can save 58% by planting Bt.[38] Evidence shows that, though Bt seed costs 2 to 3 times more than a conventional seed, savings on pesticide expenditures guarantee a much higher net return for Bt adopters. Using a household survey from 2004, 7 years after the initial commercialization of Bt cotton in China, we show that total pesticide expenditure for Bt cotton farmers in China is nearly equal to that of their conventional counterparts, about $101/ha. Bt farmers in 2004, on the average, have to spray pesticide 18.22 times, which are more than 3 times higher compared with 6 times the pesticide spray in 1999. Detailed information on pesticide expenditures reveals that, though Bt farmers saved 46% of bollworm pesticides relative to non-Bt farmers, they spend 40% more on pesticides targeted to kill an emerging secondary pest. These secondary pests, e.g., mirid bugs, were rarely found in the field prior to the adoption of Bt cotton, presumably kept in check by bollworm populations and regular pesticide spraying.

Cotton is attacked by more than 165 pests, and farmers repeatedly spray pesticides, which increase the chances of resurgence of secondary pests. In Andhra Pradesh, the number of attacks of aphids, thrips, and jassids has increased since the introduction of Bt cotton in 2002. Many diseases and pests such as tobacco leaf streak virus and tobacco caterpillars have newly emerged in Bt cotton ecosystems in this state.[11]

Pesticide Use in Weeds

Herbicide should be applied at the time when their impact on weeds is highest. If preemergence weed control is optimized, the need for pest emergence measures may be reduced. The cultivation of genetically modified herbicide-tolerant crops has the potential to reduce herbicide inputs. The world sales of agrochemicals is dominated by herbicides (46%) followed by insecticides (26%), fungicides (23%), and others (5%).[39] On the other hand, the Indian market trend indicates domination of insecticides (61.39%) followed by fungicides (19.06%), herbicides (16.75%), and others (2.80%).[40]

The large-scale adoption of dwarf high-yielding varieties (HYV) and hybrids and the increased use of irrigation, fertilizers, and monocropping have increased weed problem in agro-horticultural ecosystems, simultaneously leading to increased herbicide use. Herbicides, such as isoproturon, atrazine, alachlor, butachlor, and oxyfluorfen, are applied on agro-horticultural ecosystems for control of weeds. Globally, herbicides constitute 52% of the total pesticide sales, and in some countries such as the United States, Germany, and Australia, the figure is as high as 60%–70%.[41] According to USDA–NASS (US Department of Agriculture–National Agricultural Statistics Service) report, the use of genetically modified crops is the main reason for the rise in herbicide use.[42] For example, widespread introduction of genetically modified soybeans, cotton, and corn by Monsanto resulted in a 15-fold increase in the use of glyphosate (Roundup) from 1994 to 2005 on these three crops in the United States. The excessive use of glyphosate has resulted in resistant weeds, as a result of which the application of glyphosate, atrazine, 2,4-D, and other leading weed-killing chemicals has further increased since 2002. 2,4-D, the second most heavily used herbicide on soybeans (after glyphosate) in the United States, is associated with a number of adverse health impacts on agricultural workers. These herbicides have increased the risk of cancer, have increased the rate of birth defects in children of men who apply the herbicide, and are also a suspected endocrine disruptor. Similarly, atrazine, the most heavily used herbicide on corn, has been linked to endocrine disruption, neuropathy, and cancer (particularly breast and prostate cancer). It is regularly detected in drinking water supplies in the United States and has been associated with low sperm counts in men. Exposure to extremely low levels of atrazine can cause sex change and/or deformities in frogs, fish, and other organisms. Based on this evidence, and the widespread presence of atrazine in drinking water supplies, the European Union announced a ban on atrazine in 2006. However, the US EPA reregistered atrazine in 2003 despite objections from scientists and environmental groups.[43] Cheaper formulations of herbicides containing 2,4-D and 2-methyl-4-chlorophenoxyacetic acid (MCPA) are still used in many countries, and weeds have developed resistance, e.g., in Bulgaria, 47% of wheat and barley crops were affected by 2,4-D-resistant weeds in 2000.[44]

Pesticide Residues in Agro-Horticultural Ecosystems

Agro-horticultural produce constitutes an essential part of human diet, and as per the recommendation of World Health Organization (WHO), there should be at least 30% fruits and vegetables in our daily diet, depending on the body weight of the person. Vegetables are the major source of vital nutrients.

But it is not heartening to know that instead of fulfilling the nutritional requirements, these fruits and vegetables carry pesticide residue harmful to the health of the consumers. Main reasons of finding the pesticide residues in these food stuffs may be because of providing shiny and fresh appearance and good color to the crop. This leads to usage of synthetic chemicals beyond the safe limits. Due to the persistent nature of some of the highly toxic pesticides, or maybe due to illegal use of prohibited/banned pesticides, these compounds have been detected in the environment worldwide.[45] So, usage of restricted pesticides is a matter of concern especially in case of vegetables. There is a general belief that these food items are much harmful if impregnated with pesticides in comparison to other food stuffs, because they are generally consumed raw or semi-cooked. Many studies support the presence of pesticide residues in vegetables, on an average, this percentage is 50%–70% in India as mentioned by [46–48]. In India, 51% of the food commodities have been detected with pesticide residues.[49] A study conducted by Charan et al. in 2010[47] revealed that 67% of total contaminated samples exceeded the maximum residue limit (MRL) values recommended by the Food and Agriculture Organization (FAO)/WHO. Another study revealed the presence of monocrotophos, chlorpyrifos, cypermethrin, and endosulfan, etc. in the vegetables.[48] According to a report, over 98% of sprayed insecticides and 95% of herbicides reached nontarget destinations such as other species, air, water, and soil.

Most of the pesticides used on crops are persistent, especially organochlorines, which persist for a longer period in the environment (substrate). The organochlorine insecticides (such as DDT and BHC) that were banned still persist in soil and contaminate both organic and conventional crop produce. Baker and co-workers observed pesticide residues in organic fruit samples. The reasons for residues in organic fruit samples were in violation of organic methods of cultivation, pesticide-contaminated water used for irrigation, or pesticide residues left in the soil, if previously used to grow conventional crops.[50]

Apple fruit crop is attacked by a number of insect pests and diseases such as apple scab, San Jose scale, wooly apple aphid, fruit scrapper, defoliating beetles, and tent caterpillar. Pesticides such as chlorpyrifos, endosulfan, carbendazim, propineb, and mancozeb are applied to control these pests. After spraying/treatment, pesticide residues get deposited on the fruits and dissipate slowly depending upon the number of factors such as physiochemical characteristics of pesticide, weather conditions, and time after treatment. Preharvest or postharvest interval or waiting period between spray and harvest is required for safe consumption of fruits. Sometimes, the produce is sent to the market immediately after spraying, and consumers unknowingly consume the product and may be badly affected. Similarly, under Indian conditions, a number of synthetic pesticides, such as deltamethrin, cypermethrin, dimethoate, quinalphos, oxydemeton methyl and carbaryl, are used to control mango crop pests such as mango hopper, mango mealy bug, and fruit fly, as well as powdery mildew and malformations. Deltamethrin at 0.002% does not require any waiting period, but cypermethrin requires 11 days of waiting period. Mango is eaten after removing the peel, but the residues on its peel also find their way into the consumer by contact. Residues of mancozeb and lindane though within the permissible limit were detected in mango fruit samples.[51] The repeated spray of bifenthrin on mango from flowering to 1 month before harvest resulted in residues that persisted on the peel for more than a month, and rate of degradation was very low.[52]

The consumption of pesticide in India is low as compared to other countries, in spite of this, there is widespread contamination of food commodities with pesticide residues due to non-judicious use of pesticides. An earlier survey carried out by the Indian Council of Medical Research, New Delhi, revealed that 51% of food commodities contained pesticide residues, and out of these, 20% had pesticide residues above the MRL values, as compared to 21% contamination with only 2% of samples above the MRL on a worldwide basis.[52]

Now, the scenario in India has started changing very rapidly as new pesticide molecules, whose application rate (as well as persistence in the environment) is very low, are being introduced every year. Heavy-duty pesticides have been either banned or put under restricted use. The pesticide load on the agro-horticultural ecosystem has declined as compared to the last decades. Maximum pesticides in India are used on cotton and rice. The Malwa area of Punjab, which is famous for cotton growing, has been named as the cancer belt of Punjab because pesticides have contaminated the whole environment, including groundwater, and caused cancer among its people. Out of the total pesticides used in India, only 13%–14% is used on fruits and vegetables; despite this, half of the fruits and vegetables were found contaminated with pesticide residues.[52] Pesticide residues in 10% of the samples were above the MRL value. Residues of methyl parathion, endosulfan, chlorpyrifos, Dimethyl 2,2-Dichlorovinyl Phosphate (DDVP), dimethoate, fenitrothion, monocrotophos, cypermethrin, deltamethrin, copper, etc., were above the MRL in fruits and vegetables.[52]

Approaches for Pesticide Use Reduction

Since the excessive and indiscriminate use of pesticides has polluted every component of the environment, people all over the world have realized the need for pesticide reduction so as to prevent the environment from further deterioration. The use of pesticides in fruits and vegetable production has been developed in many countries and the range of pesticides is very large, most of them are chemicals used since the 1950s, quite often pesticides not licensed for use on food crops (typically cotton pesticides and consumers health).

However, it is not an easy task, as agrochemical market and crop protection knowledge are increasingly controlled by few multinationals. Today, more than 80% of worldwide pesticide sales fall to the share of only six companies. Presently, efforts are being made to reduce pesticide applications worldwide by organic and integrated pest management (IPM) approaches. IPM system relies on biological, cultural, and other less chemically intensive approaches to pest management and are best options to minimize residues in several horticultural commodities requiring cooperation between farming community and pesticide dealers while providing high quality and pest-free produce in developing countries. IPM strategies need to be concentrated on improving the cultural practices and reducing pesticide use, substituting less costly ecofriendly pesticide, continuous monitoring of the pest menace, and finally, training the extension officers/farming community to make IPM a successful event.

However, in countries like India, this alternative pest management approach to reduce pesticide use could not find much success due to poor farmer participation. An all-India survey confirmed that 34% of the respondents have no idea about IPM and only less than 5% of them follow complete IPM technology.[53] Nowadays in many states of India, government authorities are promoting natural farming and urging the farming community to adopt it in order to discourage the use of chemical pesticides.

However, IPM techniques are still characterized by a large amount of pesticide use and by the application of many different pesticides, e.g. organic apple production does not use any herbicides and applies only biological control, but fungal diseases like apple scab demand for the intensive use of sulfur and copper in organic apple orchards and copper has a negative impact on the environment. IPM techniques have not been widely implemented on many crops, e.g., wheat. There is a great need to modify "good agricultural practice" and change it to "pesticide avoidance practice" and to improve the education of farmers so as to promote organic farming or natural farming and IPM as the best alternatives for pesticides.

References

1. Dhaliwal, G.S., and Arora, R. *Integrated Pest Management Concepts and Approaches*. Kalyani Publishers, New Delhi, 2001, 427pp.
2. Imhoff, D. *King Cotton—Pesticide Residue is Commonin Cotton Byproducts Used in Agriculture—Brief Article*. Sierra, 1999.
3. Pesticides Industry Sales and Usage: 2000 and 2001 Market Estimates; U.S. Environmental Protection Agency, May 2004, Table 3.4, available at www.epa.gov/oppbead1/pestsales/01pestsales/market_estimates2001.pdf (accessed March 2, 2019).
4. Anonymous. *Scientific Agriculture Prevents Mass Starvation*. Herxter. J. Agvet Div., Hoechst, Australia Inc., 1992.

5. Kennedy, I.R. Pesticides in Perspective: Balancing their benefits with the need for environmental protection and remediation of their residues in seeking agricultural produce free of pesticide residues. In Kennedy, I.R., Skerritt, J.H., Johnson, G.I., and Highley, E. (eds.) *Proceedings of an International Workshop held in Yogyakarta, Indonesia, 17–19 February, 1998*; The Australian Centre for International Research (ACIAR), 1998.
6. Tata Strategic Management Group: Next generation Indian agriculture-Role of crop protection solutions. A report on Indian Agrochemical Industry, 2016.
7. FAOSTAT. 2018, available at http//www.fao.org/faostat/en/#data/EP
8. Pesticides and the Agrochemical Industry. Pesticide use reduction in Germany Pesticide Action Network Germany Pesticide Action Network Europe, available at www.pan-germany.org (accessed February 27, 2019).
9. Dhaliwal, G.S., and Koul, O. *Quest for Pest Management from Green Revolution to Gene Revolution*. Kalyani Publishers, New Delhi, 2010.
10. Anonymous. *Ministry of Statistics and Programme Implementation*. Government of India, 2018. http://www.mospi.gov.in/ Website (assessed on May 15, 2020).
11. Indira, D.P. Pesticides in agriculture—A boon or a curse? A case study of Kerala. *Economic and Political Weekly*, 2010, xlv (26 and 27) EPW, 50–52.
12. Jayakrishnan, T. Health impacts of pesticides used in agriculture. In Devi, P.L. (ed.) *Pesticide Use and Environmental Health*, compendium of papers presented in the workshop, KAU/SANDEE; Kerala Agricultural University Publishers, 2006.
13. Johansen, C.A. Pesticides and pollinators. *Annual Review of Entomology,* 1977, 22: 177–192.
14. Hardin, M.R., Benrey, B., Coll, M., Lamp, W.O., Roderick, G.K., and Barbosa, P. Arthropod pest resurgence: An overview of potential mechanisms. *Crop Protection*, 1995, 14: 3–18.
15. Ruberson, J.R., Nemotom, H., and Hirose, Y. Pesticides and conservation of natural enemies in pest management. In Barbosa, P. (ed.) *Conservation Biological Control*. Academic Press, San Diego, CA, 1998, pp. 207–220.
16. Croft, B.A. *Arthropod Biological Control Agents and Pesticides*. John Wiley & Sons, New York, 1990.
17. De Clercq, P., De Cock, A., Tirry, L., Vinuela, E., and Degheele, D. Toxicity of Diflubezuron and Pyriproxyfen to the predatory bug *Podisus maculiventris*. *Entomologia Experimentalis et Applicata*, 1995, 74: 17–22.
18. Charleston, D.S., Kfir, R., Dicke, M., and Vet, L.E.M. Impact of botanical pesticides derived from *Melia azedarach* and *Azadirachta indica* on the biology of two parasitoid species of the Diamond back moth. *Biological Control*, 2005, 33: 131–142.
19. Ware, G.W., and Whitacre, D.M. *The Pesticide Book*. Meister Pro Information Resources, Willoughby, OH, 2005.
20. Cloyd, R.A. Managing insect and mite pests. In Nau, J. (ed.) *Ball Red Book* (Vol. 2, 18th Edition). Ball Publishing, West Chicago, IL, 2011, pp. l07–119.
21. Cloyd, R.A., and Bethke, J.A. Impact of neonicotinoid insecticides on natural enemies in greenhouse and interior environments. *Pest Management Science*, 2011, 67: 3–9.
22. Cloyd, R.A. Compatibility of insecticides with natural enemies to control pests of greenhouses and conservatories. *Journal of Entomological Science*, 2006, 41: 189–197.
23. Stapel, J.O., Cortesero, A.M., and Lewis, W.J. Disruptive sublethal effects of insecticides on biological control: Altered foraging ability and life span of a parasitoid after feeding on extra floral nectar of cotton treated with systemic insecticides. *Biological Control*, 2000, 17: 243–249.
24. Jacobs, R.J. Kouskolekas, C.A., and Gross, H.R. Jr. Responses of *Trichogramma pretiosum* (Hymenoptera: Trichogrammatidae) to residues of permethrin and endosulfan. *Environmental Entomology*, 1984, 13: 355–358.
25. Elzen, G.W., O'Brien, P.J., and Powell, J.E. Toxic and behavioral effects of selected insecticides on the *Heliothis* parasitoid *Micropliti scroceipes*. *Entomophaga*, 1989, 34: 87–94.

26. Elzen, G.W. Sublethal effects of pesticides on beneficial parasitoids. In Jepson, P.C. (ed.) *Pesticides and Non-Target Invertebrates*. Intercept, Wimborne, 1990, pp. l29–150.
27. Roger, C., Vincent, C., and Coderre, D. Mortality and predation efficiency of *Coleomegilla maculata lengi* Timb. (Col., Coccinellidae) following application of neem extracts (*Azadirachta indica* A. Juss., Meliaceae). *Journal of Applied Entomology*, 1995, 119: 439–443.
28. Desneux, N., Decourtye, A., and Delpuech, J.M. The sublethal effects of pesticides on beneficial arthropods. *Annual Review of Entomology*, 2007, 52: 81–106.
29. Moriarty, F. The Sublethal effects of synthetic insecticides on insects. *Biological Reviews*, 1969, 44: 321–357.
30. Wright, D.J., and Verkerk, R.H.J. Integration of chemical and biological control systems for arthropods: Evaluation in a multitrophic context. *Pesticide Science*, 1995, 44: 207–218.
31. Gerson, U., and Cohen, E. Resurgence of spider mites (Acari: Tetranychidae) induced by synthetic pyrethroids. *Experimental and Applied Acarology*, 1989, 6(1): 29–46.
32. Beers, E.H. Integrated Mite Control: Nontarget effects on predator and prey. *84th orchard Pest and Disease Management Conference*, 13–15 January, 2010, Portland Hilton, Portland, OR, 2010.
33. Huang, J.K., Ruifa, H., Fan, C., Pray, C.E., and Rozelle, S. Bt cotton benefits, costs and impacts in China. *AgBioForum*, 2002, 5(4): 153–166.
34. Pray, C.E., Huang, J., Hu, R., and Rozelle, S. Five years of Bt cotton in China—The benefits continue. *Plant Journal*, 2002, 31(4): 423–430.
35. Qaim, M., and Zilberman, D. Yield effects of genetically modified crops in developing countries. *Science*, 2003, 299: 900–902.
36. Qaim, M., and Dejanvry, A. Genetically modified crops, corporate pricing strategies, and farmers' adoption: The case of Bt cotton in Argentina. *American Journal of Agricultural Economics*, 2003, 85(4), 814–828.
37. Traxler, G., Godoy-Avila, S., Falck-Zepeda, J., and Espinoza-Arellano, J. Transgenic cotton in Mexico: A case study of the Comarca Lagunera. In Kalaitzandonakes, N. (ed.) *The Economic and Environmental Impacts of Agrbiotech*. Kluwer, New York, 2003, pp. 183–202.
38. James, C. *Global Status of Commercialized Transgenic Crops: 2002*. ISAAA Briefs 27. Ithaca, NY, 2002.
39. Pimental, D. Area-wide pest management: Environmental, Economic and food issues. In Vreysen, M.J.B., Robinson, A.S., and Hendrichs, J. (eds.) *Area-wide Control of Insect Pests: From Research to Field Implementation*. Springer, Dordrecht, 2007, pp. 35–47.
40. Sheety, P. K., Murugan, M., and Sreeja, K. G. Crop protection stewardship in India: Wanted or Unwanted. *Current Sciences*, 2008, 95: 457–464.
41. Dixit, A. Herbicide recommendation in different crops. *Crop Care*, 2009, 35(2): 33–38.
42. Pispini, M., Schimpf, M., Lopez, J., and Chandrasekaran, K. Who Benefits from GM Crops? The Rise in Pesticide Use. Friends of the Earth International and Center for Food Safety, 2008, see especially pp. 8–12, available at www.centerforfoodsafety.org/WhoBenefitsPR2_13_08.cfm (accessed December 14, 2010).
43. Anonymous. 2011, available at www.beyondpesticides.org/pesticides/factsheets/Atrazine.pdf (accessed February 18, 2011).
44. Nikolova, S. Pesticide use, issues and how to promote sustainable agriculture in Bulgaria. Published by Pesticide Action Network Germany (PAN Germany) in co-operation with Association, Agrolink, 2004.
45. Rajendran, R.B., and Subramanian, A.N. Chlorinated pesticide residues in surface sediments from the river Kaveri, South India. *Journal of Environmental Science and Health*, 1999, 34: 269–288.
46. Karanth, N.G.K. Challenges of limiting pesticide residues in fresh/vegetables: The Indian experience. In Hanak, E.E., Boutrif, P., and Fabre, M.P. (eds.) *Food Safety Management in Developing Countries*. CIRAD-FAO, Montpellier, 2002, pp. 11–13.

47. Charan, P.D., Ali, S.F., Kachhawa, Y., and Sharma, K.C. Monitoring of pesticide residues in farm gate vegetables of central Aravalli region of Western India. American-Eurasian Journal *of* Agricultural & Environmental *Sciences*, 2010, 7: 255–258.
48. Ranga Rao, G.V., Sahrawat, K.L., Srinivasa, R.C., Binitha, D., Reddy, K.K., and Bharath, B.S. Insecticide residues in vegetable crops grown in Kothapalli Watershed, Andhra Pradesh, India: a case Study. *Indian Journal of Dryland Agricultural Research and Development*, 2009, 24: 21–27.
49. Gupta, P.K. Pesticide exposure. Indian Science. *Toxicology*, 2004, 198: 83–90.
50. Baker, B.P., Benbrook, C., Groth, E., III, and Benbrook, K.L. Pesticide residues in conventional, integrated pest management (IPM)—Grown and organic foods: Insights from three US data sets. *Food Additives & Contaminants*, 2002, 19: 427–446.
51. Bhandari, R. Pesticide residues in vegetables and fruits. *International Journal of Scientific Research in Chemical Sciences*, 2015, 2(1): 11–17.
52. Soudamini, M., and Rekha, A. Persistence of dicofol residues in/on acid lime. *Pesticide Research Journal*, 2005, 17(1): 64–65.
53. Shetty, P.K., Murugan, M., and Sreeja, K.G. Crop protection stewardship in India: Wanted or unwanted. *Current Science*, 2008, 95(4): 457.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

11

Pollution: Pesticides in Natural Ecosystems

J.K. Dubey and
Meena Thakur

Introduction

Pesticides play an important role in boosting the economy of the agricultural industry by providing effective pest control, and their continued use is essential for enhancing the productivity.[1] It is estimated that food supplies would immediately fall to 30%–40% due to the ravages of pests if pesticides are not used.[2] A United Nations report stated that population growth is a major problem facing our planet. In 1900, there were 1.6 billion people on the planet. In 1992, this has risen to 5.25 billion, and by the year 2050, it will reach 10 billion. Developing countries are more affected by this explosive increase in world population. Presently, our dependence on pesticides has increased up to the extent that if modern agriculture was operated without chemical control, the crop production will probably decline in many areas, food price will soar far higher, and food shortage will become more severe. Although pesticides have played an important role in enhancing crop yields, they have also come up with various environmental problems. When present above permissible limits, they act as pollutants, creating pesticide pollution. Many pesticides are present today in different concentrations in various components of our environment such as air, water, and soil. More than 5,00,000 people are either killed or incapacitated every year by poisoning, and most of these casualties occur in developing nations.[3]

Ecologically, however, pesticides have created two major problems that were not previously anticipated. As pollutants, they contaminate numerous natural ecosystems [terrestrial: forest, grassland, desert, etc.; aquatic: fresh water (running water such as spring, stream, or rivers or standing water such as lake, pond, pools, puddles, ditch, and swamp); and marine (deep water bodies such as ocean or shallow ones such as a sea and estuary)] not intended to be targets. Second, most of them have directly/indirectly affected human health. The objective of this entry is to provide basic knowledge on pesticide exposure and to understand issues on residues in the natural ecosystem.

History of Pesticides and Pesticide Problems

The term *pesticide* covers a wide range of compounds including insecticides, fungicides, herbicides, rodenticides, molluscicides, nematicides, plant growth regulators, and others. In the 1940s, dichlorodiphenyltrichloethane (DDT) became the first widely available synthetic insecticide. It was highly effective, but it showed signs of becoming less effective as insects became resistant to it. It accumulated in the bodies of animals and high up the food chain by biomagnifications and bioconcentrations, causing problems with reproduction. Rachel Carson's *Silent Spring* in 1962 drew the attention of environmentalists to the disaster that was gathering pace across the globe. Public awareness of problems with pesticides grew by the 1970s when DDT was banned in many countries. It is still used in some places for malaria control, and it is still present in the bodies of many animals, even hundreds of miles away from where it has not been used. The introduction of other synthetic insecticides—organophosphate (OP) insecticides in the 1960s, carbamates in the 1970s, and pyrethroids in the 1980s, as well as herbicides and fungicides in 1970s–1980s, contributed to a great extent in pest control and agricultural output. The consequences of pesticide use have resulted in serious health implications to man and his environment. There is now overwhelming evidence that some of these chemicals pose potential risk to humans and other forms of life and unwanted side effects to the environment.[4] The worldwide deaths and chronic illnesses due to pesticide poisoning numbered about 1 million per year.[5]

The problem is more serious when pesticides that are banned are used indiscriminately. Banned pesticides are still used on crops that are not consumed directly, e.g., cotton. Few people think of cotton as food, but once the fiber is removed, two-thirds of the cotton crop winds up in the food we eat. Every year in the United States, half a million tons of cottonseed oil goes into processed salad dressings, baked goods, and snacks like Fritos and Goldfish. Another 3 million tons of cottonseed is fed to beef and dairy cattle, which also eat vast amounts of the cotton by-products known as "gin trash."[6]

How Do Pesticides Sprayed on Agro-/ Horti-Ecosystems Enter Natural Ecosystems?

Almost less than 1% of the total pesticides applied actually hit the target organisms.[7] Most reach nontarget sectors of agro-ecosystems and/or spread to surrounding ecosystems as chemical pollutants. The pesticide somehow "leaks" into another ecosystem via movement of water from one body to another via outflow streams or seepage into the water table. Some pesticides might evaporate into the atmosphere and be carried elsewhere by winds. Regardless of how the leak occurs, the pesticide could affect accidental targets; e.g., a volatile insecticide used to control mosquitoes evaporates and kills bees; thus, a wide variety of plants do not get pollinated, thereby affecting their yield. The pesticide may also be taken in by migratory animals (birds in particular) and carried elsewhere; the toxin may affect the birds' reproduction in some way, or those birds might be eaten up by a higher order of predators and the toxin may inflict some injury to them. Either way, this would affect the balance of predation in some land-based ecosystem. The movement of systemic insecticides' active ingredient into floral parts may indirectly impact natural enemies that feed on plant pollen or nectar as a nutritional food source.[8,9]

Pesticide Pollution and Natural Ecosystems

Effect on the Soil Environment

Many pesticides contain chemicals that are persistent soil contaminants; their effects may last for years. Pesticides move with water in soil to groundwater and on soil to surface water. They decrease biodiversity in the soil by killing soil organisms; when life in the soil is killed off, the soil quality deteriorates and has a knock-on effect upon the retention of water. This is a problem for farmers particularly in times of drought.[10] At such times, organic farms have been found to have yields 20%–40% higher than conventional farms. Soil fertility is affected in other ways, too. When pesticides kill off most of the active soil organisms, the complex interactions that result in good fertility break down. Risk assessment of pesticide impact on human health is not an easy and particularly accurate process because of difference in the periods and the levels of exposure types of pesticides mixtures in the field and geographic and meteorological characteristics of the agricultural area of pesticide applied.[11,12] The data is usually collected from various tests conducted on metabolism patterns, acute toxicity, subchronic, chronic, carcinogenecity, genotoxicity, teratogenecity, and generation study using rat as a model mammal or dogs and rabbits.[13]

Application of systemic pesticides as drenches or granules may exhibit indirect effects on natural enemies via several mechanisms including elimination of floral parts by consumption of active ingredient while ingesting plant fluids and contamination of prey ingesting either lethal or sublethal concentration.[14–16]

Plants depend on millions of bacteria and fungi to bring nutrients to their rootlets. When these cycles are disrupted, plants become more dependent upon exact doses of chemical fertilizers at regular intervals. Even so, the incredibly rich interactions in healthy soil cannot be fully replicated by the farmer with chemicals. Hence, the soil and our nutritionist compromised. We get large but watery vegetables and fruits, which often lack natural taste and nutrients and may even contain harmful toxic pesticide residues. Studies of pesticide effects on the soil fauna have reported increased numbers of collembolan, because chemicals reduced populations of natural enemies, especially of predatory mites.[17]

Effect on the Aquatic Environment

Pesticides enter the freshwater ecological systems either from direct application of pesticides for the control of harmful aquatic fauna or as runoff from the treated areas, drift during aerial spraying, and industrial effluents from washing and spraying of equipment and containers. Several groups of organochlorine pesticides such as DDT, endosulfan, and chlorinated phenoxy acetic acid used as herbicides and fungicides such as hexachlorobenzene and pentachlorophenol are of interest in water pollution. Because of their solubility in water and tendency to be absorbed on solid surfaces, only traces of these chemicals are found in solid surfaces and treated water.[18]

Microorganisms form a vital part of the freshwater environment. Bishop[19] measured the effects of DDT on Mastigophora, Infusoria, and Sarcodina in ponds near Savannah, Georgia, and found little change in population numbers after treatment at relatively low rates. Hoffman and Olive[20] found that the growth of populations in Colorado lakes was inhibited after the addition of rotenone and toxaphene. Phytoplankton (beneficial/detrimental) can be seriously affected by agricultural chemicals. DDT sprays have caused serious reductions of bottom-dwelling invertebrates, the reductions in some cases amounting to 95% of the population. Malathion has also caused destruction of stream invertebrates. Cushing and Olive[21] found that toxaphene and rotenone reduced numbers of midge larvae in Colorado reservoirs and algae. On higher plants in the freshwater environment, adsorption of pesticides in/on the vegetation resulted in phytotoxicity, which either retarded the growth of or killed aquatic plants. Kolleru Lake is the largest natural freshwater body of Andhra Pradesh in India where agriculture and aquaculture are some of the primary activities at the lake basin. The increased use of pesticides in agriculture and aquaculture had a negative impact on the quality of water in the lake.[22,23]

A major environmental impact has been the widespread mortality of fish and marine invertebrates due to the contamination of aquatic systems by pesticides. Most of the fish in Europe's Rhine River were killed by the discharge of pesticides, and at one time, fish populations in the Great Lakes became very low due to pesticide contamination. In addition, many of the organisms that provide food for fish are extremely susceptible to pesticides, so the indirect effects of pesticides on the fish food supply may have an even greater effect on fish populations. Some pesticides, such as pyrethroids, are extremely toxic to most aquatic organisms. It is evident that pesticides cause major losses in global fish production.

Effect on the Terrestrial Environment

A wide variety of pesticides are applied on horticultural and agricultural crops. Some of them are highly specific and others are broad spectrum; both types can affect terrestrial wildlife, soil, water systems, and humans. The misuse of pesticides can cause valuable pollinators such as bees and hoverflies to be killed, and this in turn can badly affect food crops. Bees are extremely important in the pollination of crops and wild plants; about 33% of all crops require pollination. Although pesticides are screened for toxicity to bees, and the use of pesticides toxic to bees is permitted only under stringent conditions, many bees are killed by pesticides, resulting in the considerably reduced yield of crops dependent on bee pollination. Bee population has been suffering a serious decline in recent years. Without bees, many food crops would simply fail to grow.

Neonicotinoid pesticides developed in 1980 are under fire for risks to pollinators, and European Union Commission had already imposed this class of pesticides as researchers have become concerned about their potential to harmless birds, earthworms, aquatic insects, and pollinators.[24] Three neonicotinoids—thiamethoxam, clothianidin, and imidacloprid—posed an unacceptable risk to bees and were banned for use for some period on flowering crops such as corn, oilseed rape, and sunflower upon which the bees feed.[25]

It has been observed that through natural selection, some pests eventually become quite resistant to pesticides and farmers may need increasing amounts of pesticides, making the problem worse. Orchards are complex ecosystems easily perturbed by the extensive use of pesticides, and there are many instances of increased pest attacks in orchards after the use of pesticides, e.g., outbreaks of codling moth, leafrollers, aphids, scales, and tetranychid mites.[26] When pesticides were first used on tropical cotton crops, they controlled two or three important pests of the crops and greatly increased yields. Within a few seasons, however, the chemicals reduced the population of natural enemies and a number of other arthropod species became pests.[27]

Amphibians such as frogs are particularly vulnerable to concentrations of pesticides in their habitat. Atrazine, the most heavily used herbicide, is regularly detected in drinking water supplies in the Midwest, the United States, and exposure to extremely low levels of atrazine can cause sex change and/or deformities in frogs, fish, and other organisms.[28] Based on this evidence, and the widespread presence of atrazine in drinking water supplies, the European Union (EU) announced a ban on atrazine in 2006. The US Environmental Protection Agency re-registered atrazine in 2003 despite objections from scientists and environmental groups.

Pesticides have had some of their most striking effects on birds, particularly those in the higher trophic levels of food chains, such as bald eagles, hawks, and owls. These birds are often rare, endangered, and susceptible to pesticide residues such as those occurring from the bioconcentration of organochlorine insecticides through terrestrial food chains. Pesticides may kill grain- and plant-feeding birds, and the elimination of many rare species of ducks and geese has been reported. Populations of insect-eating birds such as partridges, grouse, and pheasants have decreased due to the loss of their insect food in agricultural fields through the use of insecticides. Pesticides can affect animal reproduction directly, as evident by the deleterious effect of the persistent organochlorine insecticides on reproduction in receptors and other birds. The US National Academy of Sciences stated that the DDT metabolite, dichlorodiphenyldichloroethane (DDE), causes eggshell thinning and that the bald eagle population in the United States declined primarily because of exposure to DDT and its metabolites.[29] Fish-eating birds are more severely affected than terrestrial predatory birds, because the former acquire more pesticides via their food chain.[30] Pesticides can also affect reproduction in invertebrates, e.g., sublethal doses of DDT, dieldrin, and parathion increased egg production of Colorado potato beetle after 2 weeks by 50%, 33%, and 65%, respectively.[31] Aquatic ecosystems with flowing water can usually recover their structure and function more quickly from pesticide effects than ponds with standing water.

Effects on Humans

There are a number of ways in which humans can be exposed to pesticides through the environmental route. Man's primary exposure to pesticides is probably via those used domestically in wood preservation or as household insecticides. Pesticides can endanger workers during production and transportation or during and after use. Bystanders may also be affected at times, e.g., walkers using public rights-of-way on adjacent land or families whose homes are close to crop-spraying activities. One of the main hazards of pesticide use is to farm workers and gardeners. A recent study by the Harvard School of Public Health in Boston discovered a 70% increase in the risk of developing Parkinson's disease for people exposed to even low levels of pesticides.[32]

The effects of pesticide residues in food and water probably cause a great public concern, although reports of clinical poisoning due to residues are extremely rare. Their residual population in food commodities is alarming. Leafy vegetables, cereals, fruits, rice, meat, milk, fish, and even human milk have been contaminated by various pesticides in a range of 0.1–25.7 mg/kg. The herbicide 2,4-D is identified as a carcinogen in humans and dogs. Acephate is a mutagen, carcinogen, fetotoxic, feminizes rats, and kills birds. In a multicountry study (Belgium, China, Federal Republic of Germany, India, Israel, Japan, Mexico, Sweden, the United States, and Yugoslavia) on the assessment of human exposure to selected organochlorine compounds, the residue levels for pp'-DDE and β-HCH were found to be higher in the human milk samples collected from developing countries such as China, India, and Mexico than in the participating developed countries. A higher level of these chemicals in mother's milk is a clear-cut reflection of their increased burden through their translocation passage.[33]

A number of bottles of wine were tested for pesticide residues, and 100% of conventional wines included in the analysis were found to contain pesticides, with one bottle containing ten different pesticides. On an average, each wine sample contained more than four pesticides. The analysis revealed 24 different pesticide contaminants, including five classified as being carcinogenic, mutagenic, neurotoxic, or endocrine disrupting by the EU. Human health is at risk when chemical residues are present in so much of our food supplies.

Children are particularly vulnerable to the toxic effects of pesticides. Studies have found higher rates of brain cancer, leukemia, and birth defects in children who suffered early exposure to pesticides. A survey of baby foods in 2000 showed detectable pesticide residues in nearly 50% of foods sampled. Fourteen percent of foods tested showed residues of more than one pesticide at levels 30 times the proposed limit of 0.01 mg/kg.[34] A UK government report in 2003 showed that more than 70%, 90%, 61%, 54%, and 35% of apples, lemons, bread, rice, and potatoes analyzed had pesticide residues, respectively. The main source of exposure to pesticides for most people is through diet. A study in 2006 measured organophosphorus levels in 23 school children before and after changing their diet to organic food. The levels of organophosphorus exposure dropped immediately and dramatically when the children began the organic diet.[34]

Over the last few decades, agricultural pesticides have become common household items in rural areas of the developing world. In 2014, the Natural Crime Records Bureau of India reported 5650 farmer suicides.[35] In India, the first report of poisoning due to pesticides was from Kerala in 1958, where more than 100 people died after consuming wheat flour contaminated with parathion.[36] This prompted the Special Committee on Harmful Effects of Pesticides constituted by the Indian Council of Agricultural Research to pay more attention on the problem.[37] Exposure to accidental emissions of methyl isocyanate from a pesticide factory in Bhopal, India, killed more than 5000 people, leaving more than 50,000 with permanent damage.

Alternatives for Pesticide Problems

The toxic effects of pesticides on our foods, land, and their effects on the health of human beings and their progeny make it an issue that is becoming more and more crucial. Integrated pest management (IPM; pest surveillance, use of crop varieties resistant to pest, sound cultural practices, biological control, and use of ecofriendly pesticides) emphasizes the need for simpler and ecologically safer measures for pest control to reduce environmental pollution and other problems caused by excessive and indiscriminate use of pesticides. Preference should be given to organic foods that are grown without toxic pesticides by organic methods. There are now many biological control tactics available where benign species are used to manage less benign ones. Ladybirds (ladybugs) are often introduced to control aphids (greenfly and others). Organic nontoxic sprays are used to stimulate the soil. They work by stimulating fungi in the soil that help to feed the plants and help them in developing resistance to disease and insect attack. There are also many successful barrier methods that help to deter insect attacks (the use of nets to ward off birds and larger insects). Companion planting is also used; garlic, for example, helps some plants resist insect attacks. There are a small number of organic pesticides that are legitimate to use in organic food production system. Some of these can be made at home using simple ingredients such as soap and alcohol.

It is evident that misuse, overuse, and abuse of pesticides lead to many environmental problems as discussed. Pesticides must be used as part of a planned systematic pest management program utilizing as many control techniques as applicable (IPM).

Novel pesticides' modes of action, improved safety profiles, and the implementation of alternative cropping systems which are less dependent on pesticides could minimize exposure to pesticides and undesirable exposure on human health. Moreover, the use of appropriate and well-maintained spraying equipment along with taking all the precautions required in all the stages of pesticide handling could also reduce exposure to pesticides. The overall pesticide handling according to regulations reducing the public concerns in food and drinking water minimize the effects of pesticides in human health and environment.[38]

Emphasis should be placed on using all the techniques of organic farming and supplementing these with the use of pesticides, i.e., using pesticides as part of an organic farming system. The IPM approach will help to minimize the effects of pesticide pollutants on the environment and natural ecosystems and will also help in economic and ecological sustainable food production.

References

1. Kent, J. Education and training in farm chemical management. *Proc. Conf. Agriculture, Education and Information Transfer*; Murrumbidgee College of Agriculture, Yanco, 1991.
2. Anonymous. *Scientific Agriculture Prevents Mass Starvation*; The Herxter. J. of the Agvet Division, Hoechst, Australia Inc., 1992.
3. Anonymous. Water pollution—Case study, 2010, www.environmentandpeople.org. All Rights Reserved © Environment and People. Web site designed and maintained by HS visual FX. (accessed October 8, 2010).
4. Igbedioh, S.O. Effects of agricultural pesticides on humans, animals and higher plants in developing countries. *Archives of Environmental Health* **1991**, *46*, 218.
5. Environews Forum. Killer environment. *Environmental Health Perspectives* **1999**, *107*, A62.
6. Imhoff, D. *King Cotton 1999—Pesticide Residue is Commonin Cotton Byproducts Used in Agriculture; Brief Article*; Sierra, May 1999.
7. Pimentel, D.; Levitan, L. Pesticides: Amounts applied and amounts reaching pests. *BioScience* **1986**, *36*, 86–91.
8. Kiman, Z.B.; Yeargan, K.V. Development and the reproduction of the predator *Orius insidious* (Hemiptera: Anthocoridae) reared on the diets of the selected plant material and arthropod prey. *Annals of the Entomological Society of America* **1985**, *78*, 464–467.

9. Hagen, K.S. Ecosystem analysis: Plant cultivars (HPR), Entomophagus species and food supplements. In: Boethal, D.J.; Eikenbary, R.D. (eds.) *Interactions of Plant Resistance and Parasitiods and Predators of Insects*; John Wiley & Sons, Inc., New York, 1986, pp. 151–197.
10. Roger, A.; Simpson, I.; Oficialc, R.; Ardales, S.; Jimen, R. Effects of pesticides on soil and water microflora and mesofauna in wetland ricefields: A summary of current knowledge and extrapolation to temperate environments. *Australian Journal of Experimental Agriculture* **1994**, *34*, 1057–1068.
11. Bolognesi, C. Genotoxicity of pesticides: A review of human biomonitoring studies. *Mutation Research* **2003**, *543*, 251–272.
12. Pastor, S.; Creus, A.; Parron, T.; Cebulska-Wasilewska, A.; Siffel, C.; Piperakis, S.; Marcos, R. Biomonitoring of four European populations occupationally exposed to pesticides: Use of micronuclei as biomarkers. *Mutagenesis* **2003**, *18*, 249–258.
13. Matthews, G.A. *Pesticides: Health, Safety and the Environmen*; Blackwell Publishing, Oxford, 2006.
14. Tillman, P.G.; Mullinix, B.G. Comparison of susceptibility of pest *Euschistusservus* and predator *podisus maculiventris* (Heteroptera: Pentatomidae) to selected Insecticides. *Journal of Economic Entomology* **2004**, *97*, 800–806.
15. Krischiik, V.A.; Landmark, A.L.; Heimpel, G.E. Soil applied imidacloprid is translocated to nectar and kills nectar feeding *Anagyrus pseudococci* (Girault) (Hymenoptera: Encyrtidae). *Environonmental Entomology* **2007**, *36*, 1238–1245.
16. Szczepaniec, A.; Creaky, S.F.; Laslowski, K.L.; Nyrop, J.P.; Raupp, M.J. Neonicotinoid insecticide imidacloprid causes outbreaks of spider mites on elm trees in urban landscapes. *PLoS ONE* **2011**, *6*(5), e2018.
17. Edwards, C.A.; Thompson, A.R. Pesticides and the soil fauna. *Residue Reviews* **1973**, *45*, 1–79.
18. Beitz, H.; Schmidt, H.; Herzel, F. Occurrence, toxicological and ecotoxicological significance of pesticides in ground-water and surface water. *Chemistry of Plant Protection* **1994**, *8*, 3–53.
19. Bishop, E.L. Effects of DDT mosquito larviciding on wildlife; the effects on the plankton population of routine larviciding with DDT. *Public Health Reports* **1947**, *62*(35), 1263–1268.
20. Hoffman, D.A.; Olive, J.R. The effects of rotenone and toxaphene upon plankton of two Colorado reservoirs. *Limnology and Oceanography* **1961**, *6*(2), 219–222.
21. Cushing, C.E. Jr.; Olive, J.R. Effects of toxaphene and rotenone upon the macroscopic bottom fauna of two northern Colorado reservoirs. *Transactions of the American fisheries Society* **1957**, *86*, 294–301.
22. Victor, L.L. Fleisher J.H. and B.J. Jandorf (ed.). Pesticides in sea water and the possibilities of their use in mariculture. In: *Research in Pesticides*; Chichester Academic Press, New York and London, 1965.
23. Edwards, C.A. U.S. Environmental Protection Agency Website. Pesticides, www.epa.gov/pesticides (accessed October 8, 2010).
24. Campwell, P.J. Declining European bee health: Banning the neonicotinoids is not the answer. *Outlooks on Pest Management* **2013**, *24*(2), 52–57.
25. Cressey, D. Europe debates risk to bees. *Nature* **2013**, *496* (7446), 408.
26. Brown, A.W.A. *Ecology of Pesticides*; John Wiley & Sons, New York, 1978.
27. ICAITI (Instituto Centro Americano de Investigacion y Technología Industrial). *An Environmental and Economic Study of the Consequences of Pesticide Use in Central American Cotton Production: Final Report*; Central American Research Institute for Industry (Guatemala), United Nations Environment Programme, Nairobi, Kenya, 1977.
28. Anonymous. 2011, www.beyondpesticides.org/pesticides/factsheets/Atrazine.pdf (accessed October 8, 2010).
29. Liroff, R.A. Balancing risks of DDT and malaria in the global POPs treaty. *Pesticide Safety News* **2000**, *4*, 3.
30. Pimentel, D. *Ecological Effects of Pesticides on Non-Tar-get Species*; Executive Office of the President, Office of Science and Technology, Washington, DC, 1971, 220 pp.

31. Abdallah, M.D. The effect of sublethal dosages of DDT, parathion, and dieldrin onoviposition of the Colorado potato beetle (*Leptinotarsa decemlineata* Say) (Coleoptera: Chrysomelidae). *Bulletin of the Entomological Society of Egypt, Economic Series* **1968**, *2*, 211–217.
32. Ascherio, A.; Chen, H.; Weisskopf, M.G.; O'Reilly, E.; McCullough, M.L.; Calle, E.E.; Schwarzschild, M.A.; Thun, M.J. Pesticide exposure and risk for Parkinson's disease. *Annals of Neurology* **2006**, *60*(2), 197–203.
33. UNEP/WHO. *Assessment of Human Exposure to Selected Organochlorine Compounds through Biological Monitoring*; Slorach, S.A.; Vaz, R., Eds.; Swedish National Food Administration, Uppsala, 1983, p. 49.
34. Anonymous. *Pesticide Problems are a Growing Concern*, 2011. Copyright © Greenfootsteps.com 2006–2010.
35. National Crime Reports Bureau, ADSI Report Annual-2014. Government of India, 2014, p. 242.
36. Karunakaran, C.O. The Kerala food poisoning. *Journal of Indian Medical Association* **1958**, *31*, 204.
37. ICAR. *Harmful Effects of Pesticides. Report of the Special Committee of ICAR*; Wadhwani, A.M.; Lall, I.J., Eds.; Indian Council of Agricultural Research, New Delhi, 1972, p. 44.
38. Damalas, C.A.; Eleftherohorinos, I.G. Pesticide exposure, safety and risk assessment indicators. *International Journal of Environment Research and Public Health* **2011**, *8*, 1402–1419.

12

Polychlorinated Biphenyls (PCBs)

Marek Biziuk and
Angelika Beyer

Introduction

There is growing concern about the trace quantities of highly chlorinated organic compounds (e.g., dioxins, PCBs, and certain pesticides) that exist in diverse environmental media (air, water, soils, sediments, and biota), enter the trophic chain, and reach humans and wildlife. Consequently, there is growing scientific, regulatory, and social interest in measuring the levels of chlorinated chemicals in environmental media, and in determining the environmental effects of such contamination.

Representative of these synthetic organic chlorine compounds are polychlorinated biphenyls commonly known as PCBs. PCBs are man-made chemicals that never existed in nature until the 1900s when they started to be released into the environment by manufacturing companies and consumers. Although production of PCBs was banned, when their ability to accumulate in the environment and to cause harmful effects became apparent (in 1970, Sweden; in 1972, Japan; in 1977, the United States), these chemicals still are found in the environment.[1]

PCBs make up a group of 209 individual chlorinated biphenyl rings—congeners. They were typically manufactured as mixtures of 60 to 90 different congeners and were usually contaminated with small amounts of very toxic chemicals such as polychlorinated dibenzofurans (furans) or polychlorinated dibenzodioxins (dioxins). The trade names of some commercial PCB mixtures are Aroclor (United States), Clophen (Germany), Fenclor (Italy), Kanechlor (Japan), and Phenoclor (France).[2]

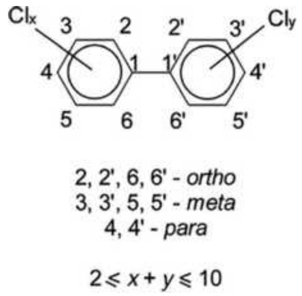


FIGURE 1 Structure of PCBs.

Chemical Identity

PCBs are a mixture of individual chemicals. The general chemical structure of chlorinated biphenyl (two benzene rings with a carbon–carbon bond between carbon 1 on one ring and carbon 1′ on the second ring) with a varying number of chlorines is shown in Figure 1. It can be seen from the structure that a large number of chlorinated compounds are possible (209 possible congeners) in which 2–10 chlorine atoms are attached to the biphenyl molecule. Chlorines can be attached to any of the carbons by removing the hydrogen from that carbon and substituting the chlorine in its place. The common nomenclature used for identifying the location of chlorine atoms on the biphenyl rings is as shown in Figure 1.

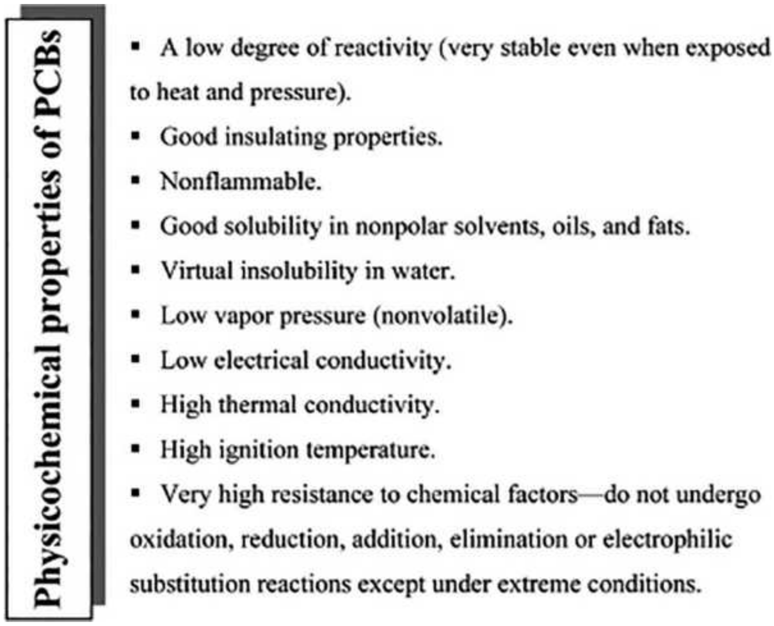
The congeners are arranged in ascending numerical order using a numbering system developed by Ballschmiter and Zell[3] that follows the International Union of Pure and Applied Chemistry (IUPAC) rules of substituent characterization in biphenyls. The resulting PCB numbers, also referred to as congener, IUPAC, or BZ numbers, are widely used for identifying individual congeners.

Properties of PCBs

Because of physical and chemical properties of PCBs (summarized in Figure 2[4–6]), these chemicals were quickly acclaimed as an industrial breakthrough. PCBs are either oily liquids or solids that are colorless to light yellow without known smell or taste. In general, PCBs are relatively insoluble in water (solubility decreases with increased chlorination) but freely soluble in nonpolar organic solvents and biological lipids.[7] Because of their thermal stability, they do not easily burn, hence their past popular use as coolants, as insulating materials, and for electrical applications. The properties of PCBs vary from one congener to the next, e.g., color of PCB mixture darkens, viscosity increases, the flash point rises, and the substance becomes less combustible with rising chlorine content. Also, as the number of chlorines in a PCB mixture increases, the mixture is more stable and thus resistant to biodegradation. The congeners with large numbers of chlorines are also proving to be the ones that present the greatest environmental and health risks.

The properties that make PCB mixtures so desirable and applicable in industry (general inertness, thermal stability) are the ones that make the mixtures so hazardous to the environment. The toxicity of a PCB congener is dependent on the number of chlorines present on the biphenyl structures and the positions of the chlorines. The congeners in which there is a coplanar confirmation with chlorine substituents on the *meta* and *para* positions of the phenyl rings are the most toxic and bioaccumulative ones. For instance, congeners with chlorines in both *para* positions (4 and 4′) and at least 2 chlorines at the *meta* positions (3, 5, 3′, 5′) are considered to be "dioxin like" and are particularly toxic.[8]

The high thermal and chemical resistance of PCB congeners means that they do not readily break down when exposed to heat or chemical treatment. However, since PCBs do not break down, they remain in the environment. Due to their persistence in the environment and the fact that they are poorly biodegraded, PCBs accumulate in the environment.



Physicochemical properties of PCBs

- A low degree of reactivity (very stable even when exposed to heat and pressure).
- Good insulating properties.
- Nonflammable.
- Good solubility in nonpolar solvents, oils, and fats.
- Virtual insolubility in water.
- Low vapor pressure (nonvolatile).
- Low electrical conductivity.
- High thermal conductivity.
- High ignition temperature.
- Very high resistance to chemical factors—do not undergo oxidation, reduction, addition, elimination or electrophilic substitution reactions except under extreme conditions.

FIGURE 2 Physicochemical properties of PCBs.

Global Distribution and Sources

PCBs were first produced in 1929 for a wide variety of uses because of their unique physical properties that made them attractive compounds for industries (see Table 1[9–12]). As more uses were found for PCBs, their production increased exponentially. In Table 1, there are identified PCB use areas based on their presence in closed, partially closed, and open systems. These designations refer to how easily the PCBs contained within a product can escape to the surrounding environment. In closed applications, PCBs are held completely within the equipment. Under ordinary circumstances, no PCBs would be available for exposure to the user or the environment. However, PCB emissions may occur during equipment servicing/repairing and decommissioning, or as a result of damaged equipment. Partially closed PCB applications, in which the PCB oil is not directly exposed to the environment, but may become so periodically during typical use, lead also to PCB emissions, through air or water discharge, whereas in open systems, PCBs are in direct contact with the environment and thereby may be easily transferred to the environment. Generally, closed and partially closed systems contain PCB oils or fluids. The PCBs in open systems take on the form (type of media) of the product they have been used in as an ingredient. Therefore, PCBs in open applications may be found in forms ranging from paint to plastic or rubber.[9]

The first indication that PCBs may be damaging to human health occurred four decades after PCBs were first introduced into the environment. Preliminary studies suggested that PCB s may pose a serious health threat to humans, and at the same time, there were indications of widespread distribution and longevity throughout the environment. As more attention was turned towards PCBs, it became clearer that PCBs were having a negative impact on many biological systems. Eventually, all production and importation of PCBs was banned in the 1970s.[13] Today, the production of PCBs has been ceased in many countries with the exception of small quantities manufactured strictly for research purposes. However, the ecotoxicological problems created by PCB contamination will be evident for many years to come, despite the restrictions of PCB utilization.

TABLE 1 Examples of Applications of PCBs

Closed Applications	Partially Closed Applications	Open Applications

• Electrical transformers
• Electrical capacitors
  • Power factor capacitors in electrical distribution systems
  • Lighting ballasts
  • Motor start capacitors in refrigerators, heating systems, air conditioners, hair dryers, water well motors
  • Capacitors in electronic equipment including television sets and microwave ovens
• Electrical motors in some specialized fluid-cooled motors
• Electrical magnets in some fluid-cooled separating magnets
	• Heat transfer fluids
	• Hydraulic fluids
	• Vacuum pumps
	• Switches
	• Voltage regulators
	• Liquid-filled electrical cables
	• Liquid-filled circuit breakers
		• Lubricants
		  • Immersion oil for microscopes (mounting media)
		  • Brake linings
		  • Cutting oils
		  • Lubricating oils
		• Casting waxes
		  • Pattern waxes for investment castings
		• Surface coatings
		  • Paints
		  • Surface treatment for textiles
		  • Carbonless copy paper
		  • Flame retardants
		  • Dust control
		• Adhesives
		  • Special adhesives
		  • Adhesives for waterproof wall coatings
		• Plasticizers
		  • Gasket sealers
		  • Filling material in joints of concrete
		  • Polyvinyl chloride plastics
		  • Rubber seals
		• Inks
		  • Dyes
		  • Printing inks
		• Other uses
		  • Insulting materials
		  • Pesticides

Although the manufacture, processing, distribution, and use of PCBs are widely prohibited, they have been released to the environment solely by human activity and still are redistributed from one environmental compartment to another.[1] There still exist a lot of different activities that generate PCB wastes. PCBs entered air, water, and soil during their manufacture, use, and disposal, mainly by leakage of supposedly closed systems, from landfill sites, incineration of waste, agricultural lands, industrial discharges, and sewage effluents. For more details, see Table 2.[14–23]

Nowadays, PCBs are present in all environmental media because of global circulation. The most important mechanism for global dispersion of these contaminations is atmospheric transport, which depends on the number of chlorines present on the biphenyl molecule:

- Biphenyls with one chlorine atom remain in the atmosphere.
- Those with one to four chlorines gradually migrate toward polar latitudes in a series of volatilization/de- position cycles between the air and the water and/or soil.
- Those with four to eight chlorines remain in midlatitudes.
- Those with eight to nine chlorines remain close to the source of contamination.[24]

There are two classic approaches to model the distribution of PCBs (and other persistent organic pollutants [POPs]).[25] Multicompartment models use just limited meteorological data but include detailed descriptions of the partitioning of the species within and between the different environmental media,[26] while chemistry transport models have a detailed treatment of transport and chemistry in

TABLE 2 Sources and Transport of PCBs within the Environment

Part of the Environment	Sources of PCBs in the Environmental System	Factors Influencing the Pattern and Rates of PCB Movement in the Media
Aquatic system— higher concentrations in the sediments of aquatic systems	• Accidental spills of PCB-containing hydraulic fluids • Improper disposal • Combined sewer overflows, or storm water runoff • Runoff and leaching from PCB- contaminated sewage sludge applied to farmland	• Properties of PCB congeners—desorption of PCBs from particulate is more likely to occur from lower-chlorinated, more water-soluble PCB congeners • Sorption reactions—with the chlorine content of PCB congener, surface area increases; with the organic content of the sediment, sorption increases • Sudden hydrographic activity like flooding or dredging causes sediments to be resuspended and redistributed and can cause the release of PCBs from sediments to overlying waters
Air system	• Volatilization from soil and water • Escape from uncontrolled landfills and hazardous waste sites • Incineration of PCB-containing wastes • Leakage from older electrical equipment • Improper waste disposal or spills • Leakage from supposedly closed systems • Incineration of waste • Industrial discharges • Sewage effluents	• Air temperature • Wind speed • Storm frequency • Rainfall rates • Volatility of individual PCB congeners
Soil system	• Accidental leaks and spills • Release from contaminated soils in landfills and hazardous waste sites • Deposition of vehicular emissions near roadway soil • Land application of sewage sludges containing PCBs	• Sorption reactions—i.e., highly chlorinated congeners are sorbed by soils and remain significantly immobile against leaching • Vapor phase transport—PCB congeners have a moderate vapor pressure, so vapor phase transport may allow for redistribution or migration through the saturated soil pores

the atmosphere but a rather simple description of the compartments other than the atmosphere.[27,28] The multicompartment models have been successful in describing the global distribution of POPs and their long-term environmental fate in the various compartments.[29,30] Unfortunately, not many direct comparisons between model results and measurement data have been made so far.[31]

Environmental Fate

PCBs, as it was indicated before, can partition between environmental media such as atmosphere, oceans, rivers, or soils. Differences in partitioning behavior among PCBs reflect differences in their physicochemical properties and persistence in the various media.[32]

PCBs do not readily break down in the environment and thus may remain there for a very long time. PCBs can travel long distances in the air and be deposited in areas far away from where they were released. In water, a small amount of PCBs may remain dissolved, but most stick to organic particles and bottom sediments. These toxic compounds can also bind strongly to soil. The degradation and transformation of PCBs entail difficult mechanisms of chemical, biochemical, or thermal destruction.[33] These substance may be (and are) accumulated through the trophic chain and reach aquatic organisms, fish, and humans. Consequently, PCBs accumulate in fish and marine mammals, reaching levels that may be many thousands of times higher than in water. That is why there is great interest in different pathways for PCB loss, such as volatilization, adsorption on organic matter, and biodegradation, which can reduce PCB bioavailability.

Volatilization

PCBs enter the atmosphere from volatilization from both soil and water surfaces.[34] It was reported over 35 years ago by Haque,[35] who found minimal PCB loss at ambient temperature. Heat, airflow (hood storage), coarse grain size, high water content, and enrichment in lower *ortho*-chlorinated congeners all were expected to increase the rate and extent of PCB volatilization. As indicated by their higher vapor pressures, the lower-chlorinated homologs in particular are subject to volatilization.[36] This can result in both a loss and a source of lighter homologs—a source because upon volatilization, the atmosphere is enriched with these homologs, which are then subject to atmospheric deposition. Once in the atmosphere, PCBs are both present in the vapor phase and sorbed to particles. PCBs in the vapor phase appear to be more mobile and are transported further than particle-bound PCBs.[24]

Because of their persistence and semi-volatility, PCBs have a great potential for long-range atmospheric transport, which enables them to migrate from the mid-latitudes to the Arctic regions, for instance.[31,37] Atmospheric transport may occur in many mechanisms, one of them is a mechanism known as cold condensation,[37] by which some PCBs are preferably removed from the atmosphere in cold regions and by which they can reach surprisingly high concentrations in the Arctic environment where they can bioaccumulate in animals and humans. Less volatile compounds that sorb strongly to atmospheric particles or that dissolve easily in rain droplets tend to have a more limited potential for long-range atmospheric transport, whereas semivolatile species can be transported over long distances in one or more steps towards the Arctic region.[37,38]

Both wet deposition and dry deposition remove PCBs from the atmosphere.[39]

Adsorption and Desorption on Organic Matter and Bioavailability

Once released into the environment, PCBs adsorb strongly to soil and sediment. As a result, these compounds tend to persist in the environment, with half-lives for most congeners ranging from months to years. Over time, contaminated sediments can be a source of hydrophobic organic contaminants (such as PCBs) and a significant health risk to aquatic food webs.[40] Leaching of PCBs from sediment and soil is slow, particularly for the more highly chlorinated congeners, and translocation to plants via soil is insignificant. Cycling of PCBs through the environment involves volatilization from land and water surfaces into the atmosphere, with subsequent removal from the atmosphere by wet or dry deposition, then revolatilization.[41]

Sorption properties of PCBs play a significant role in their mobility, ultimate fate in the sediments, and availability for degradation. The literature suggests that PCBs preferentially adsorb onto organic matter over adsorbing onto clay. Moreover, PCBs can sorb to dissolved organic matter (DOM) or particulate organic matter (POM). When associated with DOM, PCB contaminants are unavailable for uptake by organisms and, hence, become less bioavailable. In contrast, although PCBs sorbed to POM prevents or constrains direct uptake of PCBs, these contaminants are still available to the detrital food web, which is an important pathway in rivers. Planar PCBs bind strongly to POM and are less bioavailable.[42] Highly chlorinated homologs sorb strongly to POM and are not assimilated easily by detritus feeders.[43]

Bioavailability of sedimentary PCBs is traditionally assessed by measuring PCB uptake into benthic organisms over a standard exposure time. More recently, passive samplers have been used experimentally to estimate bio-availability.[44–46]

It is known that a combination of binding processes (sorption) and mass-transport processes (diffusion) is responsible for the partitioning of PCBs between aqueous and solid phases, and for their transport between these phases. These processes are also directly involved in and affect the environmental fate of PCBs. Precise quantitative predictions of phase speciation may allow an a priori estimate of the directly bioavailable, dissolved fractions of pollutants, as well as their tendency for long-term dispersion in the environment. Such predictions are critical in assessing the environmental risk from PCB contamination.[47]

The data generated during bench-scale adsorption studies and molecular-level study of the mechanism of adsorption of PCBs on substrates in the environment can help in effective desorption and the destruction of the persistent PCBs.[48]

Biodegradation and Transformation

The environmental persistence of PCBs results primarily from the inability of natural aquatic and soil biota to metabolize and/or degrade the compound at a significant rate. Studies on the biodegradation (degradation by bacteria or other microorganisms) of PCBs show that there are two biologically mediated processes for the degradation of PCBs: anaerobic and aerobic.[49] Microorganisms participate in the biodegradation by producing enzymes, which modify the organic pollutant into simpler compounds in such a way that the negative effects may be minimized. Biodegradation is of two forms:

- Mineralization—competent organisms use the organic pollutant as a source of carbon and energy resulting in the reduction of the pollutant to its constituent elements.
- Cometabolism—it requires a second substance as a source of carbon and energy for the microorganisms, but the target pollutant is transformed at the same time.[50,51]

If the products of cometabolism are amenable to further degradation, they can be mineralized; otherwise, incomplete degradation occurs. This may result in the formation and accumulation of metabolites that are more toxic than the parent molecule requiring a consortium of microorganisms, which can utilize the new substance as source of nutrients.[49]

The effectiveness of biodegradation depends on many factors, which are summarized in Table 3.[51–53]

Biodegradation is the only process known to degrade PCBs in soil or aquatic systems. Theoretically, the biological degradation of PCBs should result in CO_2, chlorine, and water. This process involves the removal of chlorine from the biphenyl ring (anaerobic reductive dechlorination) followed by cleavage and oxidation of the resulting compound (aerobic oxidative degradation).[54] The anaerobic process removes chlorine atoms of highly chlorinated PCBs, those with five or more chlorine atoms, which are then mineralized under aerobic condition.[49]

Under anaerobic condition, reductive dechlorination of PCBs occurs in soils and sediments. Different microorganisms with distinct dehalogenating enzymes, each exhibiting a unique pattern of congener selectivity resulting in various patterns of PCB dechlorination, exist in PCB- contaminated sites,[55] including the following isolated bacteria: *Desulfomonile tiedjei*,[56] *Desulfitobacterium, Dehalobacter restrictus, Dehalospirillum multivorans, Desulforomonas chloroethenica, Dehalococcoides ethenogenes,* and the facultative anaerobes *Enterobacter* strain MS-1 and *Enterobacter agglomeram*.[57] The rate, extent, and route of dechlorination is dependent on the composition of the active microbial community, which in turn are influenced by environmental factors such as availability of carbon sources, hydrogen or other electron donors, the presence or absence of electron acceptors other than PCBs, temperature, and pH.[58] However, a similarity between degradation patterns exists. The position of chlorine atoms on the rings affects the rate of biodegradation. Not only are PCBs with *para*- and meta-substituted rings more easily degraded than the *ortho*-substituted compounds, as shown in Figure 3,[59] but PCBs containing all chlorines on one ring are biodegraded faster than those that contain chlorines throughout both rings.

Persistence of PCBs in the environment increases with the degree of chlorination of the congener, i.e., compounds with a high degree of chlorination are resistant to biodegradation and degrade slowly in the environment. Anaerobic PCB dechlorination reduces the potential risk and potential exposure to PCBs because it significantly reduces the bioconcentration potential of the PCB mixture through conversion to

TABLE 3 Environmental Factors That Affect the Biodegradation of PCBs

Factor	How It Affects Biodegradation
Structure of the compound, i.e., the presence of substituents and their position in the molecule	• A high degree of halogenation requires high energy from the microorganisms to break the stable carbon-halogen bonds • Chlorine as the substituent alters the resonant properties of the aromatic substance as well as the electron density of specific sites; it may result in the deactivation of the primary oxidation of the compound by microorganisms • The positions occupied by substituted chlorines have stereochemical effects on the affinity between enzymes and their substrate molecule
Solubility of the compound	• Microorganisms easily access compounds with high aqueous solubility • Highly chlorinated congeners that are very insoluble in water are also very resistant to biodegradation
Concentration of the pollutant	• At a low concentration range, degradation increases linearly with increase in concentration until such time that the rate essentially becomes constant regardless of further increase in pollutant concentration • In general, a low pollutant concentration may be insufficient for the induction of degradative enzymes or to sustain growth of competent organisms • A very high concentration may render the compound toxic to the organisms
Temperature pH Presence of toxic or inhibitory substance and competing substances Availability of suitable electron acceptors Interactions among microorganisms	• The conditions should be optimal for the microorganism

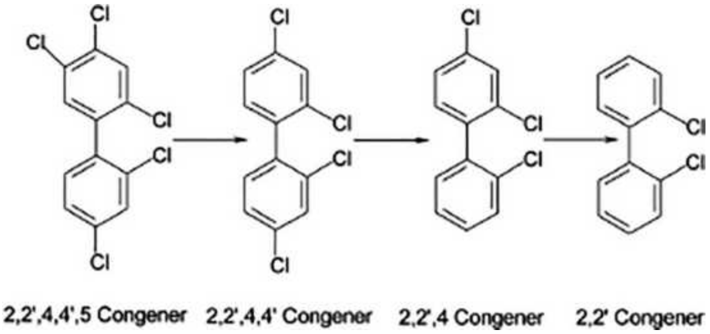


FIGURE 3 A potential pathway for anaerobic degradation of highly chlorinated PCB congeners to less chlorinated ones.

congeners that do not significantly bioaccumulate in the trophic chain.[60] Moreover, lightly chlorinated congeners produced by dechlorination can be readily degraded by indigenous bacteria.[61,62]

Aerobic biodegradation involving biphenyl ring cleavage is restricted to the lightly chlorinated PCB congeners, those with four or less chlorine atoms, resulting from the dechlorination of highly chlorinated congeners.[63,64] Aerobic oxidative destruction involves two clusters of genes. The first one is responsible for the transformation of PCB congeners to chlorobenzoic acid, and the second cluster is responsible for the degradation of the chlorobenzoic acid. A common growth substrate for

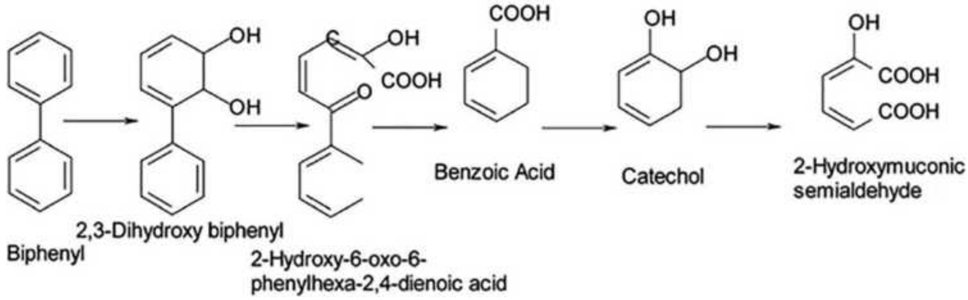


FIGURE 4 A possible pathway for the aerobic oxidative dehalogenation of PCBs.

PCB-degrading bacteria is biphenyl or monochlorobiphenyl. When biphenyl is utilized by bacteria, *meta-ring* cleavage product is produced. This has been observed in most bacteria studied especially in *Pseudomonas* sp.,[54] as well as in *Micrococcus* sp.[65] The metabolic pathway used by this family of bacteria is illustrated in Figure 4.[65]

Both anaerobic and aerobic metabolism modes transform PCBs. Different microorganisms show preferential attack on the PCB molecule, resulting in different patterns of biodegradation. The degree of chlorination of the congener and environmental factors influence the degradation potential of the compound.[49] Higher-chlorinated biphenyls therefore are potentially fully biodegradable in a sequence of anaerobic reductive dechlorination followed by aerobic mineralization of the lower-chlorinated products.[66]

Bioaccumulation, Bioconcentration, and Biomagnification

Organisms can accumulate high concentrations of PCBs relative to concentrations of these substances in non-biotic portions of the environment. This phenomenon is variously referred to as bioconcentration, bioaccumulation, and biomagnification.[67] Because some confusion exists in the literature about these definitions, we try to explain these more precisely following the terms set out by Gobas and Morrison.[68]

Bioaccumulation is a selective process that causes an increased chemical concentration in an organism compared to that in the surrounding medium and results from uptake by all exposure routes including transport across respiratory surfaces, dermal absorption (bioconcentration), and dietary absorption (biomagnification). Bioaccumulation can thus be viewed as a combination of bioconcentration and biomagnification.

The bioaccumulation factor (BAF) in fish is the ratio of the concentration of the chemical in the organism C_B to that in water, similar to that of the bioconcentration factor (BCF).

$$\mathrm{BAF} = C_B / C_{WT} \text{ or } C_B / C_{WD},$$

where BAF is the bioaccumulation factor and $C_B/C_{Wt(WD)}$ is the concentration of the chemical in the organism/in water.

The most common approach for evaluating levels of bioaccumulation is to compare the levels retained by the organism with levels in the contaminated medium in which they live.[69]

Bioconcentration results from uptake of chemicals from water (usually under laboratory conditions). Uptake occurs via the respiratory surface and/or skin and results in the chemical concentration in an organism being greater than that in the surrounding medium.

BCF is defined as the ratio of the chemical concentration in an organism, C_B, to the total chemical concentration in the water, C_{WT}, or to the freely dissolved chemical concentration in water, C_{WD} (it only takes into account the fraction of the chemical in the water that is biologically available for uptake). The BCF is expressed as follows:

$$\mathrm{BCF} = C_\mathrm{B}/C_\mathrm{WT} \text{ or } C_\mathrm{B}/C_\mathrm{WD},$$

where BCF is the bioconcentration factor and $C_\mathrm{B}/C_\mathrm{WT}(_\mathrm{WD})$ is the concentration of the chemical in the organism/in water.

Although sometimes applied to other aquatic species, the principal target organism for BCF assessment tends to be fish, primarily because of their importance as food for many species, including humans.

Biomagnification, on the other hand, is the bioaccumulation of a substance up the trophic chain when residues are transferred from consumption of smaller organisms by larger ones in the chain. It generally refers to the sequence of processes that produces higher concentrations in organisms at higher levels in the food chain (at higher trophic levels). These processes always results in an organism having higher concentrations of a substance than is present in the organism's food. Biomagnification also results in higher concentrations of the substance than would be expected if water were the only exposure mechanism.[70] A biomagnification factor (BMF) can be defined as the ratio of the concentration of chemical in the organism (C_B) to that in the organism's diet (C_A), and can be expressed as

$$\mathrm{BMF} = C_\mathrm{B}/C_\mathrm{A},$$

where BMF is the biomagnification factor, C_B is the concentration of chemical in the organism, and C_A is the concentration of chemical in the organism's diet.

This is the simplest definition of a BMF. It can also be described as the ratio of the observed lipid-normalized BCF to K_ow, which is the theoretical lipid-normalized BCF. This is equivalent to the multiplication factor above the equilibrium concentration. If this ratio is equal to or less than one, then the compound has not been biomagnified. If the ratio exceeds one, then the chemical is biomagnified by that factor.

The mechanism of biomagnification is not completely understood. Achieving a concentration of a chemical greater than its equilibrium value indicates that the elimination rate is slower than for chemicals that reach equilibrium. Transfer efficiencies of the chemical would affect the relative ratio of uptake and elimination. There are many factors that control the uptake and elimination of a chemical after contaminated food is consumed; these include factors specific to the chemical (solubility, K_ow, molecular weight and volume, and diffusion rates between organism gut, blood, and lipid pools), as well as factors specific to the organism (the feeding rate, diet preferences, assimilation rate into the gut, rate of chemical's metabolism, rate of egestion, and rate of organism growth). Because humans occupy a very high trophic level, we are particularly vulnerable to adverse health effects from exposure to chemicals that biomagnify.[71]

Chemicals that bioaccumulate do not necessarily biomagnify, although many papers report that PCB congeners do in fact biomagnify.[72,73] Some early bioaccumulation models used the concept of a food-chain multiplier, which is now considered excessively simplistic.[74] Exposure of PCBs solely from one source only occurs in laboratory experiments.[75] In nature, organisms are always exposed to different sources of contaminants, and therefore, what happens in the field is more complex than reflected in laboratory studies and cannot easily be emulated by laboratory studies. Mass balance models are simple tools that allow evaluation of various uptake and loss processes.[67] A variety of mass balance models have been developed to address water quality issues in lakes, estuaries, and slow-flowing water bodies.[76] The simpler models only consider advection and an overall loss due to the combined processes of volatilization, net transfer to sediment, and degradation. The rate constant for the overall loss is derived from fugacity calculations for a single segment system. The more rigorous models perform fugacity calculations for each segment and explicitly include the processes of advection, evaporation, water–sediment exchange, and degradation in both water and sediment. In this way, chemical exposure in all compartments (including equilibrium concentrations in biota) can be estimated.[77,78] In general, these models consider the organism to be a single "box."[67,79] These models require information about the chemicals, the organism, and associated environmental parameters.[79]

PCB congeners with less *ortho*-substitution are accumulated up the trophic chain at a greater rate than other congeners in their homolog group.[74] Non-*ortho*- substituted congeners, especially those that lack adjacent unsubstituted *meta* and *para* sites and unsubstituted *ortho* and *meta* sites, are undoubtedly metabolically recalcitrant in invertebrate and vertebrate tissues.[80] Changes in distributions of congeners are mainly caused by transfers among biotic compartments. There is no enrichment in higher trophic levels of mono- and non-*ortho*-substituted congeners. However, many coplanar congeners, especially very toxic PCB 77, are depleted with increasing trophic levels; PCB 77 is therefore almost certainly metabolized.[74]

Exposure of PCBs solely from one source only occurs in laboratory experiments. In nature, there are always multiple sources of contaminants, and therefore, field results must be studied carefully. Moreover, the properties of individual PCB congeners substantially affect accumulation or degradation pathways. Empirical models only reflect one of several possible mechanisms.[81]

Health Effects

As PCBs persist in the environment, the general population is potentially exposed to a variety of PCBs via food (especially fish caught in contaminated lakes or rivers, meat, and dairy products), air, surface soils, drinking water, and groundwater. In the workplace, people might be exposed to PCBs during repair and maintenance of PCB transformers and other old electrical devices, and disposal of PCB materials. Although mixtures used in industry are not identical to the combinations of PCBs present in the environment (or in breast milk), these mixtures have been found to have similar harmful effects.

The health effects of PCBs have been very widely studied in people (studies of industrial workers exposed to PCB-containing mixtures in the course of their work, as well as studies of adults and children exposed to PCBs as a result of consuming contaminated fish), laboratory animals, and wildlife in contaminated areas. These studies indicate that people who are regularly exposed to PCBs are at greater risk for a variety of health problems.[82] Moreover, evidence on the health effects of exposure to PCBs has been obtained from two episodes of mass poisoning that occurred in Japan (the 1968 Yusho incident) and Taiwan (the 1979 Yu-Cheng incident). Some of the most important findings are summarized below.

Research shows that PCBs cause a variety of adverse health effects depending on the route of exposure, age, sex, and area of the body where PCBs are concentrated. Studies on animals show conclusive evidence that PCBs are carcinogenic. Animals that ate food containing large amount of PCBs for short periods of time had mild liver damage and some died. PCBs have also been implicated as a cause of mass mortalities in seabirds.[49]

Moreover, a number of epidemiological studies of workers exposed to PCBs have been performed. The Department of Health and Human Services (DHHS) has concluded that PCBs may reasonably be anticipated to be carcinogens. Also, the Environmental Protection Agency (EPA) and the International Agency for Research on Cancer (IARC) have determined that PCBs are probably carcinogenic to humans.[83,84] Research also shows that exposure to PCBs in high concentration can have various acute effects including a skin disease known as chloracne (skin lesions), liver damage, other non-cancer short-term effects like body weight loss, impaired immune function, and clinically diagnosable damage to the central nervous system, causing headaches, dizziness, depression, nervousness, and fatigue. Other adverse health effects of PCBs are liver, stomach, and thyroid gland injuries; behavioral alterations; and impaired reproduction.[82]

The EPA has set a limit of 0.0005 milligrams of PCBs per liter of drinking water (0.0005 mg/L). Moreover, the Food and Drug Administration (FDA) requires that infant foods; eggs, milk, and other dairy products; fish and shellfish; poultry; and red meat contain no more than 0.2–3 parts of a PCB per million (0.2–3 ppm).[84]

Regulations

After the impact of PCBs on the environment was recognized, in 1976, the U.S. Congress charged the EPA with regulating the issue of PCBs. The ban on the manufacturing, processing, distribution in commerce, and use of PCBs, as well as the PCB disposal and marking regulations, was enclosed in the Toxic Substances Control Act (TSCA) of 1976.[85] In 1979, after subsequent amendments, the regulations stipulate that the production of PCBs in the United States is generally banned, the use of PCB-containing materials still in service is restricted, the discharge of PCB- containing effluents is prohibited, the disposal of materials contaminated by PCBs is regulated, and the import or export of PCBs is only permitted through an exemption granted from EPA.

In the European community, the use of PCBs in open applications such as printing inks and adhesives was banned in 1976 (Directive 76/403/EEC[86]). Use of PCBs as a raw material or chemical intermediate has been banned in the European Union (EU) since 1985.[87] In 1996, the 1976 directive was replaced by Directive 96/59/EC,[88] which set a deadline of 2010 for complete phase out or decontamination of equipment containing PCBs. However, the United Nations Environment Programme (UNEP) global treaty adopted at the Stockholm Convention on Persistent Organic Pollutants (May 2001) stipulates that the use in equipment shall be eliminated by 2025.[89] This date is a minimum requirement and does not prevent individual governments, or groups of governments, from maintaining earlier phase-out dates. However, the most important regulations—Council Directive 96/59/EC[88] and HELCOM Recommendation 6/1[90]—concerning total banning of PCBs have been fully implemented only by EU countries. Furthermore, the Commission has adopted community strategy for dioxins, furans, and PCBs aimed at reducing as far as possible the release of these substances in the environment and their introduction in the trophic chains.[91]

Despite the existing regulations, there is still a substantial amount of PCBs in use, because exemption has been given in many countries for contained use in existing equipment with long lifetimes, at least for an initial period after a production ban was decided. There are also quantities in storage awaiting disposal. The chemical industry is currently making a proposal to solve the disposal problem.[92]

Disposal of PCBs from the Environment

PCBs and PCB-contaminated equipment and oil are required to be properly disposed of in a manner similar to that of hazardous waste.

Much effort has been directed towards the selection of technology options for the disposal of PCBs from the environment. Although the baseline remediation technology for PCBs is incineration, other options do exist. Destruction of PCBs requires the breaking of molecular bonds by an input of thermal or chemical energy. The main features of combustion and non-combustion processes are summarized in Table 4.[93]

While destruction is to be preferred, some PCBs from a range of consumer goods are likely to enter landfills accepting municipal waste. PCBs deposited in landfills may therefore contaminate groundwater and surface water following migration into leachate. The behavior of PCBs in landfills is far from fully understood and a precautionary approach is recommended.

TABLE 4 Destruction Processes for PCB Wastes

Process		Waste Types Accepted	Advantages	Disadvantages
Incineration		• Liquids and dilute slurries • PCB-containing waste equipment (may require preprocessing)	• High destruction efficiencies (99.9999% or more) • Meeting legal requirements • Facilities can treat a range of wastes, both chlorinated and non-chlorinated	• Risk of emission of harmful substances if inadequately controlled • Careful process control is required to maintain important parameters (residence time, temperature, turbulence, and oxygen concentration) at the desired level and to ensure the effectiveness of the gas cleaning system • Costly, especially if wastes have to be shipped off-site • Some equipment may require preprocessing by mechanical alteration—e.g., shredding to expose contents of capacitors, draining and disassembly of transformers, cutting large transformers to size, or packing solids and sludges in drums and feeding via a chute
Dechlorination processes	Gas phase chemical reduction (GPCR)	• PCB-contaminated liquids, soils, sediments, equipment, and material	• High destruction efficiencies (99.9999%) • Modular, transportable, or fixed configurations • The expected throughput of the main reactor system is 1000 to 3000 tons/mo	• Need to establish treatment conditions for individual components
	Base catalyzed decomposition (BCD)	• – Liquids • – Soil and building rubble contaminated by POPs	• High destruction efficiencies (99.9999%) • Modular, transportable, or fixed plants • The process can tolerate inorganic and organic debris provided this material is smaller than 50 mm or can be shredded down to this size	• The first step of BC, designed to treat solid matrices, requires mechanical pretreatment • Need to establish treatment conditions for individual components

(*Continued*)

TABLE 4 (*Continued*) Destruction Processes for PCB Wastes

Process	Waste Types Accepted	Advantages	Disadvantages
Sodium reduction	Oils with a PCB content of up to 10,000 pg/dm^3	Transportable and fixed plants Widely used for in situ removal of PCBs from active transformers	Need to establish treatment conditions for individual components
Supercritical water oxidation (SCWO)	Liquid wastes or solids less than 200 microns in diameter, and an organic content of less than 20%	A compact, totally enclosed system All emissions and residues may be captured for assay and reprocessing if needed	Need to establish treatment conditions for individual components
Plasma arc	Liquid waste streams of any concentration (the most cost-effective method is to treat concentrated wastes) Solids in the form of a pumpable fine slurry	Various plasma reactors developed for the thermal destruction of hazardous waste Transportable and fixed units The system can treat its own fly ash plus filtration media, minimizing secondary wastes	Contaminated soil, very viscous liquids or sludges, other equipment (capacitors and transformers) can be treated after pretreatment
Pyrolysis	Solid, liquid and gaseous wastes	Transportable and fixed configurations Off gases can be reused as synthesis gas	Need to establish treatment conditions for individual components
Molten salt oxidation	Liquids Solids—only if reduced to small particle sizes for pneumatic conveying	The reaction takes place within the salt bath, virtually eliminating the fugitive inventories found in incineration	Need to establish treatment conditions for individual components
Solvated electron technology	Liquids Solid materials (up to 45 cm diameter)	POP wastes are reduced to metal salts and simple hydrocarbon compounds; PCBs are reduced to petroleum hydrocarbons, sodium chloride, and sodium amide	Material with a high water content (>40% w/w) must be dewatered prior to treatment

Conclusion

For decades, PCBs have been recognized as important and potentially harmful environmental contaminants. The intrinsic properties of PCBs, such as high environmental persistence, resistance to metabolism in organisms, and tendency to accumulate in lipids have contributed to their ubiquity in environmental media and have induced concern for their toxic effects after prolonged exposure.

PCBs are bioaccumulated mainly by aquatic and terrestrial organisms and thus enter the food web. Humans and wildlife that consume contaminated organisms can also accumulate PCBs in their tissues. Such accumulation is of concern, because it may lead to body burdens of PCBs that could have adverse health effects in humans and wildlife.

Moreover, PCBs are slower to biodegrade in the environment than are many other organic chemicals. The low water solubility and the low vapor pressure of PCBs, coupled with air, water, and sediment transport processes, mean that they are readily transported from local or regional sites of contamination to remote areas.

PCBs are transformed mainly through microbial degradation, and particularly reductive dechlorination via organisms that take them up. Metabolism by microorganisms and other animals can cause relative proportions of some congeners to increase while others decrease. Because the susceptibility of PCBs to degradation and bioaccumulation is congener specific, the composition of PCB congener mixtures that occur in the environment differs substantially from that of the original industrial mixtures released into the environment. Generally, the less-chlorinated congeners are more water soluble, more volatile, and more likely to biodegrade. On the other hand, higher-chlorinated PCBs are often more resistant to degradation and volatilization and sorb more strongly to particulate matter. Some higher-chlorinated PCBs tend to bioaccumulate to greater concentrations in tissues of animals than do lower- molecular-weight ones. The higher-chlorinated PCBs can also biomagnify in food webs.

There is still much to be learned about the chemistry of PCBs. Current research focuses on finding ways to break the molecules down into harmless compounds. The biodegradation of PCBs utilizing microorganisms presently appears to be our best hope for removing PCBs from the environment.

Glossary

BAF: Bioaccumulation factor
BCF: Bioconcentration factor
BMF: Biomagnification factor
BZ number: A system of sequential numbers for the 209 PCB congeners introduced in 1980 by Ballschmiter and Zell that identifies a given congener simply and precisely. Also referred to as congener, IUPAC, or PCB number.
DHHS: Department of Health and Human Services
DOM: Dissolved organic matter
EPA: Environmental Protection Agency
FDA: Food and Drug Administration
IARC: International Agency for Research on Cancer
PCBs: Polychlorinated biphenyls
POM: Particulate organic matter
POPs: Persistent organic pollutants
TSCA: Toxic Substances Control Act
UNEP: United Nations Environment Programme

Acknowledgments

This research was financially supported by the Polish Ministry of Science and Higher Education (grant no. N N312 300535). A. Wilkowska is grateful for financial support from the Human Capital Programme (POKL.04.01.01-00-368/09).

References

1. US Department of Health and Human Services. Toxicological profile for polychlorinated biphenyls (update). Agency for Toxic Substances and Disease Registry: Atlanta, 2000, available at http://www.atsdr.cdc.gov/toxprofiles/tp17.pdf (accessed January 2010).

2. De Voogt, P.; Brinkman, U.A. Production, properties and usage of polychlorinated biphenyls. In *Halogenated Biphenyls, Terphenyls, Naphthalenes, Dibenzodioxins and Related Products,* 2nd Ed.; Kimbrough, R.D., Jensen, *A.A.*, Eds.; Elsevier Science Publishers: Amsterdam, 1989; 3–45.
3. Ballschmiter, K.; Zell, M. Analysis of polychlorinated biphenyls (PCB) by glass capillary gas chromatography: Composition of technical Aroclor and Clophen-PCB mixtures. Fresenius Z. Anal. Chem. **1980**, *302,* 20–31.
4. Dunnivant, F.M.; Elzerman, A.W. Aqueous solubility and Henry's law constant data for PCB congeners for evaluation of quantitative structure–property relationships (QSPRs). Chemosphere **1988**, *17,* 525–541.
5. Hutzinger, O.; Safe, S.; Zitko, V. Photochemical degradation of chlorobiphenyls (PCBs). Environ. Health Perspect. **1972**, *1*, 15–20.
6. Beyer, A.; Biziuk, M. Environmental fate and global distribution of polychlorinated biphenyls. Rev. Environ. Contam. Toxicol. **2009**, *201,* 137–158.
7. Hazard waste generation and commercial hazardous waste management capacity: An assessment, SW-894. Prepared by Booz-Allen and Hamilton, Inc. and Putnam, Hayes and Barlett, Inc. for the Office of Planning and Evaluation and the Office of Solid Waste. U.S. Environmental Protection Agency: Washington, DC, 1980; D-4.
8. Barbalace, R.C. The Chemistry of Polychlorinated Biphenyls. EnvironmentalChemistry.com. 2003, available at http://EnvironmentalChemistry.com/yogi/chemistry/pcb.html (accessed January 2010).
9. Guidelines for the Identification of PCBs and Materials Containing PCBs. 1999, available at http://www.chem.unep.ch/pops/pdf/PCBident/pcbid1.pdf (accessed January 2010).
10. Durfee, R.L. Production and usage of PCB's in the United States. In Proceedings of the National Conference on Polychlorinated Biphenyls, EPA-560/6–75–004.U.S. Environmental Protection Agency: Washington, DC, 1976; 103–107.
11. Orris, P.; Kominsky, J.R.; Hryhorczyk, D.; Melius, J. Exposure to polychlorinated biphenyls from an overheated transformer. Chemosphere **1986**, *15,* 1305–1311.
12. Welsh, M.S. Extraction and gas chromatography/electron capture analysis of polychlorinated biphenyls in railcar paint scrapings. Appl. Occup. Environ. Hyg. **1995**, *10* (3), 175–181.
13. Boate, A.; Deleersnyder, G.; Howarth, J.; Mirabelli, A.; Peck L. Chemistry of PCBs, available at http://wvlc.uwaterloo.ca/biology447/modules/intro/assignments/Introduction2a.htm (accessed January 2010).
14. Eisenreich, S.J.; Baker, J.E.; Franz, T.; Swanson, M.; Rapaport, R.A.; Strachan, W.M.J.; Hites, R.A. Atmospheric deposition of hydrophobic organic contaminants to the Laurentian Great Lakes. In *Fate of Pesticides and Chemicals in the Environment;* Schnoor, J.L., Ed.; John Wiley and Sons, Inc.: New York, 1992; 51–78.
15. Pham, T.T.; Proulx, S. PCBs and PAHs in the Montreal urban community (Quebec, Canada) wastewater treatment plant and in the effluent plume in the St. Lawrence River. Water Res. **1997**, *31*(8), 1887–1896.
16. Hansen, L.G.; O'Keefe, P.W. Polychlorinated dibenzofurans and dibenzo-*p*-dioxins in subsurface soil, superficial dust, and air extracts from a contaminated landfill. Arch. Environ. Contam. Toxicol. **1996**, *31*(2), 271–276.
17. Hansen, L.G.; Green, D.; Cochran, J.; Vermette, S.; Bush, B. Chlorobiphenyl (PCB) composition of extracts of subsurface soil, superficial dust and air from a contaminated landfill. Fresenius J. Anal. Chem. **1997**, *357,* 442–448.
18. Swackhamer, D.L.; Armstrong, D.E. Estimation of the atmospheric and nonatmospheric contributions and losses of polychlorinated biphenyls for Lake Michigan on the basis of sediment records of remote lakes. Environ. Sci. Technol. **1986**, *20,* 879–883.

19. Wallace, J.C.; Basu, I.; Hites, R.A. Sampling and analysis artifacts caused by elevated indoor air polychlorinated biphenyl concentrations. Environ. Sci. Technol. **1996**, *30*(9), 2730–2734.
20. Ohsaki, Y.; Matsueda, T. Levels, features and a source of non-ortho coplanar polychlorinated biphenyl in soil. Chemosphere **1994**, *28*(1), 47–56.
21. Blumbach, J.; Nethe, L.P. Organic components reduction (PCDD/PCDF/PCB) in flue-gases and residual materials from waste incinerators by use of carbonaceous adsorbents. Chemosphere **1996**, *32*(1), 119–131.
22. Gunkel, G.; Mast, P.G.; Nolte, C. Pollution of aquatic ecosystems by polychlorinated biphenyls (PCB). Limnologica **1995**, *25*(3/4), 321–331.
23. Alcock, R.E.; Bacon, J.; Bardget, R.D.; Beck, A.J.; Hay-garth, P.M.; Lee, R.G.M.; Parker, C.A.; Jones, K.C. Persistence and fate of polychlorinated biphenyls (PCBs) in sewage sludge-amended agricultural soils. Environ. Pollut. **1996**, *93* (1), 83–92.
24. Wania, F.; Mackay, D. Tracking the distribution of persistent organic pollutants. Environ. Sci. Technol. **1996**, *30,* 390A-396A.
25. Hansen, K.M.; Prevedouros, K.; Sweetman, A.J.; Jones, K. C.; Christensen, J.H. A process-oriented intercomparison of a box model and an atmospheric chemistry transport model: Insights into model structure using alpha-HCH as the modelled substance. Atmos. Environ. **2006**, *40,* 2089–2104.
26. Mackay, D. *Multimedia Environmental Models. The Fugacity Approach*, 2nd Ed.; Lewis Publishers: Boca Raton, FL, 2001.
27. Gong, S.L.; Huang, P.; Zhao, T.L.; Sahsuvar, L.; Barrie, L. A.; Kaminski, J.W.; Li, Y.F.; Niu, T. GEM/POPs: A global 3-D dynamic model for semi-volatile persistent organic pollutants—Part 1: Model description and evaluations of air concentrations. Atmos. Chem. Phys. **2007**, *7,* 4001–4013.
28. Huang, P.; Gong, S.L.; Zhao, T.L.; Neary, L.; Barrie, L.A. GEM/POPs: A global 3-D dynamic model for semi-volatile persistent organic pollutants—Part 2: Global transports and budgets of PCBs. Atmos. Chem. Phys. **2007**, *7,* 4015–4025.
29. Wania, F.; Daly, G.L. Estimating the contribution of degradation in air and deposition to the deep sea to the global loss of PCBs. Atmos. Environ. **2002**, *36,* 5581–5593.
30. Wania, F.; Su, Y.S. Quantifying the global fractionation of polychlorinated biphenyls. Ambio **2004**, *33*, 161–168.
31. Eckhardt, S.; Breivik, K.; Li, Y.F.; Mano, S.; Stohl, A. Source regions of some persistent organic pollutants measured in the atmosphere at Birkenes, Norway. Atmos. Chem. Phys. **2009**, *9*, 6597–6610.
32. Li, N.Q.; Wania, F.; Lei, Y.D.; Daly, G.L. A comprehensive and critical compilation, evaluation, and selection of physical chemical property data for selected polychlorinated biphenyls. J. Phys. Chem. Ref. Data **2003**, *32,* 1545–1590.
33. Erickson, M.D. *Analytical Chemistry of PCBs*; Butterworth Publishers: Stoneham, Massachusettes, 1986.
34. Hansen, L.G. *The Ortho Side of PCBs: Occurrence and Disposition;* Kluwer Academic Publishers: Boston, 1999.
35. Haque, R.; Schmedding, D.; Freed, V. Aqueous solubility, adsorption and vapor behavior of polychlorinated biphenyl Aroclor 1254. Environ. Sci. Technol. **1974**, *8,* 139–142.
36. Chiarenzelli, J.; Scrudato, R.; Arnold, G.; Wunderlich, M.; Rafferty, D. Volatilization of polychlorinated biphenyls during drying at ambient conditions. Chemosphere **1996**, *33,* 899–911.
37. Wania, F.; Mackay, D. Global fractionation and cold condensation of low volatility organochlorine compounds in polar regions. Ambio **1993**, *22*, 10–18.
38. Wania, F. Potential of degradable organic chemicals for absolute and relative enrichment in the Arctic. Environ. Sci. Technol. **2006**, *40,* 569–577.
39. Nelson, E.D.; McConnell, L.L.; Baker, J.E. Diffusive exchange of gaseous polycyclic aromatic hydrocarbons and polychlorinated biphenyls across the air-water interface of the Chesapeake Bay. Environ. Sci. Technol. **1998**, *32* (7), 912–919.

40. Luthy, R.G.; Aiken, G.R.; Brusseau, M.L.; Cunningham, S.D.; Gschwend, P.M.; Pignatello, J.J.; Reinhard, M.; Traina, S.J.; Weber, W.J.; Westall, J.C. Sequestration of hydrophobic organic contaminants by geosorbents. Environ. Sci. Technol. **1997**, *31* (12), 3341–3347.
41. ATSDR—Agency for Toxic Substances and Disease Registry. Toxicological profile for polychlorinated biphenyls (update). Atlanta: US Department of Health and Human Services, 2000.
42. Van Bavel, B.; Andersson, P.; Wingfors, H.; Ahgren, J.; Bergqvist, P.A.; Norrgren, L.; Rappe, C.; Tysklind, M. Multivariate modeling of PCB bioaccumulation in three-spined stickleback (*Gasterosteus aculeatus*). Environ. Toxicol. Chem. **1996**, *15,* 947–954.
43. Boese, B.L.; Winsor, M.; Lee, II, H.; Echols, S.; Pelletier, J.; Randall, R. PCB Congeners and hexachlorobenzene biota sediment accumulation factor for *Macoma nasuta* exposed to sediments with different total organic carbon contents. Environ. Toxicol. Chem. **1995**, *14,* 303–310.
44. Friedman, C.L.; Burgess, R.M.; Perron, M.M.; Cantwell, M.G.; Ho, K.T.; Lohmann R. Comparing polychaete and polyethylene uptake to assess sediment resuspension effects on PCB bioavailability. Environ. Sci. Technol. **2009**, *43*(8), 2865–2870.
45. Trimble, T.A.; You, J.; Lydy, M.J. Bioavailability of PCBs from field-collected sediments: Application of Tenax extraction and matrix-SPME techniques. Chemosphere **2008**, *71*, 337–344.
46. Verweij, F.; Booij, K.; Satumalay, K.; van der Molen, N.; van der Oost, R. Assessment of bioavailable PAH, PCB and OCP concentrations in water, using semipermeable membrane devices (SPMDs), sediments and caged carp. Chemosphere **2004**, *54* (11), 1675–1689.
47. Gdaniec-Pietryka, M.; Wolska, L.; Namiesnik, J. Physical speciation of polychlorinated biphenyls in the aquatic environment. Trends Anal. Chem. **2007**, *26,* 1005–1012.
48. Adsorption and Desorption of Organic Contaminants to Predict Fate, Transport, and Propensity to Electrochemically Degrade, available at http://www.epa.gov/nrmrl/lrpcd/wm/projects/135925.htm (accessed February 2010).
49. Borja, J.; Taleon, D.M.; Auresenia, J.; Gallardo, S. Polychlorinated biphenyls and their biodegradation. Process Biochem. **2005**, *40,* 1999–2013.
50. Dobbins, D.C. Biodegradation of pollutants. In *Encyclopedia of Environmental Biology*; Academic Press Inc.: New York, 1995.
51. McEldowney, S.; Hardman, D.J.; Wait, S. *Pollution: Ecology and Biotreatment*; Longman Scientific and Technical: New York, 1993.
52. Furukawa, K. Modification of PCBs by bacteria and other microorganisms. In *PCBs and the Environment;* Waid, J.S., Ed.; CRC Press: Florida, 1986; 89–100.
53. Sylvestre, M.; Sandossi, M. Selection of enhanced PCB- degrading bacterial strains for bioremediation: Consideration of branching pathways. In *Biological Degradation and Remediation of Toxic Chemicals;* Chaudhry, G.R., Ed.; Chapman and Hall: New York, 1994.
54. Boyle, A.W.; Silvin, C.J.; Hassett, J.P.; Nakas, J.P.; Tanenbaum, S.W. Bacterial PCB biodegradation. Biodegradation **1992**, *3*, 285–298.
55. Alder, A.C.; Haggblom, M.M.; Oppenheimer, S.R.; Young, L.Y. Reductive dechlorination of polychlorinated biphenyls in anaerobic sediments. Environ. Sci. Technol. **1993**, *27,* 530–538.
56. Mohn, W.W.; Tiedje, J.M. Microbial reductive dechlorination. Microbiol. Rev. **1992**, *56,* 482–507.
57. Holliger, C.; Wohlfarth, G.; Diekert, G. Reductive dechlorination in the energy metabolism of anaerobic bacteria. FEMS Microbiol. Rev. **1998**, *22*, 383–398.
58. Wiegel, J.; Wu, Q. Microbial reductive dehalogenation of polychlorinated biphenyls. FEMS Microbiol. Ecol. **2000**, *32,* 1–15.
59. Fish, K.M.; Principe, J.M. Biotransformations of Arochlor 1242 in Hudson River test tube microcosms. Appl. Environ. Microbiol. **1994**, *60*, 4289–4296.
60. Ye, D.; Quensen, III, J.F.; Tiedje, J.M.; Boyd, S.A. Anaerobic dechlorination of polychlorinated biphenyls (Arochlor 1242) by pasteurized and ethanol-treated microorganisms from sediments. Appl. Environ. Microbiol. **1992**, *58* (4), 1110–1114.

61. Moore, J.A. *Reassessment of Liver Findings in PCB Studies for Rats*; Institute of Evaluating Health Risks: Washington, 1991.
62. Safe, S. Toxicology, structure-function relationship, and human and environmental health impacts of polychlorinated biphenyls: progress and problems. Environ. Health. Perspect. **1992**, *100,* 259–268.
63. Cookson, Jr., J.T. *Bioremediation Engineering: Design and Application;* McGraw Hill: New York, 1995.
64. Kuipers, B.; Cullen, W.R.; Mohn, W.W. Reductive dechlorination of nonachloro biphenyls and selected octachloro biphenyls by microbial enrichment cultures. Environ. Sci. Technol. **1999**, *33*, 3579–3585.
65. Benvinakatti, B.G.; Ninnekar, H.Z. Degradation of biphenyl by a *Micrococcus* species. Appl. Microbiol. Biotechnol. **1992**, *38,* 273–275.
66. Field, J.A.; Sierra-Alvarez, R. Microbial transformation and degradation of polychlorinated biphenyls. Environ. Pollut. **2008**, *155,* 1–12.
67. Mackay, D.; Fraser, A. Bioaccumulation of persistent organic chemicals: Mechanism and models. Environ. Pollut. **2000**, *110,* 375–391.
68. Gobas, F.A.P.C.; Morrison, H.A. Bioconcentration and biomagnification in the aquatic environment. In *Handbook of Property Estimation Methods for Chemicals;* Boethling, R.S., Mackay, D., Eds.; CRC Press: Boca Raton, FL, 2000; 189–231.
69. Kucklick, J.; Harvey, H.R.; Ostrom, P.; Ostrom, N.; Baker, J. Organochlorine dynamics in the pelagic food web of Lake Baikal. Environ. Toxicol. Chem. **1996**, *15,* 1388–1400.
70. Glossary for Chemists of Terms Used in Toxicology: Pure and Applied Chemistry, V. 65, no. 9; International Union of Pure And Applied Chemistry (IUPAC): Bethesda, Maryland, 1993; 2003–2122, available at (http://sis.nlm.nih.gov/enviro/glossarymain.html—online version posted by the U.S. National Library of Medicine).
71. Bierman, Jr., V.J. Equilibrium partitioning and biomagnification of organic chemicals in benthic animals. Environ. Sci. Technol. **1990**, *24,* 1407–1412.
72. Burreau, S.; Zebuhr, Y.; Broman, D.; Ishaq, R. Biomagnification of PBDEs and PCBs in food webs from the Baltic Sea and the northern Atlantic Ocean. Sci. Total Environ. **2006**, *366,* 659–672.
73. Nfon, E.; Cousins, I.T.; Broman, D. Biomagnification of organic pollutants in benthic and pelagic marine food chains from the Baltic Sea. Sci. Total Environ. **2008**, *397,* 190–204.
74. Campfens, J.; Mackay, D. Fugacity-based model of PCB bioaccumulation in complex aquatic food webs. Environ. Sci. Technol. **1997**, *31,* 577–583.
75. Pelka, A. Bioaccumulation models and applications: Setting sediment cleanup goals in the Great Lakes. National Sediment Bioaccumulation Conference Proceedings, EPA 823-R-98–002U.S.; Environmental Protection Agency Office of Water: Washington, DC, 1998; 5–9-5–30.
76. Chapra, S.C.; Reckhow, K.H. *Engineering Approaches for Lake Management. Mechanistic Modeling*; Butterworth Publishers/Ann Arbor Science: Woburn, MA, 1983; Vol. 2.
77. Warren, C.; Mackay, D.; Whelan, M.; Fox, K. Mass balance modelling of contaminants in river basins: A flexible matrix approach. Chemosphere **2005**, *61,* 1458–1467.
78. Warren, C.; Mackay, D.; Whelan, M.; Fox, K. Mass balance modelling of contaminants in river basins: Application of the flexible matrix approach. Chemosphere **2007**, *68,* 1232–1244.
79. Arnot, J.A.; Gobas, F. A food web bioaccumulation model for organic chemicals in aquatic ecosystems. Environ. Toxicol. Chem. **2004**, *23,* 2343–2355.
80. Bright, D.A.; Grundy, S.L.; Reimer, K.J. Differential bioaccumulation of non-ortho-substituted and other PCB congeners in coastal arctic invertebrates and fish. Environ. Sci. Technol. **1995**, *29*, 2504–2512.
81. Antunes, P.; Gil, O.; Reis-Henriques, M.A. Evidence for higher biomagnification factors of lower chlorinated PCBs in cultivated seabass. Sci. Total Environ. **2007**, *377,* 36–44.

82. Available at http://www.atsdr.cdc.gov/DT/pcb007.html (accessed February 2010).
83. U.S. Environmental protection agency. PCBs: A cancer dose-response assessment and applications to environmental mixtures, EPA/600/P-96/001F; 1996, available at http://cfpub.epa.gov/ncea/CFM/recordisplay.cfm?deid=12486 (accessed February 2010).
84. Available at http://www.atsdr.cdc.gov/tfacts17.pdf (accessed February 2010).
85. Online version of TSCA, from the Government Printing Office, available at http://frwebgate.access.gpo.gov/cgi-bin/usc.cgi?ACTION=BROWSE&TITLE=15USCC53 (accessed January 2010).
86. Council Directive 76/403/EEC of 6 April 1976 on the disposal of polychlorinated biphenyls and polychlorinated terphenyls.
87. Council Directive 85/467/EEC of 1 October 1985 amending for the sixth time Directive 76/769/EEC on the approximation of the laws, regulations and administrative provisions of the Member States relating to restrictions on the marketing and use of certain dangerous substances and preparations (PCBs/PCTs).
88. Council Directive 96/59/EC of 16 September 1996 on the disposal of polychlorinated biphenyls and polychlorinated terphenyls (PCBs/PCTs).
89. Manila Bulletin. Global chemical treaty (opinion/editorial). Manila Bulletin Publishing Corp.: Manila, 2001.
90. Helcom Recommendation 6/1, Recommendation regarding the elimination of use of PCBs and PCTs—adopted 13 March 1995, having regard to article 13, Paragraph b of the Helsinki Convention.
91. Communication from the Commission to the Council, the European Parliament and the Economic and Social Committee. Community strategy for dioxins, furans and polychlorinated biphenyls (2001/C 322/02), available at http://ec.europa.eu/environment/waste/pcbs/pdf/en.pdf (accessed January 2010).
92. ICCA/WCC Position Paper (1999): Best Available Techniques for Destruction of PCBs-Incineration Technology, presented at the UNEP-INC 3 meeting on POPs in Geneva, September 6–10, 1999.
93. Inventory of World-wide PCB Destruction Capacity, Second Issue; United Nations Environment Programme: Switzerland, 2004; Prepared by UNEP Chemicals, available at http://www.chem.unep.ch/pops/pcb_activities/pcb_dest/PCB_Dest_Cap_SHORT.pdf (accessed January 2010).

13

Toxic Substances

Sven Erik Jørgensen

Introduction

This entry gives a short overview of the toxic contaminants that we most often detect in the environment and consider the most threatening compounds. As we are using about 100,000 different chemicals in modern society, we can expect to find 100,000 different compounds plus intermediates of these compounds resulting from a wide spectrum of decomposition processes that take place in the environment. To be able to give an overview of many chemical compounds, their fates, effects, and associated risks, it is necessary to make a classification of the compounds. Below is a brief treatment of the properties, effects, characteristic processes, and particular risks of the following classes of contaminants that represent the most toxic compounds that we are using and the most harmful compounds for the environment.

1. Petroleum hydrocarbons
2. PCBs and dioxins
3. Pesticides
4. Polycyclic aromatic hydrocarbons (PAHs)
5. Heavy metals
6. Detergents
7. Synthetic polymers and xenobiotics applied in the plastics industry

Only the organometallic compounds are considered in this overview, while several other entries cover contamination by heavy metals in general. The seven classes are presented in the succeeding sections.

For more comprehensive coverage of toxic chemical compounds in the environment, refer to Loganathan and Lam,[1] Newman and Unger,[2] Hoffman et al.,[3] and Schuurmann and Markert.[4]

Petroleum Hydrocarbons

Petroleum hydrocarbons include a variety of organic compounds. Hydrocarbons (compounds composed of carbon and hydrogen) constitute only 50%–90% of petroleum. They are *n*-alkanes, branched alkanes, cycloalkanes, and aromatics. Cycloalkanes comprise usually the largest portion of hydrocarbons in petroleum, while aromatics are usually present to the extent of 20% or less. Characteristic compounds are benzene, alkyl-substituted benzenes, and fused-ring PAHs (see also group D). Some of these aromatics are carcinogenic, and this is probably the group that is of greatest environmental concern. In addition to hydrocarbons, petroleum contains sulfur and nitrogen compounds, such as thiophene, ethanethiol, and pyridine derivatives.

Petroleum compounds are emitted or discharged into all spheres.

Evaporation removes the lower molecular weight of the more volatile components of the petroleum mixture. Hydrocarbons with vapor pressures equal to that of *n*-octane (0.019 atm at room temperature) or greater will be lost quickly via evaporation.

The lower-molecular-weight hydrocarbons tend also to be the most water soluble, but for the same molecular weight, aromatic hydrocarbons are more soluble than cycloalkanes, which are more soluble than branched alkanes, with the *n*-alkanes being the most insoluble in water. Petroleum products are discharged directly to the sea by accidents or by violation of international regulations. In a massive discharge of petroleum products, which has been recorded in several ship accidents, most of the petroleum will initially float on the surface of marine waters as a slick. Eventually, most slicks are dispersed widely and form a 0.1 mm thick layer on the water. If drift results in landfall, or contact with coral or mangrove communities, a disastrous environmental impact may occur.

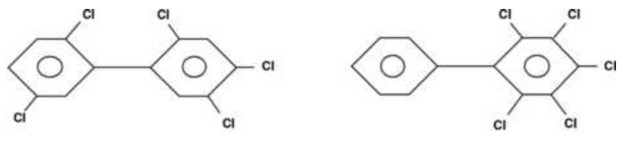
Fortunately, chemical transformation and degradation processes act on petroleum compounds in the environment. Microbial transformation and photo oxidation are of particular importance.

The aromatic compounds of petroleum are the most toxic substances present. They are lethal to crustaceans and fish with LC_{50} in the range of 0.1–10 mg/L.

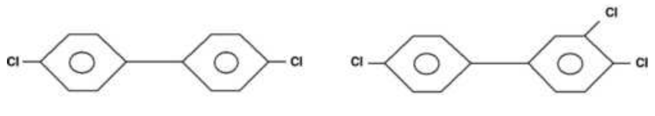
PCBs and Dioxins

PCBs and dioxins are characterized as aromatic compounds with a high content of chlorine. As the names cover polychlorinated biphenyls and polychlorinated dibenzo(1,4)dioxins, there are many different individual compounds under these labels. For instance, 209 different PCB compounds are known, although only about 130 are found in commercial mixtures. Figures 1 and 2 show the molecular structures and names of some of the most common PCBs and dioxins.

The applications of PCBs have been quite diverse (capacitor oil, plasticizers, printer's ink, etc.), but due to investigations in the 1960s and 1970s, in which it was found that PCBs occurred widely in the environment and significant bioaccumulation took place, voluntary restrictions were introduced and all "open" applications were banned.

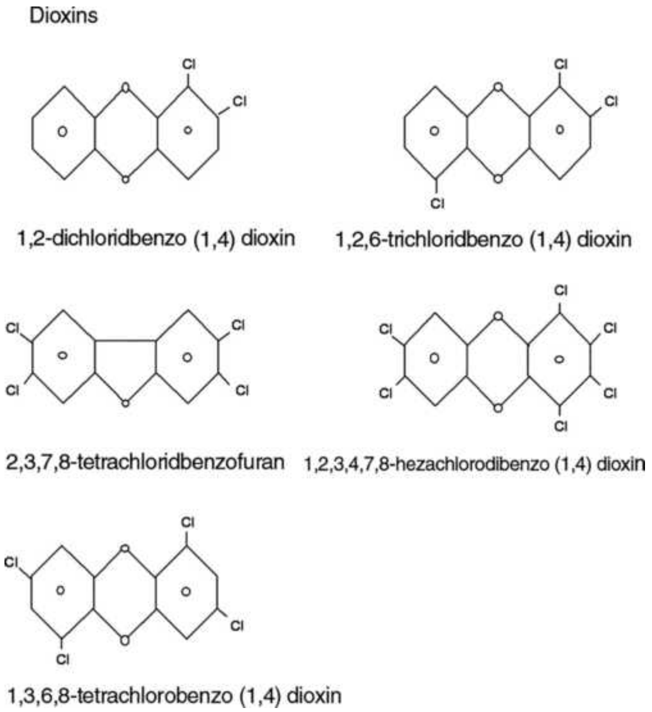


2,2',4,5,5'-pentachlorobiphenyl 2,3,4,5,6-pentachlorobiphenyl



4,4'-dichlorobiphenyl 3,4,4'-trichlorobiphenyl

FIGURE 1 The molecular structure and names of four common PCBs.



1,2-dichloridbenzo (1,4) dioxin

1,2,6-trichloridbenzo (1,4) dioxin

2,3,7,8-tetrachloridbenzofuran

1,2,3,4,7,8-hezachlorodibenzo (1,4) dioxin

1,3,6,8-tetrachlorobenzo (1,4) dioxin

FIGURE 2 Molecular structure and names of five common dioxins.

Dioxins are not deliberately produced, but are by-products of chemical processes involving chlorine, for instance, production of various organochlorine and bleaching of pulp and of combustion processes if chlorine-containing compounds are present.

Both PCBs and dioxins are characterized by a low water solubility and a high K_{ow} (most components have log $K_{ow} > 5$). Both groups of compounds are very persistent to decomposition processes, which explains why they are strong bioaccumulators, although dioxins have a UV–VIS absorption spectrum that results in significant absorption from solar radiation. Some dioxins have a half-life time in the troposphere of a few days.

Pesticides

Pesticides are used to remove pests and, due to their direct use in nature, have probably been the most criticized environmental contaminants. Usage of dichlorodiphenyltrichloroethane (DDT) and relatedinsecticides accelerated during the 1940s and the subsequent decades until environmental doubt occurred in the mid-1960s. Since 1970, DDT has been banned in most industrialized countries, but it is still used in developing countries, for instance, India, where it has resulted in very high body concentrations in the Indian population. All chlorinated hydrocarbon insecticides are banned in most industrialized countries due to their persistence and ability to bioaccumulate (K_{ow} is high).

Pesticides can be divided into the following classes depending on their use and their chemical structure:

- Herbicides comprise carbamates, phenoxyacetic acids, triazines, and phenylureas.
- Insecticides encompass organophosphates, carbamates, organochlorines, pyrethrins, and pyrethroids.
- Fungicides are dithiocarbamates, copper, and mercury compounds. See also the entries covering these two metals.

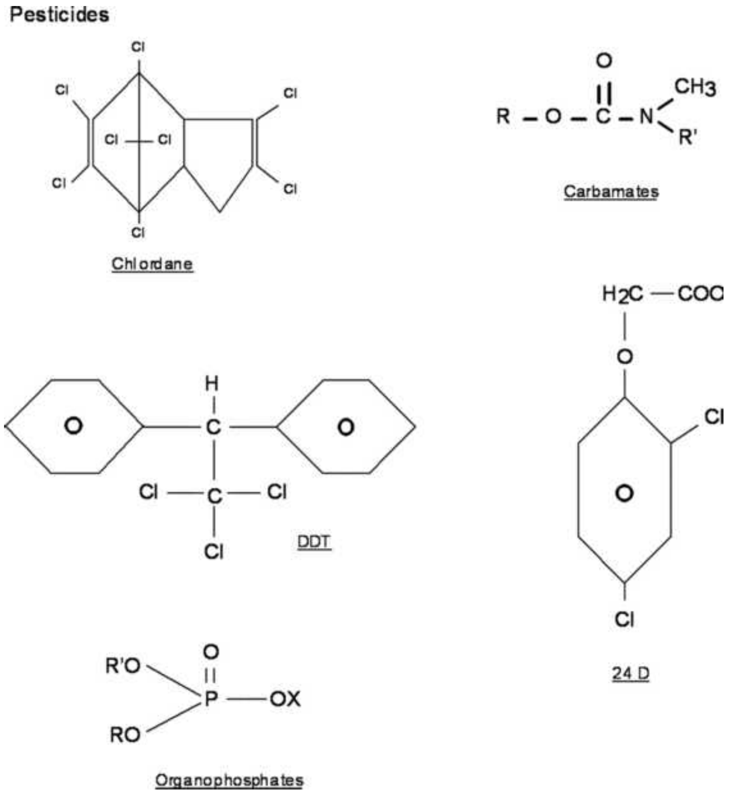


FIGURE 3 Molecular structure and names of five common pesticides.

The pesticides are an extremely chemically diverse group of substances, as they only have in common their toxicity to pests. A few of the most important molecules are shown in Figure 3. They are mostly produced synthetically although the natural pesticide pyrethrin has achieved commercial success.

Chlorohydrocarbons are strongly bioconcentrated, as already emphasized. In addition, they are very toxic to a wide range of biota, particularly to aquatic biota.

Organophosphates are almost equally toxic to biota, but due to these compounds' lack of persistence, higher solubility in water, and bioaccumulation capacity, they are still in use.

Carbamates are relatively water soluble and have limited persistence. They are, however, toxic to a wide range of biota. They act by inhibiting cholinesterase.

The pyrethins have a complex chemical structure and a high molecular weight. Thus, they are poorly soluble in water and tend to be lipophilic. They are, however, readily degraded by hydrolysis. They are more attractive to use than most of the other pesticides due to their very low mammalian toxicity.

Phenoxyaceticacidisaveryeffectiveherbicidebutcontainstraceamountsoftetrachloro-dibenzo-dioxin.

Pesticides are banned in organic agriculture where they are replaced by other methods, for instance, mechanical and biological methods (use of predator insects).

Polycyclic Aromatic Hydrocarbons

PAHs are molecules containing two or more fused 6C aromatic rings. They are ubiquitous contaminants of the natural environment, but growing industrialization has increased environmental concern about these components. Two common members are naphthalene and benzo(*a*)pyrene (see Figure 4). PAHs are usually solids, with naphthalene (lowest molecular weight) having a melting point of 81°C.

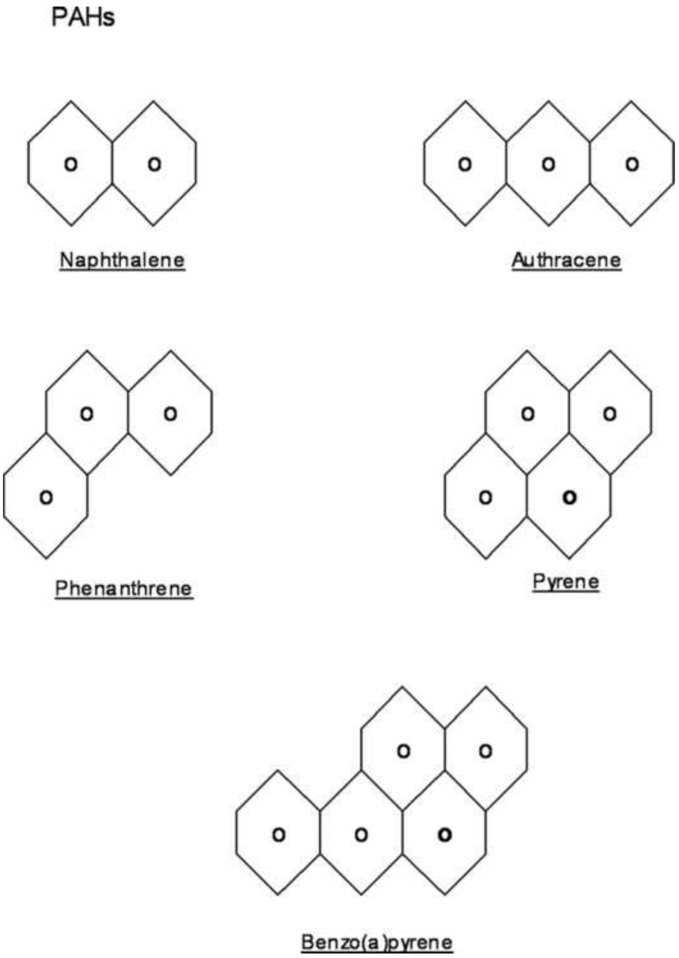


FIGURE 4 Molecular structure and names of five common PAHs.

The natural sources of PAHs in the environment are forest fires and volcanic activity. The anthropogenic sources are coal-fired power plants, incinerators, open burning and motor vehicle exhaust. As a result of these sources, PAHs commonly occur in air, soil, and biota. They are lipophilic compounds able to bioaccumulate. The low-molecular- weight compounds are moderately persistent, while for example benzo(*a*)pyrene, with a higher molecular weight, persists in aquatic systems for up to about 300 weeks.

They are relatively toxic to aquatic organisms and have LC_{50} values for fish in the range of 0.1–10 mg/L. The major environmental concern of PAHs is that many PAHs are carcinogenic. It is has been shown[5] that benzo(*a*)pyrene is an endocrine disrupter and that many more PAHs have the environmental adverse effect of disturbing the hormone balance of nature.

Human exposure to PAHs occurs through tobacco smoking as well as through compounds in food and the atmosphere.

Organometallic Compounds

Organometallic compounds are compounds having metal carbon bonds, where the carbon atoms are part of an organic group. The best known example is probably tetraethyl lead, which is used as an additive to gasoline. It has now been phased out of use in many countries—all industrialized countries—due to its environmental consequences. Organometallic compounds can be formed in nature from metal or metal ions, for instance, di- methylmercury, or are produced for various purposes, as catalysts (e.g., organo-aluminum), as pesticides (e.g., organoarsenic and organotin compounds), as stabilizers in polymers (e.g., organotin compounds), and as gasoline additives (e.g., organolead compounds). Organometallic compounds exhibit properties that are different from those of the metal itself and inorganic derivatives of the metal, for instance, a relatively higher toxicity than the metals.

Most organometallic compounds are relatively unstable and undergo hydrolysis and photolysis easily. Most organometallic compounds have weakly polar carbon–metal bonds and are often hydrophobic. They therefore only dissolve in water to a small extent and are readily sorbed onto particulates and sediments.

The most harmful organometallic compounds from an environmental point of view are organomercury, organotin, organolead, and organoarsenic, which are all very toxic to mammals.

Detergents (and Soaps)

Detergents and soaps contain surface active agents (surfactants) that are classified according to the charged nature of the hydrophilic part of the molecule:

- Anionic: negatively charged
- Cationic: positively charged
- Nonionic: neutral, but polar
- Amphoteric: a zwitterion containing positive and negative charges

They are produced and consumed in large quantities and are mostly discharged into the sewage system and end up in wastewater plants. The early surfactants contained highly branched alkyl hydrophobes that were resistant to biodegradation. These surfactants are largely obsolete today, having been replaced by linear alkyl benzene sulfonates (LAS) and other biodegradable surfactants.

The toxicity to mammals is generally low for all surfactants, while the toxicity to aquatic organisms is relatively high (LC_{50} from about 0.1 to about 77 mg/L). The toxicity will generally increase with the carbon chain length (see Figure 5).

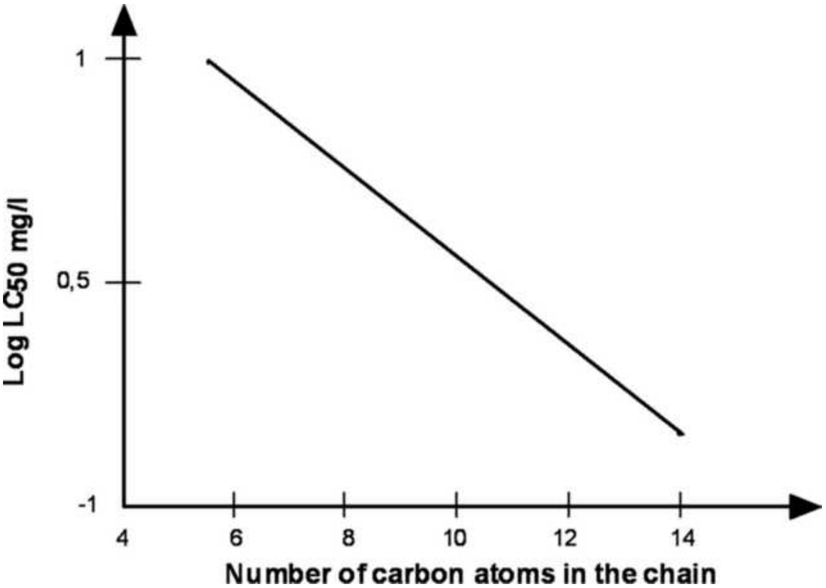


FIGURE 5 Log LC_{50} is plotted versus the number of carbon atoms in the chain for LASs. As seen, increased chain length implies increased toxicity.

Many surfactants bind strongly to soils and sediments, which implies that, to the extent that they are not biodegraded in a biological treatment plant, they will mainly be found in the sludge phase.

Synthetic Polymers and Xenobiotics Applied in the Plastics Industry

Synthetic polymers and xenobiotics applied in the plastics industry form a very diverse group of compounds from a chemical viewpoint. Synthetic polymers are useful (plumbing, textiles, paint, floor, covering, and as the basic material for a wide spectrum of products) because they are resistant to biotic and abiotic processes of transformation and degradation. These properties, however, also cause environmental management problems associated with the use of these components. In addition, several xenobiotic compounds are used as additives, softeners, stabilizers, and so on in synthetic polymer to improve their properties. Some of these additives are very toxic and may cause other and additional environmental problems; for instance, phthalates are widely used in the plastics industry and it has been demonstrated that phthalates have effects as endocrine disrupters.

After use, synthetic polymers are usually incinerated together with industrial and household garbage (solid waste). The presence of poly-vinyl-plastic (PVC) will imply that hydrochloric acid and, to a certain extent, dioxins are formed, but this is strongly dependent on the incineration conditions. As it is difficult to separate different types of plastics, it has been discussed to phase out the use of PVC, but due to the unique properties of PVC, this has not yet been decided.

References

1. Loganathan, B.G.; Lam, P.K.S. *Global Contamination Trends of Persistent Organic Chemicals;* CRC Press: Boca Raton, London, and New York, 2012; 639 pp.
2. Newman, M.C.; Unger, M.A. *Fundamentals of Ecotoxicology,* 2nd Ed.; CRC Press: Boca Raton, London, and New York, 2003.
3. Hoffman, D.J.; Rattner, B.A.; Burton, G.A., Jr.; Cairns, J., Jr. *Handbook of Ecotoxicology;* Lewis Publ./CRC Press: Boca Raton, London, and New York, 1994.
4. Schuurmann, G.; Markert, B. *Ecotoxicology;* John Wiley: New York, 1998; 902 pp.
5. Andersen, H.R. Examination of Endocrine Disruptors. Thesis, DFU, Copenhagen University, 1998.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

II

Natural Elements and Chemicals

14

Allelochemics

John Borden

Terminology

To clarify in part the emerging maze of newly discovered message-bearing chemicals, or *semiochemicals* (Gk. *semeion,* sign or signal), the term *allelochemic* (Gk. *allelon,* of each other) was coined in 1970 to embrace any semiochemical with interspecific activity.[1] Thus allelochemics are distinguished from *pheromones* (Gk. *pherum,* to carry; *horman,* to excite) that convey a message between organisms of the same species. Three categories of allelochemics are commonly recognized.

Kairomones (Gk. *kairos,* opportunistic) are allelochemics that provide an adaptive advantage to the perceiver. In most cases there is no benefit, or even harm, to the emitter, for example, the attraction of predators to the odor of their prey. The evolution of such chemicals as true biological signals would be disadaptive, and therefore unlikely. Therefore some semantic purists remove all evolutionary implications pertaining to kairomones by referring to them as infochemicals.

Allomones (Gk. *allos,* other) are allelochemics that convey an adaptive advantage to the emitter. The repellent odor of an alarmed skunk is often used as an example. However, in an evolutionary sense it may also be adaptive for the receiver to be able to detect and avoid the skunk's odor, for example, for a predator not to be "tagged" with an aroma that warns potential prey of its presence. Therefore, a skunk's odor is more aptly termed a *synomone* (Gk. *syn,* with), an allelochemic that conveys a mutual advantage to both the emitter and the receiver.

Most allelochemics have a *releaser* effect, in which behavioral responses are evoked. However, they may also have a *primer* effect, in which a physiological or biochemical function is stimulated or inhibited.

Natural Occurrence

Table 1 provides a small window on the thousands of allelochemic interactions that occur in nature. The compounds that mediate these interactions are equally diverse (Figure 1). Very few interactions are mediated by a single compound; most involve relatively simple blends; and some, for example, floral fragrances, comprise dozens of compounds in a single blend. Although Table 1 provides examples of allelochemic interactions among terrestrial plants, arthropods, and vertebrates, many are also found among aquatic organisms, and examples occur in all Kingdoms and Phyla.

Kairomonal interactions include the attraction of many species of phytophagous insects to their host plants, entomophagous insects to their prey or to insects that they parasitize, and blood-feeding diptera to their vertebrate hosts.[1–3] They also include the avoidance by prey species of odors associated

TABLE 1 Examples of Natural Occurrence and Function of Allelochemics[a]

Type of Allelochemic and Source	Example
Kairomone:	
Plants	Attraction of Colorado potato beetles, Leptinotarsa decemlineata, to 6-carbon leaf volatiles, e.g., (E)-2-hexen-1-ol-(2), from solanaceous plants
	Attraction of ambrosia beetles, Trypodendron lineatum, Gnathotrichus sulcatus and G. retusus, to ethanol (1) from moribund coniferous trees, logs, and stumps
	Attraction and stimulation of oviposition by onion maggots, Delia antigua, in response to mono- and disulfides, e.g., dipropyl disulfide (3) from onions, Allium cepa
	Employment of host tree kairomones, e.g., a-pinene (8) from conifers and a-cubene (10) from elms as synergists of aggregation pheromones that mediate mass attack of trees by bark beetles e.g., the Douglas-fir beetle, Dendroctonus pseudotsugae, and the smaller European elm bark beetle, Scolytus multistriatus, respectively
Insects	Stimulation in a parasitic chalcidoid wasp, Trichogramma evanescens, of searching for and oviposition in corn earworm eggs, Helicoverpa zea, by tricosane (15) in moth scales adhering to newly laid eggs
	Attraction of predaceous clerid beetles, Thanasimus and Enoclerus spp. to aggregation pheromones, e.g., ipsenol (19), ipsdienol (20), and frontalin (22) of their bark beetle hosts
Vertebrates	Attraction of blood-feeding mosquitoes to CO_2 (23) exhaled by mammals
	Attraction of tsetse flies, Glossina spp., to volatiles, e.g., acetone (24) and 1-octen-3-ol (25) from bovine animals on which they feed
	Avoidance by voles, Microtus spp., of volatile chemicals in the urine of mustellid predators, e.g., 2-propylthiotane (26) and 3-propyl-1,2-dithiolane (27) from short-tailed weasels, Mustella erminea
Allomone:	
Plants	Inhibition of germination of growth of one species of plant by allelopathic chemicals produced in the leaves, roots, or other tissues of another plant, e.g., by juglone (7) leaching from the leaves of black walnut trees, Juglans nigra
	Disruption of growth, metamorphosis, or reproduction of insect herbivores by producing insect hormones or hormone analogues, e.g., juvabione (11), a juvenile hormone mimic in true firs, Abies spp.
Spiders	Emission of moth sex pheromones, e.g., (Z)-9-tetradecenyl acetate (18) by bolas spiders, Mastophora spp., to attract male moths as prey
Insects	Mimicking the cuticular recognition compounds, e.g., 11-methyl pentacosane (16) of larval ants by caterpillars of lycenid butterflies, syrphid fly larvae, and scarab beetle larvae, thereby gaining entry into and acceptance within ant nests, where they prey on ant brood
Synomone:	
Plants	Avoidance of nonhost plants by insects in response to volatiles emitted by the plants, e.g., repellency of coniferophagous bark beetles, e.g., the mountain pine beetle, Dendroctonus ponderosae, to conophthorin (13) (also a repellent pheromone of cone and twig beetles) in the bark of birches, Betula spp.
	Repellency of black bean aphids, Aphis fabae, to methyl salicylate (4) in the volatiles of nonhost plants
	Tritophic interaction in which corn plants, Zea mays, respond to volicitin (14) (a kairomone) in the saliva of beet armyworm caterpillars, Spodoptera exigua, feeding on them by producing specific blends of volatiles (synomones), e.g., (E)-4, 8-dimethyl-1,3,7-nonatriene (9) that attract females of a parasitic wasp, Cotesia marginiventris, which in turn oviposit on the feeding caterpillars
	Attraction of honey bees, Apis mellifera, to multicomponent blends of floral volatiles, e.g., geraniol (12), of many species of angiosperm plants, with mutual benefit of pollination to the plant and a pollen and nectar source to the bees
Insects	Antagonists in the blends of moth sex pheromones that repel males of related species, ensuring reproductive isolation even though the major pheromone components are attractive to males of both species, e.g., (Z)-9-tetradecanal (17), a pheromone component of threelined leafrollers, Pandemis limitata, that inhibits response of oblique banded leafroller males, Choristoneura rosaceana, to threelined leafroller females
	Mutual repellency between aggregation pheromones of two species of bark beetles, e.g., (R)-(–)-ipsdienol (20) produced by pine engravers, Ips pini, and (S)-(–)-ipsenol (19) produced by California fivespined ips, I. paraconfusus, reserving the host phloem resource for the first-arriving species, and avoiding interspecific exploitative competition

[a] *Numbers in parentheses correspond with numbered compounds in Figure 1.*

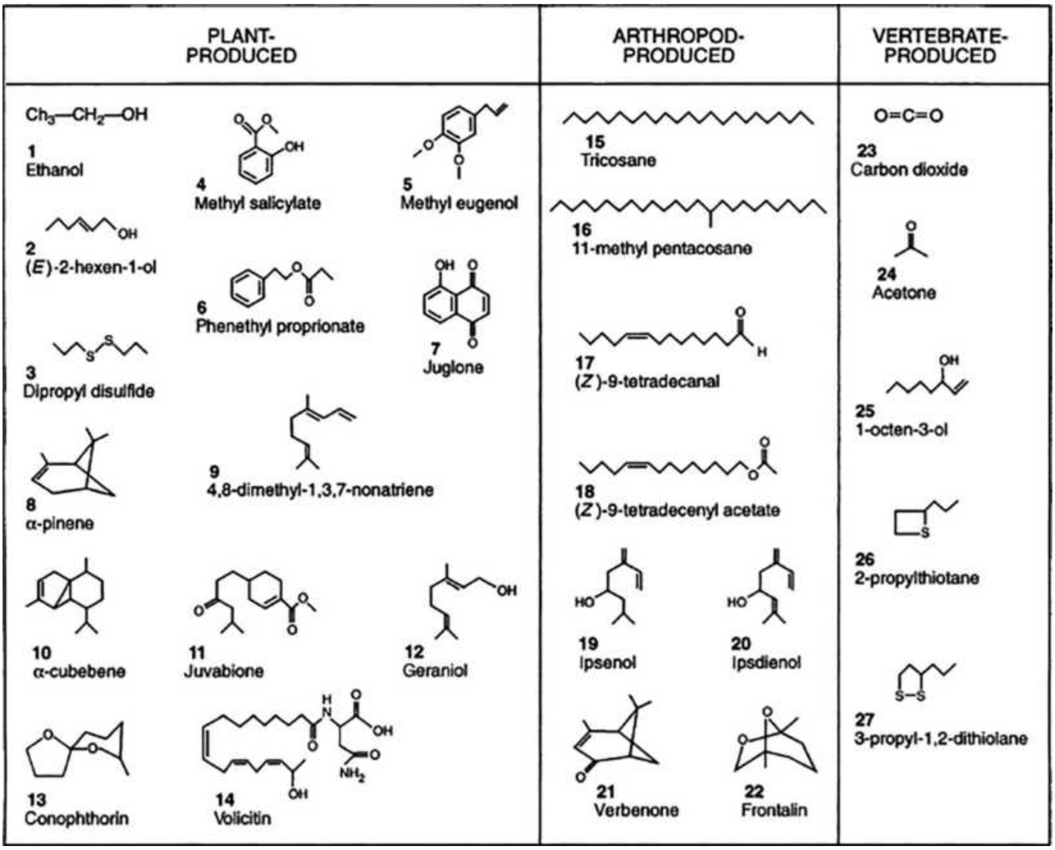


FIGURE 1 Structural formulae of compounds given in Table 1 exemplifying some of the chemical diversity among allelochemics as follows: primary alcohol (**1**, **2**), secondary alcohol (**25**), disulfide (**3**), aromatic ester (**6**), unsaturated ester (**18**) monoterpene (**8**), sesquiterpene (**9**, **10**), sesquiterpenoid (**11**), terpene alcohol (**12**, **19**, **20**), terpene ketone (**21**), spiroacetal (**13**), fatty acid derivative conjugated to an amino acid (**14**), straight chain hydrocarbon (**15**), branched hydrocarbon (**16**), unsaturated aldehyde (**17**), bicyclic ketal (**22**), atmospheric gas (**23**), ketone (**24**), thiotane (**26**), and thiolane (**27**).

with predators, a phenomenon found in five animal phyla, but curiously not yet among terrestrial insects.[4] Sometimes more than one type of semiochemical may be involved; for example, the aggregation of bark beetles necessary to mass attack and kill a tree is mediated by a blend of aggregation pheromones synergized by host tree kairomones.

Allomonal interactions may employ “trickery”,[5] for example, bolas spiders that emit moth sex pheromones that lure mate-seeking male moths to their death, and many species of myrmecophiles (ant lovers) that gain access to ant nests by chemically mimicking the cuticular recognition compounds of ants on which they prey. The most well-known allomones are released by or are contained in plants, and have a primer effect. Allelopathic allomones are often leached from the leaves (or other parts) of plants of one species, and inhibit the germination of seeds or growth of plants in other species that could be potential competitors.[6] Some plants may also produce defensive allomones against insect herbivores, for example, hormones or analogues of hormones that disrupt the growth and metamorphosis of their insect enemies.

Among the many examples of synomones are repellents that ensure reproductive isolation between closely related species of insects, or mitigate against the occurrence of interspecific exploitative competition for a limited host resource. Often synomones are the same compounds as one or more of the components that convey a pheromonal message, for example, to attract mates or to aggregate on or near a food source. There is increasing evidence that host-seeking phytophagous insects not only use kairomones to find their host plants, but also use synomones to avoid nonhost plants on which their fitness would be greatly reduced. Synomonal floral scents provide a mutual benefit to flowering plants that gain from pollination by insects that in turn are attracted to a nutritious nectar or pollen source.

Another type of allelochemic interaction involves three trophic levels and the action of both kairomonal and synomonal stimuli.[3] In one remarkable example of this type of tritrophic interaction, corn plants being fed on by beet armyworm caterpillars are exposed to minute amounts of a kairomone called volicitin in the insect's saliva. Volicitin has a primer effect, eliciting the plant to synthesize a specific blend of volatile synomonal compounds that attract females of a parasitic wasp. The wasp oviposits in the beet armyworm larvae, benefiting by finding its host, and in turn providing an advantage to the plant by parasitizing and killing the caterpillar.

Practical Applications

Knowledge about the natural occurrence and role of allelochemics opens up a huge, but relatively untapped, potential for exploiting them as pest management tools.[2] In some cases the knowledge itself is important. Plant breeders may seek varieties of plants that contain or release chemicals that deter feeding or development by phytophagous insects. Plants with allelopathic characteristics, for example, *Eucalyptus* spp., may be useful in landscaping to reduce weed problems. Species or varieties rich in attractive kairomones may be used as trap crops for various insect pests. In the production of transgenic agricultural crops it is critical not to lose the capacity for tritrophic interaction that will ensure parasitism of herbivorous insects, lest the genetically modified plants be more vulnerable to insect pests than unmodified plants.

In other cases, the capacity to use allelochemics as pest management tools may be demonstrated, but technological, economic, or social limitations may prevent their use. Allelopathic allomones from plants are under consideration for development as a new class of biodegradable herbicides.[7] But to date none can compete with conventional chemical herbicides with regard to ease of synthesis, capacity for formulation, efficacy, and/or safety. If used widely, some allelo-pathogens may pose an unacceptable threat to environmental or human health. Recent investigations show considerable promise for using nonhost volatiles to "disguise" herbivorous host plants or trees as nonhosts, but commercial formulations have not yet appeared on the market, in part because of the challenge and expense of registering those allelochemic products as pesticides. Similar problems beset the development and use of predator volatiles to protect plants from damage by herbivorous vertebrates such as deer and voles.

Despite the above limitations, a few kairomones have found widespread commercial use.[2,8] Among them is methyl engenol, which is used worldwide for capturing tephritid fruit flies, both for detection of unwanted introductions and for direct suppression of populations in a lure and kill tactic employing an insecticide-laced substrate baited with methyl engenol. A lure and kill tactic is also used effectively for control of tsetse flies that are drawn by acetone and 1-octen-3-ol baits to insecticide-treated "target" traps that simulate the silhouette of a large vertebrate. Other applications combine kairomones with pheromones. A combination of phenethyl proprionate and methyl engenol with the sex pheromone of the Japanese beetle is used in many thousands of traps in the United States each year. Similarly the kairomones ethanol and α-pinene have been used since 1981 in combination with aggregation pheromones in commercial mass trapping programs for three species of ambrosia beetles in British Columbia.

Allelochemics may also find use in the future in the application of "push-pull" tactics, in which one repellent volatile treatment is used to protect a plant or group of plants from attack by insects (push), and another attractive treatment is used in baited traps or trap plants to pull the insects away. One outstanding example of a successful push-pull application saved a rare stand of endangered Torrey pines in California from being killed by the California five spined ips. Two repellent synomones, verbenone, produced by western pine beetles, and (–)-ipsdienol, produced by pine engravers, were deployed inside the uninfested portion of the stand, and traps baited with attractive aggregation pheromone were arrayed in an adjacent area of beetle-killed pines. Over 86 weeks beginning in May 1999, 330,717 beetles were captured.

Conclusion

Despite many studies, most natural allelochemic interactions are yet to be discovered. The adoption of allelochemics as pest management tools has been limited. However, there is great potential for judicious selection and commercial development of allelochemics, particularly in integrated pest management programs that will combine a number of alternative ecologically based tactics with the reduced use of conventional chemical pesticides.

References

1. Whittaker, R.; Feeney, P. Allelochemics: chemical interactions between species. Science **1971**, *171,* 757–770.
2. Metcalf, R.; Metcalf, E. *Plant Kairomones in Insect Ecology and Control;* Chapman and Hall: New York, 1992; 168.
3. Vet, L.; Dicke, M. Ecology of infochemical use by natural enemies in a tritrophic context. Annu. Rev. Entomol. **1992**, *37,* 141–172.
4. Kats, L.; Dill, L. The scent of death: chemosensory assessment of predation risk by prey animals. Ecoscience **1998**, *5,* 361–394.
5. Stowe, M.; Turlings, T.; Loughrin, J.; Lewis, W.; Tumlinson, J. The Chemistry of Evesdropping, Alarm, and Deceit. In *Chemical Ecology: The Chemistry of Biotic Interaction*; Eisner, T., Meinwald, J., Eds.; National Academy Press: Washington, DC, 1995; 51–65.
6. Zindahl, R. *Fundamentals of Weed Science;* Academic Press: New York, 1993; 135–146.
7. Cutler, G. Allelopathy for Weed Suppression. In *Pest Management: Biologically Based Technologies*; Lumsden, R., Vaughn, J., Eds.; American Chemical Society: Washington, DC, 1993; 290–302.
8. Borden, J. Disruption of Semiochemical-Mediated Aggregation in Bark Beetles. In *Insect Pheromone Research: New Directions;* Cardé, R., Minks, A., Eds.; Chapman and Hall: New York, 1997; 421–438.

15

Aluminum

Johannes Bernhard Wehr, Frederick Paxton Cardell Blamey, Peter Martin Kopittke, and Neal William Menzies

Introduction

Aluminum (Al) is a metallic element characterized by its low density (2.7 g/cm^3) and resistance to corrosion due to the formation of a protective oxide layer on the surface. It is the most common (8%) metallic element of the earth's crust and the third most common element (after oxygen and silicon) on earth (Wehr, Blamey, and Menzies 2007). The element reacts strongly with oxygen-containing ligands and never occurs naturally as pure metal. It forms both octahedral and tetrahedral coordination compounds and occurs only in the trivalent oxidation state. In soil, Al-containing minerals dissolve at low pH, especially below pH 4.5, and release Al into the soil solution and aquatic environment. Al can also occur in the atmosphere as aeolian dust or ash.

Soluble forms of Al can have toxic effects on plants, humans, as well as soil and aquatic organisms, but alleviation of Al toxicity can be achieved by increasing the pH to above 4.5. Aluminum present in water and food is not readily taken up by humans. Nano-sized Al particles in the environment may result in some effects on plants and animals but more research is needed.

The chemistry and uses of Al are discussed, as well as the sources and alleviation of Al toxicity in soils, water, and atmosphere. Finally, the environmental toxicology of Al to plants, microorganisms, and animals is briefly discussed.

Chemistry of Al

Dissolution of Al minerals at low pH releases octahedral Al^{3+}, which is hexa-coordinated with water molecules (Martin 1996). Soluble Al^{3+} is classified as a hard acid due to its small ionic radius of 0.053 nm and reacts strongly with "hard" ligands such as oxygen and fluoride (Martin 1996, Martell et al. 1996). It also forms stable complexes with didentate and multidentate ligands (chelate effect) (Martell et al. 1996, Salifoglou 2002). As the pH of an acidic Al solution is increased, especially >4, hydrolysis of Al occurs, giving rise to a series of Al-hydroxy species [$AlOH^{2+}$, $AlOH_2^{+}$, and $Al(OH)_3$] (Martin 1996), which may undergo aggregation (polymerization) via OH bridges if the Al concentration is sufficiently high (>0.3 mg/L) (>10 µM) (Furrer, Trusch, and Muller 1992). As the pH is raised above pH 6, Al changes its coordination number to 4 and yields tetrahedral aluminate ($AlOH_4^{-}$) (Martin 1996). Depending on the pH and Al concentration in solution, hydrolyzed Al species may aggregate and form polymeric (polycationic) Al species such as $Al_2(OH)_2^{4+}$, $Al_3(OH)_4^{5+}$, and $Al_8(OH)_{20}(H_2O)x^{4+}$, the "gibbsite fragment" model forms, $Al_6(OH)_{12}(H_2O)_{12}^{6+}$ through $Al_{54}(OH)_{144}(H_2O)_{36}^{18+}$, $Al_{13}O_4(OH)_{24}(H_2O)_{12}^{7+}$ (Al_{13}), and $Al_2O_8Al_{28}(OH)_{56}(H_2O)_{26}^{18+}$ species (Bertsch and Parker 1996, Brown et al. 1985, Orvig 1993, Sarpola et al. 2006, Wang and Muhammed 1999). The Al_{13} polycation in solution is highly toxic to root elongation (Bertsch and Parker 1996). While many more Al polymers have been proposed, the experimental evidence in support of these species is limited.

Al Uses and Production

Due to its corrosion resistance, light weight, and excellent thermal and electrical conductivity, the metal is extensively used in the building and construction industries (window frames, doors, external cladding, A/C ducts, thermal insulation), automotive (engine blocks, car bodies), shipping (hulls) and aerospace industries (aircraft bodies), power lines, and food packaging (cans and other containers, foil). Aluminum salts and compounds are used for water purification, in which alum ($KAl(SO_4)_2.12H_2O$) is important, as catalysts in the chemical industry, and as ingredients in cosmetics (antiperspirants), pharmaceuticals (antacids, vaccine adjuvant), and foods (baking powder, spreading agent) (Table 1).

The main producer of bauxite ore is Australia (30% of world production, 88 Mt in 2017), with China, Brazil, Guinea, Jamaica and India producing lesser quantities (Resources-and-Energy-Quarterly 2018) (Table 2). The main producers of alumina in 2017 were Australia and China (World-Aluminum.org 2019), whereas smelting (refining of bauxite) is occurring in regions with cheap electrical energy (Table 2).

TABLE 1 Global Uses of Al (2018)

Use	(%)
Transport and manufacturing	41
Packaging	20
Construction	14
Electrical	8
Consumer durables	7
Machinery	7
Other	3

Source: https://archive.industry.gov.au/Office-of-the-Chief-Economist/Publications/ResourcesandEnergyQuarterlyMarch2018/documents/Resources-and-Energy-Quarterly-March-2018-Aluminium-alumina-and-bauxite.pdf (accessed 29 March 2019).

TABLE 2 Worldwide Production of Alumina and Primary Al Metal in 2017

Geographic Region	Alumina (kt)	Primary Al Metal (kt)
Africa and Asia (excl China)	8,382	10,776
North America	3,033	3,950
South and Central America	12,713	1,378
East and Central Europe	4,499	3,999
Western Europe	5,890	3,776
China	70,699	35,905
Oceania	20,783	1,817
Rest of World	6,391	1,800
Total	132,400	63,404

Source: www.world-aluminium.org/statistics/primary-aluminium-production (accessed 29 March 2019).

Sources of Al in Soil

Al Minerals

The main Al-containing primary minerals are feldspars and micas. Several gemstones (ruby, sapphire, tourmaline) also contain Al. The primary Al-containing minerals weather initially to 2:1 layer aluminosilicate clay minerals (e.g., vermiculite, montmorillonite, smectite), which upon further weathering form 1:1 layer aluminosilicates such as kaolinite. Further weathering of clay minerals leads to leaching of silica and base cations (calcium and magnesium), leaving behind hydrous aluminum oxide (e.g., gibbsite, boehmite, and diaspore) in the form of bauxite, which is an important aluminum ore (Wehr, Blamey, and Menzies 2007).

Forms of Soluble Al in Soil

In acidic soils ($pH < 4.5$), aluminosilicate clay minerals are unstable and dissolve, releasing the trivalent Al^{3+} cation. The solubility of Al increases three orders of magnitude for every unit decrease in solution pH: at pH 4.5 the soluble Al concentration is around 0.09 mg/L (3.5 μM), decreasing to 95 ng/L (3.5 nM) at pH 5.5 (Menzies, Bell, and Edwards 1994). The trivalent Al^{3+} ion is toxic to plant roots at concentrations of 0.1–1 mg/L (5–50 μM) (Brown et al. 2008, Horst, Wang, and Eticha 2010, Poschenrieder et al. 2008). The Al^{3+} ion is readily complexed by soil organic matter and the soil organic matter controls the Al availability in most soils containing sufficient organic matter (Adams et al. 2000, Brown et al. 2008, Lofts et al. 2001, Simonsson 2000, Skyllberg 1999, Guo et al. 2007). In mineral soils low in organic matter, the availability of Al is determined by the cation exchange capacity, ionic strength, and pH (Guo et al. 2007). The polymeric Al species (e.g., Al_{13}) are metastable with a half-life for Al_{13} of several hundred hours (Etou et al. 2009, Furrer, Gfeller, and Wehrli 1999) which implies that the species undergo depolymerization or crystallization with time. The Al_{13} species is also rhizo-toxic to plant roots, with concentrations between 0.1 and 2 mg/L (0.1–2 μM) inhibiting root growth (Kopittke, Menzies, and Blamey 2004, Bertsch and Parker 1996). It is generally accepted that sulfate, phosphate, fluoride, silicate, and organic acids prevent or reverse Al_{13} formation (Bertsch and Parker 1996, Casey 2005, Masion et al. 1994, Yamaguchi et al. 2003, Kerven, Larsen, and Blamey 1995) and crystalline gibbsite may induce crystallization and depolymerization of Al_{13} (Sanjuan and Michard 1987). Since sulfate and silicate ions are prevalent even in acid soils, the natural occurrence of Al_{13} in soil solution is considered unlikely (Bertsch and Parker 1996, Gerard, Boudot, and Ranger 2001, Hiradate, Taniguchi, and Sakurai 1998). However, there is evidence that Al_{13} may form in the cell wall of plant roots (Kopittke, Menzies, and Blamey 2004, Masion and Bertsch 1997, Xia and Rayson 1998) even if the conditions in bulk solution are not favorable for Al_{13} formation. A report by Hunter and Ross (Hunter and Ross 1991) claiming that up to 30% of acid forest soils contained Al_{13} could not be replicated by other research groups. In alkaline soils, the aluminate anion $[Al(OH)_4^-]$ can form but this species is not very toxic to plants (Kinraide 1991, Kopittke, Menzies, and Blamey 2004) but may be toxic to aquatic organisms (Griffitt et al. 2008, Sjostedt et al. 2009).

Mobilization of Al in Soil

Aluminosilicate and aluminum sesquioxide clay minerals slowly dissolve at $pH < 4.5$ and release Al into the soil solution. Currently, more than 30% of potentially arable soils are acidic and affected by Al toxicity (von Uexküll and Mutert 1995). Soil acidification and consequential Al toxicity can be exacerbated by anthropogenic acid deposition (acid rain), use of ammonium-containing fertilizers, use of legumes in crop rotations, and the removal of crops from agricultural land (von Uexküll and Mutert 1995, Lesturgez et al. 2006). In unfarmed soils, atmospheric acid inputs are the main causes of acidification (Lapenis et al. 2004, Bergholm, Berggren, and Alavi 2003), whereas acidification in fertilized soils is mainly caused by nitrification of ammonium N-fertilizer, especially when the N is not taken up by the plant and is leached from the soil (Bergholm, Berggren, and Alavi 2003, Guo et al. 2010).

Aluminum in soil solution decreases turn-over of soil organic matter by limiting microbial degradation of the stable Al-organic matter complexes formed in soil, leading to an increase in soil organic matter (Scheel, Dorfler, and Kalbitz 2007, Scheel et al. 2008, Rasmussen, Southard, and Horwath 2006, Takahashi and Dahlgren 2016, Miyazawa et al. 2013).

Aluminum can also be released from soil minerals in strongly acidic conditions found in acid sulfate soils (Dent and Pons 1995, Faltmarsch, Astrom, and Vuori 2008), monosulfidic black ooze (Bush, Fyfe, and Sullivan 2004) and acid mine drainage (Liang and Thomson 2009, Pu et al. 2010). Acid sulfate soils are limited in occurrence to permanently or temporarily waterlogged low-lying areas (up to 5 m above sea level) (Dent and Pons 1995), whereas organic matter in irrigation channels can lead to the formation of monosulfidic black ooze (Bush, Fyfe, and Sullivan 2004). Exposure of these sulfide-containing materials to air leads to the formation of sulfuric acid and release of Al (Dsa et al. 2008, Liang and Thomson 2009, Pu et al. 2010, Soucek, Cherry, and Zipper 2003).

Fluoride ions (F^-), present as an impurity in phosphate fertilizers, can form strong complexes with Al and the resultant lesser charged complexes, AlF^{2+}, AlF_2^+, and AlF_3, are mobile in soil solution (Martin 1996, Martinent-Catalot et al. 2002) and can be either removed from soil by drainage or seepage water, or taken up by plants (Manoharan et al. 2007) and alleviate Al toxicity (Yang et al. 2016).

Organic acids form complexes with Al and can mobilize Al in the soil (Takahashi et al. 2008, Lange, Solberg, and Clarke 2006a). In the field, movement of organic acid–Al complexes gives rise to podzolization of soils and downward movement of Al until conditions are favorable for dissociation of Al complexes deeper in the profile or export to waterways (Bardy et al. 2007).

Alleviation of Al in Soils

Since the solubility of Al in soil solution is lowest around pH 6–7 (McBride 1994), controlling the pH is the most effective way in reducing potential Al toxicity in ecosystems. The pH value to which acid soils need to be limed depends upon the plant species, but is often near 5.5. Compounds to increase the pH of acid soils include calcitic ($CaCO_3$) or dolomitic ($CaMgCO_3$) limestone (Machacha 2004), slaked lime $[Ca(OH)_2]$ (Sun et al. 2000), wood ash (Materechera and Mkhabela 2002), alkaline poultry manure (Mokolobate and Haynes 2002, Tang et al. 2007), charcoal (Steiner et al. 2007), biochar (Dai et al. 2017), fly ash (Morikawa and Saigusa 2002, Tarkalson et al. 2005), ground basalt (Panhwar et al. 2016), alkaline slag (Shi et al. 2017, Li et al. 2015) or waste cement (Morikawa and Saigusa 2002). Any waste product with high ash alkalinity can be used to ameliorate acid soil by increasing the pH (Naramabuye and Haynes 2006). Surface application of liming agents with subsequent incorporation into the soil (ploughing, harrowing, etc.,) is commonly used. The mobility of limestone in the soil profile is very low (Conyers et al. 2003, Godsey et al. 2007, Scott et al. 2007). Therefore, rapid alleviation of subsoil acidity can only be achieved by deep placement of limestone (deep ripping). This, however, is currently not economically possible for most low-value broadacre crops. Depending on the severity of soil acidity, and the buffer capacity of the soil as determined by the cation exchange capacity of the soil, between 0.1 and 1 ton lime-stone/ha/yr are needed to maintain soil pH in agricultural production systems (Scott et al. 2007, Machacha 2004). Numerous methods have been developed to determine the required liming rate based on laboratory soil analyses. The liming of soil needs to be repeated every few years depending on the land management system. To overcome the low mobility of limestone in soils, higher rates of gypsum can be used instead. The Ca from gypsum moves more readily down the soil profile (Liu and Hue 2001) and the increase in Ca and ionic strength can lower the Al toxicity, despite increasing the overall concentration of Al (as $AlSO_4^+$, which is less toxic than Al^{3+}). Liming has greater beneficial effect on soil bacterial communities than on soil fungal communities (Mota et al. 2008, Nelson and Mele 2006).

Application of organic matter (green manure, farm yard manure) can lower Al toxicity by complexing Al (Qin and Chen 2005, Vieira et al. 2008). However, application of organic matter requires higher application rates (>10 tons/ha) than limestone (1 ton/ha) and is, therefore, not as effective as limestone (Raboin et al. 2016). Humic acid has been shown to either precipitate or bind Al_{13} (Yamaguchi et al. 2004) and Al^{3+}, (Matthias, Maurer, and Parlar 2003, Shoba and Chudnenko 2014), thereby reducing root growth inhibition. Biochar can also adsorb Al-OH complexes and reduce rhizotoxicity in crops (Qian, Chen, and Hu 2013). Sewage sludge can also be alkaline and the presence of organic acids in sludge can bind and immobilize Al (Lopez-Diaz, Mosquera-Losada, and Rigueiro-Rodriguez 2007).

Alkaline bauxite refinery waste (red mud) poses a challenge for plant growth due to the high Fe oxide concentration leading to phosphate immobilization, absence of diverse microbial communities, and lack of organic matter (Wehr, Menzies, and Fulton 2006). However, the presence of aluminate in these waste materials is not limiting to plant root growth (Kopittke, Menzies, and Blamey 2004).

Sources of Al in Water

Acid Deposition

Atmospheric gases such as NO_2, SO_2, and CO_2 dissolve in rainwater and form acid rain (Lapenis et al. 2004, Lawrence 2002). The atmospheric acid inputs in Europe are considered to be around 0.2–4 kmol/ha/yr (de Vries, Reinds, and Vel 2003). The acids can either dissolve Al minerals in soil (see above) or dissolve sediments in water bodies receiving acidic water. Furthermore, acid deposition can release Al complexed to organic matter. Introduction of gaseous emission standards in the 1980s in Europe, and later in the United States, has lowered acid rain, with a consequential decrease in Al in surface water (Lange, Solberg, and Clarke 2006b, Skjelkvale et al. 2001, Vuorenmaa et al. 2018). Podzolization of lateritic soils was found to export vast quantities of Al to waterways (Bardy et al. 2007). The concentration of Al in drainage water from forest watersheds is highly variable, however (Lange, Solberg, and Clarke 2006b). In coastal areas subjected to atmospheric salt deposition, the salt can displace Al from cation exchange sites, leading to an increase in Al in drainage water (Lange, Solberg, and Clarke 2006b). Concentrations of Al in groundwater are governed by pH since Al solubility is pH dependent: less than 0.01 mg/L Al is observed at pH 7, which increases to 51 mg/L at pH < 4 (Fest et al. 2007).

Acid Mine Drainage and Acid Sulfate Soils

Sulfidic mine wastes and acid sulfate soils release acid when exposed to air, resulting in leaching of acid drainage water and soluble Al into water bodies. While there are indications that the high Fe concentration in acid mine drainage can protect aluminosilicate clay mineral from dissolution and minimize Al release (Dubikova et al. 2002), no such mechanism has been reported for acid sulfate soils. At Trinity inlet in North Queensland, Australia, Al concentrations in drainage water from acid sulfate soils were measured in the range 4.2–10 mg/L (Hicks, Bowman, and Fitzpatrick 1999).

Aluminum from Water Purification

Water clarification involving the settling of fines (suspended colloids and microorganisms) may be accomplished with alum (potassium aluminum sulfate). Since drinking water will be near neutral (pH 6.5–8.5 (NHMRC 2004)), solubility of Al is low. The European, Australian, New Zealand, and WHO regulations for drinking water propose a threshold concentration of Al of 0.1-0.2 mg/L (Schaefer and Seifert 2006). Traces of Al are found in raw and alum-treated potable water (Cech and Montera 2000, Srinivasan and Viraraghavan 2002). The amount of Al taken up from drinking water has been estimated at 1%–2% of total Al uptake per day (Yokel, Hicks, and Florence 2008, Stauber et al. 1999, Willhite, Ball, and McLellan 2012), and less than 2% of Al in drinking water is bioavailable (Willhite, Ball, and McLellan 2012, Yokel and McNamara 2001).

Water treatment residues (alum sludges) have been used to immobilize phosphate and heavy metals in soils and wastes (Mahdy et al. 2009, Oladeji, Sartain, and O'Connor 2009). Land application of these sludges has not resulted in elevated Al concentrations in plants grown in these soils (Oladeji, Sartain, and O'Connor 2009), especially when the sludges were aged to reduce Al availability (Agyin-Birikorang and O'Connor 2009).

Alleviation of Al in Water

Raising the pH of water to pH 6–7 will precipitate Al and establish an equilibrium concentration of Al determined by the solubility product of Al-hydroxide in water. At pH < 4.5, solubility of Al minerals increases rapidly, releasing Al^{3+} into the waterbody and at pH > 8, aluminate anions are formed (Wehr, Blamey, and Menzies 2007). Therefore, pH adjustment of water to near neutrality is the most effective method to decrease Al concentrations in water (Fest et al. 2007). The Al remaining in potable water after purification can only be removed by ion exchange resins (Othman, Abdullah, and Abd Aziz 2010).

Sources of Al in the Atmosphere

Atmospheric dust derived from raised soil particles is the main source of Al in the atmosphere, but volcanic activity can be a source also. In the atmosphere, Al occurs generally as silicate, sulfate, or oxide compounds. Flue ash and flue gases can also contribute Al to the atmosphere (Kabata-Pendias and

TABLE 3 Concentrations of Al in the Atmosphere

Region	Al Concentration (μg/m³)
South Pole	0.0003–0.0008
Greenland	0.24–0.38
Shetland Islands	0.06
Norway	0.03
Germany	0.16–2.90
Japan	0.04–10.60
China	2.00–6.00
North America	0.60–2.33
Central America	0.76–0.88
South America	0.46–15.00

Source: Adapted from Kabata-Pendias and Kabata 2001 and Wang et al., 2007.

Kabata 2001). Dust contains around 2.6% Al (Schussler, Balzer, and Deeken 2005). Concentrations of Al in the atmosphere vary between geographic locations (Table 3). Inhalation of Al from the atmosphere has been estimated as 4–20 μg Al/day (Yokel, Hicks, and Florence 2008).

Sources of Al in Foodstuff, Cosmetics, Pharmaceuticals, and Workplaces

The concentration of Al in foodstuff is generally very low since little Al is taken up by plants and translocated to the aboveground parts (Table 4) (Chen et al. 2008, Fung et al. 2009, Mueller, Anke, and Illing-Guenther 1998, Ertl and Goessler 2018, Schaefer and Seifert 2006). However, the concentration of Al in black tea is high (Table 4), but most of the Al in infused tea is complexed with phenolics and organic acids, though addition of lemon or lime juice may increase Al availability (Fung et al. 2009, Krewski et al. 2007). The bioavailability of Al from black tea is around 0.4% (Yokel and Florence 2008).

Preparation of acidic foodstuff in Al containers or covering with Al foil may increase Al in foodstuff (Ertl and Goessler 2018, Turhan 2006, Ranau, Oehlenschlager, and Steinhart 2001). The addition of sodium aluminum phosphate as a spreading agent to processed foodstuff (e.g., cheese spread) can also contribute to the dietary Al intake (Yokel, Hicks, and Florence 2008, Schaefer and Seifert 2006). Salts of Al are also used as a food additive for clarification of beverages, as anticaking agent, as baking powder, and to enhance color stability (Stahl et al. 2018, Aguilar et al. 2008, Schaefer and Seifert 2006). The Al in food contributes around 95% of daily oral Al uptake and the daily dietary Al intake has been estimated as 4–16 mg, of which 0.1%–0.3% is bioavailable (Krewski et al. 2007, Yokel, Hicks,

TABLE 4 Concentration of Al in Foodstuff

Foodstuff	Al Concentration (mg/kg DW)
Grains	1–135
Pulses	3–16
Fruit	0.5–20
Vegetables	1–270
Herbs	8–280
Spices	6–695
Meat	0.5–16
Fish	0.5–10
Dairy	1–16
Tea	900–1000

Source: Adapted from Kabata-Pendias and Kabata 2001, Yi and Cao 2008, Fung et al. 2009, Ertl and Goessler 2018, Mueller, Anke, and Illing-Guenther 1998, Schaefer and Seifert 2006, Chen et al. 2008.

TABLE 5 Estimated Sources and Daily Exposure to Al, its Bioavailability, and Estimated Daily Intake

Environmental Exposure	Exposure (μg/day)	Availability (%)	Intake (μg/day)
Air	2–200	2	0.04–4
Industrial air	25,000	2	
Water	200–1,000	<2	2–10
Food	8,000–16,000	<0.3	8–48
Cosmetics	<70,000	0.2	140
Vaccines	1–8	95	1–8
Antacids	<5,000,000	<2	100,000

Source: Adapted from Yokel, Hicks, and Florence 2008, Krewski et al. 2007, Gourier-Frery and Frery 2004.

and Florence 2008, Stahl et al. 2018) (Table 5). The tolerable weekly intake of Al has been set by the European Food Safety Authority to 1 mg Al per kilogram of bodyweight per week (Stahl et al. 2018, Aguilar et al. 2008).

Cosmetics, especially antiperspirants, contain 25% Al as Al-chlorohydrate (Shen and Nardello-Rataj 2009). Daily antiperspirant use can contribute around 70 mg Al per day, yet dermal absorption of Al is only 0.02%–0.2% (Yokel, Hicks, and Florence 2008, Flarend et al. 2001). Buffered aspirin and antacids often contain Al hydroxide suspensions and can contribute up to 5 g Al/day (Yokel, Hicks, and Florence 2008).

Al-phosphate and Al-hydroxide are used as adjuvants to enhance vaccine effectiveness (Exley, Siesjo, and Eriksson 2010). It has been estimated that the amount of Al taken up from Al-containing adjuvant in vaccine is in the range 1–8 µg Al/day (Yokel, Hicks, and Florence 2008), with a lifetime intake of 15 mg, of which most is likely bioavailable (Yokel and McNamara 2001).

Environmental Toxicology of Al

Al Toxicity in Plants

The leading cause of poor fertility of acidic soils is Al toxicity (Asher, Grundon, and Menzies 2002, Horst, Wang, and Eticha 2010, Poschenrieder et al. 2008). Toxicity of Al in plants is mainly manifested by an inhibition of root growth, with Al^{3+} and Al_{13} considered the most toxic species (Kinraide 1991, 1997, Poschenrieder et al. 2008, Singh et al. 2017). The action of Al on plant cells is both in the apoplast (binding to cell wall components) (Kopittke et al. 2015) and in the cytoplasma (formation of reactive oxygen species and oxidative stress responses (Yamamoto et al. 2003, Singh et al. 2017)). The correlation between Al concentration and root grow inhibition is sometimes poor. This is due to the several Al-hydroxy species that coexist within a narrow pH band and cannot be investigated in isolation. Furthermore, the activities of individual species must be calculated from equilibrium data that may be uncertain (Boudot et al. 1996). The critical Al concentration at which root elongation is inhibited is as low as 0.1–0.5 mg/L (5–20 µM) in solutions of low ionic strength, representative of acid soils. Exact critical values depend on the plant species and the conditions in which the plants are grown. Generally, root hairs are more sensitive to Al toxicity than roots (Brady et al. 1993). Root growth inhibition results in short stubby roots and absence of root hairs, leading to poor water and nutrient utilization by plants. The main site affected by Al is near the elongation zone of roots (1–4 mm behind the root tip) (Blamey, Nishizawa, and Yoshimura 2004, Sivaguru and Horst 1998, Kopittke et al. 2015). The reaction of aluminum with phosphate anions in the soil may result in the precipitation of Al-phosphate minerals (e.g., variscite), lowering phosphorus availability to plants. This effect is accentuated by the lower phosphorus uptake capacity of Al-damaged roots.

In plants, detoxification of Al is generally achieved by organic acid exudation (Chen, Wang, and Yeh 2017, Ryan, Delhaize, and Randall 1995), but co-precipitation of Al with Si in the cell wall to form insoluble phytolith has also been reported (Hodson and Sangster 1999, Sangster and Hodson 2001).

Many *Rhizobium* strains necessary for nodulation in legumes are inhibited by soil Al, with 0.03 mg/L Al at pH 5 being toxic (Kinraide and Sweeney 2003). Therefore, nodulation and nitrogen fixation are decreased in acidic soils (Jaiswal, Naamala, and Dakora 2018, Ramirez et al. 2018). Soil bacteria are considered to be more sensitive to low pH and Al than soil fungi (Shirokikh et al. 2004). Swarming of some bacteria can also be inhibited by Al (Illmer and Schinner 1997).

Nano-sized aluminum oxide particles are more toxic than micro-sized Al oxide particles but soluble Al^{3+} is most toxic (Shabnam and Kim 2018). In whole plants, nanoparticles are adsorbed onto roots and not translocated, thus cause no toxicity (Shabnam and Kim 2018), but are toxic when applied to callus cultures (Poborilova, Opatrilova, and Babula 2013). The effect of nanoparticles on soil microbial process is still unclear (Doshi et al. 2008, Mishra and Kumar 2009, Jiang, Mashayekhi, and Xing 2009) and requires more research.

Al Toxicity in Aquatic Organisms

In solution, Al is highly toxic to aquatic organisms in fresh waters (Gensemer and Playle 1999) with >0.4 mg/L Al in freshwater streams decreasing macroinvertebrates (Waters and Webster-Brown 2013) and >0.02 mg/L causing mortality in brown trout (Andren and Rydin 2012). In seawater, awhich is alkaline, presence of aluminate anions can affect marine organisms (Golding et al. 2015).

In freshwater fish and crayfish, Al impairs gill function by inducing secretion of mucus and causes oxygen stress (Ward, McCrohan, and White 2006, Griffitt et al. 2011). No biomagnification of Al was observed in aquatic food webs (Winterbourn, McDiffett, and Eppley 2000).

Maximal Al accumulation occurs on gills at pH 6–8 (Ward, McCrohan, and White 2006, Gensemer and Playle 1999) but complexation with organic matter can prevent Al binding to gills and decrease toxicity (Trenfield et al. 2012).

Nanoparticulate Al is less toxic than Al^{3+} in aquatic ecosystems and AlNP did not cause gill modification and no mortality up to 12 mg/L in moderately hard water (Griffitt et al. 2011). Freshwater algae showed toxicity at 40–45 mg/L after 3 day exposure to nano Al (Sadiq et al. 2011) The freshwater Daphnids are sensitive to nano-sized Al (Strigul et al. 2009, Pakrashi et al. 2013) Toxicity of nanoparticles appears to be caused both by triggering oxidative stress responses in the target and by dissolution of Al^{3+} (Pakrashi et al. 2013).

Al Toxicity in Humans

Aluminum has been implicated as a possible risk factor in Alzheimer's disease and dialysis dementia in humans (Nayak 2002, Yokel 2000, Krewski et al. 2007). Alzheimer's disease is a multifactorial disease, with environmental and genetic factors playing a role in the pathogenesis (Mutter et al. 2007, Frisardi et al. 2010, Liang 2018).

It has been shown that Al exerts its cytotoxic effects by disrupting lipid membrane fluidity, and disturbing iron (Fe), magnesium, and calcium homeostasis, and causing oxidative stress (Tönnies and Trushina 2017, Mailloux, Lemire, and Appanna 2011, Yokel 2000, Michalke, Halbach, and Nieschwitz 2009). The oxidative stress response triggered by Al may affect mitochondria (Kumar and Gill 2009) and triggering neurodegenerative disorders (Campbell and Bondy 2000).

Parenterally and intramuscularly applied Al is 100% bioavailable (Schafer and Jahreis 2009). By contrast, availability of Al via oral, inhalation, or dermal routes is less than 2% (Schafer and Jahreis 2009). In plasma, Al is bound to transferrin (Schafer and Jahreis 2009) and citric acid (Yokel and McNamara 2001). Most of the Al taken up by humans is excreted in the urine (Yokel and McNamara 2001) and feces (Yokel and McNamara 2001, Schafer and Jahreis 2009). Ionic Al^{3+} may form in lysosomes from Al adjuvants (Mold, Shardlow, and Exley 2016).

Bauxite mining in open-cut mines does not appear to affect the respiratory health of mine workers (Wesdock and Arnold 2014). Bauxite mine and Al refinery workers have no increased risk of cancer (Fritschi et al. 2008, Wesdock and Arnold 2014), but some bauxite ores may contain radioactive elements (radium and thorium), which may pose a long-term radiation exposure risk· In Al smelter workers, exposure to fluoride may affect lung function more than Al itself (Wesdock and Arnold 2014) and no congenital abnormalities have been observed in Al smelter workers (Kiesswetter et al. 2009, Wesdock and Arnold 2014). Likewise, Al welders are not showing increased incidence of neurobehavioral problems (Kiesswetter et al. 2009).

Nanoparticles of Al oxide have been shown to cause neurotoxicity, and they can cross the blood–brain barrier (Jiang, Mashayekhi, and Xing 2009, Sharma et al. 2009). Nano-sized Al is more biotoxic than micro-sized Al (Jiang, Mashayekhi, and Xing 2009, Park et al. 2015, Ates et al. 2015). However, more research is needed into the environmental threats posed by nanoparticles of Al and to determine the threhold levels for various shapes and sizes of Al nanoparticles in humans. Limited studies with human cell cultures in-vitro showed detrimental effects when applied at 50 mg/L (Song et al. 2017).

Conclusion

Aluminum has no known biological benefit to living organisms and is toxic in micromolar concentrations. Since solubility of Al is minimal at circumneutral pH, alleviation of Al toxicity can be achieved by pH adjustment. Complexation of Al by organic acids and humic substances also lowers biotoxicity of Al. While Al toxicity in soil can be easily overcome by incorporating lime into the soil by tilling, new management strategies need to be developed to counteract soil acidification and consequent Al toxicity in zero-till farming systems. The decrease in atmospheric acid inputs (acid rain) over the last two decades has resulted in lower Al solubility and less Al damage to aquatic ecosystems. Uptake of Al in humans through foodstuff and water poses an uncertain risk, but pollution with nano-particulate Al may result in human health problems. The effect of nano-sized Al particles on the functioning of soil and water ecosystems, as well as human health, needs more research.

Acknowledgments

This research was supported under the Australian Research Council's Discovery Projects Funding scheme (project number DP0665467).

References

Adams, M. L., D. J. Hawke, N. H. S. Nilsson, and K. J. Powell. 2000. The relationship between soil solution pH and Al^{3+} concentrations in a range of South Island (New Zealand) soils. *Australian Journal of Soil Research* 38 (1):141–153.

Aguilar, F., H. Autrup, S. Barlow, L. Castle, R. Crebelli, W. Dekant, K. H. Engel, N. Gontard, D. Gott, S. Grilli, R. Gurtler, J. C. Larsen, C. Leclercq, J. C. Leblanc, F. X. Malcata, W. Mennes, M. R. Milana, I. Pratt, I. Rietjens, P. Tobback, F. Toldra, and Members Panel. 2008. Safety of aluminium from dietary intake. *EFSA Journal* 6 (7). doi:10.2903/j.efsa.2008.754.

Agyin-Birikorang, S., and G. A. O'Connor. 2009. Aging effects on reactivity of an aluminum-based drinking-water treatment residual as a soil amendment. *Science of the Total Environment* 407 (2):826–834.

Andren, C. M., and E. Rydin. 2012. Toxicity of inorganic aluminium at spring snowmelt--in-stream bioassays with brown trout (Salmo trutta L.). *Science of the Total Environment* 437:422–32. doi:10.1016/j.scitotenv.2012.08.006.

Asher, C. J., N. J. Grundon, and N. W. Menzies. 2002. *How to Unravel and Solve Soil Fertility Problems*, *ACIAR Monograph No. 83*. Canberra: Australian Centre for International Agricultural Research.

Ates, M., V. Demir, Z. Arslan, J. Daniels, I. O. Farah, and C. Bogatu. 2015. Evaluation of alpha and gamma aluminum oxide nanoparticle accumulation, toxicity, and depuration in *Artemia salina* larvae. *Environmental Toxicology* 30 (1):109–118. doi:10.1002/tox.21917.

Bardy, M., C. Bonhomme, E. Fritsch, J. Maquet, R. Hajjar, T. Allard, S. Derenne, and G. Calas. 2007. Al speciation in tropical podzols of the upper Amazon basin: a solid-state Al-27 MAS and MQMAS NMR study. *Geochimica et Cosmochimica Acta* 71 (13):3211–3222.

Bergholm, J., D. Berggren, and G. Alavi. 2003. Soil acidification induced by ammonium sulphate addition in a Norway spruce forest in Southwest Sweden. *Water Air and Soil Pollution* 148:87–109.

Bertsch, P. M., and D. R. Parker. 1996. Aqueous polynuclear aluminum species. In *The Environmental Chemistry of Aluminum*, edited by G. Sposito, 117–168. Boca Raton: CRC Lewis Publishers.

Blamey, F. P. C., N. K. Nishizawa, and E. Yoshimura. 2004. Timing, magnitude, and location of initial soluble aluminum injuries to mungbean roots. *Soil Science and Plant Nutrition* 50:67–76.

Boudot, J. P., O. Maitat, D. Merlet, and J. Rouiller. 1996. Occurrence of non-monomeric species of aluminium in undersaturated soil and surface waters: consequences for the determination of mineral saturation indices. *Journal of Hydrology* 177 (1–2):47–63.

Brady, D. J., D. G. Edwards, C. J. Asher, and F. P. C. Blamey. 1993. Calcium amelioration of aluminium toxicity effects on root hair development in soybean *Glycine max*. *New Phytologist* 123:531–538.

Brown, T. T., R. T. Koenig, D. R. Huggins, J. B. Harsh, and R. E. Rossi. 2008. Lime effects on soil acidity, crop yield, and aluminum chemistry in direct-seeded cropping systems. *Soil Science Society of America Journal* 72 (3):634–640.

Brown, P. L., R. N. Sylva, G. E. Batley, and J. Ellis. 1985. The hydrolysis of metal ions. Part 8. Aluminium (III). *Journal of the Chemical Society, Dalton Transactions* 1985:1967–1970.

Bush, R. T., D. Fyfe, and L. A. Sullivan. 2004. Occurrence and abundance of monosulfidic black ooze in coastal acid sulfate soil landscapes. *Australian Journal of Soil Research* 42 (5–6):609–616.

Campbell, A., and S. C. Bondy. 2000. Aluminum induced oxidative events and its relation to inflammation: a role for the metal in Alzheimer's disease. *Cellular and Molecular Biology* 46:721–730.

Casey, W. H. 2005. Large aqueous aluminum hydroxide molecules. *Chemical Reviews* 106:1–16.

Cech, I., and J. Montera. 2000. Spatial variations in total aluminum concentrations in drinking water supplies studied by geographic information system (GIS) methods. *Water Research* 34 (10):2703–2712. doi:10.1016/s0043-1354(00)00026-9.

Chen, R. F., R. F. Shen, P. Gu, H. Y. Wang, and X. H. Xu. 2008. Investigation of aluminum-tolerant species in acid soils of South China. *Communications in Soil Science and Plant Analysis* 39 (9–10):1493–1506.

Chen, Y.-T., Y. Wang, and K.-C. Yeh. 2017. Role of root exudates in metal acquisition and tolerance. *Current Opinion in Plant Biology* 39:66–72. doi:10.1016/j.pbi.2017.06.004.

Conyers, M. K., D. P. Heenan, W. J. McGhie, and G. P. Poile. 2003. Amelioration of acidity with time by limestone under contrasting tillage. *Soil & Tillage Research* 72 (1):85–94. doi:10.1016/s0167-1987(03)00064-3.

Dai, Z. M., X. J. Zhang, C. Tang, N. Muhammad, J. J. Wu, P. C. Brookes, and J. M. Xu. 2017. Potential role of biochars in decreasing soil acidification – a critical review. *Science of the Total Environment* 581:601–611. doi:10.1016/j.scitotenv2016.12.169.

de Vries, W., G. J. Reinds, and E. Vel. 2003. Intensive monitoring of forest ecosystems in Europe 2: atmospheric deposition and its impacts on soil solution chemistry. *Forest Ecology and Management* 174 (1–3):97–115. doi:10.1016/s0378-1127(02)00030-0.

Dent, D. L., and L. J. Pons. 1995. A world perspective on acid sulfate soils. *Geoderma* 67 (3–4):263–276.

Doshi, R., W. Braida, C. Christodoulatos, M. Wazne, and G. O'Connor. 2008. Nano-aluminum: transport through sand columns and environmental effects on plants and soil communities. *Environmental Research* 106 (3):296–303.

Dsa, J. V., K. S. Johnson, D. Lopez, C. Kanuckel, and J. Tumlinson. 2008. Residual toxicity of acid mine drainage-contaminated sediment to stream macroinvertebrates: relative contribution of acidity vs. metals. *Water Air and Soil Pollution* 194 (1–4):185–197.

Dubikova, M., P. Cambier, V. Sucha, and M. Caplovicova. 2002. Experimental soil acidification. *Applied Geochemistry* 17 (3):245–257. doi:10.1016/s0883-2927(01)00081-6.

Ertl, K., and W. Goessler. 2018. Aluminium in foodstuff and the influence of aluminium foil used for food preparation or short term storage. *Food Additives and Contaminants: Part B* 11:153–159.

Etou, A., S. Q. Bai, T. Saito, H. Noma, Y. Okaue, and T. Yokoyama. 2009. Formation conditions and stability of a toxic tridecameric Al polymer under a soil environment. *Journal of Colloid and Interface Science* 337 (2):606–609.

Exley, C., P. Siesjo, and H. Eriksson. 2010. The immunobiology of aluminium adjuvants: how do they really work? *Trends in Immunology* 31:103–109.

Faltmarsch, R. M., M. E. Astrom, and K. M. Vuori. 2008. Environmental risks of metals mobilised from acid sulphate soils in Finland: a literature review. *Boreal Environment Research* 13 (5):444–456.

Fest, E. P. M. J., E. J. M. Temminghoff, J. Griffioen, B. Van Der Grift, and W. H. Van Riemsdijk. 2007. Groundwater chemistry of Al under Dutch sandy soils: effects of land use and depth. *Applied Geochemistry* 22 (7):1427–1438.

Flarend, R., T. Bin, S. D. Elmore, and S. L. Hem. 2001. A preliminary study of the dermal absorption of aluminium from antiperspirants using aluminium-26. *Food and Chemical Toxicology* 39:163–168.

Frisardi, V., V. Solfrizzi, C. Capurso, P. G. Kehoe, B. P. Imbimbo, A. Santamato, F. Dellegrazie, D. Seripa, A. Pilotto, A. Capurso, and F. Panza. 2010. Aluminum in the diet and Alzheimer's disease: from current epidemiology to possible disease-modifying treatment. *Journal of Alzheimers Disease* 20:17–30.

Fritschi, L., J. L. Hoving, M. R. Sim, A. Del Monac, E. MacFarlane, D. McKenzie, G. Benke, and N. de Klerk. 2008. All cause mortality and incidence of cancer in workers in bauxite mines and alumina refineries. *International Journal of Cancer* 123 (4):882–887.

Fung, K. F., H. P. Carr, B. H. T. Poon, and M. H. Wong. 2009. A comparison of aluminum levels in tea products from Hong Kong markets and in varieties of tea plants from Hong Kong and India. *Chemosphere* 75 (7):955–962.

Furrer, G., M. Gfeller, and B. Wehrli. 1999. On the chemistry of the Keggin Al_{13} polymer: kinetics of proton-promoted decomposition. *Geochimica et Cosmochimica Acta* 63:3069–3076.

Furrer, G., B. Trusch, and C. Muller. 1992. The formation of polynuclear aluminum under simulated natural conditions. *Geochimica et Cosmochimica Acta* 56:3831–3838.

Gensemer, R. W., and R. C. Playle. 1999. The bioavailability and toxicity of aluminum in aquatic environments. *Critical Reviews in Environmental Science and Technology* 29 (4):315–450.

Gerard, F., J. P. Boudot, and J. Ranger. 2001. Consideration on the occurrence of the Al-13 polycation in natural soil solutions and surface waters. *Applied Geochemistry* 16 (5):513–529.

Godsey, C. B., G. M. Pierzynski, D. B. Mengel, and R. E. Lamond. 2007. Management of soil acidity in no-till production systems through surface application of lime. *Agronomy Journal* 99 (3):764–772. doi:10.2134/agronj2006.0078.

Golding, Lisa A., Brad M. Angel, Graeme E. Batley, Simon C. Apte, Rick Krassoi, and Chris J. Doyle. 2015. Derivation of a water quality guideline for aluminium in marine waters. *Environmental Toxicology and Chemistry* 34 (1):141–151. doi:10.1002/etc.2771.

Gourier-Frery, C., and N. Frery. 2004. Aluminium. *EMC-Toxicologie Pathologie* 1:79–95.

Griffitt, R. J., A. Feswick, R. Weil, K. Hyndman, P. Carpinone, K. Powers, N. D. Denslow, and D. S. Barber. 2011. Investigation of acute nanoparticulate aluminum toxicity in zebrafish. *Environmental Toxicology* 26 (5):541–551. doi:10.1002/tox.20669.

Griffitt, R. J., J. Luo, J. Gao, J. C. Bonzongo, and D. S. Barber. 2008. Effects of particle composition and species on toxicity of metallic nanomaterials in aquatic organisms. *Environmental Toxicology and Chemistry* 27 (9):1972–1978.

Guo, J. H., X. J. Liu, Y. Zhang, J. L. Shen, W. X. Han, W. F. Zhang, P. Christie, K. W. T. Goulding, P. M. Vitousek, and F. S. Zhang. 2010. Significant acidification in major Chinese croplands. *Science* 327:1008–1010.

Guo, J. H., X. S. Zhang, R. D. Vogt, J. S. Xiao, D. W. Zhao, R. J. Xiang, and J. H. Luo. 2007. Evaluating main factors controlling aluminum solubility in acid forest soils, southern and southwestern China. *Applied Geochemistry* 22:388–396.

Hicks, W. S., G. M. Bowman, and R. W. Fitzpatrick. 1999. Environmental impact of acid sulfate soils near Cairns, QLD. In *CSIRO Land and Water Technical Report 15/99*. Canberra: Commonwealth Scientific and Industrial Research Organisation.

Hiradate, S., S. Taniguchi, and K. Sakurai. 1998. Aluminum speciation in aluminum-silica solutions and potassium chloride extracts of acidic soils. *Soil Science Society of America Journal* 62 (3):630–636.

Hodson, M. J., and A. G. Sangster. 1999. Aluminium/silicon interactions in conifers. *Journal of Inorganic Biochemistry* 76 (2):89–98. doi:10.1016/s0162-0134(99)00119-1.

Horst, W. J., Y. Wang, and D. Eticha. 2010. The role of the root apoplast in aluminium-induced inhibition of root elongation and in aluminium resistance of plants: a review. *Annals of Botany* 106:185–197.

Hunter, D., and D. S. Ross. 1991. Evidence for a phytotoxic hydroxy-aluminum polymer in organic soil horizons. *Science* 251 (4997):1056–1058.

Illmer, P., and F. Schinner. 1997. Influence of aluminum on motility and swarming of *Pseudomonas* sp. and *Arthrobacter* sp. *FEMS Microbiology Letters* 155 (1):121–124.

Jaiswal, S. K., J. Naamala, and F. D. Dakora. 2018. Nature and mechanisms of aluminium toxicity, tolerance and amelioration in symbiotic legumes and rhizobia. *Biology and Fertility of Soils* 54 (3):309–318. doi:10.1007/s00374-018-1262-0.

Jiang, W., H. Mashayekhi, and B. S. Xing. 2009. Bacterial toxicity comparison between nano- and micro-scaled oxide particles. *Environmental Pollution* 157 (5):1619–1625.

Kabata-Pendias, A., and H. Kabata. 2001. *Trace elements in soils and plants*. 3rd ed. Boca Raton: CRC Press.

Kerven, G. L., P. L. Larsen, and F. P. C. Blamey. 1995. Detrimental sulfate effects on formation of Al-13 tridecameric polycation in synthetic soil solutions. *Soil Science Society of America, Journal* 59:765–771.

Kiesswetter, E., M. Schaper, M. Buchta, K. H. Schaller, B. Rossbach, T. Kraus, and S. Letzel. 2009. Longitudinal study on potential neurotoxic effects of aluminium: II. Assessment of exposure and neurobehavioral performance of Al welders in the automobile industry over 4 years. *International Archives of Occupational and Environmental Health* 82 (10):1191–1210.

Kinraide, T. B. 1991. Identity of the rhizotoxic aluminium species. In *Plant-soil interactions at low pH*, edited by R. J. Wright, V. C. Baligar and R. P. Murrmann, 717–728. Dordrecht: Kluwer Academic Publishers.

Kinraide, T. B. 1997. Reconsidering the rhizotoxicity of hydroxyl, sulphate, and fluoride complexes of aluminium. *Journal of Experimental Botany* 48 (310):1115–1124.

Kinraide, T. B., and B. K. Sweeney. 2003. Proton alleviation of growth inhibition by toxic metals (Al, La, Cu) in rhizobia. *Soil Biology & Biochemistry* 35 (2):199–205.

Kopittke, P. M., N. W. Menzies, and F. P. C. Blamey. 2004. Rhizotoxicity of aluminate and polycationic aluminium at high pH. *Plant and Soil* 266 (1–2):177–186.

Kopittke, P. M., K. L. Moore, E. Lombi, A. Gianoncelli, B. J. Ferguson, F. P. C. Blamey, N. W. Menzies, T. M. Nicholson, B. A. McKenna, P. Wang, P. M. Gresshoff, G. Kourousias, R. I. Webb, K. Green, and A. Tollenaere. 2015. Identification of the primary lesion of toxic aluminum in plant roots. *Plant Physiology* 167 (4):1402–1411. doi:10.1104/pp.114.253229.

Krewski, D., R. A. Yokel, E. Nieboer, D. Borchelt, J. Cohen, J. Harry, S. Kacew, J. Lindsay, A. M. Mahfouz, and V. Rondeau. 2007. Human health risk assessment for aluminium, aluminium oxide, and aluminium hydroxide. *Journal of Toxicology and Environmental Health-Part B-Critical Reviews* 10:1–269.

Kumar, V., and K. D. Gill. 2009. Aluminium neurotoxicity: neurobehavioural and oxidative aspects. *Archives of Toxicology* 83 (11):965–978.

Lange, H., S. Solberg, and N. Clarke. 2006a. Aluminum dynamics in forest soil waters in Norway. *Science of the Total Environment* 367:942–957.

Lange, H., S. Solberg, and N. Clarke. 2006b. Aluminum dynamics in forest soil waters in Norway. *Science of the Total Environment* 367 (2–3):942–957. doi:10.1016/j.scitotenv.2006.01.033.

Lapenis, A. G., et al. 2004. Acidification of forest soil in Russia: From 1893 to present. *Global Biogeochemical Cycles* 18(1): GB1037.

Lawrence, G. B. 2002. Persistent episodic acidification of streams linked to acid rain effects on soil. *Atmospheric Environment* 36 (10):1589–1598. doi:10.1016/s1352-2310(02)00081-x.

Lesturgez, G., R. Poss, A. Noble, O. Grunberger, W. Chintachao, and D. Tessier. 2006. Soil acidification without pH drop under intensive cropping systems in Northeast Thailand. *Agriculture Ecosystems & Environment* 114 (2–4):239–248.

Li, J. Y., Z. D. Liu, W. Z. Zhao, M. M. Masud, and R. K. Xu. 2015. Alkaline slag is more effective than phosphogypsum in the amelioration of subsoil acidity in an Ultisol profile. *Soil & Tillage Research* 149:21–32. doi:10.1016/j.still.2014.12.017.

Liang, R. 2018. Cross talk between aluminum and genetic susceptibility and epigenetic modification in Alzheimer's disease. In *Neurotoxicity of Aluminum*, edited by Qiao Niu, 173–191. Singapore: Springer Singapore.

Liang, H. C., and B. M. Thomson. 2009. Minerals and mine drainage. *Water Environment Research* 81 (10):1615–1663.

Liu, J., and N. V. Hue. 2001. Amending subsoil acidity by surface applications of gypsum, lime, and composts. *Communications in Soil Science and Plant Analysis* 32 (13–14):2117–2132. doi:10.1081/css-120000273.

Lofts, S., C. Woof, E. Tipping, N. Clarke, and J. Mulder. 2001. Modelling pH buffering and aluminium solubility in European forest soils. *European Journal of Soil Science* 52 (2):189–204.

Lopez-Diaz, M. L., M. R. Mosquera-Losada, and A. Rigueiro-Rodriguez. 2007. Lime, sewage sludge and mineral fertilization in a silvopastoral system developed in very acid soils. *Agroforestry Systems* 70 (1):91–101. doi:10.1007/s10457-007-9046-9.

Machacha, S. 2004. Comparison of laboratory pH buffer methods for predicting lime requirement for acidic soils of eastern Botswana. *Communications in Soil Science and Plant Analysis* 35 (17–18):2675–2687. doi:10.1081/lcss-200030441.

Mahdy, A. M., E. A. Elkhatib, N. O. Fathi, and Z. Q. Lin. 2009. Effects of co-application of biosolids and water treatment residuals on corn growth and bioavailable phosphorus and aluminum in alkaline soils in Egypt. *Journal of Environmental Quality* 38 (4):1501–1510.

Mailloux, R. J., J. Lemire, and V. D. Appanna. 2011. Hepatic response to aluminum toxicity: dyslipidemia and liver diseases. *Experimental Cell Research* 317 (16):2231–2238. doi:10.1016/j.yexcr.2011.07.009.

Manoharan, V., P. Loganathan, R. W. Tillman, and R. L. Parfitt. 2007. Interactive effects of soil acidity and fluoride on soil solution aluminium chemistry and barley (*Hordeum vulgare* L.) root growth. *Environmental Pollution* 145:778–786.

Martell, A. E., R. D. Hancock, R. M. Smith, and R. J. Motekaitis. 1996. Coordination of Al(III) in the environment and in biological systems. *Coordination Chemistry Reviews* 149:311–328.

Martin, R. B. 1996. Ternary complexes of Al^{3+} and F^- with a third ligand. *Coordination Chemistry Reviews* 141:23–32.

Martinent-Catalot, V., J.-M. Lamerant, G. Tilmant, M.-S. Bacou, and J. P. Ambrosi. 2002. Bauxaline: a new product for various applications of Bayer process red mud. *Light Metals* 2002:125–131.

Masion, A., and P. M. Bertsch. 1997. Aluminium speciation in the presence of wheat root cell walls: a wet chemical study. *Plant, Cell and Environment* 20:504–512.

Masion, A., F. Thomas, D. Tchoubar, J. Y. Bottero, and P. Tekely. 1994. Chemistry and structure of Al(OH)/organic precipitates. A small-angle X-ray scattering study. 3. Depolymerization of the Al_{13} polycation by organic ligands. *Langmuir* 10:4353–4356.

Materechera, S. A., and T. S. Mkhabela. 2002. The effectiveness of lime, chicken manure and leaf litter ash in ameliorating acidity in a soil previously under black wattle (Acacia mearnsii) plantation. *Bioresource Technology* 85 (1):9–16. doi:10.1016/s0960-8524(02)00065-2.

Matthias, A., M. Maurer, and H. Parlar. 2003. Comparative aluminium speciation and quantification in soil solutions of two different forest ecosystems by Al-27-NMR. *Fresenius Environmental Bulletin* 12 (10):1263–1275.

McBride, M. B. 1994. *Environmental chemistry of soils*. New York: Oxford University Press.

Menzies, N. W., L. C. Bell, and D. G. Edwards. 1994. Exchange and solution phase chemistry of acid, highly weathered soils. 2. Investigations of mechanisms controlling Al release into solution. *Australian Journal of Soil Research* 32:269–283.

Michalke, B., S. Halbach, and V. Nieschwitz. 2009. JEM Spotlight: metal speciation related to neurotoxicity in humans. *Journal of Environmental Monitoring* 11:939–954.

Mishra, V. K., and A. Kumar. 2009. Impact of metal nanoparticles on the plant growth promoting *Rhizobacteria*. *Digest Journal of Nanomaterials and Biostructures* 4 (3):587–592.

Miyazawa, M., T. Takahashi, T. Sato, H. Kanno, and M. Nanzyo. 2013. Factors controlling accumulation and decomposition of organic carbon in humus horizons of Andosols. *Biology and Fertility of Soils* 49:929–938.

Mokolobate, M. S., and R. J. Haynes. 2002. Comparative liming effect of four organic residues applied to an acid soil. *Biology and Fertility of Soils* 35 (2):79–85. doi:10.1007/s00374-001-0439-z.

Mold, M., E. Shardlow, and C. Exley. 2016. Insight into the cellular fate and toxicity of aluminium adjuvants used in clinically approved human vaccinations. *Scientific Reports* 6:31578. doi:10.1038/srep31578.

Morikawa, C. K., and M. Saigusa. 2002. Si amelioration of Al toxicity in barley (*Hordeum vulgare* L.) growing in two Andosols. *Plant and Soil* 240 (1):161–168. doi:10.1023/a:1015804401190.

Mota, D., F. Faria, E. A. Gomes, I. E. Marriel, E. Paiva, and L. Seldin. 2008. Bacterial and fungal communities in bulk soil and rhizospheres of aluminum-tolerant and aluminum-sensitive maize (*Zea mays* L.) lines cultivated in unlimed and limed Cerrado soil. *Journal of Microbiology and Biotechnology* 18 (5):805–814.

Mueller, M., M. Anke, and H. Illing-Guenther. 1998. Aluminium in foodstuffs. *Food Chemistry* 61:419–428.

Mutter, J., J. Naumann, R. Schneider, and H. Walach. 2007. Mercury and Alzheimer's disease. *Fortschritte der Neurologie und Psychiatrie* 75:528–540.

Naramabuye, F. X., and R. J. Haynes. 2006. Effect of organic amendments on soil pH and AI solubility and use of laboratory indices to predict their liming effect. *Soil Science* 171 (10):754–763. doi:10.1097/01.ss.0000228366.17459.19.

Nayak, P. 2002. Aluminum: impacts and disease. *Environmental Research* 89 (2):101–115.

Nelson, D. R., and P. M. Mele. 2006. The impact of crop residue amendments and lime on microbial community structure and nitrogen-fixing bacteria in the wheat rhizosphere. *Australian Journal of Soil Research* 44 (4):319–329. doi:10.1071/sr06022.

NHMRC. 2004. Australian drinking water guidelines 6. National water quality management strategy. Canberra, ACT, Australian Government - National Health and Medical Research Council. 615p.

Oladeji, O. O., J. B. Sartain, and G. A. O'Connor. 2009. Land application of aluminum water treatment residual: aluminum phytoavailability and forage yield. *Communications in Soil Science and Plant Analysis* 40 (9–10):1483–1498.

Orvig, C. 1993. The aqueous coordination chemistry of aluminum. In *Coordination chemistry of aluminum*, edited by G. H. Robinson, 85–121. New York: VCH.

Othman, M. N., M. P. Abdullah, and Y. F. Abd Aziz. 2010. Removal of aluminium from drinking water. *Sains Malaysiana* 39 (1):51–55.

Pakrashi, S., S. Dalai, A. Humayun, S. Chakravarty, N. Chandrasekaran, and A. Mukherjee. 2013. *Ceriodaphnia dubia* as a potential bio-indicator for assessing acute aluminum oxide nanoparticle toxicity in fresh water environment. *PLoS ONE* 8 (9). doi:10.1371/journal.pone.0074003.

Panhwar, Q. A., U. A. Naher, J. Shamshuddin, R. Othman, and M. R. Ismail. 2016. Applying limestone or basalt in combination with bio-fertilizer to sustain rice production on an acid sulfate soil in Malaysia. *Sustainability* 8 (7). doi:10.3390/su8070700.

Park, E. J., G. H. Lee, J. H. Shim, M. H. Cho, B. S. Lee, Y. B. Kim, J. H. Kim, Y. Kim, and D. W. Kim. 2015. Comparison of the toxicity of aluminum oxide nanorods with different aspect ratio. *Archives of Toxicology* 89 (10):1771–1782. doi:10.1007/s00204-014-1332-5.

Poborilova, Z., R. Opatrilova, and P. Babula. 2013. Toxicity of aluminium oxide nanoparticles demonstrated using a BY-2 plant cell suspension culture model. *Environmental and Experimental Botany* 91:1–11. doi:10.1016/j.envexpbot.2013.03.002.

Poschenrieder, C., B. Gunse, I. Corrales, and J. Barcelo. 2008. A glance into aluminum toxicity and resistance in plants. *Science of the Total Environment* 400 (1–3):356–368.

Pu, X. X., O. Vazquez, J. D. Monnell, and R. D. Neufeld. 2010. Speciation of aluminum precipitates from acid rock discharges in Central Pennsylvania. *Environmental Engineering Science* 27 (2):169–180.

Qian, L. B., B. L. Chen, and D. F. Hu. 2013. Effective alleviation of aluminum phytotoxicity by manure-derived biochar. *Environmental Science & Technology* 47 (6):2737–2745. doi:10.1021/es3047872.

Qin, R. J., and F. X. Chen. 2005. Amelioration of aluminum toxicity in red soil through use of barnyard and green manure. *Communications in Soil Science and Plant Analysis* 36 (13–14):1875–1889. doi:10.1081/css-200062480.

Raboin, L. M., A. H. D. Razafimahafaly, M. B. Rabenjarisoa, B. Rabary, J. Dusserre, and T. Becquer. 2016. Improving the fertility of tropical acid soils: Liming versus biochar application? A long term comparison in the highlands of Madagascar. *Field Crops Research* 199:99–108. doi:10.1016/j.fcr.2016.09.005.

Ramirez, M. D. A., J. D. Silva, N. Ohkama-Ohtsu, and T. Yokoyama. 2018. In vitro rhizobia response and symbiosis process under aluminum stress. *Canadian Journal of Microbiology* 64:511–526.

Ranau, R., J. Oehlenschlager, and H. Steinhart. 2001. Aluminium levels of fish fillets baked and grilled in aluminium foil. *Food Chemistry* 73 (1):1–6. doi:10.1016/s0308-8146(00)00318-6.

Rasmussen, C., R. J. Southard, and W. R. Horwath. 2006. Mineral control of organic carbon mineralization in a range of temperate conifer forest soils. *Global Change Biology* 12:834–847.

Resources-and-Energy-Quarterly. 2018. https://archive.industry.gov.au/Office-of-the-Chief-Economist/Publications/ResourcesandEnergyQuarterlyMarch2018/documents/Resources-and-Energy-Quarterly-March-2018-Aluminium-alumina-and-bauxite.pdf.

Ryan, P. R., E. Delhaize, and P. J. Randall. 1995. Malate efflux from root apices and tolerance to aluminium are highly correlated in wheat. *Australian Journal of Plant Physiology* 22:531–536.

Sadiq, I. M., S. Pakrashi, N. Chandrasekaran, and A. Mukherjee. 2011. Studies on toxicity of aluminum oxide (Al2O3) nanoparticles to microalgae species: Scenedesmus sp and Chlorella sp. *Journal of Nanoparticle Research* 13 (8):3287–3299. doi:10.1007/s11051-011-0243-0.

Salifoglou, A. 2002. Synthetic and structural carboxylate chemistry of neurotoxic aluminum in relevance to human diseases. *Coordination Chemistry Reviews* 228:297–317.

Sangster, A.G., and Hodson, M.J. 2001. Silicon and aluminium codeposition in the cell wall phytoliths of gymnosperm leaves. *In Phytoliths - Applications in earth science and human history* (eds. Meunier, J.D., Colin F.). A.A. Balkema, Lisse, The Netherlands. pp. 343–355.

Sanjuan, B., and G. Michard. 1987. Aluminum hydroxide solubility in aqueous solutions containing fluoride ions at 50°C. *Geochimica et Cosmochimica Acta* 51:1823–1831.

Sarpola, A. T., V. K. Hietapelto, J. E. Jalonen, J. Jokela, and J. H. Ramo. 2006. Comparison of hydrolysis products of AlCl3.6H2O in different concentrations by electrospray ionization time of flight mass spectrometer (ESI TOF MS). *International Journal for Environmental Analytical Chemistry* 86:1007–1018.

Schaefer, U., and M. Seifert. 2006. Oral intake of aluminum from foodstuffs, food additives, food packaging, cookware and pharmaceutical preparations with respect to dietary regulations. *Trace Elements and Electrolytes* 23:150–161.

Schafer, U., and G. Jahreis. 2009. Update on regulations of aluminium intake - biochemical and toxicological assessment. *Trace Elements and Electrolytes* 26:95–99.

Scheel, T., C. Dorfler, and K. Kalbitz. 2007. Precipitation of dissolved organic matter by aluminum stabilizes carbon in acidic forest soils. *Soil Science Society of America Journal* 71:64–74.

Scheel, T., K. Pritsch, M. Schlater, and K. Kalbitz. 2008. Precipitation of enzymes and organic matter by aluminum-impacts on carbon mineralization. *Journal of Plant Nutrition and Soil Science-Zeitschrift für Pflanzenernährung und Bodenkunde* 171 (6):900–907.

Schussler, U., W. Balzer, and A. Deeken. 2005. Dissolved Al distribution, particulate Al fluxes and coupling to atmospheric Al and dust deposition in the Arabian Sea. *Deep-Sea Research Part Ii-Topical Studies in Oceanography* 52 (14–15):1862–1878. doi:10.1016/j.dsr2.2005.06.005.

Scott, B. J., I. G. Fenton, A. G. Fanning, W. G. Schumann, and L. J. C. Castleman. 2007. Surface soil acidity and fertility in the eastern Riverina and western slopes of southern New South Wales. *Australian Journal of Experimental Agriculture* 47 (8):949–964.

Shabnam, N., and H. Kim. 2018. Non-toxicity of nano alumina: a case on mung bean seedlings. *Ecotoxicology and Environmental Safety* 165:423–433. doi:10.1016/j.ecoenv.2018.09.033.

Sharma, H. S., S. F. Ali, S. M. Hussain, J. J. Schlager, and A. Sharma. 2009. Influence of engineered nanoparticles from metals on the blood-brain barrier permeability, cerebral blood flow, brain edema and neurotoxicity. An experimental study in the rat and mice using biochemical and morphological approaches. *Journal of Nanoscience and Nanotechnology* 9 (8):5055–5072.

Shen, J. S., and V. Nardello-Rataj. 2009. Deodorants and antiperspirants: chemistry under arms. *Actualite Chimique* (331):8–18.

Shi, R. Y., J. Y. Li, N. Ni, K. Mehmood, R. K. Xu, and W. Qian. 2017. Effects of biomass ash, bone meal, and alkaline slag applied alone and combined on soil acidity and wheat growth. *Journal of Soils and Sediments* 17 (8):2116–2126. doi:10.1007/s11368-017-1673-9.

Shirokikh, I. G., A. A. Shirokikh, N. A. Rodina, L. M. Polyanskaya, and O. A. Burkanova. 2004. Effects of soil acidity and aluminum on the structure of microbial biomass in the rhizosphere of barley. *Eurasian Soil Science* 37 (8):839–843.

Shoba, V. N., and K. V. Chudnenko. 2014. Ion exchange properties of humus acids. *Eurasian Soil Science* 47 (8):761–771. doi:10.1134/s1064229314080110.

Simonsson, M. 2000. Interactions of aluminium and fulvic acid in moderately acid solutions: stoichiometry of the H^+/Al^{3+} exchange. *European Journal of Soil Science* 51:655–666.

Singh, S., D. K. Tripathi, S. Singh, S. Sharma, N. K. Dubey, D. K. Chauhan, and M. Vaculík. 2017. Toxicity of aluminium on various levels of plant cells and organism: A review. *Environmental and Experimental Botany* 137:177–193. doi:10.1016/j.envexpbot.2017.01.005.

Sivaguru, M., and W. J. Horst. 1998. The distal part of the transition zone is the most aluminum-sensitive apical root zone of maize. *Plant Physiology* 116:155–163.

Sjostedt, C., T. Wallstedt, J. P. Gustafsson, and H. Borg. 2009. Speciation of aluminium, arsenic and molybdenum in excessively limed lakes. *Science of the Total Environment* 407 (18):5119–5127.

Skjelkvale, B. L., K. Torseth, W. Aas, and T. Andersen. 2001. Decrease in acid deposition – Recovery in Norwegian waters. *Water Air and Soil Pollution* 130 (1–4):1433–1438. doi:10.1023/a:1013956829092.

Skyllberg, U. 1999. pH and solubility of aluminium in acidic forest soils: a consequence of reactions between organic acidity and aluminium alkalinity. *European Journal of Soil Science* 50:95–106.

Song, Z. M., H. Tang, X. Y. Deng, K. Xiang, A. N. Cao, Y. F. Liu, and H. F. Wang. 2017. Comparing toxicity of alumina and zinc oxide nanoparticles on the human intestinal epithelium in vitro model. *Journal of Nanoscience and Nanotechnology* 17 (5):2881–2891. doi:10.1166/jnn.2017.13056.

Soucek, D. J., D. S. Cherry, and C. E. Zipper. 2003. Impacts of mine drainage and other nonpoint source pollutants on aquatic biota in the upper Powell River system, Virginia. *Human and Ecological Risk Assessment* 9 (4):1059–1073.

Srinivasan, P. T., and T. Viraraghavan. 2002. Characterisation and concentration profile of aluminium during drinking-water treatment. *Water SA* 28 (1):99–106.

Stahl, T., S. Falk, H. Taschan, B. Boschek, and H. Brunn. 2018. Evaluation of human exposure to aluminum from food and food contact materials. *European Food Research and Technology* 244 (12):2077–2084. doi:10.1007/s00217-018-3124-2.

Stauber, J. L., T. M. Florence, C. M. Davies, M. S. Adams, and S. J. Buchanan. 1999. Bioavailability of Al in alum-treated drinking water. *Journal American Water Works Association* 91 (11):84–93.

Steiner, C., W. G. Teixeira, J. Lehmann, T. Nehls, J. L. V. de Macedo, W. E. H. Blum, and W. Zech. 2007. Long term effects of manure, charcoal and mineral fertilization on crop production and fertility on a highly weathered Central Amazonian upland soil. *Plant and Soil* 291 (1–2):275–290. doi:10.1007/s11104-007-9193-9.

Strigul, N., L. Vaccari, C. Galdun, M. Wazne, X. Liu, C. Christodoulatos, and K. Jasinkiewicz. 2009. Acute toxicity of boron, titanium dioxide, and aluminum nanoparticles to *Daphnia magna* and *Vibrio fischeri*. *Desalination* 248 (1–3):771–782.

Sun, B., R. Poss, R. Moreau, A. Aventurier, and P. Fallavier. 2000. Effect of slaked lime and gypsum on acidity alleviation and nutrient leaching in an acid soil from Southern China. *Nutrient Cycling in Agroecosystems* 57 (3):215–223. doi:10.1023/a:1009870308097.

Takahashi, T., and R. A. Dahlgren. 2016. Nature, properties and function of aluminum-humus complexes in volcanic soils. *Geoderma* 263:110–121.

Takahashi, T., A. Mitamura, T. Ito, K. Ito, M. Nanzyo, and M. Saigusa. 2008. Aluminum solubility of strongly acidified allophanic Andosols from Kagoshima prefecture, southern Japan. *Soil Science and Plant Nutrition* 54 (3):362–368.

Tang, Y., H. Zhang, J. L. Schroder, M. E. Payton, and D. Zhou. 2007. Animal manure reduces aluminum toxicity in an acid soil. *Soil Science Society of America Journal* 71 (6):1699–1707. doi:10.2136/sssaj2007.0008.

Tarkalson, D. D., G. W. Hergert, W. B. Stevens, D. L. McCallister, and S. D. Kackman. 2005. Fly, ash as a liming material for corn production. *Soil Science* 170 (5):386–398. doi:10.1097/01.ss.0000169910.25356.3a.

Tönnies, E., and E. Trushina. 2017. Oxidative stress, synaptic dysfunction, and Alzheimer's disease. *Journal of Alzheimer's Disease* 57 (4):1105–1121.

Trenfield, Melanie A., Scott J. Markich, Jack C. Ng, Barry Noller, and Rick A. van Dam. 2012. Dissolved organic carbon reduces the toxicity of aluminum to three tropical freshwater organisms. *Environmental Toxicology and Chemistry* 31 (2):427–436. doi:10.1002/etc.1704.

Turhan, S. 2006. Aluminium contents in baked meats wrapped in aluminium foil. *Meat Science* 74 (4):644–647. doi:10.1016/j.meatsci.2006.03.031.

Vieira, F. C. B., Z. L. He, C. Bayer, P. J. Stoffella, and V. C. Baligar. 2008. Organic amendment effects on the transformation and fractionation of aluminum in acidic sandy soil. *Communications in Soil Science and Plant Analysis* 39 (17–18):2678–2694. doi:10.1080/00103620802358813.

von Uexküll, H. R., and E. Mutert. 1995. Global extent, development and economic impact of acid soils. In *Plant soil interactions at low pH*, edited by R. A. Date, N. J. Grundon, G. E. Rayment and M. E. Probert, 5–19. Dordrecht: Kluwer Academic.

Vuorenmaa, J., A. Augustaitis, B. Beudert, W. Bochenek, N. Clarke, H. A. de Wit, T. Dirnbock, J. Frey, H. Hakola, S. Kleemola, J. Kobler, P. Kram, A. J. Lindroos, L. Lundin, S. Lofgren, A. Marchettom, T. Pecka, H. Schulte-Bisping, K. Skotak, A. Srybny, J. Szpikowski, L. Ukonmaanaho, M. Vana, S. Akerblom, and M. Forsius. 2018. Long-term changes (1990–2015) in the atmospheric deposition and runoff water chemistry of sulphate, inorganic nitrogen and acidity for forested catchments in Europe in relation to changes in emissions and hydrometeorological conditions. *Science of the Total Environment* 625:1129–1145. doi:10.1016/j.scitotenv.2017.12.245.

Wang, M., and M. Muhammed. 1999. Novel synthesis of Al_{13}-cluster based alumina materials. *NanoStructured Materials* 11:1219–1229.

Wang, P., et al. 2007. Study of aluminium distribution and speciation in atmospheric particles of different diameters in Nanjing, China. *Atmospheric Environment* 41(27): 5788–5796.

Ward, R. J. S., C. R. McCrohan, and K. N. White. 2006. Influence of aqueous aluminium on the immune system of the freshwater crayfish *Pacifasticus leniusculus*. *Aquatic Toxicology* 77 (2):222–228.

Waters, A. S., and J. G. Webster-Brown. 2013. Assessing aluminium toxicity in streams affected by acid mine drainage. *Water Sci Technol* 67 (8):1764–72. doi:10.2166/wst.2013.051.

Wehr, J. B., F. P. C. Blamey, and N. W. Menzies. 2007. Aluminum. In *Encyclopedia of Soil Science*, edited by R. Lal, 1–6. New York: Taylor and Francis.

Wehr, J. B., N. W. Menzies, and I. Fulton. 2006. Revegetation strategies for bauxite refinery residue: a case study of Alcan Gove in Northern Territory, Australia. *Environmental Management* 37:297–306.

Wesdock, James C., and Ian M. F. Arnold. 2014. Occupational and environmental health in the aluminum industry: key points for health practitioners. *Journal of occupational and environmental medicine* 56 (5 Suppl):S5–S11. doi:10.1097/JOM.0000000000000071.

Willhite, C. C., G. L. Ball, and C. J. McLellan. 2012. Total allowable concentrations of monomeric inorganic aluminum and hydrated aluminum silicates in drinking water. *Critical Reviews in Toxicology* 42 (5):358–442. doi:10.3109/10408444.2012.674101.

Winterbourn, M. J., W. F. McDiffett, and S. J. Eppley. 2000. Aluminium and iron burdens of aquatic biota in New Zealand streams contaminated by acid mine drainage: effects of trophic level. *Science of the Total Environment* 254 (1):45–54.

World-Aluminum.org. 2019. www.world-aluminium.org/statistics/primary-aluminium-production.

Xia, H., and G. D. Rayson. 1998. Investigation of aluminum binding to a *Datura innoxia* material using ^{27}Al NMR. *Environmental Science and Technology* 32:2688–2692.

Yamaguchi, N. U., S. Hiradate, M. Mizoguchi, and T. Miyazaki. 2003. Formation and disappearance of Al tridecamer in the presence of low molecular weight organic ligands. *Soil Science and Plant Nutrition* 49:551–556.

Yamaguchi, N., S. Hiradate, M. Mizoguchi, and T. Miyazaki. 2004. Disappearance of aluminum tridecamer from hydroxyaluminum solution in the presence of humic acid. *Soil Science Society of America Journal* 68 (6):1838–1843. doi:10.2136/sssaj2004.1838.

Yamamoto, Y, Y. Kobayashi, S. R. Devi, S Rikiishi, and H Matsumoto. 2003. Oxidative stress triggered by aluminum in plant roots. *Plant and Soil* 255:239–243.

Yang, Y., Y. Liu, C. F. Huang, J. de Silva, and F. J. Zhao. 2016. Aluminium alleviates fluoride toxicity in tea (Camellia sinensis). *Plant and Soil* 402 (1–2):179–190. doi:10.1007/s11104-015-2787-8.

Yi, J., and J. Cao. 2008. Tea and fluorosis. *Journal of Fluorine Chemistry* 129 (2):76–81.

Yokel, R. A. 2000. The toxicology of aluminum in the brain: a review. *Neurotoxicology* 21 (5):813–828.

Yokel, R. A., and R. L. Florence. 2008. Aluminum bioavailability from tea infusion. *Food and Chemical Toxicology* 46 (12):3659–3663. doi:10.1016/j.fct.2008.09.041.

Yokel, R. A., C. L. Hicks, and R. L. Florence. 2008. Aluminum bioavailability from basic sodium aluminum phosphate, an approved food additive emulsifying agent, incorporated in cheese. *Food and Chemical Toxicology* 46:2261–2266.

Yokel, R. A., and P. J. McNamara. 2001. Aluminium toxicokinetics: an updated mini review. *Pharmacology and Toxicology* 88:159–167.

16

Boron: Soil Contaminant

Rami Keren

Introduction

Boric acid is moderately soluble in water. Its solubility increases markedly with temperature due to the large negative heat of dissolution. Boron is considered as a typical metalloid having properties intermediate between the metals and the electronegative non-metals. Boron has a tendency to form anionic rather than cationic complexes. Boron chemistry is of covalent B compounds and not of B^{3+} ions because of its very high ionization potentials. Boron has five electrons, two in the inner spherical shell ($1s^2$), two in the outer spherical shell ($2s^2$), and one in the dumbbell shaped shell $\left(2p_x^1\right)$.[1] In the hybrid orbital state, the three electrons in the 2s and 2p orbitals form a hybrid orbital state $\left(2s^1 2p_x^1 2p_y^1\right)$, where each electron is alone in an orbit whose shape has both spherical and dumbbell characteristics. Each of these three orbits can hold one electron from another element to form a covalent bond between the element and B (BX_3). This leaves one 2p electron orbit that can hold two electrons, which if filled would completely fill the eight electron positions (octet) associated with the second electron shell around B. BX_3 compounds behave as acceptor Lewis acids toward many Lewis bases such as amines and phosphines. The acceptance of two electrons from a Lewis base completes the octet of electrons around B. Boron also completes its octet by forming both anionic and cationic complexes.[1] Therefore, tri-coordinate B compounds have strong electron-acceptor properties and may form tetra- coordinate B structures. The charge in tetra-coordinate derivatives may range from negative to neutral and positive, depending upon the nature of the ligands.

For the unshared oxygen atoms bound to B, they are, probably, always OH groups. Thus, in accordance with the electron configuration of B, boric acid acts as a weak Lewis acid:

$$B(OH)_3 + 2H_2O = B(OH)_4^- + H_3O^+ \qquad (1)$$

The formation of borate ion is spontaneous. The first hydrolysis constant of $B(OH)_3$, K_{h1}, is 5.8×10^{-10} at 20°C,[2] and the other K_{h2} and K_{h3} values are 5.0×10^{-13} and 5.0×10^{-14}, respectively.[3] A dissociation beyond $B(OH)_4^-$ is not necessary to explain the experimental data, at least below pH 13.[4,5] Boron species other than $B(OH)_3$ and $B(OH)_4^-$, however, can be ignored in soils for most practical purposes. The first hydrolysis constant of $B(OH)_3$ varies with temperature from 3.646×10^{-10} at 178 K to 7.865×10^{-10} at 318 K.[6]

Both $B(OH)_3$ and $B(OH)_4^-$ ion species are essentially monomeric in aqueous media at low B concentration ($\le 0.025 \text{mol} L^{-1}$). However, at high B concentration, polyborate ions exist in appreciable amount.[7] The equilibria between boric acid, monoborate ions, and polyborate ions in aqueous solution are rapidly reversible. In aqueous solution, most of the polyanions are unstable relative to their monomeric forms $B(OH)_3$

and $B(OH)_4^-$.[8] Results of nuclear magnetic resonance[9] and Raman spectrometry[10] lead to the conclusion that B(OH)3 has a trigonal-planar structure, whereas the $B(OH)_4^-$ ion in aqueous solution has a tetrahedral structure. This difference in structure can lead to differences in the affinity of clay for these two B species.

Boron–Soil Interaction

The elemental form of boron (B) is unstable in nature and found combined with oxygen in a wide variety of hydrated alkali and alkaline earth-borate salts and borosilicates as tourmaline. The total B content in soils, however, has little bearing on the status of available B to plants.

Boron can be specifically adsorbed by different clay minerals, hydroxy oxides of Al, Fe, and Mg, and organic matter.[11] Boron is adsorbed mainly on the particle edges of the clay minerals rather than the planar surfaces. The most reactive surface functional group on the edge surface is the hydroxyl exposed on the outer periphery of the clay mineral. This functional group is associated with two types of sites that are available for adsorption: Al(III) and Si(IV), which are located on the octahedral and tetrahedral sheets, respectively. The hydroxyl group associated with this site can form an inner sphere surface complex with a proton at low pH values or with a hydroxyl at high pH values. The B adsorption process can be explained by the surface complexation approach, in which the surface is considered as a ligand.[12] Such specific adsorption, which occurs irrespective of the sign of the net surface charge, can occur theoretically for any species capable of coordination with the surface metal ions. However, because oxygen is the ligand commonly coordinated to the metal ions in clay minerals, the B species $B(OH)_3$ and $B(OH)_4^-$ are particularly involved in such reactions. Possible surface complex configurations for B—broken edges of clay minerals—were suggested by Keren, Grossl, and Sparks.[12]

Keren and Bingham[11] reviewed the factors that affect the adsorption and desorption of B by soil constituents and the mechanisms of adsorption. Soil pH is one of the most important factors affecting B adsorption. Increasing pH enhances B adsorption on clay minerals, hydroxy-Al and soils, showing a maximum in the alkaline pH range (Figure 1).

The response of B adsorption on clays to variations in pH can be explained as follows. Below pH 7, $B(OH)_3$ predominates and since the affinity of the clay for this species is relatively low, the amount of

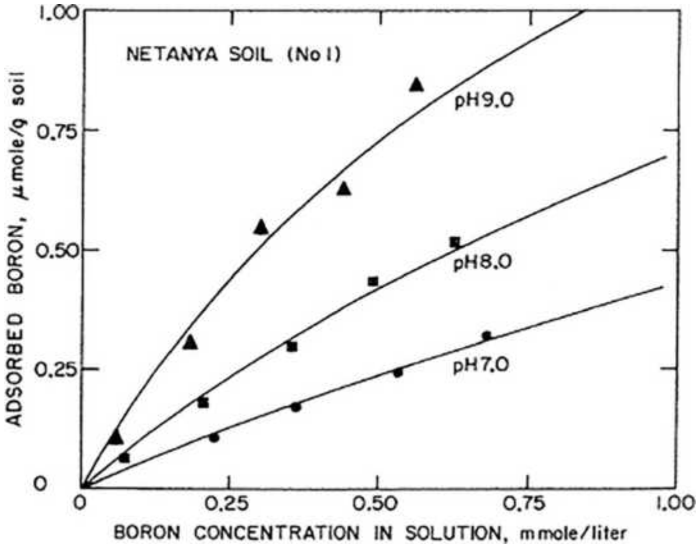


FIGURE 1 Boron adsorption isotherms for a soil as a function of solution B concentration and pH. Bold lines—calculated values.
Source: Mezuman and Keren.[28]

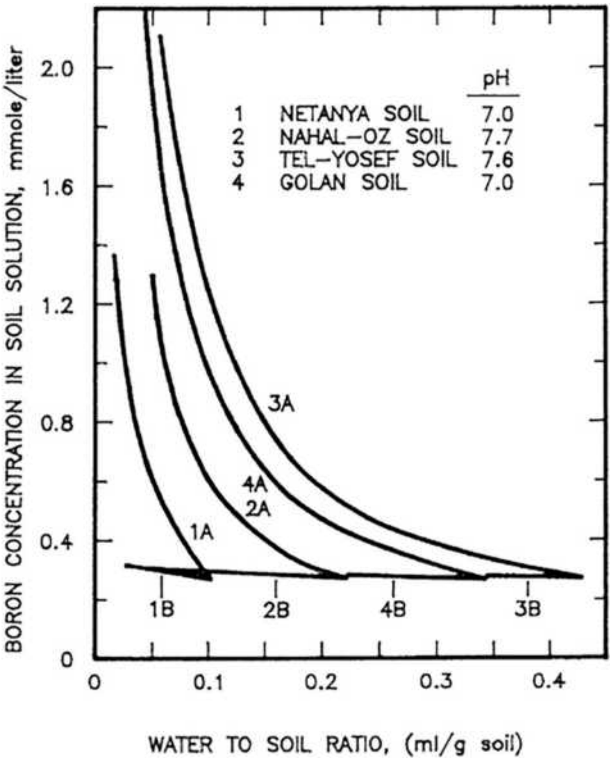


FIGURE 2 Boron concentration in soil solution as a function of solution-to-soil ratio for a given total amount of B. (a) No interaction between B and soil, (b) Boron adsorption account for.
Source: Mezuman and Keren.[28]

adsorption is small. Both $B(OH)_4^-$ and OH^- concentrations are low at this pH; thus, their contribution to total B adsorption is small despite their relatively strong affinity for the clay. As the pH is increased to about 9, the $B(OH)_4^-$ concentration increases rapidly. Since the OH^- concentration is still low relative to the B concentration, the amount of adsorbed B increases rapidly. Further increases in pH result in an enhanced OH^- concentration relative to $B(OH)_4^-$, and B adsorption decreases rapidly due to the competition by OH^- at the adsorption sites. Adsorption models for soils, clays, aluminum oxide, and iron oxide minerals have been derived by various workers.[13–17]

In assessing B concentration in irrigation water, however, the physicochemical characteristics of the soil must be taken into consideration because of the interaction between B and soil. Boron sorption and desorption from soil adsorption sites regulate the B concentration in soil solution depending on the changes in solution B concentration and the affinity of soil for B. Thus, adsorbed B may buffer fluctuations in solution B concentration, and B concentration in soil solution may change insignificantly by changing the soil-water content (Figure 2). When irrigation with water high in B is planned, special attention should be paid to this interaction because of the narrow difference between levels causing deficiency and toxicity symptoms in plants.

Boron–Plant Interaction

Boron is an essential micronutrient element required for growth and development of plants.

Many of the experimental data suggest that B uptake in plants is probably a passive process. There are clear evidences, however, that B uptake differs among species.[18] Several mechanisms have been postulated to explain this apparent paradox.[18–20] Boron deficiency in plants initially affects meristematic tissues, reducing or terminating growth of root and shoot apices, sugar transport, cell-wall synthesis and structure, carbohydrate metabolism and many biochemical reactions.[21,22] Tissue B concentrations associated with the appearance of vegetative deficiency symptoms have been identified in many crop species. It is essential to remember that for B, as for phosphorus and several other plant nutrient elements, deficiency may be present long before visual deficiency symptoms occur.

Excess and toxicity of boron in soils of semi-arid and arid areas are more of a problem than deficiency. Boron toxicity occurs in these areas either due to high levels of B in soils or due to additions of B in irrigation water. A summary of B tolerance data based upon plant response to soluble B is given by Maas.[23] Bingham et al.[24] showed that yield decrease of some crops (wheat, barley, and sorghum) due to B toxicity could be estimated by using a model for salinity response, suggested by Maas and Hoffman.[25]

There is a relatively small difference between the B concentration in soil solution causing deficiency and that resulting in toxicity symptoms in plants.[11] A consequence of this narrow difference is the difficulty posed in management of appropriate B levels in soil solution.

The suitability of irrigation water has been evaluated on the basis of criteria that determine the potential of the water to cause plant injury and yield reduction. In assessing the B in irrigation water, however, the physicochemical characteristics of the soil must be taken into consideration because the uptake by plants is dependent only on B activity in soil solution.[26,27] Boron uptake by plants grown in a soil of low-clay content is significantly greater than that of plants grown in a soil of high-clay content at the same given level of added B (Figure 3). This knowledge may improve the efficacy of using water of different qualities, whereby water with relatively high B levels could be used to irrigate B-sensitive crops in soils that show a high affinity to B. Such water can be used for irrigation as long as the equilibrium B concentration in soil solution is below the toxic concentration threshold (the maximum permissible concentration for a given

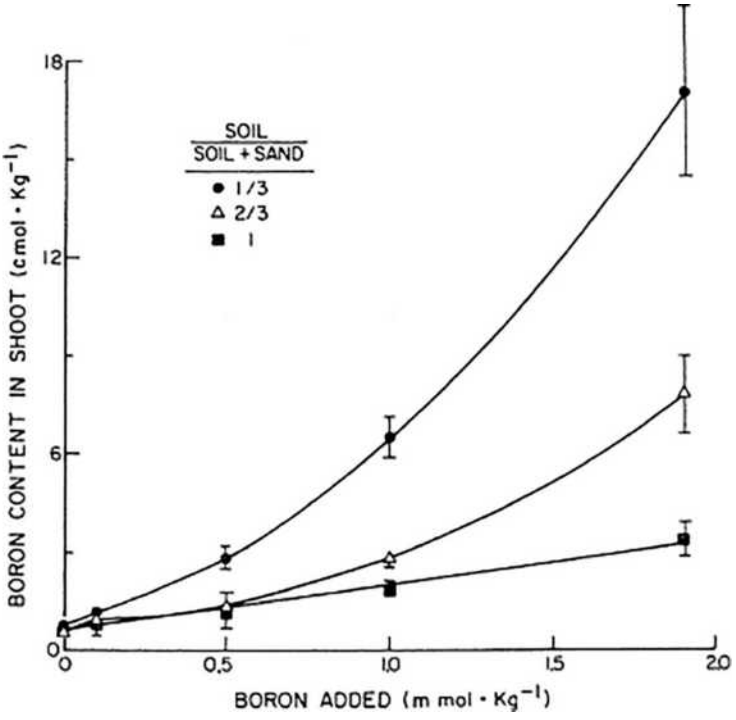


FIGURE 3 Relationship between B content in wheat shoot and the amount of B added to soil, for three ratios of soil-sand mixtures.
Source: Keren et al.[26]

crop species that does not reduce yield or lead to injury symptoms) for the irrigated crop. The existing criteria for irrigation water, however, make no reference to differences in soil type.

References

1. Cotton, F.A.; Wilkinson, G. *Advanced Inorganic Chemistry,* 5th Ed.; Wiley and Sons: New York, NY, 1988.
2. Owen, B.B. The dissociation constant of boric acid from 10 to 50°. J. Am. Chem. Soc. **1934**, *56,* 1695–1697.
3. Konopik, N.; Leberl, O. Colorimetric determination of PH in the range of 10 to 15. Monatsh **1949**, *80,* 420–429.
4. Ingri, N. Equilibrium studies of the polyanions containing B^{III}, Si^{IV}, Ge^{IV} and V^{V}. Svensk. Kem. Tidskr. **1963**, *75*, 199–230.
5. Mesmer, R.E.; Baes, C.F., Jr.; Sweeton, F.H. Acidity measurements at elevated temperature. VI. Boric acid equilibria. Inorg. Chem. **1972**, *11,* 537–543.
6. Owen, B.B.; King, E.J. The effect of sodium chloride upon the ionozation of boric acid at various temperatures. J. Am. Chem. Soc. **1943**, *65,* 1612–1620.
7. Adams, R.M. *Boron, Metallo-Boron Compounds and Bo- ranes*; John Wiley and Sons: New York, 1964.
8. Onak, T.P.; Landesman, H.; Williams, R.E.; Shapiro, I. The B^{II} nuclear magnetic resonance chemical shifts and spin coupling values for various compounds. J. Phys. Chem. **1959**, *63,* 1533–1535.
9. Good, C.D.; Ritter, D.M. Alkenylboranes: II. Improved preperative methods and new observations on methylvinylboranes. J. Am. Chem. Soc. **1962**, *84,* 1162–1166.
10. Servoss, R.R.; Clark, H.M. Vibrational spectra of normal and isotopically labeled boric acid. J. Chem. Phys. **1957**, *26*, 1175–1178.
11. Keren, R.; Bingham, F.T. Boron in water, soil and plants. Adv. Soil Sci. **1985**, *1,* 229–276.
12. Keren, R.; Grossl, P.R.; Sparks, D.L. Equilibrium and kinetics of borate adsorption-desorption on pyrophyllite in aqueous suspensions. Soil Sci. Soc. Am. J. **1994**, *58,* 1116–1122.
13. Keren, R.; Gast, R.G.; Bar-Yosef, B. pH-dependent boron adsorption by na-montmorillonite. Soil Sci. Soc. Am. J. **1981**, *45,* 45–48.
14. Keren, R.; Gast, R.G. pH dependent boron adsorption by montmorillonite hydroxy-aluminum complexes. Soil Sci. Soc. Am. J. **1983**, *47,* 1116–1121.
15. Goldberg, S.; Glaubig, R.A. Boron adsorption on aluminum and iron oxide minerals. Soil Sci. Soc. Am. J. **1985**, *49,* 1374–1379.
16. Goldberg, S.; Glaubig, R.A. Boron adsorption on California soils. Soil Sci. Soc. Am. J. **1986**, *50,* 1173–1176.
17. Goldberg, S.; Forster, H.S.; Heick, E.L. Boron adsorption mechanisms on oxides, clay minerals and soils inferred from ionic strength effects. Soil Sci. Soc. Am. J. **1993**, *57,* 704–708.
18. Nable, R.O. Effects of B toxicity amongst several barley wheat cultivars: a preliminary examination of the resistance mechanism. Plant Soil **1988**, *112,* 45–52.
19. Nable, R.O.; Lance, R.C.M.; Cartwright, B. Uptake of boron and silicon by barley genotypes with differing susceptibilities to boron toxicity. Ann. Bot. **1990**, *66*, 83–90.
20. Brown, P.H.; Hu, H. Boron uptake by sunflower, squash and cultured tobacco cells. Physiol. Plant **1994**, *91,* 435–441.
21. Loomis, W.D.; Durst, R.W. Chemistry and biology of boron. BioFactors **1992**, *3,* 229–239.
22. Marschner, H. *Mineral Nutrition of Higher Plants,* 2nd Ed.; Academic Press: London, 1995.
23. Maas, E.V. Salt tolerance of plants. In *Handbook of Plant Science in Agriculture;* Christie, B.R., Ed.; CRC Press, Inc.: Cleveland, Ohio, 1984.
24. Bingham, F.T.; Strong, J.E.; Rhoades, J.D.; Keren, R. An application of the Maas-Hoffman salinity response model for boron toxicity. Soil Sci. Soc. Am. J. **1985**, *49,* 672–674.

25. Maas, E.V.; Hoffman, G.J. Crop salt tolerance—current assessment. ASCE J. Irrig. Drainage Div. **1977**, *103,* 115–134.
26. Keren, R.; Bingham, F.T.; Rhoades, J.D. Effect of clay content on soil boron uptake and yield of wheat. Soil Sci. Soc. Am. J. **1985**, *49,* 1466–1470.
27. Keren, R.; Bingham, F.T.; Rhoades, J.D. Plant uptake of boron as affected by boron distribution between liquid and solid phases in soil. Soil Sci. Soc. Am. J. **1985**, *49,* 297–302.
28. Mezuman, U.; Keren, R. Boron adsorption by soils using a phenomenological adsorption equation. Soil Sci. Soc. Am. J. **1981**, *45,* 722–726.

17

Cadmium: Toxicology

Sven Erik Jørgensen

Introduction: Dispersion and Application

From 1940 to 1960, Japanese in the Toyama Prefecture were poisoned by cadmium in their rice, because the river water used for irrigation was contaminated by cadmium from a cadmium mine.[1] Cadmium replaces calcium in the bones, which makes them soft. It causes extreme pain and the disease was named *itai-itai,* which means "ouch- ouch." When it was discovered that cadmium was causing the disease, the mine waste was controlled, but by then, several thousands (mainly farmers) were already suffering from the very painful disease.

Cadmium is used for metal surface treatment, as a stabilizing agent in plastic and many alloys. Phosphorus fertilizers have a relatively high cadmium concentration—in the order of 10–80 mg/kg. Together with the cadmium emitted from coal-fired power plants, it is the most important source of global cadmium dispersion. More than 1000 tons of cadmium is globally dispersed from coal-fired power plants. Cadmium was previously applied in ceramic, but this use of the very toxic metal is now banned in most industrialized countries.

Like other heavy metals, cadmium shows biomagnification and bioaccumulation and is accumulated in the sediment of aquatic ecosystems. It entails that the dispersion of cadmium in aquatic ecosystems can, in most cases, best be determined by analyses of the sediment where relatively higher concentrations are found.[2] Similar to lead (*Lead: Ecotoxicology,* p. 1651), it is possible by analyses of sediment cores to determine the history of cadmium emission. The phosphorus fertilizer industry has, for several decades, discharged cadmium-containing production waste to the Little Belt in Denmark. An environmental impact assessment was carried out for Little Belt by analyses of a large number of sediment cores, whereby the history of the contamination was determined. Fortunately, cadmium can form complexes with chloride, which are less toxic than the cadmium ions.[3,4] The complexes are at the same time more soluble, which means that less cadmium is transferred to the sediment and more to the open sea where it is diluted significantly. The discharge of cadmium waste was stopped as a result of the investigation, but it was also concluded that the contamination of the Little Belt was less than expected, probably due to the formation of cadmium–chloride complexes.

Due to the high cadmium concentration in phosphorus fertilizers, there is a risk for cadmium contamination of agricultural land, particularly by the application of intensive agriculture. Cadmium is taken up by plants (see also *Bioremediation*, p. 408). A cadmium model has been developed to relate the cadmium concentration in crops as a function of the cadmium contamination by the use of fertilizers.[5]

It is beneficial for all toxic substances and particularly for the most toxic heavy metals to make a regional mass balance to identify the sources and the dispersion to assess whether a specified toxic substance would reach an unacceptably high concentration and thereby do the most harm. Jørgensen and Fath[5] have exemplified regional mass balances. Figure 1 shows one example, namely, a cadmium balance for Danish agricultural land, but similar balances can also be found for lead, mercury, and copper, as well as in other regions. The cadmium pollution of the agricultural land comes from the use of fertilizers (1.7 g/ha/yr), from the use of sludge (0.18 g/ha/yr), and from air pollution, dry deposition, and rainwater (1.7 g/ha/yr). A total of 1.7 g/ha/yr is taken up from the soil by the plants and 0.2 g/ha/yr contaminates vegetable products, while only 0.01 g of cadmium per year contaminates the animal food, even when the cadmium in imported fodder is considered. By far, most of the cadmium-contaminating domestic animals go back to the agricultural land by animal waste. In total, about 3.8 g of cadmium is added to the agricultural land per year, but 0.7 g of cadmium per hectare per year is transported by drainage water and groundwater to other sites (ecosystems) and 0.2 g/ha/yr will be removed by the harvest of vegetables. The net accumulation in the agricultural soil is therefore 3.1 g/ha/yr. The mass balance can be used to calculate the effect of using fertilizers with less cadmium or the result of a reduction of cadmium discharged to the atmosphere. If more complex pollution abatement strategies are applied, it is necessary to apply an ecotoxicological model with the state variables, processes, and transfers as shown in Figure 1 as the core of the model.

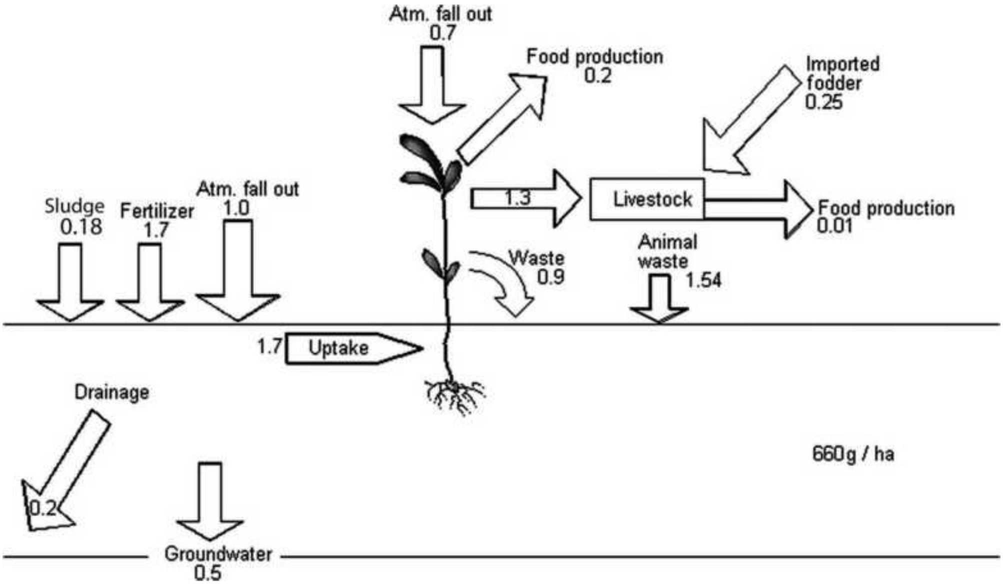


FIGURE 1 Cadmium balance for Danish agricultural land. The accumulated amount is about 660 g/ha, while the annual input by fertilizers is 1.7 g/ha and by air pollution another 1.7 g/ha, of which 1 g is accumulated directly on the bare soil and the 0.7 g on the plants. The amount from fertilizers corresponds to the amount taken up by plants. 1.3 g/ha of the cadmium contamination of plants and 0.25 g/ha of cadmium in imported fodder give an accumulation of 1.55 g of cadmium per hectare per year in domestic animals, but only the 0.01 g/ha will end up as cadmium contamination of animal products.

Toxicity and Ecotoxicity

Cadmium has carcinogenic and teratogenic effects and is also highly toxic, as indicated by the LD_{50} value for rats—70–90 mg/kg. LD_{50} for cadmium in smoke (as cadmium oxide) by inhalation is 500 min/mg/m^3. The cadmium concentration in rice that caused the itai-itai disease in Japan was about 1 mg/kg.

Uptake from food of cadmium is about the same as for other heavy metals—7%–10%. Cadmium is accumulated mainly in the kidneys. The total cadmium concentration in the body increases with age, as the uptake is bigger than the excretion, mainly by the urine. The excretion follows a first-order reaction with a coefficient of 0.0001 1/24*h*. At an age of 50, the cadmium content in the body will be about 25 mg, of which the 8–9 mg is accumulated in the kidney and the 3 mg is accumulated in the liver.[6] The daily accumulation is about 2–3 μg.

It is possible to express cadmium accumulation in the body as a function of time by the following differential equation:

$$\mathrm{dCd}/\mathrm{d}t = \text{daily uptake} - \text{excretion coefficient} * \mathrm{Cd}$$

Cd is the amount of cadmium in the body and the daily uptake is found as about 10% of the cadmium content in the food and 50% of the cadmium in the 20 m^3 of air used for respiration per 24 hr. The intake by food is about 25 μg/24 hr and the intake by respiration is very minor. Therefore, the equation can be written as:

$$\mathrm{dCd/dt} = 0.025*10/100 - 0.0001*\mathrm{Cd}\ \mathrm{mg/24hr}$$

At steady state, the cadmium concentration becomes 0.0025/0.0001 = 25 mg, which (of course) is close to the average value at the age of 50 years.

References

1. Newman, M.C.; Unger, M.A. *Fundamentals of Ecotoxicology,* 2nd Ed.; CRC Press: Boca Raton, London, and New York, 2003.
2. Jørgensen, S.E. *Principles of Pollution Abatement;* Elsevier: Amsterdam, 2000; 520 pp.
3. Jørgensen, S.E.; Halling-Sørensen, B.; Mahler, H. *Handbook of Estimation Methods in Ecotoxicology and Environmental Chemistry;* Taylor and Francis Publ.: Boca Raton, FL; 230 pp.
4. Jørgensen, S.E.; Jørgensen, L.A.; Nors Nielsen, S. *Handbook of Ecological and Ecotoxicological Parameters;* Elsevier: Amsterdam, 1991; 1380 pp.
5. Jørgensen, S.E.; Fath, B. *Fundamentals of Ecological Modelling,* 4th Ed.; Elsevier: Amsterdam, Oxford, 2011; 396 pp.
6. *Om Metaller,* 2nd Ed.; Statens Naturvørdsverk: Sweden, 1980.

18

Carbon: Soil Inorganic

Donald L. Suarez

Introduction

Concerns related to global climate change and greenhouse gas warming have led to extensive interest in changes in carbon storage and release of carbon dioxide into the atmosphere. While the burning of fossil fuels was a major contributor to the increase in atmospheric carbon dioxide over the last 200 years, there has been extensive research on the impact of other processes, such as changes in the organic carbon (OC) stocks in soils. Inorganic carbon (IC) storage and changes associated with land use have received relatively little attention, despite the fact that IC is a major carbon pool in the near-surface environment. At the global scale, total soil OC is estimated at 1550 Pg and soil IC at 950 Pg.[1] In addition, in arid regions, IC can comprise more than 90% of the total C in the soil. The IC in the soil is present in the minerals calcite and dolomite. Both minerals are relatively insoluble; however, dolomite dissolution is much slower than calcite dissolution at the intermediate pH values relevant to most soils. Also, dolomite does not generally precipitate under earth surface conditions. As a result, it can be considered that over the time frames of current interest related to climate change, the dolomite component in soils remains constant or decreases due to dissolution, while calcium carbonate content may increase or decrease.

This entry considers the soil processes and land-use changes that impact changes in soil IC and its implication for the atmospheric carbon dioxide budget. Among the land-use changes considered are land clearing, tillage, cropping, fertilization, irrigation in both humid and arid environments, irrigation management (leaching), and composition and source of the irrigation water. The impacts of these changes on atmospheric carbon dioxide concentrations are more difficult to analyze and differ with regard to long-term and short-term effects.

Soil Processes

In terms of OC at the land surface and in the soil, a decrease in OC is directly related to either a net removal of OC from the site such as timber or crop harvesting or input of carbon dioxide to the atmospheric in the case of oxidation of soil organic matter. In contrast, for dissolution of the carbonate minerals, a decrease in IC in the soil can result in either a release or an uptake of carbon dioxide from the atmosphere. The impact of these processes is entirely dependent on the local environmental and hydrologic conditions.

In the instance of dissolution under non-acidic conditions, the net reaction for dissolution of calcium carbonate is:

$$C_aCO_3 + H_2O + CO_2 \rightarrow C_a^{2+} + 2HCO_3^-. \tag{1}$$

Thus, dissolution of calcium carbonate results in a sink for atmospheric carbon. Increased soil carbon dioxide concentrations also lead to a net increase in carbonic acid in solution (H_2CO_3). This carbonic acid is partially recharged to the groundwater and is thus a sink for atmospheric CO_2 that is greater than that indicated by Eq. 1 and generally not considered when evaluating organic matter decomposition.

Dissolution of calcium carbonate under acidic conditions results in the following net reaction:

$$C_aCO_3 + 2H^+ \rightarrow C_a^{2+} + H_2O + CO_2, \tag{2}$$

thus leading to a net release of carbon dioxide to the atmosphere. In the case of dissolution of IC, the residence time in the groundwater will determine the net impact on atmospheric carbon dioxide in the time scale of most concern—tens to hundreds of years.

The majority of the alkalinity in the earth's near-surface environment is derived from dissolution of carbonate rocks. The net long-term impact is uptake of carbon dioxide from the atmosphere, dissolution of IC, movement of alkalinity in the surface and subsurface, and discharge to the oceans via rivers.

In terms of the overall C cycle, most of the carbon dioxide released from precipitation of carbonates in the soil (an important process in arid regions) results in a return of C to the atmosphere, as seen by the reverse reaction of Eq. 1. Precipitation of IC results in a net release of 1 mol of carbon dioxide for each mole of IC produced. Sequestration of C in the form of soil IC is carbon storage but the net result of the process is a net release of C to the atmosphere. The quantity released is less than indicated by Eq. 1 since a small amount of the carbon dioxide generated may be recharged with the groundwater, a process that occurs in the absence of precipitation of soil IC.

An important factor to consider when evaluating the implication of an increase in carbon storage is the origin of the calcium and bicarbonate ions. The above instance considered dissolution of IC from the soil. Weathering of silicate minerals is important over long time scales, resulting in release of calcium, sodium, magnesium, and potassium ions as well as bicarbonate. In this instance, the net reaction for dissolution is consumption of 2 mol of carbon dioxide to produce 2 mol of bicarbonate ions, in addition to the release of an equivalent quantity of cations into solution. Thus, silicate dissolution in the soil results in sequestering atmospheric carbon in the form of dissolved bicarbonate. It may be possible in arid regions that the resultant soil solution becomes concentrated, resulting in precipitation of IC. In this process, 1 mol of C is released during IC precipitation and 2 mol are sequestered from silicate dissolution. In this instance, the net result of the silicate weathering is still carbon sequestration from the atmosphere since not all of the alkalinity generated by silicate dissolution is reprecipitated. In terms of IC, the result is an increase in soil IC. Silicate dissolution has a relatively small role in terms of changes related to human activity.

Dissolved IC Cycling

The net impact of soil IC changes on atmospheric carbon dioxide is critically related to the fate of alkalinity (primarily bicarbonate ions) transported into the subsurface below the root zone. If we consider that most of the recharge is ultimately discharged to surface systems, then the potential exists for reprecipitation in these systems, which are at lower carbon dioxide concentrations than the subsurface. It is estimated[2] that inland waters receive approximately 1.9 Pg C/yr, of which 0.8 is returned to the atmosphere as carbon dioxide while 0.9 Pg C/yr is discharged to the oceans, of which 0.45 Pg C/yr is IC. The direct degassing of groundwater is estimated at 0.01 Pg/yr with an uncertainty of 0.003–0.03 Pg/yr.[2] Combining soil liming as well as irrigation practices and other land-use changes, it is likely that the IC flux to groundwater has increased due to human activities. It is estimated that the dissolved IC (mostly bicarbonate) flux into groundwater is on the order of 0.2 Gg/yr.[3] The impact of changes in this subsurface flux, however, is likely relatively long term as the groundwater residence times are mostly on the order of hundreds to thousands of years.[3]

Land Use

Land use, among other environmental factors, affects changes in soil IC with the full impact on atmospheric carbon dioxide expressed only over the long term. Both increases and decreases in IC storage are possible as a result of various management practices. In irrigated lands, there are various factors that cause either a decrease or an increase in soil IC. In most instances, multiple land-use changes have occurred as a result of human activity, and it is difficult to isolate the specific process that is dominant in explaining changes in soil IC. It is calculated that China has a total of 53 Pg of soil IC and that human activity has resulted in a net loss of 1.6 Pg of IC.[4]

Soil respiration processes include plant respiration and release of carbon dioxide via the roots, as well as microbial decomposition of soil organic matter. These processes result in an elevated carbon dioxide concentration in the soil relative to the atmospheric condition. The soil CO_2 concentration is typically in the range of 0.1%–5%, orders of magnitude higher than the atmospheric value. Increased plant biomass productivity due to crop selection irrigation and/or fertilization results in increased soil carbon dioxide production and concentration, thus contributing to increased weathering of both silicates and carbonates. Inadvertent or deliberate introduction of non-native species also impacts biomass productivity, soil CO_2 concentrations, water use, recharge, and erosion with subsequent impact on soil IC.

Land Clearing

Land clearing generally results in increased water runoff and soil erosion. This process or any other process such as tillage that increases erosion serves to remove the surface soil horizons. Since these horizons are generally depleted in IC relative to less weathered, deeper horizons, erosion causes an apparent increase in the IC content of the top 1–2 m of carbonate-containing soils. This may result in erroneous estimates of changes in soil IC when comparing the top 1 m of disturbed vs. native vegetation sites. In terms of carbonate dissolution, the impact of land clearing is not certain. After clearing, there is increased runoff, thus decreased surface infiltration, favoring less dissolution of carbonates from the surface horizons. This effect may be compensated by the decreased water consumption (lower evapotranspiration) after clearing, resulting in increased deep recharge and possibly greater carbonate dissolution or less reprecipitation at depth, depending on the volumes of water. Depending on how much biomass remains after clearing, there is likely a short-term increase in soil CO_2 followed by a longer-term reduction, favoring less carbonate dissolution in the soil.

Cropping and Tillage

The impact of cropping and tillage on IC storage is not certain and there are limited studies, but the sum of the studies suggests net IC accumulation. The impact of tillage and cropping on IC in the Northern Great Plains of the United States has been evaluated by comparing long-term cultivation (likely >80 years) with uncultivated grassland.[5] There was a statistically significant increase in IC with cropping for one soil type (at the 90% confidence level but not at the 95% confidence level) but no statistically significant differences for the other two soil types examined. The lack of a clear impact is likely related to the fact that these soils did not have significant quantities of IC in the upper portion of the profile where most of the water transport and root water uptake would occur.

In a study of Russian Chernozem soils, the soil IC was greater on a continuously cropped field and on a fallow field relative to a native grassland, with most of the approximately 100 Mg/ha increase in IC being in the upper portion of the profile.[6] A grassland field where hay was harvested showed no significant changes over the native vegetation (grassland) field. The authors attributed the increase in the cropped and fallow fields to cultivation, irrigation, and fertilization, with the calcium being supplied by fertilizer and manure additions. The manure also supplied alkalinity necessary for carbonate formation. The effects in this study may be more related to fertilizer application than changes due to cultivation per se.

When comparing IC in cropped soils to native grassland, it needs to be considered that grassland and native vegetative soils can also have net accumulation or loss of IC with net increase in pedogenic and recently formed IC. We cannot assume that increases in pedogenic carbonate necessarily correspond to increases in IC storage. For example, it is reported that for grasslands and forested soils in Saskatchewan, Canada, there is an increase in pedogenic IC of 1.2–1.8 kg/ha/yr,[7] and this has been cited as an increase in IC. However, the soils contain dolomite and re-cystallization of carbonates is listed by the authors as a major process producing the pedogenic carbonate.[7] Only one of the soils (Black Chernozem) had net IC accumulation based on comparison of the pedogenic carbonate accumulated and the depletion of the lithogenic carbonate.[7]

Land clearing, tillage, and overgrazing in arid lands serve to increase wind erosion and to redistribute soil in the landscape. In this manner, non-calcareous soils receive inputs of carbonates. This process likely increases net dissolution of carbonates, as it spreads carbonates across the landscape and into areas with non-calcareous soils.

In humid environments, IC is leached from the soil. The elevated CO_2 concentrations in the soil enhance IC solubility relative to earth surface conditions. In humid environments, carbonates are successively leached from the upper portions of the soil profile. Agricultural practices may serve to enhance or reduce the net removal of carbonates. Removal of vegetation from a site with cropping practices such as tree harvesting or crop or forage harvesting serves to remove base cations and causes net acidification of the upper portions of the soil profile. If carbonates are present deeper in the soil, this acidification increases IC dissolution. The impact of removal of vegetation in humid environments with carbonates in the subsoil can be calculated by assuming that the net harvested alkalinity is compensated by an equal increase in carbonate dissolution in the subsurface.

Fertilization and Liming

Since optimum plant growth is generally at a pH lower than that observed in untreated calcareous soils, acid fertilizers are commonly applied in arid regions. Use of sulfur with subsequent oxidation to sulfate results in acid release to the soil (2 mol of protons per mole of sulfur). Application of ammonia salts with subsequent fixation into organic matter or oxidation to nitrous oxide or elemental nitrogen also releases protons (2 and 1 mol of protons per mole of ammonia ions, respectively). This acidification will increase carbonate dissolution proportionately and has a significant effect since ammonia salts are widely applied as fertilizers.

Application of fertilizer as urea or ammonia gas should have no net effect on carbonate dissolution (upon oxidation to nitrous oxide or elemental nitrogen) other than the indirect impact on soil CO_2. In contrast, use of nitrate fertilizers serves to increase pH and thus reduce carbonate dissolution. Generally, nitrate is not utilized on calcareous soils, so the impact on IC storage in soil is slight. The quantitative impact of fertilization on changes in IC is not easily calculated, as it depends on the extent of N incorporation into organic matter, mineralization, the extent to which the harvested biomass is removed from the site, and the occurrence of carbonates in the subsurface.

The practice of liming soils is very prevalent for crop production and urban landscaping in humid environments and represents a major anthropogenic input. The liming is done by application of calcium carbonate or dolomite to the soil. Addition of liming products, primarily calcite, are reported as 3.7 Tg C/yr in the United States for 1978.[8] This is a significant but temporary addition to the soil IC pool. Since liming is not generally needed in arid regions, it is reasonably assumed that the majority of the material is applied to acid soils and thus it is readily dissolved. The liming serves two functions, to increase the pH of acid soils and as an inexpensive calcium source in regions with leached, calcium-deficient soils. The net impact of liming is neutralization of most of the carbonate in the mineral and release of the generated carbon dioxide to the atmosphere, as represented in Eq. 2. The amount of carbon dioxide release to the atmosphere is less than indicted by Eq. 2 for all but the highly acidic soils, as some of the carbon is sequestered in the form of dissolved bicarbonate. In humid environments, most of the bicarbonate produced will likely not reprecipitate, even in the long term; instead, it will be discharged into the ocean via river drainage systems.

Humid Region Irrigation

Large quantities of water applied in excess of plant transpiration needs result in leaching and maintenance of elevated water content at or near the soil surface. The additional water applied as irrigation enhances dissolution of IC as does the increased soil carbon dioxide concentrations due to increased water content as in the soil. Irrigation in humid environments thus serves to increase the net recharge through the soils and thus removal of carbonates. These changes may be relatively difficult to detect in view of the limited amount of irrigation water added and the fact that irrigation in humid environments, although increasing rapidly, was very limited in the past. Field studies are needed to determine the impact of irrigation on changes in IC in humid environments.

Arid and Semiarid Region Irrigation

Arid zone soils usually contain at least minor amounts of carbonates, even if they are not classified as calcareous. In the absence of irrigation, there may be redistribution of carbonates within the soil but likely little net precipitation. The majority of the pedogenic calcite is reprecipitated calcite with relatively small amounts added as a result of silicate mineral weathering. This process is of geological significance and central to explaining soil formation but of less importance to the time frames of interest with regard to recent changes in atmospheric CO_2. Significant amounts of carbonates are added to the surface of arid land soils as dust. IC is leached from the upper part of the soil profile by dissolution into the infiltrating rain and is mostly reprecipitated at depth after plant extraction of the available water.

Irrigation in arid and semiarid environments may result in a net increase or decrease in soil IC, depending on the water source and fraction of water applied that is leached (leaching fraction). There are various opposing effects. First, elevated CO_2 concentrations in the root zone relative to the atmospheric condition results in enhanced calcite and dolomite solubility and dissolution and, hence, depletion of soil IC. Second, plant water extraction and evaporation concentrate the soluble salts into a smaller volume of water and enhance calcite precipitation, thus increasing IC. Plant roots extract relatively dilute water from the soil, leaving most of the salts behind. At low leaching fractions (where the quantity of water applied plus rain is only 2%–10% greater than the amount of water transpired), the effect of concentration of salts due to plant water extraction and evaporation is greater than the enhanced CO_2 effect and there is net precipitation of calcite.

For a calcite-supersaturated surface water such as the Colorado River, it is estimated[9] that at a leaching fraction of 0.1, there is net precipitation of 125 kg/ha/yr of C, based on water consumption of 1.2 m/yr and an average soil CO_2 partial pressure of 3 kPa (approximately 3% of the soil atmosphere). Model simulations indicate that net precipitation of calcite occurs in the soil profile, with loss of IC in the upper portion of the root zone and precipitation of IC in the lower portion. At high leaching fractions, there is net dissolution of carbonates. Using a predictive calcite precipitation model that accounts for 3-fold calcite supersaturation in precipitating environments, it is predicted that at a leaching fraction of 0.4, there will be a net dissolution of IC of 70 kg/ha/yr of C. Again, this calculation is based on water consumption of 1.2 m/yr and soil CO_2 partial pressure of 3 kPa.[10] In all instances, there is a prediction of net dissolution in the upper portion of the soil root zone and net precipitation in the lower portions of the profile. Using average leaching fractions for the western United States, it is estimated that irrigation with surface waters on 12 million ha[11] results in an increase in soil IC of 1 Tg/yr,[10] or 80 kg/ha/yr. Considering the reverse reaction in Eq. 1, this increase in soil IC corresponds to a net release of almost 1 Tg/yr of C to the atmosphere.

Irrigation with groundwater saturated with respect to calcite will result in precipitation of carbonates at almost all leaching fractions, since the irrigation water is equilibrated at the groundwater CO_2 partial pressure and is highly supersaturated with respect to calcite upon degassing and application to the soil surface. Calcite-saturated groundwater is used for irrigation on an estimated 3.12 million ha in the United States.[11] It is estimated that irrigation on these soils results in a net IC precipitation of 1.3 Tg/yr or 420 kg/ha/yr.[10] This corresponds to slightly less than 1.3 Tg/yr of carbon dioxide release to the atmosphere.

The model calculations of IC precipitation are dependent on several assumptions regarding calcite precipitation and soil CO_2 concentrations and do not consider the effect of fertilizers. Generally, acidifying fertilizers are applied in arid regions, thus reducing the extent of expected increases in IC with irrigation. A partial validation of the modeling is available by comparing modeling results with data from Palo Verde, California. Using measured CO_2 partial pressure in the groundwater in Palo Verde Valley, there is prediction of no net change in IC at a leaching fraction of 0.5, obtained from the measured water diversions and drainage volumes. Consistent with these predictions, the groundwater composition draining back to the Colorado River from Palo Verde Valley shows no evidence for net precipitation or dissolution of carbonates in the soil.

There is some direct field evidence for the influence of irrigation on soil IC, but as expected from the above discussion, the results are site specific depending on water source, water composition, and leaching, as well as quantity and type of fertilizer used. It is also difficult to be certain that differences among sites are only related to changes in management. Researchers[12] observed a net decrease in the calcium carbonate content of three pairs of soil profiles taken from sites in Israel irrigated for approximately 40 years as compared to non-irrigated sites. The estimated input of 4.40 m of water per year at those sites is contrasted with the yearly potential evapotranspiration of 1.93 m. The observed trend is qualitatively consistent with model predictions if we account for the input of rain and the estimated high leaching fraction of 0.56. Isotopic evidence indicated that there was precipitation of pedogenic carbonate at depth despite a net decrease in IC content at depth.[12] This suggests solubilization and reprecipitation, but the measured net impact was still loss of IC in the soil profile, as expected from modeling this water budget data using surface water for irrigation.

In a study in the San Joaquin Valley in California, researchers compared samples of a soil taken from irrigated and native vegetation sites.[13] They also measured a net loss of carbonates attributed to 8 years of irrigation. Net carbonate loss was estimated as 7×10^3 kg/ha/yr (800 kg C/ha/yr). Leaching fractions at the site were not reported, but this value corresponds to approximately 10 times greater dissolution than expected based on model simulations. However, another study by the same author[14] found no change in total carbonate when comparing pedons with native vegetation and those irrigated for 5–25 years.

In this instance, both gypsum and sulfur were applied as amendments for reclamation. Gypsum would tend to greatly increase precipitation of carbonates and thus increase IC, while sulfur would acidify the soil and cause net dissolution of carbonates.

In other studies, the results are also mixed as expected if we consider variations in irrigation water composition and leaching fractions. In a semiarid northwestern U.S. study, irrigated crops showed a net increase in IC while irrigated pasture showed a net decrease relative to native sagebrush vegetation.[15] In the San Joaquin Valley, California, there was a net gain in IC of 1.8 kg/m^2 after 30 years of irrigation and a net loss of IC after 55 years of irrigation at another site, while in Imperial Valley, there was a net gain of IC of 4 kg/m^2 or 40 Mg/ha after 85–90 years of irrigation.[16] Another, more extensive study in Imperial Valley[17] on paired soil cores (irrigated and adjacent non-irrigated sites) observed no changes in IC storage after 90 years of irrigation and no isotopic C shifts indicative of recrystallization. Accurately modeling the net global impact of irrigation on soil IC based on water types, water application quantities, and evapotranspiration will first require the ability to accurately simulate the impacts on specific locations.

Sodic Soil Reclamation

Reclamation of sodic soils can result in either an increase or a decrease in soil IC. Gypsum application to a sodic and alkaline soil will increase the soil carbonate content, as the increased Ca will precipitate most of the soluble bicarbonate and carbonate. This will increase IC sequestration in the soil but have a net effect of also increasing the carbon dioxide flux to the atmosphere (and decreasing the net flux of C and alkalinity to the groundwater). Gypsum is also very commonly used in arid regions to improve water infiltration. At least some of the applied calcium from the gypsum application is reprecipitated as calcium carbonate, increasing the soil IC. Application of sulfuric acid, sulfur, or green manuring all serve to dissolve soil carbonates, thus depleting IC stocks in the soil. Addition of acid will result in an increase in atmospheric carbon dioxide equal to the IC removed.

Green manuring is a management practice of adding fresh plant organic matter to a calcareous soil, enhancing CO_2 concentrations in the soil and thus enhancing carbonate dissolution, providing a calcium source to replace the soil exchangeable sodium. This process will decrease soil IC and result in net release of CO_2 to the atmosphere. The dissolution process will deplete atmospheric carbon, but the effect is negated by the larger increase in carbon dioxide release related to the oxidation of the OC. It is estimated that this process can dissolve on the order of 400 to 800 kg/ha during a year of reclamation. Use of acid is currently a widespread and generally recommended practice to prevent emitter clogging in drip irrigation systems. This practice may result in total removal of carbonates within 10–20 years, for soils with less than 3% carbonate content.

Impact of IC on Atmospheric Carbon Dioxide

Dissolution of carbonates in neutral to alkaline environments results in consumption of CO_2 gas and formation of aqueous bicarbonate $\left(HCO_3^-\right)$, while precipitation of carbonates results in release of CO_2. The net effect of dissolution or precipitation of soil IC on atmospheric CO_2 depends on the solution flow path. In regions irrigated with surface water, the dissolution of carbonates results in a net C sink. The high alkalinity drainage water usually flows back to the river, but even in shallow groundwater systems, this would take on the order of tens to several hundred years. The resultant degassing of carbonic acid in arid environments will result in reprecipitation of carbonate in the river or reservoir and releases CO_2 back into the atmosphere. If the water is recharged into deep aquifers, the net soil flux of dissolved IC is preserved as an IC sink. In acid environments, liming of soils results in CO_2 release to the atmosphere as there is little or no net alkalinity produced. Examination of records of volumes of water and concentrations of alkalinity (mostly bicarbonate) revealed that within the past 50 years, there has been a dramatic increase in net alkalinity export from the Mississippi River to the Gulf of Mexico.[18] The 17.7 Tg/yr increase in IC is attributed to increased rainfall in the basin and increased proton delivery to the land due to enhanced carbon dioxide production in the soil, resulting in increased weathering. It seems reasonable to consider that management changes for increased crop productivity may have increased soil liming and dissolution of IC. Such a drastic change does not likely reflect a response to a changing groundwater composition but is likely related to increased alkalinity in surface overland flow. The impact of a drier climate on the net export of alkalinity was not evaluated. However, it is reasonable to assume that decreased runoff would result in decreased mass of alkalinity transported to surface waters.

Conclusion

Land-use practices have a long-term impact on soil IC. Due to the large C pools in the soil, these impacts are not generally observed in short-term studies. Use of acidifying fertilizers such as ammonia and sulfur reduces soil IC. Practices that favor dissolution of carbonates include irrigation with surface waters and irrigation with water in large excess of plant transpiration. Practices that favor accumulation of IC in the soil include lower water applications relative to transpiration when irrigating in arid and semiarid regions (leaching less than 30% of the applied water), irrigation with groundwaters at elevated CO_2 concentrations, application of gypsum to alkaline soils, and use of nitrate fertilizer. Other factors that affect soil carbonate content include land clearing, cropping practices, and erosion.

In humid environments, the major anthropogenic impacts on inorganic C are liming of surface soils, use of fertilizers, removal of vegetation, and soil erosion. In semiarid and arid environments, increased IC is favored by use of groundwater vs. surface water for irrigation and application of gypsum amendments used to improve water infiltration. Decreased IC is favored by inefficient irrigation with surface water and application of NH_4 fertilizer. The net effect of irrigation on a global scale, neglecting the effects of fertilizer addition, is an estimated increase in soil inorganic C by 30 Tg/yr as well as a release of an almost equal amount of C to the atmosphere. Liming practices in humid regions throughout the world are estimated to have no net effect on inorganic soil C but release up to 85 Tg C/yr to the atmosphere. Of course this is the net effect from IC processes and does not include sequestration of OC by crop production.

References

1. Batjes, N.H. Total carbon and nitrogen in soils of the world. Eur. J. Soil Sci. **1966**, *47*, 151–163.
2. Cole, J.J.; Prairie, Y.T.; Caraco, N.F.; McDowell, W.H.; Tranvik, L.J.; Striegl, R.G.; Duarte, C.M.; Kortelainen, P.; Downing, J.A.; Middelburg, J.J.; Melack, J. Plumbing the global carbon cycle: Integrating inland waters into the terrestrial carbon budget. Ecosystems **2007**, *10*, 171–184.
3. Kessler, T.J.; Harvey, C.F. The global flux of carbon dioxide into groundwater. Geophys. Res. Lett. **2009**, *28*, 279–282.
4. Wu, H.; Guo, Z.; Gao, Q.; Peng, C. Distribution of soil inorganic carbon storage and its changes due to agricultural land use activity in China. Agric. Ecosyst. Environ. **2009**, *129*, 413–421.
5. Cihacek, L.J.; Ulmer, M.G. Effects of tillage on inorganic carbon storage in soils of the Northern Great Plains of the U.S. In *Agricultural Practices and Policies for Carbon Sequestration in Soil;* Kimble, J.M., Lal, R., Follett, R.F., Eds.; CRC Press: Boca Raton, 2002; 63–69.
6. Mikhailov, E.; Post, C.A. Effects of land use on soil inorganic carbon stocks in the Russian Chernozem. J. Environ. Qual. **2006**, *35*, 1384–1388.
7. Landi, A.; Mermut, A.R.; Anderson, D.W. Origin and rate of pedogenic accumulation in Saskatchewan soils, Canada. Geoderma **2003**, *117*, 143–156.
8. Voss, R.D. What constitutes an effective liming material. In *National Conference on Agricultural Limestone*; National Fertilizer Development Center: Muscle Shoals, AL, 1980; 52–61.
9. Suarez, D.; Rhoades, J. Effect of leaching fraction on river salinity. J. Irrig. Drain. Div., Am. Soc. Civ. Eng. **1977**, *103* (2), 245–257.

10. Suarez, D. Impact of agriculture on CO_2 as affected by changes in inorganic carbon. In *Global Climate Change and Pedogenic Carbonates;* Lal, R., Kimble, J., Eswaran, H., Stewart, B., Eds.; Lewis: Boca Raton, 1999; 257–272.
11. Solley, W.B.; Pierce, R.R.; Perlman, H.A. *Estimated Use of Water in the United States in 1990;* U.S. Geol. Surv. Circular 1004; U.S. Gov. Printing Office: Washington, DC, 1993.
12. Magaritz, M.; Amiel, A. Influence of intensive cultivation and irrigation on soil properties in the Jordan Valley, Israel: Recrystallization of carbonate minerals. Soil Sci. Soc. Am. J. **1981**, *45,* 1201–1205.
13. Amundson, R.G.; Smith, V.S. Effects of irrigation on the chemical properties of a soil in the Western San Joaquin Valley, California. Arid Soil Res. Rehabil. **1988**, *2*, 1–17.
14. Amundson, R.D.; Lund, L. The stable isotope chemistry of a native and irrigated Typic Natrargid in the San Joaquin Valley of California. Soil Sci. Soc. Am. J. **1987**, *51,* 761–767.
15. Entry, J.A.; Sojka, R.E.; Shewmaker, G.E. Irrigation increases inorganic carbon in agricultural soils. Environ. Manage. **2004**, *33,* s309–s317.
16. Wu, L.; Wood, Y.; Jiang, P.; Li, L.; Pan, G.; Lu, J.; Chang, A.C.; Enloe, H.A. Carbon sequestration and dynamics of two irrigated agricultural soils in California. Soil Sci. Soc. Am. J. **2007**, *72*, 808–814.
17. Suarez, D.L. Impact of agriculture on soil inorganic carbon. Annual Meetings, Baltimore, MD, Oct. 18–23, 1998; Agronomy Abstracts; Soil Science Society of America: Madison, WI, 1998; 258–259.
18. Raymond, P.A.; Cole, J.J. Increase in the export of alkalinity from North America's largest river. Science **2003**, *301,* 88–91.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

19

Chromium

Bruce R. James and
Dominic A. Brose

Introduction

Chromium is a heavy metal that is essential for human health as a cofactor of insulin in its trivalent form [(Cr(III)] but may cause lung cancer if inhaled in the hexavalent oxidation state [Cr(VI)]. Trivalent Cr is only sparingly soluble in neutral to alkaline natural waters as $Cr(OH)_3$ or Cr_2O_3, but it can be oxidized to Cr(VI) by naturally occurring manganese (III,IV) (hydr)oxides and by hydrogen peroxide, ozone, chlorine gas, hypochlorite, and other electron acceptors used in the environmental remediation of water, sediments, and soils. Hexavalent Cr can be reduced to Cr(III) by elemental Fe and Fe(II), sulfides (H_2S and HS^-), easily oxidized organic C compounds, and other electron donors. The rates and extent of oxidation and reduction reactions of Cr are governed by the redox potential (Eh) and acidity (pH) of natural soil–water systems and of engineered environments, such as anthropogenic wetlands and wastewater treatment facilities.

Health Effects

Concerns surrounding the presence of chromium (Cr) in natural waters and drinking water supplies must address the contrasting solubilities and toxicities of its common oxidation states in natural environments: Cr(III) and Cr(VI). Currently, Cr regulation principally focuses on “total Cr,” without distinguishing the oxidation states in a solid or water sample. The U.S. Environmental Protection Agency’s (EPA) national standard for total Cr in drinking water is 100 μg/L (100 ppb), except that California has set its current drinking water standard at 50 μg/L.[1,2] In 1999, California set a Public Health Goal of 2.5 μg/L, based on a 1968 study in Germany that found stomach tumors in animals that drank chromium-enriched water. The U.S. EPA rejected that study as flawed and determined that there was no evidence it was carcinogenic in water, which resulted in California rescinding its goal in 2001 and reverting back to the 50 μg/L standard.[2] The point of contention regarding the 1968 study was whether Cr(VI) is reduced to Cr(III) in the stomach by gastric acids.

Chromium(VI) is genotoxic in a number of in vitro and in vivo toxicity tests.[3] Because the mechanisms of geno-toxicity and carcinogenicity are not fully understood, the National Toxicology Program (NTP) conducted animal tests to assess the potential for cancer due to ingestion of Cr(VI).[4] Reduction of Cr(VI) to Cr(III) is hypothesized to occur primarily in the stomach, as a mechanism of detoxification. In this 2 years NTP study, no neoplasms or non-neoplastic lesions were observed in the forestomach or glandular stomach of rats or mice. However, observed increases in neoplasms, or abnormal growths of tissue, in the small intestine of mice, toxicity to red and white blood cells and bone marrow, and uptake of Cr(VI) into tissues of rats and mice suggest that at least a portion of the administered Cr(VI) was not reduced in the stomach.[4] The stepwise reduction of Cr(VI) to Cr(III) through Cr(V) and Cr(IV) creates short-lived oxidation states that may be the forms of Cr that are the actual carcinogens.

This finding, in addition to the absence of increases in neoplasms or non-neoplastic lesions of the small intestine in rats or mice exposed to chromium picolinate monohydrate (CPM), an organically bound form of Cr(III),[5] provides evidence that Cr(VI) is not completely reduced in the stomach and is responsible for these carcinogenic effects. Additionally, it should be noted that Cr(III), like that found in CPM, is essential for human health in trace amounts as an activator of insulin,[6] but exists predominantly in nature in cationic forms that are typically only sparingly soluble in near-neutral pH soils, plants, cells, and natural waters.[7]

Sources of Chromium and Occurrence in Natural Waters and Water Supplies

When soluble Cr is detected in natural waters, especially at high concentrations, it is usually Cr(VI) derived from industrial wastes containing Cr(VI) or possibly resulting from the oxidation of certain forms of Cr(III) in soils or sediments.[8,9] Chromium is the seventh most abundant metal on earth with an average content of 100 mg/kg in the earth's crust and 3700 mg/kg for the earth as a whole, principally as Cr(III) in unreactive, insoluble miner-als, such as chromite ($FeO.Cr_2O_3$).[10] Roasting chromite ore under alkaline, high-temperature conditions oxidizes Cr_2O_3 to soluble Cr(VI), a widely used starting material for production of stainless steel, pressure-treated lumber, chrome-tanned leather, pigments, chrome-plated metals, and other common products used in modern societies.[11] As a result, Cr(VI) remaining in industrial by-products, such as chromite ore processing residue, chrome plating bath waste, paint aerosols, and other industrial wastes, may enrich soils and contaminate surface waters and groundwater that are supplies for domestic uses, irrigation, and industrial processes.

In contrast to these concentrated, anthropogenic sources of Cr(VI); naturally occurring sources of Cr are predominantly Cr(III) and occur at low concentrations. Ultramafic and basaltic rocks (and soils developed from these parent materials), however, may contain up to 2400 mg Cr/kg, and can release small fractions of the Cr contained in them as Cr(VI), either through dissolution of Cr(VI) minerals or possibly via oxidation of Cr(III) by Mn(III,IV) (hydr)oxides. As a result, Cr(VI) has been detected in groundwater (0.05–0.5 mg/L) in arid regions dominated by these alkaline, Cr-rich rocks and soils. A concentration of Cr(VI) of 7.5 mg/L in pH 12.5 groundwater from Jordan is the highest known level that is not due to human influence. Naturally occurring Cr in alkaline, aerobic ocean water exists principally as Cr(VI) at concentrations in the range of 3–7.3 nM (0.16–0.38 μg/L).[12]

The balance of the different forms and the solubilities of Cr(III) and Cr(VI) in natural waters is governed by pH, aeration status, and other environmental conditions (Table 1). Understanding and predicting the oxidation state, solubility, mobility, and bioavailability of Cr in water are further complicated by the fact that Cr(III) can be oxidized (lose three electrons) to form Cr(VI), whereas Cr(VI) can gain three electrons and be reduced to Cr(III).[13,14] Natural variation and human-induced changes in pH and the oxidation–reduction status of soil and water can control the solubility of Cr. As a result, purification of drinking water supplies and treatment of wastewaters contaminated with Cr are possible through chemical and microbiological processes that modify the acidity and the relative abundance of oxidizing and reducing agents for Cr.[15,16]

Solubility Controls of Chromium Concentrations in Water

Most inorganic compounds of Cr(III) are less soluble in water than are those of Cr(VI) because Cr(III) cations have high ionic potentials (charge-to-size ratio) and hydrolyze to form covalent bonds with OH^- ions (Table 1). When three OH^- anions surround the Cr^{3+} cation, it is particularly stable in water as the sparingly soluble compound, $Cr(OH)_3$ (Table 2). Upon aging and dehydration, $Cr(OH)_3$ slowly converts to the more crystalline, less soluble Cr_2O_3.[12] Incorporation of Fe(III) or Fe(II) into solid phases and precipitates containing Cr(III) renders the Cr(III) less soluble, often by a factor of 10^3 in the

TABLE 1 Oxidation States and Forms of Chromium in Natural Waters

Oxidation State	Form	Name	Chemical Conditions of Water under Which It Is Found and Pertinent Reaction in Natural Waters
Chromium (III) (trivalent chromium)	$Cr(H_2O)_6^{3+}$	Hexaquochromium(III)	pH < 3.5; strong affinity for negatively charged ions (e.g., phosphate) and colloid surfaces (e.g., living cells and phyllosilicate clays or fulvic and humic acids); green color
	$Cr(H_2O)_5OH^{2+}$	Monohydroxychromium(III)	First hydrolysis product formed at pH > 3.5 upon dilution of or addition of base to solution of Cr(III); green
	$Cr(H_2O)_4(OH)_2^+$	Dihydroxychromium(III)	Second hydrolysis product of Cr(III); may dimerize and polymerize to form high molecular weight cations in planes of octahedron; green
	$Cr(H_2O)_3(OH)_3^0$	Chromium hydroxide	Metastable, uncharged hydrolysis product that precipitates as the sparingly soluble $Cr(OH)_3$
	$Cr(H_2O)_2(OH)_4^-$	Hydroxochromate	Fourth hydrolysis product of Cr(III) that may form at pH > 11; may oxidize to Cr(VI) by O_2
	Cr(III)–organic acid complexes and chelates	For example: chromium citrate, chromium picolinate, chromium fulvate	Soluble complexes and chelates in which water molecules of hydration surrounding $Cr(H_2O)_6^{3+}$ are displaced by carboxylic acid and N-containing ligands; formation is pH and concentration dependent; blue–green–purple colors, depending on ligand binding Cr(III)
Chromium (VI) (hexavalent chromium)	H_2CrO_4	Chromic acid	Fully protonated form of Cr(VI) formed at pH < 1; see Figure 2 for key Eh values for redox
	$HCrO_4^-$	Bichromate	Form of Cr(VI) that predominates at 1 < pH < 6.4; yellow; see Figure 2 for key Eh values for redox
	CrO_4^{2-}	Chromate	Form of Cr(VI) that predominates at pH > 6.4; yellow; see Figure 2 for key Eh values for redox
	$Cr_2O_7^{2-}$	Dichromate	Form of Cr(VI) that predominates at pH < 3 and in concentrated solutions (>1.0 mM); rapidly reverts to $HCrO_4^-$ or CrO_4^{2-} upon dilution or pH change

TABLE 2 Solubility in Water at pH 7 of Selected Chromium Compounds

Oxidation State	Compound Name	Formula	Approximate Solubility (mol Cr/L)
Chromium (III) (trivalent chromium)	Chromium(III) hydroxide	$Cr(OH)_3$ (am)	10^{-12}
	Chromium(III) oxide	Cr_2O_3 (Cr)	10^{-17}
	Chromite	$FeO{\cdot}Cr_2O_3$ (Cr)	10^{-20}
	Chromium chloride	$CrCl_3$	Highly soluble
	Chromium sulfate	$Cr_2(SO_4)_3$	Highly soluble
	Chromium phosphate	$CrPO_4$	10^{-10}
	Chromium fluoride	CrF_3	1.2×10^{-3}
	Chromium arsenate	$CrAsO_4$	10^{-10}
Chromium (VI) (hexavalent chromium)	Potassium chromate	K_2CrO_4	3.2
	Sodium chromate	Na_2CrO_4	5.4
	Calcium chromate	$CaCrO_4$	0.14
	Barium chromate	$BaCrO_4$	1.7×10^{-3}
	"Zinc yellow" pigment	$3ZnCrO_4{\cdot}K_2CrO_4{\cdot}Zn(OH)_2H_2O$	8.2×10^{-3}
	Strontium chromate	$SrCrO_4$	5.9×10^{-3}
	Lead chromate	$PbCrO_4$	1.8×10^{-6}
	Chromium(VI) jarosite	$KFe_3(CrO_4)_2(OH)_6$ (Cr)	10^{-30}

solubility product (Ksp).[17,18] In the pH range of 5.5–8.0, Cr(III) reaches minimum solubility in water due to this hydrolysis and precipitation reaction, an important process that controls the movement of Cr(III) in soils enriched with industrial wastewaters and solid materials. Under strongly acidic conditions ($pH < 4$), unhydrolyzed $Cr(H_2O)_6^{3+}$ cations exist in solution, while $Cr(OH)4^-$ forms under strongly alkaline conditions ($pH > 11$), particularly in response to adding base to solutions of soluble salts of Cr(III), e.g., $CrCl_3$, $Cr(NO_3)_3$, or $Cr_2(SO_4)_3$.

Other anions besides OH^- coordinate with $Cr(H_2O)_6^{3+}$ and displace water molecules of hydration to form sparingly soluble compounds and soluble chelates (Table 2). In water treatment facilities and in natural waters, phosphate ($H_2PO_4^-$, HPO_4^{2-}, PO_4^{3-}), arsenate ($H_2AsO_4^-$, $HAsO_4^{2-}$) and fluoride (F^-) may form low solubility compounds with Cr(III). Organic complexes of Cr(III) with carboxylic acids (e.g., citric, oxalic, tartaric, fulvic) remain soluble at pH values above which $Cr(OH)_3$ forms. By increasing the solubility of Cr(III) in neutral and alkaline waters, such organic complexes enhance the potential for absorption of Cr(III) by cells. Stable, insoluble complexes of Cr(III) also form with humic acids and other high molecular aggregate weight organic moieties in soils, sediments, wastes, and natural waters.[19]

With the exception of chromium(VI) jarosite (Table 2), Cr(VI) compounds are more soluble over the pH range of natural waters than are those of Cr(III), thereby leading to the greater concern about the potential mobility and bioavailability of Cr(VI) than Cr(III) in natural waters. The alkali salts of Cr(VI) are highly soluble, $CaCrO_4$ is moderately soluble, and $PbCrO_4$ and $BaCrO_4$ are only sparingly soluble. In colloidal environments containing aluminosilicate clays and (hydr)oxides of Al(III), Fe(II,III), and Mn(III,IV) (e.g., in soils and sediments), Cr(VI) anions may be adsorbed similarly to SO_4^{2-}. Low pH and high ionic strength promote retention of $HCrO_4^-$ and CrO_4^{2-} on positively charged sites, especially those associated with colloidal surfaces dominated by pH-dependent charge. Such electrostatic adsorption may be reversible, or the sorbed Cr(VI) species may gradually become incorporated into the structure of the mineral surface (chemisorption). Recently precipitated $Cr(OH)_3$ can adsorb Cr(VI) or incorporate Cr(VI) within its structure as it forms, thereby forming a Cr(III)–Cr(VI) compound.[20]

Oxidation-Reduction Chemistry of Chromium in Natural Waters

The paradox of the contrasting solubilities and toxicities of Cr(III) and Cr(VI) in natural waters and living systems is complicated by two reduction–oxidation (electron transfer) reactions: Cr(III) can oxidize to Cr(VI) in soils and natural waters, and Cr(VI) can reduce to Cr(III) in the same systems, and at the same time. Understanding the key electron transfer processes (redox) and predicting environmental conditions governing them are central to treatment of drinking water, wastewaters, and contaminated soils, and to predicting the hazard of Cr in natural systems.[21] The metaphor of a seesaw (Figure 1) is useful in picturing the undulating nature of the changes in Cr speciation in water due to oxidation of Cr(III) and reduction of Cr(VI). A balance for the two redox reactions is achieved in accordance with the quantities and reactivities of reductants and oxidants in the system [e.g., organic matter and Mn(III,IV) (hydr)oxides, as modulated by pH and pe as a master variables].[15,22]

The thermodynamics (energetics predicting the relative stability of reactants and products of a chemical reaction) of interconversions of Cr(III) and Cr(VI) compared to other redox couples can be used to predict the predominance of Cr(III) or Cr(VI) in water supplies (Figure 2). Certain electron-poor species may act as oxidants (electron acceptors) for Cr(III), especially soluble forms of Cr(III), in the treatment of water supplies or in soils enriched with Cr(III) (Nieboer and Jusys, Figure 3).[19] Examples are those above the bold line for Cr(VI)–Cr(III) on the Eh–pH diagram: Cl_2, OCl^-, H_2O_2, O_3, and MnOOH. In contrast, electron-rich species may donate electrons to electron-poor Cr(VI) and reduce it to Cr(III): Fe^{2+} [or Fe(0)], H_2S, H_2, ascorbic acid (and organic compounds, generally), and SO_2. Sunlight may affect the kinetics of both oxidation and reduction reactions for Cr, a relevant fact for natural processes in lakes and streams and for treatment technologies for drinking water purification. Depending on pH, temperature, and the concentrations of oxidants and reductants, Cr(VI)-to-Cr(III) ratio in natural waters may be predicted.

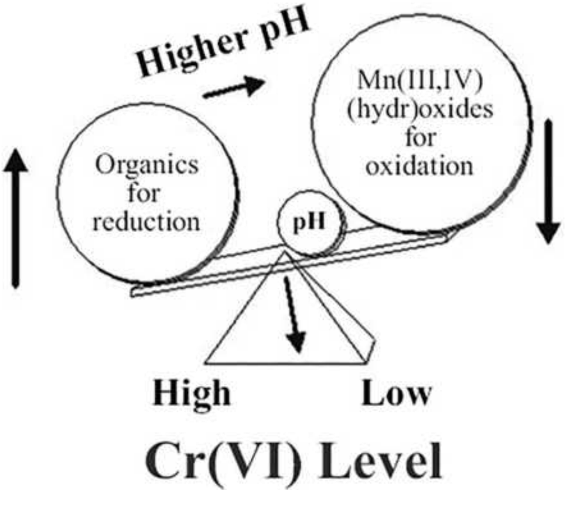


FIGURE 1 Seesaw model depicting a balance of the oxidation of Cr(III) by Mn(III,IV) (hydr)oxides and the reduction of Cr(VI) by organic compounds, with the pH acting as a sliding control (master variable) on the seesaw to set the redox balance for given quantities and reactivities of oxidants and reductants. The equilibrium quantity of Cr(VI) in the water is indicated by the pointing arrow from the fulcrum.

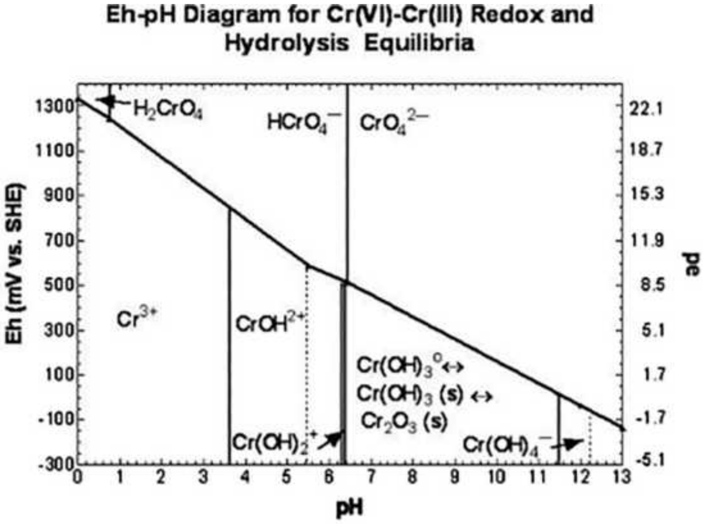


FIGURE 2 Eh–pH diagram illustrating the stability field defined by Eh (redox potential relative to the standard hydrogen electrode, SHE) and pH for Cr(VI) and Cr(III) at 10^{-4} M total Cr. The vertical dashed lines indicate semiquantitatively the pH range in which $Cr(OH)_3$ is expected to control Cr(III) cation activities in the absence of other ligands besides OH^-.

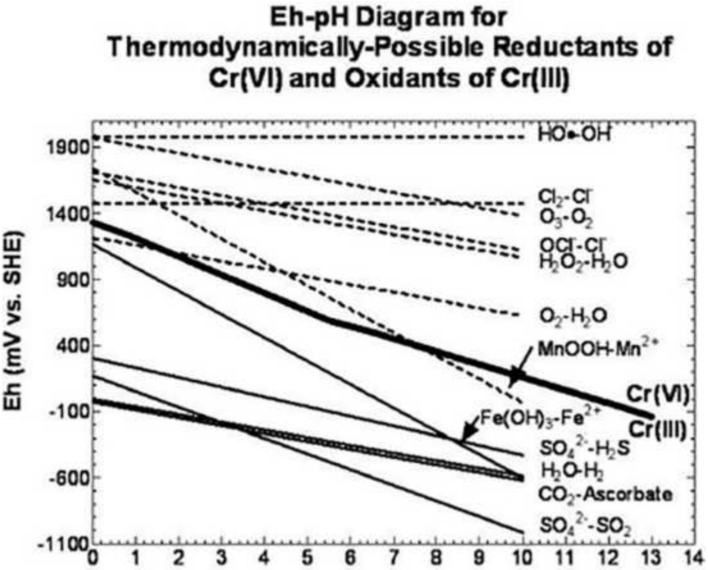


FIGURE 3 Eh–pH diagram showing potential oxidants for Cr(III) in natural waters as dashed lines above the bold Cr(VI)–Cr(III) line and potential reductants for Cr(VI) below the line. Each line for an oxidant (first species of the pair) and reductant (second species) combination represents the reduction potential (in mV) at a given pH established by that oxidant–reductant pair (e.g., O_3-O_2). The oxidant member of a pair for a higher line is expected to oxidize the reductant member of the lower line, thereby establishing the area and species between the lines as thermodynamically favored to exist at chemical equilibrium.[23]

Conclusion

Hexavalent chromium remaining in industrial by-products may contaminate soils, surface waters, and groundwater that are supplies for domestic uses, irrigation, and industrial processes. Prediction of the likelihood of Cr(III) oxidation and Cr(VI) reduction occurring is important for water treatment and for establishing health-based regulations and allowable limits for Cr(VI) and Cr(III) in water supplies. In agricultural soil–plant–water systems, Cr(VI) added in irrigation water or formed via oxidation of Cr(III) will reduce to Cr(VI) if electron donors (e.g., Fe^{2+}, H_2S, and organic matter) and Eh–pH conditions are sufficiently reducing (Bartlett and James[8] and James and Bartlett[20]; Figure 2). If not reduced, Cr(VI) may leach from surface soils to subsoils and groundwater. Therefore, prediction of Cr bioavailability and mobility in natural waters must consider redox reactions of this heavy metal.

References

1. US Environmental Protection Agency (EPA). Basic Information about Chromium in Drinking Water, 2010. Available at http://www.epa.gov/safewater/contaminants/basicinformation/chromium.html#four. (Accessed on May 13, 2012).
2. California Department of Public Health (DPH). *Chromium-6: Timeline for Drinking Water Regulations,* 2009. Available at: http://www.cdph.ca.gov/certlic/drinkingwater/Pages/Chromium6timeline.aspx. (Accessed May 13, 2012).
3. International Agency for Research on Cancer (IARC). Chromium and chromium compounds. IARC Monogr. Eval. Carcinog. Risks Hum. **1990**, *49,* 49–256.
4. Stout, M.D.; Herbert, R.A.; Kissling, G.E.; Collins, B.J.; Travlos, G.S.; Witt, K.L.; Melnick, R.L.; Abdo, K.M.; Malarkey, D.E.; Hooth, M.J. Hexavalent chromium is carcinogenic to F344/N rats and B6C3F1 mice after chronic oral exposure. Environ. Health Perspect. **2009**, *117* (5), 716–722.
5. National Toxicology Program (NTP). Toxicology and carcinogenesis studies of chromium picolinate monohydrate (CAS No. 27882-76-4) in F344/N rats and B6C3F1 mice (feed studies). TR556. 2010. Research Triangle Park, N.C.
6. Anderson, R.A. Essentiality of chromium in humans. Sci. Total Environ. **1989**, *86,* 75–81.
7. Kimbrough, D.E.; Cohen, Y.; Winer, A.M.; Creelman, L.; Mabuni, C.A. Critical Assessment of Chromium in the Environment. Crit. Rev. Environ. Sci. Technol. **1999**, *29* (1), 1–46.
8. Bartlett, R.; James, B. Behavior of chromium in soils: III. Oxidation. J. Environ. Qual. **1979**, 8, 31–35.
9. James, B.R.; Petura, J.C.; Vitale, R.J.; Mussoline, G.R. Oxidation–reduction chemistry of chromium: Relevance to the regulation and remediation of chromate-contaminated soils. J. Soil Contam. **1997**, *6* (6), 569–580.
10. Nriagu, J.O. Production and uses of chromium. In *Chromium in the Natural and Human Environments*; Nriagu, J.O., Nieboer, E., Eds.; Wiley-Interscience: New York, **1988**; 81–103.
11. Barnhart, J. Chromium chemistry and implications for environmental fate and toxicity. J. Soil Contam. **1997**, *6* (6), 561–568.
12. Ball, J.W.; Nordstrom, D.K. Critical evaluation and selection of standard state thermodynamic properties for chromium metal and its aqueous ions, hydrolysis species, oxides, and hydroxides. J. Chem. Eng. Data **1998**, *43,* 895–918.
13. Bartlett, R.J. Chromium redox mechanisms in soils: Should we worry about Cr(VI)? In *Chromium Environmental Issues;* FrancoAngeli: Milan, 1997; 1–20.
14. James, B.R. Remediation-by-reduction strategies for chromate-contaminated soils. Environ. Geochem. Health **2001**, *23* (3), 175–179.
15. James, B.R. The challenge of remediating chromium-contaminated soils. Environ. Sci. Technol. **1996**, *30,* 248A–251A.

16. Fendorf, S.; Wielenga, B.W.; Hansel, C.M. Chromium transformations in natural environments: The role of biological and abiological processes in chromium(VI) reduction. Int. Geol. Rev. **2000**, *42,* 691–701.
17. Rai, D.; Saas, B.M.; Moore, D.A. Chromium(III) hydrolysis constants and solubility of chromium(III) hydroxide. Inorg. Chem. **1987**, *26*, 345–369.
18. Sass, B.M.; Rai, D. Solubility of amorphous chromium(III)–Iron(III) hydroxide solid solutions. Inorg. Chem. **1987**, *26,* 2228–2232.
19. Nieboer, E.; Jusys, A.A. Biologic chemistry of chromium. In *Chromium in the Natural and Human Environments;* Nriagu, J.O., Nieboer, E., Eds.; Wiley-Interscience: New York, 1988; 21–80.
20. James, B.R.; Bartlett, R.J. Behavior of chromium in soils: VII. Adsorption and reduction of hexavalent forms. J. Environ. Qual. **1983**, *12,* 177–181.
21. James, B.R. Redox phenomena. In *Encyclopedia of Soil Science,* 1st Ed.; Lal, R., Ed.; Marcel Dekker, Inc.: New York, 2002; 1098–1100.
22. James, B.R.; Brose, D.A. Oxidation - Reduction Phenomena. In *Handbook of Soil Sciences: Properties and Processes.* 2nd edn.; Huang, P.M., Li, Y., Summer, M.E., Ed.; CRC Press: Boca Raton, 2012; 14-1–14-24.
23. James, B.R. Chromium. In *Encyclopedia of Water Science;* Stewart, B.A.; Howell, T., Ed., Marcel Dekker: New York, 2003; 75–79.

20

Cobalt and Iodine

Ronald G. McLaren

Soil Cobalt

The Co concentration in soils depends primarily on the parent materials (rocks) from which they were formed and on the degree of weathering undergone during soil development.[2] Cobalt tends to be most abundant as a substituent ion in ferromagnesian minerals, and therefore has relatively high concentrations in mafic and ultramafic rocks (rocks containing high or extremely high proportions of ferromagnesian minerals). Conversely, Co concentrations are relatively low in felsic rocks (rocks containing large amounts of silica-rich minerals) such as granite, and in coarse-textured quartz-rich sedimentary rocks (sandstones). Higher concentrations of Co may be associated with finer textured sediments (shales) in which Co has become surface adsorbed by, or incorporated into, secondary layer silicates by isomorphous substitution.[3] Typical concentrations of Co reported in different rock types are shown in Table 1. As a result of the large range in Co concentrations of soil parent materials, and variation in the degree of weathering, total soil Co concentrations also vary widely. However, the mean values reported for agricultural soils from many countries appear to have a somewhat restricted range of between approximately 2 and 20 mg/kg (Table 1).

Forms of Cobalt in Soils

Cobalt in soils, whether released from parent materials during soil development, or derived from anthropogenic contaminant sources, occurs in several different forms or associations. Cobalt may be present as 1) the simple Co^{2+} ion, or as complexes with various organic or inorganic ligands in the soil solution; 2) exchangeable Co^{2+} ions; 3) specifically adsorbed Co, bound to the surfaces of inorganic soil colloids (clays and oxides/hydrous oxides of Al, Fe, and Mn); 4) Co complexed by soil organic colloids; 5) Co occluded by soil oxide materials; and 6) Co present within the crystal structures of primary and secondary silicate minerals.[6] In some soils, there appears to be a particularly strong association between Co and manganese (Mn) oxides, especially in soils where Mn oxides occur as distinct nodules or coatings.[7,8]

TABLE 1 Cobalt and Iodine Concentrations in Rocks and Soils

	Co Concentration (mg/kg)	I Concentration (mg/kg)
Rock type		
Ultramafic (e.g., serpentinite)	100–300	0.01–0.5
Mafic (e.g., basalt, gabbro)	30–100	0.08–50
Intermediate (e.g., diorites)	1–30	0.3–0.5
Felsic (e.g., granites, gneiss)	<1–10	0.2–0.5
Sandstones	0.3–10	0.5–1.5
Shales/argillites	11–40	2–6
Limestones	0.1–3.0	0.5–3.0
Soils		
Complete range	0.1–300	<0.1–25.4
Range of mean values	2–21.5	1.1–13.1

Source: Kabata-Pendias and Pendias,[1] Aubert and Pinta,[4] and Swaine.[5]

Plant Availability of Soil Cobalt

The immediate source of Co for plant uptake is the soil solution. However, Co concentrations in the soil solution are extremely low, generally much less than 0.1 μg/mL.[1] Soil solution Co appears to be in equilibrium with Co adsorbed at the surfaces of soil colloids, and the distribution between solution and surface phases is strongly influenced by soil pH. As pH increases, soluble Co decreases.[6] Similarly, soils with high capacities to adsorb Co, particularly those soils with high Mn oxide contents, also have low solution Co concentrations.[9] Thus, in addition to soils with low total Co concentrations, the plant availability of Co may be low in soils with high pH or high Co adsorption capacities. Cobalt availability is also influenced by soil moisture status, availability increasing under waterlogged conditions.[10]

Determination of soil Co dissolved by various extractants is the most common way of assessing the plant availability of soil Co. Ideally, the forms of Co extracted should include any soluble and exchangeable Co, together with any solid-phase forms of Co that are able to move readily into the soil solution in response to changes in solution Co concentrations.[6] The two extractants used most commonly for this purpose are 2.5% acetic acid and solutions of ethylenediaminetetraacetic acid (EDTA). However, the ability of these extractants to accurately assess Co availability appears to be somewhat limited.[11,12]

Cobalt Deficiency

Cobalt deficiency in grazing sheep and cattle was first diagnosed in the 1930s, initially in New Zealand, Australia, and Scotland.[13–15] The condition causes animals to lose their appetite, become weak and emaciated, suffer severe anemia, and eventually die. Subsequently, it was shown that Co is a constituent of both vitamin B_{12} and a closely related coenzyme, and that Co deficiency is in effect a deficiency of vitamin B_{12}.[16] It has been concluded from field studies that pasture containing Co below 0.08 mg/kg for sheep or below 04 mg/kg for cattle is unlikely to meet nutritional requirements in terms of maintaining adequate serum and liver vitamin B_{12} concentrations, and healthy growth rates.[17,18]

Cobalt deficiency is commonly treated by applying Co-containing fertilizers to pastures, usually at very low rates, e.g., 350 g/ha/yr of $CoSO_4 \cdot 7H_2O$. However, on some soils such treatments appear to be ineffective.[11] Alternative treatments for Co deficiency involve injecting the animal with vitamin B_{12}, the use of Co drenches, or the insertion of slowly dissolving Co “bullets” in the animal’s rumen.

Soil Iodine

The range of I concentrations found in rocks and soils is shown in Table 1. Iodine occurs as a minor constituent in various minerals, where it can replace anions such as OH^-, SiO_4^{4-}, and CO_3^{2-}, and has relatively low concentrations in most types of rock.[19] Highest I concentrations are generally found in fine-grained sedimentary rocks (Table 1). However, soil I concentrations are generally higher than in the rocks from which they have been derived, a fact attributed predominantly to atmospheric accessions of I.[20] Iodine is known to be present in the atmosphere in vapor form and associated with particles of dust. In coastal areas, accession of I may also be related to sea spray.[1]

Forms of Iodine in Soils

Information of the forms of I in soils is limited, most published analyses of soils have determined total I concentrations only. Of the three most common forms of elemental (I_2), iodide (I^-), and iodate (IO_3^-), it seems likely that most I in soils is present as iodide or possibly as elemental I. The presence of iodate in soils has also been postulated, but would require high oxidation conditions in neutral or alkaline soils.[20] Indeed, there is some evidence that when iodate is added to soils it is rapidly reduced to elemental I or iodide by soil organic matter.[21] There is also evidence that, under some conditions, elemental I can be volatilized from soils.[22]

Most I in soils appear to be associated with soil organic matter and iron (Fe) and aluminium (Al) oxides, materials by which both iodide and elemental I are known to be strongly adsorbed.[21,23] Indeed the distribution of I in soil profiles appears to follow the distribution of these soil components.[24] The atmospheric accessions of I and the affinity between I and organic matter often result in maximum concentrations of I in the surface horizons of soils.[20,24]

Plant Availability of Soil Iodine

Interest in soil analysis as a means of assessing plant availability of I has been minimal,[20] and relationships between the I-status of soils and the concentrations of I in plants appear to be poor.[1] Even soil extractants designed to determine the most soluble forms of I in soils do not provide a good indication of I availability to plants.[25] Plants are capable of absorbing I directly from the atmosphere,[1] and plant species or varietal differences appear to have a greater influence on plant I concentration than soil I status.[20,26] Dicotyledonous pasture species (e.g., clovers) generally have higher I contents than do grasses.[26] Concentrations of I in plants may be reduced by liming,[25] by the application of N fertilizer,[26] and by the application of farmyard manure.[25] Seasonal effects on pasture I concentrations have also been observed, with decreases in the summer, and slight increases in the autumn.[26]

Iodine Deficiency

Low concentrations of I in food and water have been associated with the occurrence of endemic goitre (enlargement of the thyroid gland) in humans and farm livestock.[27] Early work suggested a close relationship between goitre incidence and low soil levels of I, however, it is now recognized that other factors are also involved. In particular, the presence of a group of substances known as goitrogens, which occur in various plant species, has been shown to reduce thyroid hormone synthesis and metabolism.[20] In the absence of goitrogens, it is considered that diets containing 0.5mg I/kg DM will more than adequately meet the I requirements of all classes of animals, while levels as low as 0.15 mg/kg might be sufficient to meet the requirements for growing animals.[28] In the presence of goitrogens, I requirements may be as high as 2 mg I/kg DM.[28] Attempts to increase pasture I concentrations with I-containing fertilizers have been generally ineffective, with very low recoveries of the applied I.[25,26] Direct supplementation of livestock is normally the preferred way to increase I intakes.

Conclusions

Cobalt and Iodine deficiencies in livestock result from low soil concentrations and/or low availability of these trace elements for uptake by pasture plants. Plant availabilities of Co and I are determined by several factors including soil forms, soil sorption properties, soil pH, soil moisture status, season, plant species, and fertilizer applications. Deficiencies of Co and I can be prevented by application of fertilizers (Co) to the soil, or by direct treatment of livestock (Co and I).

References

1. Kabata-Pendias, A.; Pendias, H. *Trace Elements in Soils and Plants*; CRC Press: Boca Raton, 1984; 238–246.
2. Mitchell, R.L. Cobalt in soil and its uptake by plants. In *Atti del IX Simposio Internazionale di Agrochimica su La Fitonutrizione Oligominerale*; Punta Ala: Italy, 1972; 521–532.
3. Hodgson, J.F. Chemistry of micronutrient elements in soils. Adv. Agron. **1963**, *15,* 119–159.
4. Aubert, H.; Pinta, M. *Trace Elements in Soils, Development in Soil Science 7;* Elsevier Scientific Publishing Co.: Amsterdam, 1977; 395 pp.
5. Swaine, D.J. *The Trace Element Content of Soils;* CAB: Harpenden, England, 1955; 157 pp.
6. McLaren, R.G.; Lawson, D.M.; Swift, R.S. The forms of cobalt in some Scottish soils as determined by extraction and isotopic exchange. J. Soil Sci. **1986**, *37,* 223–234.
7. Taylor, R.M.; McKenzie, R.M. The association of trace elements with manganese minerals in Australian soils. Aust. J. Soil Res. **1966**, *4,* 29–39.
8. Jarvis, S.C. The association of cobalt with easily reducible manganese in some acidic permanent grassland soils. J. Soil Sci. **1984**, *35,* 431–438.
9. Tiller, K.G.; Honeysett, J.L.; Hallsworth, E.G. The isotopically exchangeable form of native and applied cobalt in soils. Aust. J. Soil Res. **1969**, *7*, 43–56.
10. Adams, S.N.; Honeysett, J.L. Some effects of soil waterlogging on Co and Cu status of pasture plants grown in pots. Aust. J. Agric. Res. **1964**, *15,* 357–367.
11. McLaren, R.G.; Lawson, D.M.; Swift, R.S.; Purves, D. The effects of cobalt additions on soil and herbage cobalt concentrations in some S.E. Scotland pastures. J. Agric. Sci. Camb. **1985**, *105,* 347–363.
12. McLaren, R.G.; Lawson, D.M.; Swift, R.S. The availability to pasture plants of native and applied soil cobalt in relation to extractable soil cobalt and other soil properties. J. Sci. Food Agric. **1987**, *39,* 101–112.
13. Grange, L.I.; Taylor, N.H. Bush sickness. Part IIA. The distribution and field characteristics of bush-sickness soils. Bull. N. Z. Dept. Sci. Ind. Res. **1932**, *32,* 21.
14. Underwood, E.J.; Filmer, J.F. The determination of the biologically potent element (cobalt) in limonite. Aust. Vet. J. **1935**, *11*, 84–92.
15. Corner, H.H.; Smith, A.M. The influence of cobalt on pine disease in sheep. Biochem. J. **1938**, *32,* 1800–1805.
16. Smith, R.M.; Gawthorne, J.M. The biochemical basis of deficiencies of zinc, manganese, copper and cobalt in animals. In *Trace Elements in the Soil-Plant-Animal System;* Nicholas, D.J.D., Egan, A.R., Eds.; Academic Press: New York, 1975; 243–258.
17. Andrews, E.D. Observations on the thrift of young sheep on a marginally cobalt deficient area. N. Z. J. Agric. Res. **1965**, *8*, 788–817.
18. Gardner, M.R. *Cobalt in Ruminant Nutrition: A Review;* Department of Agriculture: Western Australia, 1977; 602–620.
19. Goldschmidt, V.M. *Geochemistry;* Oxford University Press: Oxford, 1954.
20. Fleming, G.A. Essential micronutrients. II. Iodine and selenium. In *Applied Soil Trace Elements*; Davies, B.E., Ed.; John Wiley and Sons, Ltd.: Chichester, 1980; 199–234.

21. Whitehead, D.C. The influence of organic matter, chalk, and sesquioxides on the solubility of iodide, elemental iodine, and iodate incubated with soil. J. Soil Sci. **1974**, *25,* 461–470.
22. Whitehead, D.C. The volatilisation, from soils and mixtures of soil components, of iodine added as potassium iodide. J. Soil Sci. **1981**, *32*, 97–102.
23. Whitehead, D.C. The sorption of iodide by soils as influenced by equilibrium conditions and soil properties. J. Sci. Food. Agric. **1973**, *24,* 547–556.
24. Whitehead, D.C. Iodine in soil profiles in relation to iron and aluminium oxides and organic matter. J. Soil Sci. **1978**, *29*, 88–94.
25. Whitehead, D.C. Uptake by perennial ryegrass of iodide, elemental iodine and iodate added to soil as influenced by various amendments. J. Sci. Food. Agric. **1975**, *26,* 361–367.
26. Hartmans, J. Factors affecting the herbage iodine content. Neth. J. Agric. Sci. **1974**, *22,* 195–206.
27. Underwood, E.J. *Trace Elements in Human and Animal Nutrition,* 3rd Ed.; Academic Press: New York, 1971; 543 pp.
28. Agricultural Research Council. In *The Nutrient Requirements of Ruminant Livestock;* Commonwealth Agricultural Bureau: Slough, U.K., 1980; 351 pp.

21

Copper

David R. Parker and
Judith F. Pedler

Copper in Soils

The average total concentration of Cu in the Earth's crust is estimated to be 70 mg/kg, although levels of 20–30 mg/kg are prevalent in average soils.[1] Common primary minerals include Cu sulfides, with Cu largely in the +I oxidation state, which dissolve by weathering processes. Secondary minerals of Cu(II) include oxides, carbonates (malachite), silicates, sulfates, and chlorides, most of which are relatively soluble. Copper(II) may substitute for Fe, Mg, and Mn in an assortment of minerals, especially silicates and carbonates.[1]

The Cu^{2+} ion can form strong inner-sphere complexes, and is thus immobilized by carboxylic, carbonyl, or phenolic functional groups, even at low pH. Exchangeable and weak acid-extractable Cu represent a small percentage of total Cu in most soils. The bulk of the Cu is complexed by organic matter, occluded in oxides, and substituted in primary and secondary minerals. Organic matter and Mn oxides are the most likely materials to retain Cu in a nonexchangeable form in soils. Alkali extraction techniques that remove organic matter from soils usually release large fractions of the total soil Cu.[2] Addition of organic matter to soils, and biological exudation of organic acids may both increase dissolved organic carbon, thus solubilizing Cu from mineral forms, increasing the total dissolved Cu in soil solution,[3] but predictive models of humic acid binding of Cu in soil solution are generally inadequate.[4] Overall, Cu is one of the least mobile of the trace elements, maintained in a form sufficiently available to plants but relatively resistant to movement by leaching.[1]

Free Cu^{2+} in soil solution decreases with increasing pH, reaching a minimum above pH 10. In the absence of organic ligands, Cu speciation is dominated by free Cu^{2+}, and increasingly by carbonate and hydroxy complexes as pH rises above 6.5.[1] The dissolved organic carbon found in most surface soils has a strong affinity for Cu, but estimates of the percentage of soluble Cu that is organically complexed can vary widely.

Copper and Plants

Plant uptake of Cu appears to be directly related to the concentration of the free ion, $Cu^{2}+$, but may also be influenced by the total concentration in soil solution, including organic complexes.[5] As with most trace metals, it is not known whether Cu is passively absorbed or actively taken up across the root-cell membrane.[6] Rates of absorption are generally low, on the order of 1 nmole h^{-1} (g root dry weight)$^{-1}$.[7] The activity of free Cu^{2}+ required in nutrient solution for optimal plant growth is just 10^{-14}–10^{-16} M.[5] Copper absorption is generally halted by metabolic inhibitors and uncoupling agents which disrupt the normal transmembrane potential.[7]

Uptake of Cu is strongly affected by pH: increasing concentration of hydrogen ions decreases the absorption of Cu ions by plant roots.[8] Uptake is also affected by the presence of Ca, and to a lesser extent by Mg, both of which compete with Cu for binding sites at the root plasmalemma.[9] The effects of other trace metals on Cu uptake have frequently been seen as inhibitory (Zn), or stimulatory (Mn), but not necessarily under well-defined conditions.[2]

Copper as an Essential Element

Copper is an essential element for plant growth and is a component of many enzymes, including plastocyanin, and thus is an indispensable prosthetic group in Photosystem 2. Cu-containing proteins are also important in respiration (cytochrome c oxidase is the terminal oxidase of the mitochondrial electron chain), in detoxification of superoxide radicals (superoxide dismutase), and in lignification (polyphenol oxidase). Ascorbate oxidase, which contains eight Cu^{2+} ions, has been proposed as an indicator of plant Cu status, although its relevant biological function has yet to be determined.[10]

The critical concentration of Cu in shoot tissue for optimal growth does not vary greatly between plant species, ranging from 1 to 6 μg/g dry weight of young leaf tissue.[11] Most crops are recorded as having a requirement of 3–5 μg/g.[11] The average concentration of Cu in plant parts varies with age and with the level of Cu and N supply. Translocation of Cu to plant shoots increases with an increasing supply of N. In the xylem and phloem saps, Cu is probably complexed by amino acids.[2] Copper is usually described as having "variable" phloem mobility in plants,[2] as the retranslocation of Cu from older tissues is regulated by both Cu-supply and by N-status. Lack of a sufficiently long-lived radioisotope makes study of Cu transport and translocation problematic.

Plant Growth on Copper-Deficient Soils

Copper deficiency most often occurs on organic soils where excessive leaching has occurred, or on calcareous sands. In general, crops grown on mineral soils with Cu contents less than 4 to 6 μg/g, or on organic soils with less than 20 to 30 μg/g are the most likely to suffer Cu deficiency,[3] although this varies with specific soil type and the crop grown.

Symptoms of Cu-deficiency include chlorosis, necrosis, leaf distortion and terminal dieback, and are most evident in new leaf growth. Wilting can also occur, indicating structural weaknesses due to reduced lignification of the xylem elements. These symptoms are not entirely specific to Cu deficiency, and can be observed in plants under a variety of stresses. The most profound symptoms of Cu deficiency are those seen in the reproductive cycles of many sensitive species: delay of flowering, and/or reductions in seed and fruit yield as a consequence of sterile pollen or reduced floret numbers. Because these latter symptoms are not observed until maturity or harvest, Cu deficiency is often termed a "hidden hunger." Rice, citrus and cassava are sometimes referred to as indicator species that are more sensitive to Cu deficiency, but are still not reported to require more than 5–6 μg/g Cu to avoid Cu deficiency. Cereal rye and canola are crops more tolerant of Cu deficiency, requiring only 1–2 μg/g Cu.[11]

Copper deficiency in legumes depresses nodulation and N2 fixation, leading to N deficiencies. Unlike Mo and Co, there seems to be no specific Cu requirement for N2 fixation in nodules beyond that required for plant growth and the production of carbohydrates.[12]

Copper deficiency decreases polyphenol oxidase activity and thus lignification.[14] Susceptibility of Cu-deficient plants to pathogenic attack may be increased due to reduced lignification of xylem elements, or due to impaired lignification in response to pathogenic invasion (wounding response). Application of Cu to soil, at rates too low to directly affect the pathogen, has controlled powdery mildew in wheat.[15]

It has also been suggested that, where Cu in soils is more than sufficient, the accumulation of additional Cu in roots provides a fungistatic defense against pathogens.[16] Conversely, where Cu is deficient, roots are more vulnerable to pathogenic invasion.

There are genetic differences in the absorption of Cu by plant roots. Rye is able to take up significantly more Cu from soil than wheat, and is thus viewed as being more Cu-efficient. Tritcale, the wheat-rye hybrid, inherits the efficiency factor.[7] Copper efficiency could be a useful trait in breeding crops for regions where soils are commonly Cu-deficient.[17] However, the mechanism of the efficiency factor is not clear.

Plant Growth on High-Copper Soils

As some plant species have adapted to soils of low Cu status, others have evolved tolerance to Cu-toxic conditions. The 16th century author, Agricola, wrote of indicator plants that grow on naturally Cu-rich soils. Natural revegetation of mine spoils has been shown to reflect rapid genetic evolution of Cu tolerance by grasses and other plants.[14] There seem to be several possible mechanisms of Cu tolerance, although exclusion from shoots is a common feature. The exceptions are a few Cu-accumulator species which may contain in excess of 1000 μg/g Cu in shoot tissue.[12] In other Cu-tolerant species, root compartmentation or immobilization of Cu may be achieved by immobilization in cell walls, by complexation with intracellular proteins, or by removal of Cu to the vacuole.[14]

With nontolerant taxa, plant growth is likely to be depressed when Cu concentrations in the whole shoots exceed ~20 μg Cu g^{-1}. Symptoms of Cu toxicity include poorly developed and discolored root systems, reduced shoot vigor, and leaf chlorosis. Toxicity thresholds (e.g., for a 10% yield reduction) seem to vary widely, probably because of the low translocation of Cu from roots to shoots. Only when roots are overwhelmed by Cu rhizotoxicity does sufficient Cu reach shoots to affect growth and function. Other syndromes, such as Fe deficiency, can readily occur as secondary consequences of excess Cu.[14]

Exclusion of Cu from the shoot protects photosynthetic activity, which is highly sensitive to excess Cu. Photosynthetic electron transport is blocked by high levels of free Cu, at the oxidizing side of Photosystem 2, and at the reducing side of Photosystem 1. Excess Cu supply results in reduced lipid content and noticeable changes in the fatty-acid composition of tomato roots and primary leaves, indicating enhanced activity of enzymes which catalyze lipid peroxidation.[18]

Concentrated Cu sprays have historically been used to control foliar pathogens, especially in vineyard and orchard crops. These fungicidal sprays often included limestone to reduce their phytotoxicity, and to make them more rainfast. The accumulation of Cu in the soils under these crops has not regularly caused Cu toxicity, which indicates the remarkable ability of most plant roots to accumulate Cu, while regulating its flow to the shoots. Both Cu deficiency and toxicity may result in non-specific symptoms of plant stress. Assessment of the Cu status of a soil, or of crop plants, is most accurate when soil type, soil history, and soil and plant analyses for Cu are all considered.

References

1. McBride, M.B. Forms and distribution of copper in solid and solution phases of soil. In *Copper in Soils and Plants*; Loneragan, J.F., Robson, A.D., Graham, R.D., Eds.; Academic Press: Sydney, Australia, 1981; 25–45.
2. Loneragan, J.F. Distribution and movement of copper in plants. In *Copper in Soils and Plants*; Loneragan, J.F., Robson, A.D., Graham, R.D., Eds.; Academic Press: Sydney, Australia, 1981; 165–188.
3. Sanders, J.R.; McGrath, S.P. Experimental measurements and computer predictions of copper complex formation by soluble soil organic matter. Environ. Pollut. **1988**, *49,* 63–79.
4. Robertson, A.P.; Leckie, J.O. Acid/base copper binding, and Cu^{2+}/H^+ exchange properties of a soil humic acid, an experimental and modelling study. Environ. Sci. Technol. **1999**, *33*, 786–795.

5. Bell, P.F.; Chaney, R.L.; Angle, J.S. Free metal activity and total metal concentrations as indices of micronutrient availability to barley *(Hordeum vulgare* L. Cv Klages). Plant Soil **1991**, *130,* 51–62.
6. Strange, J.; Macnair, M.R. Evidence for a role for the cell membrane in copper tolerance of Mimulus guttatus fischer Ex DC. New Phytol. **1991**, *119,* 383–388.
7. Graham, R.D. Absorption of copper by plant roots. In *Copper in Soils and Plants*; Loneragan, J.F., Robson, A.D., Graham, R.D., Eds.; Academic Press: Sydney, Australia, 1981; 141–163.
8. Lexmond, Th.M.; van der Vorm, P.D.J. The effect of pH on copper toxicity to hydroponically grown maize. Neth. J. Agric. Sci. **1981**, *29,* 217–238.
9. Parker, D.R.; Pedler, J.F.; Thomason, D.T.; Li, H. Alleviation of copper rhizotoxicity by calcium and magnesium at defined free metal-ion activities. Soil Sci. Soc. Am. J. **1998**, *62,* 965–972.
10. Maksymiec, W. Effect of copper on cellular processes in higher plants. Photosynthetica **1997**, *34,* 321–342.
11. Reuter, D.J.; Robinson, J.B. *Plant Analysis: an Interpretation Manual*; CSIRO Press: Melbourne, Australia, 1997; 83–566.
12. Römheld, V.; Marschner, H. Function of micronutrients in plants. In *Micronutrients in Agriculture*, 2nd Ed.; Mortvedt, J.J., Cox, F.R., Shuman, L.M., Welch, R.M., Eds.; SSSA: Madison, WI, 1991; 297–328.
13. Jarvis, S.C. Copper concentrations in plants and their relationship to soil properties. In *Copper in Soils and Plants*; Loneragan, J.F., Robson, A.D., Graham, R.D., Eds.; Academic Press: Sydney, Australia, 1981; 265–285.
14. Woolhouse, H.W.; Walker, S. The physiological basis of copper toxicity and copper tolerance in higher plants. In *Copper in Soils and Plants;* Loneragan, J.F., Robson, A.D., Graham, R.D., Eds.; Academic Press: Sydney, Australia, 1981; 235–262.
15. Graham, R.D. Susceptibility to powdery mildew (*Erisiphe graminis)* of wheat plants deficient in copper. Plant Soil **1980**, *56*, 181–185.
16. Graham, R.D.; Webb, M.J. Micronutrients and disease resistance and tolerance in plants. In *Micronutrients in Agriculture;* Mortvedt, J.J., Cox, F.R., Shuman, L.M., Welch, R.M., Eds.; Soil Science Society of America: Madison, WI, 1991; 329–370.
17. Owuoche, J.O.; Briggs, K.G.; Taylor, G.J. The efficiency of copper use by 5A/5RL wheat-rye translocation lines and wheat (*Triticum aestivum* L.) cultivars. Plant Soil **1996**, *180,* 113–120.
18. Ouariti, O.; Boussama, N.; Zarrouk, M.; Cherif, A.; Ghorbal, M.H. Cadmium-and copper-induced changes in tomato membrane lipids. Phytochemistry **1997**, *45,* 1343–1350.

22

Globalization

Alexandru V. Roman

Introduction

Within the last two decades, the world has witnessed an extraordinary increase in the flow of resources, goods, services, knowledge, and ideas among countries. Technological developments and competition imperatives have led to higher levels of interconnectedness between nations and their economies. Today, it would be impossible to discuss major innovations, products, services, or companies only within the contexts of state boundaries. Globalization has also paved the way for the evolution of a global community—one that acts and functions in a coordinated manner in regard to common dilemmas. However, for all the benefits that can be directly or indirectly associated with economic growth and globalization, many argue that it imposes great burdens on the environment.

This entry discusses the challenges associated with understanding environmental impacts and tracing policy priorities within the framework and dynamics of globalization. The discussion revolves around four dominant themes: the globalization debate, the concepts of sustainable economic growth, global governance, and the expectations about the future developments within environmental management and globalization. The entry makes two holistic arguments: 1) the understanding of the environmental and governance impacts of globalization is evolving, but it is still seriously hindered by the conflict and competing discourses; and 2) an adequate response to the environmental consequences of current trends has to be built on collaborative efforts that are comprehensive in nature and do not identify the environmental impacts as mere externalities.

Understanding Globalization: The Debate, Perspectives, and the Complexity of the Environmental Challenges

The term globalization became an essential part of governance discourse in the 1990s. Academics have not fared well in their attempts to reach a single definition of globalization. The phenomenon is very complex, and its ramifications are still not fully understood and induce a great deal of debate.

How globalization is defined will depend on the assumptions that one makes about its nature and dynamics. The term is commonly used to describe the increased flow of knowledge, labor, resources, goods, services, and policy considerations, but also the standardization of values across societies, within a broader context of global awareness, and social transformational influences that lead to higher degrees of interconnectedness and interdependence. In essence, globalization is a description of the social, financial, economic, and political conditions that became pronounced in the second part of the 1980s and intensified after the end of the Cold War.

To a certain extent, the roots of globalization can be traced back to the 19th century,[1] while the underlining trade patterns possibly as far back as several thousand years. The actual concept entered the academic vocabulary in the 1960s but at the time it failed to attract any serious attention from the literature or the social context. With time, the idea of globalization started to symbolize the rise of consciousness of the world society in regard to normative developments.[2] As patterns of interconnectedness evolved, they led to manifold pressure but also opportunities to restructure or to regiment societies through a possible homogenization of beliefs and perspectives.[2]

Currently, globalization is primarily associated with the increased levels of financial and economic interdependencies. These dynamics are expected to remain an important part of the future economic landscape.[3] The Organisation for Economic Co-operation and Development identified three leading forces behind the phenomenon: 1) information and communication technologies; 2) liberalization of capital movements and deregulation; and 3) increased access to markets for trade and investments.[4] Others discern only two broad categories of motivators: on the one hand, the technological changes led to reduction in transport and communication costs, and on the other hand, policy decisions prodded regional and global integration initiatives.[5] While economic dimensions are important drivers behind globalization, it would be misleading to pick them as the only consequential aspects. It is probably best to think of globalization as a broad set of social processes or as an interpretative framework that aside from economic interdependencies also acknowledges associations within the world's political systems, beliefs, cultures, and individuals.

The concept of globalization and the consequences associated with it are believed to be doomed to controversy and inimical debate.[6,7] The dispute over whether globalization is "good" or "bad" can ignite anger, which can easily spill into violent street manifestations, as was the case in Seattle in 1999. Challenges and issues caused by integration trends will probably continue to define the future of global political and economic discourses. Historically, there always existed a significant conceptual gap between the advocates and opponents of globalization. The fissure in the perspectives of the two camps appears to be highly dependent on and correlated with the economic and political states of the world. At the same time, it has also been argued that the two opposing views are nothing more than reflections of other historical factious debates such as pro/anti-capitalism or pro/anti-American.[6]

The supporters believe that amidst economic growth, globalization also improves social equality and standards of living, helps the environment, and provides a platform for developing sensitivity for global issues. The globalization advocates construct their support on three main grounds. Foremost, it is argued that technological and knowledge spillovers that result from globalization lead to improvements in the socioeconomic status of individuals around the globe.[7] In addition, it is thought that globalization might provide the solutions to wicked problems that otherwise would not have been possible to envision or develop. Finally, based on the assumption of the rationality of the individual behavior within the market frameworks, it is expected that globalization will create the conditions for improved democratic outcomes within the design of domestic policies but also in international politics.

The critics of globalization condemn it on the basis of two overarching ideas. First, its opponents claim that the concept is an evolutionary form of a domination discourse promoted by the powerful, both within the national and the international arenas. This discourse demonstrates limited sensitivity to or acceptance of the competing views of marginalized groups, and that inexorably worsens the economic, environmental, and human conditions of the global community. Second, many point out that the economic rewards and the environmental burdens are not fairly distributed among the established economies and the less developed ones, with the former receiving the greater proportions of the benefits.[8] It is commonly suggested that globalization leads to harmful economic and social behaviors, such as "race to the bottom"—corporations seeking lax regulatory environments in order to achieve lower production costs at the expense of the environment. In the same vein, some are concerned that even developed countries might bid down their regulatory standards with the intent to attract investments.[9,10] Most importantly, it appears that the ecological consequences have been systematically understated within economic analyses and when discussing long-term strategic developments.[11]

Both camps, as well as those caught somewhere in between, are able to present well-constructed arguments. At the moment neither of the two extremes can claim an overwhelming advantage in the debate nor that the globalization dynamics have been overly responsive to any single perspective. Using the KOF index of globalization, which measures globalizations along political, social, and economic components, scholars have shown that the degree to which countries become interdependent and interconnected has been increasing constantly, with developing Asia experiencing the most rapid rates of globalization.[5] While social and political globalization patterns are similar to those of economic globalization, the former are less stable and follow a more volatile path than the latter.[5] As a result it is almost impossible to identify a community, agency, network, country, or society that is immune to the effects of globalization. One simply cannot avoid being part of the globality—the social condition.[12] Even the antiglobalizers are an essential part of globalization.[6,12]

Therefore, one could easily argue that the nature of globalization is far more complex than it is habitually considered within the contexts of partisan rhetoric. Adding environmental impacts to the picture complicates matters even further. First, as has already been pointed out, we have yet to reach an adequate understanding of globalization on its own. Attempting to account for its long-term environmental implications will most likely reduce any rational analysis to idealized guesstimations. Second, globalization is routinely associated with global climate change. The latter concept has generated cult-like reactions ever since its rise to agenda frontline status. The increased saliency of climate change is primarily due to awareness-building efforts, such as Al Gore's *An Inconvenient Truth* (2006) documentary. The belief that there is a direct causal relationship between globalization and global warming leaves little time for authentic and reasonable dialogue among pro- and antiglobalizers. Finally, the majority of current academic efforts in the area are fragmented, and there are not many credible and coherent interdisciplinary efforts to adequately understand the nature of globalization. Researchers frequently address only specific aspects of globalization. More often than not, these academic accounts are biased in some way and rarely allow for the possibility of differing interpretations.

Globalization, Sustainable Economic Growth, and Environmental Impacts

To globalize or not to globalize is not the question. What is of import is "how" to globalize.[13] Quintessentially, the matter is that of sustainable economic growth—maximizing economic development, but only within the framework imposed by environmental constraints.[14] While at times, globalization is too closely associated or even believed to be synonymous with international trade, thus underestimating other facets of the phenomenon, for the purpose of this entry, the economic hyperactivity of the last few decades warrants principal attention since it impacts the environment the most.

The globalization-driven changes have created real and perceived shifts in terms of the ecological, sociopolitical, and cultural systems. In many ways, the understanding of the impacts of global economic development and the inertia dynamics that come with it is lagging the development itself. From a broad perspective, the environmental challenges associated with globalization can be grouped within two categories: 1) direct and interdependent issues; and 2) transformational and wicked issues.

Direct and interdependent challenges are the ones that arise from the scale of human activities or from the incremental evolution of the existing conditions. For example, much of the growth in demand for trade will be fueled from the demand side of the equation rather than the supply side of it. Existing consumption habits will stimulate additional demand. Technological changes and competition will cause short life cycles for products and increased interdependency among producers and consumers. Once certain consumption dynamics are set in place, they become part of the larger culture and are eventually institutionalized. Population growth will lead to higher demand for food, and the global community will have to face the pollution and ecological trade-offs associated with meeting that demand.[15]

The transformational and wicked challenges represent the environmental impacts caused by the lack of a complete understanding of the nature of the ecological system but also as a direct consequence of the efforts to mitigate current environmental pressures. The ecological dilemma associated with the production of nuclear power is one challenge that can be categorized as wicked. Such a form of energy production is a highly needed solution for alleviating the pressures caused by high demand for energy. However, the events of Chernobyl in 1986 and Fukushima in 2011 serve as painful reminders of the vulnerability of this form of energy production. At the same time, once an industry or a belief is established within a policy system, it becomes very problematic to transform it or terminate it, even if a more appropriate alternative is available. Essentially, "older" responses to the environmental dilemmas, once institutionalized, could stymie the development of improved solutions when the "new" solutions endanger their own survival.

Other wicked challenges are those associated with the transformational impacts of the efforts to meet the global demand for resource and mainly alimentary demands.[15] According to some estimates, it would take 1.5 "planets earth" to sustain the current demand for resources.[16] The extra 50% in resource demand has to be covered through increased efficiency, genetic manipulations, and gains in productivity. The consequences of many of these transformations are yet to be fully understood, and on many occasions, they are either speeding up the rates of ecological damages (e.g., farming) or generating new environmental challenges that have yet to make it on the policy agenda. Solutions for some of the issues within the latter category are beyond our present-day technological or cognitive capacity (e.g., accurately tracing the influences of genetic changes in food).

The liberalization of the markets of the recent years has led to an increase in trade, which stimulated economic growth[5,17–19] and possibly led to convergence in terms of economic development and income levels.[5,20] Higher levels of trade will constitute a significant part of the future global developments. With the exception of 2009, for the last two decades, international trade experienced average annual growth rates above 6%, and at the same time, international travel and transport grew, on average, 14% a year.[20,21] Research by the Hamburg Institute of International Economics suggests that the values of trade among countries will continue to see larger annual growths until 2030—among East Asia and Pacific countries, 12.6%; South Asia, 10.9%; Latin America, 8.5%; and industrialized countries, 5.7%.[22] These economic patterns are not expected to change,[23] and they will continue to place unprecedented levels of pressure on the environment.[4] Here it is important to note that it has not yet been conclusively proven whether the degradation of the environment would have been considerably different under scenarios of zero or minimal international trade. Globalization might have simply increased the speed of proliferation of otherwise inevitable scale consequences of growth.[6]

The majority of the reports on environmental impacts of trade liberalization is not positive. A study that looked at the data from 63 countries for the period of 1960 to 1999 finds that increased trade has harmful effects on the environment; for every percent increase in trade openness, there is a 0.579% increase in carbon dioxide.[24] Global transportation is responsible for 14% of the total greenhouse emissions; of this, trucks account for 3.2%, ships 1.4%, and aviation 0.98%.[3] It is estimated that in the United States, the environmental damages from transportation using trucks are \$0.0023/mi,[25] while in Europe, these costs are between 0.0209 and 0.0746 EUR/km.[26] Damages from air transport are believed to be between 0.01 and 0.05 EUR per passenger/km.[27] Greenhouse gases emitted by air transports leave deeper pollution footprints since their emissions happen at high altitudes. Unfortunately, this will also be the type of emission that will grow the fastest among the group.[3]

In addition to the direct harms to the environment from transportation of goods, there are also significant concerns about oil spills. In the period of 2000 to 2010, there has been an average of 3.3 major oil spills (700 tons or more) per year.[28] In spite of constant growth in oil trade, the above numbers actually represent a significant improvement from the average of 7.8 spills/year in the 1990s, 9.3 spills/year in the 1980s, and 25.3 spills/year in the 1970s.[28]

It is important to note that many of the highest-polluting and most environmentally unfriendly industries, such as ocean fisheries and industries using coal or water, are currently overproducing as a result of the governmental subsidies that they receive in order to meet inflated market demands.[14,15] It is believed that the aggregate and continuous pressures from global economic activities will lead to the fact that by 2100, the global temperature will rise anywhere from 1.1°C to 6.3°C.[29] By 2050, the outcomes or higher temperatures could displace up to 200 million people (e.g., due to droughts, floods, rising sea levels).[3]

Higher levels of trade and economic growth as a result of globalization are not perceived by everyone as being first and foremost negative for environment. At the moment, the available empirical data are not sufficient to reach a decisive conclusion on whether economic development harms or benefits the environment,[30] as many empirical studies reach mixed results.[20,31,32] Some scholars suggest that free trade benefits the environment—more trade leads to higher individual incomes, which motivate the demand for a cleaner environment and lower levels of sulfur dioxide pollution concentrations.[6,14,33–35] It is expected that the world will grow cleaner as it grows richer.[6,33,35] The latter belief, which assumes the existence of the environmental Kuznets curve (EKC), is highly dependent on institutional arrangements.[34] Countries that are able to institutionally enforce environmental concerns will be able to devise the contexts necessary for economic growth and ecological needs to coexist.

The possibility of the existence of EKC was met by a lot of enthusiasm from the academic community when the idea was introduced in the early 1990s. However, the promises associated with the EKC research have not materialized. Recent studies have questioned the validity of the earlier empirical results, which were highly dependent on the specifications of the econometric models and on the quality of the data.[20,30] While one can still make a case for an EKC effect being present in regard to certain pollutants, there is strong evidence that there is no EKC relationship present in terms of CO_2 emissions.[20]

The large range of disagreements within the academic research is primarily due to the failures to emphasize the role of the institutional contexts. Institutions appear to be much more important in determining the degree of environmental effects of globalization than was previously thought. At least in the case of poor countries, it is doubtful that, at this point in time, they would have the strong institutional frameworks necessary for the growing-cleaner- while-growing-richer effect to become a reality. Corruption and unstable regulative and enforcement structures within these nations would probably preclude, at least in the short run, any significant environmental improvements expected from market liberalization or from growing richer.

Global Governance: Organizations, Policy Efforts, and Environmental Justice

It is often the case that when discussing globalization, one unavoidably encounters politically laden narratives that at times can reduce otherwise constructive debates to rhetoric- driven exchanges. In order to permit constructive progress within the global environmental management efforts, it is vital to acknowledge and understand the forces behind the dominant globalization metaphors. Only those initiatives that will be identified as legitimate and will rely on citizens' participation and input in the design process will be successful in the long run.

In a globalized world, almost all policy actors are global actors. Not so long ago, one could have argued that national governments, the International Monetary Fund, the World Bank, the United Nations, the World Trade Organization, and several of the larger corporations were the principal players one should consider in the global policy arena. This is not the case any longer. Information communication technologies have reduced transaction costs and barriers to entry and increased the possible impact and saliency of the actions by smaller policy actors. These developments extended the set of participants to thousands of nongovernmental organizations (NGOs) and networks. According to United Nations' estimates, there are more than 40,000 internationally active NGOs, while at the national levels, these numbers are in the millions. These institutions have become regular actors in defining, designing, and implementing environmental and global governance policies. The power of the NGOs cannot be easily dismissed, as many of them have levels of influence and public acceptance that can frequently challenge the validity of governmental actions.[36] A large number of actors in the global governance arena is welcomed in terms of offering a diversification of perspectives, but at the same time, this makes coordinated efforts less feasible because of the following: 1) NGOs hold competing views and interests that preclude effective action when a collaborative framework is not established; and 2) the presence of a multitude of actors with quasi-veto powers increases the instability and unpredictability of the policy systems. The latter enforces a short-term strategic outlook, as actors choose to emphasize short-range progress rather than facing the complexity of the long-run issues.

Desire for fast economic growth at times entices countries to choose economic benefits over environmental considerations.[37] Within this context, developing a unified global environmental policy framework becomes critical but problematic. The inability to generate an encompassing global policy approach is associated with the fragmentation of interests of the global actors and with the convoluted nature of pollution and its narratives. Greenhouse gas (carbon dioxide, nitrous oxide, and methane) emissions from increased trade and economic activity do not stay within borders, and climate change does not discriminate in its effects.[14] The location of their emission is irrelevant since their effects are global. Meanwhile, the positive environmental impacts from increased trade, such as changes in demand patterns due to higher incomes and technological spillovers, might be only localized in nature.

While economic theory can suggest many viable regulatory solutions that have been tested in practice, more often than not, these solutions are not employed.[14,15] The failure to do so is primarily due to the difficulties in parsing an acceptable understanding of the existing condition within the competing discourses but also due to the inability to break down the politics of financial interests and certain undesired consequential habits of the markets. The economic and social activities that cause the highest damages to the environment are still receiving the bulk of all governmental subsidies, and governments appear unwilling to promote strict environmental standards and technology.[15]

Some of the leading social concerns associated with globalization are in regard to the ethic of pollution distribution and cost. There is a significant unease in regard to the migration of polluting industries toward poor countries (pollution haven hypothesis) with weak institutional and social structures. Is it appropriate to concentrate polluting industries in certain areas? Should countries that do not significantly benefit from trade be recompensed for the increased pollution from transportation emissions? These are just a few of the questions that need to be answered.

As is the case with most globalization dimensions, the implications of such dynamics are not necessarily clear. Researchers suggest that given that high-polluting industries are generally capital intensive,[31,32] they would in fact concentrate in countries with strong financial and economic systems.[14,34] It is believed that the industrial composition of United States has shifted to relatively higher-polluting industries as a result of globalization.[38] Also, at least in the short run, the economic costs of stricter environmental standards are disproportionally borne by the low-income individuals[39]—that is, the poor will have to pay a higher relative price for a cleaner environment.

Future of the Global Environment: Research, Controversies, Current Successes, and Future Expectations

The complexity of the globalization dilemmas requires coordinated efforts on the part of the global community. The task of reaching a common denominator on policy actions is by no means a trivial one. Some suggest that the efforts to generate constructive solutions and progress in the globalization debate are frustrated by many concerns, with the inability to define, assess, and understand globalization acting as the primary stymieing barriers.[5,40] The fact that the study of globalization has not settled into a disciplinary field can be both beneficial and detrimental toward generating an adequate understanding of globalization. It is valuable because such a multifaceted matter would require interdisciplinary collaboration, but it is disadvantageous for at least two reasons. First, the research fragmentation reduces the probability of coherent and focused knowledge creation. Second, globalization research efforts might have a difficult time generating the prestige and respectability that comes with a recognized field of study. Nevertheless, the progress made within the academic community and in terms of political awareness of the environmental implications of globalization warrants hope that viable solutions to environmental dilemmas can be generated.

Currently, concerns regarding climate change are high on the political agendas of most countries, as it becomes evident that environmental degradation is a negative-sum game—if adequate actions are not undertaken, everyone will lose. Questions such as the availability of clean water, health impacts of changing environment, and loss of biodiversity, directly or peripherally linked to the globalization dynamics, are also becoming more salient. Even with an incomplete understanding of the environmental ramifications of economic growth and globalization, there is a silent confidence that there are possible ways to remedy the condition.[6,15] The necessary policies and solutions are available, realistic, and at a reasonable cost.[40] It is very important that corrective actions be taken immediately[41]- since the costs of the same policies might become prohibitive in 10 to 20 years.[3]

Policies that are restricted jurisdictionally to state or regional delineations will not be effective in addressing the nature of the environment impacts or in dealing with transportation fleets that operate under flags of convenience, thus not subject or responsive to operational regulations. In order to make a difference, environmental considerations need to be global in character and intertwined with all financial and economic policies domestically and internationally. Environmental policies should not be limited to the efforts of environmental agencies. Every contract, every school curriculum, every professional meeting, and every legislative act should have a subsection acknowledging or addressing environmental considerations. Research suggests that stricter regulations on emissions of nitrogen and sulfur oxides would lower such discharges by one- third as soon as 2030.[41] The trade-off of achieving this milestone in 2030 would be a global gross domestic product of 97% higher than today, rather than 99% higher. By employing an optimal pollution fee (marginal social cost equals marginal benefits), the greenhouse gas emissions can be reduced by 25% by 2050 and up to 45% by 2100.[42] Pollution-curtailing policies will be more effective if they tax the "bad" (e.g., pollution emissions, farming runoffs), which can be measured relatively more easily, rather than if they subsidize the "good," which is harder to identify and measure.[41]

In order to successfully manage the environmental impacts of globalization, changes at the cultural levels are unavoidable. The impacts that consumerist-type financial and economic systems have on the environment through increased pressures from the demand side are not minor. Efficiency gains in terms of energy consumptions obtained through the production of more environmentally friendly products and through ecologically conscious designs or eco-labeling are a start. Current products from lightbulbs to cars can be made to be more efficient to the point where the energy savings can reach 90% of the amount some of the existing products use.[15] Interestingly, it has been identified that "going green" is good for business—as such, industries prefer and lobby for stricter environmental regulations once trade is liberalized.[43] Foreign industries are generally more environmentally conscious than similar domestic firms operating in the same market.[44] Still, it is doubtful that such efficiency gains can provide sufficiently fast ecological relief without being supplemented by significant changes in social behavior and dominant environmental management discourse.

Based on what has been discussed thus far in the entry it becomes clear that there is no simple solution to managing the environmental changes caused by economic activity. Climate change is considered to be the greatest market failure in history.[3] An effective response would have to account for a multitude of dimensions:

1. Environmental ownership—environmental impacts need to be directly accounted for within the production costs and also clearly delineated within educational efforts. Environmental consequences, positive or negative, should be reinterpreted outside the traditional language of the "externality" narrative, and emissions accounting should become *a core consideration* of policy design and implementation.
2. "Environmentalization" of perspectives—redefining the environment as a local, domestic, and global priority in terms of both social norms and economic and security demands. The designs of environmental regulations should move past the traditional command- and-control policies and encourage more innovation.
3. Sustained collaboration—holistic, continuous, and coordinated global cooperative efforts.

Embracing these approaches does not necessarily mean an acceptance of lower levels of economic growth. By some accounts, stricter environmental regulations might not even hinder economic activity; on the contrary, such approaches lead to more innovative and productive economies.[45]

Conclusions

It is hard to foresee how globalization dynamics and clash of national values will affect democracy and the administrative approaches toward the understanding and the design of environmental policy, specifically in the long run.[46] It is certain that the answers cannot be identified within traditional and simplistic dichotomies like unique vs. universal or us vs. the other. Ironically, even antiglobalizers, in order to oppose globalization, need mechanisms, tools, and communication channels that are part of the globalization-driven transformative patterns. This is just one of many paradoxical complications found in the debate over the nature and interconnectedness of the environmental impacts of globalization. What has become clear is that in order for any policy to effectively address environmental challenges, it would have to incorporate both micro (individual) and macro (societal) imperatives, since managing the impacts of globalization calls for fundamental changes in interpretations and regulations.

Any viable and practical solution would have to attend to the complexity of globalization as well as the unpredictable character of the dynamics of ecological shifts. The 2009 and 2010 United Nations Conventions on Climate Change (UNCC) held in Copenhagen and Cancun, respectively, have not been as successful as one might have wished in legally binding governments in terms of serious actions. The Kyoto Protocol of 1997 that went into effect in 2005, despite its rather modest character, is still probably the main regulatory success at a global level. The scope and outcomes of the protocol would be significantly more consequential should the United States decide to ratify. The participation of United States is crucial since it will provide an incentive for nonparticipating developing countries to follow. Nevertheless, the dialogue during the 2009 and 2010 conventions suggests that participants understand the urgency of the situation and are ready to engage in global long-term environmental measures. Climate change and sustainable growth are top priorities for the United Nations.[47] The 2011 UNCC in Durban, South Africa, should succeed in having the parties move past nonbinding dialogue and into real action above what was achieved by the Kyoto Ptorocol.

Lester Brown[15] argues for the need for two major redefinitions of traditional policies and interpretations in order to manage the ecological challenges and avoid the "perfect storm" or the "ultimate recession." First, taxes on income should be decreased, while those on carbon emissions should be raised. Second, there is a need for redefining the idea of security as environmental security and to allocate resources toward addressing these challenges. According to Brown, the global community needs to reinterpret climate change, population growth, water shortages, poverty, rising food prices, and failing states rather than military aggression as the main threats to global well-being.[15] In an encompassing sense, global governance is in dire need of what some have called "clumsy solutions"— policies that creatively integrate opposing or competing understandings and solutions to address the common problem of climate change.[48]

Economists love to cite Milton Friedman stating that there is no such a thing as a free lunch. The deterioration of the global environment is the price that the global community is paying for its habits and economic expectations. Overlooking the environmental costs, either on purpose or due to the inability to quantify them, does not make those costs any less serious or real. Removing or reversing globalization trends will not solve the environmental challenges. The scale effects from economic and population growth would have resulted in ecological challenges regardless of international trade. The latter has intensified the criticality of the condition, but it is doubtful that stopping trade or the exchange of knowledge, services, or cultures would reduce the strain on the environment. Paradoxically, globalization is exactly what is needed to manage the environmental impacts of globalization.

References

1. Masson, P. Globalization: Facts and figures, IMF Policy Discussion Paper 2001, 01/04.
2. Meyer, J. Globalization. Int. J. Comp. Sociol. **2007**, *48* (4), 261–273.
3. Stern, N. *The Economics of Climate Change: The Stern Review*; Cambridge University Press: Cambridge, 2007.
4. OECD. *Measuring Globalisation: OECD Handbook on Economic Globalisation. Indicators 2005*; OECD Publishing, 2005.
5. Villaverde, J.; Maza, A. Globalisation, growth and convergence. World Econ. **2011**, *34* (6), 952–971.
6. Bhagwati, J. *In defense of globalization*; Oxford University Press: New York, 2007.
7. Kwong, J. Globalization's effects on the environment. Society **2005**, *42* (2), 21–28.
8. Stiglitz, J. *Globalization and its discontents*; Norton: New York, 2002.
9. Oates, W.; Schwab, R. Economic completion among jurisdictions: Efficiency enhancing or distortion inducing? J. Pub. Econ. **1998**, *35* (3), 333–354.
10. Levison, A. Environmental regulatory competition: A status report and some new evidence. Natl. Tax J. **2003**, *56* (1), 91–106.
11. Weitzman, M. On modeling and interpreting the economics of catastrophic climate change. Rev. Econ. Stat. **2009**, *91* (1), 1–19.
12. Steger, D.M. *Globalization: A Very Short Introduction*; New York: Oxford University Press, 2009.
13. Rodrick, D. Globalisation, social conflict and economic growth. World Econ. **1998**, *21* (2), 143–158.
14. Frankel, J. The environment and globalization, Working Paper 10090; NBER Working Paper Series, available at http://www.nber.org/papers/w10090 (accessed August 2011).
15. Brown, L.B. *World on the Edge: How to Prevent Environmental and Economic Collapse*; New York: W. W. Norton and Company, 2011.
16. Wackernagel, M.; Schulz, N.B.; Deumling, D.; Linares, A.C.; Jenkins, M.; Kapos, V.; Chad Monfreda, C.; Loh, J.; Norman Myers, N.; Norgaard, R.; Randers, J. *Tracking the Ecological Overshoot of the Human Economy*, Proceedings of the National Academy of Science, 99 (14), July 2002; 9266–9271.
17. Crafts, N. Globalisation and economic growth: A historical perspective. World Econ. **2004**, *27* (1), 45–58.
18. Frankel, J.; Romer, D. Does trade cause growth? Am. Econ. Rev. **1999**, *89* (3), 379–399.
19. Singh, T. Does international trade cause economic growth? A survey. World Econ. **2010**, *33* (11), 1517–1164.
20. Frankel, J; Rose, A. Is trade good or bad for the environment? Sorting out the causality. Rev. Econ. Stat. **2005**, *87* (1), 85–91.
21. WTO. *World Trade Report 2008—Trade in a Globalizing World*; WTO: Geneva, 2008.
22. Berenberg Bank and Hamburg Institute for International Economics. *Strategy 2030—Maritime Trade and Transport Logistics*. Berenger Bank: Hamburg, 2006.
23. WTO. *World Trade Report 2011—The WTO and Preferential Trade Agreements: From Co-existence to Coherence*; WTO: Geneva, 2011.

24. Magani, S. Trade liberalization and the environment: Carbon dioxide for 1960–1999. Econ. Bull. **2004**, *17* (1), 1–5.
25. Forkenbrock, D. Comparison of external costs of rail and truck freight transportation. Transp. Res. Part A **2001**, *35* (4), 321–337.
26. Bickel, P.; Schmid, S.; Friedrich, R. Environmental costs. In *Measuring the Marginal Social Cost of Transport Research in Transportation Economics*; Nash, C., Ed.; Elsevier: Oxford, 2005; Vol. 14, 185–210.
27. Dings, J.; Wit, R.; Leurs, B.; Davidson, M. *External Costs of Aviation*, Research Report 299 96 1006; Federal Ministry of the Environment, Nature Conservation and Nuclear Safety, 2003.
28. Huijer, K. Trends in oil spill from tanker ships. The International Tanker Owners Pollution Federation Limited, 2011, available at http://www.itopf.com/information-services/data-and-statistics/statistics/ (accessed August 2011).
29. Intergovernmental Panel on Climate Change. *Climate Change 2007: Synthesis Report. Contribution of Working Groups I, II and III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*; IPCC: Geneva, 2007.
30. William, H.T.; Levinson, A.; Wilson, D.M. Reexamining the empirical evidence for an environmental Kuznets curve. Rev. Econ. Stat. **2002**, *84* (3): 541–551.
31. Cole, M.; Elliott, R. FDI and the capital intensity of "dirty" sectors: A missing piece of the pollution haven puzzle. Rev. Dev. Econ. **2005**, *9* (4), 530–548.
32. Shen, J. Trade liberalization and environmental degradation in China. Appl. Econ. **2008**, *40* (8), 997–1004.
33. Grossman, G; Kruger, A. Economic growth and the environment. Q. J. Econ. **1995**, *110* (2), 353–77.
34. Antweiler, W.; Copeland, B.; Taylor, S. Is free trade good for the environment? Am. Econ. Rev. **2000**, *91* (4), 877–908.
35. Bhagwati, J. *Free Trade Today;* Princeton University Press: Princeton, 2002.
36. Koppell, J. *World Rule*: *Accountability, Legitimacy and the Design of Global Governance*; The University of Chicago Press: Chicago, 2010.
37. Thai, K.; Rahm, D.; Coggburn. J. *Handbook of Globalization and Environment*; Taylor and Francis: Boca Raton, 2007.
38. Ederington, J.; Levinson, A; Miner, J. Footloose and pollution-free. The Review of Economics and Statistics **2005**, *87* (1), 92–99.
39. Hasset, K.; Mathur, A.; Metcalf, G. The incidence of a US carbon tax: A lifetime and regional analysis. Energy J. **2009**, *30* (2), 155–177.
40. Scholte, J. A. Defining globalization. World Econ. **2008**, *31* (11), 1299–1313.
41. OECD. *Globalisation in OECD,* OECD Environmental Outlook to 2030; OECD Publishing, 2008.
42. Nordhaus, W. *A Question of Balance: Weighing the Options on Global Warming Polices;*Yale University Press: New Haven, 2008.
43. McAusland, C. Trade, politics, and the environment: Tailpipe vs. smokestack. J. Environ. Econ. Manage. **2008**, *55* (1), 52–71.
44. Eskeland, G.; Harrison, A. Moving to greener pastures? Multinationals and the pollution haven hypothesis. J. Dev. Econ. **2003**, *70* (1), 1–23.
45. Porter, M; van der Linde, C. Toward a new conception of the environment-competitiveness relationship. J. Econ. Perspect. **1995**, *9* (4), 97–118.
46. Box, R.; Marshall, G.; Reed, B.; Reed, M. New public management and substantive democracy. Pub. Admin. Rev. **2001**, *61* (5): 608–619.
47. United Nations. UN chief sees sustainable development as top priority in his second term, 2011, available at http://www.un.org/apps/news/story.asp?NewsID=39288 (accessed September 2011).
48. Verweij, M.; Douglas, M.; Ellis, R.; Engel, C.; Hendriks, F.; Lohmann, S.; Ney, S.; Rayner, S.; Thompson, M. Clumsy solutions for a complex world: The case of climate change. Pub. Admin. **2006**, *84* (4), 817–843.

23

Heavy Metals

Mike J. McLaughlin

Introduction

It is generally accepted that metals having a specific gravity (weight per unit volume) greater than 5 Mgm^{-3} are termed heavy metals. In soils, these elements include cadmium (Cd), cobalt (Co), chromium (Cr), copper (Cu), iron (Fe), mercury (Hg), manganese (Mn), molybdenum (Mo), nickel (Ni), lead (Pb) and zinc (Zn)[1] (http://www.soils. org/sssagloss; accessed February 2001). The term “heavy metals” is often used synonymously with the term “trace elements,” but this is incorrect as trace elements are generally defined as those elements normally occurring in soil at concentrations less than 100 mg kg^{-1},[2] which precludes several heavy metals, e.g., Cr, Fe, and Mn. Arsenic (As) is often included in the group “heavy metals,” but is more correctly classified as a metalloid.

General Chemistry

Some heavy metals are essential for either plant or animal survival on land (Co, Cr, Cu, Fe, Mn, Mo, Ni, and Zn) while others are nonessential and toxic at low concentrations (Cd, Hg, and Pb). All the heavy metals, except Pb, are transition elements, belonging to the d-block in the periodic table. Many of these elements differ from the alkaline earth metals (e.g., Na, Ca, and Mg) in that they can exist in several valence states in soil (Table 1). In particular, Cr, Fe, and Mn are markedly affected by soil redox potential and undergo both oxidation and reduction depending on soil conditions. This has important implications for the availability and toxicity of many heavy metals. As both Fe and Mn are major structural metals in soil minerals, reduction of the Fe^{3+} and Mn^{4+} ions may lead to a change in soil mineral surfaces important for retention of many elements, including other metals. For Cr, oxidation converts the nontoxic Cr^{3+} ion to the toxic and carcinogenic Cr^{6+} ion. This reaction has even more significance in soils as a strongly sorbed or precipitated cation (Cr^{3+}) is converted into a poorly sorbed or soluble anion (CrO_4^{2-}). However, even in aerobic soils, Cr^{3+} is the thermodynamically stable state, so added Cr^{6+} ion is rapidly converted to Cr^{3+} in most soils.

Solubility and availability/toxicity to organisms of heavy metal cations (Cd^{2+}, Cr^{3+}, Fe^{n+}, Pb^{2+}, Mn^{n+}, Hg^{2+}, Ni^{2+}, and Zn^{2+}) decrease as soil pH increases. This is due to the increase in negative charge on variable charge surfaces in soil[4] and the propensity for these metals to precipitate as sparingly soluble compounds (phosphates, carbonates, and hydroxides) as soil pH increases.[5] On the other hand, solubility and availability/toxicity to organisms of anionic heavy metals (CrO_4^{2-}, MoO_4^{2-}) may increase as soil pH increases, again due to increases in surface negative charge on soil particles affecting sorption.

TABLE 1 Physician and Chemical Properties of the Heavy Metals

										Dominant Species in Soil Solution[b]	
Element	Symbol	Atomic No.	Molecular Weight	Valence	Natural Isotopes	Density (MgnT–3)	Melting Point (°C)	First Ionization Potential (eV)	Dominant Species in Soil[a]	pH 3.5–6.0	pH 6.0–8.5
Cadmium	Cd	48	112.41	2	8	8.65	321	8.96	Cd^{2+}	Cd^{2+}, $CdCl^{+}$, $CdSO_4^0$	Cd^{2+}, $CdCl^{+}$, $CdSO_4^0$
Chromium	Cr	24	52.01	2, 3, 6	4	7.19	1875	6.76	Cr^{3+}, CrO_4^{2-}	Cr^{3+}, $CrOH^{2+}$	$Cr(OH)_{4-}$
Cobalt	Co	27	58.94	2,3	1	8.90	1493	7.86	Co^{2+}	—	—
Copper	Cu	29	63.54	1, 2	2	8.94	1083	7.73	Cu^{2+}	Cu^{2+}, Cu-org.	Cu-hydroxy species. $CuCO_3^0$, Cu-org.
Iron	Fe	26	55.85	2,3	4	7.87	1536	7.87	Fe^{2+}, Fe^{3+}	Fe-hydroxyspecies, Fe-org.	Fe-hydroxy species. Fe-org.
Lead	Pb	82	207.19	2,4	4	11.35	327	7.42	Pb^{2+}	Pb^{2+}, $PbSO_4^0$, Pb-org.	Pb-hydroxy and carbonate species, Pb-org.
Manganese	Mn	25	54.94	2, 3,4, 7	1	7.44	1244	7.44	Mn^{2+}, Mn^{4+}	Mn^{2+}, $MnSO_4^0$, Mn-org.	Mn^{2+}, $MnSO_4^0$, $MnCO_3^0$
Mercury	Hg	80	200.61	1, 2	7	13.54	–39	10.44	Hg^{2+}, $(CH_3)_2Hg$	—	—
Molybdenum	Mo	42	95.94	6	7	10.22	2610	7.10	MoO_4^{2-}	—	—
Nickel	Ni	28	58.71	2,3	5	8.91	1453	7.64	Ni^{2+}	Ni^{2+}, $NiSO_4^0$, Ni-org.	Ni^{2+}, $NiHCO_3^{+}$, $NiCO_3$
Zinc	Zn	30	65.37	2	5	7.14	420	9.39	Zn^{2+}	Zn^{2+}, $ZnSO_4^0$, Zn-org.	Zn^{2+}, Zn-hydroxy and carbonate species, Zn-org.

[a] From Logan.[2]

[b] From Ritchie and Sposito.[3]

Under reducing conditions in soil, many of the metals form insoluble metal sulfides, therefore reducing availability to plants and animals.

Heavy metals are also subject to complexation reactions in soil with both inorganic and organic ligands, which may markedly increase mobility in soil. This may be used to good effect by plants in scavenging essential metals from soil, e.g., phytochelatins (see below). However, for nonessential metals, complexation and increased mobility also may lead to increased environmental risks through leaching and plant uptake, e.g., chloride-induced uptake of Cd by food crops in saline soils.[6]

Abundance in Rocks and Soils

Heavy metals become incorporated into primary minerals in igneous rocks through isomorphous substitution. In sedimentary rocks, heavy metals are incorporated as constituents of minerals, or are removed from the water column and trapped in sediments by adsorption or precipitation processes.[7] Apart from Cr, Fe, and Mn, heavy metals are generally present at trace concentrations (<100 mg kg^{-1}) in most soils, with the exception of soils developed over mineralized parent materials or soils developed through biological enrichment.[8] Background concentrations in soil usually reflect the composition of the parent rock material. Background concentrations of heavy metals are usually determined where no known anthropogenic inputs have occurred, but this assessment is often problematic due to the global spread of anthropogenic emissions through atmospheric transport processes.[9] The typical range in abundances of heavy metals in unpolluted soils is shown in Figure 1.

Concentrations of heavy metals in soil can be significantly increased through human activity. A number of primary and secondary sources have been identified as contributing to enhanced concentrations of heavy metals in soil[12] (Table 2). Many countries have introduced legislation to minimize the amount of metals accumulating in soil, and set ceiling concentrations above which further metal additions must stop.

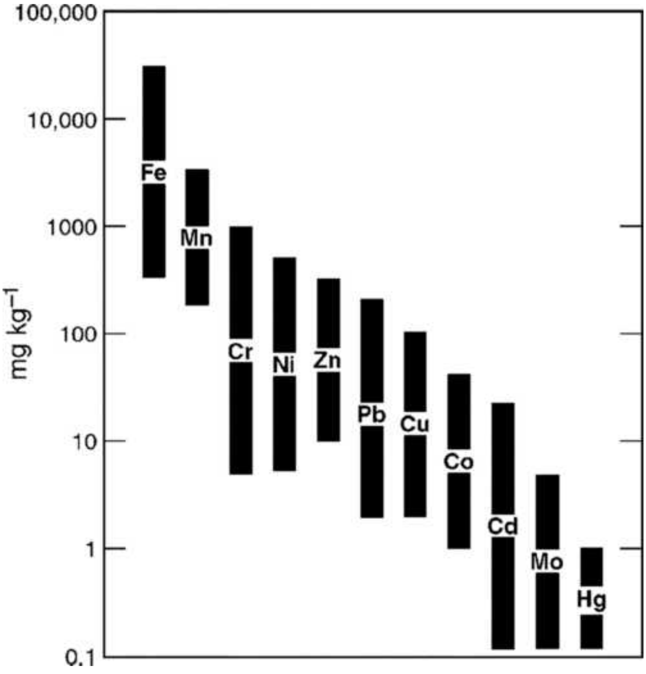


FIGURE 1 Range of heavy metal concentrations of soil. Bars represent commonly found values.

Source: Logan,[2] Alloway,[7] Barry and Rayment,[8] Lantzy and Mackenzie,[9] Kabata-Pendias and Pendias,[10] Swaine,[11] and Ferguson.[12]

TABLE 2 Sources of Heavy Metal Contamination in Soils

Source	Main Heavy Metals
Primary sources	
Fertilizers	Cd, Cu, Mo, Pb, Zn
Irrigation water	Cd, Fe
Manures and composts	Cd, Cr, Cu, Fe, Hg, Mn, Mo,Ni, Pb, Zn
Pesticides	Cu, Hg, Pb, Zn
Sewage biosolids (sludges)	Cd, Cr, Cu, Fe, Hg, Mn, Mo, Ni, Pb, Zn
Soil amendments (lime, gypsum, etc.)	Cu, Mn, Pb, Zn
Secondary sources	
Automobile aerosols	Pb
Coal combustion	Pb
Mine waste and effluents	Cd, Cu, Fe, Hg, Mn, Ni, Pb, Zn
Nonferrous smelter waste	Cd, Cu, Hg, Mn, Ni, Pb, Zn
Paint dispersal	Cd, Pb
Tire wear	Cd, Zn
Waste combustion	Cd, Pb

Source: Barry and Rayment.[8]

Biological Effects

Co, Cr, Cu, Fe, Mn, Mo, Ni, and Zn are all essential for either healthy plant or animal functioning in soil and all these metals, except Ni, are used in fertilizers or animal stock supplements to ensure efficient agricultural production on soils deficient in essential metals. Due to the low solubility of essential heavy metals in neutral and alkaline soils, many plants have developed strategies to mobilize solid phase forms to facilitate uptake by roots (e.g., Fe, Mn, and Zn). Nonproteinogenic amino acids, or phytochelatins, secreted by actively growing roots are important in acquisition of Cu, Fe, Zn, and possibly Mn from deficient soils.[13] At high concentrations in soil, all the essential elements may pose risks to microorganisms, plants, animals, or humans (Figure 2). Nonessential heavy metals in soil (Cd, Hg, and Pb) have no beneficial effects at low concentrations (Figure 2) and may be toxic at even trace concentrations (e.g., Cd, Hg).

Critical exposure pathways for expression of heavy metal toxicity in soil have been examined as part of regulation governing re-use of biosolids (sewage sludge) on soil,[14] but act as a general risk

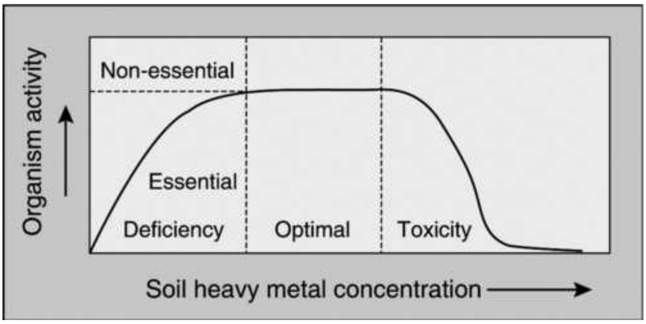


FIGURE 2 Typical concentration–response relationship for essential and nonessential heavy metals in soil.

TABLE 3 Critical Risk Pathway Assessment for Heavy Metal Pollution of Soils

Metal	Dominant Risk Pathway	Secondary Risk Pathway	Most Important Predictor Required
Cd	Food chain transfer	Phyto- and ecotoxicity	Soil–plant uptake
Co	Food chain transfer	Phyto- and ecotoxicity	Soil–plant uptake
Cr	Phyto- and ecotoxicity	Metal leaching	Toxic threshold definition, speciation
Cu	Phyto- and ecotoxicity	Soil ingestion by animals/humans	Toxic threshold definition
Fe	Phyto- and ecotoxicity[a]	None	Toxic threshold definition
Hg	Soil ingestion by animals/humans	Metal leaching	Toxic threshold definition, speciation
Mn	Phyto- and ecotoxicity	Soil ingestion by animals/humans	Toxic threshold definition
Ni	Phyto- and ecotoxicity	Soil ingestion by animals/humans	Toxic threshold definition
Pb	Soil ingestion by animals/humans	Phyto- and ecotoxicity	Oral bioavailability assessment
Zn	Phyto- and ecotoxicity	Food chain transfer	Soil-plant uptake

[a] Only in acid soils under reducing condition.
Source: McLaughlin.[15]

analysis template for all types of pollution of soil by heavy metals.[15] For some elements, e.g., Cd and Co, food chain transfer is the main risk pathway as these elements are easily accumulated by plants in edible tissues. For other elements, sorption in soil is strong, bioaccumulation by microorganisms low, and plant uptake and translocation so low that the dominant risk pathway is (for higher animals and humans) through direct ingestion of soil, e.g., Hg and Pb. For the other heavy metals, behavior in soil and bioaccumulation characteristics result in toxicity to plants and microorganisms (phyto- and ecotoxicity) being the dominant risk pathway at high concentrations, e.g., Cr, Cu, Mn, Ni, and Zn. Thus, prediction of risks from heavy metal pollution of soils through soil testing requires a different emphasis depending on the metal considered (Table 3).

References

1. *Internet Glossary of Soil Science Terms*; Soil Science Society of America: Madison, WI.
2. Logan, T.J. Soils and environmental quality. In *Handbook of Soil Science*; Sumner, M.E., Ed.; CRC Press: Boca Raton, FL, 2000; G155–G163.
3. Ritchie, G.S.P.; Sposito, G. Speciation in soils. In Chemical *Speciation in the Environment*; Ure, A.M.E., Davidson, C.M., Eds.; Blackie Academic and Professional: London, 1995; 201–233.
4. Barrow, N.J. *Reactions with Variable Charge Soils;* Marti- nus Nijhoff: Dordrecht, The Netherlands, 1997; 191 pp.
5. Lindsay, W.L. *Chemical Equilibria in Soils;* Wiley: New York, 1979; 449 pp.
6. McLaughlin, M.J.; Tiller, K.G.; Smart, M.K. Speciation of cadmium in soil solution of saline/sodic soils and relationship with cadmium concentrations in potato tubers. Aust. J. Soil Res. **1997,** *35,* 1–17.
7. Alloway, B.J. The origins of heavy metals in soils. In *Heavy Metals in Soils,* 2nd Ed.; Alloway, B.J., Ed.; Blackie Academic and Professional: London, 1995; 38–57.
8. Barry, G.; Rayment, G.E. Heavy metals and nutrients in soils and sediments of Raine island, great barrier reef. Land Contam. Reclam. **1997,** *5* (4), 281–285.
9. Lantzy, R.J.; Mackenzie, F.T. Atmospheric tracemetals: global cycles and assessment of man's impact. Geochim. Cosmochim. Acta **1979,** *43,* 511–525.
10. Kabata-Pendias, A.; Pendias, H. *Trace Elements in Soils and Plants,* 2nd Ed.; CRC Press: Boca Raton, FL, 1992; 365 pp.
11. Swaine, D.J. *The Trace Element Content of Soils;* Commonwealth Bureau of Soil Science Technical Communication No. 48; Commonwealth Agricultural Bureaux: Farnham Royal, U.K., 1955; 157 pp.

12. Ferguson, J.E. The Heavy Elements: Chemistry, Environmental Impact and Health Effects; Pergamon Press: Oxford, 1990; 614 pp.
13. Römheld, V. Different strategies for iron acquisition in plants. Physiol. Plant. **1987,** *70,* 231–234.
14. USEPA—United States environment protection agency. *Part 503—Standards for the Use and Disposal of Sewage Sludge;* Federal Register 58, 1993; 9387–9404.
15. McLaughlin, M.J.; Zarcinas, B.A.; Stevens, D.P.; Cook, N. Soil testing for heavy metals. Commun. Soil Sci. Plant Anal. **2000**, *31* (11–14), 1661–1700.

24

Inorganic Carbon: Composition and Formation

Larry P. Wilding and
H. Curtis Monger

Introduction

It has become increasingly common for "soil inorganic carbon" to mean soil carbonate mineral carbon, mainly $CaCO_3$. In the strict sense, however, inorganic carbon not only encompasses carbon in carbonate minerals, but also carbon in the carbonic acid system.[1] The carbonic acid system includes gaseous carbon dioxide ($CO_{2(g)}$), aqueous carbon dioxide ($CO_{2(aq)}$), carbonic acid ($H_2CO_{3(aq)}$), bicarbonate ion ($HCO_{3(aq)}{}^-$), and carbonate ion ($CO_{3(aq)}{}^{2-}$).

Composition

In the soil solution, as with other solutions, the interaction of these species can be represented by the following reaction:[1]

$$CO_{2(g)} + H_2O_{(1)} = H_2CO_{3(aq)} = HCO_{3(aq)^-} + H_{(aq)^+}$$
$$= CO_{3(aq)^{2-}} + 2H_{(aq)^+} \tag{1}$$

Cations, such as Ca^{2+}, Mg^{2+}, Fe^{2+}, Mn^{2+}, and Na^+, precipitate with $HCO_3{}^-$ (which is the dominant anion between pH 6.5 and 10.5) and $CO_3{}^{2-}$ (which is the dominant anion above pH 10.5) to form a variety of carbonate minerals. The reaction of $Ca_{(aq)}{}^{2+}$ with $HCO_{3(aq)}{}^-$ to form calcite is illustrated below:

$$Ca_{(aq)^{2+}} + 2HCO_{3(aq)^-} = CaCO_{3(s)} + CO_{2(g)} + H_2O_{(1)} \tag{2}$$

There are about 60 carbonate minerals, which in addition to calcite, include aragonite ($CaCO_3$), dolomite [$CaMg(CO_3)_2$], siderite ($FeCO_3$), magnesite ($MgCO_3$), rhodocrosite ($MnCO_3$), cerussite ($PbCO_3$), and malachite [$CuCO_3Cu(OH)_2$]. In soil, the overwhelmingly abundant carbonate mineral is calcite.[2] In unique soil environments, however, other carbonate minerals have been found, such as pedogenic siderite and dolomite.[3]

Individual carbonate crystals of pedogenic origin are generally too small to be seen with the unaided eye. Yet when concentrated together, their combined presence takes on a white color with a variety of macroscopic forms. These forms include carbonate filaments (also called my- celia, pseudomycelia, and threads), films, coatings, soft spheroidal segregations (white eyes), nodules, cylindroids, concretions, glaebules, and veins. Soil fabric which is impregnated with carbonate to the point that it occurs as an essentially continuous medium has been termed "k-fabric."[4] Stages of morphogenetic carbonate accumulation, in which progressively greater amounts of carbonate occur in progressively older soils, are important chronologic indicators. The calcic and petrocalcic horizons are diagnostic horizons in soil taxonomy.[5] Calcic horizons generally contain greater than 15% carbonate by weight. Petrocalcic horizons are indurated forms of calcic horizons. Examples of these horizons are shown in Figures 1A and B. Examples of carbonate crystals as viewed with optical microscopy and scanning electron microscopy are shown in Figures 1C and D.

Dissolution of carbonates in soil systems can be represented by the following reaction (Eq. 3). In humid regions, soluble products of this weathering reaction flux through the vadose zone into groundwater, or precipitate as pedogenic carbonates deep in the soil or geologic system. In arid

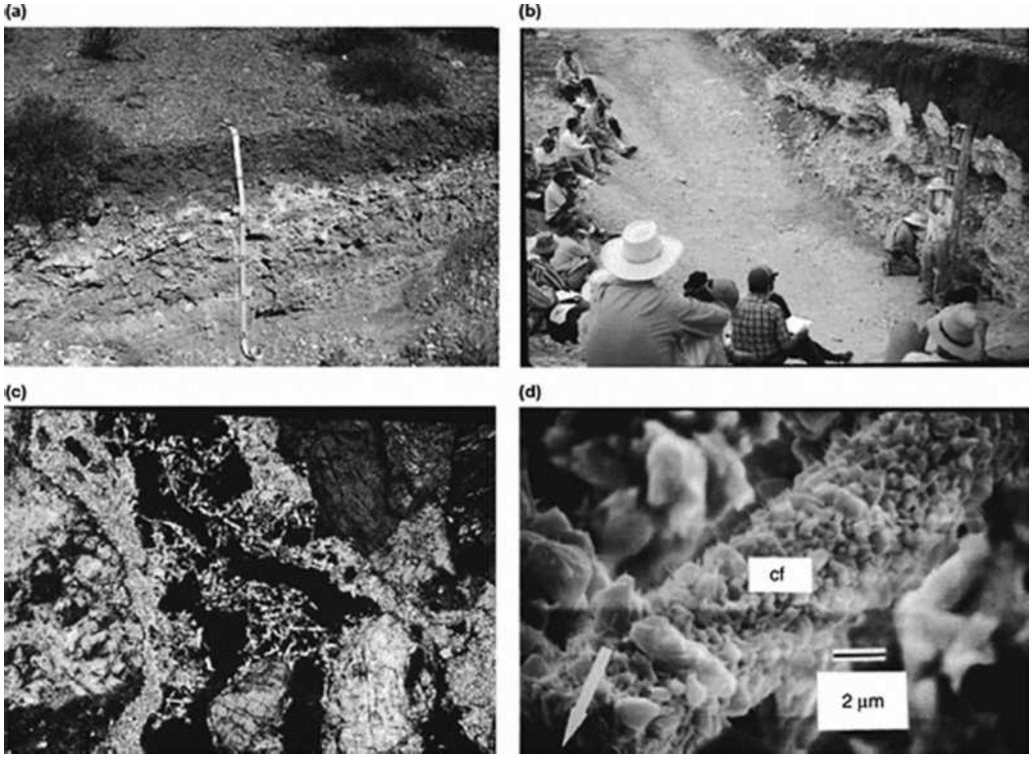


FIGURE 1 Examples of inorganic carbon as it exists in the field and under the microscope. The white horizon is a calcic horizon in (a), and petrocalcic horizon in (b). The small golden crystals in (c) are calcite crystals coating sand grains; the black region is a pore space as it appears in cross-polarized light. A calcified fungal filament (cf) viewed with scanning electron microscopy is shown in (d).

regions, soluble products precipitate at relatively shallow depths as a result of sparse rainfall and insufficient leaching.

$$HCO_{3(aq)^-} + H_{(aq)^+} + CaCO_{3(s)} = Ca(HCO_3)_{2(aq)} \tag{3}$$

Formation

Being located in an arid, semiarid, or subhumid climate is the primary factor that controls carbonate formation. In many areas, the boundary between carbonate-accumulating soil and noncarbonate-accumulating soil is about 500 mm (20 in.) mean annual rainfall.[6] This relationship is confounded, however, by the effects of soil temperature, soil drainage, nature and properties of the parent material (e.g., soil texture, carbonate content, carbonate mineralogy, and porosity), soil drainage, landform position, geomorphic stability, and effectiveness of precipitation (rainfall intensity and duration). Hence, there are many examples in humid and subhumid environments where soil carbonate persists in the soil system at depths inconsistent with regional models. In humid regions, for example, inorganic carbon persists as calcite or dolomite detritus in soils derived from certain parent materials (e.g., calcareous loess, till, outwash, alluvial deposits, sedimentary and metamorphic rocks). In seasonally wet soils, carbonate can accumulate in upper subsoils from capillary rise of bicarbonates via evaporative pumping from shallow groundwater.[7] In addition, carbonate minerals can occur in wetland soils which commonly contain soluble carbonates, bicarbonates or carbonic acid depending on the pH of the local environment

Pedogenic vs. Geogenie Carbonate

Many soils develop in calcareous parent materials. For these soils, it has been a challenge quantifying carbonate that formed in the soil profile vs. carbonate mechanically inherited from parent material. Carbonate formed in the soil profile has been termed "secondary," "authigenic," or "pedogenic."[4,8] On the other hand, carbonate mechanically inherited from parent material has been termed "primary," "geogenic," or "lithogenic."[2,9] Criteria for distinguishing pedogenic from geogenic carbonates involve the scrutiny of both field and laboratory evidence. Field evidence, for example, includes differences in the presence of marine fossils, carbonate morphology (such as nodules, pendants, and laminar caps which indicate pedogenic), and distribution patterns with depth, where, for example, a carbonate horizon of pedogenic origin is overlain and underlain by noncalcareous soil. Laboratory evidence includes comparing mineralogy, particle size, microfabric, and $^{13}C/^{12}C$ ratios of carbonate with unknown origin to those of carbonate with known geogenie origin.[10–12]

Models of Carbonate Formation

There are several processes that cause carbonates to form in soil. Excluding geologic processes, such as lacustrine and deep groundwater cementation that preserves the original sedimentary structure, the formation pedogenic carbonate can broadly be placed into four models—per descensum, per ascensum, in situ, and biogenic models.

Per Descensum Model

The per descensum model accounts for carbonate formation resulting from downward moving meteoric water and can be subdivided into three types. First is the dissolution of pre-existing carbonates in the upper profile, their vertical translocation, and their precipitation in the subsoil. This model was invoked to explain why progressively shallower carbonates occur in progressively drier climates.[13] Later, this per descensum model was used as the basis for calculating the number of wetting-fronts required to leach carbonates to a particular depth.[14] In both cases, it was assumed that carbonate was uniformly distributed in parent material at the beginning of pedogenesis.

Second is the case in which pedogenic carbonate forms in soils with noncalcareous parent materials. Unlike the model described before, noncalcareous parent material does not have carbonate uniformly distributed throughout the profile at the beginning of pedogenesis. In southern New Mexico, for example, prominent calcic and petrocalcic horizons occur in soils with rhyolite alluvium as parent material. This alluvium would yield low amounts of calcium if the rhyolite particles were thoroughly decomposed, which they were not.[15] Therefore, atmospheric additions, another per descensum model, was judged to be the source of carbonates.[15] Initially calcareous dust was measured and considered to be the source of carbonate. Later it was realized that Ca^{2+} in rain was an additional, and even larger source of Ca^{2+} for reaction with soil HCO_3^- to form carbonates.[15] Building on these per descensum concepts, compartmental models have been constructed that compute the depth, amount, and distribution of pedogenic carbonate as a function of climate and time.[16,17]

Third, in addition to vertical illuviation within a soil profile, lateral, downslope migration of the soil solution containing soluble products of carbonate is another per descensum model. In this case, carbonate is thought to precipitate after carbonate-charged waters migrate from upslope positions to lower landscape positions.[18]

Per Ascensum Model

The per ascensum model accounts for carbonate formation resulting from bottom–up movement. A primary example is the capillary rise of Ca^{2+} and bicarbonate from shallow water tables by evaporative pumping, which leads to the precipitation of carbonates in the upper subsoil.[7] Moreover, chemical studies have shown that in some environments plants promote carbonate formation by transporting Ca^{2+} upward to the land surface from subsoil, rock, and groundwater.[19]

In Situ Model

Third, in the in situ model, pedogenic carbonate is the result of in-place dissolution and reprecipitation of bedrock composed of marine carbonate.[20] Limestone, for instance, is progressively transformed into pedogenic carbonate as a result of short-range carbonate dissolution and reprecipitation proximal to the depth of the upper contact with limestone. This is a rather unique method to form pedogenic carbonates where the total carbonate content of the zone of enriched pedogenic products is less than the carbonate content of the limestone originally. These pedogenic zones have a much higher macro- and micro-porosity than the limestone.

In addition to marine carbonates, the in situ model also includes carbonate formation resulting from in-place chemical weathering of Ca-bearing igneous rock. Upon release into the soil solution by weathering, Ca^{2+} precipitates with bicarbonate formed from the reaction of water with CO_2 generated by root and microbial respiration. In many cases, however, igneous parent material has been considered as an insufficient source of Ca^{2+} and hence external sources, such as atmospheric additions of Ca^{2+}, have been sought.[6]

Biogenic Model

Fourth, some plants, microorganisms, and termites produce calcium carbonate. Evidence that various plants play a direct role in carbonate formation comes from the presence of euhedral calcite crystals on plant roots.[21] Moreover, several references in the Russian literature note carbonate formation by plant tissue and the downward translocation of these carbonates with wetting fronts.[22] Evidence that some microorganisms precipitate carbonates is based on observations of calcified bacteria and fungal hyphae with electron microscopy and in vitro laboratory experiments.[23,24] Evidence that termites precipitate carbonate in certain environments is based on the studies of termite mounds in Africa and southeast Asia.[25] Such mounds can be calcareous even though surrounding soils are noncalcareous, making the mounds attractive to native farmers who spread them over their agricultural fields.[25]

Conclusions

The formation of pedogenic carbonate may be dominated by one of the models listed before or may involve several of the models working together in different magnitudes. Understanding pedogenic carbonate formation has been extremely useful for understanding relative ages of geomorphic surfaces and landscape evolution.[15] A knowledge of pedogenic carbonate formation has also been useful for soil classification. Marbut,[26] for instance, used the presence of carbonate as a criterion for the highest category of his soil classification—Pedocal (soils with carbonate accumulation) and Pedalfers (soil with Al and Fe accumulation). Today, studies of pedogenic carbonate have expanded to include questions about paleoclimate, paleoecology, paleoatmospheric composition, global carbon cycles and the greenhouse effect.

References

1. Morse, J.W.; Mackenzie, F.T. *Geochemistry of Sedimentary Carbonates. Developments in Sedimentology* 48; Elsevier: Amsterdam, 1990.
2. Doner, H.E.; Lynn, W.C. Carbonate, halide, sulfate, and sulfide minerals. In *Minerals in Soil Environments,* 2nd Ed.; Dixon, J.B., Weed, S.B., Eds.; Soil Science Society of America: Madison, WI, 1989; 279–330.
3. Capo, R.C.; Whipkey, C.E.; Blachère, J.R.; Chadwick, O.A. Pedogenic origin of dolomite in a basaltic weathering profile, Kohala Peninsula, Hawaii. Geology **2000**, *28,* 271–274.
4. Gile, L.H.; Peterson, F.F.; Grossman, R.B. The K horizon—a master soil horizon of carbonate accumulation. Soil Sci. **1965**, *99,* 74–82.
5. Soil Survey Staff. *Soil Taxonomy—A Basic System of Soil Classification for Making and Interpreting Soil Surveys,* 2nd Ed.; USDA Agriculture Handbook Number 436; U.S. Govt. Printing Office: Washington, DC, 1999.
6. Birkeland, P.W. *Soils and Geomorphology,* 3rd Ed.; Oxford University Press: New York, 1999; 430.
7. Sobecki, T.M.; Wilding, L.P. Formation of calcic and argil-lic horizons in selected soils of the texas coast prairie. Soil Sci. Soc. Am. J. **1983**, *47,* 707–715.
8. Pal, D.K.; Dasog, G.S.; Vadivelu, S.; Ahuja, R.L.; Bhat- tacharyya, T. Secondary calcium carbonate in soils of arid and semiarid regions of India. In *Global Climate Change and Pedogenic Carbonates*; Lal, R., Kimble, J.M., Eswaran, H., Stewart, B.A., Eds.; Lewis Publishers: London, 2000; 149–185.
9. West, L.T.; Wilding, L.R.; Rabenhorst, M.C. Differentiation of pedogenic and lithogenic carbonate forms in texas. Geoderma **1987**, *43,* 271–287.
10. Rabenhorst, M.C.; Wilding, L.P.; West, L.T. Identification of pedogenic carbonates using stable carbon isotopes and microfabric analysis. Soil Sci. Soc. Am. J. **1984**, *48,* 125–132.
11. Drees, L.R.; Wilding, L.P. Micromorphic record and interpretations of carbonate forms in the rolling plains of texas. Geoderma **1987**, *40,* 157–175.
12. Nordt, L.C.; Hallmark, C.T.; Wilding, L.P.; Boutton, T.W. Quantifying pedogenic carbonate accumulations using stable carbon isotopes. Geoderma **1998**, *82,* 115–136.
13. Jenny, H.; Leonard, C.D. Functional relationships between soil properties and rainfall. Soil Sci. **1934**, *38,* 363–381.
14. Arkley, R.J. Calculation of carbonate and water movement in soil fromclimatic data. Soil Sci. **1963**, *96,* 239–248.

15. Gile, L.H.; Hawley, J.W.; Grossman, R.B. *Soils and Geomorphology in the Basin and Range Area of Southern New Mexico—Guidebook to the Desert Project*; New Mexico Bureau of Mines and Mineral Resources, Memoir 39: Socorro, New Mexico, 1981; 222 pp.
16. McFadden, L.D.; Tinsley, J.C. Rate and depth of pedogenic- carbonate accumulation in soils: formation and testing of a compartment model. In *Soils and Quaternary Geology of the Southwestern United States;* Weide, D.L., Ed.; Special Paper 203; Geological Society of America: Boulder, CO, 1985; 23–42.
17. Marion, G.M.; Schlesinger, W.H. Quantitative modeling of soil forming processes in deserts: the CALDEP and CALGYP models. In *Quantitative Modeling of Soil-Forming Processes;* Bryant, R.B., Arnold, R.W., Eds.; Soil Sci. Soc. Am. Spec. Publ. 39: Madison, WI, 1994; 129–145.
18. Scharpenseel, H.W.; Mtimet, A.; Freytag, J. Soil inorganic carbon and global change. In *Global Climate Change and Pedogenic Carbonates*; Lal, R., Kimble, J.M., Eswaran, H., Stewart, B.A., Eds.; Lewis Publishers: London, 2000; 27–42.
19. Goudie, A. *Duricrusts in Tropical and Subtropical Landscapes;* Oxford Univ. Press: London, 1973; 174.
20. Rabenhorst, M.C.; Wilding, L.P. Pedogenesis on the Edwards plateau, Texas: III. A new model for the formation of petrocalcic horizons. Soil Sci. Soc. Am. J. **1986**, *50,* 693–699.
21. Monger, H.C.; Gallegos, R.A. Biotic and abiotic processes and rates of pedogenic carbonate accumulation in the southwestern United States—relationship to atmospheric CO_2 sequestration. In *Global Climate Change and Pedogenic Carbonates;* Lal, R., Kimble, J.M., Eswaran, H., Stewart, B.A., Eds.; Lewis Publishers: London, 2000; 273–290.
22. Labova, E. *Soils of the Desert Zone of the USSR;* Israel Program for Scientific Translation: Jerusalem, Israel, 1967.
23. Phillips, S.E.; Milnes, A.R.; Foster, R.C. Calcified filaments: an example of geological influences in the formation of calcretes in South Australia. Aust. J. Soil Res. **1987**, *25,* 405–428.
24. Monger, H.C.; Daugherty, L.A.; Lindemann, W.C.; Liddell, C.M. Microbial precipitation of pedogenic calcite. Geology **1991**, *19,* 997–1000.
25. Thorp, J. Effects of certain animals that live in soils. Sci. Monthly **1949**, *68,* 180–191.
26. Marbut, C.F. Subcommission II. Classification, nomenclature, and mapping of soils in the United States. Soil Sci. **1928**, *25*, 61–71.

25

Lead: Ecotoxicology

Sven Erik Jørgensen

Introduction: Dispersion and Application

Lead is found in all environmental components—both living and non-living. The use of lead has dispersed the metal worldwide due to its long-term use in gasoline, batteries, solders, pigments, ammunition, paint, ceramic, and even piping. It is found for instance on the glacial ice and snow of Greenland, one of the most uncontaminated places on the Earth. Many toxic substances are generally widely dispersed and a global increase in the concentration of heavy metals and pesticides has been recorded, as exemplified for lead in Figure 1. The concentration of lead is shown in micrograms per ton (μg/t) as a function of time, which can be found by the use of analyses of ice cores. As it can be seen, the lead concentration has, since the mid-eighth century, increased 10 times—from about 10 μg/t to about 200 μg/t of snow. The dispersion of lead is caused by the many uses of this metal: in mining and smelting, in batteries, in lead-based paints, in electronic devices, in leaded gasoline, and in shots applied in hunting

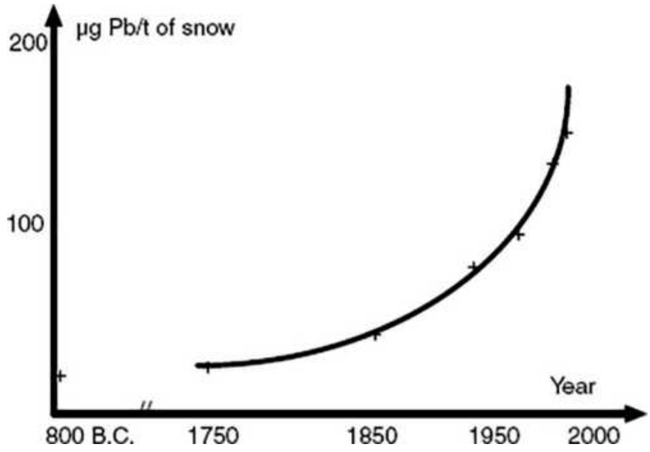


FIGURE 1 The lead concentration in the glacial ice, Greenland, as a function of time, from 800 B.C. to year 2000. The lead concentration has increased tenfold during the last 250 years, mainly due to combustion of leaded gasoline. **Source:** After Jørgensen[2] and Chemistry.[3]

and target shooting, The last two applications were phased out in most industrialized countries more than 25 years ago, but lead gasoline is still in use in many developing countries. In the United States, the manufacture of batteries is the dominant use of lead today.[1] The global dispersion of lead is particularly caused by the combustion of leaded gasoline, which today is inconceivable because the organic lead chemicals were applied in the gasoline to obtain a sufficiently high octane number at the lowest cost, and the less harmful alternatives could only have increased the gasoline price by about a quarter of an American cent per liter.

The concentrations of lead in food items are shown in Table 1 to illustrate the presence of lead in our food—an-other illustration of the consequences of the global dispersion of lead. The concentrations in the table are taken from the mid-1980s, i.e., before the introduction of lead-free gasoline had shown any significant effect. The differences between the three countries are explained by the differences in the traffic (and population) densities.

The concentration of lead in completely uncontaminated water is about 1 ng/L, while concentrations of 20 ng/L are often found when only minor discharge of lead has taken place. In contaminated and very contaminated water, a lead concentration of 100–200 ng/L is often found.

Heavy metals are dispersed globally, but the regional concentrations of most heavy metals may of course be much higher regionally than globally. The relationship between a global and a regional pollution problem and the role of dilution for this relationship are illustrated in Table 2, where the ratios of heavy metal concentrations in the River Rhine and the North Sea are shown. Notice that the amount of nickel and lead used in the region of the River Rhine is the same but the ratio is 70 times higher for

TABLE 1 Lead in Food (see Jørgensen[2])

	Typical Lead Concentration (mg/kg fresh weight)		
Food Items	England	Holland	Denmark
Milk	0.03	0.02	0.005
Cheese	0.10	0.12	0.05
Meat	0.05	<0.10	<0.10
Fish	0.27	0.18	0.10
Eggs	0.11	0.12	0.06
Butter	0.06	0.02	0.02
Oil	0.10	–	–
Corn	0.16	0.045	0.05
Potatoes	0.03	0.1	0.05
Vegetables	0.24	0.065	0.15
Fruits	0.12	0.085	0.05
Sugar	–	0.01	0.01
Soft drinks	0.12	0.13	–

TABLE 2 Heavy Metal Pollution in the River Rhine (from 1985) (see Jørgensen[2])

	River Rhine (t/yr)	Ratio: Conc. in the Rhine/Conc. in the North Sea
Cr	1,000	20
Ni	2,000	10
Zn	20,000	40
Cu	200	40
Hg	100	20
Pb	2,000	700

lead than for nickel due to the application of leaded gasoline, which disperses the lead uncontrolled, while nickel has more closed applications, which allow recycling of the metal. Notice also that lead is transported in the atmosphere primarily in the particulate phase.[4]

Ecotoxicity and Environmental Problems of Lead

The toxicity of Pb is mainly associated with the free ions Pb^{2+} (the +2 oxidation state), but lead can form complexes with hydroxide ions, carbonate, chloride ions, and many organic compounds, for instance, humic acid and amino acids. The toxicity of the complexes is generally lower than the toxicity of the free ions, because the uptake of the complexes is slower than the uptake of the free ions. It implies that it is necessary in every case study to determine by analytical methods or chemical calculation the concentrations of lead as free ions and in the form of the various complexes to determine the toxicity. The toxicity of lead declines with higher concentrations of the hardness ions, calcium, and magnesium. Lead shows, as other heavy metals, bioaccumulation and biomagnification, which is more pronounced at lower pH, because the solubility and the relative concentration of the free ions are increasing with decreasing pH.

Contaminated aquatic ecosystems have a significantly elevated concentration of lead in the sediment. Generally, the sediment has higher concentrations of heavy metals than the water, and as it is possible to analyze sediment core,[5,6] it is possible to find the contamination of heavy metals as functions of the time, provided the settling rate is known or can be estimated. This is particularly informative in the case of lead, because the use of leaded gasoline started shortly after the Second World War and was banned in industrialized countries before or around the mid-1980s. It entails that the sediment, from the approximately 40 years when leaded gasoline was used, will show a particularly high lead concentration, which of course facilitates the dating of the sediment. The lead concentration in sediment is usually 10–200 mg/kg dry weight, but as much as 3000–10,000 mg/kg dry weight can be found in contaminated areas.

A filter feeding bivalve mollusc shows a contamination of lead (and other heavy metals) that is proportional to the concentration in the sediment. The proportional constant is dependent on the composition of the sediment, but it is frequently between 0.01 and 0.05—the highest values for sediment with a high concentration of organic matter.[7]

A major source of lead exposure and toxicity for wild birds is the ingestion of lead-based ammunition. For birds, concentration of lead is in the order of 0.2 mg/kg dry weight, while a toxic effect would correspond to 100 times as much and death to 250 times as much.[5] LD_{50} for rats is 130 mg/kg dry weight (see *Inorganic Compounds: Eco-Toxicity,* p. 1479). There is a primary lead shot poisoning from direct ingestion of lead-based ammunition and a secondary lead shot poisoning when birds (and of course also other animals) ingest lead shotgun pellets and bullet fragment embedded in the flesh of dead or wounded animals shot with lead-based ammunition.

Lead is bound to SH- groups in the proteins and can therefore generally be taken up by plants with high concentrations of proteins more effectively than by plants with low protein concentrations; see also the entry about "Bioremediation." Removal of lead from areas that have been applied for target shooting with ammunition containing lead is possible by bioremediation.[8]

With other heavy metals, the uptake of lead from food is relatively low—about 7%–10%,[9,10] while lead is taken up from the atmosphere by the lungs with a higher efficiency. Figure 2 shows a steady-state model of the uptake for an average European in the mid-1980s, just before the use of leaded gasoline was banned for all new cars. The uptake from food is as seen 10%, namely, 30 μg/day out of the 300 mg/day that is in the food (compare with Table 1). The amount of lead in the 20 m^3 of air that is daily used for respiration is about 25 mg, but as it is taken up with an efficiency of 50%, as much as 12.5 μg lead per day was accumulated in the body of an average European 25 years ago due to direct atmospheric pollution. The 30 mg/day is excreted mainly through the urine and therefore the 12.5 mg lead per day is accumulated in the body—mainly in the bones, where their effect fortunately is very minor. Due to the reduced use of leaded gasoline, the average European will today have less lead in the body. Food contains roughly half as much lead today than 25 years ago and the atmospheric pollution is also one half of the level

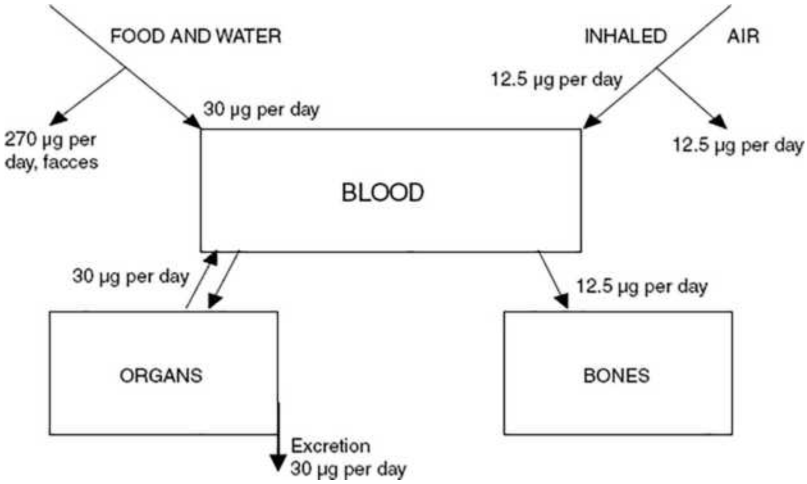


FIGURE 2 A steady-state model of an average European in 1985. The uptake from food was 30 µg/day out of 300 µg/day (10% efficiency) and 12.5 µg lead per day by respiration (uptake efficiency, 50%). The excretion balances the uptake from the food, and the lead taken up by respiration is accumulated in the bones.

today compared with 1985. It means that the amount of lead from food today is 15 mg lead per day and that from respiration is 6.25 µg lead per day. Excretion is also reduced to a level about 15–20 µg lead per day, which implies that less lead is accumulated in the bones today. The indicated amounts of lead for an average European today are found by use of a model that was calibrated and validated by the use of the amounts from 1985—it means the values that are shown in Figure 2.

A conceptual diagram for a lead model of a food chain in an aquatic ecosystem is shown in Figure 3. The boxes represent state variables, which in this case are lead in water, lead in phytoplankton, lead in zooplankton, lead in planktivorous fish (fish I), lead in carnivorous fish (fish II), and lead in the sediment. The two latter concentrations are the highest in most cases. At each level in the food chain, lead is taken up from the water and from the food. The process rates are dependent on the organisms, the temperature, the pH, and the concentrations of the free lead ions and lead complexes. Quantification of the all the processes will reveal that the lead concentration is increasing at a factor of 10–100 through the food chain. This is in contrast to a factor of about 10,000 for DDT. The difference is due to the less effective uptake of lead from the food—as mentioned, the uptake efficiency for heavy metals in food items is only about 7%–10%. DDT is taken up from food with an efficiency of 90% due to the low solubility in water and high solubility in fat tissue.

The spectrum of toxicological and ecotoxicological effects of lead is very wide. Acute toxicity of lead can cause headaches, irritability, and loss of appetite. Chronic toxicity can cause brain damage, reduced memory, anemia, liver and kidney damage, and possibilities of cancerous tumors of the kidney. It has been shown that if children are exposed to high lead concentrations, it will have a pronounced effect on their learning ability.[11] Elevated lead concentrations have also teratogenic effects,[12] and prenatal lead exposure has been demonstrated to be associated with an increased risk of malformations.[13] Genetic effects on animals have also been observed.[12] Epidemiological studies have shown that lead is related to the risk of elevated blood pressure.[14]

Significantly different lead concentrations have been found in the blood of people living far from towns and those in urban areas. Indians in the Amazons have about 8 ng lead per gram of blood. Farmers living in the European countryside have about 50 ng/g of blood and people living in very big cities (New York for instance) have about 150 ng/g of blood.[15] The highest concentrations of lead have been found in policemen who regulated traffic in industrialized countries before lead was phased out: 150–700 ng/g blood. Lead in clean air is as low as 0.0005 ng/L, while in urban areas, it is mostly between 2 ng/L and 25 ng/L.[16] In very polluted urban areas, the concentration may reach 50 ng/L, which could

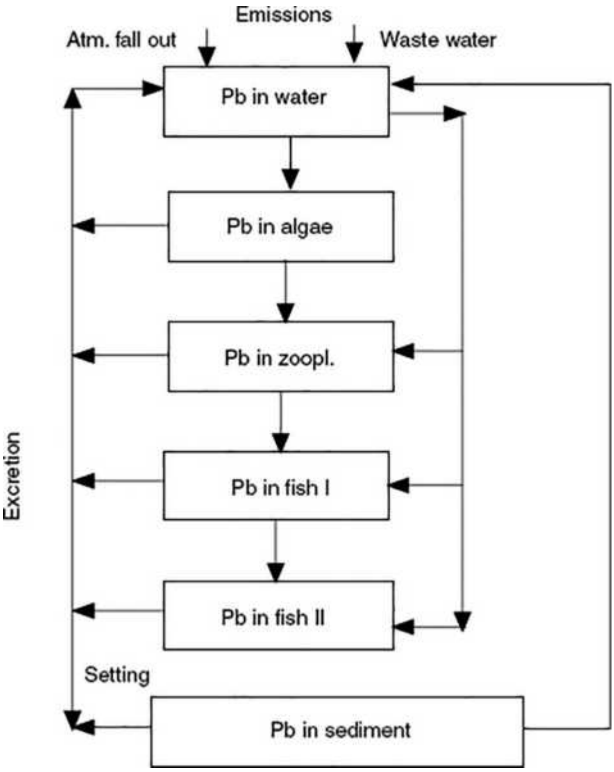


FIGURE 3 Conceptual diagram of biomagnifications of lead through the food chain in an aquatic ecosystem. Notice that each level in the food chain takes lead up from the water and from the food, but each level also excretes lead to the water.

give a daily uptake of 0.345 mg per person and a lead concentration in blood about 720 ng/g.[16] Figure 2 shows a model that could be used to find these values in the blood, when the air pollution level is known.

Lead in soil shows a similar wide range of values. Generally, a lead concentration of 0.2–5 mg/kg dry matter in agricultural areas is found,[17] but as a concentration as high as 100 mg/kg dry matter is found in particularly contaminated soil,[18] and in some extreme cases, concentrations as high as 10,000 mg/kg dry matter have been found.[19] A typical average soil in an industrialized country contains about 20 mg/kg dry matter of lead, and 95% of soil samples randomly sampled will show concentrations between 10 and 60 mg/kg dry matter. Lead concentrations above 1000 mg/kg dry weight in soil can kill earthworms and springtails.

The half-life of lead in human is about 6 years (whole body) and about 3 times as much for the skeletal system. It has been shown that skeletal burdens of lead increase linearly with age, while the non-skeletal burden is eliminated faster (as indicated, whole-body lead has a half-life of 6 years, which means that the non-skeletal lead is exchanged faster than once every 6 years). It is possible to reduce the body burden for patients clinically affected by lead by ethylene-diamine-tetra-acetate (EDTA).[20]

Abatement Methods for Reduction of Lead Pollution

Two major sources of lead pollution, namely, the combustion of leaded gasoline and the use of lead ammunition, have been eliminated in most industrialized countries by environmental legislation. The use of lead in ceramic and paints has similarly been reduced significantly by legislation or by agreement between environmental agencies and the industry.

Lead pollution has been reduced considerably during the last 25–30 years due to environmental legislation and the treatment of wastewater and contaminated smoke and air. Environmental technological solutions have been increasingly applied for the treatment of water and air. A description of the environmental technological methods can be found in the following entries of this encyclopedia:

- *Ion Exchange Application for Treatment of Water and Wastewater*
- *Wastewater Treatment: Overview of Conventional Methods*
- *Municipal Wastewater and Its Treatment*
- *Air Pollution and Environmental Technology: An Overview*

Contaminated land has also been treated, both by environmental technological methods and by bioremediation (see *Bioremediation*, p. 408).

A few cases of successful application of cleaner technology (replacing lead by less harmful components) have also been reported in the journal *Cleaner Technology*. The results of these efforts are encouraging for the use of a consequent environmental management. When environmental legislation, environmental technology, cleaner technology, and ecotechnology are working hand in hand, it is possible to achieve good pollution abatement results, as the lead pollution problem has demonstrated.

Use of Integrated Environmental Management in the Case of Lead Contamination

As with all pollution problems, lead contamination is complex. This section briefly discusses how to go around this complexity to propose an integrated and holistic environmental management and thereby solve the problems properly. The discussion in this section is in principle valid for all pollution problems (compare to the Topical Table of Contents and the How to Use This Encyclopedia in the frontmatter) although particularly for all heavy metal contaminations. The following crucial questions require answers in the case of lead contamination:

1. Which forms have the lead-free ions, or which complexes in which concentrations?
2. What are the sources to the problem? Quantitatively?
3. What possibilities do we have to eliminate which sources? Will that be sufficient to solve the problem?
4. How can we best combine the methods to solve the problem?

It is recommended to consider the following points to be able to arrive at an answer for the crucial questions:

1. The forms of lead can be found either analytically or by chemical calculations;[21] where straightforward chemical calculations are clearly shown for heavy metals with illustrative examples.
2. It is advantageous to set up a mass balance. Figure 4 shows a mass balance for the lead contamination of 1 ha Danish agricultural land. The mass balance clearly reveals the important sources. Atmospheric fallout is the dominant source, although it may also be beneficial to reduce the lead contamination coming from sludge and fertilizers. Particularly the first one of these two could be eliminated.
3. It is possible to eliminate the air pollution of lead as it has been discussed by phasing out the use of lead in gasoline. The second most important air pollution source, in many countries, is coal-fired power plants, which can be eliminated by changing to other forms of fossil fuel or to alternative energy sources. This change of the energy policy will, however, often be prohibitively expensive, and it is therefore as indicated a political question. The phasing out of lead in gasoline may be sufficient. The core question is this: how much will the lead contamination in food be reduced if we reduce the atmospheric fallout so and so much? The answer requires calculations and sometimes the use of models. Lead models have been developed by Jørgensen.[7,9] Lead models considering

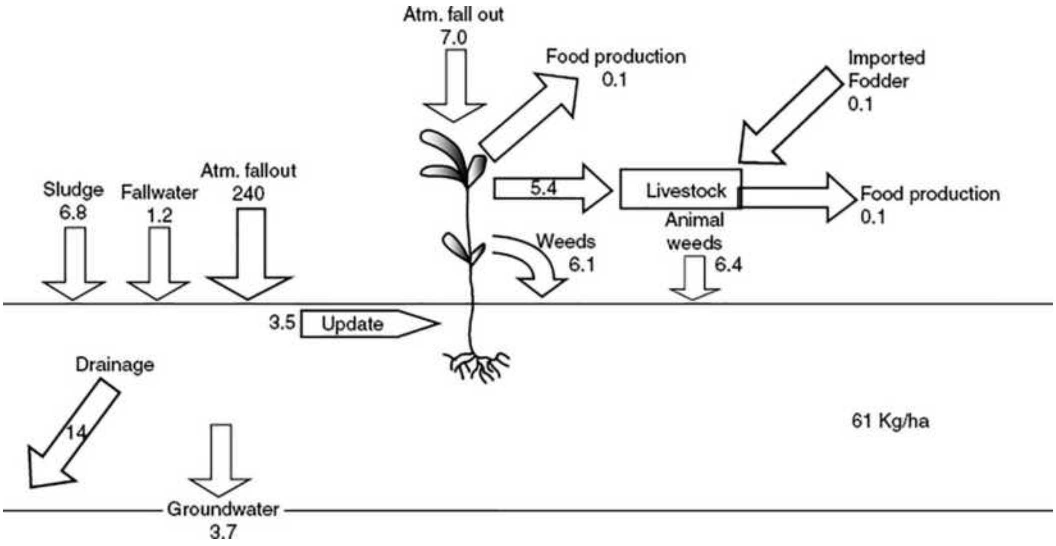


FIGURE 4 Lead balance of average Danish agriculture land. All rates are expressed as grams of Pb per hectare per year.

hydrodynamics, bioconcentration, bioaccumulation, excretion, and sedimentation can furthermore be found in Lam and Simons[22] and Aoyama et al.[23] Another core question is: if we are able to reduce the concentration in food, sediment, or soil to desired levels, would this reduction be sufficient to reduce or even eliminate the effect? This question will require a comprehensive overview of the toxicological and ecotoxicological literature about lead and its effect. If information about the effect is not available, it is necessary to perform bioassay.[24,25] These references describe in detail how to perform bioassay for heavy metals including lead.

4. A discussion of the answer to the fourth question is covered in “The Contents of the Encyclopedia and How to Use the Encyclopedia for Integrated Ecological and Environmental Management.” Other examples are given in Jørgensen et al.[21]

References

1. Laws, E.A. *Aquatic Pollution: An Introductory Text,* 2nd Ed.; John Wiley and Sons: New York, 1993; 380 pp.
2. Jørgensen, S.E. *Principles of Pollution Abatement;* Elsevier: Amsterdam, 2000; 520 pp.
3. Chemistry. The lead we breathe. Chemistry **1968**, *41,* 7–12.
4. Haygarth, P.M; Jones, K.C. Atmospheric deposition of metals to agricultural surfaces. In *Biochemistry of Trace Metals;* Adriano, D.C., Ed.;. Lewis Publ.: Boca Raton, 1992; 423 pp.
5. Jørgensen, S.E.; Fath, B. *Ecotoxicology;* Elsevier: Amsterdam, Oxford, 2010; 390 pp.
6. Forstner, U.; Wittmann, G.T.W. *Metal Pollution in the Aquatic Environment;* Springer Verlag: Heidelberg, Berlin, New York, 1979; 490 pp.
7. Jørgensen, S.E.; Fath, B. *Fundamentals of Ecological Modelling,* 4th Ed.; Elsevier: Amsterdam, Oxford, 2011; 396 pp.
8. Mitsch, W.J.; Jørgensen, S.E. *Ecological Engineering and Ecosystem Restoration*; John Wiley: New York, 2004; 410 pp.
9. Jørgensen, S.E.; Bendoricchio, G. *Fundamentals of Ecological Modelling*, 3rd Ed.; Elsevier: Amsterdam, Oxford, 2001; 530 pp.

10. Newman, M.C.; Unger, M.A. *Fundamentals of Ecotoxicology*; 2nd Ed.; CRC Press: Boca Raton, London and New York, 2003.
11. Bellinger, D.; Leviton, A.; Needleman, H.L.; Waternaux, C.; Rabinowitz, H.B. Low-level lead exposure and infant development in the first year. Neurobehav. Toxicol. Teratol. **1986**, *8,* 151–161.
12. Swedish EPA. *Om Metaller,* 2nd Ed.; Statens Naturvårdsverk: Sweden, 1980.
13. Needleman, H.L.; Rabinowitz, M.; Leviton, A.; Linn, S.; Schoenbaum, S. The relationship between prenatal exposure to lead and congenital anomalies. J. Am. Med. Assoc. **1984**, *251,* 2956.
14. Schwartz, J.; Angle, C.; Pitcher, H. Relationship between childhood blood lead level and stature. Pediatrics **1986**, *77*, 281–288.
15. Jørgensen, S.E.; Jorgensen, L.A.; Nors Nielsen, S. *Handbook of Ecological and Ecotoxicological Parameters*; Elsevier: Amsterdam, 1991; 1380 pp.
16. Francis, B.M. *Toxic Substances in the Environment*; John Wiley and Sons: New York, 1994; 362 pp.
17. Prost, R. *Contaminated Soil*; Institut National de la Recherche Agronomique: Ann Arbor, MI, 1995; 525 pp. +CD.
18. Hutschinson, T.C.; Gordon, C.A.; Meema, K.M. *Global Perspectives on Lead Mercury and Cadmium Cycling in the Environment*; Wiley Eastern Limited: Ann Arbor, MI, 1994; 412 pp.
19. Kabata-Pendias, A. *Effects of Trace Metals Excess in Soils and Plants*; CRC: Boca Raton, 1986; 330 pp.
20. Nigel, B. *Environmental Chemistry*; 2nd and 3rd Eds.; Wuerz Publ. Ltd.: Winnepeg, 1994 and 2000; 378 pp. and 420 pp.
21. Jørgensen, S.E.; Tundisi, J.G.; Tundisi, T. *Handbook of Inland Aquatic Ecosystem Management*; CRC: Boca Raton, 2012; 480 pp.
22. Lam, D.C.L.; Simons, T.J. Computer model for toxicant spills in Lake Ontario. In *Metals Transfer and Ecological Mass Balances, Environmental Biochemistry*; Nriago, J.O., Ed.; Ann Arbor Science; Ann Arbor, MI, 1976; Vol. 2, 537–549.
23. Aoyama, I.; Inoue, Yos; Inoue, Yor. Simulation analysis of the concentration process of trace heavy metals by aquatic organisms from the viewpoint of nutrition ecology. Water Res. **1978**, *12,* 837–842.
24. Newman, M.C.; Jagoe, C.H. *Ecotoxicology. A Hierarchical Treatment;* CRC: Boca Raton, 1996; 412 pp.
25. Landis, W.G.; Yu, M.H. *Introduction to Environmental Toxicology;* Lewis Publ.: Boca Raton, 1995; 330 pp.

26

Lead: Regulations

Lisa A. Robinson

Introduction

Phasing out lead in gasoline is often identified as one of the U.S. Environmental Protection Agency's (USEPA's) greatest achievements.[1] As part of the regulatory process, USEPA economists developed innovative approaches for expressing the benefits of the resulting risk reductions in monetary terms, so that they could be directly compared to the costs of different control options.[2] As evidence mounts that lead may pose hazards even at relatively low exposure levels,[3,4] the methods used to estimate these monetary benefit values are in need of continual improvement and updating for application in future regulatory assessments. This entry reviews how these benefit values have been developed and identifies challenges for subsequent work. While it focuses on the approaches used in U.S. regulatory analyses, these approaches can be adapted for use in other countries and have global relevance.

The entry first reviews previous assessments, summarizing the framework for regulatory benefit-cost analysis and providing examples of how this framework has been implemented to assess the benefits of lead regulations. Because these analyses indicate that the most significant quantifiable benefits include the relationship between lead-related neurological damage and future earnings (for children) and between lead exposure and cardiovascular disease (for adults), it next discusses these benefit categories in more detail. It then describes benefits that have not been fully incorporated into previous regulatory analyses, but that may be desirable to explore in the future. The concluding section summarizes the discussion and highlights related challenges.

Review of Previous Analyses

The USEPA's pathbreaking work on valuing lead benefits was initiated in the late 1970s and early 1980s, as the United States moved to phase out lead in gasoline.[2,5] Lead was completely banned from gasoline in 1996. The valuation approaches evolved as the USEPA subsequently considered regulating lead in drinking water, sewage sludge, and paint, as well as ambient air. Typically, the quantified benefits are dominated by the effect of lead on earnings (due to its effect on IQ) and, in some cases, by its effects on cardiovascular disease, particularly the risks of premature mortality. However, analysts have been unable to quantify many other potential impacts, due largely to gaps or inconsistencies in the research on the association between lead exposure and additional health outcomes.

General Framework

When the USEPA began assessing the benefits of reducing lead exposure, requirements for conducting regulatory benefit–cost analysis were in their infancy.[6] The first comprehensive Executive Order mandating such analysis was issued by President Reagan in 1991.[7] The requirements for these analyses have since been refined, updated, and expanded, although the basic principles remain the same. Thus, the early work on lead was both innovative and pioneering, evolving into approaches used in numerous subsequent analyses.

Currently, benefit–cost analysis is required for major rules under President Clinton's Executive Order 12866,[8] as supplemented by President Obama's Executive Order 13563.[9] These executive orders mandate that Federal agencies assess alternative policies for actions that may be economically significant, i.e., that may lead to a rulemaking that has an annual economic effect of $100 million or more or has important adverse effects. To support implementation of these requirements, the U.S. Office of Management and Budget (OMB) issued Circular A-4, *Regulatory Analysis*,[10] in 2003 and summarized and clarified related requirements in a 2010 checklist for agencies.[11]

Typically, the resulting analyses contain five major components, in addition to sections discussing the rationale for the rulemaking and the regulatory options considered, as summarized in Table 1. These analyses also include both quantitative and qualitative assessment of related uncertainties, as well as information on effects that could not be quantified or valued.

This entry is primarily concerned with the economic valuation of benefits, under step 4b in Table 1. As discussed later, the ability to carry out such valuation is dependent on the extent to which the changes in risks can be assessed under step 4a. Building a strong research base that allows analysts to estimate the changes in risks of different types associated with various changes in lead exposures has been a major challenge.

Ideally, economic values reflect individuals' willingness to pay (WTP) for the risk reductions or other benefits they would receive from a regulation or policy. This approach is consistent with the theoretical framework underlying benefit–cost analysis, which is based on respect for individual preferences (often referred to as "consumer sovereignty"), assuming that each individual is the best judge of his or her own welfare. The consideration of WTP also reflects the types of trade-offs being considered in policy decisions. Given constrained resources, pollution abatement policies inevitably require trading off increased expenditures on risk reductions against decreased expenditures on other desired goods

TABLE 1 Overview of Analytic Components

1. Estimate current and potential future *baseline conditions* in the absence of government intervention.
2. Predict *responses* to each policy option under consideration.
3. Estimate the *national costs* associated with each option, summing the costs of the predicted compliance actions across those subject to the provisions and accounting for resulting market impacts (e.g., changes in consumption due to price increases).
4. Estimate the *national benefits* associated with each option, including the effects on human health and the natural and built environment. For environmental hazards, this generally consists of two steps:
   a. A *risk assessment,* which considers the link between changes in pollution levels and each health or environmental outcome of concern.
   b. An *economic analysis,* which includes monetary valuation of each outcome to the extent possible.
5. Assess the distribution of the impacts across subpopulations of concern. Such subpopulations typically include small businesses as well as sensitive and/or vulnerable subgroups (such as children or low-income individuals) whom policymakers wish to protect against disproportionate adverse effects.

and services on a society-wide level. WTP, the maximum amount of income (or wealth) an individual is willing to exchange for a beneficial outcome, represents this type of exchange. Willingness to accept compensation (WTA), or the smallest amount an individual would accept to forego the improvement, is also consistent with this framework. WTA is used infrequently in regulatory analysis, however, both because it can be difficult to measure and because regulations often involve paying for improvements rather than compensating for harms.

In some cases, individual WTP can be estimated based on consumer demand for market goods. For outcomes not directly bought and sold in the marketplace (such as health risk reductions), WTP is instead estimated from revealed or stated preference studies. Revealed preference studies use data from market transactions or observed behaviors to estimate the value of related non marketed goods or outcomes. For example, the value of mortality risk reductions is often estimated from the relationship between earnings and job-related risks, controlling for other influencing factors. Alternatively, stated preference studies rely on responses to survey questions or similar approaches. For example, a survey respondent may be asked whether he or she would be willing to pay “$X” for a l-in-10,000 reduction in the risk of death associated with decreased exposure to an air pollutant. Each approach has advantages and limitations. In particular, revealed preference studies rely on data from actual markets but often address scenarios that differ from those of concern in policy analysis. Stated preference studies enable researchers to better tailor the scenario to the risks of concern, but the responses are hypothetical and must be carefully elicited.

For many outcomes, estimates of WTP are lacking, and analysts often rely instead on averted costs as rough proxies. In particular, health risk reductions may be valued using cost of illness estimates, including expenditures on medical treatment and often lost productivity.[12,13] The latter is typically valued using the human capital approach, which assumes that workers are paid the value of their marginal product.[12–16] Thus, compensation data are used to value illness-related lost work time. The human capital approach also can be used to assess the value of changes in unpaid work time (i.e., household production and volunteer work), based on estimates of either the wages foregone when an individual chooses to engage in unpaid rather than paid work or the cost of replacing the unpaid worker with one who is compensated.

Another, less commonly used approach is the friction cost method, which assumes that productivity will decrease temporarily while the employer implements measures to replace the absent individual rather than over the full course of the illness.[14,16,17] The “friction period” is defined as the time it takes to find and train a new employee or reallocate duties among existing employees. However, this approach is likely to understate the loss particularly during periods of full employment, because it does not take into account the additional loss that accrues if the new employee was previously working in a different job or involved in nonmarket production.

WTP to reduce the risk of illness may differ from these averted costs. Costs reflect incurred cases, not expected risk reductions, and being ill and treated is typically worse than not being ill at all. Medical costs can be difficult to estimate accurately due to the distorting effects of insurance and other third-party payments, and analysts often rely on average per case costs, which may differ from the marginal costs of small changes in the risk of illness.

Approaches for valuing time losses also often rely on a number of simplifying assumptions regarding the functioning of the labor market and individual choices between paid and unpaid work and leisure. In reality, productivity losses at a societal level will depend on unemployment rates and on the extent to which coworkers compensate for the loss of the ill worker’s time. At an individual level, sick leave and disability insurance may reduce the impact of illness on earnings. In addition, inflexible work hours and other factors limit workers’ ability to choose jobs that reflect their willingness to trade off paid and unpaid work and leisure time; hence, wages may over- or understate their WTP for changes in time use.

For these and other reasons, estimates of averted costs may be less than, equal to, or greater than WTP to avoid the illness. For example, medical costs may frequently understate WTP because they do not reflect the value of avoiding pain and suffering, but may overstate WTP in cases where the availability of insurance leads individuals to seek treatment that they would not be willing to pay for themselves. In addition, although average costs per case are often reported on an yearly basis, few studies track individuals longitudinally and provide data on the lifetime costs associated with particular illnesses. Thus, while the concept of averted costs seems straightforward, the direct and indirect costs of illness can be difficult to estimate. However, they provide a reasonable and widely-used proxy when WTP estimates are not available.

As discussed below, the valuation approaches used in lead regulatory analyses rely largely on averted costs, due to the lack of suitable WTP estimates. Very few WTP studies directly address lead exposures[18] or IQ-related decrements;[19] more generally, WTP for morbidity risks has not been well-studied.[20] In particular, the approach typically used to value lead-related IQ decreases is an averted cost approach, focusing on lost earnings; some analyses also consider related educational costs. The logic is that individual WTP to avoid the IQ loss would be at least equal to the income foregone. Medical costs are typically not included, because no fully effective treatment now exists. While chelation therapy has been used to address high blood lead levels, its effectiveness has been questioned in recent years.[21]

Averted costs are often also used to estimate the value of morbidity risk reductions, including reductions in nonfatal cardiovascular disease associated with decreased lead exposure. WTP estimates are available for mortality risk reductions, however.

Types of Benefits Assessed

The approaches used to value lead-related lost earnings and cardiovascular effects in regulatory analysis have evolved over the past 30 years, reflecting new data as well as increasingly sophisticated methods. One of the first comprehensive benefit analyses of reduced lead exposures was published in 1985 for USEPA regulations developed under the Clean Air Act, addressing the phase down of the amounts of lead allowed in gasoline.[5] The outcomes assessed include reductions in the costs of lead screening and treatment; compensatory education; the risks of hypertension, myocardial infarction, stroke, and premature mortality; and damages to pollution control equipment, vehicle maintenance, and fuel economy; as well as health and welfare effects associated with other pollutants. While the authors indicate that the majority of the benefits (75%) stem from the reduced risks of cardiovascular disease (with mortality risk reductions dominating the results), these effects were not considered in the USEPA's decision making because the underlying epidemiological studies had not yet undergone widespread review. However, even if the benefits of cardiovascular risk reductions are not counted, the costs of the regulations are less than the benefits.

Soon after, the USEPA's 1986 analysis of a drinking water rule added an important new benefit category, the effects of changes in IQ on earnings, while also assessing many of the other benefit categories included in the 1985 analysis of lead in gasoline.[22] Rather than summing the estimates of lost earnings and educational costs, this analysis compares them as alternative measures of cognitive damage. For this rulemaking, the majority of the benefits result from reduced water system corrosion. Of the health-related benefits, lost earnings and total cardiovascular risk reductions are similar in magnitude.

These approaches were refined in subsequent analyses. A particularly significant example is the USEPA's 1997 *Retrospective Analysis of the Clean Air* Act.[23] That analysis covers the effects of the full set of regulations implemented under the Clean Air Act over the years 1970 to 1990. For lead, the benefits assessed include compensatory education and other educational costs, future earnings, neonatal mortality, and cardiovascular disease (hypertension, coronary heart disease, stroke, and premature mortality). The results in this case are dominated by the value of reducing the mortality risks associated with cardiovascular disease.

More recent USEPA analyses focus on the effects of lead on IQ and earnings, including its 2008 assessment of the Lead Renovation, Repair, and Painting Program rule developed under the Toxic Substances Control Act (TSCA) and its 2008 analysis of the lead National Ambient Air Quality Standards (NAAQS) under the Clean Air Act.[24,25] The benefits associated with reducing cardiovascular risks were not quantified in these analyses, because the USEPA found that data on adult blood lead levels were outdated and that more information on the association between these levels and related risks was needed. For IQ and earnings, the two analyses use somewhat different assumptions. For the renovation rule, USEPA relies solely on 1995 estimates of the relationship of IQ to earnings developed by Salkever, while for the NAAQS rule, it also considers the effects of a lower 1994 estimate developed by Schwartz, because recent research suggests that the Salkever estimates may be overstated. These estimates are discussed in more detail in the next section.

In general, these regulatory analyses indicate that the monetized benefits of reductions in lead exposure often far exceed the costs. The most significant benefit categories are the impact of lead on future earnings (for children) and its impact on mortality risks from cardiovascular disease (for adults). Regulations that focus on lead in ambient air or drinking water will affect both adults and children, while those that primarily address the ingestion of lead paint or contaminated soil will largely affect children.

The benefit categories addressed are similar across analyses in part because of the epidemiological research available. The benefit assessments focus on those areas where the risk assessments have found the most consistently strong associations, and where the outcomes are amenable to monetary valuation. The number of lead epidemiological studies is large, and they consider a much wider range of endpoints than reflected in these benefit analyses.[4] Some outcomes are excluded from the benefit-cost analysis because the risk data are weak or inconsistent; for others, the risk assessment shows strong associations but addresses endpoints that are difficult to value in monetary terms. For example, small changes at the cellular level cannot be easily monetized unless information is also available on the likelihood that such changes will lead to noticeable health impairments. As a result, these analyses are likely to understate benefits, but the degree of understatement is uncertain.

Valuing IQ-Related Benefits

The benefit category most frequently assessed in lead-related regulatory analyses is the effect of IQ on earnings. This assessment generally includes two components: determining the percent change in lifetime earnings associated with each one-point change in IQ and estimating the dollar value of lifetime earnings for each cohort affected. Because the first component is far more difficult to address, the discussion that follows focuses on how the percent change in earnings is estimated. The second component, lifetime earnings, can be estimated based on readily accessible data on earnings and survival probabilities, as demonstrated by Grosse et al.,[15] who provide estimates by year of age as of 2007.

Describing these benefits as the effect of lead-associated IQ decrements on earnings is somewhat misleading, however. The underlying epidemiological studies report associations between lead and varying measures of cognitive abilities, as well as measures of behavioral and other problems that may affect both school- and work-related achievement. Thus, this category may include a number of interrelated neurological effects that impact educational attainment, the likelihood of employment, the amount of earned income, and household production.

This entry does not discuss the effects of lead or IQ on school-related costs in detail, because they tend to be a much smaller proportion of total benefits. Reductions in lead exposure may have counterbalancing impacts on these costs. To the extent that reductions improve IQ and other aspects of neurological functioning, the years of education may increase and the need for compensatory education may decrease. These changes may not be completely offsetting, however, due to differences in per-student costs and in the number of students affected. The change in the need for compensatory education is likely to affect a relatively small number of very low IQ children, while the increase in regular education may affect a larger number of children throughout the IQ spectrum. In addition, to the extent that IQ gains increase the number of years of schooling, they will also defer full employment while increasing compensation once employed.

Previous Regulatory Analyses

In simple terms, expected earnings are the product of the likelihood of employment (i.e., labor force participation) and the wages earned if employed. Cognitive ability can affect both wage levels and the probability of employment directly. In addition, cognitive ability can affect the number of years of schooling, which in turn also affects wages and participation rates. Again, the effects of lead may be greater than the effects of cognitive ability alone, because of associated behavioral problems, attentional difficulties, and other neurological effects that can influence functioning in school and at work. These relationships are illustrated conceptually in Figure 1.

Assessing the pathways that lead from blood lead levels to changes in earnings is a complicated task. The foundation for the approach used in USEPA regulatory analyses is described in a 1994 article by Schwartz.[26] These relationships were then re-estimated in 1995 by Salkever,[27] whose estimates have been used in many, if not most, subsequent analyses. However, more recent work by Grosse[28,29] suggests that Salkever may overstate these effects. Disentangling the direct and indirect effects on earnings, to avoid double counting while capturing lead's impacts on earnings to the fullest extent possible, has been a major challenge.

Schwartz focuses on the relationships marked as solid lines (a) through (g) in Figure 1: the direct effect of lead on IQ and earnings (a → b → e), the indirect effect of lead on earnings through schooling and wages (c → d → e), and the indirect effect of lead on earnings through schooling and labor force participation (c → f → g). While these effects are interrelated, they can be summed as long as the underlying studies control for each effect separately. In total, Schwartz estimates that these factors lead to a 1.76% change in earnings for each one-point change in IQ. However, because his analysis includes estimates of the direct effects of lead on schooling, it likely captures some of the effects of lead on behavior (e.g., on attention span) rather than purely measuring the effects of IQ. While the percent change may appear small on an individual basis, the large number of individuals affected by many regulations leads to high total benefits.

Salkever then revisits the approach developed by Schwartz, using data from the National Longitudinal Study of Youth (NLSY) to extend and re-estimate many of the components of the analysis. Salkever adds the two dashed arrows marked (x) and (y) in Figure 1, estimating the effect of IQ on schooling and on labor force participation. He develops estimates separately for women and men.

When weighted to reflect the relative contribution of women and men to total earnings, the Salkever estimates average about a 2.4% change in earnings per IQ point, significantly higher than the Schwartz estimate (1.76%). For example, Schwartz calculated that a 1 μg/dl reduction in1984 childhood blood lead concentrations would yield increased earnings of $5.1 billion annually (1989 dollars). If we instead apply the factors from Salkever, this amount increases by more than 30%, to close to $7 billion. While lead

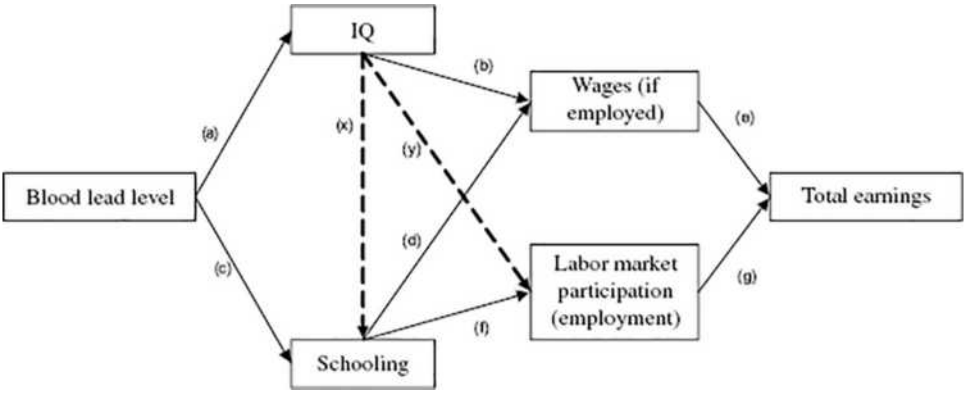


FIGURE 1 Conceptual model of relationship between lead and earnings.

exposure has dropped in recent years, newer epidemiological studies suggest that the effects of lead on IQ may be larger than suggested by this older research, which would further increase these 1984 benefits estimates.

Recent Research Findings

Linking lead-related IQ changes to changes in earnings is difficult for two reasons. First, the longitudinal databases used in the econometric analyses often rely on tests of cognitive ability that differ from those used in epidemiological studies of the associations between lead and cognitive ability. The epidemiological studies generally rely on the Wechsler Intelligence Scale for Children (WISC), which is a standard and widely used test of IQ administered in childhood. Many of the econometric studies (including the Salkever analysis) rely on data from the NLSY, which instead includes results from the Armed Forces Qualification Test (AFQT) administered in 1980 when NLSY participants were between ages 15 and 23. The AFQT measures achievement as well as ability. Thus, while IQ tends to be relatively stable after early childhood, the inclusion of achievement means that AFQT scores can change as a result of subsequent education and work experience.

Second, the causal relationships are complex and not necessarily well understood, posing challenges in terms of both measurement and statistical analysis. A wide range of personal, family, and community factors (including education system and job market characteristics) may affect earnings, and also may be affected by IQ. The causal relationships can be somewhat circular. Green and Riddell[30] provide a simple example: if we are trying to estimate the independent effects of IQ and years of schooling on earnings, but higher-IQ individuals stay in school longer because their psychic costs of schooling (e.g., the difficulty of studying) are lower, then IQ is in part causing the in-creased schooling and the contributions of the two factors to earnings are difficult to separate.

More recent evidence on the relationship between IQ and earnings is reviewed by Grosse.[28,29] Some of these studies use the same NLSY data set as the Salkever study but implement various types of adjustments. The NLSY began tracking individuals in 1979 when they were between ages 14 and 22; Salkever and the more recent studies discussed below use 1990 NLSY data, which include information on educational attainment, hourly wages, and hours worked as of ages 25 to 33. As noted earlier, the measure of cognitive ability in the NLSY is the AFQT, administered in 1980, when many respondents were already working or pursuing postsecondary education.

Recognizing that postsecondary job experience and education are likely to increase AFQT scores, in 1996, Neal and Johnson[31] re-estimated the relationship of IQ to earnings as part of their study of earnings differentials between blacks and whites. They restrict their NLSY sample to those who were age 18 or younger when they took the AFQT, so that their estimates of the relationship between ability and earnings are less inflated by the effects of further education and experience than the estimates used by Salkever.

In 2002, Grosse, Schwartz, and colleagues[28] used the results from the Neal and Johnson study to assess the economic gains associated with the U.S. reduction in blood lead levels that occurred from 1976 to 1999. According to Grosse et al., the Neal and Johnson study indicates that a one-point IQ loss results in a 1.15% difference in earnings for men and a 1.52% difference for women. These results are from models that do not control for schooling; hence, Grosse et al. do not add separate estimates of the effects of schooling on earnings to avoid double counting. However, because Neal and Johnson do not estimate the impact of ability on labor force participation, Grosse et al. add the estimate of this effect from Salkever. Once weighted to reflect the relative share of earnings by men and women in the labor force, Grosse et al. find that a one-point IQ loss is associated with a 1.66% change in earnings, slightly lower than the Schwartz estimate of 1.76% and substantially less than the Salkever estimate of 2.4%.

In his 2007 review, Grosse also discusses a newer study by Heckman et al.[32] These authors rely on the same NLSY data as discussed previously, considering the factors that influence wages at age 30. The modeling is substantially more sophisticated than prior studies, considering a wide array of personal, family, and community characteristics, including standard demographic variables as well as behavioral choices (such as smoking and participation in crime). It takes into account latent (i.e., unobserved) factors representing cognitive and non-cognitive ability and considers how these factors influence schooling choices as well as labor market choices and earnings.

Comparing the Heckman et al. findings to the findings from other studies requires some conversion. Heckman et al. multiply their coefficients by the standard deviation of the associated variable in reporting their results. Because IQ is standardized so that one standard deviation equals 15 IQ points, this entry follows Grosse[29] and divides the Heckman et al. coefficients on cognitive ability by 15 to determine the effect per IQ point. However, it uses results from Heckman et al.'s model 2 (with corrected AFQT scores) in Table 5 of Heckman[32] rather than their model 3 (as in Grosse, which relies on latent measures of ability), because the latter are more difficult to interpret. Depending on the controls included in the statistical modeling, the results suggest that a one-unit change in cognitive ability raises wages for males by 0.44% to 0.95% per IQ point.

This result is lower than the results from Neal and Johnson (1.15% for men and 1.52% for women), which do not control for schooling. Adding the pathways not included in the Heckman et al. analysis, i.e., the effect of ability on labor force participation, would increase the Heckman et al. estimates somewhat, as would including women (for whom returns to ability appear higher than for men). Perhaps more importantly, Heckman et al. find that the effects of non-cognitive factors are significant. Lead has many behavioral effects not captured in the epidemiological studies of the association between lead exposure and IQ; considering the effect of these factors could noticeably increase the impact of lead exposure on earnings.

Grosse also identifies another study, by Zax and Rees,[33] that uses an alternative data source: the Wisconsin Longitudinal Study (WLS) of Social and Psychological Factors in Aspiration and Attainment. This study includes IQ scores based on the Henmon–Nelson Test of Mental Ability administered in the 11th grade. The authors explore the effects of factors measured at age 17 (in 1957) on male earnings at ages 35 and 53. Zax and Rees find that household, community, and peer characteristics all affect earnings at both ages. Depending on the extent to which the model controls for these factors, the effect of a one-IQ-point change on earnings ranges from 0.36% to 0.74% at age 35, rising to 0.90%–1.39% at age 53, for the subsample who had earnings at both ages (Tables 5 and 6 of Zax and Rees[33]). The effect of IQ on earnings for the full sample at age 35 was almost identical to the effect for this subsample at that age.

These findings have four important implications for considering the effects of lead on earnings. First, the socioeconomic status of those affected by reduced lead exposures may affect the relationship between IQ and earnings. The Zax and Rees results for age 35 are somewhat lower than the factors found by Heckman et al. at age 30. However, the Zax and Rees sample is from a relatively affluent state, which may not be representative of the national population, and they rely on data for high school graduates. In contrast, the NLSY data used by Heckman et al. represent a wider population. Different relationships between IQ and earnings may hold if lead exposures disproportionately affect lower-income individuals, who may be less likely to complete high school and for whom the returns to IQ may be influenced dissimilarly by household and community characteristics.

Second, these findings suggest that using a constant factor, based on the IQ-to-earnings relationship for ages around 30 to 35, will understate the effect of IQ on earnings at older ages. The effect of this changing relationship on the present value of lifetime earnings is unclear, however, because it depends on three factors: the extent to which the IQ-and-earnings relationship varies across all years of age, the extent to which earnings change with age, and the discount rate used to estimate the present value of future earnings.

Third, the Zax and Rees analysis further emphasizes some of the difficulties inherent in sorting out causal links. Their model with the most controls (which leads to the lowest proportional changes; i.e., the 0.36% and 0.90% above) includes variables such as whether parents encouraged the individual to attend college. Parents may be more encouraging if their child has a higher IQ, or this encouragement may be related to other factors (such as whether the parents attended college or believe that college is affordable). To the extent that IQ leads to such encouragement, then controlling for this encouragement will understate the effects of IQ.

Finally, applying a longitudinal study to the present time increases uncertainty. To track individuals from childhood to their peak earning years requires starting at least 30 years in the past. Whether these past relationships apply to children exposed to lead in the current year is uncertain, given changes in the educational system, labor market, and other factors over time.

In sum, newer research suggests that both the Salkever estimate of a 2.4% change in earnings per one-point change in IQ and the older Schwartz estimate of 1.76% are overstated. The degree of overstatement is uncertain, however, due to the difficulties inherent in disentangling the effects represented in Figure 1. In the near term, analysts may want to test the sensitivity of their results to a range of factors, as does the USEPA in its 2008 NAAQS analysis.[25]

In the longer term, more detailed review of recent research on the relationship between IQ and earnings would be desirable, given both the complexity of the relationships and their importance in assessing the benefits of lead abatement. This review could include determining whether it is possible to use the data that underlie the reported results to estimate factors that are more directly applicable to the populations most affected by changes in lead exposures. New research may be needed that directly addresses these populations and focuses on the data needed to estimate earnings-related benefits. The research cited above suggests that the impacts of lead on non-cognitive functioning may also have important effects on education and earnings that should be further investigated. In addition, more work is needed to estimate individual WTP to avoid these risks, given that WTP is a more conceptually correct indicator of the value of related benefits.

Valuing Cardiovascular Risk Reductions

For cardiovascular risks, suitable WTP estimates that incorporate both morbidity and mortality are not currently available; thus, these two components are typically valued separately in regulatory analysis and then summed. Well-established approaches exist for valuing mortality risk reductions, based on estimates of WTP, but few studies address WTP for morbidity risks.[20] As a result, cost-of-illness estimates are generally used to value reductions in nonfatal cardiovascular risks.

The suite of lead-related cardiovascular conditions assessed varies across USEPA's benefit analyses but has at times included hypertension, stroke, and myocardial infarction or coronary heart disease more generally. The approaches for valuing these conditions have evolved over time, reflecting changes in the available data.

For mortality risk reductions, the USEPA currently follows a valuation approach based on a review conducted in the early 1990s and described in both its 2000 and 2010 *Guidelines for Preparing Economic Analyses*.[34,35] The estimates are typically expressed as the "value per statistical life" or VSL, to parallel how risk reductions are typically expressed in policy analyses.[36,37] Most regulations lead to small risk changes at the individual level, which are often presented as "statistical" cases aggregated across the affected population. For example, a 1-in-10,000 risk reduction affecting 10,000 individuals can be expressed as a statistical case (1/10,000 risk reduction × 10,000 individuals = 1 statistical case), as can a 1-in-100,000 risk reduction affecting 100,000 individuals (1/100,000 risk reduction × 100,000 individuals = 1 statistical case). For most regulations, the specific individuals who would avoid illness or whose lives would be extended by the policy cannot be identified in advance. A regulation that is expected to "save" a statistical life is one that is predicted to result in one less death in the affected population during a particular time period. "Saving" a statistical life is not the same as saving an identifiable individual from certain death.

The VSL is typically calculated by dividing individual WTP for a small risk change (in a defined time period) by the size of the risk change. For example, if an individual is willing to pay $700 for a 1-in-10,000 reduction in his or her risk of dying in the current year, the VSL is $7 million ($700 ÷ 1/10,000 = $7 million). Alternatively, individual WTP for small risk reductions could be aggregated across a population. A \$7 million VSL also results if each member of a population of 10,000 is willing to pay an average of \$700 for a 1-in-10,000 annual risk reduction (\$700 × 10,000 = \$7 million).

The USEPA has devoted considerable attention to research on the value of mortality risk reductions. At present, it recommends applying a mean VSL of 7.4 million dollars with a standard deviation of \$4.7 million (in 2006 dollars), updated to the appropriate dollar year.[35] Analysts typically adjust this value to reflect changes in real income over time as well as any significant delays between changes in exposure and changes in mortality incidence (i.e., latency or cessation lag).

This approach is likely to evolve in the near future, as a result of new research and expert review. The USEPA asked its Science Advisory Board to review a 2010 White Paper it drafted on changes in its approach.[38] In 2011, the Science Advisory Board recommended that EPA 1) change the VSL terminology to avoid confusion between the value of small risk reductions and the value of saving an individual's life; 2) recognize that values vary across contexts and, to the extent possible, use values tailored to the particular risk to be regulated; 3) apply enhanced criteria to determine which valuation studies are of sufficient quality for application in policy analysis; and 4) conduct additional research on topics such as whether individuals are willing to pay differing amounts for risk reductions with differing characteristics.[39] In the near term, the extent to which these recommendations will significantly change the values used in USEPA regulatory analyses is unclear; over the longer term, an ambitious research agenda will be needed to fully meet the challenges raised by the report.

For reductions in morbidity risks, the USEPA has long relied on cost-of-illness estimates due to the lack of suitable estimates of WTP. These estimates have changed over time, but typically include the direct costs of a wide range of inpatient and outpatient services, supplies, and medications. For indirect costs, the estimates include lost market work time and, at times, lost household production.

The most recent (2008) USEPA analyses do not include cardiovascular risks in the monetized benefits estimates, because of concerns about the data available on adult blood lead levels and on the effects of lead on these risks. Previous lead analyses, such as a 2006 assessment of the USEPA's proposed renovations rule,[40] relied on relatively old cost data to value nonfatal risks, collected in the 1970s and 1980s. Given that medical treatment and labor market characteristics have changed substantially since that time, more review of research is needed to determine the appropriate values for these impacts, regardless of whether WTP or cost-of-illness methods are used.

Other Challenges and Opportunities

While there are many opportunities to improve the approaches used previously to assess the benefits of increased regulation of lead exposure, perhaps the largest challenge is expanding the types of outcomes considered. When comparing benefits to costs, policy analysts and decision makers understand that monetized benefits are likely to understate total benefits, because they exclude the value of many potential outcomes. However, without more data and analysis, it is difficult to determine whether the excluded outcomes would have a relatively small or relatively large impact on the benefit estimates, and hence whether they suggest that a substantial increase in spending on these regulations might be worthwhile.

Expanding the types of benefits that are quantified poses challenges in terms of both the risk assessment and the valuation approach. As illustrated in Box 1, estimating the benefits of lead regulations first requires data on the associated risk reductions (step 4a). The discussion of previous regulatory analyses indicates how concerns about the quality of the risk data can limit the inclusion of some benefit categories (e.g., cardiovascular risk reductions), despite the availability of a valuation method. In addition, double counting may be a concern when various epidemiological studies cover overlapping outcomes, as illustrated by the discussion of the relationship between IQ, schooling, labor force participation, and earnings, as well as the relationship to other types of neurological damage associated with lead.

The USEPA periodically reviews the scientific evidence linking lead exposures to different outcomes. The most recent draft of its *Integrated Science Assessment for Lead*,[4] which is now being reviewed, covers seven health-related categories: neurological effects, cardiovascular effects, renal effects, immune system effects, effects on heme synthesis and red blood cell function, reproductive effects and birth outcomes, and cancer. It concludes that there is a causal relationship for all of these outcomes except cancer, for which the causal relationship is described as "likely." In addition, the report covers a number of ecological effects and determines that lead affects growth, mortality, and other factors related to the health of terrestrial and aquatic systems. Careful review of these data is needed to develop estimates of the relationships between exposure reductions and risk reductions that are suitable for use in regulatory analysis.

Another challenge is determining which outcomes will be noticeably affected by a particular regulation. For example, while complete elimination or major decreases in lead exposure might avert the need for periodic testing of young children, an individual rulemaking may not have widespread or large enough effects to change the testing requirements, in which case related savings would not be expected. Regulations that address childhood lead exposures may affect their cardiovascular risks when they become adults; cardiovascular risks will be more immediately affected by a rule that addresses adult exposures. Careful comparison of baseline and postregulatory effects (steps 1 and 2 in Box 1) is needed to identify those outcomes that may be most important within the context of a particular analysis.

In addition, developing values for some outcomes included in the risk assessment may be difficult. Valuation requires information on discernable impacts, the effects of which can be weighed by individuals in determining how much they would be willing to pay to reduce related risks. For example, to value changes in red blood cell function, information on how these changes might ultimately affect an individual's activities and quality of life is needed. Similarly, valuing changes in blood pressure requires linking these changes to the need for treatment of hypertension and to the likelihood of stroke or other serious illnesses. Generally, in these cases, analysts must develop a model that estimates how these types of changes affect the likelihood of more significant health impairments as well as premature death.

Once data on the risk reductions attributable to a change in lead exposure are available, the effects can be valued using a number of different approaches. Scholars have assessed many lead-related effects that are not typically addressed in the USEPA's regulatory analyses, due in part to differences in the nature of the risk change and in part to concerns about the quality of the data.

Many of these analyses focus on elimination of lead-related hazards, rather than the smaller marginal changes likely to be associated with a federal regulation. For example, Landrigan et al.[41] assess the effects of eliminating all childhood asthma, cancer, and developmental disabilities (as well as lead poisoning) associated with environmental pollutants, using an averted cost approach that could be updated and adapted for use in other contexts. Another example is an analysis by Gould,[42] based in part on earlier work by Korfmacher, which explores the costs and benefits of eliminating the hazards that household lead paint poses to young children. She explores lead's effects on crime and attention deficit-hyperactivity disorder (ADHD) as well as other outcomes, using an averted cost approach. For example, for ADHD, Gould notes that the effects are lifelong and can include a range of types of misconduct and antisocial behavior, including drug use and crime. Considering only medical treatment and parental lost work time, she estimates that lead-associated ADHD costs $267 million annually (1996 dollars). She also estimates the effects of reducing average preschool blood lead levels on the incidence of burglaries, robberies, aggravated assaults, rapes, and murders. She concludes that the total direct cost of these crimes is approximately $1.8 billion, including victim costs, costs of legal proceedings and incarceration, and lost earnings for both the criminal and victim. Finally, she estimates that an additional $11.6 billion is lost in indirect costs, including psychological and physical damage that requires treatment and measures taken to prevent criminal actions. In addition to suggesting that the benefits associated with these large changes in exposure are significant, these and other studies indicate that data are available that can be used to estimate the value of additional benefit categories.

Conclusions

The types and magnitude of benefits from lead regulations are determined by the complex interactions of a number of factors, such as the characteristics and size of the exposed population, the magnitude of the averted exposures, the types and severities of the health effects, and the monetary values of the impacts. However, this entry indicates that the effects of lead on future earnings, and on premature mortality from cardiovascular disease, are the most significant benefit categories monetized in previous regulatory analyses. It further indicates that these analyses could be expanded to include numerous other potentially significant benefit categories.

For childhood exposures, the impact of lead on earnings (due to its effects on IQ and schooling) dominates the results. On a per-person basis, the changes in earnings are small. However, because USEPA regulations often affect a substantial number of individuals, the earnings-related benefits become large when aggregated nationally. More work is needed, however, to update and refine the estimates of the effect of a change in IQ on earnings, as well as to better understand the relationship between lost earnings and WTP for IQ changes.

For adult cardiovascular effects, premature mortality dominates the results, due in part to the fact that the value of mortality risk reductions is generally measured in millions of dollars while morbidity values are often at least an order of magnitude less. While the value of mortality risk reductions is well studied and current practices are evolving as a result of recent research and expert review, WTP estimates for many types of morbidity risks are not currently available.[20] Work currently underway (e.g., by Hammitt and Haninger[43] and Cameron and DeShazo.[44]) may eventually provide useful estimates of WTP for cardiovascular risks. More research on the lifetime costs of cardiovascular disease would also be informative.

Regulatory analyses currently do not include quantified and monetized estimates for several other benefit categories. The expansion of the epidemiological research base, represented by USEPA's recent *Integrated Science Assessment*, provides opportunities for better characterizing the relationships between lead and those outcomes traditionally included in the monetary valuation of benefits, as well as opportunities for expanding the analysis to include important additional outcomes. Some areas worthy of additional exploration include lead's non-cognitive effects on educational attainment and earnings and its effects on delinquent behavior and crime. Analysts will need to pay careful attention to the interrelationships between outcomes as the number of endpoints considered expands. Tracing these relationships is important both to avoid double counting and to ensure that the most significant benefit categories are included in the analysis to the greatest extent possible.

Acknowledgments

I would like to thank Scott D. Grosse (Centers for Disease Control and Prevention), James E. Neumann (Industrial Economics, Incorporated), and two anonymous reviewers for their helpful comments on earlier drafts of this entry.

References

1. Aspen Institute. *40 Years: EPA 40th Anniversary*; Aspen Institute: Washington, DC, 2010.
2. Robinson, L.A. *Benefits of Reduced Lead Exposure: A Review of Previous Studies.* Prepared for the U.S. Environmental Protection Agency under subcontract to Industrial Economics, Incorporated, 2007.
3. Lanphear, B.P.; Hornung, R.; Khoury, J.; Yolton, K.; Baghurst, P.; Bellinger, D.C.; Canfield, R.L.; Dietrich, K.N.; Bornschein, R.; Greene, T.; Rothenberg, S.J.; Needleman, H.L.; Schnaas, L.; Wasserman, G.; Graziano, J.; Roberts, R. Low-level environmental lead exposure and children's intellectual function: An international pooled analysis. Environ. Health Perspect **2005**, *113* (7), 894–899.

4. U.S. Environmental Protection Agency. Integrated Science Assessment for Lead. U.S. Environmental Protection Agency: Washington, DC, forthcoming. Available at http://www.epa.gov/ncea/isa/index.htm.
5. U.S. Environmental Protection Agency. *Costs and Benefits of Reducing Lead in Gasoline*; EPA 230–05–85–006. U.S. Environmental Protection Agency: Washington, DC, 1985.
6. *U.S. Office of Management and Budget. Report to Congress on the Costs and Benefits of Federal Regulations, 1997.*
7. Reagan, R. Executive Order 12291: Federal regulation. Fed. Regist. **1981**, *46* (190), 13193–13198.
8. Clinton, W.J. Executive Order 12866: Regulatory planning and review. Fed. Regist. **1993**, *58* (190), 51735–51744.
9. Obama, B. Executive Order 13563: Improving regulation and regulatory review. Fed. Regist. **2011**, *76* (14), 3821–3823.
10. U.S. Office of Management and Budget. *Circular A-4: Regulatory Analysis*; U.S. Office of Management and Budget: Washington, DC, 2003.
11. U.S. Office of Management and Budget. *Agency Checklist: Regulatory Impact Analysis*; U.S. Office of Management and Budget: Washington, DC, 2010.
12. Rice, D.P. Estimating the cost of illness. Am. J. Public Health **1967**, 57, 424–440.
13. Yabroff, K.R. et al., Eds. Health care costing: Data, methods, future directions. Med. Care **2009**, *47* (7), Sup 1.
14. Grosse, S.D.; Krueger, K.V. The income-based human capital valuation methods in public health economics used by forensic economics. J. Forensic Econ. **2011**, *22* (1), 43–57.
15. Grosse, S.D.; Krueger, K.V.; Mvundura, M. Economic productivity by age and sex: 2007 estimates for the United States. Med. Care **2009**, *47* (7), Sup. 1.
16. Zhang, W.; Bansback, N.; Anis, A.H. Measuring and valuing productivity loss due to poor health: A critical review. Soc. Sci. Med. **2011**, 72, 185–192.
17. Koopmanschap, M.A.; Rutten, F.F.H.; van Ineveld, B.M.; van Roijen, L. The friction cost method for measuring indirect costs of disease. J. Health Econ. **1995**, *4*, 171–189.
18. Agee, M.D.; Crocker, T.D. Parental altruism and child lead exposure: Inferences from demand for chelation therapy. J. Hum. Resour. **1996**, *31* (3), 677–691.
19. von Stackelberg, K.; Hammitt, J.K. Use of contingent valuation to elicit willingness-to-pay for the benefits of developmental health risk reductions. Environ. Resour. Econ. **2009**, *43*, 45–61.
20. Robinson, L.A.; Hammitt, J.K. Valuing health and longevity in regulatory analysis: Current issues and challenges (Ch. 30). In *Handbook on the Politics of Regulation*; Levi-Faur, D., Ed.; Edward Elgar: Cheltenham, U.K. and Northampton, 2011.
21. Rischitelli, G.; Nygren, P.; Bougatsos, C.; Freeman, M.; Helfand, M. Screening for elevated lead levels in childhood and pregnancy: An updated summary of evidence for the U.S. Preventive Services Task Force. Pediatrics **2006**, *118*, 1867–1895.
22. U.S. Environmental Protection Agency. *Reducing Lead in Drinking Water: A Benefit Analysis;* EPA 230–09–86-019. U.S. Environmental Protection Agency: Washington, DC, 1986 (revised 1987).
23. U.S. Environmental Protection Agency. *The Benefits and Costs of the Clean Air Act: 1970–1990*; EPA 410-R-97-002. U.S. Environmental Protection Agency: Washington, DC, 1997.
24. U.S. Environmental Protection Agency. *Economic Analysis for the TSCA Lead Renovation, Repair, and Painting Program Final Rule for Target Housing and Child-Occupied Facilities;* U.S. Environmental Protection Agency: Washington, DC, 2008.
25. U.S. Environmental Protection Agency. *Regulatory Impact Analysis of the Proposed Revisions to the National Ambient Air Quality Standards for Lead;* U.S. Environmental Protection Agency: Washington, DC, 2008.
26. Schwartz, J. Societal benefits of reducing lead exposure. Environ. Res. **1994**, *66*, 105–124.
27. Salkever, D.S. Updated estimates of earnings benefits from reduced exposure of children to environmental lead. Environ. Res. **1995**, *70*, 1–6.

28. Grosse, S.D.; Matte, T.D.; Schwartz, J.; Jackson, R.J. Economic gains resulting from reduction in children's exposure to lead in the United States. Environ. Health Perspect. 2002, *110* (6), 563–569.
29. Grosse, S.D. How much does IQ raise earnings? Implications for regulatory impact analyses. AERE Newsl. **2007**, *27* (2), 17–21.
30. Green, D.A.; Riddell, W.C. Literacy skills, non-cognitive skills, and earnings: An economist's perspective. In *Towards Evidence-Based Policy for Canadian Education*; de Broucker, P., Sweetman, A., Eds.; McGill-Queens University Press: Montreal and Kingston, 2002.
31. Neal, D.A.; Johnson, W.R. The role of premarket factors in black-white wage differences. J. Political Econ. **1996**, *104*, 869–895.
32. Heckman, J.J.; Stixrud, J.; Urzua, S. The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. J. Labor Econ. **2006**, *24*, 411–482.
33. Zax, J.S.; Rees, D.I. IQ, academic performance, environment, and earnings. Rev. Econ. Stat **2002**, *84*, 600–616.
34. U.S. Environmental Protection Agency. *Guidelines for Preparing Economic Analysis*; EPA 240-R-00–003. U.S. Environmental Protection Agency: Washington, DC, 2000.
35. U.S. Environmental Protection Agency (EPA). *Guidelines for Preparing Economic Analysis*; EPA 240-R-10–001.U.S. Environmental Protection Agency: Washington, DC, 2010.
36. Hammitt, J.K. Valuing mortality risk: Theory and practice. Environ. Sci. Technol. **2000**, *34*, 1396–1400.
37. Robinson, L.A. How US government agencies value mortality risk reductions. Rev. Environ. Econ. Policy. **2007**, *1* (2), 283–299.
38. U.S. Environmental Protection Agency. *Valuing Mortality Risk Reductions for Environmental Policy: A White Paper (Review Draft).* Prepared by the National Center for Environmental Economics for consultation with the Science Advisory Board-Environmental Economics Advisory Committee, U.S. Environmental Protection Agency: Washington, DC, 2010.
39. Kling, C. et al. *Review of Valuing Mortality Risk Reductions for Environmental Policy: A White Paper (December 10, 2010);* EPA-SAB-11-011. U.S. Environmental Protection Agency: Washington, DC, 2011.
40. U.S. Environmental Protection Agency. *Economic Analysis for the Renovation, Repair, and Painting Program Proposed Rule*; U.S. Environmental Protection Agency: Washington, DC, 2006.
41. Landrigan, P.J.; Schechter, C.B.; Lipton, J.M.; Fahs, M.C.; Schwartz, J. Environmental pollutants and disease in American children: Estimates of morbidity, mortality, and costs for lead poisoning, asthma, cancer, and developmental disabilities. Environ. Health Perspect. **2002**, *110* (7), 721–728.
42. Gould, E. Childhood lead poisoning: Conservative estimates of the social and economic benefits of lead hazard control. Environ. Health Perspect. **2009**, *117* (17), 1162–1167.
43. Hammitt, J.K.; Haninger, K. Valuing morbidity risk: Willingness to pay per quality-adjusted life year. Toulouse School of Economics, LERNA Working Paper Series, 2011.
44. Cameron, T.A.; DeShazo, J.R. Demand for health risk reductions. Unpublished manuscript, 2010.

27

Mercury

Sven Erik Jørgensen

Introduction

Mercury is an extremely toxic element. It has no role as biological element beyond its toxicity and is not essential for any organism. It is element number 80 in the periodic table and has an atomic weight of 200.59. Mercury caused one of the most significant environmental catastrophes in history. Mercury was discharged with the wastewater from a chemical factory in Minamata Bay, Japan, in the 1950s.[1] Mercury accumulated in the sediment, where it could react microbiologically with the organic matter and form methyl mercury ions and dimethyl mercury, which has a low boiling point and can be transferred from the sediment to the water phase or even to the atmosphere. Organic mercury compounds can be taken up by fish, where concentrations 3000 times higher than in water can be recorded. These processes explain why the fish caught in the Minamata Bay had a very high mercury concentration. As a result, hundreds of people died, and over a thousand became invalids due to mercury contamination. In the 1950s, the victims were considered to have a new, previously unknown disease, the Minamata disease. Later in the 1960s, it was detected that the disease was caused by high mercury concentration in fishermen's families. The accumulation in the sediment, the biomagnifications through the food chain (water–phytoplankton–zooplankton–fish–fisherman), and the ability of mercury to form organic compounds by microbiological reactions explain the emergence of Minamata disease and mercury poisoning.

Sources of Mercury Pollution

The use of mercury has declined during the last decades due to its extreme toxicity. Mercury compounds were previously applied as fungicides, as dyestuff (cinnabar red), and for the production of chlorine, but these applications are now banned in all industrialized countries. Chlorine and sodium hydroxide are unfortunately still produced in some developing countries by a method based on a dripping mercury electrode that causes mercury to be discharged with wastewater to the environment. Mercury still is applied by dentists and is used for some electrical instruments. Today's major mercury contamination is, however, caused by coal-fired power plants due to a small concentration of mercury in all coal. Each year, 3300 tons of mercury is discharged as air pollution to the atmosphere from coal-fired power plants and the incineration of solid waste.[2] The natural emission of mercury is about 1500 tons/yr, mainly from volcanic activity.

Pollution Effects

The most important acute effects of mercury are on the lungs and the central nervous system. The chronic effect is complete damage of the central nervous system, which was the main symptom of the Minamata fishermen. In addition to central nervous system symptoms, the victims also lose weight and appetite. The LD50 value for mice is 5 mg/kg, and the World Health Organization recommends a maximum concentration of 1 ppm in food items. (Mercury has further teratogenic and genetic effects. For details about all effects, see Jørgensen et al.[3–5])

Important Mercury Processes

Evaluation of mercury pollution requires that environmental management consider the environmental processes in which mercury can participate. It is clear from Section 1 that the formation of organic mercury compounds mainly in the form of methyl and dimethyl mercury is extremely important, and these processes should therefore be included in the environmental considerations, included when an ecotoxicological model for the distribution and effect of mercury is developed.

Figure 1 shows a conceptual diagram of mercury distribution and effects in Mex Bay, close to Alexandria. The details of the model can be found in Jørgensen and Bendoricchio.[6] From the figures, it is clear that the methylation processes are included and that the uptake from water and biomagnifications through the food chain are included to account for the most important state variable of the model: the mercury concentration in the top carnivorous fish (tuna fish), which is an important consumer fish in the Mediterranean region. Figure 2 shows the input/output processes that are included in the model.

The uptake of mercury by fish is highly dependent on the pH, due to the pH dependence of mercury solubility. This dependence is of course of interest mainly when the mercury pollution is inorganic. Figure 3 shows the pH dependence for the mercury in fish when exposed to 1.5 ppm inorganic mercury.[7]

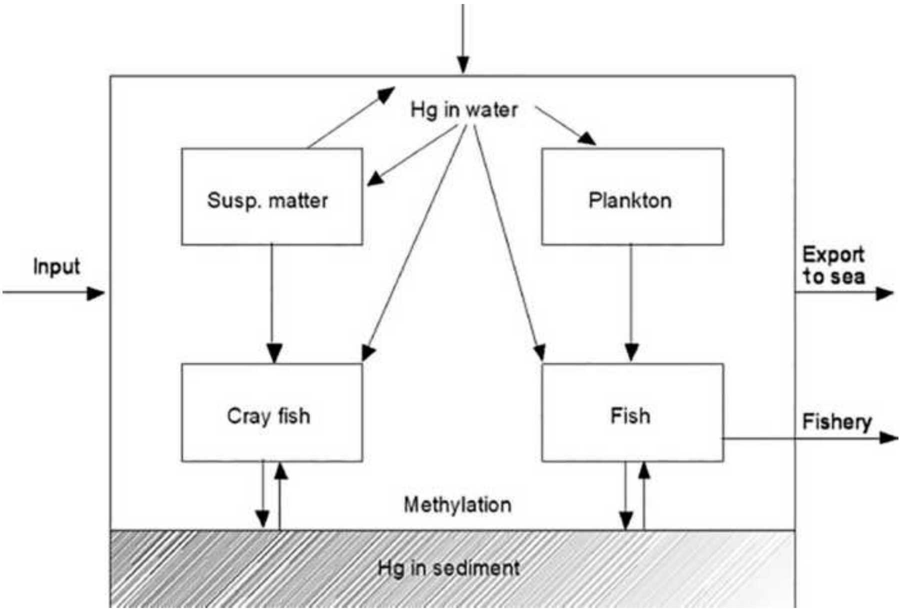


FIGURE 1 Conceptual diagram of the mercury model for Mex Bay, Egypt. Notice that methylation is considered for the release of mercury from the sediment. Notice also that the fish mercury is determined by uptake directly from the water (including the concentration of organic mercury) and through the food chain—here indicated only as water–phytoplankton–fish, but one to two additional steps could be considered.

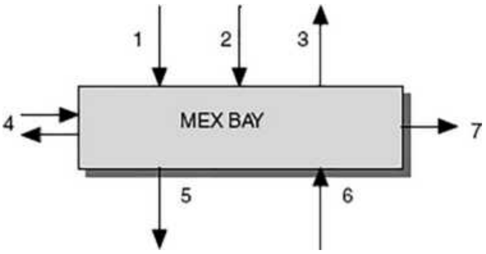


FIGURE 2 A model of the distribution of mercury in an aquatic ecosystem is a bio-geo-chemical model that must include all inputs and outputs. The numbers indicate the following: 1) discharge from waste and tributary to the bay; 2) deposition of mercury from the atmosphere (for instance, from coal-fired power plants); 3) evaporation of mercury; 4) input and output from the open sea; 5) sedimentation; 6) release from the sediment; and 7) fishery.

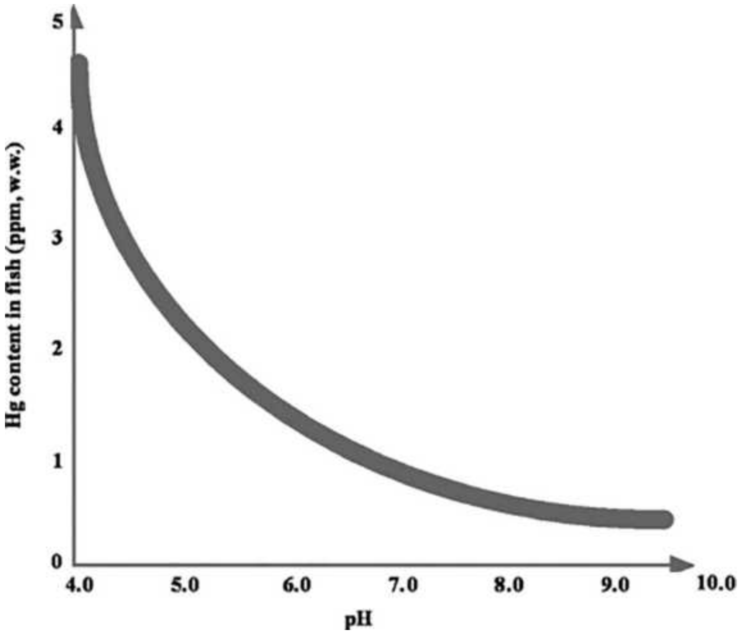


FIGURE 3 The concentration of mercury in fish expressed as ppm wet weight versus pH. The concentrations are measured at steady state and for a concentration of mercury chloride in the water of 1.5 ppm (mg/L). The values are based on many measurements for a wide spectrum of different fish species, and the concentrations in the fish are therefore indicated as a range.

The uptake of mercury from food is very different for organic mercury and inorganic mercury, namely, 90% and 20%, respectively.[7] This difference entails that a mercury model should follow and include the processes for both inorganic and organic mercury to be able to account for the final concentration, for instance, in fish as a result of a mercury contamination, whether the discharge is inorganic or organic mercury. The transfer between the two forms should of course be included, too.

Abatement of Mercury Pollution

The mercury pollution caused by coal-fired power plants is reduced for the part of the mercury adsorbed to particulate matter by the methods used for treatment of air pollutants to the extent that they are removing particulate matter. However, oil and natural gases are depleted within 60–80 years, which implies that a coal-fired power plant will still emit mercury if the power plant is using coal as fossil fuel. A significant part of mercury emission is not adsorbed to the particulate matter. The question is if we, due to emission of carbon dioxide, can continue to use coal-fired power plants after years 2050–2080. Today's emission of mercury is, as indicated above, 3300 tons/yr, and it is therefore a question of whether we can use coal-fired power plants after oil and natural gas have been depleted, not only due to the continuous emission of carbon dioxide but also due to the continuous emission of mercury, which is a consequence of the use of coal. This problem has not yet been examined in sufficient detail to give a clear answer—so this question is still open and requires further attention. It will probably be necessary, due to the greenhouse effects of carbon dioxide, to shift partially or completely to alternative energy sources before year 2080.

The mercury pollution causing wastewater problems can be solved by ion exchange—and ion exchangers with a particularly high efficiency of mercury removal are available. Furthermore, adsorption on activated carbon (see *Wastewater Treatment: Conventional Methods,* p. 2677) removes mercury, even organic mercury compounds, rather effectively. In this context, it should be mentioned that bioremediation methods are also able to reduce mercury concentration in contaminated soil (see *Bioremediation,* p. 408).

In conclusion, mercury pollution problems can be solved by today's technology, except for the emission of mercury from coal-fired power plants, which requires particular attention if the use of coal as an energy source continues to grow, unless the problem is solved by shifts to alternative energy. It is furthermore expected that general restrictions in the use of mercury compounds will be tightened in the years to come.

References

1. Newman, M.C.; Unger, M.A. *Fundamentals of Ecotoxicology,* 2nd Ed.; CRC Press: Boca Raton, London and New York, 2003.
2. *NESCAUM. Mercury Emissions from Coal-Fired Power Plants, The Case for Regulatory Action; 2003.*
3. Jørgensen, S.E.; Halling-Sprensen, B.; Mahler, H. *Handbook of Estimation Methods in Ecotoxicology and Environmental Chemistry*; Taylor and Francis Publ: Boca Raton, Boston, London, New York, Washington, D.C.; 1998. 230 pp.
4. Jørgensen, S.E.; Jørgensen L.A.; Nors Nielsen, S. *Handbook of Ecological and Ecotoxicological Parameters.* Elsevier: Amsterdam, 1991; 1380 pp.
5. Jørgensen, L.A.; Jørgensen, S.E.; Nors Nielsen, S. *Ecotox,* CD; Elsevier: Amsterdam, 2000; 4000 pp.
6. Jørgensen, S.E.; Bendoricchio, G. *Fundamentals of Ecological Modelling,* 3rd Ed.; Elsevier: Amsterdam; 530 pp.
7. Jørgensen, S.E. *Principles of Pollution Abatement;* Elsevier: Amsterdam, 2000; 520 pp.

28

Mycotoxins

J. David Miller

Introduction

The agriculturally important mycotoxins represent a major challenge for food and feed safety. They are unavoidable and no technology has been developed to make them disappear from our major food crops and none are on the horizon. This briefly reviews the major toxins and provides an insight into their major effects on human and animal health in fully developed market economies, the impact of fungicides and, in one case, pesticides on their reduction, and the impact on occupational health. This is one of the most intensively studied of any area of science because of the economic impact on western economies.[1,2] In addition, there is a massive toll in human suffering in parts of Africa, Latin America and East Asia.[3,4] At the time of writing, a PubMed search of the term "mycotoxins" resulted in 47,100 entries, which underscores the importance of the issue.

Five Important Mycotoxins

Although there are hundreds of fungal metabolites that are toxic in experimental systems, there are only five that are of major agricultural importance: deoxynivalenol, aflatoxin, fumonisin, zearalenone, and ochratoxin.[5,6] A number of other mycotoxins that can be important in some areas & in some years including ergot alkaloids, T-2/HT-2 toxins and some *Penicillium* toxins that can occur in animal feed.[7,8]

Deoxynivalenol occurs when wheat, barley, maize, and sometimes oats and rye are infected by *Fusarium graminearum, F. asiaticum* or *F. culmorum*. These species are the primary cause of Fusarium head blight in small grains. This disease is a major agricultural problem worldwide, as well as a similar disease in maize called Gibberella ear rot. Disease incidence is most affected by moisture at flowering, and many cultivars and hybrids used lack genetic tolerance to the disease.[9–11] *F. graminearum* is common in wheat from North America and China, where *F. asiaticum* is also important. Thirty years ago, *F. culmorum* had been the dominant species in cooler wheat-growing areas, such as Finland, France, Poland, and the Netherlands, but this trend has changed in recent years as European summers have reached record warm temperatures such that *F. graminearum* largely dominates.[9,12,13] Fusarium head blight is caused by *F. asiaticum* in warmer regions.[14]

Within these taxa, there are a number of chemotypes. Most strains fall into one of two chemoyptes, one producing deoxynivalenol by a pathway involving 3-acetyl deoxynivalenol and the other by

15-acetyl deoxynivalenol. Some strains of *F. graminearum* and *F. asiaticum* produce nivalenol.[15] Some strains of *F. graminearum* produce a trichothecene reported in 2015, NX.[16] Finally, some strains of *F. graminearum* produce both deoxynivalenol and NX.[17]

Humans are not less and are probably more sensitive to deoxynivalenol than swine.[7] Swine are the most sensitive domestic animal species to the effects of deoxynivalenol. Several mechanisms account for the toxicity of deoxynivalenol including affecting different aspects of the appetite suppression, modulation of the insulin-like growth factor acid-labile subunit and alters brain chemistry.[15,18] In addition, deoxynivalenol causes changes in immune system function in male mice, occurring at dietary concentrations often encountered by humans. As with other trichothecenes, high exposures increase susceptibility to facultative pathogens such as *Listeria*.[19] Cattle, cows, and poultry species are tolerant to deoxynivalenol and milk production is not affected at typical field concentrations.[20] The Provisional Maximum Tolerable Daily Intake (PMTDI) of the Joint Expert Committee on Food Additives and Contaminants (JECFA) of the World Health Organization (WHO) and Food and Agriculture Organization (FAO) for this toxin was revised in 2010 by making a group PMTDI for deoxynivalenol and the two acetylated precursors. These can co-occur with the parent compound in grains.[21]

Crops that are contaminated by deoxynivalenol can often contain zearalenone albeit at a lower frequency. Zearalenone is more common in maize than in small grains. Zearalenone is an estrogen analogue and causes hyperestrocism in female pigs at low levels; the dietary no-effect level is less than 1 mg/kg. Cows and sheep are also sensitive to the estrogenic effects of this toxin with depressed ovulation and lower lambing percentages.[20] Non-human primates and possibly humans are also sensitive to the estrogenic effects of zearalenone.[22,23]

Aflatoxin is mainly produced by *Aspergillus flavus,* which is a problem in many commodities, but most human exposure comes from contaminated maize, groundnuts, and rice. *Aspergillus parasiticus* is uncommon outside North and South America, and they are more associated with peanuts. There are also some less common species that produce aflatoxin. *A. flavus* contamination of maize or peanuts occurs in two basic ways. Either airborne or insect-transmitted conidia contaminate the silks and grow into the ear when the maize is under high-temperature stress or (more commonly) insect- or bird-damaged kernels become colonized with the fungus and accumulate aflatoxin. In maize and peanut plants, drought-, nutrient-, or temperature-stressed plants are more susceptible to colonization by *A. flavus* or *A. parasiticus*.[5,24]

In poultry, aflatoxin exposure results in liver damage, impaired productivity and reproductive efficiency, decreased egg production in hens, inferior egg-shell quality, inferior carcass quality, and increased susceptibility to disease. The effects of acute and chronic exposure in swine are largely attributable to liver damage. In cattle, the primary symptoms are reduced weight gain, liver and kidney damage, and reduced milk production.[5] Aflatoxin is also immunotoxic in domestic and laboratory animals with oral exposures in the microgram-per-gram range. Cell-mediated immunity (lymphocytes, phagocytes, mast cells, and basophils) is more affected than humoral immunity (antibodies and complement).[19] Naturally-occurring mixtures of aflatoxins including aflatoxin B1 are human liver carcinogens. In much of Africa as well as parts of Latin America and East Asia, aflatoxin exposure can be multiples, sometimes orders of magnitude above tolerable levels. Aside from increased rates of liver cancer, aflatoxin is associated with child stunting.[3,24-26]

Fumonisins are produced by *Fusarium verticillioides, Fusarium proliferatum,* and a number of less common fusaria. Fumonisins are a very common contaminant of maize-based food and feed in the United States, China, Europe, southern Africa, South America, and Southeast Asia.[27,28] *F. verticillioides* and *F. proliferatum* can be recovered from virtually all maize kernels including those that are healthy, which suggests that it may be an endophyte, i.e., a mutualistic relationship.[5,28,29] *F. verticillioides* and *F. proliferatum* cause a "disease" called Fusarium kernel rot. In parts of the United States and lowland tropics, this is one of the most important ear diseases and is associated with warm, dry years and insect damage. Maize plant disease stress also promotes the growth of *F. verticillioides* and fumonisin formation.[28]

Fumonisin causes equine leukoencephalomalacia (ELEM), which involves a liquefactive necrosis of the cerebral hemisphere. Clinical manifestations include abnormal movements, aimless circling, lameness, etc., followed by death. In swine, high exposures of fumonisin results in porcine pulmonary edema (PPE) caused by fumonisin-induced heart failure.[20] At lower exposures, both liver damage and kidney damage have been reported in swine. Fumonisin causes feed refusal and changes in carcass quality at dietary concentrations in the low milligram-per-kilogram range.[20,30] Fumonisin exposure results in impaired growth rates in a variety of domestic and laboratory animals.[31] Fumonisin exposure results in neural tube birth (NTD) defects in rodents.[32,33] The unifying mechanism for these toxicities is that fumonisin interferes with sphingolipid metabolism unleashing a cascade of harmful cellular effects.[34]

Exposure to *F. verticillioides*-contaminated maize has been linked to the elevated rates of esophageal cancer in the Transkei and this has since been linked to fumonisin exposure.[30] Fumonisin caused tumors in mice and rats in the U.S. National Toxicology Program study and is an IARC class 2B carcinogen (possible human carcinogen).[27] People exposed to dietary fumonisin above the Tolerable Daily Intake demonstrate fumonisin-induced sphigolipid alteration.[32,33] As with aflatoxin, exposure to fumonisin in Africa has been associated with child stunting.[27,31]

Human exposure to ochratoxin A in temperate zones mainly comes from eating foods made from contaminated small grains.[35] Ochratoxin production in small grains results from the growth of *Penicillium verucosum* typically in small pockets of grain that get wet or go in the bin wet.[36] A small number of highly contaminated kernels in a portion of cereal (~1/5000) can put the lot over the regulatory limit in the EU.[37] Ochratoxin exposure can arise from *Aspergillus carbonarius* and some strains of *A. niger*, and several related species also produce ochratoxin on grapes and coffee.[38] In Europe, minor exposures occur from meat, especially pork, from animals fed contaminated grain and from the consumption of moldy sausage, etc. resulting from the growth of *P. nordicum*.[39]

Ochratoxin is a potent nephrotoxin in swine and causes kidney cancer in male Fisher 344 rats. Pigs appear to be the most sensitive species with respect to the nephrotoxic effects.[20,40] Low exposures result in kidney damage in swine but, typically there are no overt signs. At higher concentrations (>2 μg/g), decreased weight gains occur. Poultry are affected, showing reduced growth rate and egg production at low ochratoxin concentrations >2 μg/g. Cattle are resistant to ochratoxin concentrations found in naturally contaminated grain.[20]

At the expense of competing hypotheses, ochratoxin was suspected as the cause of Balkan Endemic Nephropathy.[6] This a syndrome manifesting as late life urinary tract cancers and kidney damage in areas where ochratoxin exposure was believed to be high in parts of Eastern Europe.[5,6,40] In recent years, a compelling case for human disease seen has been made for exposure to the human carcinogen aristolochic acid. Dietary exposure occurs from *Aristolochia* seeds contaminating harvested wheat and from crops that take up aristolochic acids from the soil[41–43] As noted by JECFA,[40] evidence for the carcinogenicity of ochratoxin remains confined to laboratory strains of mice and rats. Mechanisms that explain the rodent carcinogenicity have been proposed unrelated to the production of ochratoxin-DNA adducts.[40,44]

Mycotoxins and Pesticides

Insect damage is known to promote accumulation of deoxynivalenol, fumonisin, aflatoxin, and probably zearalenone. Steps taken to reduce insect herbivory or drought stress on crops have the general effect of reducing mycotoxin contamination. Bt corn has lowered fumonisin content compared to similar non-Bt hybrids.[45–47] The situation with respect to lower aflatoxin and Bt is not clear[48] Potential reductions in deoxynivalenol in Bt hybrids depends on the degree of insect damage.[46,49]

Modeling of mycotoxin accumulations in small grains based on weather at anthesis is a part of the risk management strategy for Fusarium head blight. If the weather is conducive to disease development, farmers are able to make informed decisions about fungicide application[11,50] This has the effect of restricting the prophylactic use of fungicides when not warranted. This is an important area of research for food protection and food security.[51]

Occupational Exposures

Grain dusts represent an occupational hazard for a number of reasons including the possibility of the allergic disease hypersensitivity pneumonitis and organic dust toxic syndrome, a poorly understood disease. Inhalation of mycotoxins contained in airborne dusts is also a potential health risk. Inhalation of mycotoxins in spores and dust affects macrophage function and other aspects of lung biology. Inhalation of mycotoxins is a more potent route of exposure for systemic toxicity for aflatoxin, trichothecenes, and ochratoxin. Workplace exposure to airborne aflatoxin results in increased relative risk of liver cancer. However, the practical concern is the impact that fungal glucan and the aflatoxin has on the function of lung macrophages.[52–54]

Risk characterizations of inhalation exposure to grains containing high concentrations of deoxynivalenol exposures indicate that aside from the allergic and inflammatory diseases resulting from exposure to grains, the presence of dexoynivalenol is a concern.[54] Urinary biomarkers for deoxynivalenol (DON, DOM-1) were higher in active farmers in France, particularly from larger farms, than from retired farmers where exposure was from the diet.[55] Workplace inhalation exposure-outcome data for other toxins including zearalenone, fumonisin and ochratoxin are limited.[54] In modern grain handling facilities, dust control is a priority not only for workplace safety but also to prevent grain dust explosions. Thus, inhalation exposure to mycotoxins is likely now a minor problem but remains a potential concern in informal grain handing systems.

Conclusion

As noted, mycotoxin contamination of field crops is a problem that cannot be “solved” in the normal use of that term, results in important losses of human life, and has major economic and social consequences. Recognition that fungi produced toxins in crops began with ergot in Europe during the Middle Ages and has been recognized as low-molecular-weight chemicals since the 1880s and solving their structures and understanding appropriate management strategies took agonizingly long periods of study.[5,6] The current strategies for managing mycotoxins in the diet require modeling, crop breeding, better detection systems, and awareness.

References

1. Mitchell, N. J., Bowers, E., Hurburgh, C., Wu, F. Potential economic losses to the US corn industry from aflatoxin contamination. Food Add. Contam. Part A **2016**, *33,* 540–550.
2. Wilson, W., Dahl, B., Nganje, W. Economic costs of fusarium head blight, scab and deoxynivalenol. World Mycotoxin J **2018**, *11*, 291–302.
3. Wild, C.P.; Gong, Y.Y. Mycotoxins and human disease: A largely ignored global health issue. Carcinogenesis **2010**, *31,* 71–82.
4. Logrieco, A., Miller, J.D., Eskola, M., Krska, R., Ayalew, A., Bandyopadhyay, R., & Li, P. The mycotoxin charter: increasing awareness of, and concerted action for, minimizing mycotoxin exposure worldwide. Toxins **2018**, *10*, 149.
5. Miller, J.D. Fungi and mycotoxins in grain: Implications for stored product research. J. Stored Prod. Res. **1995**, *31,* 1–6.
6. Pitt, J. I., Miller, J. D. A concise history of mycotoxin research. J. Agric. Food Chem. **2016**, *65,* 7021–7033.
7. Miller, J.D. Mycotoxins in small grains and maize: Old problems, new challenges. Food Addit. Contam. **2008**, *25,* 219–230.
8. Miller, J.D. Changing patterns of fungal toxins in crops: challenges for analysts. J AOAC Int. **2016**, *99*, 837–841.

9. Miller, J.D. Epidemiology of *Fusarium* ear diseases. In *Mycotoxins in Grain: Compounds Other Than Aflatoxin*; Miller, J.D., Trenholm, H.L., Eds.; Eagan Press: St. Paul, MN, 1994; 19–36.
10. Boutigny, A.L. Richard-Forget, F. Barreau, C. Natural mechanisms for cereal resistance to the accumulation of *Fusarium* trichothecenes. Eur. J. Plant Pathol. **2008**, *121,* 411–423.
11. Bianchini, A., Horsley, R., Jack, M. M., Kobielush, B., Ryu, D., Tittlemier, S., Miller, J. D. DON occurrence in grains: A North American perspective. Cereal Foods World **2015**, *60,* 32–56.
12. Pasquali, M., Beyer, M., Logrieco, A., Audenaert, K., Balmas, V., Basler, R., González-Jaén, M. T. A European database of *Fusarium graminearum* and *F. culmorum* trichothecene genotypes. Frontiers in Microbiology **2016**, *7,* 406.
13. Xu, X.M. Nicholson, P. Ecology of fungal pathogens causing wheat head blight. Annu. Rev. Phytopathol. **2009**, *47,* 83–103.
14. Backhouse, D. Global distribution of *Fusarium graminearum*, *F. asiaticum* and *F. boothii* from wheat in relation to climate. European J. Plant Pathol. **2014**, *139,* 161–173.
15. Miller, J. D. Mycotoxins in Food and Feed: A Challenge for the Twenty-First Century. In *Biology of Microfungi*; Li, D., Ed.; Springer, Basel, 2016; 469–493.
16. Aitken, A., Miller, J. D., McMullin, D. R. Isolation, chemical characterization and hydrolysis of the trichothecene 7α-hydroxy, 15-deacetylcalonectrin (3ANX) from *Fusarium graminearum* DAOMC 242077. Tetrahedron Letters **2019**, *60*, 852–856.
17. Crippin, T., Renaud JB, Sumarah MW, Miller JD (2019) Comparing genotype and chemotype of *Fusarium graminearum* from cereals in Ontario, Canada. PLoS ONE **2019** *14(5)*, e0216735
18. Pestka, J. J. Deoxynivalenol: mechanisms of action, human exposure, and toxicological relevance. Archives of Toxicology **2010**, *84*, 663–679.
19. Bondy, G.S. Pestka, J.J. Immunomodulation by fungal toxins. J. Toxicol. Environ. Health, Part B **2000**, *3*, 109–143.
20. Prelusky, D.B. Rotter, B.A.; Rotter, R.G. Toxicology of mycotoxins. In *Mycotoxins in Grain: Compounds Other Than Aflatoxin*; Miller, J.D., Trenholm, H.L., Eds.; Eagan Press: St. Paul, MN, 1994; 359–404.
21. Bulder, A. S., DiNovi, M., Kpodo, K. A., Leblanc, J. C., Resnik, S., Shephard, G. S., Wolterink, G. Deoxynivalenol (addendum). WHO Food Additives Series: 63 FAO JECFA Monographs **2011**, *8*, 317.
22. Kuiper-Goodman, T., Scott, P., & Watanabe, H. Risk assessment of the mycotoxin zearalenone. Reg. Tox. Pharm. **1987**, *7*, 253–306.
23. EFSA Panel on Contaminants in the Food Chain. Scientific Opinion on the risks for public health related to the presence of zearalenone in food. EFSA Journal **2011**, *9*, 2197.
24. Doerge DR, Shephard GS, Adegoke GO, Benford D, Bhatnagar D, Bolger M, Boon PE, Cressey P, Edwards S, Hambridge, Miller JD, Mitchell RT, Riley RT, Wheeler MW Aflatoxins. Safety evaluation of certain contaminants in food: prepared by the eighty-third meeting of the Joint FAO/WHO Expert Committee on Food Additives (JECFA). WHO Food Additives Series, **2018**, *74*, 3–279.
25. Wild, C., Miller J.D., Groopman, J.D. Mycotoxin control in low and middle income countries. IARC Working Group Report #9, International Agency for Research on Cancer, Lyon, France. ISBN 978–92–832–2510–2 **2015**
26. McMillan, A., Renaud, J.D., Burgess, K.M.N., Orimadegun, A.E., Akinyinka, O.O., Allen, S.H., Miller, J.D., Reid, G., Sumarah, M.W. Aflatoxin exposure in Nigerian children with severe acute malnutrition. Food Chem. Tox. **2018**, *111,* 356–362.
27. Riley, R.T., Edwards, S.G., Aidoo, K., Alexander, J., Bolger, M., Boon, P.E., Cressey, P.E., Doerge, D.R., Edler, L., Miller, J.D., Zhang, Y. Fumonisins. Safety evaluation of certain contaminants in food: prepared by the eighty-third meeting of the Joint FAO/WHO Expert Committee on Food Additives (JECFA). WHO Food Additives Series, **2018**, *74,* 415–571.
28. Miller, J.D. Factors that affect the occurrence of fumonisin. Environ. Health Perspect. **2000**, *109 (s.2),* 321–324.

29. Yates, I.E., Sparks, D. *Fusarium verticillioides* dissemination among maize ears of field-grown plants. Crop Prot. **2008**, *27*, 606–613.
30. Marasas, W.F.O.; Miller, J.D.; Riley, R.T.; Visconti, A. Fumonisins B1. Environmental Health Criteria 219; International Program for Chemical Safety, World Health Organization: Geneva, 2000.
31. Chen, C., Riley, R. T., Wu, F. Dietary fumonisin and growth impairment in children and animals: A review. Comp. Rev. Food Science Food Safety **2018**, *17*, 1448–1464
32. Riley, R.T., Torres, O., Matute, J., Gregory, S.G., Ashley-Koch, A.E., Showker, J.L., Gelineau-van Waes, J.B. Evidence for fumonisin inhibition of ceramide synthase in humans consuming maize-based foods and living in high exposure communities in Guatemala. Mol. Nutr. Food Res. **2015**, *59*, 2209–2224.
33. Voss, K.A., Riley, R.T., Gardner, N.M., Gelineau-van Waes, J. Fumonisins. In: *Reproductive Developmental Toxicology*, 2nd edition, Gupta, R.C., Ed.; Academic Press, New York; 2017, 925–943.
34. Voss, K.A., & Riley, R.T. Fumonisin toxicity and mechanism of action: overview and current perspectives. Food Safety **2013**, *1*(1), 2013006-2013006.
35. Lee, H. J., Ryu, D. Worldwide occurrence of mycotoxins in cereals and cereal-derived food products: public health perspectives of their co-occurrence. J Agric Food Chem. **2017**, *65*, 7034–7051.
36. Limay-Rios, V., Miller, J.D., Schaafsma, A.W. Occurrence of *Penicillium verrucosum*, ochratoxin A ochratoxin B and citrinin in on-farm stored winter wheat from the Canadian Great Lakes Region. PLoS One **2017**, *12,* e0181239.
37. Whitaker, T. B., Slate, A. B., Nowicki, T. W., Giesbrecht, F. G. Variability and distribution among sample test results when sampling unprocessed oat lots for ochratoxin A. World Mycotoxin J. **2015** *8*, 511–524.
38. Cabañes, F. J., Bragulat, M. R. Black aspergilli and ochratoxin A-producing species in foods. Current Opinion Food Science **2018**, *23*, 1–10.
39. Magistà, D., Susca, A., Ferrara, M., Logrieco, A. F., & Perrone, G. Penicillium species: crossroad between quality and safety of cured meat production. Current Opinion Food Sci. **2017** *17*, 36–40.
40. Benford, D., Boyle, C., Dekant, W., Fuchs, R., Gaylor, D. W., Hard, G., Solfrizzo, M. Ochratoxin A. In: Safety evaluations of certain mycotoxins in food. WHO Food Additives Series, WHO, Geneva; **2001**, 47, 281–387
41. Bui-Klimke, T., Wu, F. Evaluating weight of evidence in the mystery of Balkan endemic nephropathy. Risk Analysis **2014**, *34*, 1688–1705.
42. Stiborová, M., Arlt, V. M., Schmeiser, H. H. Balkan endemic nephropathy: an update on its aetiology. Archives Toxicology **2016**, *90*, 2595–2615.
43. Li, W., Chan, C. K., Liu, Y., Yao, J., Mitić, B., Kostić, E. N., Dedon, P. C. Aristolochic acids as persistent soil pollutants: Determination of risk for human exposure and nephropathy from plant uptake. J. Agric. Food Chem. **2018**, *66*, 11468–11476.
44. Mally, A. Dekant, W. Mycotoxins and the kidney: Modes of action for renal tumor formation by ochratoxin A in rodents. Mol. Nutr. Food Res. **2009**, *53,* 467–478.
45. De La Campa, R., Hooker, D.C., Miller, J.D., Schaafsma, W.A., Hammond, B.G Modelling effects of environment, insect damage and BT genotypes on fumonisin accumulation in maize in Argentina and the Philippines. Mycopathologia **2005**, *159,* 539–552.
46. Folcher, L., Delos, M., Marengue, E., Jarry, M., Weissenberger, A., Eychenne, N., Regnault-Roger, C. Lower mycotoxin levels in Bt maize grain. Agronomy Sust. Dev. **2010**, *30*, 711–719.
47. Bowers, E., Hellmich, R., Munkvold, G. Comparison of fumonisin contamination using HPLC and ELISA methods in Bt and near-isogenic maize hybrids infested with European corn borer or Western bean cutworm. J. Agric. Food Chem. **2010**, *62*, 6463–6472.

48. Abbas, H. K., Zablotowicz, R. M., Weaver, M. A., Shier, W. T., Bruns, H. A., Bellaloui, N., Abel, C. A. Implications of Bt traits on mycotoxin contamination in maize: overview and recent experimental results in Southern United States. J. Agric. Food Chem. **2013**, *61*, 11759–11770.
49. Schaafsma, A. W., Hooker, D. C., Baute, T. S., Illincic-Tamburic, L. Effect of Bt-corn hybrids on deoxynivalenol content in grain at harvest. Plant Disease **2002**, *86*, 1123–1126.
50. Schaafsma, A.W., Hooker, D.C. Climatic models to predict occurrence of *Fusarium* toxins in wheat and maize. Int. J. Food Microbiol. **2007**, *119,* 116–125.
51. Van der Fels-Klerx, H. J., C. Liu, P. Battilani, P. Modelling climate change impacts on mycotoxin contamination. World Mycotoxin J. **2016**, *9,* 717–726.
52. Miller, J.D. Mycotoxins. In: *Handbook of Organic Dusts*; Rylander, R., Pettersen, Y., Eds.; CRC Press: Boca Raton, FL, 1994; 87–92.
53. Liao, C-M.; Chen, S-C. A probabilistic modeling approach to assess human inhalation exposure risks to airborne aflatoxin B1 (AFB1). Atmos. Environ. **2005**, *39,* 6481–6490.
54. Pitt, J.I., Wild, C.P., Baan, R.A., Gelderblom, W.C.A., Miller, J.D., Riley, R.T., Wu, F. Improving public health through mycotoxin control. International Agency for Research on Cancer Scientific Publications Series, No. 158. Lyon, France; 2012.
55. Turner, P.C., Hopton, R.P., Lecluse, Y., White, K.L., Fisher, J., Lebailly, P. Determinants of urinary deoxynivalenol and de-epoxy deoxynivalenol in male farmers from Normandy, France. J. Agric. Food Chem. **2010**, *58,* 5206–5212.

29

Nitrogen

Oswald Van Cleemput and Pascal Boeckx

Introduction

Nitrogen (N) is essential to all life. It is the nutrient that most often limits biological activity. In agricultural and natural ecosystems, N occurs in many forms covering a range of valence states from –3 to +5. The change from one valence state to another depends primarily on environmental conditions. The transformations and flow from one form to another constitute the basics of the soil N cycle (Figure 1). The use of N fertilizers has become essential to increase the productivity of agriculture, and has resulted in an almost doubling of the global food production in the past 50 years. However, this also implies that the natural N cycle has substantially been disturbed. In the following paragraphs an overview of the different N transformation processes in the soil is given.

Nitrogen Cycle: General

Atmospheric N_2 gas (valence 0) can be converted by lightening to various oxides and finally to nitrate (NO_3^-) (valence +5), which can be deposited and taken up by growing plants. Also N_2 gas can be converted to ammonia (NH3, valence –3) by biological N_2 fixation, with the NH_3 participating in a number of biochemical reactions in the plant. When plant residues decompose the N-compounds undergo a series of microbial conversions (mineralisation) leading first to the formation of ammonium (NH_4^+) (valence –3) and possibly ending up in NO_3^- (nitrification). Under anaerobic conditions NO_3^- can be converted to various N-oxides and finally to N_2 gas (denitrification). When mineral or organic N fertilizers are used they also undergo the same transformation processes and influence the rate of other N-transformations. In considering the soil compartment, there can be N gains (such as biological N_2 fixation) as well as N losses (such as leaching and denitrification). Furthermore N can be exported from the soil via harvest products, or immobilized in soil organic matter.

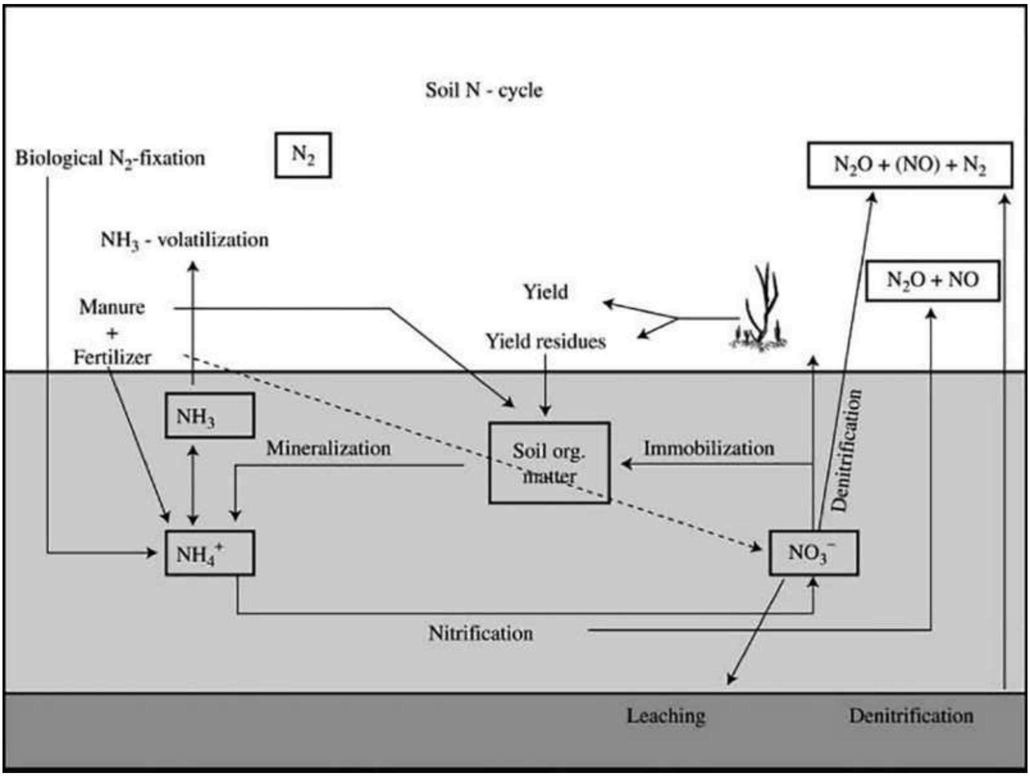


FIGURE 1 The soil N cycle. The white compartment represents the atmosphere; the light gray compartment represents the biosphere and the dark gray compartment the subsoil.

Nitrogen Transformations in the Soil

The principal forms of N in the soil are NH_4^+, NO_3^- or organic N-substances. At any moment, inorganic N in the soil is only a small fraction of the total soil N. Most of the N in a surface soil is present as organic N. It consists of proteins (20%–40%), amino sugars, such as the hexosamines (5%–10%), purine and pyrimidime derivates (1% or less), and complex unidentified compounds formed by reaction of NH_4^+ with lignin, polymerization of quinones with N compounds and condensation of sugars and amines. In the subsoil, an important fraction of the present N can be trapped in clay lattices (especially *illitic* clays) as nonexchangeable NH_4^+ and is consequently largely unavailable. Organic substances slowly mineralize by microorganisms to NH_4^+, which could be converted by other microorganisms to NO_3^-(see further).

The NH_4^+ can be adsorbed to negatively charged sites of clay minerals and organic compounds. This reduces its mobility in the soil compared to the more mobile NO_3^- ion. Microorganisms can use both NH_4^+ and NO_3^- to satisfy their need for N. This type of N transformation is called microbial immobilization.

The ratio between carbon (C) and N (C : N ratio) in organic matter determines whether immobilization or mineralization is likely to occur. When utilizing organic matter with a low N content, the microorganisms need additional N, decreasing the mineral N pool of the soil. Thus, incorporation of organic matter with a high C:N ratio (e.g., cereal straw) results in immobilization. Incorporation of organic matter with a low C : N ratio (e.g., vegetable or legume residues) results in N-mineralization. A value of the C:N ratio of 25 to 30 is often taken as the critical point toward either immobilization or mineralization.

Nitrification is a two-step process. In the first step NH_4^+ is converted to nitrite (NO_2^-) (valence +3) by a group of obligate autotrophic bacteria known as *Nitrosomonas* species. The second step is carried out by another group of obligate autotrophic bacteria known as *Nitrobacter* species. Also a few heterotrophs can carry out nitrification, usually at much lower rates.

Soil water and aeration are crucial factors for nitrification. At a water potential of 0 kPa (saturation), there is little air in the soil and nitrification stops, due to oxygen limitation; nitrification is greatest near field capacity (–33 kPa in medium- to heavy-textured soils, to 0 to –10 kPa in light sandy soils). Also in dry soils NH_4^+ and sometimes NO_2^- accumulate presumably because *Nitrobacter* species are more sensitive to water stress than the other microorganisms.

Nitrification is slow in acid conditions with an increasing rate at increasing pH. Mainly under alkaline conditions, nitrite is also accumulating, because *Nitrobacter* is known to be inhibited by ammonia, which is formed under alkaline conditions. Nitrification is a process that acidifies the soil as protons (H^+) are liberated:

$$NH_4^+ + 2O_2 \rightarrow NO_3^- + 2H^+ + H_2O$$

During nitrification minor amounts of nitrous oxide (N_2O) (valence +1) and nitric oxide (NO) (valence +2) are formed. Both compounds have environmental consequences, discussed below.

The effect of temperature on nitrification is climate dependent. There is a climatic selection of species of nitrifiers, with those from cooler regions having lower temperature optima and less heat tolerance than species from warmer regions. All above-mentioned factors influencing nitrification also influence the nitrifying population. The population and activity of nitrifiers can be reduced by the use of nitrification inhibitors, such as dicyanodiamide, nitrapirin and neem (*Azadirachta indica*) seed cake. They are used mostly to retard the nitrification of manure; otherwise their practicality is controversial and they are not extensively used. More details about nitrification and nitrification inhibitors can be found in McCarty[1] and Prosser.[2]

Nitrogen Input Processes

Atmospheric Nitrogen Deposition

The total atmospheric N (NH_4^+ and NO_3^-) deposition is in the order of 10–40 kg N ha^{-1} yr^{-1} in much of north-western and central Europe and some regions in North America. It ranges from 3–5 kg N ha^{-1} yr^{-1} in pristine areas.[3] It is originating from previously emitted NH_3 and NO_x from agricultural and industrial activities or traffic.

Biological Nitrogen Fixation

Rhizobium species living in symbiotic relationship in root nodules of legumes, e.g. clover (*Trifolium*), lucerne (*Medicago*), peas (*Pisum*) and beans (*Faba*)—can convert atmospheric N_2 gas to NH_3, which is further converted to amino acids and proteins. Parallel to this process, the *rhizobium* species receive from the legume the energy they need to grow and to fix N_2. Photosynthetic cyanobacteria are also N-fixing organisms and are especially important in paddy rice (*Oryza*). The amount of N fixed varies greatly from crop to crop, ranging from a few kg to a few hundred kg N ha^{-1} yr^{-1}. The process is depressed by ample N supply from other sources, and it is sensitive to lack of phosphorus. The amount of globally fixed N is almost the double the amount of applied fertilizer N. Next to symbiotic N fixing bacteria also non-symbiotic species (e.g. *Azotobacter*) occur in soils. In general, free-living diazotrophs make a small but significant contribution to the soil N status. Some nonleguminous trees and plants (e.g. alder (*Alnus*), sugarcane (*Saccharum*) host N-fixing bacteria as well. Much uncertainty exists about the association of N fixing bacteria with non-legumes (so called associative N fixing bacteria).

Mineral and Organic Nitrogen Fertilization

Theoretically plants should prefer NH_4^+ above NO_3^-, because NH_4^+ does not need to be reduced before incorporation into the plant. In most well-drained soils oxidation of NH_4^+ is fairly rapid and therefore most plants have developed to grow better with NO_3^-. However, a number of studies have shown that plants better develop when both sources are available. Rice, growing under submerged conditions must grow in the presence of NH_4^+ as NO_3^- is not stable under flooded conditions. When urea is applied it rapidly hydrolyzes under well-drained conditions, unless a urease inhibitor is being added; under submerged conditions rice plants may also absorb N directly as molecular urea. Organic manure can be of plant or animal origin or a mixture of both. However, most comes from dung and urine from farm animals. It exists as farmyard or stable manure, urine, slurry or as compost. Because its composition is not constant and because plant material (catch or cover crops, legumes) is often added freshly (green manure) to the soil, less than 30% of its nutrients becomes available for the next crop.

Nitrogen Uptake by Plants

Growing plants get their N from fertilizer N as well as from organic soil N upon mineralization. Plants take up N compounds both as NO_3^- and as NH_4^+. In general, NO_3^- is the major source of plant N. There is some evidence that small amounts of organic N (urea or amino acids) can be taken up by plants from the soils solution. Plant uptake of N can be studied through the use of mineral fertilizers or organic matter labeled with the stable N isotope^{15}N. The proportion of applied N taken up by the crop is affected by many factors, including crop species, climate and soil conditions. Above ground parts of the crop can recover 40%–60% of the fertilizer N applied.

Nitrogen Loss Processes

Ammonia Volatilization

Losses of N from the soil by NH_3 volatilization amount globally to 54 Mt (or 10^{12} g) NH_3^-N yr^{-1} and 75% is of anthropogenic origin.[4] According to the ECETOC,[5] the dominant source is animal manure and about 30% of N in urine and dung is lost as NH_3. The other major source is surface application of urea or ammonium bicarbonate and to a lesser degree other ammonium-containing fertilizers. As urea is the most important N fertilizer in the world, it may lead to important NH_3 loss upon hydrolysis and subsequent pH rise in the vicinity of the urea till. The transformation of NH_4^+ to the volatile form NH_3 increases with increasing pH, temperature, soil porosity, and wind speed at the soil surface. It decreases with increasing water content and rainfall events following application. Ammonia losses from soils can be effectively reduced by fertilizer incorporation or injection instead of surface application.

Emission of Nitrogen Oxides (N_2O, NO) and Molecular Nitrogen (Nitrification and Denitrification)

Microbial nitrification and denitrification are responsible for the emission of NO and N_2O.[6] They are by-products in nitrification and intermediates during denitrification. Probably about 0.5% of fertilizer N applied is emitted as NO[7] and 1.25% as N_2O.[8] However, wide ranges have been reported. Intensification of arable agriculture and of animal husbandry has made more N available in the soil N cycle increasing the emission of N oxides. The relative percentage of NO and N_2O formation very much depends on the moisture content of the soil. At a water-filled pore space (WFPS, or the fraction of total soil pore space filled with water) below 40% NO is produced mainly from nitrification. Between a WFPS of 40% and 60% formation of NO and N_2O from nitrification occurs. Between a WFPS of 60% and 80% N_2O is predominantly produced from denitrification and the formation of NO is decreasing sharply. At a WFPS above 80% the formation of N_2 by denitrification is dominant. In practice these WFPS ranges will overlap and depend on the soil type.[9] Next to water content, also temperature, land use and availability of N and decomposable organic matter are important determining factors for N_2O formation. Nitrous oxide is a greenhouse gas contributing 5%–6% to the enhanced greenhouse effect. Increased concentrations are also detrimental for the stratospheric ozone layer.[10] In the presence of sunlight, NO_x (NO and NO_2) react with volatile organic compounds from evaporated petrol and solvents and from vegetation and forms tropospheric ozone which is, even at low concentration, harmful to plants and human beings. The major gaseous end-product of denitrification is N_2. The ratio of N_2O to N_2 produced by denitrification depends on many environmental conditions. Generally themore anaerobic the environment the greater the N_2 production. Denitrification is controlled by three primary factors (oxygen, nitrate and carbon), which in turn are controlled by several physical and biological factors. Denitrification N loss can reach 10% of the fertilizer N input—more on grassland and when manure is also applied.[11] Chemical denitrification is normally insignificant and is mainly related to the stability of NO_2^- and acid conditions.[12] It is more difficult to reduce N_2O and NO from soils then NH_3 losses. A general principle is to minimize N surpluses in the soil profile via careful fertilizer adjustment, corresponding to the actual crop demands.

Leaching

Applied NO_3^-or NO_3^-, formed through nitrification from mineralized NH_4^+ or from NH_4^+ from animal manure, can leach out of the rooting zone. It is well possible that this leached NO_3^- can be denitrified at other places and returned into the atmosphere. The amount and intensity of rainfall, quantity and frequency of irrigation, evaporation rate, temperature, soil texture and structure, type of land use, cropping and tillage practices and the amount and form of fertilizer N are all parameters influencing the amount of NO_3^- leaching to the underground water. Nitrate leaching should be kept under control as it may influence the nitrate content in drinking water influencing human health and in surface water, causing eutrophication. Nitrate losses can be minimized by reducing the mineral N content in the soil profile during the winter period by careful fertilizer adjustment, growing of cover crops or riparian buffer areas.

References

1. McCarty, G.W. Modes of action of nitrification inhibitors. Biol. Fert. Soils **1999**, *29*, 1–9.
2. Prosser, J.I., Ed. *Nitrification, Special Publications of the Society of General Microbiology;* IRL Press: Oxford, 1986; 20 pp.
3. Lagreid, M.; Bockman, O.C.; Kaarstad, O. *Agriculture, Fertilizers and the Environment*; CIBA Publishing: Oxon, U.K., 1999; 294 pp.
4. Sutton, M.A.; Lee, D.S.; Dollard, G.J.; Fowler, D. International conference on atmospheric ammonia: emission, deposition and environmental impacts. Atmospheric Environment **1998**, *32,* 1–593.
5. ECETOC *Ammonia Emissions to Air in Western Europe, (No. 62)*; European Centre for Ecotoxicology and Toxicology of Chemicals: Brussels, 1994; 196 pp.
6. Bremner, J.M. Sources of nitrous oxide in soils. Nutr. Cycl. Agroecosys. **1997**, *49,* 7–16.
7. Veldkamp, E.; Keller, M. Fertilizer-induced nitric oxide emissions from agricultural soils. Nutr. Cycl. Agroecosys. **1997**, *48,* 69–77.
8. Mosier, A.; Kroeze, C.; Nevison, C.; Oenema, O.; Seitsinger, S.; Van Cleemput, O. Closing the global N_2O budget: nitrous oxide emissions through the agricultural N cycle. Nutr. Cycl. Agroecosys. **1998**, *52,* 225–248.
9. Davidson, E.A. Fluxes of Nitrous Oxide and Nitric Oxide from terrestrial ecosystems. In *Microbial Production and Consumption of Greenhouse Gases: Methane, Nitrogen Oxides, and Halomethanes;* Rogers, J.E., Whitman, W.B., Eds.; American Society for Microbiology: Washington, DC, 1991; 219–235.

10. Crutzen, P.J. The influence of nitrogen oxides on the atmospheric ozone content. Quat. J. Royal Meteor. Soc. **1976**, *96*, 320–325.
11. von Rheinbaben, W. Nitrogen losses from agricultural soils through denitrification—A critical evaluation. Z. Pflanzenern. Bodenk. **1990**, *153*, 157–166.
12. Van Cleemput, O. Subsoils: chemo- and biological denitrification, N_2O and N_2 emissions. Nutr. Cycl. Agroecosys. **1998,** *52*, 187–194.

30

Phenols

Leszek Wachowski
and Robert Pietrzak

Introduction

In 1832, German analytical chemist F.F. Runge extracted from coal a tar substance, which a year later, F. Gerhard called phecolenolem (Latin *phenolum, carbolum*). The old name for benzene was phene, and its hydroxyl derivative came to be called phenol. It is the simplest compound from the numerous class generally called phenols, comprising hydroxyl derivatives of aromatic compounds (benzene, naphthalene).

The compound whose correct chemical name according to the International Union of Pure and Applied Chemistry (IUPAC), nomenclature should be benzenol can be treated as benzene in which one of the hydrogen atoms has been replaced by a hydroxyl group. Phenols are aromatic alcohols with one or more hydroxyl groups directly bonded to the aromatic ring, which differentiates them from alcohols. Their structure is reflected in their properties, e.g., in enhanced acidic properties.[1]

Phenols make up a large group of compounds of natural, anthropogenic, or endogenic origin found in the biosphere. Because of a number of physical and chemical properties attractive from the viewpoint of practical use, phenol is produced on a large scale, about 7 billion kg/yr, as a precursor to many useful compounds. The greatest producer of phenol is the United States. Phenol is one of the oldest chemical intermediates in organic industry, and its economical importance continuously increases. The greatest amounts of phenol are used in organic synthesis, including manufacturing of phenolic resins and nylon, in which the chemical intermediates are bisphenol-A and ε-caprolactam,[2] respectively. Bacterial and fungicidal properties of phenol were the first recognized and were the reason for its wide application. The employment of phenol for sterilization has been of great importance not only in medicine but also in the history of humanity. Sir J. Lister, a Scottish surgeon, introduced the method of disinfection of surgery instruments and hands of surgeons with a 5% water solution of phenol, known at that time as carbolic acid.[3]

As a result of natural transformations taking place in the natural environment and the processes generating various phenolic compounds performed by man and their widespread use, the phenolic compounds have penetrated the biosphere. They are met more often as micropollutants of water than the atmospheric air. As the exposure of living organisms to these toxic compounds can be harmful, their presence has been monitored, and certain measures have been undertaken to restrict their amount introduced to the natural environment.[1,2,4]

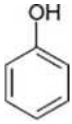
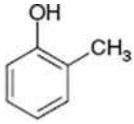
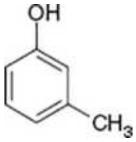
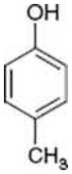
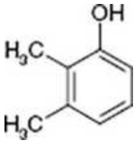
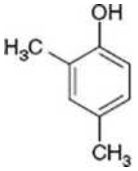
Origin, Properties, and Application of Phenol and Its Derivatives

Phenol and phenolic derivatives met in the biosphere are released from natural and anthropogenic sources. Their amounts entering the natural environment and coming from anthropogenic sources have been estimated as much higher.[5–7]

The class of compounds known as phenols plays a great role in many areas of our lives. Both the synthetic and natural phenols have been raw products or intermediates for the production of, e.g., plastics, detergents, or cosmetic products. Because of their biological activity and, in particular, antioxidant properties, they are essential components of our diet and substrates of a large number of therapeutic drugs.[8–15]

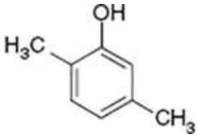
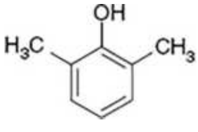
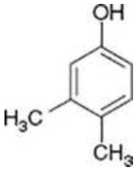
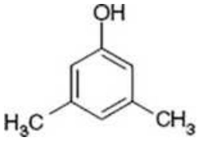
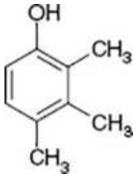
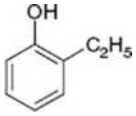
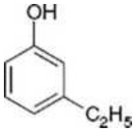
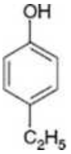
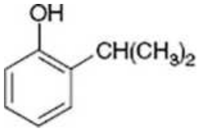
The large diversity of chemical compositions and structures of phenol compounds is illustrated in Tables 1–7.

TABLE 1 Name, Molecular Formula, and Structures of Selected Groups of Phenolic Compounds

No.	Name of Compound	Molecular Formula	Structural Formula
	Methyl derivatives of one-hydroxyl phenols		
1.	*Phenol **Carbolic acid, hydroxybenzene, benzenol, phenylic acid, phenic acid	C_6H_6O ***C_6H_5OH	
2.	o-Cresol *2-Methylphenol **2-Hydroxytoluene, 2-methylbenzenol	C_7H_8O ***$(CH_3)C_6H_4OH$	
3.	*m*-Cresol *3-Methylphenol **3-Hydroxytoluene, 3-methylbenzenol	C_7H_8O ***$(CH_3)C_6H_4OH$	
4.	*p*-Cresol *4-Methylphenol **4-Hydroxytoluene, p-hydroxytoluene, p-methylphenol, 4-methylbenzenol	C_7H_8O ***$(CH_3)C_6H_4OH$	
5.	2,3-Xylenol **2,3-Dimethylphenol, 1-hydroxy-2,3-methylbenzene, 3-hydroxyl-o-xylene	$C_8H_{10}O$ ***$(CH3)_2C_6H_3OH$	
6.	2,4-Xylenol **2,4-Dimethylphenol, 1-hydroxy-2,4-methylbenzene, 4-hydroxy-m-xylene	$C_8H_{10}O$ ***$(CH_3)_2C_6H_3OH$	

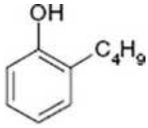
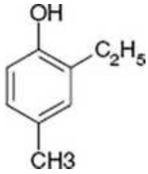
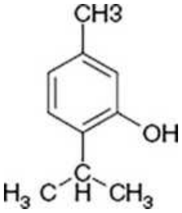
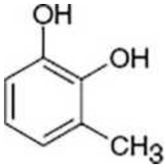
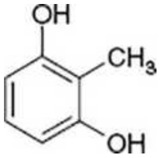
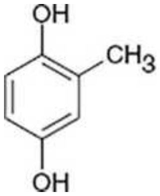
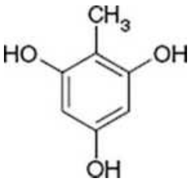
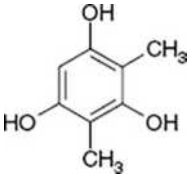
(*Continued*)

TABLE 1 (*Continued*) Name, Molecular Formula, and Structures of Selected Groups of Phenolic Compounds

No.	Name of Compound	Molecular Formula	Structural Formula
7.	2,5-Xylenol **2,5-Dimethylphenol, 2-hydroxy-*p*-xylene, *p*-xylenol, 1-hydroxy-2,5-methylbenzene	$C_8H_{10}O$ ***$(CH_3)_2C_6H_3OH$	
8.	2,6-Xylenol **2,6-Dimethylphenol, 2-hydroxy-rn-xylene, 1-hydroxy-2,6-methylbenzene	$C_8H_{10}O$ ***$(CH_3)_2C_6H3OH$	
9.	3,4-Xylenol **3,4-Dimethylphenol, 4-hydroxy-o-xylene, 1-hydroxy-3,4-methylbenzene	$C_8H_{10}O$ ***$(CH_3)_2C_6H_3OH$	
10.	3,5-Xylenol **3,5-Dimethylphenol, 5-hydroxy-rn-xylene, 1-hydroxy-3,5-methylbenzene	$C_8H_{10}O$ ***$(CH_3)_2C_6H_3OH$	
11.	2,3,4-Trimethylphenol ***1-Hydroxy-2,3,4-methylbenzene	$C_9H_{12}O$ ***$(CH_3)_3C_6H_2OH$	
	One core phenols with aliphatic side-chain		
12.	2-Ethylphenol **(1-Hydroxy-2-ethylbenzene, o-ethylphenol)	$C_8H_{10}O$ ***$(C_2H_5)C_6H_4OH$	
13.	3-Ethylphenol **(1-Hydroxy-3-ethylbenzene, *m*-ethylphenol)	$C_8H_{10}O$ ***$(C_2H_5)C_6H_4OH$	
14.	4-Ethylphenol **(1-hydroxy-4-ethylbenzene, p-ethylphenol, 1-hydroxy-4-ethylbenzene, 4-hydroxyphenylethane)	$C_8H_{10}O$ ***$(C_2H_5)C_6H_4OH$	
15.	2-Isopropylphenol **(*o*-Isopropylphenol, 1-hydroxy-2-isopropylbenzene, o-hydroxycumene; 2-hydroxycumene)	$C_9H_{12}O$ ***$(C_3H_7)C_6H_4OH$	

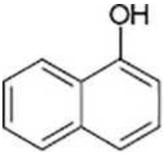
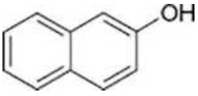
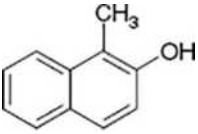
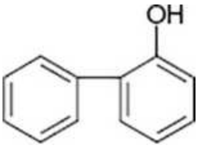
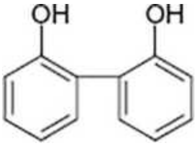
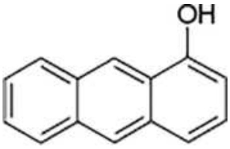
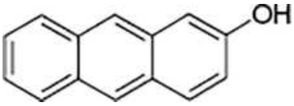
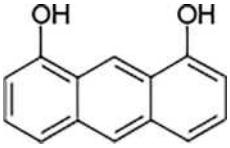
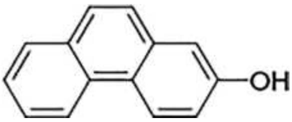
(*Continued*)

TABLE 1 (*Continued*) Name, Molecular Formula, and Structures of Selected Groups of Phenolic Compounds

No.	Name of Compound	Molecular Formula	Structural Formula
16.	o-Butylphenol *2-Butylphenol **(1-Hydroxy-2-butylbenzene)	$C_{10}H_{14}O$ ***$C_6H_4(OH)(C_4H_9)2$	
17.	*2-Ethyl-4-methylphenol **(2-Ethyl-4-Methyl-phenol, 2-ethyl-p-cresol)	$C_9H_{12}O$ ***$(C_2H_5)(CH_3)C_6H_3OH$	
18.	Thymol *2-Isopropyl-5-methylphenol **(IPMP)	$C_{10}H_{14}O$ ***$2\text{-}[(CH_3)_2CH]C_6H_3\text{-}5\text{-}(CH_3)$ OH	
	Methyl derivatives of di- and trihydroxyl phenols		
19.	3-Methylpyrocatechol **(Pyrocatechol 1,2-dihydroxy-3-methylbenzene, 2,3-dihydroxytoluene, 2-hydroxy-3-methylphenol)	$C_7H_8O_2$ ***$(CH_3)C_6H_3(OH)_2$	
20.	2-Methylresorcinol **(2,6-Dihydroxy, 2,6-toluenediol, toluene-2,6-diol)	$C_7H_8O_2$ ***$(CH_3)C_6H_3(OH)_2$	
21.	2-Methylhydroquinol *2-Methylbenzene-1,4-diol **(2-Methyl-1,4-benzenediol,2-methyl-1,4-hydroquinone, 2-methylbenzene-1,4-diol)	$C_7H_8O_2$ ***$(CH_3)C_6H_3(OH)_2$	
22.	2,4,6-trihydroxytoluene *2-Methylbenzene-1,3,5-triol **(2-Methylphloroglucinol, Toluene-2,4,6-triol	$C_7H_8O_3$ ***$C_6H_2(OH)_3(CH_3)$	
23.	2,4-Dimethylphloroglucinol *2,4-Dimethylbenzene-1,3,5-triol	$C_8H_{10}O_3$ ***$(CH_3)_2C_6H(OH)_3$	

(*Continued*)

TABLE 1 (*Continued*) Name, Molecular Formula, and Structures of Selected Groups of Phenolic Compounds

No.	Name of Compound	Molecular Formula	Structural Formula
		Di-co,re phenols	
24.	1-Naphthol *Naphthalen-1-ol **(1-hydroxynaphthalene, 1-Naphthalenol; alpha-Naphthol a-Naphtol, naft-1-ol)	$C_{10}H_8O$ **$C_{10}H_7OH$	
25.	2-Naphthol *Naphthalen-2-ol **(2-hydroxynaphthalene, 2-Naphthalenol; beta-Naphthol β-Naphtol, naft-2-ol)	$C_{10}H_8O$ **$C_{10}H_7OH$	
26.	Methyl-2-naphtol **(1-methyl-2-hydroxynaphthalene, 1-methylnaphthalen-2-ol)	$C_{11}H_{10}O$ ***$(CH_3)C_{10}H_6OH$	
27.	*2-phenylphenol **(2-Hydroxybiphenyl, o-phenylphenol, biphenylol, orthophenyl phenol, o-xenol, orthoxenol)	$C_{12}H_{10}O$ ***$C_{12}H_9OH$	
28.	2,2'-Biphenol **(2,2'-Dihydroxybiphenyl, 2,2'-biphenyldiol)	$C_{12}H_{10}O$ ***$C_{12}H_8(OH)_2$	
		Thre -core phenols	
29.	1-Anthranol *Anthracen-9-ol **(1-Hydroxyanthracene, 9-anthracenol, 9-anthranol)	$C_{14}H_{10}O$ ***$C_{14}H_9OH$	
30.	2-Anthranol *Anthracen-2-ol **(2-Hydroxyanthracene, 2-anthracenol, 2-anthranol, beta-hydroxyanthracene)	$C_{14}H_{10}O$ ***$C_{14}H_9OH$	
31.	1,9-Dihydroxyanthracene *Anthracene-1,9-diol **1,9-Anthracenediol	$C_{14}H_{10}O_2$ ***$C_{14}H_8(OH)_2$	
32.	2-Phenantrenol	$C_{14}H_{10}O_2$ ***$C_{14}H_9(OH)$	

*, IUPAC name; **, synonyms; ***, other formula.

Sources of Phenol and Phenolic Compounds in the Biosphere

Natural Origin

Phenol and phenolic derivatives from natural sources are commonly met organic micropollutants. The precursors of the majority of them are phenylalanine and tyrosine, from which as a result of deamination, cinnamic acid and its hydroxy derivatives are formed.[4,8,13–16]

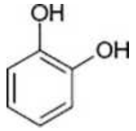
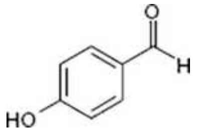
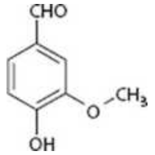
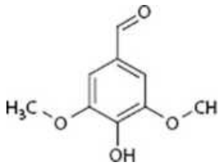
Usually, these compounds are the following:

- Intermediate products of natural decay of organic matter such as proteins, humic compounds, and lignin[8]
- Products of metabolic processes taking place in living organisms[1,2,9]
- Secondary metabolic plant product[1,2,9]
- Products formed as a consequence of forest fires[17]
- Products of natural decay of original plant matter![1,2,9,16]

In Table 2, molecular and structural formulae of the natural derivatives of phenol most often met in the natural environment are presented.

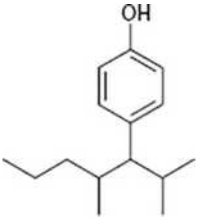
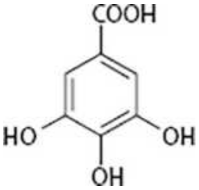
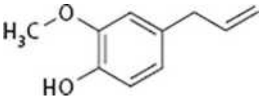
Lignin is a complex chemical compound most commonly found in wood and all vascular plants, localized not only between the cells but also in the cell walls. In lignin, two groups of phenolic compounds dominate. The first comprises phenyl acid derivatives, occurring in plants mostly in the bound form as components of lignin and hydrolyzing tannins in the form of esters and glycosides.[13,15]

TABLE 2 Phenol Compounds of Natural Origin Most Commonly Found in the Biosphere

No.	Name of Compound	Molecular Formula	Structural Formula
1.	*Phenol **Carbolic acid, hydroxybenzene, benzenol, phenylic acid, phenic acid	(See Table 1 no.1)	
2.	Catechol *Benzene-1,2-diol **(Pyrocatechol, 1,2-benzenediol, 2-hydroxyphenol, 1,2-dihydroxybenzene)	$C_6H_6O_2$ ***$C_6H_5(OH)2$	
3.	*4-Hydroxybenzaldehyde ***(*p*-Hydroxybenzaldehyde, 1-hydroxybenzaldehyde, p-formylphenol)	$C_7H_6O_2$ ***$(CHO)C_6H_4OH$	
4.	*o*-Cresol *2-Methylphenol **2-Hydroxytoluene, 2-methylobenzenol	(See Table 1 no. 2)	
5.	Vanillyn *4-Hydroxy-3-methoxybenzaldehyde **(Methyl vanillin, vanillin, vanillic aldehyde, 3-methoxy-4-hydroxybenzaldehyde)	C8H8O3 ***$(OCH_3)(CHO)$ C_6H_3OH	
6.	Syringaldehyde *4-Hydroxy-3,5-dimethoxybenzaldehyde (3,5-Dimethoxy-4-hydroxybenzaldehyde, **3,5-dimethoxy-4-hydroxybenzene carbonal, gallaldehyde 3,5-dimethyl ether, 4-hydroxy-3,5-dimethoxybenzaldehyde, syringic aldehyde)	C9H10O4 ***(OCH3)2(CHO) C6H2OH	

(*Continued*)

TABLE 2 (*Continued*) Phenol Compounds of Natural Origin Most Commonly Found in the Biosphere

No.	Name of Compound	Molecular Formula	Structural Formula
7.	Alkyl phenols With saturated alkyl groups at para position C_nH_{2n}+i, usually $n = 1\text{-}3$	C6H4(OH)(CnH2n+1) $n = 1\text{-}12$	
8.	Gallic acid *3,4,5-Trihydroxybenzoic acid **(Gallic acid, gallate, 3,4,5-trihydroxybenzoate)	$C_7H_6O_5$ ***$C_6H_2(OH)_3(COOH)$	
9.	Eugenol *4-Allyl-2-methoxyphenol **[4-Allyl-2-methoxyphenol, 2-methoxy-4-(2- propenyl)phenol, eugenic acid, caryophyllic acid, 1-allyl-3-methoxy-4-hydroxybenzene, allylguaiacol, 2-methoxy-4-allylphenol,-4-allylcatechol-2-methyl ether, 2-methoxy-4-(2-propen-1-yl)phenol]	$C_{10}H_{12}O_2$ ***$(OCH_3)(C_3H_4)$ C_6H_3OH	
10.	Thymol *2-Isopropyl-5-methylphenol **(IPMP)	(See Table 1 no. 18)	
11.	1-Naphthol *Naphthalen-1-ol **(1-Hydroxynaphthalene, 1-naphthalenol; *alpha-naphthol*, α-Naphtol, naft-1-ol)	(See Table; 1 no. 24)	

*, IUPAC name; **, synonyms; ***, other formula.

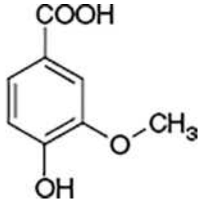
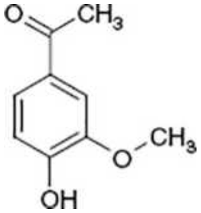
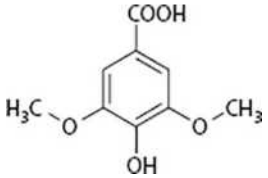
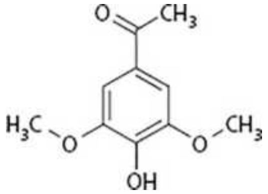
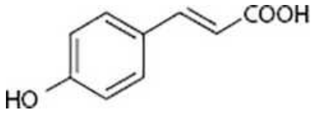
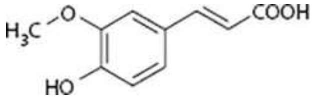
The second group comprises hydroxycinnamic acids present most often in the form of esters, while hydroxybenzoic acids are mainly present in plants in the form of glycosides. Moreover, in plant tissues, phenyl acids have been identified, besides other naturally occurring compounds, like flavonoids, fatty acids, sterols, or cell wall polymers.[15]

It has been established that gymnospermous (*Gymnospermae*) plants like pine, spruce, and fir contain vanillyl, but they do not contain siryngyl and cinnamyl phenols. In the angiospermous *(Angiospermae)* plants, both vanillyl and siryngyl compounds are present, but cinnamyl compounds are not found.[12–16] Some examples of natural phenolic compounds used as indicators of original plant matter are given in Table 3.

Phenols of natural origin, generally called photochemical, include a large and very chemically diverse group of polyphenols characterized by the presence of large multiples of phenol units.

The above phenolic compounds, of which more than 8000 are known, embrace a wide range of plants' secondary metabolites, possessing in common an aromatic ring substituted by one or more hydroxyl groups.[13–16] They are the most widely distributed secondary metabolites ubiquitously present in the plant kingdom. Simple phenols are relatively rare in plants.[14] Some polyphenolic substructures are secondary plant substances called quasi-vitamins or natural nonnutritious substances showing great biological activity. This group of compounds is sometimes referred to as quasi-vitamin or vitamin P. They are mainly found in the bound form as components of lignin and hydrolyzing tannins, in the form of esters

TABLE 3 Phenolic Derivatives Used as Indicators of Original Plant Matter

No.	Name of Compound	Molecular Formula	Structural Formula
1.	Vanillyn	(See Table; 2 no.5)	
2.	Vanillic acid *4-Hydroxy-3-methoxybenzoic acid **(4-Hydroxy-rn-anisic acid, vanillate)	$C_8H_8O_4$ **$C_6H_3(OH)(OCH_3)(COOH)$	
3.	Apocynin *1-(4-Hydroxy-3-methoxyphenyl)ethanone **4-Hydroxy-3-methoxyacetophenone, acetovanillone	$C_9H_{10}O_3$ ***$C_6H_3(OH)(OCH_3)$ $(COCH_3)$	
4.	Syringaldehyde	(See Table ; 2 no. 6)	
5.	Syringic acid *4-Hydroxy-3,5-dimethoxybenzoic acid **(Gallic acid 3,5-dimethyl ether)	$C_9H_{10}O_5$ $C_6H_2(OH)(OCH_3)(COOH)$	
6.	Acetosyringone *4'-Hydroxy-3',5'-dimethoxyacetophenone **Acetosyringenin	$C_{10}H_{12}O_4$ ***$C_6H_2(OH)$ $(OCH_3)_2(COCH_3)$	
7.	*p*-Coumaric acid *3-(4-Hydroxyphenyl)-2-propenoic acid **para-Coumaric acid, 4-hydroxycinnamic acid, β-(4-hydroxyphenyl)acrylic acid]	$C_9H_8O_3$ ***$C_6H_4(OH)$ (CH=CH-COOH	
8.	Ferulic acid *(E)-3-(4-hydroxy-3-methoxy-phenyl) prop-2-enoic acid **[2-Propenoic acid, 3-(4-hydroxy-3-methoxyphenyl)-ferulic acid, 3-(4-hydroxy-3-methoxyphenyl)-2-propenoic acid, 3-(4- hydroxy-3-methoxyphenyl) acrylic acid, 3-methoxy-4-hydroxycinnamic acid, 4-hydroxy-3-methoxycinnamic acid, (2E)-3-(4-hydroxy-3-methoxyphenyl)-2-propenoic acid, ferulate coniferic acid, trans-ferulic acid, (E)-ferulic acid]	$C_{10}H_{10}O_4$ ***$C_6H_3(OH)(OCH_3)$ (CH=CH-COOH)	

Note: * IUPAC name; **, synonyms; ***, other formula.

Source: Adapted from Kroon and Wiliamson,[12] Robinson,[13] and Duke.[14]

or glycosides, in all parts of plants, i.e., in the leaves, seeds, flowers, fruit, roots, bark, and wooden parts. Their chemical composition depends on the species and variety of a plant, climatic conditions, and agrotechnological procedures applied. Some polyphenols occurring in plants belong to phytoalexins, so the substances involved in the mechanisms protecting against the attack of insects, fungi, or viruses.

To sum up, polyphenols do the following:

- Play important roles in plant metabolism[13–16]
- Display high bioactivity; such natural products with healing and nutritional values are called nutraceuticals or dietary supplements[18]
- Participate in the healing and adaptive processes in living organisms, can protect against development of cancer, or can have therapeutic effect in treatment of different diseases[10,12]
- Participate in regulation of growth and reproduction of plants[11–13]
- Contribute in determination of sensory features of fruit, vegetables, and processed food[10–14]
- Endow plants and fruits with a specific tart and bitter taste[13,16]
- Responsible for the color and fibrous nature of plants and fruit[13]
- Participate in morphogenesis, energy flow, sex determination, photosynthesis, respiration, regulation of gene expression, and regulation of synthesis of growth hormones[19]
- Have a protecting effect against ultraviolet irradiation and against stress[14,15]

Humic substances are the end products of decaying natural organic matter in the microbial process taking place with involvement of edaphone (mainly fungi and actinomycetes) called humification. They are major organic constituents of soil (humus), peat bogs, coal, sewage, compost heaps, carbonaceous shales, lignites, and all types of natural waters (aquatic humic substances) and can form complex ions that are commonly found in the natural environment. A typical humic substance is a mixture of many molecules, some of which are based on a motif of aromatic nuclei with phenolic and carboxylic substituents (see Figure 1).[20–22]

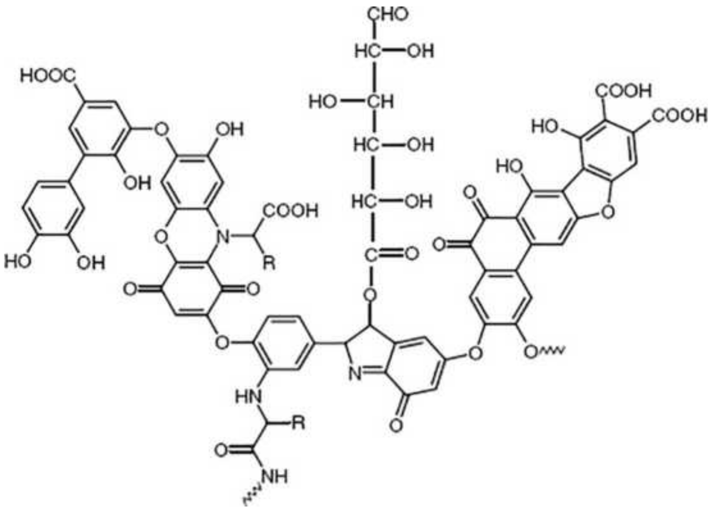


FIGURE 1 The hypothetical structure of humic acid, having a variety of components including quinone phenol, cresol, and sugar moieties. Source: Adapted from Stevenson,[20] Muscola and Soidari,[21] and Hessen and Tranvik.[22]

There are three types of humic substances, which differ slightly in acidity and chemical composition. They are humic acid, fulvic acid, and humins.

Aquatic humus substances bear about 40%-60% of dissolved organic carbon and make up the largest fraction of natural organic matter in water. The major functional groups include carboxylic acids, phenolic hydroxyl, and carbonyl and hydroxyl groups.[22] Phenol hydroxyl groups are usually in amounts of about 1 μeq/mg C of humic material or 94 |μag/mg of elemental carbon as humic matter.[5]

As mentioned earlier, humin substances are also constituents of coal, which is a fossil phytomaterial. It is assumed that lignin, which is one of the most important chemical compounds in plants, is one of the parent substances of coals having aromatic acidic character. It is a consequence of the presence of phenolic hydroxyls, which are the most characteristic oxygen group of coals. The degree of carbonization of coal is measured by the content of hydroxyl groups. In coals with a low degree of carbonization, hydroxyl groups comprise up to 90% of all oxygen groups. Their content decreases with increasing degree of carbonization, and they disappear in coal with a high degree of carbonization.[23]

Anthropogenic Origin

It has been established that a fundamental part of phenol present in the natural environment is of anthropogenic origin. Considerable sources of phenol and its derivatives are the processes of production of intermediate semiproducts and their further processing. Three main sources of phenol and its derivatives released to the natural environment are distinguished: industrial processes, nonindustrial processes and endogenous sources. The greatest amounts of phenol and its derivatives come from industrial processes, mainly production and application of different kinds of phenolic resins and phenolic plastics, generally called phenolics, and caprolactam.[24] Great demand for phenolic resins, widely applied, e.g., as a binding agent in insulating materials, chipboards, shatterproof glass, paints, and casting molds, makes them a profound source of phenolic compounds. The emission of these compounds from these materials is measured in terms of a concentration of free phenol occurring in the monomeric form in the resins in the amount of 1%-5% wt.[24] In foundries, phenolic compounds are emitted in processes of mold production and in casting.[26] Phenolic compounds occur in large amounts in the products coming from the chemistry of coal and chemistry of coke. Processing of raw benzenol obtained as a result of coal coking leads to formation of phenol and xylems. Phenolic compounds are met in wastewater from industrial plants in which phenol and its derivatives are used as raw materials in amounts to a few g/dm^3. Table 4 gives a list of processes that are the most abundant sources of phenol released to the natural environment and values of its emission coefficients.

The most important nonindustrial source of phenolic compounds is fuel combustion in motor vehicles. The exhaust gases from motor vehicles contain from about 0.3 to 1.4–2.0 ppm of phenol, which corresponds to the amounts from 1.2 to 5.4–7.7 mg/m^3.[24,28,29]

TABLE 4 Selected Industrial Processes Leading to Release of Phenol to the Atmospheric Air and Estimated Values of Its Emission Coefficients

Type of Process	Phenol Emission Coefficient
Production of phenol resins	0–0.5 g of phenol emitted per kg of resin
Production of phenol and its derivatives	–
Production of caprolactam	0.2–0.05 g of phenol emitted per kg of cyclohexanol (semiproduct)
Production of coke	–
Production of insulating materials	–
Emission from phenol processing	–

Source: RIVM Criteria Document: Phenol, Bilthoven, the Netherlands.[27]

Phenolic compounds in considerable concentrations are released in volatile form during combustion of other fuels such as wood, coal, and mazout, in heating chambers, house furnaces, and fireplaces.[30] It should be mentioned that phenolic compounds are commonly met in cigarette smoke, in which the estimated average amount of phenol is 0.4 mg per cigarette.[31] Phenol can be also found in smoked food products.[32]

Endogenous sources of phenol include its synthesis in vivo from different xenobiotics released to the natural environment, e.g., from benzene. It has been established that benzene and its phenolic derivatives in the in vivo conversion can be an endogenous source of phenol.[33] Among phenol derivatives, particular attention has been paid to pentachlorophenol (PCP) and its salts, because of its properties and low cost of production. This material has been widely applied in industry (mainly for wood impregnation) and agriculture and as a component of household products (disinfectant). In agriculture, PCP is used as a pesticide of a broad spectrum of activity as it acts against algae, bacteria, fungi, weeds, insects, and mollusks. Pentachlorophenol has been recognized as hazardous for human health because of its toxicity and widespread use. It is used in leather tanning and finishing. Monochlorophenols are used as synthetic intermediates for dyes and chlorinated phenols. Pentachlorophenol is used as a denaturant for alcohol, an antiseptic, and a selective solvent for refining minerals. [34–36]

Nomenclature, Structure, and Chemical Composition

Phenol is a common name for the simplest and most common aromatic alcohol, labeled with CAS-RegistrySM –The world's largest substance database number 108–95-2, in which the hydroxyl group, known as a phenolic hydroxyl, is attached to the phenyl group. By definition, phenol is a hydroxybenzene, but according to the IUPAC, its correct name should be benzenol.

In literature, phenol is referred to by many other names, such as benzyl, carbolic acid, carbol, phenol liquefied, phenolic acid, phenyl hydroxide, phenic acid, phenylic acid, oxybenzene, monophenol, monohydroxybenzene, and phenyl hydrate.[1,2,24] It also has quite a few commercial names, such as carbol liquor and phenyl liquor (Netherlands), kristallliertes Kreozot or Steinkohlenteerkreosot (Germany), Venzenol (France), or code label ENT1814. The class of phenolic compounds, whose name comes from phenol, includes cresols (aromatic compounds from the group of phenols that are derivatives of toluene).

In contrast to alcohols, in phenol, the hydroxyl group is attached directly to the aromatic ring, which gives enhancement of its acidic properties. Phenols are the compounds in which the ring (benzene) is substituted with one or a few hydroxyl groups. If the compounds are built with aromatic rings (naphthalene, anthracene), with one or more hydroxyl substituents at the benzene ring, then they are called naphthols (hydroxynaphthalenes) or hydroxyanthracenes.[1,2]

According to the IUPAC nomenclature, in naming substitution products of these compounds, the numbering starts at the group already present and is done in the direction that gives the lowest numbers to other groups on the ring. Sometimes, the benzene ring is treated as a substituent (for hydrogen) on another molecule. In that case, the C_6H_5- group of benzene is called "phenyl." Substituents are cited in alphabetical order. Carboxyl and acyl groups take precedence over the phenolic hydroxyl in determining the base name. The hydroxyl group is treated as a substituent. Higher substituted compounds are named as derivatives of phenol.

Structural diversity of phenol compounds makes their systematization rather difficult. With respect to the structure of the carbon skeleton, phenolic compounds can be divided into the following groups:

- Phenylcarboxylic acids—derivatives of benzoic acid of the carbon skeleton C_6–C_i (where C_6 denotes a benzene ring), whose structure can be described by the general formula in Figure 2 (see also Table 5).
- Phenylpropenic acids—derivatives of cinnamic acid of the carbon skeleton C6–C3, whose structure can be described by the general formula in Figure 3 (see also Table 6).

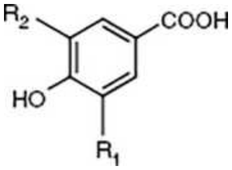
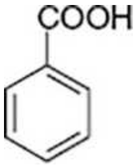
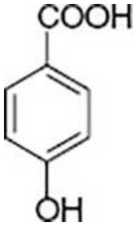
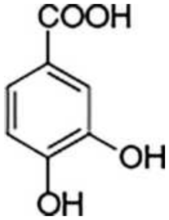


FIGURE 2 The general formula of aromatic hydroxyacid derivatives of benzoic acid.

TABLE 5 Exemplary Hydroxyacids That Are Benzoic Acid Derivatives, Occurring in Plants

No.	Name of Hydroxyacid	Molecular Formula	Structural Formula
1.	*Benzoic acid **(Benzenecarboxylic acid, carboxybenzene, E210, dracylic acid)	$C_7H_6O_2$ ***$C_6H_5(COOH)$	
2.	*4-Hydroxybenzoic acid **(*p*-Hydroxybenzoic acid, para-hydroxybenzoic acid)	$C_7H_6O_3$ ***$C_6H_5(COOH)$	
3.	Protocatechuic acid *3,4-Dihydroxybenzoic acid **Protocatechuic acid (PCA)	$C_7H_6O_4$ **$C_6H_4(OH)_2(COOH)$	
4.	Gallic acid	(See Table 2 no. 8)	
5.	Vanillic acid	(See Table 3 no. 2)	
6.	Syringic acid	(See Table 3 no. 5)	

Note: *, IUPAC name; **, synonyms; ***, other formula

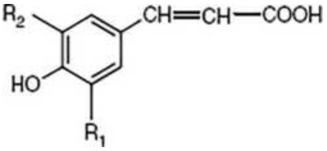
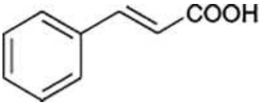
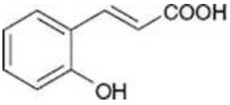
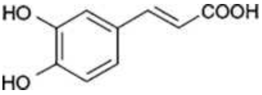
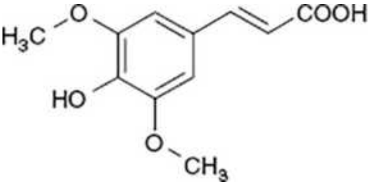


FIGURE 3 The general formula of aromatic hydroxyacids that are derivatives of cinnamic acid.

- Stilbene (trans)—of the carbon skeleton C_6–C_2–C_6, naturally occurring compounds, found in a wide range of plants, aromatherapy products, and dietary supplements, whose structure can be illustrated by the following general formula in Figure 4. Exemplary compounds from this group are resveratrol, piceatannol, rhapontigerin, and pterostilbene. The first mentioned stilbene,

TABLE 6 Some Hydroxyacids That Are Cinnamic Acid Derivatives, Occurring in Plants

No.	Name of Compound	Molecular Formula	Structural Formula
1.	Cinnamic acid *(E)-3-phenylprop-2-enoic acid **(Cinnamic acid, trans-cinnamic acid, phenylacrylic acid, cinnamylic acid, 3- phenylacrylic acid, (E)-cinnamic acid, benzenepropenoic acid, isocinnamic acid)	$C_9H_8O_3$ ***$C_6H_4(OH)(C_3H_3O_2)$	
2.	o-Coumaric acid **(2-Hydroxycinnamic, o-coumaric acid, 2-coumaric acid, 2-coumarate 2-hydroxycinnamate, trans-2-hydroxycinnamic acid trans-2-hydroxycinnamate	$C_9H_8O_3$ ***$C_6H_4(OH)(C_3H_3O_2)$	
3.	p-Coumaric acid	See Table 3 no. 7	
4.	Caffeic acid *[3-(3,4-Dihydroxyphenyl 2-propenoic acid, 3,4-dihydroxy-cinnamic acid trans- caffeate, 3,4-dihydroxy-trans-cinnamate), (E)-3-(3,4-dihydroxyphenyl)-2-propenoic acid 3,4-dihydroxybenzeneacrylicacid 3-(3,4- dihydroxyphenyl)-2-propenoic acid]	$C_9H_8O_4$ ***$C_6H_3(OH)_2(C_3H_3O_2)$	
5.	Ferulic acid	(See Table 3 no 8)	
6.	Sinapinic acid *3-(4-Hydroxy-3,5-dimethoxyphenyl) prop-2- enoic acid **(Sinapinic acid, sinapic acid, 3,5-dimethoxy- 4-hydroxycinnamic acid, 4-hydroxy-3,5- dimethoxycinnamic acid)	$C_{11}H_{12}O_5$ ***$C_6H_2(OH)$ $(OCH_3)_2(C_3H_3O_2)$	

Note: *, IUPAC name; **, synonyms; ***, other formula

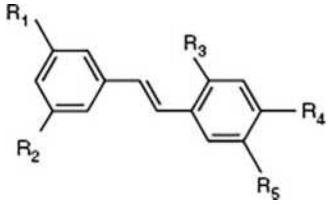


FIGURE 4 The general formula of trans stilbenes (1,2- diphenylethylene; R_1, R_2, R_3, R_4, and R_5 denote OH, OCH_3, or –glucose group).

resveratrol, has been intensely studied and proved to have potent anticancer, anti-inflammatory, and antioxidant activities.[37]

- Polyphenols known also as flavonoids or bioflavonoids, of the carbon skeleton C_6–C_3–C_6, whose general structure is shown in Figure 5. This group includes the following subgroups: flavons, the compounds responsible for the yellow color of plants or their parts (Latin *flavus* means yellow); flavones; flavanones; flavanols; anthocyanes (proanthocyanes); and isoflavones. They constitute the largest class of phenolic compounds, with more than 3000 structures.[38,39] Table 7 presents representative compounds of the above-mentioned subgroups of polyphenols.

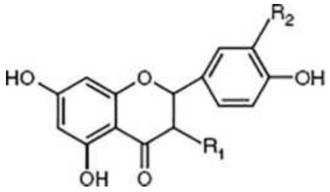


FIGURE 5 The general structure of flavonoids (R_1, R_2 denote H or OH group).

TABLE 7 Selected Examples of Chemical Composition and Structures of Subgroups among Polyphenols

(Polyphenols (flavonoids))		
Name of Compounds	Molecular Formula	Occurrence
Flavons		
Quercetin *2-(3,4-Dihydroxyphenyl)-3,5,7-trihydroxy-4H-chromen-4-one **(Sophoretin, meletin, quercetine, xanthaurine, quercetol, quercitin, quertine, flavin meletin)	$C_{15}H_{10}O_7$	Green tea, grape skin, ginkgo leaves, apples
Flavonols		
Rutin *2-(3,4-Dihydroxyphenyl)-5,7-dihydroxy-3-[α-L-rhamnopyranosyl-(1→6)-β-D-glucopyranosyloxy1-4H-chromen-4-one **(Rutoside, phytomelin, sophorin, birutan, eldrin, birutan forte, rutin trihydrate, globularicitrin, violaquercitrin)	$C_{27}H_{30}O_{16}$	Pogoda tree buds (Sophora japonica Fabaceae), apple skin
Flavanones		
Hesperidin *(2S)-5-hydroxy-2-(3-hydroxy-4-methoxyphenyl)-7-[(2S,3R,4S,5S,6R)-3,4,5-trihydroxy-6-[[(2Æ,3Æ,4Æ,5Æ,6S)-3,4,5-trihydroxy-6-methyloxan-2-yl1oxymethyl1oxan-2-yl1oxy-2,3-dihydrochromen-4-one	$C_{28}H_{34}O_{15}$	Citrus fruit skin
Flavanols (Catechins)		
Epicatechol **(epi-Catechinepi-catechol, epicatechin, l-acacatechin)	$C_{15}H_{14}O_6$	Green tea, grape seeds
Anthocyanes (Proanthocyanes)		
Cyanidin *2-(3,4-Dihydroxyphenyl) chromenylium-3,5,7-triol **Cyanidine	$C_{15}H_{11}O_6^+$	Bilberry, red and black grapes, red wine
Isoflavones		
Genistein *5,7-Dihydroxy-3-(4-hydroxyphenyl)chromen-4-one **4',5,7-Trihydroxyisoflavone	$C_{15}H_{10}O_5$	Leguminous plants, soybean, cereal grains, fruits, vegetables

Note: *, IUPAC name; **, synonyms

- Other phenolic compounds, e.g., tannins (proanthocyanides) of the carbon skeleton Cn>-12, comprising the following:
- Hydrolyzable tannins, consisting of several gallic acid units bound through ester linkages to a central glucose (see Figure 6a); these types of tannins are quite water soluble and are part of the water products of plants.
- Condensed tannins, which are a diverse group of poly- phenolic compounds of plant origin called flavonoids or bioflavonoids (see Figure 6b).

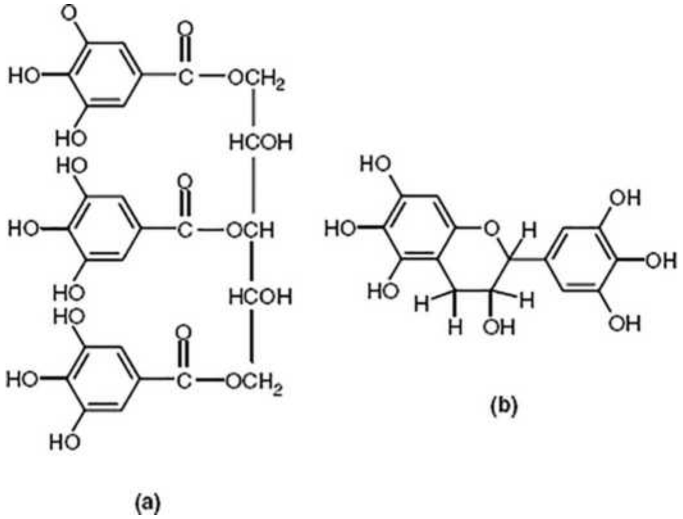


FIGURE 6 The hypothetical structures of hydrolyzable (a) and condensed (b) types of tannins.

Polyphenols make a large group of natural substances occurring in many plants. They are found in greatest amounts in fruit (chokeberry, blueberry, grapes, nuts, garlic) and vegetables (cabbage, cereal seeds).[40–42] Polyphenols are mostly distributed in the external parts of the fruit. In drinks, polyphenols are found in green tea, red wine, coffee, and beer. These compounds make up the largest group of antioxidants supplied in the diet, and hence, they are often called bioflavonoids.[43–45]

Physical, Chemical, and Organoleptic Properties

Phenol is a colorless to light pink crystalline (long needlelike crystallites) solid. When molten, it becomes a bright colorless liquid of low or high viscosity. It attains pink color on exposure to air and light as a result of partial oxidation. In the atmosphere of wet air, the crystals deliquesce. Crude product can be pink, brown, or black. The physicochemical properties of phenols differ significantly from those of aliphatic or unsaturated alcohols; [1,2,7,24] some of them are given in Table 8.

Phenols have hydroxyl groups that can participate in intermolecular hydrogen bonding with other phenol molecules or other H-bonding systems, e.g., water. Hydrogen bonding results in higher dipole moment and melting points and much higher boiling points for phenols than those of hydrocarbons of similar molecular weight. The ability of phenols to form strong hydrogen bonds also enhances their solubility in water.[1,24]

Phenol dissolves to give a 9.3% solution in water, compared with a 3.6% solution of cyclohexanol in water. This water solution of phenol is called phenol liquefied. The water solubility of phenol increases with temperature, and above 68.4°C, both substances become fully miscible.

Phenol is more readily soluble in most organic solvents and in water solutions of soaps. An example of phenol in a water solution of soap is Lysol, being a mixture of a water solution of a potassium soap and cresols, used mainly in the veterinary field for disinfection of rooms and vessels. Its solubility in aliphatic solvents is limited.[1,2]

In contrast to neutral alcohols, water solutions of phenol show weak acidic properties; hence, phenol has been referred to as carbolic acid. Phenol dissociates with the formation of phenolate (C_6H5O^-), also called phenoxide ion, and proton (H3O)$^{+:}$

$$C_6H_5OH_{(aq)} + H_2O_{(1)} \leftrightarrow C_6H_5O^-{}_{(aq)} + H_3O^+{}_{(aq)} \tag{1}$$

TABLE 8 Some Physical and Chemical Properties of Phenol

Molecular weight	94.11 g/mol
Boiling point	181.75°C (101.3 Pa)
Melting point	43.0°C 40.9°C (pure substance)
Relative density	1071
Relative vapor density (air = 1)	3.24 mm Hg
Vapor pressure (20°C)	0.357 mm Hg
(50°C)	2.48 mm Hg
(100°C)	41.3 mm Hg
Concentration of saturated vapor (20°C) in air	0.77 g/m3
Water solubility (16°C)	67 g/L
above 68.4°C	Fully soluble
Partition coefficient n-octanol/ water (logPow)	1.46
Dissociation constant in water Ka (20°C)	1.28 x 10-10
Ignition point:	
• Closed crucible	80°C
• Open crucible	79°C
Acidity	Ka = 1.3 x 10-10 (pKa = 9.55)

Source: Adapted from Rappoport,[1] Tyman,[2] and Weber et al.[24]

Dissociation occurs to only a slight extent; phenol is a very weak acid with $pK_a = 9.55$. Phenols are more acidic than alcohols of $pK_a = 16–20$ but less acidic than carboxylic acids of $pK_a = 5$.[2] Introduction of substitutes, particularly those in ortho or para position to the –OH group, can dramatically influence the acidity of phenol due to resonance and/or inductive effects. In reactions with strong bases (NaOH, KOH), phenols create phenates (C_6H_5OMe), which are more stable than alcoholates:

$$C_6H_5OH + Na^+OH^- \rightarrow C_6H_5O^-Na^+ + H_2O \tag{2}$$

This increased stability is a consequence of the mesomeric effect stabilizing the phenate ion owing to the delocalization of the negative charge on the aromatic ring of phenol.[1,2]

Phenol has a peculiar characteristic smell and a strong corrosive effect on skin. It is poisonous in nature but acts as a powerful antiseptic.[24,41,46]

Phenol is susceptible to oxidants (e.g., CrO_3, $K_2Cr_2O_7$). The hydrogen abstraction from phenolic hydroxyl is accompanied by the resonance stabilization of the formed phenoxy radical ($C_6H_5O^{\cdot}$)**, which can be further oxidized.[46]

Phenols are highly reactive toward electrophilic substitution because the nonbonding electron on oxygen stabilizes the intermediate cation.[1,2,7]

Phenol reacts with carbonyl compounds both in acidic and basic media. In the presence of formaldehyde, it easily undergoes hydroxymethylation followed by condensation of resin. Condensation with acetone gives bisphenol A, a key building block of polycarbonates. Conversion with formaldehyde produces phenolic resins, the best known of which is Bakelite.[1,2,25,47] The product of catalytic hydrogenation of phenol, performed on a large scale, is cyclohexanol. Another industrially important process is alkylation of phenol, which takes place in the presence of an acid catalyst or a Friedl–Crafts catalyst.

In large amounts, phenol is also used to make different derivatives containing chlorine, usually coming from NaOCl or Cl_2. The most important from among these compounds is PCP, manufactured usually by direct chlorination in the presence of metallic nickel as a catalyst or by hydrolysis of hexachlorobenzene.[1,2] Phenol also reacts directly with alkyl halides in alkali solutions to form phenyl ethers. The phenate ion shows nucleophilic character and replaces halogen from alkyl halide:

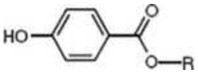


FIGURE 7 General formula of paraben, with R standing for methyl, ethyl, n-propyl, isopropyl, n-butyl, isobutyl, or benzyl group.

$$C_6H_5OH + NaOH \rightarrow C_6H_5ONa + H_2O \tag{3}$$

$$C_6H_5ONa + CH_3Cl \rightarrow C_6H_5OCH_3 + NaCl \tag{4}$$

They are also formed when vapors of phenol and an alcohol are heated over ThO_2.

Phenol can also make esters, among them an interesting group called parabens (esters of *p*-hydroxybenzoic acid with aliphatic alcohol), whose general formula is shown in Figure 7.[48]

Phenol has a characteristic acidic smell and pungent taste. It shows antiseptic activity because of the ability to coagulate protein. It is active against a wide range of microorganisms including some fungi and viruses but is only slightly effective against spores. Although phenol was the first antiseptic used on wounds and in surgery, it is a protoplasmic poison that damages all kinds of cells and is alleged to have caused an astonishing number of poisonings since it came into general use.[2,6,32] Generally, it is a strong and violent poison attacking the nervous system, alimentary track, and circulatory system. It has been estimated that approximately 1 g of phenol is enough to cause death.[2,6,7,27,46] Phenol and its vapors are corrosive to the eyes, skin, and respiratory tract.[49] There is no evidence to claim that phenol causes cancer in humans. Besides its hydrophobic effects, another mechanism of its toxicity is via formation of phenoxyl radicals. Poisoning can also result from inhalation of phenol in the form of atmospheric aerosol, which quickly coagulates in cold air.[50]

Chlorophenols are characterized by a 100–1000 times more intense smell than the parent phenol. During chemical wastewater treatment (chlorination), the water contaminated with phenols acquires a repulsive taste as a result of formation of phenol chloroderivatives.[5] It has been established the limiting concentrations of phenol perceptible though the sense of smell, taste and touch are: 0.021–20 mg/m^3 in air[51] and 0.3 mg/L in water, respectively.[52]

Uses of the Phenol and Its Derivatives

Phenol belongs to the 50 most abundantly produced chemicals. It is one of the oldest chemical intermediates in organic chemistry that still plays a very important role. Both phenol and its numerous derivatives of natural, synthetic, and semisynthetic origin are basic raw materials for industry.

The applications of phenol and its derivatives can be divided into industrial, nonindustrial, and niche.

The largest single use of phenol in industry is as an intermediate in the production of phenolic resins or related materials in the reaction of polycondensation of phenol with formaldehyde. Thus, produced phenyl aldehyde resins are low cost, thermo set, and versatile. They are used as resin glues for plywood in plywood adhesive, cast resins, molding resins, and varnish resins; moreover, they are applied in construction, automotive, and appliance industries.[2,24]

Alkylphenols introduced on the market in the 1940s have been widely used as paint components, herbicides, pesticides, some nonionic detergents, components of cosmetics, composite materials, and media used for removal of fat from the surface of wool, leather, and metal finishing.[1,21]

Non-anionic detergents are produced by alkylation of phenol to give alkylphenols, which are then subjected to ethoxylation.[21] Alkyl phenols with branched chains are used as supplements in oil lubricants, antioxidants (stop chain reactions), and auxiliary substances in conversion of rubbers and plastics.[1,2,24,25] Chemical activity of alkyl derivatives of phenol as well as antioxidants results from the facility of hydrogen abstraction from the phenol group, initiated by the interaction of oxygen and light.

In order to hinder the process of aging, alkyl derivatives are introduced to many different organic products in amounts of 0.1%-2% wt. For example, they can be found in gasoline, lubricating oils, lipids, polymers and plastics, rubber, soaps, cosmetics, and pharmaceuticals.

The chlorinated phenols that are used in largest quantities include PCP and the trichlorophenol isomers used as wood preservatives. Their main components are cresols known as creosotes.[2]

Higher chloroderivatives of phenol, such as 2,4-dichlorophenoxyacetic acid and 2,4,5-trichlorophenoxyacetic acid act, as selective herbicides. The first of the compounds mentioned is one of the most important herbicides against dicotyledons in crops of monocotyledonous plants (cereals, linen, grass). The second of the above compounds acts against perennial weeds, bushes, and coppices in grasslands. Problems that these compounds impose on the natural environment have prompted gradual limitation on their production. Ammonolysis of phenol in gas phase in the presence of aluminosilicate catalysts (zeolites) is one of the common methods for the synthesis of aniline, which is an intermediate for manufacturing salicylic acid. It is also used to make pharmaceuticals, aromatic esters, synthetic dyes, and food preservatives.[2,3,241

All industrial applications of phenol and its derivatives are too numerous to be mentioned.

The major nonindustrial applications of phenol and its derivatives include the following.

- Medicine; as an antiseptic (water solution of 2%–10% is a very effective antiseptic medium called phenolated water; and in a concentration of 0.2%, it shows bacteriostatic action, while above 2%, it is bactericidal);[1] to relieve itching; as an anesthetic in medicinal preparations (ointments, ear and nose drops, cold sore lotions, throat lozenges, and antiseptic lotions). They are active against a wide range of microorganisms including some fungi and viruses but are only slowly effective against spores. They have been used to disinfect skin, but recently, phenol (Phenolum FP VII) and liquid phenol (Phenolum liquefactum FP IV) have been sporadically used, taking into account their toxicity.[1,2,6,24,25]
- Salicylic acid is an intermediate in production of drugs, flavor esters, dyestuffs, and food preservatives.[2,53]
- Pentachlorophenol can be applied as a fungicide for wood solution of phenol.[1,2,53–55]
- The niche applications of phenol and its derivatives include the following.
- Diphenyl ether is used as material for the production of phenoxathiin. It is an intermediate for polyamide and polyimide, and a processing aid in the production of polyesters. Some polybrominated biphenyl ethers are flame retardants and are used in soap perfumes.[56]
- Phenol is used for embalming bodies for study because of its ability to preserve tissues for extended periods of time.
- Production of cosmetics such as sunscreens,[53] hair dyes, and skin lightening preparations.[54]
- Phenol is used as an exfoliant in cosmetic surgery.
- Phenol and its derivatives are used for phenolization, which is a surgical procedure that serves to treat ingrown nails.
- Phenol is used to denature and remove protein when purifying nucleic acids in molecular biology procedures.
- Phenol and its derivatives are added to polyimide to make antispastic and painkilling drugs used in cancer therapy.[54]
- They are used as slimicides, i.e., substances used to kill slime-producing organisms including bacteria and fungi.[1,2,24,26,55]

Distribution, Transport, and Conversion of Phenol and Its Derivatives in the Biosphere

A phenolic compound occurs in the biosphere in concentrations from trace to high orders of a few milligrams per cubic decimeter or even a few grams per cubic meter in the industrial emissions. Phenol evaporates more slowly than water and can remain in the air, soil, and water for long periods of time if large amounts of it are released at one time or if it is constantly released to the environment from a source.[57,58] Coefficients for conversion of phenol concentration are the following: 1 mg/m^3 = 0.26 ppm, 1 ppm = 3.84 mg/m^3.[2]

In the atmosphere, phenol exists predominantly in the vapor phase.[57,58] No reliable information has been given hitherto on the background concentration of phenol in the atmospheric air far from the sources of its emission. It has been a priori assumed to be low. It has been estimated that the background level of phenol in the atmospheric air is below 1 ng/m^3.[59] Higher concentrations are to be expected above cities, mainly as a result of its emission in gas exhausts of motor vehicles.[50,60,61]

According to estimations, the mean day concentration in urban and suburban atmosphere reaches 0.12 μg/m^3, and in the atmosphere of large urban agglomerations, it can vary within the range of 0.1–8 p.g/m^3.[62] In the atmosphere over the industrial sources of phenol emission, its concentration can be two orders of magnitude higher. In the neighborhood of a phenol resin–producing plant, the concentration of phenol was at a level of 190 p.μg/m^3.[63,64] Still higher values were recorded directly at the workplace in the gases released by iron casting plants; in 1980, the concentration of phenol reached at such a place was 0.8–3.5 mg/m^3.[65,66]

It has been estimated that half-life of phenol in air generally varies depending upon atmospheric conditions, and values ranging from 2.28 to 22.8 hr for reaction with hydroxyl radicals **(*OH and HO$_2$)*** have been reported in literature. It is the reason why a small amount of phenol does not remain in the air for longer than a day.[59]

The presence of phenol was detected in rain and surface water as well as in underground water, but relevant information is scarce. Phenolic compounds react as a weak acid in water, and they are not expected to dissociate in the pH range typical in the natural environment, i.e., 6.5–8.5.[5]

For example, the concentration of phenol dissolved in rainwater collected in Portland, United States, was 0.08–1.2 |μg/L, but the mean concentration was 0.26 μg/L.[67] In diluted water solutions, phenol undergoes conversion to dihydroxybenzenes, nitrophenols, nitrosophenols, and nitrochinone, probably according to the radical mechanism with nitrate(V) ions with the use of hydroxyl radicals and phenoxyl.[68,69] The compound of 2,4,6-trichlorophenol occurs in water of different types as a result of exposure to chlorine and its effect on organic precursors. The mean concentration of chlorophenols in posttreated water varies from 0.003 to 1 μg/dm^3.[69–71] If chlorine is used as a water disinfectant, different chlorophenols are formed,[72] while if chlorine dioxide is used, a variety of *p*-benzoquinones are formed.[73]

In wastewater or sewage, phenol can react with nitric(V) acid to form toxic cyanides.[74]

No information on the content of phenol in the soil has been found; however, its presence in the soil is rather unlikely taking into regard its fast biodegradation and transport in underground water or air.[75] It has been suggested on the basis of some studies[76] that release of phenol from the dry near-surface layer of soil should be rather fast.

The partition (K_{oc}) coefficient for phenol for two types of loamy soil was 39 and 91 dm^3/kg, which implies high mobility of phenol in the soil and its easy penetration to underground water.[60,76,77]

As follows from the logarithm of the phenol dissociation constant pK$_a$ (log 1.28×10^{-10}) in water and in moist soil, phenol occurs in a partly dissociated form, so its transport and reactivity can depend on pH of the environment.[1,2,60,77] Phenol reacts as a weak acid in water, so it is not expected to dissociate in the range typical of the natural environment.[1,2]

Determination of phenol in drinking water performed in the United States has shown that its content is close to 1 lμg/L but usually below the detection limit.[60] Higher content of phenol was found in samples of underground water contaminated with the wastewater left after coal gasification.[60,77]

Because of its short half-life, phenol is not expected to be transported over a great distance in the atmosphere.[50] In soil, phenol is highly mobile, and its mobility, likewise its reactivity, depends on pH of the environment.

In the main elements of the biosphere, the transport of phenol takes place via its washing out by precipitations or by leaching from the soil. It is highly unlikely to be a stable and durable reagent in the natural environment.[1,2,60,76,77] The hitherto state of our knowledge on the harmful effect of phenol in the biosphere does not permit a reliable evaluation of the potential threat it may pose. Nevertheless, in view of the calculated value of the maximum admissible critical concentration of phenol in water ecosystems, e.g., 0.5 μg/L, it cannot be excluded that its presence can be related to some risk for water fauna and flora.[2,76,77]

Photooxidation, photodegradation in air, and biodegradation in water and soil are expected to be the major removal processes of phenol and its derivatives from the biosphere.[78] Phenol can potentially be removed from the atmosphere via photooxidation by reaction with hydroxyls and nitrate radicals, photolysis, and wet and dry deposition.[6,50,79] The majority of phenol from the atmosphere undergoes photodegradation to dihydroxybenzenes, nitrophenols, and products of ring cleavage, and the rest is washed out with precipitates.[50] Biodegradation is a major process for the removal of phenol from surface waters. Phenol generally reacts with hydroxyl and peroxyl radicals and singlet oxygen in sunlit surface waters.[5,50,76] The half-life of phenol in the air evaluated in photochemically active conditions in a smog chamber was 4–5 hr and was consistent with the value determined on the basis of the rate of its reaction with hydroxyl radicals.[50,78,79] In natural sunlit water reservoirs, phenols usually react with hydroxyl radicals formed as a result of photochemical re- actions.[77,79]

Phenols easily undergo biodegradation in water if they do not occur in the concentrations toxic for microorganisms, as the latter play the main role in phenol degradation in the soil, water, water sediment, and bottom sediment.[76–79] The amount of bacteria capable of phenol degradation makes up a very small percentage of the total population of bacteria present in the soil.[80] Chemical composition of the products of bacteria-assisted degradation of phenols depends on the environmental conditions. In aerobic conditions, the products include carbon dioxide,[81] while in anaerobic conditions, they include carbon dioxide and/or methane.[82]

Phenol can undergo degradation in the free form as well as in the adsorbed form, in the soil or sediment, although the presence of the sorbent decreases the rate of its biodegradation.[76–79]

Environmental Impact of Phenol and Its Derivatives

For the general population, the most important sources of exposure to phenol are those coming from the air (emission from motor vehicles and product of photooxidation of benzene) and those from cigarette smoking and consumption of smoked food products.[50] The available information on the degree of exposure to phenol is insufficient. To evaluate the risk of harmful exposure, the value of total daily intake by a man weighing 70 kg was estimated to be 0.1 mg/kg of body mass per day.[2,6] The danger of poisoning by phenol vapor has been known for a long time, although no lethal cases have been reported. 1,5,6,78] The symptoms of poisoning by phenol inhalation include anorexia, loss of body mass, headache and vertigo, hypersalivation, and dark-colored urine. The risk of poisoning by intake of poisoned water or food is highly unlikely because of the very unpleasant taste and smell of phenol.[2,77] There is some anxiety over the reports on the possible genotoxic effect of phenol and there are controversies over its carcinogenic effect.[83–85]

The available information on the health effects of phenol exposure to humans is almost exclusively limited to case reports of acute effects of oral exposure.[86]

Most of the simple phenolic compounds have similar features and are toxic to aquatic organisms. The presence of a hydroxyl group intensifies their toxicity, which is why phenol is more toxic than benzene. 2,5,86] The acute toxicological effects of phenol are predominantly upon the central nervous system, and death can occur as soon as one-half hour after exposure. Acute poisoning by phenol can also cause severe gastrointestinal disturbances, kidney malfunction, circulatory system failure, lung edema, and convulsions. Fatal doses of phenol may be adsorbed through the skin. Key organs damaged by chronic exposure to phenol include the spleen, pancreas, and kidney. The toxic effects of other phenols resemble those of phenol.

Water or wastewater treatment by exposure to chlorine (Cl_2) leads to formation of chlorinated phenols, in particular, PCP, responsible for the revolting smell and taste. Although exposure to chlorophenols has been correlated with liver malfunction and dermatitis, polychlorinated dibenzodioxins also may have caused some of the observed effects.[2,6,86] Moreover, phenol and chlorophenols occurring in water are bioaccumulated by water fauna, which is responsible for the unpleasant smell and taste of fish meat.[5]

The most important organs involved in phenol metabolism are the liver, lungs, and mucous membrane of alimentary track; their involvement depends on the mode of exposure and the dose absorbed. Results of in vivo and in vitro studies have proven the covalent bonding of phenol with proteins from human tissues and blood plasma. Some metabolites of phenol also react with proteins.[1,2,86] The main route of phenol elimination from humans and animals is with urine. The rate of elimination with urine depends on the mode of exposure and dose absorbed. Phenol is also eliminated with stool and exhaled air.

Identification and Quantification of Phenol and Its Derivatives

Phenol gives violet coloration with ferric chloride solution (the test reaction of phenol) due to the formation of colored iron complex, which is a characteristic of the existence of keto-enol tautomerism in phenols:

$$6C_6H_5 + FeCl_3 \rightarrow 3H^+Fe\left[O\left(C_6H_4-\right)_6\right]^{3-} + 3HCl \qquad (5)$$

In many countries, the content of phenol and its derivatives is controlled and standardized in all elements of the ecosphere, i.e., atmosphere, surface water, drinking water, soil, and food products.[87]

Development of analytical methods based on physicochemical phenomena and processes taking place in the natural environment and in living organisms has brought a decrease in the detection level and increase in accuracy of measurements. The range of applicability of analytical methods is related not only to the properties and type of substance studied but also to the selectivity and reproducibility of the method for isolation and enrichment of the analyte. From the point of view of routine analysis, the development of selective and reproducible procedures is difficult. Therefore, the application of coupled methods (off-line and/or on-line) is recommended. Because of their diversity and richness of forms of presence, phenolic compounds create many analytical problems, especially in routine analyses.[88]

Phenols react with 4-aminoantipyrine at pH 10 in the presence of potassium ferrocyanide, forming antipyrine dye, which is extracted into pyridine and measured at 460 nm.[89] This method is, however, charged with large error for the following reasons:

- Phenols with para substituents do not react with a reagent (the exceptions are para derivatives substituted with carboxyl, methoxyl, sulfonic groups, or halides).
- The yields of coupling with 4-aminoantipyrine for phenol, cresoles, xylenols, and other relevant compounds are significantly different.
- Maxima of absorption of the reaction products with the reagent are not the same for all phenols; some organic compounds that are not phenols also react with 4- aminoantipyrine.

The above-mentioned drawbacks, the lack of possibility to predict the qualitative and quantitative composition of phenolic compounds in a water sample studied, and problems with the choice of standard mixtures have prompted the introduction of the so-called phenol index describing the total concentration of phenols determined spectrophotometrically and expressed in milligrams per liter of phenol. It is in agreement with the international standard (ISO 6439) recommending also a procedure of colorimetric determination of phenolic compounds with the help of 4-aminoantipyrine, in which the phenol index is defined as the total concentration of phenolic compounds expressed in milligrams per liter phenol. The limit of determination of the phenol index is 0.010 milligrams per liter for the spectrophotometric method with 4-aminoantipyrine (nonspecific method) and 0.002 mg/L for the extraction-spectrophotometric method (single extraction with chloroform).[90]

From water and wastewater, phenolic compounds are separated by distillation with steam from an acidic solution, and then they are subjected to a reaction with 4-aminoantipyrine (1-fenylo-2,3-dimetylo-4-aminopirazolon) in an alkaline environment at a pH 9.8 in the presence of potassium ferrocyanide as an oxidant. This reaction leads to formation of an indophenolic dye of green-yellow to orange color depending on the concentration of phenols, extracted with chloroform. The content of phenols is determined spectrophotometrically by measurement of absorbance of the colored solution at the analytical wavelength, usually $\lambda = 460$ nm.[89]

The diversity of analytical tasks implies the need to choose the most appropriate analytical method for a specific problem. As follows from the survey of methods used for determination of phenols, presented by Thielemann,[91] all of them are used for the purposes.

Determination of trace amounts of phenols is realized by colorimetric and chromatographic method, while determination of phenols with content above 1% is performed by chemical bromatometric and iodometric methods.[89]

Conclusion

Phenol is one of the first compounds inscribed into the list of priority pollutants by the U.S. Environmental Protection Agency (EPA). It is commonly used in different branches of industry, including chemical production of alkylphenols, cresols, xylenols, phenolic resins, aniline, and other compounds.

Phenol is produced as an intermediate in the preparation of other chemicals and can be released as a by-product or contaminant. The economic importance of phenol and its derivatives has been significant since coal tar was used as a rubber solvent in the beginning of the 19th century. The current use of this group of compounds as pure products includes the chemical synthesis of plastics, synthetic rubber, paints, dyes, explosives, pesticides, detergents, perfumes, and drugs.

These compounds are used mainly as mixtures in solvents and constitute a variable fraction of gasoline. They have been used extensively for a wide variety of applications. They have been found to be among the most economical and effective surfactants, e.g., phenol ethoxylate and alkyl phenol ethoxylates. Production of their anionic derivatives has been severely restricted because they or their degradation products have been found to be estrogenic.

There is increasing environmental concern worldwide regarding the disposal of wastewater containing nonbiodegradable organic compounds. Since most pollutants do not respect national borders, a world effort to monitor their movement and to develop tools to prevent them from polluting environmental components or to remediate consequent pollution is desirable.

Phenol has antiseptic and germicidal properties, which increase to a maximum as the length of an alkyl side-chain substituted reaches about six carbon atoms. The lowest member, phenol, is highly toxic and caustic, but these properties progressively diminish with the higher members of the series. Polyhydric phenols are still markedly toxic but less caustic than the monohydric compounds (such as phenol).

The antiseptic activity of phenol is a property that accounts for much of its industrial use. In dilute solutions (of 1%–2%), it also finds application as an agent against skin itching. Undiluted phenol is highly corrosive to mucous membranes and skin, and it is considered a nerve poison. It may enter the human system by absorption through skin, oral ingestion, and vapor inhalation.

Phenol and three cresol isomers are of about the same order of toxicity and produce identical symptoms in poisoned animals.

Chemical and pharmaceutical industries are large users of phenol for conversion to many different products. The salicylates (carboxyphenols) are starting materials for the preparation aspirin and flavor.

Chlorophenols and their derivatives find application as fungicides, bactericides, and selective weed killers. Alkyl phenols make an important group synthetic tanning agent, and triphenyl phosphate is a plasticizer. Phenolic derivatives are among the most important contaminants in the environment. These compounds are used in several industrial processes to manufacture pesticides, explosives, drugs, and dyes. They are also used in the bleaching process of paper manufacturing. Apart from these sources, phenolic compounds have substantial applications in agriculture as herbicides, insecticides, and fungicides. However, phenolic compounds are not only generated by human activity but also formed naturally, e.g., in the process of decomposition of leaves or wood.

As a result of these applications, they are found in soils and sediments, and this often leads to wastewater and groundwater contamination. Owing to their high toxicity and persistence in the environment, both the U.S. EPA and the European Union have included some of them in their list of priority pollutants.

Current standard methods of phenolic compound analysis in water samples are based on liquid–liquid extraction, while Soxhlet extraction is the most used technique for isolating phenols from solid matrices.

Phenols and their derivatives are common in the natural environment. These compounds are used as the component of dyes, polymers, drugs, and other organic substances. The presence of phenol substances in ecosystems is also related with production and degradation of numerous pesticides and the generation of industrial and municipal sewages. Some phenols are also formed during natural processes. These compounds may be substituted with chlorine atoms, nitrated, methylated, or alkylated. Both phenols and catechols are harmful ecotoxins.

Toxic action of these compounds stems from unspecified toxicity related to hydrophobicity and also to generation of organic radicals and reactive oxygen species. Phenols and catechols reveal peroxidative capacity, are hematoxic and hepatoxic, and provoke mutagenesis and carcinogenesis toward humans and other living organisms.

References

1. Rappoport, Z. *The Chemistry of Phenols*; Wiley-Inter- science: New York, 2003.
2. Tyman, J.H.P. *Synthetic and Naturals Phenols;* Elsevier Science: New York, 1996.
3. McTavish, D. *Joseph Lister*; Bookwright Press: New York, 1992.
4. Breinholt, V. Desirable versus harmful levels of intake of flavonoids and phenolic acids. In *Natural Antioxidants and Anticarcinogens in Nutrition, Health and Disease*; Kum- pulainen, J.T., Salonen, J.T., Eds.; The Royal Society of Chemistry: London 1999; 93–99.
5. Thurman E.M. *Organic Geochemistry of Natural Waters*; Martinus Nijhoff/Dr W. Junk Publishers: Dordrecht, 1985.
6. Manahan, S.E. *Toxicological Chemistry and Biochemistry,* 3rd Ed.; CRC Press LLC: New York, 2003.
7. Verschueren, K. *Handbook of Environmental Data on Organic Chemicals*, 2nd Ed.; van Nostrand Reinhold Company: New York, 1983.
8. Spoelstra S.F. Degradation of tyrosine in anaerobically stored piggery wastes and in pig feces. Appl. Environ. Microbiol. 1978, 361–638.
9. Boskou, D. *Natural Antioxidant Phenols: Source, Structure-Activity, Relationship, Current Trends in Analysis and Characterisation*; Research Signpost: New York, 2006.
10. Glade, M.J. Dietary phytochemicals in cancer prevention and treatment. Book Rev. Nat. Rev. 1997, *13* (4), 394–397.
11. Friedman, M. Chemistry, biochemistry and dietary role of potato polyphenols. J. Agric. Food Chem. 1997, *45,* 1523–1540.
12. Kroon, P.A.; Wiliamson, G. Hydroxycinnamates in plants and food: Current and future perspectives. J. Sci. Food Agric. 1999, *79,* 335–361.
13. Robinson, T. *The Organic Constituents of Higher Plants,* 4th Ed.; Cordus Press: North Amherst, MA, 1980.
14. Duke J.A. *Handbook of Physicochemical Constituents of GRAS Herbs and Other Plants*; Ed. Academic Press; New York, 1992.

15. Hedges, J.I.; Mann, D.C. The characterization of plant tissues by their lignin oxidation products. Geochim. Cosmochim. Acta 1979, *43,* 1803–1807.
16. Herman, K. Occurrence and content of hydroxycinnamic acid and hydroxybenzoic acid compounds in foods. Crit. Rev. Food Sci. Nutr. 1989, *28,* 315–347.
17. Hubble, B.R.; Stetter, J.R.; Gebert, E.; Harkness, J.B.L.; Flotard, R.D. Experimental measurements of emissions from residential wood-burning stoves. In *Residential Solid Fuels: Environmental Impacts and Solutions;* Cooper, J.A., Malek, D., Eds.; Oregon Graduate Center: Beaverton, OR, 1981; 79–138.
18. Hardy, G. Nutraceuticals and functional foods: Introduction and meaning. Nutrition 2000, *16* (7–8), 688.
19. Aherne, S.A.; O'Brien, N. Dietary flavonols, chemistry, food content, occurrence and intake. Nutrition 2002, *18*, 75–81.
20. Stevenson, F.J. *Humus Chemistry, Genesis, Composition, Reactions;* John Wiley & Sons: New York, 1994.
21. Muscola, A.M.; Soidari M. *Soil Phenols;* Nova Science Publishers Inc.: New York, 2010.
22. Hessen, D.O.; Tranvik, L.J. *Aquatic Humic Substances: Ecology and Biogeochemistry;* Springer Verlag: Berlin, 1998.
23. Larsen, J.W.; Nadar, P.A.; Mohammadi, H.; Montano P.A. Spatial distribution of oxygen in coals. Development of a tin labelling reaction and Mossbauer studies. Fuel, 1982, *62,* 889–893.
24. Weber, M.; Weber, M.; Kleine-Boyman, M. Phenol. In *Ull- mann's Encyclopedia of Industrial Chemistry;* Wiley-VCH: New York, 2004.
25. Gardziela, A.; Pilato, L.; Knop, A. *Phenolic Resins: Chemistry, Applications, Standardization, Safety and Ecology*; Springer Verlag: New York, 2000.
26. Ryser, S.; Ulmer, G.; Etude de la pollution par les resines syn- thetiques utylisees en fonderie. Fonderie 1980, 33, 313–324.
27. *RIVM Criteria Document: Phenol, Bilthoven, the Netherlands*, National Institute of Public Health and Environmental Protection (Document No. 738513002); 1986.
28. Kuwata, K.; Uebori, M.; Yamazaki, Y. Determination of phenol in polluted air as p-derivative nitrobenzene azophenol by reversed phase high performance liquid chromatography. Anal. Chem. 1980, *52* (6), 867–860.
29. Verschueren, K. *Handbook of Environmental Data on Organic Chemicals,* 2nd Ed.; van Nostrand Reinhold Company: New York, 1983.
30. Den Boeft, J.; Kruiswijk, F.J.; Shulting, F.L. Air pollution by combustion of solid fuels; The Hague Ministry of Housing, Physical Planning and Environment (Publication Lucht No. 37), 1984.
31. Groenen, P.J. Components of tobacco smokes. Nature and quantity, potential influence on health; Zeist, the Netherlands, CIVO-TNO Institute Report No R/5787, 1978.
32. Gosselin, R.E.; Smith, R.P.; Hodge, H.C. Phenol. In *Clinical Toxicology of Commercial Products,* 5th Ed.; Wiliams & Wilkins: Baltimore, 1984; III-344–III-348.
33. Pekari, K.; Vainotalo, S.; Heikkila, P.; Palotie, A.; Luotamo, M.; Rihimaki, V. Biological monitoring of occupational exposure to low levels of benzene. Scand. J. Work Environ. Health 1992, *18,* 317–322.
34. Ahlborg, U.G.; Thunberg T.M. Chlorinated phenols: Occurrence, toxicity, metabolism, and environmental impact. CRC Crit. Rev. Toxicol. 1980, 7, 1–35.
35. Carrey, F.A. Chlorophenol. In *Encyclopaedia Britannica Online;* Encyclopaedia Britannica, 2001.
36. Mycke, B. Preliminary results on free dissolved phenolic compounds in natural Waters. In *SCOPE/UNEP Transport of Carbon and Minerals in Major World Rivers, 52;* Degenes, E.T., Ed.; University of Hamburg: Hamburg, 1982; 571–574.
37. Roupe, K.A.; Remsberg, C.M.; Yanez J.A.; Davies, N.M. Pharmacometrics of stilbenes: Segueing towards the clinic. Curr. Clin. Pharmacol. 2006, *1,* 81–101.
38. Begon, M.; Harper, J.L.; Townsend C.R. *Ecology, Individuals Populations and Communities*; Black Science Ltd.: Edinburgh, 1996.

39. Smith; M.B.; March, J. *Advanced Organic Chemistry. Reactions, Mechanisms, and Structure*, **6**th Ed.; Wiley-Inter- stice: New York, 2007.
40. Rice-Evans, C. *Screening of Phenolic and Flavonoids for Antioxidant Activity, Antioxidant Food Supplements in Human Health*; Academic Press: New York, 1999.
41. Golovanow, I.B.; Zhenodarova, S.M. Structure-property correlations. XV. Properties of phenol derivatives. Russ. J. Gen. Chem, 2008, 73 (10), 1603–1607.
42. Herman, K. Occurrence and content of hydroxycinnamic acid and hydroxybenzoic acid compounds in foods. Crit. Rev. Food Sci. Nutr. 1989, *28,* 315–347.
43. Dragsted, L.O. Antioxidant actions of polyphenols in humans. Int. J. Vitam. Natur. Res. 2003,73, 112–119.
44. Budryn, G.; Nebesny, E. Phenylacids. Properties, occurrence in plants and metabolic conversion. Bromat. Chem. Toksycol. 2006, *102,* 547–554.
45. Foley, S.; Navartman, S.; MvGarvey, D.J.; Land, E.J.; Truscott G.; Rice-Evans, C.A. Singlet oxygen quenching and the redox properties of hydroxycinnamic acids. Free Radic. Biol. Med. 1999, *26* (9/10), 1202–1208.
46. Kirk, R.E.; Othmer, D.F. *Encyclopedia of Chemical Toxicology,* 3rd Ed.; John Wiley and Sons: New York, 1980; Vol. 17, 373–379.
47. Bolling, F.J.; Decker, K.H. *Phenols Resin;* Kunstshoffe: Berlin, 1980.
48. Ying, G.G.; Wiliams, B.; Kookana R. Environmental fate of alkylphenol and alkyl ethoxylates—A review. Environ. Int. 2002, *28*, 215.
49. Budavari, S. *The Meck Index. An Encyclopedia of Chemical Drugs and Biologicals*; Whitehouse Station: New York, 1996.
50. Seinfeld, J.H.; Pandis S.N. *Atmospheric Chemistry and Physics;* John Wiley & Sons, Inc.: New York, 1996.
51. van Gemert, L.J. *Flavour Thresholds. Compilations of Flavour Threshold Values in Water and Other Media*, (second edition) Ed. Oliemans Punter & Partners BV; the Netherlands, 2011.
52. US EPA. Risk Management Air Research; 1982.
53. DeSelms, R.H. *UV-Active Phenol Ester Compounds*; Eni- gen Science Publishing: Washington, DC, 2008.
54. Svobodova, A.; Psotova, J.; Walterova D. Natural phenolics in the prevention of UV-induced skin damage. A review. Biomed. Papers 2003, *147* (2), 137–145.
55. Fiege, H.; Voges H.M.; Umemura S.; Iwata, T; Miki, H.; Fujita, Y.; Buysch, H.J.; Garbe, D.; Paulus, W. Phenol derivatives. In *Ullmann's Encyclopedia of Industrial Chemistry*; Wiley-VCH: Weinheim, 2000.
56. Ueoda, M.; Aizawa T.; Imai, Y. Preparation and properties of polyamides containing phenoxathiin units. J. Polym. Sci. Chem. Ed. 1977, *15* (11), 2739–2747.
57. Eisenreich, S.J.; Looney, B.B. Thornton, J.D. Water solubility enhancement of pyrene in the presence of humic substances. Environ. Sci. Technol. 1981, *15,* 30.
58. Schnitzer, M.; Khan S.U. *Humic Substances in the Environment;* Marcel Dekker: New York, 1972; 1–7.
59. Web-based Archive of the Dutch National Institute for Public Health and the Environment (RIVM), Final report 2002.
60. Howard, P.H. *Handbook of Environmental Fate Exposure Data of Organic Chemicals*; Lewis Publishers: New York, 1991; Vol. 1.
61. Kawamura, K.; Kaplan, I.R. Stabilities of carboxylic acids and phenols in Los Angeles rain waters during storage. Water Res. 1990, *24* (11), 1419–1423.
62. Brodzinski, R.; Singh H.B. Volatile organic chemicals in the atmosphere: An assessment of available data, EPA 600383o27(A); US Environmental Agency, Office of Research and Development: Research Triangle Park, NC., 1983.
63. Tesarova, E.; Paczkowa, W. Gas and high-performance liquid chromatograph of phenols. Review paper. Chromatographia 1983, *17* (5), 269–284.

64. Grosjean, D. Atmospheric fate of toxic aromatic compounds. Sci. Total Environ. 1991, *100,* 367–414.
65. Kuwata, K.; Uebori, M.; Yamazaki, Y. Reversed-phase liquid chromatographic determination of phenols in auto exhaust and tobacco smoke as *p*-nitrobenzenephenol derivatives. Anal. Chem. 1981, *53* (9), 1531–1534.
66. Bruno T.J.; Svoronos P.D.N. *Basic Tables for Chemical Analysis,* 3rd Ed.; CRC Press: Boca Raton, 2011.
67. Rogge, W.F. Molecular tracers for sources of atmospheric carbon particles: Measurements and model predictions. Ph.D. Thesis, California Institute of Technology: Pasadena, 1993.
68. Niessen, R.; Lenoir, D.; Boule, P. Phototransformation of phenol induces by excitation of nitrate ions. Chemosphere 1988, 17, 1977–1984.
69. Louw, R.; Santoro D. Comment on formation of nitroaromatic compounds in advanced oxidation processing: photolysis versus photocatalysis. Environ. Sci. Technol. 1999, *33*, 3281.
70. Kunte, V.H.; Slemrova, J. Gaschromatographische und massenspektrometrische Identifizierung phenolischer substanzen aus ober-flachenwassern. Zeitschrift fuhr Wasser und Abwasser Forschung 1975, *8*, 176–182.
71. World Health Organization (WHO). *Guidelines for Drinking Water Quality,* 2nd Ed.; Geneva, 1984, Vol. 1.
72. Knoevenagel, K.; Himmelreich, R. Degradation of compounds containing carbon atoms by photooxidation in the presence of water. Arch. Environ. Contam. Toxicol. 1976, *4* (3), 324–33.
73. Jarvis, S.N; Straube, R.C.; Wiliams, A.L.J.; Bartlett C.L.R. Illness associated with contamination of drinking water supplies with phenol. Brit. Med. J. 1985, *20,* 1800.
74. Tratnyek, P.G.; Hoigne J. Kinetics of reactions of chlorine dioxide (OClO) in water (II). Quantitative structure–activity relationships for phenolic compounds. Water Res. 1994, *28* (1), 57–66.
75. Nakagawa, N.; Mizukoshi, O.; Maeda, Y. Sonochemical degradation of chlorophenols in water. Ultrason. Sonochem. 2000, 7 (3), 115–120.
76. Faust, S.D.; Stutz, H.; Aly, O.M.; Andersen, P.W. Recovery, separation, and identification of phenolic compounds from polluted waters. Part I—Occurrence and distribution of phenolic compounds of the surface and ground waters of New Jersey. Geological Survey Open- File Report, 1970.
77. Simonow, L.; Sargsyan, V. *Soil Chemical Pollution, Risk Assessment, Remediation and Security*; Springer: Dordrecht, 2007.
78. Yong, R.N.; Mohamed, A.M.O.; Warkentin, B.P. *Principles of Contaminant Transport in Soils*; Elsevier: New York, 1992.
79. Gilman, A.P. *Chlorophenols and Their Impurities: Health Hazard Evaluation*; Environmental Health Directorate, Health Protection Branch: New York, 1988.
80. Schwarzenbach, R.P.; Gschwend, P.M.; Imboden, D.M. *Environmental Organic Chemistry,* 2nd Ed.; Wiley-Interstice: New York, 2003.
81. Hickman, G.T.; Novak, J.T. Relationship between subsurface biodegradation rates and microbial density. Environ. Sci. Technol. 1989, *23,* 525–532.
82. Boyd, J.T.; Carlucci, A.F. Degradation rates of substituted phenols by natural populations of marine bacteria. Aquat. Toxicol. 1993, *25* (1), 71–82.
83. Dobbins, D.C.; Thornton, J.; Jones, D.D.; Fedale, T.W. Mineralization potential for phenol in subsurface soils. J. Environ. Qual. 1987, *16* (1), 54–58.
84. Bukowska, B., Kowalska S. Phenol and catechol induce prehemolytic and hemolytic changes in human erythrocytes. Toxicol. Lett. 2004, *152* (1), 73–84.
85. Bruce, R.M.; Santodonado, J.; Neal, M.W. Summary review of the health effects associated with phenol. Toxicol. Indust. Health 1987, 3, 535–568.
86. Oikawa, S.H.; Hirosawa, K.; Kawanishi S. Site specificity and mechanism on oxidative DNA damage induced by carcinogenic catechol. Carcinogenesis 2001, 22, 1239–1245.
87. WHO. IPCS Environmental Health Criteria for Phenol (161). First draft prepared by MS G.K. Monitzan; WHO: Finland, 1994.

88. Thompson, R.D. Determination of phenolic disinfectant agents in commercial formulations by liquid chromatography. J. AOAC Int. 2001, *84* (3), 815.
89. Buszewski, B.; Buszewska, T.; Szumski, M.; Siepak, J. Simultaneous determination of phenols and polyaromatic hydrocarbons isolated from environmental samples by SFE–SPE-HPLC. Chem. Anal. 2003, *48,* 13–25.
90. *Standard Methods for Examination of Water and Wastewater.* APHA, AWWA, WPCF: Washington, DC, 1992.
91. International Standard ISO 6439. Water quality—Determination of phenol index—4-Aminoantipyrine spectrometric methods after distillation, Prepared by Technical Committee ISO/TC 147; Geneva (Switzerland) 1990, p. 7.
92. Thielemann, H. Thin layer chromatographic separation of phenol, cresols, dimethylphenols and naphthols. Die Pharmazie 1970, *25* (5), 365–366.

31

Phosphorus: Agricultural Nutrient

John Ryan,
Hayriye Ibrikci,
Rolf Sommer, and
Abdul Rashid

Introduction

Never in the history of mankind have there been such widespread societal concerns about the capacity of our planet to sustain its projected population growth. This concern was first widely articulated in the late 19th century by Thomas Malthus. However, his prognosis of widespread starvation has been challenged by advances in agricultural technology and medicine. These revolutionary developments led to not only a better-fed population but also a vastly expanded population in the past century. While population growth has stabilized in developed countries, it is continuing unabated in many areas of the lesser-developed world. World population is now projected to increase from its current 6.8 billion people to over 9 billion by midcentury.[1] In addition to such increases, world food production will have to be doubled as a result of increased affluence in some developing countries such as India and China.[2] The threat of the "population monster" that Norman Borlaug, Nobel laureate and father of the Green Revolution, railed against is now more credible,[3] with current predictions of a world population of more than 9 billion people by the middle of the century.

The implications for mankind of a more crowded world, where the earth's natural resources are increasingly threatened and even the future of mankind is called into question, are both ominous and daunting.[4,5] The "grand challenges" that our society faces embrace all aspects of agriculture and the environment. A litany of such challenges includes both biophysical (land degradation, water scarcity, loss of biodiversity) and socioeconomic (population growth, poor nutrition and health care, poverty, inadequate research investment and information infrastructure) difficulties. Looming large over these concerns are the implications of climate change, with areas of the world such as the Mediterranean and much of Africa likely to become drier and others to endure climate-related stresses.[6] While much research attention is now focused on adaptation strategies, the predicted change scenario is likely to exacerbate the world's already precarious food production capacity. Such developments have posed an extraordinary challenge to the global scientific community[7] and especially the international network of agricultural research centers.[8] Chemical fertilizers are vital to world food production and food security,[9] with over half of global output attributed to applied fertilizer nutrients and future food supplies being even more dependent on fertilizers.[10] Consequently, fertilizer management, in a sustainable manner that protects the environment and the natural resource base, is central to the issue of balancing the world's food supply–demand equation.

Among the nutrients that are essential to plants but potentially damaging to the environment are nitrogen (N) and phosphorus (P). Large amounts of N are used in modern agriculture; in fact, N fertilizer is the driver of the phenomenal increases in global crop production, especially cereals, in the past half century.[11] Data on current fertilizer N suggest that the trend in N use is set to continue, especially in countries such as China and India, which are now the leading producers of wheat in the world. However, as N is either applied as nitrate (NO_3^-) form as a fertilizer or rapidly converted biologically to this mobile form of N, losses of N from agricultural environments through leaching and runoff became associated with high N fertilizer use.[11,12] Enrichment of water sources with NO_3 not only represents economic losses in terms of fertilizer costs but also causes a potential health hazard in drinking water.[13] More widespread effects of excess N were manifested in terms of water quality only when combined with P,[14] an element that was initially thought to be immobile and subject to minimal loss from the point of application in farmers' fields. Gradually, the catalytic effect of P combined with N and eutrophication or the development of anoxic conditions in water bodies were recognized by agricultural scientists.[15] This brings us to a consideration of the role of P in agriculture and the inadvertent effects of its misuse on the environment,[16] as well as management strategies, which were recognized by agricultural scientists.[17] A brief overview of the various aspects of P in soils and the relationship of P to crop production is a prerequisite to any discussion of P in relation to the environment.

Perspective on Phosphorus

As with carbon and N, the overall nature of P in terrestrial and aquatic ecosystems and the atmosphere can best be described as a cycle, which indicates changes between one state and another, i.e., a state of flux. Various depictions of such a cycle have been made, but the one adapted by Tunney et al.[15] gives a good representation of the inputs and outputs involving P and the relationship with the various P phases in nature (Figure 1). As plant uptake occurs from the soluble P fractions—and as losses occur from that fraction as well—solution P is a central component of the P cycle and is closely linked to microbial P. Inputs to the system are primarily fertilizers (and plants) to the solution phase and manures to the pool involving microbial forms of P; both forms are being replenished by or contributing to somewhat less soluble inorganic and inorganic phases, i.e., labile or moderately labile forms, and constitute the more rapidly changing part of the P cycle.

The behavior of P in nature is also influenced by very slow reactions that occur over several years and do not have an immediate impact during one cropping season. The slow inorganic component of the cycle involves primary and secondary P-bearing minerals that can be weathered to increase solution P; some inorganic forms (occluded) have no direct effect on soluble P. Similarly, some organic P forms have little immediate effect on the more reactive P phases in the cycle. While the depiction in Figure 1 does not specifically indicate the avenue of losses, the soluble P phase is most susceptible to loss by leaching and runoff in addition to the actual transport of P-enriched soil particles by erosive processes. The fundamental problem is related to excessive inputs to the system in terms of chemical fertilizers and manures, thus exceeding the crop's needs and the soil's capacity to retain the P against loss.

No element can rival P in terms of its chemical complexity, its dynamic equilibrium, and the diversity of inorganic and organic P compounds. The addition of P fertilizers to soils and its uptake and

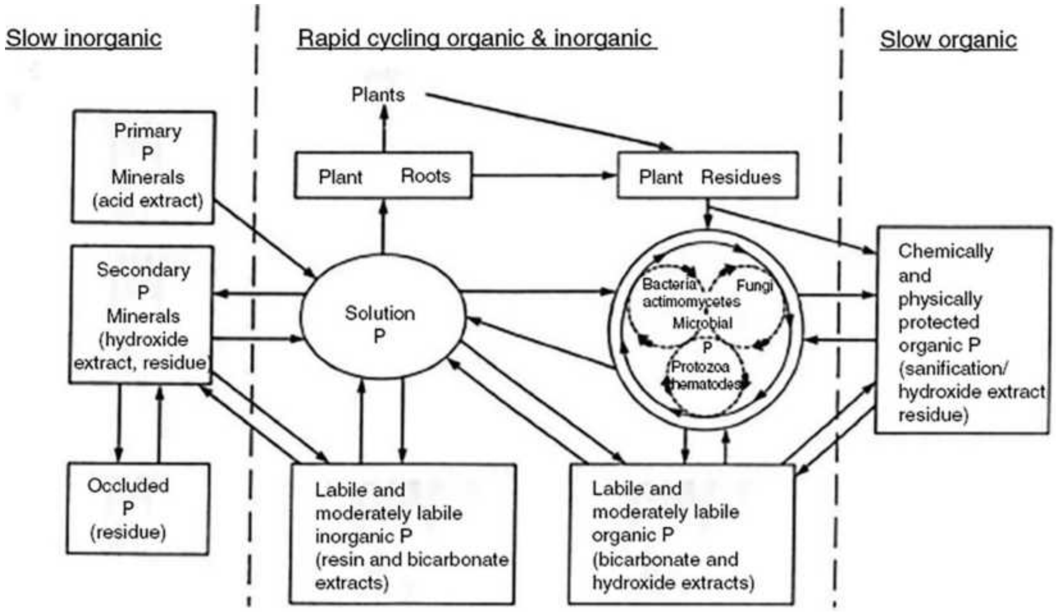


FIGURE 1 The soil P cycle in nature and its reaction phases.

utilization by crops pose further challenges to soil and crop scientists. Much has been written about P from the basic and applied perspectives, especially during the past half century. The thousands of research papers dealing with various aspects of P have gradually revealed a better understanding of this tantalizing element. Progress in P research has been documented in broad-based discipline reviews[18–20] in addition to regional[21–22] and country-level[23–24] reviews. While such references are selective and subjective, they broadly capture the slow but inexorable unraveling of the many aspects of P in soils, fertilizers, and crops.

Shifting Emphasis

A brief survey of the recent literature on P indicated a revision of concepts related to P availability and a shift in emphasis from production-related P research to that related to the environment. For instance, while the early P review of Larsen[18] did not refer to the environment aspect of P use, the wide-ranging P review of Khasawneh et al.[20] addressed the issue in only one brief chapter[25] of the 29 chapters in the volume. That review addressed the emerging concerns of P in the environment, geological influences on soluble P, as well as agricultural losses of P through surface flow and sediment transport; P losses from animal wastes and urban runoff were also briefly addressed.

While the subsequent comprehensive review of P still focused on the agricultural aspect of P,[26] it devoted a major part of the volume (five chapters) to P in relation to the environment. The rapidly expanding awareness of the environmental implications of P led to the establishment of a series of international workshops with focus on the environment.[15] As an illustration of the current research emphasis on P in the environment, keynote addresses at the recent meeting of the workshop series[27] dealt with issues such as P mobilization at the plot and field scales, field-scale indices of P loss, P dynamics and impact in water bodies, and implementation of mitigation options. Reviews such as that of Delgado and Scalenghe[16] focused on Western Europe as one of the world's hot spots for P pollution, where the consequence of overuse of P fertilizers and animal manures had attracted broad societal and legislative concern.

Phosphorus: Potentially Harmful Effects

Though vital to all forms of life on earth, P has no known toxic effects on humans or animals unlike some other elements. The potential effect of P on the environment is related to an excess of P in the terrestrial ecosystems with a carryover to the aquatic environment. Eutrophication of water bodies involves the excessive growth of undesirable algae and aquatic weeds and a consequent depletion of such organisms.[28] The process of eutrophication is defined as an increase in the fertility status of natural waters that causes accelerated growth of algae or water plants. Eutrophication is due not solely to P but to a complex interaction between N, P, environmental conditions (temperature, salinity, light), and the physical and hydrologic characteristics of surface waters, i.e., streams, lakes, and estuaries.[29] Under natural conditions, growth of aquatic organisms is limited by the normally low levels of P in the water. External inputs of P from urban wastewater, surface runoff, or subsurface flow can upset the delicate P balance in water bodies and stimulate growth of aquatic organisms to ecologically undesirable levels.

Critical levels for soluble and total P associated with eutrophication are in the order of 10 and 20 mg L^{-1},[17] while others have indicated a range of values from 20 to 100 mg L^{-1},[29] which is still considerably less than such thresholds for N. The main driver of eutrophication is depletion of dissolved oxygen in the water due to the excessive growth of aquatic organisms and microbial decomposition of such biota. This leads to a chain reaction that causes a deterioration in the water's ecology, e.g., decreased light penetration, surface algal scums, foul odors, impeded water flow, and increased turbidity and sedimentation. Associated with such pollution is the occurrence of surface blooms of cyanobacteria that can reduce water palatability and even kill livestock and pose a threat to human health. Advanced eutrophication of surface water thus has serious implications for fisheries, recreation, and human or animal consumption.

The primary cause of eutrophication is the P that originates from agricultural land as nonpoint source or from urban and industrial effluents. The problem of P-induced pollution was first noted in the Great Lakes region of North America in the 1960s. The phenomenon was later seen in P-enriched lakes in Western Europe. Initially, the problem was attributed to point sources, i.e., sewage-derived inputs. Control measures included sewage diversion, removing P from sewage effluents, and reducing P contents of detergents. However, the persistence of eutrophication in many water bodies following the control of plant-source P emissions led to a focus on loss from agriculture as the primary cause of continued eutrophication. While the problem may appear to lend itself to a simple solution of avoiding use of excess P fertilizer or P-containing manures, Tunney et al.[15] argued that the process of P transport is anything but simple as it depends on soil properties, flow pathways, and bioavailability. Development of a strategy to control or mitigate P losses from fertilized fields or a watershed requires an understanding of the processes and mechanisms that govern P loss.[17] Thus, it is appropriate in this review to provide a brief background on the behavior of P in soils and to outline how such factors impinge on the environmental dimension of P.

Phosphorus Behavior in Soils

In order to establish the significance of P for the environment, there are a number of key aspects that need to be considered: the need for P for crops, forms of P in soils, reactions in soil, P cycling, and P availability. In addition, consideration has to be given to global fertilizer use as well as reserves of P for future use. This brief discussion is extracted from several sources.[15,26,30,31] Much of the information presented is now firmly part of the accepted literature on P in agriculture and the environment.

Phosphorus in Soil: Forms and Amounts

The essentiality of P for plant growth has been established for almost a century. In contrast to other elements such as N and potassium, the total amount of P in soils is relatively low, generally in the range of 0.01%–0.3% (100–3000 mg P kg^{-1}). In its native state, P solubility is extremely low, i.e., less than 0.01 mg L^{-1}. This soluble P pool, from which plants derive their P for plant uptake and growth, has to be replenished many times during the growing season in order to sustain growth. Thus, a sort of equilibrium exists between P in solution and solid-phase P. In natural ecosystems, the supply of P to plant roots is governed by reactions involving sorption, desorption, and precipitation. In such conditions, the small amount of P in solution is maintained by weathering and dissolution of rocks and minerals of low solubility. Without fertilization, which greatly increases the soluble P pool, few soils can maintain an adequate P supply to meet the needs of modern high yielding agricultural crops.

Prior to considering the soluble or "available" P fractions further, it is necessary to appreciate the forms of P in soils, which can be categorized into inorganic and organic P forms, the former being dominant in most soils except those high in organic matter. Inorganic P compounds range widely in number and type depending on soil properties but are generally associated with amorphous and crystalline sesquioxides, i.e., iron (Fe) and aluminum (Al), mainly in acid soils, and calcium (Ca) compounds in calcareous soils. The relative distribution of P compounds is largely influenced by the degree of soil weathering, which in turn influences the reactivity of these compounds.

In acid soils, hydrous oxides of Fe and Al readily react with added soluble P, resulting in precipitation, while in neutral or calcareous soils, Ca controls P reactions. In Ca dominated systems, precipitates in the following order of solubility have been identified: monocalcium phosphate, dicalcium phosphate dihydrate, and hydroxyapatite or fluorapatite; the latter compound has extremely low solubility and thus controls solution P concentration. In acid soils, few well-crystallized P compounds have been observed. Reactions of soluble P with solid-phase components also involve sorption onto adsorbing surfaces. In effect, it is difficult to separate the process of precipitation from sorption. However, there is evidence that regardless of the mechanism involved, these reactions can be reversed. Precipitation/dissolution processes differ from sorption/desorption in that the solubility product of the least soluble P compound governs solution P concentration, whereas solution P controls the amount of P sorbed. In essence, retention of soluble P by soil components is a continuum between precipitation and surface reactions. In practice, P sorption curves or isotherms have been used extensively to describe the relationship between the amount of P sorbed or removed from solution by soils and that remaining in solution. Such isotherms have been used to identify sorption maxima, bonding energy between soils, and the equilibrium P concentration at which no sorption or desorption occurs.

Regardless of the type of P reactions and the soil components involved, addition of soluble P to soils results in an immediate increase in soluble or "plant-available" P. Numerous studies have documented a subsequent rapid decrease in P availability followed by a slower rate of decline in solubility. The initial rapid phase was invariably attributed to surface reactions and/or precipitation, while the slow decline phase was attributed to diffusive penetration of adsorbed P into soil components ("absorption") and increased crystallization of precipitated P compounds. Despite the dominance of Ca in the chemical reactions of P in calcareous soils, Fe oxides (both amorphous and crystalline) have been shown to have a disproportionate role in the initial and subsequent reactions in soils.[32–34] Such oxides accentuated the rate of decline of soluble P and decreased the desorption rate. From the earliest years of P experimentation, this decline in P solubility, following reactions of soluble P with soils in laboratories, was accompanied by decreased plant P availability in greenhouse and field studies. The process was initially considered to be irreversible and often termed "fixation," retention, or immobilization.

Phosphorus Availability: Critical Concentration, Fertilization

A prerequisite to rational use of P fertilizers is knowledge of what are the critical levels of plant-available P in the soil. Much research has been conducted, involving correlation and calibration to identify appropriate indices of availability and critical test values in the field beyond which no response to added fertilizer would occur. A wide range of chemicals have been tested for availability indices based on acid dissolution, anion exchange, cation complexation, and cation hydrolysis.[35] Depending on the country or soil type and region, tests such as the Mehlich I, Bray I, and Olsen procedures are widely adopted. Multinutrient extractants such as the Mehlich III and ammonium bicarbonate-diethylenetriaminepentaacetic acid (AB-DTPA) tests have increased in popularity. Regardless of the extractant used, each should theoretically extract a fraction of the relatively soluble phase of soil P that showed a close relationship or correlation with plant uptake. Subsequently, field studies are required to establish a critical P value and thus distinguish between sufficiency and excess of soil P.[36] Thus, the amount of P fertilizer required is inversely related to soil P test values. Such guidelines are essential in order to ensure that fertilizer is applied in situations where it is needed and where it will evoke an economic crop response; conversely, they will prevent unnecessary use of fertilizer when the soil has adequate amounts of the nutrient for the season's crop, thus reducing input costs and avoiding the potential of loss of the excess P to the environment.

Estimates of critical soil values are dependent on the specific test and are influenced by soil factors; for instance, P values of 10 and 30 mg kg^{-1} were established for the Olsen and Bray I, respectively, and 10 and 25 mg kg^{-1} for clay and sandy soils, respectively. In addition to specific tests and soil type, many other factors influence P availability: temperature, compaction, moisture, aeration, pH, and type of clay. Once the need for P fertilizer is established, and a basis for the required amount indicated by critical test values that are calibrated for field conditions has been determined, other factors have to be considered in the actual fertilization process. In contrast to N, timing of P fertilizer application is less important. As an immobile nutrient, P is applied and mixed with the soil for cultivated crops, but broadcast applications are feasible only for established perennial forage crops. Regardless of whether P is applied as inorganic fertilizer or as animal manures, such additions invariably increase available P levels in the soil.

Due to the immobility of P applied to soils, efficiency of uptake per unit of P applied is always higher when the material is placed close to the seed compared with broadcasting.[24] Such placement puts the P fertilizer close enough to the seedling roots, while minimizing contact immobilization or relation. Such positional availability is also influenced by crop type, specifically its rooting pattern. In order for banding or restricted P placement to enhance P uptake by roots, the rate of absorption and root growth in the P-enriched soil zone must compensate for the limited impact on root growth in the unfertilized portion of the root zone. Much research has been done on P placement with the common range of field crops with respect to the permissible amounts of P fertilizer that can be used without damaging the seedlings and with respect to establishing optimum distances for the fertilizer band to the side or below the seed, usually about 2–5 cm. Regardless of the method of P application, the question of residual P availability arises.

Residual Phosphorus Availability: Changing Concept

Early years of P research focused on laboratory studies of P reactions in previously unfertilized soils, along with short-term greenhouse and one-season field studies. The outcome of such investigations led to the notion of P being "fixed" and of limited effectiveness following the first year of fertilization. Gradually, as commercial P fertilizer use became almost universal in agriculture, especially in the United States, Europe, and other developed countries, the concept of residual P availability emerged.[31] Evidence in support of this new concept came from long-term agronomic trials in which treatments involved initial one-time P dressings as well as current or yearly applications.

A few examples will suffice to illustrate residual availability. For instance, Halvorson and Black[37] showed that Olsen P test levels were increased for 16 years following P fertilization with one application. Following the initial increase, P levels declined slowly for 12 years and then leveled off at a level higher than initial values, suggesting a new equilibrium of available P. The elevated soil P levels in this study were also accompanied by corresponding increases in crop yields. Similar soil and crop yield responses were observed from even longer-duration trials.[38] The general acceptance of long-term effects of P fertilization was reflected in a entry on "Evaluation and Utilization of Residual Phosphorus in soils" by Barrow[39] in the comprehensive milestone on *The Role of Phosphorus in Agriculture* by Khasawneh et al.[20] With buildup of residual P, availability cropping can occur without any P fertilizer. In studies cited by Sharpley,[30] the role of decrease in available P by depletion with cropping is inversely related to the soils' buffering capacity or the P sorption saturation (available P/P sorption maximum). In shorter-term, multiyear trials under dryland Mediterranean conditions in Turkey[40] and Syria,[41] a relatively quick buildup of available P was observed following modest fertilization of calcareous soils.

The evaluation of P availability with progress in accumulated research, especially at the field level, brought with it a changed concept of efficiency. Based on early studies that evaluated P growth and uptake efficiency in terms of differences between fertilized and control plots, P fertilizer use efficiency was generally believed to be less than 20%. However, the concept of efficiency differed depending on the methods used to evaluate it. The direct method uses a radioactive isotope of P (^{32}P), but due to its short half-life, it is amenable only to short-term studies. The difference method considers differences between fertilized and unfertilized crops, i.e., "agronomic efficiency" for yield and "apparent recovery" for P uptake. On the other hand, the balance method considers only yield and P uptake relative to the amount of P fertilizer applied. In a recent comprehensive review of P efficiency supported by several case studies, Syers et al.[31] clearly demonstrated significant and consistently higher efficiency values for the balance method compared with the difference method. The key issue in such a method is the time scale. Thus, when a residual effect of P is considered over several years, efficiency values in the order of 90% can occur. This revised concept of P use efficiency represents a paradigm shift in soil fertility assessment and crop nutrition.

Dynamic Nature of Phosphorus in Soils

Of many and varied aspects of P, including broad categories such as P fertilizer sources, reactions of P fertilizers in soils, plant nutrition and crop management, animal nutrition, management practices, and the environment,[26] few are as intriguing as the dynamic or changeable nature P in soils, which was addressed recently by Condron et al.[42] for P organic P and by Tiessen[43] for P in tropical soils. In an earlier brief review, Ryan and Rashid[44] outlined the inputs to the soil solution phase of P, potential losses or withdrawals from that small fraction, and the extent to which various categories exist based on solubility. The dynamic nature of P in soils has been schematically depicted by Johnston[45] and is a simplified version of the P cycle (Figure 1), which represents the big picture with respect to P. In contrast to the P cycle, the diagram of Johnston[45] does indicate losses of soluble P, while its unique feature is the representation of the availability pools that impact the soluble P pool. The "readily available" pool represents a proportion of soil P that can be estimated by soil tests and can contribute substantially to P uptake by the crop. This pool is in apparent or pseudo-equilibrium with a less soluble fraction of soil P. A major fraction of soil P is in the very-slowly available pool. There is some evidence that this fraction can be slowly dissolved or weathered to sustain the less-readily available P pool.

In essence, these solubility categories or phases represent suites of inorganic and organic P compounds. Nevertheless, despite its simplicity, the schematic of Johnston[45] captures the essence of P dynamics and is centered on the soluble P pool from which plant uptake by roots occurs; this phase can be depleted by loss in drainage or greatly increased by the addition of soluble P fertilizers. With continued crop uptake, this tiny pool is replenished by the readily available P pool, which in turn is in "equilibrium" with a less soluble P phase. While the overall direction of P reactions is toward decreased solubility, the reverse can also occur under some circumstances. In terms of the plant roots for P uptake, accessibility is immediate in the solution phase but decreases with decreasing solubility to the right of the sketch.

Global Fertilizer Use and Phosphorus Sources

A snapshot of P in relation to both agriculture and the environment would not be complete without reference to P fertilizer usage at the global level, followed by a notion of the world's reserves of P to sustain the future of life on this planet. Production of phosphate rock, the raw material for P manufacture, has consistently expanded to meet demand primarily for P fertilizer (80%) as well as for industry. With the projected increase in world population—and consequently, crop production—fertilizer use is set to continue.[46] In essence, as more P is used, the potential for impact on the environment increases.

The continued use of phosphate rock raises the issue of the world's reserve of this finite resource,[47] which is vital to food production—and, indeed, to life on earth. That P is limited underlines the need for efficient use as well as recycling, furthering the argument for reducing or eliminating unintended deleterious effects on the global environments. While P rock is mined in many countries, the world's supply is dominated by relatively few countries, i.e., the United States, Russia, China, and Morocco. Much debate has centered on the longevity of P deposits. Many factors have to be considered, establishing estimates such as the likelihood of new discoveries, energy costs associated with P mining, market prices for P fertilizers, and continued demand. While estimates of P reserves and resources range from 105 to 470 years,[48] more recent predictions by the International Fertilizer Development Center[49] set the figure between 300 and 400 years but cautioned that many factors, known and unknown, can greatly influence such estimates. Coinciding with such estimates, Cordell et al.[50] argued that "peak" consumption will occur in the next few decades. Recognizing the finiteness of P reserves, as there is no alternative source of P, Withers[51] cautioned that protecting future food security requires a radical rethink of how P is managed from field to global scales; he further argued the need for closing gaps in the P cycle to minimize wastage and stressed the need to recover or recycle P. More effective use of P already in circulation is a prerequisite to reducing society's dependence on inorganic fertilizers and minimizing the environmental footprint of P fertilizer manufacture.

Implications for Phosphorus Loss

Without P fertilization of soils in their natural state, the potential for movement of P in its soluble state is essentially insignificant. Even when P is lost in sediments due to erosion, the impact on water bodies is minimal due to the low concentrations of P in unfertilized soils. The problem is one that accompanies application of commercial fertilizers and manures, involving excess fertilization beyond the needs of the crop. However, as economic crop production is unthinkable without P fertilizer use, the problem is one that society has to live with. However, there are guidelines and a code of fertilizer management practices that can help farmers avoid excess P use while maintaining adequate P for optimum crop production.[17] Despite the propensity of soil to react with P fertilizers, recent research has shown that excess P can overcome or saturate soil's capacity to retain P tightly. As more P is added, the soluble P fraction increases, and that is where P "leakage" occurs. Fertilization increases the P fractions attached to soil particles and increases their potential for loss when the soil is transported as sediments. An overriding condition for loss of P from fertilized land is heavy rainfall; this induces more leaching and runoff and, where appropriate soil conservation measures are not in place, promotes P loss in sediments. Accordingly, many of the hot spots of P pollution of water bodies occur in intensively formed areas or watersheds with high rainfall, e.g., Western Europe, Mississippi Valley and the Gulf of Mexico, and Chesapeake Bay in northeastern United States.

Phosphorus Losses and Mitigation Strategies

Anthropogenic or man's activities have been responsible for the accelerated cycling and fluxes of P at global and regional scales in the past few decades, resulting in eutrophication of terrestrial and aquatic ecosystems together with biodiversity loss and human health risk.[51] Current estimates indicate that fluvial transport of dissolved (about 5 Tg yr^{-1}) and particulate P (20 Tg yr^{-1}) to oceans (a permanent P sink) is at least double that of preindustrial times.[52,53] Without concentrated efforts to increase P use in agriculture through interventions, such losses are likely to continue with projected increases in commercial fertilizer consumption, higher-yielding crops, bioenergy crops, urbanization, and economic growth.[51] Recognition of the threat that such developments posed for the environment has led to expanded funding to mitigate such an environmental threat, especially in areas where the problem of P loss is acute. For instance, from 2000 to 2008, the European Research Framework Program invested over €6 million to support P-centered projects, most of which had an environmental focus.[16] Such emphasis on surplus P in the environment is in stark contrast to the situation ongoing in many developing countries, especially those in Africa, where available P is a major crop-limiting factor and where fertilizer use is dismally low or nonexistent.

However, with respect to the issue of transfer of P from agricultural watersheds, resulting in eutrophication of surface waters, a number of mitigation strategies were advanced by Sharpley and Tunney.[17] Soil P testing for environmental risk assessment was put forward as a means of establishing acceptable P losses compatible with economic farming; an issue is the appropriateness of a field/ plot or watershed scale for such studies. Recently, Maguire et al.[54] reviewed approaches to using soil testing to delineate threshold levels above which P fertilizer use is limited or not allowed. While soil testing was traditionally used for diagnosing P deficiency for crop production, it could also define an upper limit for soil P with respect to potential P losses, but many issues still need to be clarified by research. Threshold P levels should especially consider site vulnerability to P loss. An analysis of the pathways of P transport is needed at the watershed scale.

The current approach to fertilizer management in general using best management practices is one that is eminently applicable to mitigating P loss; such approaches are designed to bring P inputs as fertilizers and manures in synchrony with crop demands for P. Conventional erosion control measures form a key element in this approach. The current drive toward minimum- or no-till is a significant development in reducing P losses from cultivated land. While much remains to be learned technically in the area of P loss mitigation, a lot of existing knowledge is already available for adaptation in tackling P loss. Strategic initiatives are needed to bring about lasting change in land management targeting consumer-supported programs that encourage better standards of land management by farmers, inculcating a sense of stewardship with societal-accepted environmental protection goals.

Conclusions

The indispensability of P fertilizers for global agricultural production is undisputed, except for some adherents to organic agriculture; in fact, both commercial fertilization and organic farming should be complementary. With increasing demand on our land resources to feed the world's growing population, fertilizer use, including P, is set to continue. Research over the past century has done much to elucidate the complex nature of P in soils and plants. However, the world now faces a dilemma, with some developed countries with intensive agriculture having used excessive amounts of P, as commercial fertilizer and animal manure originating from confined animal feeding operations, while many countries of the developing world are hampered by use of little or no fertilizer. International policies have recognized the need to expand fertilizer P use in developing countries and curtail or eliminate P in developed areas with excess P. Overuse of P not only represents an economic loss to farmers but also contributes to transport of P from agricultural land to surface water bodies, where it causes eutrophication and its attendant consequences in terms of deteriorated ecology and health hazards for humans and animals.

Consequently, a major paradigm shift in soil research has been away from production agriculture and toward the environment. The change in emphasis comes at a time of general and growing societal awareness of the finite nature of rock phosphate that underpins the commercial fertilizer industry, thus making efficient use of P a major priority along with emphasis on recycling P from wastes. Much has been learned about the mechanisms of P loss from cropland and the implications of land use management in accelerating such losses; the current expansion of conservation agriculture, especially with minimum- or no-till, can greatly contribute to minimizing P losses. Tests are well established to monitor excess P in soils. However, as is explicit from chapters on the environmental aspects of P use in the most recent comprehensive treatise on P,[26] much remains to be known to provide a more effective basis for technical remediation.

The momentum of societal concern has resulted in policies restricting excess nutrient use on agricultural land in the interest of environmental protection. In essence, technology can—and will—underpin societal action to reconcile the needs for agriculture and protection of the environment.

References

1. FAO. *World Agriculture towards 2035/2050;* Food and Agriculture Organization of the United Nations: Rome, Italy, 2006.
2. Cribb, J. *The Coming Famine: The Global Food Crisis and What We Can Do to Avoid It;* CSIRO Publ.: Collingwood, Victoria, Australia, 2010.
3. Borlaug, N.E. Feeding a hungry world. Science **2007**, *318,* 359.
4. Friedman, T.L. *Hot, Flat, and Crowded: Why We Need a Green Revolution and How It Can Renew America*; Farrar, Strauss, and Giroux: New York, USA, 2008.
5. Diamond, J. *Collapse: How Societies Chose to Fail or Succeed;* Penguin Group: New York, USA, 2005.
6. Inter-Governmental Panel on Climate Change. *Assessment of Global Climate Change;* Washington, DC, USA, **2008**.
7. Godfray, H.C.; Beddington, J.R.; Crute, I.R.; Haddad, L.; Lawrence, D.; Muir, J.M.; Pretty, J.; Robinson, S.; Thomas, S.M.; Toulmin, C. Food security: The challenges of feeding 9 billion people. Science **2010**, *327,* 812–818.
8. Deane, C.; Ejita, G.; Rabbinge, R.; Saye, T. Science for global development. Crop Sci. **2010**, *50,* 1–7.
9. Roy, R.N.; Finck, A.; Blair, G.J.; Tandon. H.S. Plant nutrition for food security: a guide to integrated nutrient management. FAO Fertilizer & Plant Nutrition Bulletin No. 16; Food and Agriculture Organization of the United Nations: Rome, Italy, 2006.
10. Stewart, W.M.; Hammond, L.L.; Van Kauwenbergh, S.J. Phosphorus as a natural resource. In *Phosphorus: Agriculture and the Environment;* Sims, J.T., Sharpley, A.N., Eds; American Society of Agronomy, Crop Science Society of America, Soil Science Society of America: Madison, WI, USA, 2005; 1–22.
11. Mosier, A.R.; Syers, J.K.; Freney, J.R. *Agriculture and the Nitrogen Cycle,* SCOPE 65 (Scientific Committee on Problems of the Environment); Island Press: Washington, DC/London, 2004.
12. Stevenson, F.J. *Nitrogen in Agricultural Soils;* American Society of Agronomy, Crop Science Society of America, Soil Science Society of America: Madison, WI, USA, 1982.
13. Scheppers, T.S.; Raun, W.A., Eds. *Nitrogen in Agricultural Systems*, Agronomy Monograph No. 49; American Society of Agronomy, Crop Science Society of America, Soil Science Society of America: Madison, WI, USA, 2008.
14. Sharpley, A.N.; Smith, S.J.; Namey, J.W. Environmental impact of agricultural nitrogen and phosphorus use. J. Agric. Food Chem. **1987**, *35,* 812–817.
15. Tunney, H.; Carton, O.T.; Brookes, P.C.; Johnston, A.E. *Phosphorus Loss from Soil to Water*; CAB International; Wallingford, U.K., 1997.
16. Delgado, A.; Scalenghe, R. Aspects of phosphorus transfer from soils in Europe. J. Plant Nutr. Soil Sci. **2008**, *171,* 552–575.
17. Sharpley, A.N.; Tunney, H. Phosphorus research strategies to meet agricultural and environmental challenges of the 21st century. J. Environ. Qual. **2000**, *29,* 176–181.
18. Larsen, S. Soil phosphorus. Adv. Agron. **1967**, *19,* 151–210.

19. Dalal, R.C. Soil organic phosphorus. Adv. Agron. **1977**, *29,* 83–117.
20. Khasawneh, F.E.; Sample, E.C.; Kamprath, E.J. *The Role of Phosphorus in Agriculture*; American Society of Agronomy, Crop Science Society of America, Soil Science Society of America: Madison, WI, USA, 1980.
21. Ryan, J. Phosphorus in soils of arid regions. Geoderma **1983** *19,* 341–356.
22. Matar, A.; Torrent, J.; Ryan, J. Soil and fertilizer phosphorus and crop responses in the dryland Mediterranean zone. Adv. Soil Sci. **1992**, *18,* 82–146.
23. Ryan, J. Phosphorus fertilizer use in dryland agriculture: The perspective from Syria. In Proceedings of the OECD workshop on "Innovative Soil and Plant Systems for Sustainable Practices," Izmir, Turkey, June 3–7, 2002; Abstracts, 45.
24. Rashid, A.; Awan, Z.I.; Ryan, J.; Rafique, E.; Ibrikci, H. Strategies for phosphorus nutrition of dryland wheat in Pakistan. Commun. Soil Sci. Plant Anal. **2010**, *41,* 2555–2567.
25. Taylor, A.W.; Kilmer, V.J. Agricultural phosphorus in the environment. In *The Role of Phosphorus in Agriculture*; Khasawneh, F.E., Sample, E.C., Kamprath, E.J., Eds.; American Society of Agronomy, Crop Science Society of America, Soil Science Society of America: Madison, WI, USA, 1980; 545–557.
26. Sims, T.J.; Sharpley, A.N., Eds. *Phosphorus: Agriculture and the Environment*; American Society Agronomy, Crop Science Society of America, Soil Science Society of America: Madison, WI, USA, 2005.
27. IWP-6. Sixth International Phosphorus Workshop, Seville, Spain, Sept 27–Oct 1, 2010.
28. Sharpley, A.N.; Rekolainen, S. Phosphorus in agriculture and its environmental implications. In *Phosphorus Loss from Soil to Water;* Tunney, H., Carton, O.T., Brookes, P.C., Johnston, A.E., Eds.; CAB International: Wallingford, U.K., 1997; 1–53.
29. Pierzinski, G: Sims, J.T; Vance, G.F. *Soils and Environmental Quality,* 2nd Ed.; CRC Press: Boca Raton, FL, USA, 2000.
30. Sharpley, A.N. Phosphorus availability. D18–38. In *Handbook of Soil Science;* Sumner, M.E., Ed.; CRC Press: Boca Raton, FL, USA, 2000.
31. Syers, J.K.; Johnston, A.E.; Curtin, D. Efficiency of soil and fertilizer phosphorus use. Reconciling changing concepts of soil phosphorus behaviour with agronomic information. FAO Fertilizer and Plant Nutrition Bulletin No. 18; Food and Agriculture Organization of the United Nations: Rome, Italy, 2008.
32. Ryan, J.; Curtin, D.; Cheema, M.A. Significance of iron oxides and calcium carbonate particle size in phosphorus sorption and desorption in calcareous soils. Soil Sci. Soc. Am. J. **1985**, *49,* 7476.
33. Ryan, J.; Hassan; H., Bassiri, M., Tabbara, H.S. Availability and transformation of applied phosphorus in calcareous soils. Soil Sci. Soc. Am. J. **1985**, *51,* 1215–1220.
34. Torrent, J.A. Rapid and slow phosphate sorption by Mediterranean soils: Effect of iron oxides. Soil Sci. Soc. Am. J. **1987**, *53,* 78–82.
35. Fixen, P.E.; Grove, J.H. Testing soils for phosphorus. In *Soil testing and Plant Analysis;* Westermann, R.L., Ed.; American Society of Agronomy: Madison, WI, USA, 1990; 141–180.
36. Brown, J.R. Soil testing: Sampling, correlation, calibration and interpretation. Special Publication No. 21; Soil Science Society of America: Madison, Wisconsin, USA, 1987.
37. Halvorson, A.D.; Black, A.L. Long-term dryland crop responses to residual phosphorus fertilizer. Soil Sci. Soc. Am. J. **1985**, *49,* 928–933.
38. McCollum, R.E. Buildup and decline is soil phosphorus: 30-year trends in a Typic Umbraquult. Agron. J. **1991**, *83,* 77–85.
39. Barrow, N.T. Evaluation and utilization of residual phosphorus in soils. In *The Role of Phosphorus in Agriculture*; Khasawneh, F.E., Sample, E.C., Kamprath, E.J. Eds.; American Society of Agronomy, Crop Science Society of America, Soil Science Society of America: Madison, WI, USA, 1980; 333–360.

40. Ibrikci, H.; Ryan, J.; Ulger, A.C.; Buyuk, G.; Cakir, B.; Korkmaz, K.; Karnez, E.; Ozgenturk, G.; Konuskan, O. Maintenance of phosphorus fertilizer and residual phosphorus effect on corn production. Nutr. Cycling Agroecosyt. **2005**, 72, 279–286.
41. Ryan, J.; Ibrikci, H.; Singh, M.; Matar, A.; Masri, S.; Rashid, A.; Pala, M. Response of residual and currently applied phosphorus in dryland cereal/legume rotations in three Syrian Mediterranean agroecosystems. Eur. J. Agron. **2008**, *28,* 126–137.
42. Condron, L.M.; Turner, B.L.; Cade-Menum, B.J. Chemistry and dynamics of soil organic phosphorus. In *Phosphorus: Agriculture and the Environment;* Sims, J.T., Sharpley, A.N., Eds.; American Society of Agronomy, Crop Science Society of America, Soil Science Society of America American Society of Agronomy: Madison, WI, USA, 2005; 87–122.
43. Tiessen, H. Phosphorus dynamics in tropical soils. In *Phosphorus: Agriculture and the Environment*; Sims, J.T., Sharpley, A.N., Eds; American Society of Agronomy, Crop Science Society of America, Soil Science Society of America: Madison, WI, USA, 2005; 253–262.
44. Ryan, J.; Rashid, A. Phosphorus. In Encyclopedia of Soil Science, 2nd Ed.; Lal, R., Ed.; Taylor & Francis: New York, 2007; available at http://www.informaworld.com/smpp/content~db=all~content=a740187227~frm=titlelink. (accessed May 9, 2012).
45. Johnston, A.E. *Soil and Plant Phosphorus;* International Fertilizer Industry Association (IFA): Paris, France, 2000.
46. Stewart, W.M.; Hammond, L.L.; Van Kauwenbergh, S.J. Phosphorus as a natural resource. In *Phosphorus: Agriculture and the Environment;* Sims, J.T., Sharpley, A.N., Eds.; American Society of Agronomy, Crop Science Society of America, Soil Science Society of America: Madison, WI, USA, 2005; 1–22.
47. van Kauwenbergh, S.J. *World Phosphate Rock Reserves and Resources*; International Fertilizer Development Center: Muscle Shoals, AL, USA, 2010.
48. Cramer, M.D. Phosphate as a limiting resource: Introduction. Plant Soil **2010**, *334*, 1–10
49. Prud'homme, M. Global fertilizers and raw materials supply and supply/demand: 2006–2010. In Proceedings of the IFA Annual Conference, Cape Town, South Africa; Paris, France, 2006; 39–42.
50. Cordell, D.; Drangert, T.; White, S. The story of phosphorus: Global food security and food for thought. Global Environ. Change **2009**, *19,* 292–305.
51. Withers, P. Global phosphorus fluxes and the threat to food security. In Proceedings of the 6th International Phosphorus Workshop (IPW-6), Seville, Spain; 2010; Abstracts, 18.
52. Filippelli, G.M. The global phosphorus cycle: Past present and future. Elements **2008**, *4*, 89–95.
53. Smit, A.L., Bindraban, P.S.; Schrober, J.J.; Conijin, J.G.; van der Meer, H.G. Plant Res. Int. Rep. 2009, 282.
54. Maguire, R., Cardon, W., Simard, R.R. Assessing potential environmental impacts of soil phosphorus by soil testing. In *Phosphorus: Agriculture and the Environment;* Sims, J.T., Sharpley, A.N., Eds.; American Society of Agronomy, Crop Science Society of America, Soil Science Society of America: Madison, WI, USA, 2005; 145–180.

32

Potassium

Philippe Hinsinger

Potassium in Plants

K is the major cation in most plants, occurring at concentrations ranging from 5 to 50 g kg^{-1}, twice as much as Ca and slightly less than N.[1–3] The etymology of its name accounts for the abundance of K in plant-derived ash material (potash). K is involved in a large number of physiological processes: osmoregulation and cation–anion balance, protein synthesis and activation of enzymes.[2,3] K is often referred to as "a cation for anions" as it balances the abundant negative charges of inorganic (nitrate) and organic anions (carboxylates) in plant cells. It therefore occurs at large concentrations, 100–200 mM in the cytosol, 5–10 times less in the vacuole. Being a major inorganic solute, it plays a key role in the water balance of plants: maintenance of the osmotic potential and turgor pressure involved in cell extension. K-controlled changes in turgor pressure in guard cells is a key process of stomatal opening and closure and hence, of the regulation of plant transpiration. Many of these physiological roles are related to the high mobility of K at all levels in the plant. This unique, considerable mobility of K in the plant is essentially due to the large permeability of cell membranes to K-ions, which arises from the occurrence of a range of highly K selective, low and high affinity ion channels and transporters. These are now being increasingly characterized at a molecular level.[4] Large rates of K uptake can thereby be achieved in plant roots. In addition, K-ions can easily be leached out of living plant tissues, as documented for tree foliage which contributes a large flux of K back to the soil via through- fall.[3] K also rapidly leaves dead roots and other plant debris compared with N and P, which require hydrolysis of organic molecules. At an agronomic level, the demand for K largely varies with plant species and productivity. The uptake of K essentially occurs during the vegetative stage and can reach values of 10 kg ha^{-1}day^{-1} and above. Depending also on the agricultural practices (removal of straw, for instance), the amount of K removed with the harvested material will range from 5–50 kg ha^{-1} for cereal grains to 50–500 kg ha^{-1} for forage, root and tuber or plantation crops.

Potassium in Soils

Among major nutrients, K is usually the most abundant in soils as total K content ranges from 0.1 to 40, with an average of 14 g kg^{-1}.[1,5] A major proportion of soil K occurs as structural K in feldspars and interlayer K in micaceous minerals (Figure 1).[1,6,7] Some minor proportion of soil K (usually much less

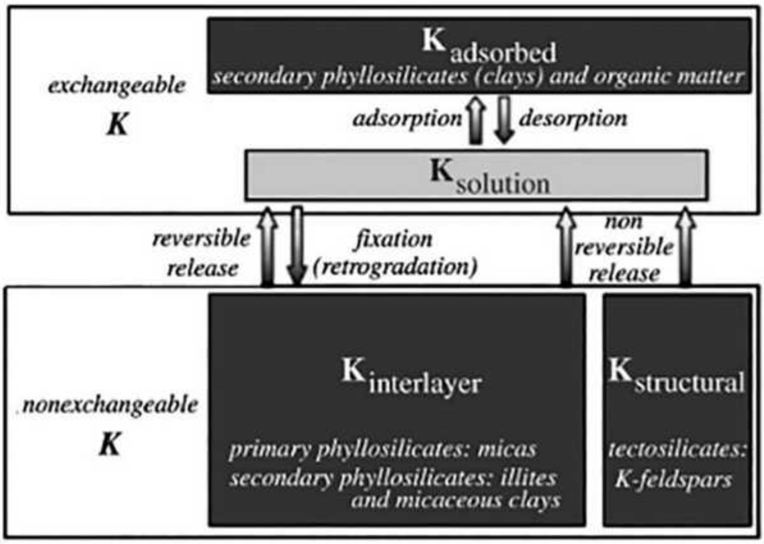


FIGURE 1 The various forms of soil K and the chemical processes involved in soil K dynamics

than 1%) is adsorbed on negatively charged soil constituents, namely clay minerals and organic matter. A marginal part is present as free K-ions in the soil solution. Bulk soil solution concentrations usually amount to 100–1000 µM (less than 0.01%–0.1% of total K). The reason for this rather low concentration of K in the soil solution and hence restricted mobility of K in soils, compared to other metal cations such as Na or Ca is related to its selective adsorption onto some clay minerals. Because of its ionic radius and small hydration energy, K-ions indeed perfectly fit into the interlayer sites of micaceous minerals (micas, illites and mica-derived clays).[8] These sites and, to a lesser degree, the sites on the frayed edges of these minerals have thus a considerably larger affinity for K than for other cations, including divalent cations such as Ca or Mg (Figure 2). Clay minerals also bear sites with larger affinity for divalent cations than for monovalent cations such as K. These sites are located on the planar faces of clay minerals and are thus dominant in clays such as kaolins and smectites. They also occur in organic compounds. K is thus much less strongly held in soils dominated by kaolins (tropical soils), sand or organic matter than in soils dominated by illite–vermiculite clay minerals. Traces of mica-derived clay minerals can dramatically influence the dynamics of soil K as evidenced in tropical soils that are apparently dominated by kaolins.[9] More generally, K dynamic is largely dependent on soil mineralogy which determines both ion exchange and release-fixation processes,[6,7] i.e., the dynamics of "nonexchangeable K." The latter is defined as that (major) portion of soil K which cannot be exchanged by NH_4-ions. NH_4-ions have the same charge and radius than K-ions and can successfully desorb K only from low K-affinity sites. Exchangeable K thus comprise soil solution and easily desorbable K-ions. Nonexchangeable K mostly comprise interlayer K (high K-affinity sites) of micaceous minerals and structural K of feldspars (Figure 1), i.e., 90%–99% of total K in many soils. The release of K from feldspars requires a complete and irreversible dissolution of the mineral and is enhanced under acidic conditions.[6,7] The release of interlayer K from micaceous minerals can proceed similarly or involve an ion exchange process (Figure 1) leading to an expansion of the phyllosilicate (Figure 2). This reversible release is essentially governed by the concentrations of K and competing cations in the outer solution.[8,10,11] Cations which can be responsible for this release, such as Ca- and Mg-ions, have a large hydration energy, contrary to K- or NH4- ions. Therefore, they remain hydrated when exchanging interlayer K-ions and expand the interlayer space (Figure 2), making it possible for the release to proceed further, whereas NH_4-ions would block the reaction.[12] However, because of the considerable affinity of these interlayer sites for K relative to Ca or Mg, the release can occur only for extremely low concentrations of K in the soil solution (Figure 3), in the micromolar range.[11,12]

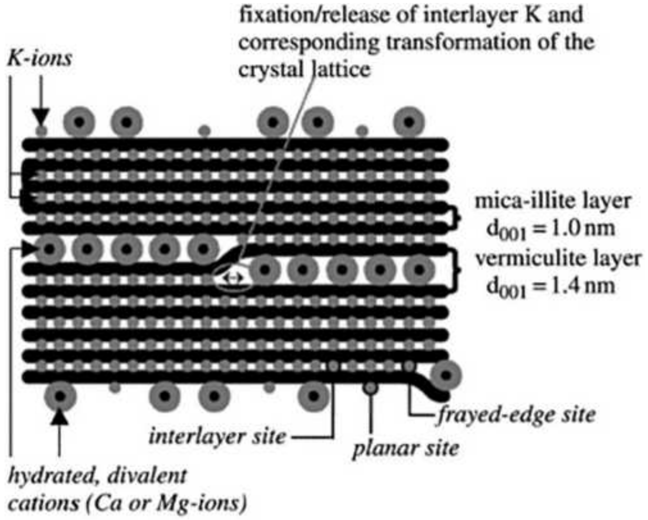


FIGURE 2 The various sites of exchange of K-ions in micaceous clay minerals and the transition between mica and vermiculite layer that occurs when interlayer K is exchanged by hydrated, divalent cations.

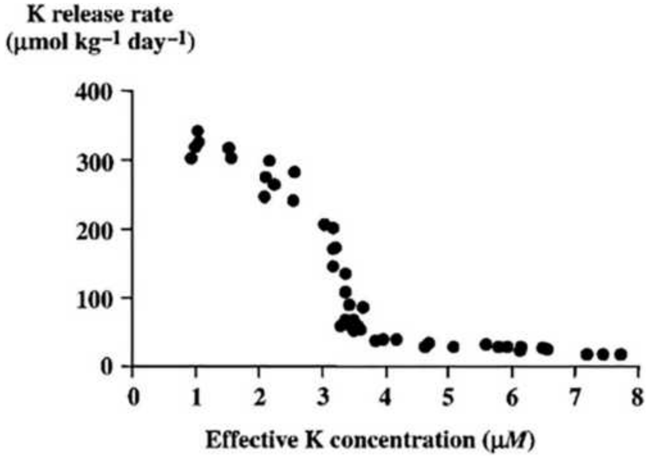


FIGURE 3 Effect of solution K concentration on the rate of release of nonexchangeable K from a soil.
Source: Springob and Richler.[11]

Conversely, elevated concentrations of K are prone to the reverse reaction of fixation, i.e., the collapse of expanded layers and concomitant increase in nonexchangeable K at the expense of readily available K. As long as exchangeable K was assumed to be the only plant-available K fraction, K fixation, i.e., the poor recovery of applied K in the exchangeable fraction was seen as a poor efficiency of K fertilization. This apparent loss of applied K has therefore received considerably more interest than the reverse release process.[7] In addition, the release of nonexchangeable K was considered unlikely to occur to any great extent, especially because K concentration in the bulk soil solution of fertilized soils is usually far above the critical concentrations that are prone to K release. However, numerous long-term fertilizer trials have shown that the release of nonexchangeable K contributes a major proportion of soil K supply in unfertilized and possibly fertilized plots too, although an overall net fixation is often found in the latter (Figure 4).[5,13] Nonexchangeable K can contribute up to 100 kg ha^{-1} yr^{-1}, 80%–100% of soil K supply.

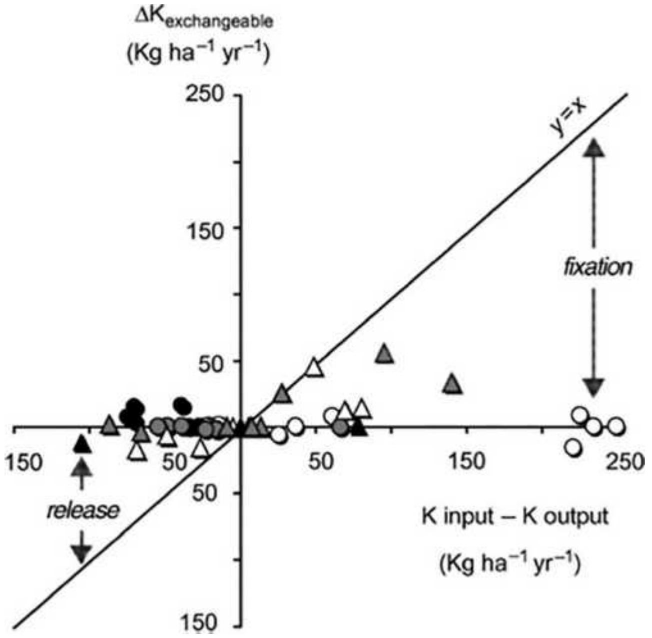


FIGURE 4 Annual change in exchangeable K ($K_{exchangeable}$) as a function of the annual K budget in various K treatment plots of long-term fertilizer trials. The K input comprehends organic and inorganic, applied K fertilizers. The K output corresponds to the offtake of K in the harvested product.
Source: Adapted from Blake[5] and Gachon.[13]

Such annual fluxes of release are fairly large compared with K dissolution rates commonly estimated by geochemists (in the order of 5–15 kg ha^{-1} yr^{-1}). However, many geochemical models do not take into account the amount of K taken off in the vegetation, leading to large underestimates of the actual dissolution rates.[14] Conversely, many K budgets provided by agronomists do not take into account atmospheric inputs and leaching of soil K, assuming that these terms are fairly negligible in most cases. This certainly holds true for atmospheric inputs which rarely exceed values in the order of a few kg ha^{-1} yr^{-1}. Leaching can, however, vary over a much wider range of values, from several kg ha^{-1} yr^{-1} in most cases up to several tens (and up to 1–3 hundreds) of kg ha^{-1} yr^{-1} in those situations that are the most prone to leaching: bare soil or poor soil coverage by the vegetation, excessive fertilizer rates, coarse-textured soils. Omitting the leaching term would, however, have led to underestimating the actual release rate.[5] Hence, considerable amounts of nonexchangeable K can be released in agricultural soils and contribute a significant proportion of plant uptake, in contradiction to the widespread viewpoint shared by numerous agronomists and soil scientists. Reasons for this can be found when considering root–soil interactions occurring in the rhizosphere.

Potassium in Plant–Soil Interactions

K occurs at rather low concentrations in the soil solution, compared to other nutrient cations and to the large requirements of plants for K. The transfer of soil K via mass-flow toward plant roots (i.e., the convective flow of solute accompanying panying transpiration-driven water flow) contributes about 1%–20% of plant demand.[3,15–17] A direct consequence is the rapid depletion of K-ions from the soil solution in the vicinity of plant roots, i.e., the rhizosphere. The resulting concentration gradient generates a diffusion of K-ions in the rhizosphere which plays a key role in the transport of K toward plant root (i.e., 80%–99% of plant demand). Such depletion results in a shift of the cation exchange equilibria which rule

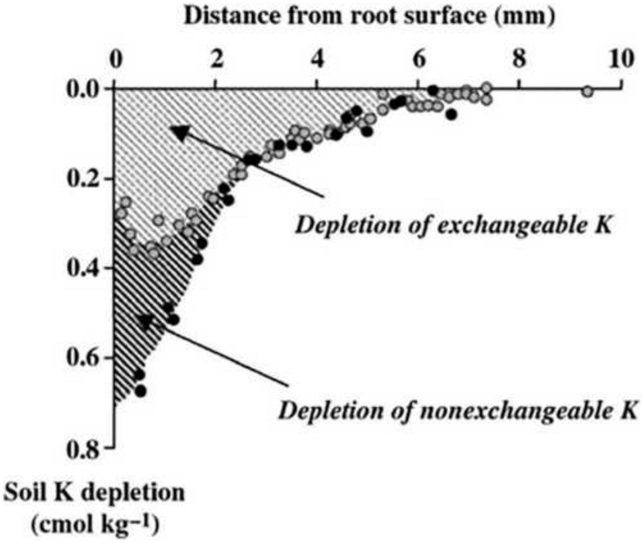


FIGURE 5 Depletion of both exchangeable K (gray dots) and HCl-extractable K (black dots) as a function of the distance from rape roots.
Source: Adapted from Jungk and Claassen.[17]

the dynamics of both exchangeable and interlayer K. This ultimately results in a desorption of exchangeable K and eventually of interlayer K,[7,16,17] as shown by their depletion in the rhizosphere (Figure 5). The extent of the depletion of exchangeable K will depend on chemical parameters such as the initial level of exchangeable K and the K buffering capacity of the soil and on physical parameters that directly determine the diffusive transport of K-ions: soil texture and structure, soil water content.[17] The K depletion zone will extend over several millimeters in clayey, dry soils up to several centimeters in wet, sandy soils.[7,16,17] The intensity of the depletion will also depend on how far the K concentration of solution is decreased, which may vary among plant species according to the K uptake ability of their root. Plants with a lower external K efficiency, i.e., with a higher affinity transport system, will have the capability to take up K at lower K concentrations and may thus deplete soil K further.[17] In the vicinity of roots, solution K concentration can indeed decrease by 2–3 orders of magnitude, down to as little as 2–3 μM.[16] At such low K concentrations, the release of nonexchangeable K can occur at large rates, whereas it would be dramatically restricted at bulk soil K concentrations of several hundreds of μM (Figure 3).[11] Plants thus play a major role in the dynamic of interlayer K via the root-induced depletion of solution K.[16] Measurements in pot experiments have indeed revealed that within several days of growth, the release of nonexchangeable K can amount up to 90% of K supplied to the plant.[7,16] Soil–root chemical interactions in the rhizosphere thus largely explain the unexpectedly large contribution of the release of nonexchangeable K to plant uptake that is found in many agricultural soils, including fertilized soils.

Assessing and Managing Potassium Fertility

Soil K fertility is most often evaluated by measuring exchangeable K[1,7] most frequently with molar NH_4 acetate in batch conditions. However, the adequacy of exchangeable K to predict plant response, i.e., the actual bioavailability of soil K, is rather poor in many soils. This arises from the major contribution of the release of nonexchangeable K in some soils, especially when exchangeable K is low and/or when large reserves of nonexchangeable K are readily available as a consequence of: 1) soil mineralogical composition or 2) fertilization history (build-up of fixed K due to excessive K-fertilizer rates). In these situations, quantitative evaluation of the potential release of nonexchangeable K would be highly recommended for a better prediction of plant response and fertilizer needs.[7] There are several methods for assessing nonexchangeable K but none of them is routinely used on a broad scale, because of their cost. These are either based on the use of 1) concentrated, strong acids that dissolve K-bearing minerals or 2) cationic resins or chemicals such as Na tetraphenylboron that can promote the release of interlayer K by removing K-ions from soil solution and by shifting the exchange equilibria.[6,7] Alternatively, correction factors can be used when interpreting exchangeable K values, which account either for the cationic exchange capacity (or clay content) or for the soil type and K release potential.[7] Exchangeable K is nonetheless often used alone for fertilizer recommendations, resulting in frequently overestimated fertilizer needs to compensate for the expected large fixation and negligible release. Many long-term fertilizer trials have shown that adequate yields of crops can be obtained at fairly low rates of K fertilizer application, or even, for the least demanding crops such as cereals, without any K fertilizer for several years or decades.[7,13] Other more demanding crops, however, require the application of K fertilizer to achieve high yield and quality in the harvested products.[18] The need for K-fertilizers will thus depend on the release potential of the soil and on the demand of the plant, the latter being now increasingly accounted for in fertilizer recommendations. Fertilizer trials have also shown that commonly used soluble K fertilizers and organic sources such as manure or crop residues have fairly comparable efficiencies. This is not surprising as K is highly mobile in organic compounds where it occurs as soluble or exchangeable K-ions. These sources are thus equally important as K-fertilizers and absolutely need to be accounted for in K budgets.

Conclusions

K is the major nutrient cation for plants and thus taken up at large rates by plant roots. These are achieved by both high and low affinity transport systems which explain the considerable mobility of K within the plant. In comparison, K is much less mobile in soils because of the strong affinity of some exchange sites of clays. The large K uptake rates achieved by roots result in a steep depletion of solution K in the rhizosphere, and hence in a shift of the equilibria of cation exchange. Exchangeable K and even nonexchangeable K can thereby be significantly depleted and contribute a substantial proportion of plant uptake. This is confirmed by K balance both in short-term pot experiments and long-term field trials. In addition to the desorption–adsorption of exchangeable K, release and fixation processes thus need to be accounted for when evaluating soil K fertility.

References

1. Munson, R.D. *Potassium in Agriculture*; American Society of Agronomy, Crop Science Society of America; Soil Science Society of America: Madison, WI, 1985; 1223 pp.
2. Mengel, K.; Kirkby, E.A.; Kosegarten, H.; Appel, T. *Principles of Plant Nutrition,* 5th Ed.; Kluwer Academic Publishers: Dordrecht, Netherlands, 2001; 673 pp.
3. Marschner, H. *Mineral Nutrition of Higher Plants,* 2nd Ed.; Academic Press: London, 1995; 889 pp.
4. Schachtman, D.P. Molecular insights into the structure and function of plant K^+ transport mechanisms. Biochim. Bio-phys. Acta **2000**, *1465,* 127–139.
5. Blake, L.; Mercik, S.; Koerschens, M.; Goulding, K.W.T.; Stempen, S.; Weigel, A.; Poulton, P.R.; Powlson, D.S. Potassium content in soil, uptake in plants and the Potassium balance in three European long-term field experiments. Plant Soil **1999**, *216,* 1–14.
6. Sparks, D.L. Potassium dynamics in soils. Adv. Soil Sci. **1987**, *6*, 1–63.
7. International Potash Institute. *Methodology in Soil-K Research*; Proceedings of the 20th colloquim of the International Potash Institute, Baden bei Wien, Austria, International Potash Institute: Bern, Switzerland, 1987; 428 pp.
8. Dixon, J.D.; Weed, S.B. *Minerals in Soil Environment;* Soil Science Society of America: Madison, WI, 1989; 1244 pp.

9. Fontaine, S.; Delvaux, B.; Dufey, J.E.; Herbillon, A.J. Potassium exchange behaviour in Carribean Volcanic ash soils under banana cultivation. Plant Soil **1989**, *120,* 283–290.
10. Schneider, A. Influence of soil solution Ca concentration on short-term release and fixation of a loamy soil. Eur. J. of Soil Science **1997**, *48,* 513–522.
11. Springob, G.; Richter, J. Measuring interlayer Potassium release rates from soil materials. II. A percolation procedure to study the influence of the variable solute K in the <1…10 μM range. Z. Pflanzen. Bodenk. **1998**, *161,* 323–329.
12. Springob, G. Blocking the release of potassium from clay interlayers by small concentrations of NH_4^+ and Cs^+. Eur. J. of Soil Sci. **1999**, *50,* 665–674.
13. Gachon, L. *Phosphore et Potassium dans les Relations Sol-Plante: Conséquences sur la Fertilisation·,* Institut National de la Recherche Agronomique: Paris, France, 1988; 566 pp.
14. Taylor, A.B.; Velbel, M.A. Geochemical mass balances and weathering rates in forested watersheds of the southern blue ridge II. Effects of botanical uptake terms. Geoderma **1991**, *51,* 29–50.
15. Barber, S.A. *Soil Nutrient Bioavailability. A Mechanistic Approach,* 2nd Ed.; Wiley: New York, 1995; 414 pp.
16. Hinsinger, P. How do plant roots acquire mineral nutrients? chemical processes involved in the rhizosphere. Adv. Agron. **1998**, *64,* 225–265.
17. Jungk, A.; Claassen, N. Ion diffusion in the soil-root system. Adv. Agron. **1997**, *61,* 53–110.
18. http://www.ipipotash.org/publications/publications.html (accessed October 2000).

33

Radionuclides

Philip M. Jardine

Introduction

Soil, the thin veneer of matter covering the Earth's surface and supporting a web of living diversity, is often abused through anthropogenic inputs of toxic waste. The disposal of radioactive waste generated at U.S. Department of Energy (DOE) facilities within the Weapons Complex has historically involved shallow land burial in unsaturated soils and sediments. Disposal methods from the 1940s to the 1980s ranged from unconfined pits and trenches to single- and double-shell buried steel tanks. Most of the below-ground burial strategies were deemed to be temporary (i.e., an average life span of several decades) until suitable technologies were developed to deal with the legacy waste issues. Technologies for retrieving and treating the below-ground radionuclide waste inventories have been slow to evolve and are often cost prohibitive or marginally effective. The scope of DOE's disposal problem is massive, with landfills estimated to contain more than 3 million cubic meters of radioactive and hazardous buried waste; a significant proportion of which migrated into surrounding soils and groundwater. It is estimated that the migration of these waste plumes contaminated over 600 billion gallons of water and 50 million cubic meters of soil.

Fate and Transport Processes

Hydrologic Processes

Soil is a complex continuum of pore regions ranging from large macropores at the mm scale to small micropores at the sub-μm scale. It is the physical properties of the media (e.g., structured or layered), coupled with the duration and intensity of precipitation events that dictates the avenues of water and radionuclide movement through the subsurface. In humid environments where structured media is commonplace, transient storm events invariably result in the preferential migration of water.[1–8] Highly conductive voids within the media (e.g., fractures, macropores) carry water around low-permeability, high-porosity matrix blocks or aggregates resulting in water bypass of the latter. In these humid regimes, recharge rates are very high with more than 50% of the infiltrating precipitation resulting in groundwater and surface water recharge. This condition promotes the formation of massive contaminant plumes in the soil since storm flow and groundwater interception with waste trenches is frequent and long-lasting. Even in semiarid environments, where recharge is typically small, subsurface preferential flow is a key mechanism controlling water and solute mobility.[9,10] Lithologic discontinuities and sediment

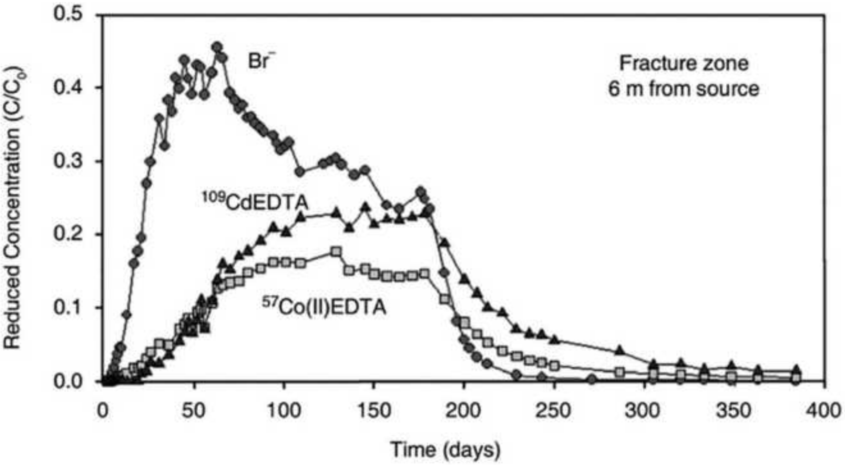


FIGURE 1 Field-scale fate and transport of nonreactive Br$^-$ and reactive ^{57}Co(II)EDTA^{2-} and ^{109}CdEDTA^{2-} in fractured subsurface media at the Oak Ridge National Laboratory. Although transport rates are rapid, geochemical reactions significantly impede the mobility of the chelated radionuclides as is indicated by their delayed breakthrough.
Source: Jardine et al.[40]

layering promote perched water tables and unstable wetting fronts that drive both lateral and vertical subsurface preferential flow.

In both humid and semiarid regimes, water that is preferentially flowing through the soil media often remains in intimate contact with the porous matrix, and physical and hydrologic gradients drive the exchange of mass from one pore regime to another. Mass exchange is time-dependent and is often controlled by diffusion to and from the matrix. Thus, a significant inventory of radionuclide waste can reside within the soil matrix. This waste source is hydrologically linked to preferred flow paths which significantly enhances the extent and longevity of subsurface contaminant plumes. This scenario is commonplace at the Oak Ridge National Laboratory, located in eastern Tennessee, U.S., where thousands of underground disposal trenches and ponds have contributed to the spread of radionuclides such as ^{137}Cs, ^{60}Co, ^{90}Sr, and $^{235/238}$U across tens of kilometers of landscape. Highly concentrated contaminant plumes move through soil and groundwater at time scales of meters per day (Figure 1) since the soils are highly structured and conducive to rapid preferential flow. However, the soil matrix, which has a high porosity and low permeability, serves as a source/sink for contaminants.[5,6,11] The preferential movement of water and radionuclides through the subsurface also significantly impacts geochemical and microbial processes by controlling the extent and rate of various reactions with the solid phase. It imposes kinetic constraints on biogeochemical reactions and limits the surface area of interaction by partially excluding water and mass from the matrix porosity.

Geochemical Processes

Radionuclide fate and transport in soil and sediments is also controlled by interfacial reactions with the soil solid phase. Most soils are a complex mixture of variably charged phyllosilicates, redox reactive Fe- and Mn-oxides, organic matter, and mineral carbonates. Radionuclides interact with these solid phases through coulombic exchange, chemisorption, redox alterations, transformation processes such as polymerization, precipitation/dissolution, and complexation reactions. Both the extent and rate of these processes can be significantly influenced by variations in water content and the degree of pore regime connectivity. To make matters worse, radionuclide waste generated at the U.S. DOE facilities was often co-disposed with various chelating agents and organic acids. These synthetic organic constituents form highly stable, water-soluble complexes with a wide variety of radionuclides.[12,13] The presence of the complexing agent significantly alters the geochemical behavior of the disposed contaminants in soils and sediments through increased solubility, accelerated redox reactions, and ionic charge reversal.

The geochemical mechanism controlling the fate and transport of chelated radionuclides has been well characterized in numerous soils and subsurface materials.[14–22] Typically, Fe(III) and Mn(IV) oxyhydroxides are the dominant subsurface mineral assemblages that catalyze co-contaminant oxidation/reduction and dissociation reactions (Figure 2). The mineral oxides have repeatedly been shown to catalyze the oxidation of ^{60}Co(II)EDTA^{2-} to ^{60}Co(III)EDTA^{-}, thereby adversely enhancing the transport and persistence of ^{60}Co in a variety of subsurface environments ranging from aquifer sands to fractured weathered shale saprolites.[15,17,18,20,22] Further, Fe(III)-oxides have also been shown to effectively dissociate a large number of chelated metal and radionuclide complexes (e.g., ^{60}Co–, ^{90}Sr–EDTA) through ligand competition.[15,20–22]

Certain radionuclides such as ^{137}Cs do not form strong bonds with many of the chelating agents and organic acids that were used during decontamination. Nevertheless, these radionuclides still interact aggressively with the soil solid phase. In the case of ^{137}Cs, 2 : 1 phyllosilicates and micas serve as excellent sorbents since the interlayer spaces of these mineral assemblages strongly attenuate the radionuclide. The migration tendency of ^{137}Cs in soils is often related to colloid mobility of contaminated sediments[23] or cation competition for surface sites in harsh environments such as those found beneath the Hanford tank farms in western Washington State, U.S.[24]

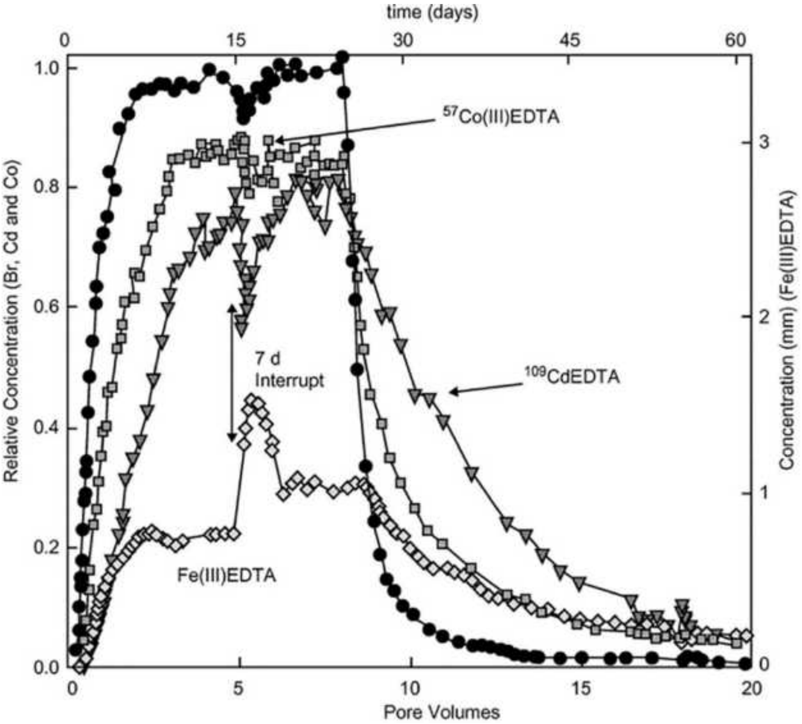


FIGURE 2 Fate and transport of nonreactive Br– and reactive ^{57}Co(II)EDTA^{2-} and 109 CdEDTA^{2-} in undisturbed soil columns of fractured weathered shale. Geochemical reactions impede the mobility of the chelated radionuclides. Co(II)EDTA^{2-} is oxidized to Co(III)EDTA^{-} where Mn-oxides serve as the oxidant. Fe-oxides effectively dissociate CdEDTA^{2-} complexes resulting in the formation of free Cd and Fe(III)EDTA^{-}. A flow interruption technique was employed to quantify the presence of physical and geochemical nonequilibrium processes. **Source:** Mayes et al.[22]

Microbial Processes

Radionuclides such as ^{60}Co and $^{235/238}U$ can exist in more than one oxidation state, and their behavior in the environment depends on their oxidation state. For example, U(VI) is soluble and mobile in the environment whereas U(IV) is much less soluble and relatively immobile. Likewise, the oxidized $^{60}Co(III)EDTA$ complexes are much more stable and exhibit greater mobility in subsurface environments than the reduced $^{60}Co(II)EDTA$.[15,17,18] Subsurface Fe- and Al-oxides can effectively dissociate the Co(II)EDTA complex to Fe(III)EDTA[20] and Al(III)EDTA,[16] respectively, and aqueous Co^{2+} is free to participate in sorption or precipitation reactions. Co(III) EDTA, on the other hand, is unaffected by Fe(III)- and Al-oxides. Therefore, the oxidized forms of these radionuclides and metals promote their undesirable enhanced migration through subsurface environments.

Numerous metal-reducing bacteria have been isolated that enzymatically reduce toxic metals and radionuclides to stable end-products. Microbial reduction of U(VI) to form the sparingly soluble U(IV) has been shown using chemostat experiments for a number of metal-reducing bacteria.[25,26] Gorby et al.[27] have also shown that certain metal-reducing bacteria can link the enzymatic reduction of $^{60}Co(III)EDTA^{-}$ to support cell growth. Recently, important advances have been made towards implementing field-scale microbially mediated metal reduction strategies in oxygen-deficient environments. Several studies have investigated contaminant reduction in the presence of solid phase material.[27–29] Gorby et al.[27] have shown that the metal-reducing bacterium *Shewanella alga* preferentially reduced $Co(III)EDTA^{-}$ to $Co(II)EDTA^{2-}$ in the presence of Mn-oxides. Likewise, Wielinga et al.[29] documented the bioreduction of U(VI) by *Shewanella alga* in the presence of various Fe-oxide mineral phases. These authors noted that the rate of U(VI) bioreduction was unaffected in the presence of goethite and only slightly diminished in the presence of poorly crystalline Fe(III)-oxides, where the latter Fe solid phase effectively competed as a terminal electron acceptor. Recent studies by Brooks, Carroll, and Jardine.[28] showed the sustained microbial reduction of $Co(III)EDTA^{-}$ under dynamic flow conditions. The net reduction of the $Co(III)EDTA^{-}$ dominated the fate and transport of the contaminant even in the presence of strong mineral oxidants such as Mn- and Fe-oxides that are known to effectively reoxidize $Co(II)EDTA^{2-}$ back to $Co(III)EDTA^{-}$.[15,18,20] The research findings of Brooks, Carroll, and Jardine[28] provide new and important information on how to successfully implement a bioreduction strategy at the field scale. Their use of a dynamic flow system with sustained bacterial growth conditions in geochemically reactive media is consistent with contaminant migration scenarios in situ.

The studies of Brooks, Carroll, and Jardine,[28] however, used uniformly packed media that contained little structure. Undisturbed subsurface soils and geologic material consist of a complex continuum of pore regions ranging from large macropores and fractures at the mm scale to small micropores at the sub-μm scale. Structured media, common to most subsurface environments throughout the world, accentuates this physical condition that often controls the geochemical and microbial processes affecting solute transport. Redox sensitive radionuclides such as U(VI), Co(III)EDTA, and Tc(VI) reside within nearly all of the pore structure of the subsurface media, with the greatest concentration of contaminants held within micropores.[2,3] Bacteria that are capable of reducing these contaminants are too big to reach a large fraction of the micropore regime and are largely restricted to macro- and mesopore domains.[30,31] Fortunately, the pore structure of the media is hydrologically interconnected, and contaminants move from one pore class to another via hydraulic and concentration gradients.[4,6,7] This process is slow, however, and is often the rate-limiting factor governing the success of contaminant bioremediation. Thus, faster-flowing fracture-dominated regimes will most likely be physically more appealing for sustained bioreduction as long as a suitable electron donor can be supplied. In contrast, bioreduction processes in slower-flowing matrix regimes will most likely be limited by rate-dependent mass transfer of contaminants from smaller pores into larger pores.

Certain bacteria are also capable of degrading chelates and thus potentially immobilizing radionuclides in situ. The biodegradation of the commonly used aminopolycarboxylate chelates NTA, EDTA, and DTPA have been studied in soil and sediment systems for many years.[32,33,35] Research has shown that NTA has the greatest potential for biodegradation in subsurface systems compared with the other aminopolycarboxylates.[34,35]

Bolton et al.[36] and Bolton and Girvin[37] have shown that the bacterial strain *Chelatobacter heintzii* (ATCC 29600) is capable of degrading NTA in the presence of many different toxic metals and radionuclides. Likewise, Payne et al.[38] and Liu et al.[39] have deciphered the mechanisms by which certain bacteria degrade radionuclide– EDTA complexes. These studies lend promise to the potential for using bacteria to biodegrade chelates and enhance the geochemical immobilization of radionuclides in situ.

Conclusions

Radionuclide fate and transport in soils is controlled by coupled time-dependent hydrologic, geochemical, and microbial processes. Hydrologic processes such as preferential flow and matrix diffusion can serve to both accelerate and impede radionuclide migration, respectively. Preferential flow results in hydraulic, physical, and geochemical nonequilibrium conditions since differences in fluid velocities and solute concentrations in different-sized pores create hydraulic and concentration gradients that drive time-dependent inter-region advective and diffusive mass transfer. Thus, in soil systems with a large matrix porosity or a significant quantity of disconnected immobile water, radionuclide migration rates can be greatly retarded due to the slow transfer of mass to actively flowing preferential flow paths. Nevertheless, the prevalence of preferential flow can greatly accelerate the transport of mass in soil systems. Geochemical processes such as sorption, redox alterations, and dissociation reactions can also serve to both accelerate and impede radionuclide migration. Sorption and radionuclide-chelate dissociation reactions almost always result in retarded radionuclide migration rates, whereas oxidation reactions often result in more soluble, and thus more mobile, radionuclide species. Microbial processes can also potentially influence the fate and transport of radionuclides in soil. Metal-reducing bacteria and chelate degraders can alter the geochemical behavior of redox sensitive radionuclides which facilitates their immobilization via solid phase sorption and precipitation reactions.

Enhanced knowledge of the coupled hydrologic, geochemical, and microbial processes controlling radionuclide migration in soils will improve our conceptual understanding and predictive capability of the risks associated with spread of radioactive material in the subsurface environment. Too often risk assessment models treat soil and bedrock as inert media or assume that the media is in equilibrium with migrating contaminants. Failure to consider the time-dependent coupled processes that control radionuclide migration will greatly over-predict the offsite contribution of contaminants from the primary waste source and thus provide an inaccurate assessment of pending risk. By recognizing the importance of soil processes on radionuclide migration, we can improve our decision making strategies regarding the selection of effective remedial actions and improve our interpretation of monitoring results after remediation is complete.

References

1. Shuford, J.W.; Fitton, D.D.; Baker, D.E. Nitratenitrogen and chloride movement through undisturbed field soil. J. Environ. Qual. **1977**, *6*, 255–259.
2. Jardine, P.M.; Wilson, G.V.; Luxmoore, R.J. Unsaturated solute transport through a forest soil during rain storm events. Geoderma **1990**, *46*, 103–118.
3. Jardine, P.M.; Wilson, G.V.; McCarthy, J.F.; Luxmoore, R.J.; Taylor, D.L. Hydrogeochemical processes controlling the transport of dissolved organic carbon through a forested hillslope. J. Contam. Hydrol. **1990**, *6*, 3–19.
4. Jardine, P.M.; O'Brien, R.; Wilson, G.V.; Gwo, J.P. Experimental techniques for confirming and quantifying physical nonequilibrium processes in soils. In *Physical Nonequilibrium in Soils: Modeling and Application;* Selim, H.M., Ma, L., Eds.; Ann Arbor Press: Chelsea, MI, 1998; 243–271.

5. Jardine, P.M.; Wilson, G.V.; Luxmoore, R.J.; Gwo, J.P. Conceptual model of vadose-zone transport in fractured weathered shales. In *Conceptual Models of Flow and Transport in the Fractured Vadose Zone;* U.S. National Committee for Rock Mechanics, National Research Council; National Academy Press: Washington, DC, 2001; 87–114.
6. Wilson, G.V.; Jardine, P.M.; O'Dell, J.D.; Collineau, M. Field-scale transport from a buried line source in variable saturated soil. J. Hydrol. **1993**, *145*, 83–109.
7. Wilson, G.V.; Gwo, J.P.; Jardine, P.M.; Luxmoore, R.J. Hydraulic and physical nonequilibrium effects on multi-region flow and transport. In *Physical Nonequilibrium in Soils: Modeling and Application;* Selim, H.M., Ma, L., Eds.; Ann Arbor Press: Chelsea, MI, 1998; 37–61.
8. Hornberger, G.M.; Germann, P.F.; Beven, K.J. Through flow and solute transport in an isolated sloping soil block in a forested catchment. J. Hydrol. **1991**, *124*, 81–97.
9. Porro, I.; Wierenga, P.J.; Hills, R.G. Solute transport through large uniform and layered soil columns. Water Re-sour. Res. **1993**, *29*, 1321–1330.
10. Ritsema, C.J.; Dekker, L.W.; Nieber, J.L.; Steenhuis, T.S. Modeling and field evidence of finger formation and finger recurrence in a water repellent sandy soil. Water Resour. Res. **1998**, *34*, 555–567.
11. Jardine, P.M.; Sanford, W.E.; Gwo, J.P.; Reedy, O.C.; Hicks, D.S.; Riggs, R.J.; Bailey, W.B. Quantifying diffusive mass transfer in fractured shale bedrock. Water Resour. Res **1999**, *35*, 2015–2030.
12. Riley, R.G.; Zachara, J.M. *Chemical Contaminants on DOE Lands and Selection of Contaminant Mixtures for Subsurface Science Research*; DOE=ER-0547T; U.S. Govt. Print. Office: Washington, DC, 1992.
13. Toste, A.P.; Osborn, B.C.; Polach, K.J.; Lechner-Fish, T.J. Organic analyses of an actual and simulated mixed waste: Hanford's organic complexant site revisited. J. Radioanal. Nucl. Chem. **1995**, *194*, 25–34.
14. Swanson, J.L. *Effect of Organic Complexants on the Mobility of Low-level Waste Radionuclides in Soils*; Status Report PNL-3927, UC-70, 1981.
15. Jardine, P.M.; Jacobs, G.K.; O'Dell, J.D. Unsaturated transport processes in undisturbed heterogeneous porous media: II. Co-contaminants. Soil Sci. Soc. Am. J. **1993**, *57*, 954–962.
16. Girvin, D.C.; Gassman, P.L.; Bolton, H. Adsorption of aqueous cobalt ethylenediaminetetraacetate by d-Al2O3. Soil Sci. Soc. Am. J. **1993**, *57*, 47–57.
17. Zachara, J.M.; Gassman, P.L.; Smith, S.C.; Taylor, D. Oxidation and adsorption of Co(II)EDTA2_ complexes in subsurface materials with iron and manganese oxide grain coatings. Geochim. Cosmochim. Acta. **1995**, *59*, 4449–4463.
18. Brooks, S.C.; Taylor, D.L.; Jardine, P.M. Reactive transport of EDTA-complexed cobalt in the presence of errihydrite. Geochim. Cosmochim. Acta. **1996**, *60*, 1899–1908.
19. Read, D.; Ross, D.; Sims, R.J. The migration of uranium through clashach sandstone: the role of low molecular weight organics in enhancing radionuclide transport. J. Contam. Hydrol. **1998**, *35*, 235–248.
20. Szecsody, J.E.; Zachara, J.M.; Chilakapati, A.; Jardine, P.M.; Ferrency, A.S. Importance of flow and particlescale heterogeneity on Co(II/III)EDTA reactive transport. J. Hydrol. **1998**, *209*, 112–136.
21. Davis, J.A.; Kent, D.B.; Coston, J.A.; Hess, K.M.; Joye, J.L. Multispecies reactive tracer test in an aquifer with spatially variable chemical conditions. Water Resour. Res. **2000**, *36* (1), 119–134.
22. Mayes, M.A.; Jardine, P.M.; Larsen, I.L.; Brooks, S.C.; Fen-dorf, S.E. Multispecies transport of metal-EDTA complexes and chromate through undisturbed columns of weathered, fractured saprolite. J. Contam. Hydrol. **2000**, *45*, 243–265.
23. Solomon, D.K.; Marsh, J.D.; Larsen, I.L.; Wickliff, D.S.; Clapp, R.B. *Transport of Contaminants During Storms in the White Oak Creek and Melton Branch Watersheds*; ORNL/TM-11360; Oak Ridge National Laboratory: Oak Ridge, TN, 1991.

24. Serne, R.J.; Burke, D.S. *Chemical Information on Tank Supernatants, Cs Adsorption from Tank Liquids onto Hanford sediments, and Field Observations of Cs Migration from Past Tank Leaks, Report No. PNNL-11495*; Pacific Northwest National Laboratory: Richland, WA, 1997.
25. Gorby, Y.A.; Lovley, D.R. Enzymatic uranium precipitation. Environ. Sci. Technol. **1992**, *26*, 205–207.
26. Francis, A.J.; Dodge, C.J.; Lu, F.; Halada, G.P.; Clayton, C.R. XPS and XANES studies of uranium reduction by Clostridium Sp. Environ. Sci. Technol. **1994**, *28*, 636–639.
27. Gorby, Y.A.; Caccavo, F.; Drektrah, D.B.; Bolton, H. Microbial reduction of Co(III)EDTA_ in the presence and absence of manganese (IV) dioxide. Environ. Sci. Technol. **1998**, *32*, 244–250.
28. Brooks, S.C.; Carroll, S.L.; Jardine, P.M. Sustained bacterial reduction of Co(III)EDTA_ in the presence of competing geochemical oxidation during dynamic flow. Environ. Sci. Technol. **1999**, *33*, 3002–3011.
29. Wielinga, B.; Bostick, B.; Rosenzweig, R.F.; Fendorf, S. Inhibition of bacterially promoted uranium reduction: ferric (hydr)oxides as competitive electron acceptors. Environ. Sci. Technol. **2000**, *34*, 2190–2195.
30. Smith, M.S.; Thomas, G.W.; White, R.E.; Ritonga, D. Transport of Escherichia Coli through intact and disturbed soil columns. J. Environ. Qual. **1985**, *14*, 87–91.
31. McKay, L.D.; Cherry, J.A.; Bales, R.C.; Yahya, M.T.; Gerba, C.P. A field example of bacteriophage as tracers of fracture flow. Environ. Sci. Technol. **1993**, *27*, 1075–1079.
32. Tiedje, J.M. Microbial degradation of ethylenedi aminetet-raacetic acid in soils and sediments. Appl. Environ. Microbiol. **1975**, *30*, 327–329.
33. Tiedje, J.M.; Mason, B.B. Biodegradation of nitrilotriacetic acid (NTA) in soils. Soil Sci. Soc. Am. Proc. **1974**, *38*, 278–283.
34. Means, J.L.; Kucak, T.; Crerar, D.A. Relative degradation rates of NTA, EDTA, and DTPA and environmental implications. Environ. Pollut. Ser. B. **1980**, *1*, 45–60.
35. Bolton, H., Jr.; Li, S.E., Jr.; Workman, D.J.; Girvin, D.C. Biodegradation of synthetic chelates in subsurface sediments from the southeast coastal plain. J. Environ. Qual. **1993**, *22*, 125–132.
36. Bolton, H.; Girvin, D.C.; Plymale, A.E.; Harvey, S.D.; Workman, D.J. Degradation of metal-nitrilotriacetate complexes by *Chelatobacter heintzii*. Environ. Sci. Technol. **1996**, *30*, 931–938.
37. Bolton, H., Jr.; Girvin, D.C., Jr. Effect of adsorption on the biodegradation of nitrilotriacetate by *Chelatobacter heintzii*. Environ. Sci. Technol. **1996**, *30*, 2057–2065.
38. Payne, J.W.; Bolton, H.; Campbell, J.A.; Xun, Y.L. Purification and characterization of EDTA monooxygenase from the EDTA-degrading bacterium BNC1. J. Bacteriology 1998, *180*, 3823–3827.
39. Liu, Y.; Louie, T.M.; Payne, J.; Bohuslavek, J.; Bolton, H.; Xun, L.Y. Identification, purification, and characterization of iminodiacetate oxidase from the EDTA-degrading bacterium BNC1. Applied Environ. Microbiol. **2001**, *67*, 696–701.
40. Jardine, P.M.; Mehlhorn, T.L.; Larsen, I.L.; Bailey, W.B.; Brooks, S.C.; Roh, Y.; Gwo, J.P. Influence of hydrological and geochemical processes on the transport of chelated-metals and chromate in fractured shale bedrock. J. Conta- min. Hydrol. **2001**, *in press.*

34

Rare Earth Elements

Zhengyi Hu,
Gerd Sparovek,
Silvia Haneklaus,
and Ewald Schnug

Introduction

The rare earth elements (REEs) comprise the elements scandium (Z/21) and yttrium (Z/ 39), and 15 lanthanides with successive atomic numbers (Z) from 57 to 71 (Table 1). Rare earth elements are applied to soils as fertilizer materials or as contaminations of industrial sludges so that an assessment of their behavior in soils is required for evaluating agro-environmental effects.[1–6] This entry summarizes the present knowledge of soil chemical properties of REEs.

TABLE 1 Symbol and Atomic Numbers of Rare Earth Elements (REEs)

Elements	Symbol	Atmomic Number	Descriptive Classification
Scandium	Sc	21	
Yttrium	Y	39	Light earths
Lanthanum	LA	57	
Cerium	Ce	58	
Praseodymium	Pr	59	
Neodymium	Nd	60	
Promethium	Pm	61	
Samarium	Sm	62	
Europium	Eu	63	
Gadolinium	Gd	64	
Terbium	Tb	65	Heavy earths
Dysprosium	Dy	66	
Holmium	Ho	67	
Erbium	Er	68	
Thulium	Tm	69	
Ytterbium	Yb	70	
Lutetium	Lu	71	

Source: Liu.[1]

TABLE 2 Concentrations and Coefficient of Variation of Rare Earth Elements in Different Sludges

	Night Soil Sludge[a] (n/10)		Sewage Sludge (n/14)		Food Industry Sludge (n/10)		Chemical Industry Sludge (n/10)	
Elements	Mean (mg kg^{-1})	CV (%)	Mean (mg kg^{-1})	CV (%)	Mean (mg kg^{-1})	CV (%)	Mean (mg kg^{-1})	CV (%)
La	3.39	37	6.70	47	0.89	72	2.46	98
Ce	6.98	44	14.10	58	1.83	77	2.69	105
Pr	0.82	38	1.48	46	0.22	82	0.48	95
Nd	3.18	34	6.00	47	0.91	82	2.04	98
Sm	0.53	36	1.02	40	0.17	81	0.36	95
Gd	0.53	34	1.18	45	0.17	79	0.48	101
Tb	0.07	45	0.16	36	0.03	81	0.06	103
Dy	0.39	53	0.93	33	0.14	76	0.39	110
Ho	0.07	54	0.19	32	0.03	79	0.09	119
Er	0.21	55	0.57	31	0.08	73	0.26	118
Tm	0.03	52	0.08	26	0.01	81	0.03	112
Yb	0.20	58	0.54	31	0.09	83	0.19	109
Lu	0.03	56	0.08	31	0.01	84	0.03	105

[a] Feces and urine of humans.
Source: Kawasaki and Kimura.[5]

Origins of Soil Rare Earth Elements

Soil REEs mainly originate from parent materials.[1] Application of phosphate fertilizer[3] and phosphogypsum[4] can supply REEs to the soil. Some of the sludges, particularly those from the chemical industry, have been contaminated with REEs (Table 2). Continuous application of sewage sludges caused an accumulation of Sc, Sm in some soils in Japan.[6] The use of REEs in agriculture is widely practiced in China. By 2001, 6.5 million hectares of land in China was treated with REE fertilizers.[2] In total, 11,000 tons of REEs was applied in agricultural production in China.[7] Besides the parent material, the application of REEs on agricultural farmland is going to be a major source of REEs if the practice of applying them regularly proceeds.

Chemical Speciation of Rare Earth Elements in Soils

Total Content

Representative background values of REEs in soils are available so far only for China and Japan.[1,2,8] The REE content strongly depends on the parent material.[1] The results of 853 soil analyses showed that the total REE content in soils varied between 18 and 583 mg kg^{-1}, with a mean value of 184 mg kg^{-1}.[2] The light REEs La, Ce, Pr, Nd, Sm and Eu account for 90% of the total REE content in soils (Table 3). On the average, the La, Ce, Nd, Sm, and Eu content was 41, 73, 7.3, 27.5, and 5.6 mg kg^{-1}, respectively, in different soils in China *(n*/467) (Table 3).

Species of REEs in Soils

Binding forms of REEs in soils may be classified according to their availability for plants.[1] Approximately 9 mg kg^{-1} is exchangeable, 2, 5, 32, and 95 mg kg^{-1} are bonded to carbonates, Mn-oxides, organic substances, and amorphous Fe-oxides, while 59 mg kg^{-1} is abundant in the form of crystal iron oxides and

TABLE 3 Mean Content of REEs in Soils Extracted by Na_2O_2/NaOH (n/467)

Light REEs	Content (mg kg^{-1})	Heavy REEs	Content (mg kg^{-1})	Total Contents (mg kg^{-1}) and ratios	
La	41.2	Gd	4.8	Total REE (T)	172.8
Ce	73.4	Tb	0.7	Light REE (L)	156.0
Pr	7.3	Dy	4.4	Heavy REE (H)	16.8
Nd	27.5	Ho	0.9	L/H	9.3
Sm	5.6	Er	2.7	L/T	0.9
Eu	1.1	Tm	0.4		
		Yb	2.5		
		Lu	0.4		

Source: Liu.[1]

TABLE 4 Mean Content of Different Species of Rare Earth Elements in Soils (mg kg^{-1})

Soils	*n*	Exchangeable	Carbonate Bonded	Mn-oxide Bonded	Organically Bonded	Amorphous Fe-oxide Bonded	Crystal ion Oxides	Residual	Total
Latosol	3	1.0	—	1.4	24.3	23.8	8.2	17.5	76.2
Red soil	5	14.8	—	8.0	30.1	108.0	33.9	199.8	394.6
Yellow brown soil	2	12.1	—	7.4	26.1	79.3	88.3	55.5	268.7
Brown soil	2	11.5	—	4.7	33.7	78.8	64.1	34.5	227.3
Black soil	2	1.4	5.3	1.2	46.8	88.8	74.8	64.0	282.3
Chernozem	2	1.8	11.4	2.8	41.2	105.3	94.1	79.0	335.5
Mean		9.1	1.9	5.3	32.2	95.4	58.8	105.3	307.9

Note: Latosol (rhodic ferralsol, FAO); red soil (ferralic cambisol); yellow brown soil (haplic luvisol); brown soil (haplic alisol); black soil (luvic phaeozems); chernozem (haplic chernozems).
Source: Liu.[1]

105 mg kg^{-1} in residual forms (Table 4). Plants utilize exchangeable REEs most easily, while the uptake of other forms is limited.[1,9] In contrast, residual forms are not plant available.[1,9]

For determining the content of plant available REEs the following extractants have been proposed: 1 M HAc–NaAc (pH 4.8),[1,9] 1.0 M NH_4NO_3 (pH 7.0),[10] 0.1 M HCl,[11] and 0.1 M malic-citric acid,[12] with 1 M HAc–NaAc being most extensively used.[1,2] Plant available REE contents are highly variable and range from <1 to >200 mg kg^{-1}, with a mean value of about 12 mg kg^{-1} (n/1790).[1] Physicochemical soil properties such as pH, Eh, CEC, clay, $H_2PO_4^-$, and carbonate content have a strong impact on the amount of exchangeable and plant available REEs.[1,9] Acid soils contain significantly higher amounts of plant available REEs than more alkaline, calcareous soils.[2] The availability of soil applied REEs is usually significantly higher than that in the original soil matrix.[1,9]

Adsorption of Rare Earth Elements in Soils

In general, 95% of the added REEs are adsorbed.[13] Rare earth elements added to soils are rapidly transformed, for example, into exchangeable, organic matter bonded, and Fe/Mn oxide bonded species.[1,9] The distribution coefficients for REEs added to red and yellow brown soils declined in the following order: residual > exchangeable > organic matter bonded > Fe/Mn oxide bonded REEs.[9] The formation of bridged hydroxo complexes is probably the dominant sorption mechanism to clay minerals.[14] Clay type, pH, CEC, organic matter, and amorphous iron content regulate the adsorption kinetics of REEs.[1,2,9] Langmuir and Freundlich equations were found to describe precisely the absorption of REEs in soils.[15]

Translocation of Rare Earth Elements in Soils

The question is still open whether the use of REEs in industry and agriculture may result in a pollution of soils, plants, and groundwater. In leaching experiments under controlled conditions with ^{141}Ce and ^{147}Nd, these elements were abundant only in the top soil layer because of their strong adsorption.[9] For field conditions a translocation depth of <1 cm has been estimated,[9] but this value still needs to be validated for different soils and on a long-term basis.

Uptake of Rare Earth Elements by Plants

The uptake of REEs has been investigated for a wide range of plants.[2,9] All rare elements except praseodymium were found in plants. Uptake of REEs was related to many factors such as element, plant type, and growth conditions.[2,9] The light REEs Ce, La, Nd were highest in plants.[2,9] The total concentration of REEs in plant tissue ranges from <0.05 to 2.58 mg kg^{-1} (Table 5). The concentration of Ce in food products is usually lower than 0.2 mg CeO_2 kg^{-1} (Table 6). More attention should be paid to research on uptake, translocation, and distribution of REEs in plants in order to follow up the biological effects of REEs and to assess agro-environmental effects

TABLE 5 Total Concentration of Rare Earth Elements in Plants (mg kg^{-1})

Species	*n*	Min (mg kg^{-1})	Max (mg kg^{-1})
Rice	319	<LLD[a]	1.17
Wheat	440	<LLD	2.58
Corn	139	<LLD	0.92
Cucumber	41	<LLD	0.70
Leek	33	0.04	0.21
Spinach	41	<LLD	0.12
Cauliflower	61	<LLD	0.60
Lotus root	31	<LLD	0.76
Tomato	64	<LLD	0.18
Chinese cabbage	67	0.05	1.01
Pepper	31	<LLD	0.40
Potato	34	0.05	0.35
Cabbage	38	<LLD	1.20
Mushroom	33	<LLD	0.45
Orange	41	0.13	0.70
Litchi	30	<LLD	
Grape	61	<LLD	0.85
Longan	30	<LLD	0.71
Banana	33	<LLD	
Apple	62	0.07	0.80
Pear	34	<LLD	0.24
Watermelon	37	<LLD	0.26
Sugarcane	27	0.05	1.25
Peach	4	<LLD	

[a] LLD: Lower limit of detection.

Source: Xiong.[9]

TABLE 6 Ce content in Grains and Seeds of Different Crops in Different Countries and Regions of China (mg CeO_2 kg^{-1})

Crops	Region/Country	Sample No.	Mean	Min	Max
Rice	Hubei/China	8	0.06	0.02	0.15
Rice	Jiangxi/China	21	0.06	<0.01	0.18
Rice	Beijing/China	17	0.04	<0.01	0.11
Rice	Nanjing/China	2	—		0.12
Wheat	Heilongjiang/China	21	0.04	<0.01	0.09
Wheat	Hubei/China	7	0.04	0.01	0.07
Wheat	Henan/China	6	0.10	<0.01	0.19
Wheat	Shandong/China	4	0.12	<0.07	0.17
Wheat	Beijing/China	10	0.03	<0.01	0.05
Wheat	Tianjin/China	1	0.04	<0.01	0.08
Maize	Hubei/China	8	0.01	<0.01	0.02
Maize	Heilongjiang/China	10	0.04	<0.01	0.10
Maize	Tianjiang/China	15	0.04	<0.01	0.00
Barley	Beijing/China	14	0.04	<0.01	0.25
Wheat	Canada	5		0.04	0.16
Wheat	USA	8		<0.01	0.20
Wheat	Australia	4		0.06	0.16
Wheat	Argentina	4		0.07	0.09
Soybean	USA	3		<0.01	0.11
Maize	USA	3		0.01	0.09
Mung bean	Thailand	1			0.09

Source: Xiong.[9]

Crop Response to Rare Earth Elements

Research with a view to the use of REEs in agriculture was predominantly carried out in China and before in Russia. A small number of studies have been carried out in Australia, too.[2,9] The increases in crop yield reported by workers from all parts of China range between 5% and 103% (Table 7), with an average response of 8%–15%.[2,9] Crop response to REEs is reported to be most probable when soils contain less than 10 mg kg^{-1} of available REEs (in 1 M HAc–NaAc, pH 4.8), while a response on soils with more than 20 mg kg^{-1} of available REEs is unlikely (Table 8). So far, there is no evidence that REEs are essential for plant growth.

TABLE 7 Yield Increase (relative to control) after Application of Rare Earth Elements to Different Crops

Crop	Country	Yield Increase (%)
Sugar beet	Bulgaria	17–24
Sugar beet	China	7
Wheat	China	6–17
Rape	China	4–48
Potato	China	5–6
Soybean	China	8–9
Cotton	China	5–12
Rice	China	7

(*Continued*)

TABLE 7 (*Continued*) Yield Increase (relative to control) after Application of Rare Earth Elements to Different Crops

Crop	Country	Yield Increase (%)
Corn	China	9–103
Barley	Australia	18–19
Peanut	China	8–12
Tobacco	China	8–10
Rubber	China	8–10
Sugarcane	China	10–15
Cabbage	China	10–20
Litschi	China	14–17
Grape	China	8–12

Source: Hu.[16]

TABLE 8 Critical Values of Plant Available Rare Earth Elements in Soils

REE Content (mg kg^{-1})	Index	Crop Response to REE
<5.0	Very low	Most likely
5–10	Low	Probable
11–15	Medium	Not expected
16–20	High	Not expected
>20	Very high	Unlikely

Source: Liu[1] and Xiong.[9]

Conclusions

Basic information about the chemistry of REEs in soils is available, but this data refers to specific regions and soil types. Besides this, the effect of REEs, for example, on soil fertility and soil biological diversity is yet unclear. Crop response to REE applications depends on various factors, including soil properties. Further studies are also required in order to elaborate the most efficient application techniques.

References

1. Liu, Z. Rare earth elements in soils. In *Microelements in Soils of China*; Liu, Z., Ed.; Jiangsu Science and Technology Publishing House: Nanjing, China, 1996; 293–329.
2. Xiong, B.K.; Chen, P.; Guo, B.S.; Zheng, W. *Rare Earth Element Research and Application in Chinese Agriculture and Forest*; Metallurgical Industry Press: Beijing, China, 2000.
3. Todorovsky, D.S.; Minkova, N.L.; Bakalova, D.P. Effect of the application of superphosphate on rare earth content in the soils. Sci. Total Environ. **1997**, *203* (1), 13–16.
4. Arocena, J.M.; Rutherford, P.M.; Dudas, M.J. Heterogeneous distribution of trace elements and fluorine in phos-phogypsum. Sci. Total Environ. **1995**, *162* (2–3), 149–160.
5. Kawasaki, A.; Kimura, R.; Arai, S. Rare earth elements and other trace elements in wastewater treatment sludges. Soil Sci. Plant Nutr. **1998**, *44* (3), 433–441.
6. Zhang, F.S.; Yamasaki, S.; Kimura, K. Rare earth element content in various waste ashes and the potential risk to Japanese soils. Environ. Int. **2001**, *27* (5), 393–398.
7. Wang, S.M.; Zheng, W. Developing technology of rare earth application on agriculture and biological field in China. *Proceedings of the 4th International Conference on* Rare Earth Development and Application, Beijing, China, June 16–18, 2001; Yu, Z.S., Yan, C.H., Xu, G.Y., Niu, J.K., Chen, Z.H., Eds.; Metallurgical Industry Press: Beijing, 237–240.

8. Yoshida, S.; Muramatsu, Y.; Tagami, K.; Uchida, S. Concentrations of lanthanide elements, Th, and U in 77 Japanese surface soils. Environ. Int. **1998**, *24* (3), 275–286.
9. Xiong, P.K.; Zheng, W.; Cheng, P.; Wang, F. *Rare Earth Elements in Agricultural Environment*; China Forestry Press: Beijing, China, 1999.
10. Zhai, H.; Yang, Y.G.; Zheng, S.J.; Hu, A.T.; Zhang, S.; Wang, L.J. Selection of the extractants for available rare earths in soils. China Environ. Sci. **1999**, *19* (1), 67–71 (in Chinese).
11. Li, F.L.; Shan, X.Q.; Zhang, S.Z. Evaluation of single extractants for assessing plant availability of rare earth elements in soils. Commun. Soil Sci. Plant Anal. **2001**, *32* (15 and 16), 2577–2587.
12. Zhang, S.Z.; Shan, X.Q.; Li, F.L. Low-molecularweight-organic-acid as extractant to predict plant bioavailability of rare earth elements. Int. J. Environ. Anal. Chem. **2000**, *76* (4), 283–294.
13. Zhu, J.G.; Xing, G.X.; Yamasaki, S.; Tsumura, A. Adsorption and desorption of exogenous rare earth elements in soils: I. Rate of forms of rare earth elements sorbed. Pedo-sphere **1993**, *3* (4), 299–308.
14. Dong, W.M.; Wang, X.K.; Bian, X.Y.; Wang, A.X.; Du, J.Z.; Tao, Z.Y. Comparative study on sorption/desorption of radioeuropium on alumina, bentonite and red earth: effects of pH, ionic strength, fulvic acid, and iron oxides in red earth. Appl. Radiat. Isotopes **2001**, *54* (4), 603–610.
15. Gao, X.J.; Zhang, S.; Wang, L.J. The adsorption of La^{3+} and Yb^{3+} on soil and mineral and its environmental significance. China Environ. Sci. **1999**, *19* (2), 149–152 (in Chinese).
16. Hu, Z.Y.; Richter, H.; Sparovek, G.; Schnug, E. Physiological and biochemical effects of rare earth elements on plants and their agricultural significance: a review. J. Plant Nutr. **2003**, in press.

35

Strontium

Silvia Haneklaus
and Ewald Schnug

Introduction

Strontium (Sr) was discovered by Adair Crawford in 1790 and named after the western Scottish village of Strontian. There are four stable Sr isotopes, ^{84}Sr, ^{86}Sr, ^{87}Sr, and ^{88}Sr, which occur naturally in a ratio of 0.56:9.87:7.00:82.58.[1] From these four stable Sr isotopes, only ^{87}Sr is radiogenic. Thus, the ^{87}Sr concentration increases over time because of the decay of ^{87}Rb to ^{87}Sr (half-life of 4.7×10^{10} years). It is, however, the radioactive isotope ^{90}Sr from anthropogenic nuclear sources that this element is commonly associated with. Plant available concentrations of Sr and Ca in soils are important factors that affect the transfer$_{\text{soil/plant}}$ of ^{90}Sr.[2] Consequently, the knowledge of the factors and processes influencing the chemical and spatial speciation of Sr in the soil is of prime interest in risk assessment of undesired ^{90}Sr contaminations.

Strontium in Soils

Strontium occurs in abundance in nature, and its concentration can be as much as 0.034% in most igneous rocks.[3] Celestite ($SrSO_4$) and strontianite ($SrCO_3$) are the predominant Sr minerals. It occurs in large concentrations and in association with K in volcanic rocks, alkali rocks, and pegmatites. Sr is removed by weathering from these and other igneous rocks and sediments. Intermediate concentrations of Sr occur in basic magmatic rocks, which contain about two times higher Sr concentrations than granite. Some ultramafic rocks have very low Sr concentrations.[4] A mean Sr concentration of 375 mg kg^{-1} is observed in igneous rocks, 300 mg kg^{-1} in shales and 20 mg kg^{-1} in sandstones.[5]

Strontium isotope ratios in parent materials and minerals vary in relation to geological age and geographical location.[6] The ^{87}Sr: ^{86}Sr ratio depends on the initial Rb:Sr ratio in the parent material, age of the rock, and the Sr isotope ratio at the onset of rock formation as ^{87}Sr is radiogenic.[6] The ^{87}Sr:^{86}Sr ratio is often used to study biogeochemical processes such as weathering of minerals. However, this ratio must be evaluated critically, as with the age of weathering, surface changes in the ^{87}Sr:^{86}Sr ratio also occur, presumably because of dynamic variations in weathering rates of different minerals.[7]

TABLE 1 Variation in the Sr Content in Dependence on Chemical Speciation and Geogenic Origin

Fraction	Country	Method	Min. ($\mu g\ g^{-1}$)	Max.	Mean	n	Reference
Sr_{total}	N. Germany[a]	X-RF	15.2	118.4	79.0	155	[2]
	N. Germany[b]	X-RF	57.0	97.0	71.3	154	[9][c]
	Finland		109	2730	744	221	[10]
	N. Europe[d]	X-RF	10	667	110[e]	773	[4]
	Scotland[f]	–	60	1000	380	43	[11]
	Russia	–	53	135	92	–	[12]
	India	$HF\text{-}HNO_3\text{-}HClO_4$	74	198	127	10	[13]
	U.S.A.	–	80	900	400	13	[11]
$Sr_{total\ exchangeable}$	N. Germany[a]	1N NH_4-Ac, pH 7	4.4	48.9	18.1	155	[2]
	Finland	NH_4-Ac, pH 4.65	9.4	16.6	12.9	1944	[14][g]
	Finland	NH_4-Ac, pH 4.65	2.4	63.8	19.2	221	[10]
	N. Europe	NH_4-Ac, pH 4.5	0.04	473	8[e]	773	[4]
	Scotland[h]	1N NH_4-Ac, pH 7	3.5	17.5	6.6	7	[11]
$Sr_{rapidly\ exchangeable}$	N. Germany[a]	0.025 N $CaCl_2$	1.5	10.2	4.8	155	[2]

[a] Schleswig-Holstein.
[b] Isle of Ruegen.
[c] Sample description (Sr data unpublished).
[d] Belarus, Estonia, Finland, Germany, Latvia, Lithuania, Norway, Poland, Russia, Sweden.
[e] Median.
[f] North-east Scotland.
[g] $mg\ L^{-1}$.
[h] Aberdeenshire/Midlothian.

In the soil profile, about 90% of the Sr is generally exchangeable and 10% in non-exchangeable form.[8] The data in Table 1 summarize total and plant available Sr concentrations in soils of different origins. Reimann et al.[4] reported distinctly lower total Sr concentrations in soils of northern Europe with a median of 110 mg kg^{-1} Sr in comparison with the world average of 240 mg kg^{-1} Sr. In several countries of northern Europe, median values for total Sr concentrations range from 45 mg kg^{-1} in Poland to 181 mg kg^{-1} in Norway. About 7% of the total Sr concentration in the topsoil is plant available.[4] The reported plant available Sr concentrations range from 4 mg kg^{-1} in Poland to 15 mg kg^{-1} in Finland.[4] The corresponding values for Germany are 8 (NH_4-Ac) and 55 (total) mg kg^{-1} Sr, and somewhat lower than those reported by Haneklaus[2] in Schleswig-Holstein with 18 (NH_4-Ac) mg kg^{-1} and 79 (total) mg kg^{-1} Sr (Table 1). The mean rapidly exchangeable Sr concentration in these studies was 4.8 mg kg^{-1} Sr.

Interaction of Strontium with Soil Matrix

The advantage of using stable Sr in research is that it provides a direct measure for the spatial variation of pedogenetic soil characteristics and agro-technical measures influencing the uptake of radioactive Sr isotopes by plants. In comparison, studies with radioactive Sr (viz ^{90}Sr, ^{89}Sr, and ^{85}Sr) provide in situ analysis of radioactive Sr behavior under current soil conditions. An important point to be noted in Sr experimentation is whether the stable or carrier-free radioactive Sr has been used, which leads to application of significant amounts of Sr in the former case, compared with only insignificant rates of Sr added to the soil in the latter case.

The correlation matrix between soil pH, clay, organic C, plant available Ca concentration, and Sr species is shown in Figure 1 for soils from the federal state of Schleswig- Holstein.[2] Correlation coefficients between total exchangeable Sr and C_{org}. were significant, but low. In comparison, distinctly close and significant relationships were observed for soil pH and clay content (Figure 1). Reimann et al.[4] and Haneklaus[2] also reported a similar correlation coefficient of $r = 0.6$ for the relationship between total Sr and NH_4-acetate extractable Sr when total Sr was determined by *aqua regia* extraction and

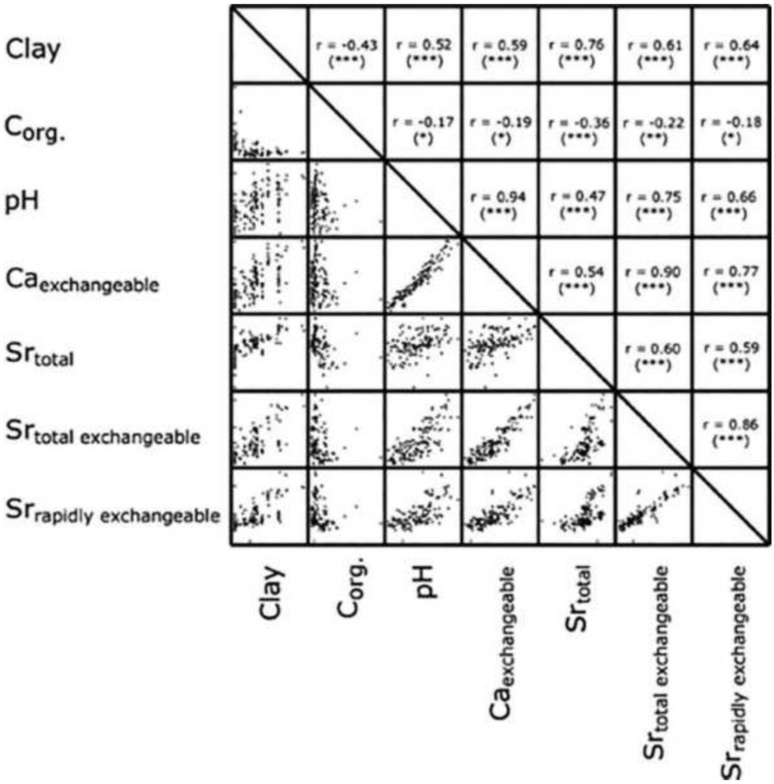


FIGURE 1 Correlation between soil parameters and Sr species in top soil samples of production fields in northern Germany (n = 155).
Source: Calculated from data in Schöller.[2]

X-RF techniques, respectively. Analogous correlation coefficients were also observed for the relationship between plant available Ca ($r = 0.78$) and plant available Sr concentrations ($r = 0.90$) in the topsoil (Figure 1).[2,4]

Rengel estimated a mean Ca loss of 300 kg ha^{-1} yr^{-1} by leaching. The Ca:Sr ratio in soils is about 140 for total[15] and may vary between 250 and 5000 for exchangeable Ca and Sr.[16] Assuming equivalent leaching rates of Ca and Sr, implies Sr losses of <1 kg ha^{-1} yr^{-1}. Wiklander[8] observed considerable leaching losses of ^{90}Sr in a lysimeter experiment after 42 months.

Strontium sorption is dominated by ion exchange and sorption/desorption is a reversible process; the organic soil fraction preferentially adsorbs Ca, while the inorganic soil fraction (e.g. the clay minerals) preferentially adsorbs Sr.[17,18] A small fraction of ^{85}Sr added to the soil is also fixed in non-exchangeable form. Valcke[17] suggested that this fixation occurs in Sr/organic matter/clay complexes.

Adsorption of Sr on different types of clay minerals has been studied intensively with the objective to decontaminate effluents from contaminated sites. Strontium is adsorbed moderately to mineral surfaces, and is desorbed when its concentration in the soil solution increases, competing ions such as Ca are present, and soil pH is low.[19] Application of even small quantities of stable Sr and Ca as amendments may increase uptake of ^{90}Sr by desorption. However, application of Ca may positively increase the Ca:Sr ratio.[20] The Sr sorption decreases in the presence of calcite and high concentrations of HCO_3 and strontianite ($SrCO_3$) may be formed at high concentrations of Sr in the soil.[19] For a Cambrian blue clay soil, the Langmuir isotherm indicated a constant of 294 and a limiting sorption of 0.034 mol kg^{-1}.[18] Also Freundlich isotherms reflected an adequate sorption of Sr with a regression coefficient of >0.99. The sorption ratio is generally higher than 0.95 for all concentrations ranging from 10^{-3} to 10^{-8} M.[21]

Strontium Uptake by Plants

Strontium uptake by plants is passive, and is thus linked to its concentration in the soil solution. Plant available Ca and Sr in the soil are important factors governing uptake of radioactive Sr. Addition of Ca often results in a negative linear relationship between Sr and Ca in plants.[2] The Sr: CEC ratio influences Sr uptake in such a way so that any increase in the cation exchange capacity decreases the Sr:uptake.

The Sr concentration in plants varies depending on the plant available concentrations in the soil, root uptake, and the growth or phenological stage. Strontium uptake by dicotyledonous crops is, for example, distinctly higher than that by monocotyledonous crops owing to the higher cation exchange capacity of the latter.[20] The Sr concentration in vegetative plant tissue varies between 1 and 150 mg kg^{-1} with a mean value of 25 mg kg^{-1}.[3] The Sr concentration in seeds is about 5–10 times lower than that in vegetative plant material, most likely because of the phloem immobility.[20] Strontium concentrations from 18 to 36 mg kg^{-1} have been reported in oilseed rape seeds, while the average concentration was 7 mg kg^{-1} in wheat.[2]

Impact of Fertilizer Use on Strontium Uptake by Plants

Gabe and Rodella[22] reported Sr concentrations in different phosphatic fertilizers, limestone, and gypsum. The mean Sr concentration (mg kg^{-1}) occurs in the following order: concentrated apatite (12121) > rock phosphates (7319) > single super phosphate (5396) > fused phosphates (4965) > gypsum (2984) > limestone (824) > ammonium phosphates (304). The variation in the Sr concentration was highest in limestone samples ranging from 36 to 4736 mg kg^{-1} Sr. Long-term application of lime significantly increased soil pH along with total Ca and Sr concentrations in an acid Albic luvisol in Romania (Figure 2).[23] Haneklaus[2] reported that the Sr uptake in plants is influenced more by the Ca concentration in soil rather than by soil pH. Liming of an acid soil can reduce the Sr uptake by 40%–45%, as was reported in a lysimeter experiment by Wiklander.[8] Under field conditions, Haneklaus[2] estimated a minimum threshold rate of lime application of 1.4 t ha^{-1} CaO for reducing the Sr uptake significantly on an acid, silty loam soil (pH 5.3). Liming combined with an additional Ca supply by gypsum can cause the maximum reduction in Sr uptake by oilseed rape probably because of the precipitation of $SrSO_4$.[2] In contrast, Anderson[20] reported that a dose of at least 10 t ha^{-1} stable Sr is required to achieve a significant reduction of^{90}Sr uptake.

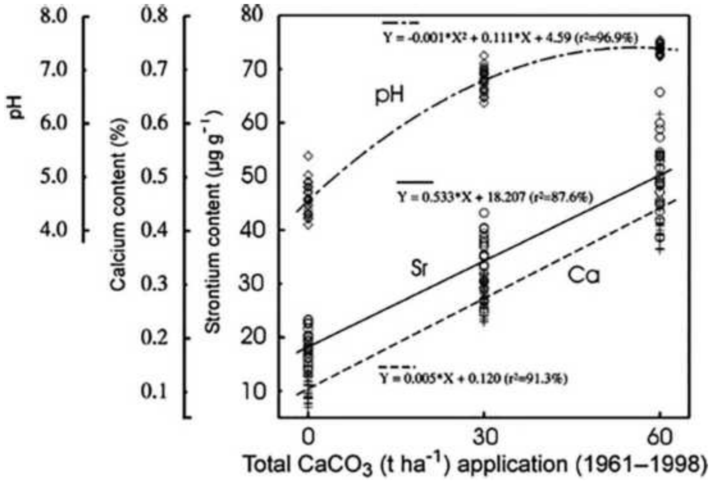


FIGURE 2 Long-term influence of regular lime applications of 0, 5, and 10 t ha^{-1} from 1961 to 1998 corresponding to total rates of 0, 30, and 60 t ha^{-1} on soil pH, and total Sr and Ca contents in the top soil of an Albic luvisol in Romania. **Source:** Adapted from Rogasik et al.[23]

Conclusion

Strontium behavior in soils and uptake by plants are similar to that of Ca. Stable Sr is usually of minor interest, but it may gain relevance with its role in reducing the transfer of radioactive Sr into the food chain. An optimum Ca saturation is a prerequisite to limit the uptake of radioactive Sr by plants. Furthermore, amendments (e.g., lime) add significant amounts of Sr to the soil and thus increase the concentration of competitive stable Sr in soils.

References

1. Weast, R.C.; Astle, M.J.; Beyer, W.H. *Handbook of Chemistry and Physics;* CRC Press Inc.: Boca Raton, Florida, 1988/89.
2. Haneklaus, S. *Strontiumgehalte in Pflanzen und Böden Schleswig-Holsteins und Bewertung von Düngungsmaßnahmen zur Verminderung der Strontiumaufnahme von Kulturpflanzen*; PhD, University, Kiel, Germany, 1989; 169 *pp*.
3. Coughtrey, P.J.; Thorne, M.C. *Radionuclide Distribution and Transport in Terrestrial and Aquatic Ecosystems. A Critical Review of Data*; Balkema: Rotterdam, 1983; Vol. 1, 93–239.
4. Reimann, C.; Siewers, U.; Tarvainen, T.; Bityukova, L.; Eriksson, J.; Gilucis, A.; Gregorauskiene, V.; Lukashev, V.K.; Matinian, N.N.; Pasieczna, A. *Agricultural Soils in Northern Europe: A Geochemical Atlas;* Bundesanstalt fuer Geowissenschaften und Rohstoffe and Staatliche Geologische Dienste in der Bundesrepublik Deutschland, Ed.; Schweizerbart'sche Verlagsbuchhandlung: Stuttgart, 2003; 279 pp.
5. Lisk, D.J. Trace metals in soils, plants, and animals. Adv. Agronom. **1972**, *24*, 267–311.
6. Barbaste, M.; Robinson, K.; Guilfoyle, S.; Medina, B.B.; Lobinski, R. Precise determination of the strontium isotope ratios in wine by inductively coupled plasma sector field multicollector mass spectrometry (ICPSF-MC-MS). J. Anal. At. Spectrom. **2002**, *17*, 135–137.
7. Ma, Y.; Liu, C. Sr isotope evolution during chemical weathering of granites. Sci. China **2001**, *44*, 726–734.
8. Wiklander, L. Uptake, adsorption and leaching of radiostrontium in a lysimeter experiment. Soil Sci. **1964**, *97*, 168–172.
9. Haneklaus, S.; Fleckenstein, J.; Schnug, E. Comparative studies of plant and soil analysis for the evaluation of the sulphur status of oilseed rape and wheat. J. Plant Nutr. Soil Sci. **1995**, *158*, 109–112.
10. Lakanen, E.; Silanpää , M. Strontium in Finnish soils. Annales Agric. Finniae **1967**, *6*, 197–207.
11. Swaine, D.J. The trace element content of soils. Commonwealth Bureau of Soil Science Technical Communications 1955, 48, 157.
12. Protasova, N.A.; Golubev, I.M.; Korobeinikov, N.I. Trace elements in landscapes of the Tambov Oblast and the biogeochemical zoning of its territory. Eurasian Soil Sci. **1996**, *29*, 1360–1366.
13. Moslehuddin, A.Z.M.; Laizoo, S.; Egashira, K. Trace elements in Bangladesh paddy soils. Commun. Soil Sci. Plant Anal. **1999**, *30*, 1975–1996.
14. Paasikallio, A. Strontium content and strontiumcalcium ratio in timothy (Phleum pratense L.) and soil in Finland. Annales Agric. Finniae **1979**, *18*, 174–181.
15. Comar, C.L. Some over-all aspects of strontiumcalcium discrimination. In *The Transfer of Calcium and Strontium across Biological Membranes*; Wassermann, R.H., Ed.; Academic Press: New York, 1963; 405–417.
16. Menzel, R.G.; Heald, W.R. Strontium and calcium contents of crop plants in relation to exchangeable strontium and calcium of the soil. Soil Sci. Soc. Am. Proc. **1959**, *23*, 110–112.
17. Valcke, E. The behaviour dynamics of radiocesium and radiostrontium in soils rich in organic matter; PhD thesis, Katholieke Universiteit Leuven, Belgium, 1993; 135 pp.
18. Chirkst, D.E.; Litinova, T.E.; Cheremisina, V.; Ivanov, M.V.; Mironenkova, N.A. Isotherm of strontium sorption on clay. Russ. J. Appl. Chem. **2003**, *76*, 727–730.

19. Mincher, B.J.; Fox, R.V.; Riddle, C.L.; Cooper, D.C.; Groe- newold, G.S. Strontium and cesium sorption to Snake river plain, Idaho soil. Radiochim. Acta **2004**, *92*, 55–61.
20. Andersen, A.J. Plant accumulation of radioactive strontium with special reference to the strontium-calcium relationship as influenced by nitrogen. Riso Report **1973**, *278*, 56.
21. Hakem, N.L.; Al Mahamid, I.; Apps, J.A.; Moridis, G.J. Sorption of cesium and strontium on Hanford soil. J. Radioanal. Nucl. Chem. **2000**, *246*, 275–278.
22. Gabe, U.; Rodella, A.A. Trace elements in Brazilian agricultural limestones and mineral fertilizers. Commun. Soil Sci. Plant Anal. **1999**, *30*, 605–620.
23. Rogasik, J.; Kurtinez, P.; Panten, K.; Funder, U.; Rogasik, H.; Schroetter, S.; Schnug, E. Kalkung und Bodenfruchtbarkeit. FAL—Agric. Res. **2005**, *286* (special issue): 71–81.

36

Sulfur

Ewald Schnug,
Silvia Haneklaus,
and Elke Bloem

Introduction

Soil fertility is generally defined as the ability of soils to yield crops and is therefore closely intertwined with plant nutrient cycles. The soil sulfur cycle is driven by biological and physico-chemical processes, which affect flora and fauna. The organic matter pool is an important source and sink for sulfur though its contribution to the mineral nutrition of high yielding crops has been overestimated in the past. Under temperate conditions it is rather the spatiotemporal variation of physico-chemical soil properties, which control the plant available sulfate-S content in the soil via the access of plant roots to sulfur rich groundwater or capillary ascending porous water. Therefore soil methods determining plant available sulfur status will always deliver only information for that instant ignoring relationships with the plant sulfur status or crop yield. A more promising way for the evaluation of the sulfur supply is a site-specific sulfur budget, which includes information about geomorphology, texture, climatic data and crop type and characteristics of the local soil water regime. The major soil properties and external sulfur sources affecting the amount of plant available sulfate in the soil are shown in Figure 1.

For a comprehensive understanding of the relationship between sulfur and soil fertility it is necessary to identify the main factors and processes controlling the plant available sulfur pool and to appraise the ecotoxicological significance of sulfur within the content of a limiting factor rather than a pollutant. Detailed reviews about the role of sulfur in agro-ecosystems already exist[1–9] and will provide further information, not explicitly mentioned here.

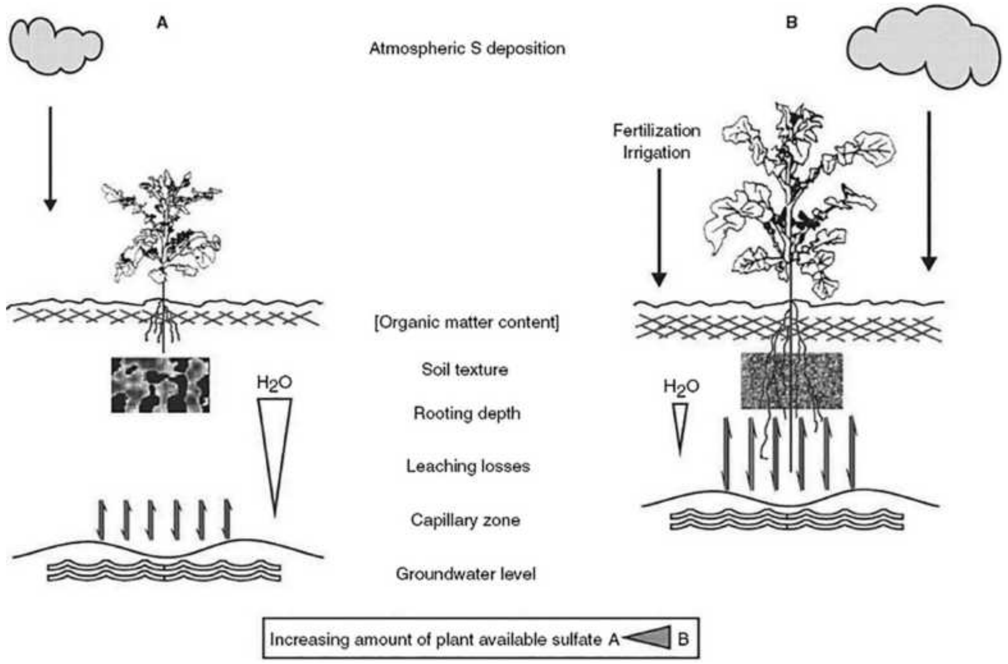


FIGURE 1 Overview and efficacy of essential impact factors modifying plant available sulfate in soils.

Sulfur in Soils

The concentration of sulfur (S) in parent materials ranges from 0.026% to 1% S with igneous rocks ≤ metamorphic rocks ≤ magmatic rocks of upper continental crust << limestones < sedimentary rocks (sulphides) ≤ shales < sedimentary rocks (sulfates) << coal.[10–12] The typical range of S in agricultural soils of humid and semi-humid regions is 100 to 500 (μg g^{-1}, or 0.01 to 0.05% S. This equals 224 to 1120 kg ha^{-1} S in the A_h-horizon.[8] The total S content of soils may be as low as 20μg g^{-1} (0.002%) in highly leached and weathered soils of humid regions or as high as 35,000 μrg g^{-1} (3.5%) in marine marsh soils and up to 50,000 μg g^{-1} (5%) in calcareous and saline soils of arid and semiarid regions.[8] Tropical soils generally contain low amounts of S due to their low organic matter content.

Biological Aspects of Sulfur in Soils

Most of the S in terrestrial soils is bound in the organic fraction, which amounts normally to more than 95% of the total S content.[2,8] Organic S in soils is a heterogeneous mixture of soil organisms, partly decomposed plant material, animal and microbial residues. Little is, however, known about the composition of individual chemical compounds. Many different approaches have been developed to separate soil organic S into major fractions. The following approaches to identify distinct forms and properties of soil organic matter were made: chemical extraction followed by physical-chemical separation into humic acids, fulvic acids and humins reactivity with reducing agents in order to separate carbon-bonded S (C–S) and sulfate esters (C–O–S); physical separation into organo-mineral size fractions and molecular weight fractionation.[13]

Sulfate-S is rapidly bound in the form of sulfate-esters, which is regarded as a short term source for S.[14] The incorporation of S into high-molecular weight fractions such as humic acids prevents the rapid mineralization of S. In litter, S is predominately carbon-bonded S (C–S) to the humin fraction to a level of about 89%. The incorporation of organic S into complex organic substances followed by association with clay minerals can lead to the relative isolation of S from decayed soil micro-organisms. This process is considered as a physical protection of organic S against degradation,[15] and it results in a decreased availability of S to plants.

Sulfur Mineralization

In soils microbial mediated processes are mainly responsible for S transformations, so that the factors affecting the microbial activity, such as temperature, moisture, pH and substrate availability will also influence the process of mineralization, immobilization, oxidation and reduction. In aerobic agricultural soils the main factor in these processes is the release of inorganic, plant-available sulfate from organic matter.

Two types of processes are involved in the mineralization of S: biological and the biochemical mineralization.[16] The biological mineralization is considered to be driven by the microbial need for organic C to provide energy, and S released as sulfate is a by-product of the oxidation of C to carbon-dioxide. This process is faster the more recently the organic matter was formed.[17] Biochemical mineralization relies on the release of sulfate from the sulfate-ester pool through enzymatic hydrolysis.

Sulfur in the microbial biomass is actively turned over in soils whilst comprising only approximately 2%–3% of the total soil S. The turnover of soil microbial bio-mass is fundamental to the incorporation of sulfate-S into soil organic matter but quantitative measures to assess this are unavailable.

The contribution of mineralization to the S supply of plants is only small,[2] because mineralization, immobilization and possible leaching of S occur concurrently.[14] The amount of S mineralized within the organic S pool in the soil ranges from 0.5% to 3.1% annually.[2] The contribution of net mineralization accounts on average for 10–30 kg ha^{-1} yr^{-1} S[18] in soils with carbon contents between 1% and 4% C. The studies of Eriksen, Murphy, and Schung[2] and Bloem[18] reveal that mineralization is an important, however not cardinal S pool for plants. High yielding crops cannot satisfy their S demand solely by mineralization and atmospheric S depositions.[7]

Crops and Crop Rotation

Dead plant material contributes to the organic matter pool but the living plant organic matter which partly remains on the field may be regarded as a storage pool for S, too. Sulfur offtake and S demand differ depending on the plant species. Sugar cane has a very low S demand with only 0.3 kg S Mg^{-1} dry matter yield,[19] cereals about 1.5–2 kg S Mg^{-1} of grain and 1–3 kg S Mg^{-1} of straw yield,[18] soybean between 4.3 and 8.8 kg S Mg^{-1},[20,21] mustard about 16 kg S Mg^{-1}[22,23] and oilseed rape up to 20 kg S Mg^{-1}.[24]

The S content of straw highly influences the mineralization of plant residues. Crops which leave residues with high S contents, may be considered as catch crops for S and are able to reduce S losses from soil by leaching. Cruciferous crops such as oilseed rape show a high S content of the residual straw,[24,25] which provide a large amount of rapidly decomposable organic material.[26] In comparison, plants without a secondary metabolism such as cereals leave smaller amounts of S. Only if the S content of for example barley straw is above 0.13% S, can a release of S from the mineralization of straw be expected.[27] If the value is below 0.13% S, a net immobilization of S takes place because the microbes need more S for the formation of their microbial biomass. This may exacerbate S deficiency of the following crop, for example oilseed rape, particularly on light soils, where macroscopic symptoms may become visible even before winter[25] because of the net immobilization of S. So far there is no evidence that winter crops benefit during their main growth from the previous crop's is plant organic matter residues with high S.

Physico-Chemical Aspects of Sulfur in Soils

Sulfate has a high mobility in soils and can be delivered from subsoil or shallow groundwater. The water soluble fraction can be leached, adsorbed, immobilized or taken up by the plant. Availability of sulfate is therefore more a question of transfer between pools in terms of space and time rather than between biological or chemical systems. The insoluble, organic fraction is only slightly leached and is not directly available for plant uptake.

Soil Water Regime

Under temperate conditions groundwater is often an important S pool and there are three ways groundwater can contribute to the S nutrition of plants. Firstly, there is a direct S input if the groundwater level is only 1–2 m below the surface. This is sufficient to cover the S requirement of most crops, as plants can utilize the sulfate in the groundwater directly by their root system. An average yielding (3 Mg ha^{-1} seeds) oilseed rape crop covers more than 50% of its S demand by shallow groundwater or soil water.[24] Secondly, groundwater contains between 5 and 100 mg L^{-1} sulfate-S which, if used for irrigation can supply up to 100 kg ha^{-1} S to the crop.[18] Thirdly, the capillary rise of groundwater under conditions of a soil moisture deficit in the upper soil layers leads to a S input. The contribution of this process to the S supply of plants depends on the soil texture within the soil profile and climatic conditions. Generally heavier soils show less frequent S deficiency than lighter ones as they retain more S-rich pore water. This is proven by a significant relationship between groundwater level and clay content, and the sulfate-S content of soils.[18]

Capillary Rise/Leaching

While rainfall water contains only about 2.5 mg S L^{-1},[28] the S content of adsorbed soil water varies between 15 and 100 mg S L^{-1}.[18] Variation of the sulfate content of water in soil pores is caused by temporal changes of the groundwater level, soil depth and differences in soil texture.[18] Sulfur-rich water in the soil pore ascends in the soil profile if the evapo-transpiration rate is higher than the amount of precipitation.

A higher groundwater level below the surface and a high soil moisture deficit at the lower periphery of the root zone will increase the extent of the capillary rise. These authors[29] calculated an average capillary rise of 0.3 mm d^{-1} on a sandy soil, 3 mmd^{-1} on a loamy sand and 5mm d^{-1} on a loamy silt with a groundwater level of 1.5m below the surface. This means that under extreme conditions on sandy soils with a groundwater level >2m below the surface virtually no sulfate will be available to the plant due to capillary rise, but on loamy and silty soils, the S supply of crops may be fully satisfied by capillary water, even if the sulfate concentration of the groundwater is low (10 mg L^{-1} S).[18] Spring crops particularly benefit from capillary ascending soil water as their main growth takes place when capillary rise dominates soil water movements.

The rate of leaching of S under temperate conditions depends on soil type, winter precipitation and the concentration of sulfate S in the soil water during the period of leaching.[18] Sulfur losses through leaching are higher over winter due to higher precipitation rates and the lack of plant uptake. As a result, S stored in soils decreases with increasing precipitation and on sandy, free-draining soils all sulfate beneath the capillary zone and rooting depth of the crop may be leached before the beginning of the vegetative period[18] because of the rapid movement of the leachate through the soil and the low adsorption capacities of such soils. Average sulfate leaching depths are 15 cm with 50mm of precipitation on a loamy soil and 25 cm on a sandy soil, respectively.[30] Additionally leaching is higher on fallow than on cropped soils[31] and is enhanced by S fertilization. Average leaching losses under temperate conditions are in the range of 30–80 kg ha^{-1} yr^{-1} S.[32,33] In addition to vertical water movements in the soil, lateral fluxes can occur particularly in the landscapes with pronounced differences in geomorphology. These fluxes are the major reason for the high spatio-temporal variability of sulfate in soils.

Soil Compaction

Soil compaction and tillage operations causing soil compaction will also decrease the amount of plant available S because of the reduced soil volume and consequent impact on soil pore space. Consequently rooting depth and density are decreased so that S rich capillary water cannot be used.[34,35] This effect of soil compaction is supported by the fact that S deficiency regularly becomes visible first in the headland and along tramlines of fields[7] and because it is sulfate-S from the sub-soil that mainly contributes to the S nutrition of crops.[18]

Sulfur Emissions

Sulfur containing atmospheric constituents are principally SO_2, SO_4^{2-}, SO_3^{2-}, H_2S, COS, DMS, CS_2 and methylmercaptan. Natural S emissions from biogenic sources and from volcanoes add around 60 Tg S yr^{-1} to the Earth's atmosphere with an upward tendency due to the increasing consumption of fossil fuels in South America, Africa and Asia.[9]

Before the industrial revolution, atmospheric S depositions were, on an average, below 10 kg S ha^{-1} yr^{-1}. As a consequence of increased burning of fossil fuels with industrial development from 1890 to 1980 atmospheric S depositions, mainly as SO_2, increased steadily by 0.47 kg S ha^{-1} yr^{-1}.[1] In some rural areas of northern Scotland, however, SO_2 depositions were zero or below 2 kg S ha^{-1} in the 1970s, while the S input in the midlands of England was about 160 kg S ha^{-1} at the same time.[36] At its peak the negative impacts of SO_2 emissions to humans, plants, soils and buildings were so serious that S became an unpopular nutrient and was called as the "yellow poison".[37] The political consequences resulted in clean air acts in European countries and North America and the desulfurization of emissions led to a drastic decrease of atmospheric S depositions.

Agriculture adapted and co-evolved to increasing S loads. Soil acidification caused by high S deposition required higher amounts of liming materials. Increasing S demands of agricultural crops due to higher yield potential as a result of plant breeding progress and production technology coincided with increased atmospheric S supply. Additionally the use of S-containing fertilizers went down nearly to zero.[25] Consequently the reduction of atmospheric S deposition had a major effect on the productivity of crops. During the 1990s, macroscopic S deficiency became the most widespread nutrient disorder in northern Europe.[7]

Sulfur Fertilizers

The worldwide demand of S fertilizers in 2010 is estimated to be about 11.3 million tons of S.[38] Numerous S fertilizers and secondary raw material fertilizer products are available for either soil or foliar application in sulfate or elemental form.[39] Arguments for and against the use of individual products depend on local farming conditions. Organic fertilizers such as manure and slurry contain about 1 kg S Mg^{-1} and 0.5 kg S Mg^{-1}, respectively.[40] This equals to 0.07 kg S per kg nitrogen.

Agro-Ecological Aspects of Sulfur

Interaction of Agro-Ecosystems with Other Ecosystems

The S input and off-take in agro-ecosystems may have a strong influence on neighboring eco-systems. Sulfur is often in excess in natural systems despite the low atmospheric S input because the turnover of S is much lower in these than in agricultural systems. Acid sub-soils, e.g., under forest vegetation, and peat soils show pH values <5, which are far below the acidity levels of fertile agricultural soils. Large storage capacities for adsorbed sulfate exist on such sites because the pH value of the soil is lower than the zero point of charge (ZPC) resulting in the hydration of metal oxides and thus a positive surface.[41] Although adsorbed sulfate plays only a minor role in the direct S nutrition of agricultural crops it may contribute positively to the S balance of the whole surrounding landscape. Natural vegetation, land without plant production and forests show a positive S balance even when atmospheric inputs are low.[2] Agricultural crops may benefit from this S pool, if groundwater reservoirs of both eco-systems are connected. This also means, however, that with an increasing share of agricultural farmland in landscapes, the risk of S deficiency overall also increases.

Sustainability

Sustainable agriculture should use soils in such a way that the present and future human needs for food or other agricultural goods are realized and the quality of the environment and the natural resources remain preserved.[42] The contamination of groundwater with nitrates is a most serious problem. Nitrogen (N) and S are both involved in protein biosynthesis and a shortage in the S supply of crops also lowers the utilization of applied fertilizer N and thus deteriorates the crop quality. Non-protein N is accumulated in plant parts and besides poor efficiency for N fertilization, S deficiencies may increase the loss of N from agricultural soils through volatilization and leaching.[43] On average each kg of S shortage to satisfy the S demand of the crop causes 15 kg of N to be lost to the environment. Such N inputs endanger strongly the stability of natural communities as for example the growth of algae in water bodies.[44] Correcting S deficiency by fertilization is environmentally safe as sulfate is, in comparison with N, geogenously abundant.

Global Change

Climate is one of the major factors involved in pedogenesis. Soil formation is directly influenced mainly by temperature and water and indirectly via the climate-depending vegetation.[45] Changes in climate may change soil types, increase erosion, affect element cycles and increase the release of greenhouse gases. Increased temperature and humidity accelerate pedogenesis. Global change thus would allow soils in the northern hemisphere to proceed faster through the individual stages of development and/or degradation associated with their individual soil series.[46] Higher temperatures accelerate the decomposition of organic materials and thus decrease the organic matter content of soils. At the same time the net mobilization of S might increase while the organic S pool decreases. This effect is expected to be more pronounced in cultivated than in range soils.[47] But at the same time under more humid climate conditions organic matter tends to accumulate in soils.[47] Therefore an expected adverse effect of global warming on soil organic matter might be at least partially compensated in those areas where this coincides with increasing humidity.[48] With increasing carbon accumulation, the other growth limiting elements, such as N, S, and P, may dilute relatively,[49] which might result in decreased soil fertility.

Plant Health

Owing to the fact that higher anthropogenic S inputs in the past decades enabled plants to adapt to increasing environmental stress, the decline in the S supply within only one decade might have serious consequences for the stability of recent ecosystems.[49] Sulfur metabolism provides several efficient mechanisms by which plants are able to tackle abiotic (e.g., xenobiotics and increasing surface ozone levels) and biotic (e.g., pests and diseases) stress, particularly via the glutathione metabolism which again is closely related to the S supply of the plants.[3] Other mechanisms involved in response to plant pathogens include the production of S containing compounds in the secondary metabolism of the agriculturally important *Brassica* species, the release of volatile S compounds, the production of S rich proteins, localized deposition of elemental S and the production of phytochelatines, which detoxify heavy metals by forming complexes.[50,51]

Certain diseases (e.g., light leaf spot in oilseed rape) occur more frequently particularly in areas with low S input in Europe[52] and an improved knowledge of the significance of S metabolites in crop resistance to diseases will be beneficial for the improvement of S fertilizer strategies and could therefore minimize the input of pesticides.

Acknowledgments

The authors cordially thank Dr. Kerr C. Walker (SAC, Aberdeen) for his linguistic efforts on our entry.

References

1. Dämmgen, U.; Walker, K.C.; Grünhage, L.; Jäger, H.-J. The atmospheric sulphur cycle. In *Sulphur in Agroecosystems*; Schnug, E., Ed.; Kluwer Academic Publ.: Dordrecht, 1998; 75–114.
2. Eriksen, J.; Murphy, M.D.; Schnug, E. The soil sulphur cycle. In *Sulphur in Agroecosystems*; Schnug, E., Ed.; Kluwer Academic Publ.: Dordrecht, 1998; 39–74.
3. Hell, R.; Rennenberg, H. The plant sulphur cycling. In *Sulphur in Agroecosystems*; Schnug, E., Ed.; Kluwer Academic Publ.: Dordrecht, 1998; 135–174.
4. Howarth, R.W.; Stewart, J.W.B. The interactions of sulphur with other element cycles in ecosystems. In *Sulphur Cycling on the Continents: Wetlands, Terrestrial Ecosystems and Associated Water Bodies, SCOPE 48;* Howarth, R.W., Stewart, J.W.B., Ivanov, M.V., Eds.; John Wiley and Sons: Chichester, 1992; 67–84.
5. Howarth, R.W.; Stewart, J.W.B.; Ivanov, M.V.; Eds. *Sulphur Cycling on the Continents: Wetlands, Terrestrial Ecosystems and Associated Water Bodies, SCOPE 48*; John Wiley and Sons: Chichester, 1992; 345 pp.
6. Janzen, H.H.; Ellert, B.H. Sulfur dynamics in cultivated, temperate agroecosystems. In *Sulfur in the Environment*; Maynard, D.G., Ed.; Marcel Dekker Inc.: New York, 1998; 11–44.
7. Schnug, E.; Haneklaus, S. Diagnosis of sulphur nutrition. In *Sulphur in Agroecosystems*; Schnug, E., Ed.; Kluwer Academic Publ.: Dordrecht, 1998; 1–38.
8. Stevenson, F.J. *Cycle of Soil. Carbon, Nitrogen, Phosphorus, Sulfur, Micronutrients*; John Wiley and Sons: New York, 1986; 285–320.
9. Whelpdale, D.M. An Overview of the atmospheric sulphur cycle. In *Sulphur Cycling on the Continents: Wetlands, Terrestrial Ecosystems and Associated Water Bodies, SCOPE 48;* Howarth, R.W., Stewart, J.W.B., Ivanov, M.V., Eds.; John Wiley and Sons: Chichester, 1992; 5–26.
10. Bowen, H.J.M. *Trace Elements in Biochemistry*; Academic Press: London, 1966.
11. Friend, J.P. The global sulfur cycle. In *Chemistry of the Lower Atmosphere*; Rasool, S.I., Ed.; Plenum Press: New York, 1973; 177–201.
12. Wedepohl, K.H. Chemical fractionation in the sedimentary environment. In *Origin and Distribution of the Elements*; Ahrens, L.H., Ed.; Pergamon Press: London, 1968; Vol. 30 of Earth Sciences.
13. Anisimova, M.; Haneklaus, S.; Schnug, E. Significance of sulfur for soil organic matter. In *Sulfur Nutrition and Sulfur Assimilation in Higher Plants;* Brunold, C., Rennenberg, H., De Kok, L.J., Stulen, I., Davidian, J.-C., Eds.; Paul Haupt Publishers: Berne, 2000; 239–244.
14. Ghani, A.; McLaren, R.G.; Swift, R.S. The incorporation and transformations of 35S in soil: effect of soil conditioning and glucose or sulfate addition. Soil Biol. Biochem. **1993**, *25*, 327–335.
15. Eriksen, J.; Lefroy, R.D.; Blair, G.J. Physical protection of soil organic S studied by extraction and fractionation of soil organic matter. Soil Biol. Biochem. **1995**, *27*, 1011–1016.
16. McGill, W.B.; Cole, C.V. Comparative aspects of cycling of organic C, N, S and P through soil organic matter. Geoderma **1981**, *26*, 267–286.
17. Ghani, A.; McLaren, R.G.; Swift, R.S. Mobilization of recently-formed soil organic sulphur. Soil Biol. Biochem **1993**, *25*, 1739–1744.
18. Bloem, E. Schwefel-bilanz von agrarökosystemen unter besonderer berücksichtigung hydrologischer und bodenphysi-kalischer Standorteigenschaften. Landbauforsch. Voelk. **1998**, *192*, 1–156.
19. Katyal, J.C.; Sharma, K.L.; Srinivas, K. *Sulphur in Indian Agriculture*; Proc. TSI/FAI/IFA Symposium on sulphur in balanced fertilisation, KS-2/1-KS-2/12; 1997.

20. Aulakh, M.S.; Pasricha, N.S.; Azad, A.S. Phosphorussulphur interrelationships for soybeans on phosphorus and sulphur deficient soils. Soil Sci. **1990**, *150*, 705–709.
21. Nambiar, K.K.M.; Gosh, A.B. Highlights of Research of a Long-Term Fertilizer Experiment in India (1971–1982); Techn. Bull. No.1 Longterm Fertilizer Experiment Project, IARI, 100, 1984.
22. Jain, G.L.; Sahu, M.P.; Somani, L.L. Balanced fertilization programme with special reference to secondary and micronutrients nutrition of crops under intensive cropping, Proc. FAI/NR Seminar. Jaipur, 1984; 147–174.
23. Aulakh, M.S.; Pastricha, N.S.; Sahota, N.S. Yield, nutrient concentration and quality of mustard crops as influenced by nitrogen and sulphur fertilizers. J. Agric. Sci. **1980**, *94*, 545–549.
24. *Schnug, E. Quantitative und Qualitative Aspekte der Diagnose und Therapie der Schwefelversorgung von Raps (Brassica napus L.) Unter Besonderer Berücksichtigung Glucosinolatarmer Sorten; DSC, Christian-Albrechts-University: Kiel, Germany, 1988.*
25. Schnug, E.; Haneklaus, S. Sulphur deficiency in brassica napus- biochemistry- symptomatology-morphogenesis. Landbauforsch. Voelk. **1994**, *144*, 1–31.
26. Wu, J.A.G.; O'Donnell, Z.L.; Syers, J.K. Microbial growth and sulphur immobilization following the incorporation of plant residues into soil. Soil Biol. Biochem. **1993**, *25*, 1567–1573.
27. Chapman, S.J. Barley straw decomposition and S immobilisation. Soil Biol. Biochem. **1997**, *29*, 109–114.
28. Dämmgen, U.; Grünhage, L.; Kü sters, A.; Jäger, H.J. Konzentrationen von luftinhaltsstoffen. I. criteria pollutants. Landbauforsch. Voelk. **1996**, *170*, 196–221.
29. Giesel, W.; Renger, M.; Strebel, O. Berechnung des kapillaren aufstiegs aus dem grundwasser in den wurzelraum unter stationären bedingungen. Z. Pflanzenern. Bodenkde **1972**, *132*, 17–30.
30. Kumar, V.; Karwasra, S.P.S.; Singh, M.; Dhankar, J.S. An evaluation of the sulphur status and crop responses in the major soils of Haryana, India. Sulphur Agric. **1994**, *18*, 23–26.
31. Kirchmann, H.; Pichlmayer, F.; Gerzabek, M.H. Sulfur balance and sulfur-34 abundance in a long-term fertilizer experiment. Soil Sci. Amer. J. **1996**, *59*, 174–178.
32. Mansfeld, T. Schwefeldynamik von böden des dithmarscher speicherkoogs und der bornhoeveder Seenkette in schleswig-holstein. PhD, Kiel, Germany, 1994.
33. Preuschoff, M. Untersuchungen zur schwefelversorgung von weiβohl an zwei löβstandorten. PhD, University Hanover, Verlag Ulrich E. GrauerStuttgart, 1995.
34. Singh, B.R. Effect of soil compaction on S availability to crop plants. Abstracts of the COST Action 829: Fundamental, Agronomical and Environmental Aspects of Sulfur Nutrition and Assimilation in Plants, Goslar, Germany, 1998.
35. Unger, P.W.; Kaspar, T.C. Soil compaction and root growth: a review. Agron. J. **1994**, *86*, 759–766.
36. Semb, A. Sulfur emissions in Europe. Atmos. Environm. **1978**, *12*, 455–460.
37. Boelsche, J. *Das gelbe Gift*; Rowohlt Verlag: Reinbek, 1994.
38. Cecotti, S.; Morris, R.J.; Messick, D. A global overview of the sulphr situation: industry's background, Market trends, and commercial aspects of sulphur fertilisers. In *Sulphur in Agroecosystems;* Schnug, E., Ed.; Kluwer Academic Publ.: Dordrecht, 1998; 175–202.
39. Paulsen, H.M. Produktionstechnische und Ökologische bewertung der landwirtschaftlichen verwertung von schwefel aus industriellen prozessen. Landbauforsch. Voelk. **1999**, *197*, 1–143.
40. Pedersen, C.A.; Knudsen, L.; Schnug, E. Sulphur Fertilisation. In *Sulphur in Agroecosystems;* Schnug, E., Ed.; Kluwer Academic Publ.: Dordrecht, 1998; 115–134.
41. Curtin, D.; Syers, J.K. Extractability and adsorption of sulfate in soils. J. Soil Sci. **1990**, *41*, 305–312.
42. Anon. *Nachhaltiges Deutschland: Wege zu Einer Dauerhaft Umweltgerechten Entwicklung*; Umweltbundesamt Berlin, Erich Schmidt Verlag: Berlin, 1997.
43. Schnug, E. Sulphur nutritional status of European crops and consequences for agriculture. Sulphur in Agriculture **1991**, *15*, 7–12.
44. Wild, A. *Umweltorientierte Bodenkunde*; Spektrum Akademischer Verlag: Heidelberg, 1993.
45. Hugget, R.J. *Geoecology;* Routledge: London, 1995.

46. Rogasik, J.; Daemmgen, U.; Luettich, M.; Obenauf, S. Wirkungen physikalischer und chemischer klimaparameter auf bodeneigenschaften und bodenprozesse. Landbauforsch. Voelk. **1994**, *148*, 107–139.
47. Burke, I.C.; Yonker, C.M.; Parton, W.J.; Cole, C.V.; Flach, K.; Schimel, D.S. Texture, climate and cultivation effects on soil organic matter content in U.S. Grassland Soils. Soil Sci. Soc. Am. J. **1989**, *53*, 800–805.
48. Schnug, E. Response of plant metabolism to air pollution and global change impact on agriculture. In *Responses of Plant Metabolism to Air Pollution and Global Change*; DeKok, L.J., Stulen, I., Eds.; Backhys Publ.: Leiden, 1998; 15–22.
49. Schnug, E.; Haneklaus, S. Ecological Aspects of Plant Sulphur Supply. Proc. 15th Int. Congr. Soil Sci. Acapulco/ Mexico, 1994; 5a: Comm. IV: Symposia, 364–371, 1994.
50. Resende, M.L.V.; Flood, J.; Ramsden, J.D.; Rowan, M.G.; Beale, M.H.; Cooper, R.M. Novel phytoalexins including elemental sulphur in the resistance of cocoa (*Theobroma cacao* L.) to verticillium wilt (*Verticillium dahliae Kleb*.). Physiol. Molec. Plant Pathol. **1996**, *48*, 347–359.
51. Schnug, E. Significance of Sulphur for the nutritional and technological quality of domesticated plants. In *Sulfur Nutrition and Sulfur Assimilation in Higher Plants*; Cram, W.J., DeKok, L.J., Stulen, I., Brunold, C., Rennberg,H., Eds.; Backhys Publ.: Leiden, 1997; 109–130.
52. Thomas, J. Watch out for light leaf spot. NIAB Fellows Newsletter 1994.

37

Sulfur Dioxide

Marianna Czaplicka
and Witold Kurylak

Introduction

Sulfur dioxide (SO_2) in large amounts is a toxic gas. The World Health Organization (WHO) recommends a concentration of no greater than 0.5 parts per million (ppm) over 24 hours for maximum exposure. A concentration of 6–12 ppm can cause immediate irritation of the nose and throat; 20 ppm can cause eye irritation; and 10,000 ppm will irritate moist skin within minutes. High concentrations of SO_2 can result in breathing problems in asthmatic children and adults who are active outdoors.[1,2] Short-term exposure has been linked to wheezing, chest tightness, and shortness of breath.

Sulfur dioxide emissions are partly responsible for acid depositions on the surface and the occurrence of winter smog episodes. It is one of the major atmospheric pollutants. In the air, SO_2 oxidizes to SO_3, which easily reacts with steam to create sulfuric acid (H_2SO_4), which then condenses rapidly in the stratosphere to form fine sulfate aerosols. These sulfate aerosols also promote complex chemical reactions on their surfaces that alter chlorine and nitrogen chemical species in the stratosphere. This effect, together with increased stratospheric chlorine levels from chlorofluorocarbon pollution, generates chlorine monoxide, which destroys ozone. Sulfate aerosols generated from volcanic emissions of sulfur-rich gas temporarily counteract some of the heating caused by elevated carbon dioxide concentrations in the atmosphere.[3] Aerosols in the atmosphere cool the climate both directly due to backscattering of sunlight and indirectly through an increase in cloud receptivity and residence time. Pollutants can be transported thousands of kilometers owing to the introduction of tall chimneys that disperse pollutants high in the atmosphere.

Acid rain causes acidification of lakes and streams, and contributes to the damage of trees at high elevations and of many sensitive forest soils. Soil acidification due to acidic deposition may contribute to forest decline[4] and can lead to the loss of plant nutrients and the accumulation of toxic aluminum, both of which may affect vegetation. In addition, acid rain accelerates the decay of building materials and paints, including irreplaceable buildings, statues, and sculptures that are part of the cultural heritage.

The presence of SO_2 was observed both in the stratosphere and in the troposphere. The lifetime of SO_2 molecules in the troposphere is a few days. In the troposphere, mainly volcanic SO_2 becomes rapidly oxidized to sulfate and washed out of the atmosphere in the presence of hydrometeors. These washout processes are not effective in the stratosphere. The lifetime of SO_2 molecules in the stratosphere is several weeks, during which time sulfate aerosols are produced. Since gaseous H_2SO_4 has a very low saturation pressure, it can be easily condensed in the stratosphere and forms aerosols of liquid hydrated H_2SO_4. Once formed, sulfate aerosols have a residence time in the stratosphere of about 3 years. Numerous measurements clearly show the importance of sulfur aerosols in modifying climate, warming the stratosphere, and cooling the troposphere. Research has also shown that the liquid drops of H_2SO_4 promote the destruction of the earth's ozone layer. Ways for SO_2 removal from the troposphere include the following:

- In gas phase, by formation of H_2SO_4, which forms condensation nuclei for aerosols and clouds and acidifies rain
- Directly, by way of uptake on aerosols and clouds, which leads to dry and wet acid depositions

SO_2 Standards in the World—Ambient Air Quality

The concentration of SO_2 in the air depends mostly on meteorological factors such as

- Transport processes by prevailing winds
- Depth of the mixing layer
- Deposition process

Levels of SO_2 in the air are higher than normal near facilities that release SO_2 through heavy industrial activities such as copper smelting or coal and oil burning or processing.

The negative influence of SO_2 on human health became a basis for the determination of air quality standards covering, among other things, maximum permissible concentration of that compound in the air. In 2005, WHO updated its air quality guidelines (AQG), an international reference on the health consequences of exposure to air pollution and a policy tool for reducing these consequences worldwide. For SO_2, AQG achieved a level 20 $\mu g/m^3$ for 24 hr and 500 $\mu g/m^3$ in the case of 10-min measurement.[5]

Air quality standards differ worldwide. Vahlsing and Smith[6] showed that average 24-hr ambient air quality standards (AAQS) in 76 countries for SO_2 is 95 $\mu g/m^3$ and the population weighted average of AAQS for SO_2 is 155 $\mu g/m^3$. In the United States, there are two existing primary national AAQS (NAAQS) for SO_2. The first is a long-term 1 year arithmetic average of 0.03 ppm. The second is a short-term 24-hr average standard where concentrations should not exceed 0.14 ppm more than once per year. The current secondary NAAQS in the United States for SO_2 is a 3-hr average concentration of 0.5 ppm not to be exceeded more than once per year. In addition, there are three standards for SO_2 defined by state law. The first is a long-term 1-year arithmetic average of 0.022 ppm. The second is a short-term 24-hr average standard of 0.088 ppm not to be exceeded, and the third is a short-term 3-hr average concentration of 0.439 ppm not to be exceeded.

In June 2010, U.S. Environmental Protection Agency (EPA) strengthened the primary NAAQS for SO_2 by establishing a new 1-hr standard designed to protect sensitive individuals from high, short-term (5-min to 24-hr) exposure.[7] The level is set at 75 parts per billion (ppb). This new 1-hr standard will replace the two existing primary standards.

The EPA's current action focuses on the primary SO_2 standard only. It will address the secondary standard for SO_2 as part of a separate review. The current secondary standard to protect public welfare is a 3-hr average set at 500 ppb, as mentioned before.

Unlike WHO, the European Union air quality standards (EU AQS)[8] and the U.S. EPA[9] continue to set a 1-hr standard for SO_2 at 350 and 196 μg/m^3 (70 ppb), respectively. According to the European standard, the permissible hourly concentration of SO_2 in the air is 350 μg/m^3 while in 24 hr it is 125 μg/m^3.[8]

Most developing countries in Asia either use the 24-hr standard and/or the annual standard. Although some countries have 1-hr guidelines, they are not comparable to the EPA or even the EU AQS except for some countries. In China, for example, there are five different standards: Grade I—applies to specially protected areas, such as natural conservation areas, scenic spots, and historical sites; Grade II—applies to residential, mixed commercial/residential, cultural, industrial, and rural areas; Grade III—special industrial areas; SAR and SAR 2—Special Administrative Region. The regulations are presented in Table 1.

Methods for Determination of SO_2 in Air

There are many techniques available for analysis of SO_2 in ambient air, including detection by spectrophotometry,[11–13] chemiluminescence,[14–16] ion chromatography,[17,18] spectrofluorometry,[19] and potentiometric[20] methods. The measurement of SO_2 in air is usually based on classic fluorescence spectroscopy principles. Sulfur dioxide exhibits a strong ultraviolet (UV) absorption spectrum between 200 and 240 nm, and when SO_2 absorbs UV, emission of photons occurs (300–400 nm). The amount of fluorescence emitted is directly proportional to the SO_2 concentration. Ultraviolet fluorescent radiation technology to detect SO_2 is characterized by a sensitivity of 0.4 ppb in the range of 0–20 ppm.[11] Chang et al.[12] described a simple fluorescence detection method for determination of SO_2 in ambient air collected by passive sampler. Ultraviolet fluorescence and pararosaniline methods are recommended by the EPA as reference methods for the determination of SO_2 in the atmosphere.[21] In the European Union, Directive 2008/50/EC of the European Parliament and of the Council recommends as reference method the procedure described in EN 14212:2005, "Ambient air quality—Standard method for the measurement of the concentration of sulfur dioxide by ultraviolet fluorescence."[8]

Sources of SO_2

Emission of SO_2 to the air results from natural processes and from anthropogenic sources. Globally, emissions due to anthropogenic activities account for approximately 70 Tg (S) of the total (natural + anthropogenic) 107 Tg (S) emissions of sulfur each year, although a greater uncertainty exists for natural sources.[22,23]

Biogenic Sources

Is well known that SO_2 from volcanoes is one of the two most important sources of stratospheric aerosols. According to Symonds et al.[24] SO_2 concentration in volcanic gases may vary from 11.8% (Kilauea Volcano) to 0.50% (Momotombo Volcano). The data on Popocatèpetl Volcano (Mexico) eruption show

TABLE 1 Standards of SO_2 in China, μg/m^3

	SO_2 1 hr	SO_2 24 hr	SO_2 Annual
China: Grade I	150	50	20
China: Grade II	500	150	60
China: Grade III	700	250	100
Hong Kong SAR	800	350	80
Hong Kong SAR 2	—	125	—

Source: Data from Air Quality in Asia: Status and Trends.[10]

that the total sulfur emitted in the years 1994–1998 achieved a level of 8,500,000 tons,[25] while the total SO_2 outgassed from Galunggung Volcano (Java) from April 1982 to January 1983 was calculated on level 2500 kt.[26] Another eruption, that of El Chichon in 1982, emitted $7 \leftrightarrow 10^{12}$ g SO_2. Halmer and coworkers[27] calculated $15 \leftrightarrow 10^{12}$ to $21 \leftrightarrow 10^{12}$ g of SO_2 as the total annual global volcanic SO_2 input into the atmosphere from both silent and explosive degassing volcanoes, corresponding to $7.5 \leftrightarrow 10^{12}$ to $10.5 \leftrightarrow 10^{12}$ g/yr S (S here as SO_2) for the period 1972 to 2000, the numbers representing an average value for volcanic gas emission over almost 30 years. The amount of volcanic gases and particles arriving in the troposphere and/or stratosphere generally depends on both volcanological and meteorological boundary conditions:

- The chemical composition of the magma depending on the tectonic environment (e.g., subduction zone, continental rift zone)
- The strength and duration of an eruption
- The location of the emitting volcano (altitude and latitude)
- Atmospheric conditions such as temperature, humidity, and wind prowls

Sulfur dioxide and other volcanic sulfur-bearing gases transform to sulfate aerosols, which are able to cool the climate directly by backscattering incoming solar radiation and indirectly through increase in cloud receptivity. The residence time of water-soluble sulfuric gases (in particular SO_2) emitted into the atmosphere depends on the height of an eruption column. The effects of SO_2 (from natural sources) on people and the environment vary widely depending on the following: 1) the amount of the gas a volcano emits into the atmosphere; 2) whether the gas is injected into the troposphere or the stratosphere; and 3) the regional or global wind and weather pattern that disperses the gas.

Recent studies have suggested oxidation of dimethyl sulfide (DMS) in the marine atmosphere as the elevated biogenic source of SO_2. Oceanic emissions of DMS are a major component of the natural sulfur budget (about 60% of total natural global sulfur emissions).[28,29] Dimethylsulfoniopropionate produced by phytoplankton in the ocean is the precursor to the DMS gas. Once in the atmosphere, DMS is oxidized by radicals to form methylsulfonic acid, dimethyl sulfoxide, dimethyl sulfone, sulfate, and SO_2. These products take part in the growth of aerosols. Gaseous H_2SO_4 is the prevailing product of DMS oxidation that can generate new cloud condensation nuclei under atmospheric conditions. One of the precursors in this formation pathway is SO_2. Seguin et al.'s[30] measurements of SO_2 concentration in the air over the Northwest Atlantic Ocean in spring and summer 2003 indicted that SO_2 concentrations were as high as 156 nmol/m^3 in the spring and from 0.55 nmol/m^3 to 120 nmol/m^3 in the summer. During the Mediterranean Intensive Oxidant Study (MINOS) experiment (July–August 2001), Mihalopoulos et al. [31] found that marine biogenic sulfur emissions (predominantly DMS) constituted 20% of total oxidized sulfur production (SO_2 and H_2SO_4). Among the natural sources of SO_2 are also biological decay and forest fires.

Measurements of Concentration of SO_2 from Natural Sources

Measurement of SO_2 emission from natural sources is a very difficult task. The emission rates of SO_2 are measured using airborne or ground-based techniques. The most common method for sampling volcanic gases is to collect them directly from fumaroles in solution-filled bottles, and then to analyze the mixtures in the laboratory.[24] Observations of SO_2 flux have been made for over 30 years by correlation spectrometers, which have been recently upgraded with compact and inexpensive powered CCD detector array–based dispersive spectrometers.[32,33] Ultraviolet spectroscopy has been widely applied to volcanoes worldwide over the last decades to investigate how SO_2 fluxes vary with activity, with heightened emissions observed during eruptions as degassing magma batches reach the surface.[34,35] Salerno et al.[36] describe the method based on differential optical absorption for automatic determination of volcanic SO_2. The other sensing techniques applied to determination of SO_2 concentration include multispectral infrared[37,38] and ship-borne differential absorption lidar (DIAL) methods,[39–41] and Correlation Spectroscopy (COSPEC).[42,43]

TABLE 2 Emission Factors of SO_2 for Different Types of Fuel (Emission Factors Are Taken from National Atmospheric Emission Inventory)

Emission Source	Emission Factor
Generators (gas oil)	$3.25 \leftrightarrow 10^{-3}$
Generators (diesel fuel)	$33 \leftrightarrow 10^{-3}$
Vehicles (principally ATF)	$1.0 \leftrightarrow 10^{-7}$
Marine gas oil (for ships)	$1.9 \leftrightarrow 10^{-2}$
Fuel oil (for ships)	$5.2 \leftrightarrow 10^{-2}$
Aviation turbine fuel (for international flights)	$7.2 \leftrightarrow 10^{-4}$
Aviation turbine fuel (for domestic flights)	$7.2 \leftrightarrow 10^{-4}$

Source: Data from Shirsat and Graf.[23]

Anthropogenic Sources

The largest anthropogenic SO_2 emissions come from fossil fuel combustion at power plants (73%) and other industrial facilities (20%).[44] Smaller sources of SO_2 emissions include industrial processes such as biomass burning, extracting metal from ore, smelting of metal sulfide ores, and burning of high-sulfur-containing fuels by locomotives, large ships, and non-road equipment. Sulfur dioxide is linked with a number of adverse effects on the respiratory system. Table 2 shows emission factors of SO_2 for various types of fuel.

Sulfur dioxide is used as fumigant, preservative, bleach, and steeping agent for grain. A small amount of SO_2 may be used, as preservative, both in the vineyard and during winemaking to protect grapes and wine from microbial attack or further oxidation. Sulfur dioxide is also a bleaching agent in the pulp and paper industry.

Corbett et al.[45] showed that SO_2 emissions from ocean-going ships significantly contribute to the global pool of sulfur and make up about 20% of biogenic DMS emissions. It has been estimated that the 1993 global SO_2 emissions from ships released 4.24 Tg of sulfur. Emissions of NO_x and SO_2 from ocean-going ships are 10%–14% and 3%–4%, respectively, of the total emissions of these species from the burning of fossil fuels, and B40% and B70%, respectively, of the total emissions of these species from the burning of biomass.

Methods of Measurements of SO_2 Concentration in Flue Gases

Recommended methodologies for emission inventory estimation are compiled in the European Monitoring and Evaluation Programme (EMEP) prepared by European Environment Agency (EEA) *Air Pollutant Emission Inventory Guidebook* (2009). Base data are available from the EEA Data Service (http://dataservice.eea.europa.eu/dataservice/metadetails.asp?id=1096) and the EMEP web site (http://www.ceip.at/).

According to the best available techniques (BAT) recommendations, measurements of SO_2 concentration in the gases for large sources are conducted by two methods[46]:

- Inline photometer (IR only), which can measure gas concentrations in gas pipes if the matrix is optically transparent (no mist).
- Online photometer with sample preparation. This is the usual method. A suitable material has to be selected for the measuring cell and the sample preparation because of the risk of corrosion.

The gas being analyzed at the outlet of a converter system has the following approximate composition: SO_2 in the range 0–1000 ppm; matrix: air, water, H_2SO_4 (30 ppm); and NO_x (50 ppm). The measurement is performed using a commercial photometer (IR units require water compensation).

In the assessment of the release to the air according to the European Pollutant Release and Transfer Register,[47] the following methods of SO_2 concentration determination in flue gases are proposed:

- EN 14791:2006. Stationary source emissions—determination of mass concentration of SO_2—reference method.
- ISO 7934:1989. Stationary source emissions—determination of the mass concentration of SO_2—hydrogen peroxide/barium perchlorate/thorin method.
- ISO 7935:1992. Stationary source emissions—determination of the mass concentration of SO_2—performance characteristics of automated measuring methods.
- ISO 11632:1998. Stationary source emissions—determination of mass concentration of SO_2—ion chromatography method.
- ISO 10396:2007. Stationary source emissions—sampling for the automated determination of gas emission concentrations for permanently installed monitoring systems.

Overview of Technologies Used to Reduce Sulfur Oxide Emissions (Removal from Anthropogenic Sources)

As mentioned before, the main anthropogenic sources of SO_2 emission are commercial power plants. Sulfur oxides are emitted from the combustion of most fossil fuels through oxidation of the sulfur contained in the fuel. According to the European Union directive Integrated Pollution Prevention and Control (IPPC) and Reference Document on Best Available Techniques,[46,48–51] application of a method for desulfurization of gases emitted to atmosphere is obligatory in all such installations.

Below, the main methods used in commercial power plants for desulfurization of flue gases are characterized.[46]

Wet Techniques

Wet scrubbers, especially the limestone–gypsum processes, are the leading flue gas desulfurization (FGD) technologies. They have about 80% of the market share and are used in large utility boilers. This is due to their high SO_2 removal efficiency and their high reliability. Limestone is used in most cases as the sorbent, as it is available in large amounts in many countries. By-products are either gypsum or a mixture of calcium sulfate/sulfite, depending on the oxidation mode. Scrubbing process is another technology used for that purpose. The sodium scrubbing process is a simple method and has been applied to a large number of small oil-fired boilers. Several magnesium-scrubbing systems are also used in relatively small industrial boilers because of the low capital costs involved. In magnesium scrubbing systems, wastewater containing magnesium sulfate can be discharged into the sea, after the removal of dust and dust-absorbed heavy metals, as magnesium sulfate is already a constituent of seawater. The process, therefore, has an advantage over other systems if a plant is located near the coast. The by-product of wet ammonia scrubbers can be used as an agricultural fertilizer.

Wet limestone scrubbers are generally divided into two categories according to the type of oxidation: forced oxidation and natural oxidation. The mode of oxidation is determined by the chemical reactions, the pH of the reagent slurry, and the resulting by-product. In the natural oxidation mode, calcium sulfite is partly oxidized by the oxygen contained in the flue gas. The main product is calcium sulfite hemihydrate. The produced mixture of calcium sulfite hemihydrate and gypsum is in a form of sludge.

At the lower pH range of 4.5 to 5.5, the chemical reaction is different. After SO_2 absorption, the primary product of the neutralization by limestone is not calcium sulfite but calcium bisulfite, $Ca(HSO_3)_2$. Calcium bisulfite is much more soluble than calcium sulfite. Hence, operation in the lower pH range involves smaller risks of scaling and plugging. Calcium bisulfite is oxidized and crystallized to form gypsum or calcium sulfate dehydrate. The by-product from the natural oxidation mode is a mixture that is difficult to dewater. This mixture is composed of calcium sulfite hemihydrate and calcium sulfate dehydrate. The configuration of wet limestone scrubbers can be generally classified into four types, which depend on the applied technical solutions.

Wet limestone FGD systems naturally suffer from an aggressive operating environment, leading to corrosion, erosion, and abrasion. The flue gas path from the inlet of the absorber to the stack discharge must be protected, for instance by using rubber or flake linings, against acid attack caused by adiabatic cooling and saturation of the gas. Wet scrubber processes have undergone considerable development in the last few decades, leading to improved reliability and removal efficiencies, as well as reduced costs. Reliability is normally more than 99% with forced oxidation, and 95%–99% with natural oxidation. The applicability may be influenced by both scrubber components and auxiliary processes connected to the absorber process. The indicative range of wet limestone scrubbing application is for SO_2 content <1000 mg/m^3 in the raw gas.

Seawater scrubbing utilizes seawater's inherent properties to absorb and neutralize SO_2 in flue gases. The flue gas is fed to the SO_2 absorber, where it comes into contact with a controlled proportion of the seawater, taken from the cooling water outflow of the steam turbine condenser. Owing to the presence of bicarbonate and carbonates in the seawater, the SO_2 of the flue gas is absorbed. The acidified absorber effluent is mixed with additional seawater to ensure that the pH is at optimal level for the oxidation process. The introduced air forces the oxidation of the absorbed SO_2 from bisulfite to bisulfate and removes dissolved CO_2.

Magnesium Wet Scrubber

In magnesium scrubbing, magnesium hydroxide is used as the reagent, produced by adding slaked lime to seawater in order to enhance alkalinity. Waste sulfate liquor is produced as a result. The magnesium wet scrubber has been mainly applied in smaller plants, i.e., less than 50 MW.

Ammonia Wet Scrubber

In the ammonia wet scrubber system, SO_2 is absorbed by aqueous ammonia, resulting in ammonium sulfate as the fertilizer by-product. The process has been chosen to satisfy the criteria for emission limits of less than 200 mg/N m^3. Besides this, the driving force to apply the wet ammonia process has been the requirement for a saleable by product.

Semidry Techniques

Spray Dry Scrubbers

Spray dry scrubbers are mostly applied for relatively small- to medium-capacity boilers using low- to mediumsulfur (1.5%) coal. The residue is normally composed of a mixture of calcium sulfite, calcium sulfate, and fly ash, which is less attractive commercially. The process consists mainly of the spray dry absorber; particulate control device, such as an electrostatic precipitator (ESP) or a fabric filter; and recycling disposal devices for the reaction products. The sorbent for SO_2 absorption is typically lime or calcium oxide. Lime is mixed with an excess of water or is slaked to produce lime slurry, which is also called lime milk. Lime slurry is atomized to a cloud of fine droplets in the spray dry absorber where SO_2 is also removed from the flue gas. Water is evaporated by the heat of the flue gas, usually with a sufficient residence time (about 10 sec) for the SO_2 and other acid gases such as SO_3 and HCl to react simultaneously with hydrated lime to form calcium sulfite/sulfate and calcium chloride. The process chemistry associated with SO_2 removal from the flue gas is a simple acid-base absorption reaction between SO_2 and hydrated lime. The absorption chemistry is strongly affected by such factors as flue gas temperature, gas humidity, SO_2 concentration in the flue gas, and atomized slurry droplet size. The by-product consists of a dry mixture of calcium sulfite, sulfate, fly ash, and non-reacted lime. Spray dry absorbers usually operate at 290–300°C above the saturation temperature, where the saturation temperature of flue gas is between 45°C and 55°C. The indicative range of the method's application is for SO_2 <1000 mg/m^3 in the raw gas.

Furnace Sorbent Injection

Furnace sorbent injection involves direct injection of a dry sorbent into the gas stream of a boiler furnace. Typical sorbents include pulverized limestone ($CaCO_3$) and dolomite ($CaCO_3 \cdot MgCO_3$). Addition of heat in the furnace results in calcination of the sorbent to produce reactive CaO particles. The surface of these particles reacts with the SO_2 in the flue gas to form calcium sulfite ($CaSO_3$) and calcium sulfate ($CaSO_4$). These reaction products are then captured along with the fly ash by the particulate control device, typically an ESP or fabric filter. The SO_2 capture process continues into the precipitator and into the filter cake of the fabric filter. The critical temperature range for the limestone reaction in furnace sorbent injection is 980–1230°C. Once a reactive lime (CaO) is produced, it must have sufficient time in the critical temperature range. Hydrated lime has two reaction windows: 980–1230°C and around 540°C. Thermochemically, $CaSO_4$ is not stable at temperatures above 1260°C in the environment typical for combustion of high sulfur fossil fuels, i.e. 2000–4000 ppm SO_2, e.g., for coal firing. The lower temperature limit for the formation of $CaSO_4$ depends on complex interactions between sulfation kinetics, crystal growth and sintering, and on the build up of a barrier layer of $CaSO_4$ on the surface of the reactive CaO.

About 50% of SO_2 removal efficiency can be achieved at a sorbent mole ratio (Ca/S) of 4–5 when limestone is injected into the boiler furnace at near optimum operation. The SO_2 removal and limestone utilization efficiencies are lower than with other FGD systems. The process is suitable for low-sulfur-containing fuels and for use in small plants. With circulating fluidized bed combustion, this technique operates under optimal temperature conditions, using low-temperature combustion in the range of 800–950°C with higher absorption efficiency for a surplus of sorbent above a factor of 2.

Duct Sorbent Injection (Dry FGD)

Duct sorbent injection means injection of a calcium- or sodium-based sorbent into the flue gas, between the air heater and the existing ESP or fabric filter. The most common types of duct sorbent injections are as follows: 1) dry hydrated lime, which requires humidification; 2) dry sodium injection, which does not require humidification; and 3) lime slurry injection or in-duct scrubbing, which does not require a separate humidification step.

After injection, the sodium bicarbonate decomposes thermally to form sodium carbonate. After the initial sorbent surface of the sodium carbonate has reacted with SO_2 to form sodium sulfite or sulfate, the reaction slows down because of pore pluggage (which resists the gas-phase diffusion of SO_2). For the reaction to continue, the sorbent particle must decompose further. This decomposition evolves H_2O and CO_2 gases into the surrounding atmosphere, creating a network of void spaces throughout the particle. This process exposes fresh reactive sorbent and allows SO_2 to diffuse once again into the particle interior. This increase in surface area is in the order of 5–20 times of the original surface area, depending on the specific sorbent considered. The rates of decomposition and subsequent sulfation of the sodium compound particle represent a complicated function of gas temperature, rate of heat transfer to the particle, H_2O and CO_2 content in flue gas, partial pressures, and effects of other flue gas components.

Hybrid Sorbent Injection

Hybrid sorbent injection is a combination of furnace sorbent injection and duct sorbent injection to improve SO_2 removal efficiency. A specific feature of hybrid sorbent injection lies in the application of limestone as a sorbent.

Dry Techniques

Circulating Fluid Bed Dry Scrubber

The circulating fluid bed (CFB) process is a type of dry scrubber, but separate from either the spray dryer scrubber or sorbent injection.

Sodium Sulfite Bisulfite Process

The sodium sulfite bisulfite process is the most widely used of the regenerable processes. Commercial plants using this process are operated for industrial boilers and power stations burning hard coal, lignite, oil, and petroleum coke. The process is based on sodium sulfite/bisulfite equilibrium. The flue gas first passes to the prescrubber unit, which saturates and cools the gas as well as removes halides and some of the remaining particulates. A variety of prescrubber systems can be used, but each includes a recirculating water stream to ensure good contact, plus a purge stream to control the concentration of chloride and solids. Absorption is countercurrent, and as the liquor travels down the column, more sulfite is converted to bisulfite until it leaves the column. The small quantity of fly ash passing through the prescrubber is also captured in the solution and is removed by filtration before the clean liquor is passed to an intermediate storage area before regeneration. Two side reactions occur in the process, one in the absorption area and the other during regeneration. In the absorption area, a small quantity of sodium sulfate, which lowers the liquor capacity for SO_2 absorption, is produced due to the oxygen content in the flue gas. During regeneration, a small quantity of sodium thiosulfate is formed by a disproportionation reaction. This reaction is autocatalytic so the thiosulfate concentration is maintained at a low level by purging. Ethylenediaminetetraacetic acid is also added to inhibit oxidation.

Magnesium Oxide Process

The magnesium oxide process is a regenerable wet scrubbing process, which uses magnesium hydroxide solution as sorbent. It is essentially the same as for the limestone wet scrubber except for the regeneration step of the spent sorbent. The flue gas then enters the scrubber where the SO_2 is absorbed by aqueous slurry of magnesium sulfate formed from the magnesium hydroxide sorbent. The magnesium sulfite/sulfate is calcined at about 900°C in the presence of carbon to regenerate magnesium oxide, which is returned to the absorption system.

According to the BAT recommendations, the concentration of SO_2 in the gases emitted to the atmosphere as a result of combustion should not exceed 50 mg/N m^3 for wet limestone scrubbing, and SO_2 <100 mg/N m^3 for semidry sorbent injection.

Conclusion

The control of emissions of SO_2 in Europe and the development of emissions regulations are guided by the IPPC BREF document, which specifies, among others, the BAT guidelines, for limiting sulfur oxide emissions from power stations, non-ferrous metals metallurgy, and other sources. In November 2010, Directive 2010/75/EU of the European Parliament and of the Council on Industrial Emissions (IPPC) was published where new standard emission for combustion plants using solid or liquid fuels with the exception of gas turbines and gas engines are presented. The emission limit values (mg/N m^3) for SO_2 for the plants are presented on Tables 3 and 4.

The cited directive also concerns other branches of industry, such as production and processing of metals, mineral industry, chemical industry, and waste management.

These emission reductions result from a combination of emission controls, adoption of advanced electric technologies, and a shift away from the direct end use of coal with increasing income levels. Only under a scenario where incomes in developing regions increase slowly do global emission levels

TABLE 3 Emission Limit Values (mg/N m³) for SO_2 for Combustion Plants Using Solid or Liquid Fuels with the Exception of Gas Turbines and Gas Engines

Total Rated Thermal Input (MW)	Coal and Lignite and Other Solid Fuels	Biomass	Peat	Liquid Fuels
50 100	400	200	300	350
100–300	250	200	300	250
>300	200	200	200	200

Source: Data from Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe.[8]

TABLE 4 Emission Limit Values (mg/N m³) for SO_2 for Combustion Plants Using Gaseous Fuels with the Exception of Gas Turbines and Gas Engines

In general	35
Liquefied gas	5
Low calorific gases from coke oven	400
Low calorific gases from blast furnace	200

Source: Data from Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe.[8]

remain at close to present levels over the next century. Under a climate policy that limits emissions of carbon dioxide, SO_2 emissions fall in a relatively narrow range. In most cases, the relative climatic effect of SO_2 emissions decreases dramatically to a point where SO_2 is only a minor component of climate forcing by the end of the century. Ecological effects of SO_2, however, could be significant in some developing regions for many decades to come.

References

1. Ng'walali, M.P.; Koreeda, A.; Kibayashi, K.; Tsunenari, S. Fatalities by inhalation of volcanic gas at Mt. Aso crater in Kumamoto. Jpn. Legal Med. **1999**, *1*, 180–184.
2. Komarnisky, L.A.; Christopherson, R.J.; Basu, T.K. Sulfur: Its clinical and toxicologic aspects. Nutrition **2003**, *19,* 54–61.
3. McGee, K.A.; Doukas, M.P.; Kessler, R.; Gerlach, T. Impacts of volcanic gases on climate, the environment, and people. *U. S. Geol. Surv. Open File Rep.* **1997**, *97–262,* 2.
4. Blank, L.W.; Roberts T.M.; Skeffington R.A. New perspectives on forest decline. Nature **1988**, *336,* 27–30.
5. Krzyzanowski, M.; Cohen, A. Update of WHO air quality guidelines. Air Qual. Atmos. Health **2008**, *1*, 7–13.
6. Vahlsing, C.; Smith K.R. Global review of national ambient air quality standards for PM10 and SO_2 (24 h). Air Qual. Atmos. Health **2011**. doi: 10.1007/s11869-010-0131–2.
7. Emery, J.C. *Report on Air Quality in the State of Maine 2006;* Department of Environmental Protection Bureau of Air Quality, Research Triangle Park, North Carolina, U.S. EPA, 2006.
8. Directive 2008/50/EC of the European Parliament and of the Council of May 21, 2008 on ambient air quality and cleaner air for Europe.
9. U.S. EPA. Environmental Protection Agency. Part II. Primary national ambient air quality standard for sulfur dioxide. Final Rule Fed. Regist. **2010**, *75* (119), 35520–35602.
10. *Air Quality in Asia: Status and Trends*; Clean Air Initiative for Asian Cities Center: Pasig City, Philippines, 2010.
11. ISO 10498:2004. Ambient air—Determination of sulfur dioxide—Ultraviolet fluorescence method.
12. Chang, W.; Okamoto, M.; Korenaga, T. A simple fluorometric method for determination of sulfur oxide in ambient air with a passive sampler. Environ. Sci. **2006**, *13,* 257–262.

13. Gayathri, N.; Balasubramanian, N. Spectrophotometric determination of sulfur dioxide in air, using thymol blue. J. AOAC Int. **2001**, *84* (4), 1065–1069.
14. ISO 10396:2001. Stationary source emissions—Sampling for the automated determination of gas emission concentrations for permanently-installed monitoring systems.
15. Zhang, D.; Maeda, Y.; Munemori, M. Chemiluminescence method for direct determination of sulfur dioxide in ambient air. *Anal. Chem.* **1985**, *57* (13), 2552–2555.
16. He, Z.; Wu, F.W.; Meng, H.; Yuan, L.J.; Luo, Q.Y.; Zeng, Y. Chemiluminescence determination of sulfur dioxide in air using tris(1,10-phenanthroline)ruthenium-potassium iodate system. Anal. Lett. **1999**, *32* (2), 401–410.
17. Salem, A.A.; Soliman, A.A.; El-Haty, I.A. Determination of nitrogen dioxide, sulfur dioxide, ozone, and ammonia in ambient air using the passive sampling method associated with ion chromatographic and potentiometric analyses. Air Qual. Atmos. Health **2009**, *2* (3), 133–145.
18. Velásquez, H.; Ramírez, H.; Díaz, J.; de Nava, M.G.; Sosa de Borrego, B.; Morales, J. Determination of atmospheric sulfur dioxide by ion chromatography in the city of Cabimas, Venezuela. J. Chromatogr. A **1996**, *739* (1–2), 295–299.
19. Sritharathikhun, P.; Oshima, M.; Motomizu, S. On-line collection/concentration and detection of sulfur dioxide in air by flow-injection spectrophotometry coupled with a chromatomembrane cell. Anal. Sci. **2004**, *20,* 113–118.
20. Myoung, S.Y.; Geun, S.C.; Meyerhoff, M.E. Differential ion-selective membrane electrode-based potentiometric gas-sensing cells with enhanced gas sensitivity. Anal. Chim. Acta **1990**, *237,* 115–125.
21. ISO 6767:1990. Ambient air—Determination of the mass concentration of sulfur dioxide—Tetrachloromercurate (TCM)/pararosaniline method.
22. Smith, S.J.; Pitcher, H.; Wigley, T.M.L. Global and regional anthropogenic sulfur dioxide emissions. Glob. Planet. Change, **2001**, *29,* 99–119.
23. Shirsat, S.V.; Graf H.F. An emission inventory of sulfur from anthropogenic sources in Antarctica Atmos. Chem. Phys. **2009**, *9*, 3397–3408.
24. Symonds, R.B.; Rose, W.I.; Bluth, G.J.S.; Gerlach, T.M. Volcanic gas studies—Methods, results, and applications. Rev. Miner. **1994**, *30,* 1–66.
25. Delgado-Granados, H.; Cárdenas González, L.; Piedad Sánchez, N. Sulfur dioxide emissions from Popocatépetl volcano (Mexico): Case study of a high-emission rate, passively degassing erupting volcano. J. Volcanol. Geotherm. Res. **2001**, *108* (1–4), 107–120.
26. Bluth, G.J.S.; Casadevall, T.J.; Schnetzler, C.C.; Doiron, S.D.; Walter, L.S.; Krueger, A.J. Evaluation of sulfur dioxide emissions from explosive volcanism: The 1982–1983 eruptions of Galunggung, Java, Indonesia. J. Volcanol. Geotherm. Res. **1994**, *63* (3–4), 243–256.
27. Halmer, M.M.; Schmincke, H.-U.; Graf, H.-F. The annual volcanic gas input into the atmosphere, in particular into the stratosphere: A global data set for the past 100 years. J. Volcanol. Geotherm. Res. **2002**, *115,* 511–528.
28. Bandy, A.R.; Thornton, D.C.; Blomquist, B.W.; Chen, S.; Wade, T.P.; Ianni, J.C.; Mitchell, G.M.; Nadler, W. Chemistry of dimethylsulfide in the equatorial Pacific atmosphere. Geophys. Res. Lett. **1996**, *23,* 741–744.
29. Cropp, R.A.; Norbury, J.; Gabric, A.J.; Braddock, R.D. Modeling dimethylsulphide production in the upper ocean. Glob. Biogeochem. Cycles **2004**, *18,* 1–21.
30. Seguin, N.; Eaton, S.; Wadleigh, M.A.; Sharma, S. Elevated biogenic sulfur dioxide concentrations over the North Atlantic. Atmos. Environ. **2010**, *44*, 1139–1144.
31. Mihalopoulos, N.; Kerminen, V.M.; Kanakidou, M.; Berresheim, H.; Sciare J. Formation of particulate sulfur species (sulfate and methanesulfonate) during summer over the Eastern Mediterranean: A modelling approach. Atmos. Environ. **2007**, *41,* 6860–6871.
32. McGonigle, A.J.S. Measurement of volcanic SO_2 fluxes with differential optical absorption spectroscopy. J. Volcanol. Geotherm. Res. **2007**, *162,* 111–122.

33. Galle, B.; Oppenheimer, C.; Geyer, A.; McGonigle, A.J.S.; Edmonds, M.; Horrocks, L.A. A miniaturised UV spectrometer for remote sensing of SO_2 fluxes: A new tool for volcano surveillance. J. Volcanol. Geotherm. Res. **2003**, *119,* 241–254.
34. Sutton, A.J.; Elias, T.; Gerlach, T.M.; Stokes, J.B. Implications for eruptive processes as indicated by sulfur dioxide emissions from Kilauea Volcano, Hawaii, 1979–1997. J. Volcanol. Geotherm. Res. **2001**, *108,* 283–302.
35. Edmonds, M.; Herd, R.A. A volcanic degassing event at the explosive-effusive transition. Geophys. Res. Lett. **2007**, *34*, L21310.
36. Salerno, G.G.; Burton, M.R.; Oppenheimer, C.; Caltabiano, T.; Tsanev, V.I.; Bruno, N. Novel retrieval of volcanic SO_2 abundance from ultraviolet spectra. J. Volcanol. Geotherm. Res. **2009**, *181* (1–2), 141–153.
37. Realmuto, V.J.; Sutton, A.J.; Elias T. Multispectral imaging of sulfur dioxide plumes from the East Rift Zone of Kilauea Volcano, Hawaii. J. Geophys. Res. **1997**, *102,* 15057–15072.
38. Oppenheimer, C.; Francis, P.; Burton, M.; Maciejewski, A.J.H.; Boardman, L. Remote measurement of volcanic gases by Fourier transform infrared spectroscopy Appl. Phys. B **1998**, *67,* 505–515.
39. Weibring, P.; Edner, H.; Svanberg, S.; Cecchi, G.; Pantani, L.; Ferrara, R.; Caltabiano, T. Volcanic sulphur dioxide emissions using differential absorption lidar (DIAL), differential optical absorption spectroscopy (DOAS) and correlation spectroscopy (COSPEC). Appl. Phys. B **1998**, *67,* 419–425.
40. Cao, H.N.; Li, S.; Fukuchi, T.; Fujii, T.; Collins, R.L.; Wang, Z.; Chen, Z. Measurement of tropospheric O_3, SO_2 and aerosol from a volcanic emission event using new multi-wavelength differential-absorption lidar techniques Appl. Phys. B **2006**, *85,* 163–167.
41. Svanberg, S. Geophysical gas monitoring using optical techniques: Volcanoes, geothermal fields and mines. Optics Lasers Eng. **2002**, *37* (2–3), 245–266.
42. Barrancos, J.; Roselló, J.I.; Calvo, D.; Padrón, E.; Melián, G.; Hernández, P.A.; Pérez, N.M.; Millán, M.M.; Galle, B. SO_2 emission from active volcanoes measured simultaneously by COSPEC and mini-DOAS. Pure Appl. Geophys. **2008**, *165* (1), 115–133.
43. Elias, T.; Sutton, A.J.; Oppenheimer, C.; Horton, K.A.; Garbeil, H.; Tsanev V.; McGonigle, A.J.S.; Williams-Jones, G. Comparison of COSPEC and two miniature ultraviolet spectrometer systems for SO_2 measurements using scattered sunlight. Bull. Volcanol. **2006**, *68,* 313–322.
44. EPA's 2008 Report on the Environment Highlights of National Trends, 2008, available at http://www.epa.gov/roe. (Accessed May 2011).
45. Corbett, J.J.; Fischbeck, P.S.; Pandis, S.N. Global nitrogen and sulfur inventories for oceangoing ships. J. Geophys. Res. **1999**, *104,* 3457–3461.
46. Integrated Pollution Prevention and Control Reference Document on Best Available Techniques for Large Combustion Plants, European Commission, July 2006.
47. Regulation (EC) No. 166/2006 of the European Parliament and of the Council of 18 January 2006 concerning the establishment of a European Pollutant Release and Transfer Register and amending Council Directives 91/689/EEC and 96/61/EC.
48. Integrated Pollution Prevention and Control (IPPC) Reference Document on Best Available Techniques in the NonFerrous Metals Industries, European Commission 2001.
49. Integrated Pollution Prevention and Control (IPPC) Reference Document on Best Available Technologies (BREF) in Pulp and Paper Industry, European Commission, 2001.
50. Integrated Pollution Prevention and Control (IPPC) Reference Document on Best Available Techniques for Mineral Oil and Gas Refineries, European Commission, 2003.
51. Srivastava, R.K. Controlling SO_2 emissions: A review of technologies, EPA/600/R-00/093 November 2000; U.S. Environmental Protection Agency Office of Research and Development: Washington, D.C.

38

Vanadium and Chromium Groups

Imad A.M. Ahmed

Introduction

The vanadium group of elements consists of vanadium (V), niobium (Nb, previously called columbium), tantalum (Ta), and dubnium (Db). The vanadium metals lie in the d-block of the periodic table and are found in Group 5. The chromium group consists of chromium (Cr), molybdenum (Mo), tungsten (W), and seaborgium (Sg). Dubnium and seaborgium are synthetic elements whose most stable isotopes are $^{268}_{108}\mathrm{Db}(t_{1/2} = 28\ \mathrm{hr})$ and $^{271}_{106}\mathrm{Sg}(t_{1/2} = 1.9\ \mathrm{min})$;[1] thus, these two elements are not discussed here. Chromium and tungsten are probably the most familiar metals of the two groups of elements. This is because of chromium's role in the manufacture of stainless steel and the famous use of tungsten as a filament in lightbulbs. This entry summarizes the impact of V, Nb, Ta, Cr, Mo, and W contaminations on the environment and human, animal, and plant health. The geochemical occurrences of these elements are first discussed to establish an understanding of their natural distribution in the environment. These elements are widely used in various industrial applications, especially in metal industries; their applications and levels in the environment are discussed next. Potential sources of these elements, which humans and the environment can be exposed to, are described in some details. Finally, the metabolism, health effects, and regulatory measures taken to reduce or prevent the release of such trace metals to the environment are also discussed.

Geochemical Occurrences

All the transition elements of the vanadium group have odd atomic numbers, which could explain their low cosmic and terrestrial abundances compared to those of the titanium (Group 4) and chromium group elements.[2] This also resulted in the presence of only one stable naturally occurring isotope of each of these elements. Vanadium (V) is the 19th most abundant element in the Earth's crust. The ores that contain vanadium include patronite (VS_4), vanadinite [$Pb_5(VO_4)_3Cl$], dechenite [$PbZn(VO_4,AsO_4)OH$],

TABLE 1 World Mining of V, NB, Ta, Cr, Mo, and W (tons)

Element	Production as of 2010[a]	Location of Major Reserves[b]
Vanadium, V	56,000	China, South Africa, and Russia
Chromium, Cr	22,000	South Africa, India, and Kazakhstan
Niobium, Nb	63,000	Brazil and Canada
Molybdenum, Mo	234,000	China, United States, Chile, Peru, Canada, Mexico, Russia, Iran, and Mongolia
Tantalum, Ta	670	Brazil, Mozambique, Rwanda, and Australia
Tungsten, W	61,000	China, Russia, Bolivia, Austria, and Portugal

[a] World total mine production (rounded).[133]

[b] Reserves are arranged in decreasing order according to their mine production as of 2010.

descloizite [$Pb(Zn, Cu)(VO_4)OH$], pucherite ($BiVO_4$), roscoelite (mica containing V), and volborthite [$Cu_3V_2O_7(OH)_2{\cdot}2H_2O$].[3,4] Vanadium (primarily as VO^{-3}) is commonly found in admixture with many minerals such as carnotite $K_2(UO_2)_2(VO_4)_2$. Niobium (Nb) and tantalum (Ta) do not occur naturally as free metals but are commonly found together in nature at trace concentrations as iron niobate [$(Fe, Mn)(NbO_3)_2$] and iron tantalate [$(Fe, Mn)(TaO_3)_2$] in the form of niobite and tantalite minerals, respectively.[5] Large deposits of Nb have been found associated with carbonatites (carbonsilicate rocks) in the form of pyrochlore [$(Na, Ca)_2Nb_2O_6(OH, F)$].[6]

Chromium (Cr) ranks 21st in abundance in the Earth's crust and is found in a number of minerals with chromite ($FeCr_2O_4$) and crocoite ($PbCrO_4$) as the principal ores of Cr. Molybdenum (Mo) ranks 54th in abundance in the Earth's crust as it is found in a limited number of ores, including wulfenite ($PbMoO_4$), molybdenite (MoS_2), and powellite ($CaMoO_4$). Molybdenum is also recovered as a by-product of copper and tungsten mining. Tungsten (W) ranks 58th in abundance and is found in a number of minerals including tungstite ($WO_3{\cdot}H_2O$), wolframite [$FeMn(WO_4)_2$], and scheelite ($CaWO_4$).[3,5,7,8] The world mining production and principal reservoirs of V, Nb, Ta, Cr, Mo, and W are given in Table 1.

Overview of Current Uses and Environmental Levels

Uses

The metals of the vanadium group are characterized by their good structural strengths and their uses in manufacturing high-strength low-alloy steels. Vanadium metal has a low-fission neutron cross section, making it useful in nuclear applications. A major application (~85% of mined metal) of vanadium is the production of heat-resistant and high-strength steels and alloys. Vanadium is commonly used in the aerospace industry as a component in jet aircraft engines. Vanadium oxides are important catalysts in inorganic and organic chemical industries such as in plastic and sulfuric acid productions.[7,9] Niobium exhibits many excellent corrosion resistance properties, which make it an important coating material in biomedical applications. It is also commonly used in manufacturing aerospace equipment and missiles due to its lightweight and strength at high temperatures. Furthermore, pure Nb exhibits excellent superconductive properties, making it the primary material in manufacturing superconducting magnets especially when alloyed with tin and zirconium.[7,9,10] Tantalum is used in manufacturing high-temperature hardware because of its very high melting point (see Table 2). Because of its excellent resistance to corrosion by hot and concentrated mineral acids (except hydrofluoric acid), Ta is widely used in various chemical industries including electrolytic capacitors.[11,12] Tantalum carbide (TaC) has a higher melting point (3985°C) than metallic Ta, which is the reason why TaC powder is added in the manufacture of hard metals. In recent years, the use of TaC has decreased due to its high cost and has been replaced by chromium and vanadium carbides. About 900,000 kg of Ta is consumed annually worldwide with the electronics industry being its largest consumer.

Chromium is mainly used to harden steel and to manufacture stainless steel and special alloys. Chromium is also used in the plating industry to produce hard and beautifully finished surfaces that

TABLE 2 Selected Physical Properties of Trace Elements of Groups 5 and 6

Element	Atomic Number	Atomic Mass[a]	Atomic Radius (pm)	Density 20°C (g cm−3)	Melting Point (°C)	Boiling Point (°C)	Oxidation States[b]
Vanadium, V	23	50.9	192	6.11	1910	3407	2,3,4,[5]
Chromium, Cr	24	51.9	185	7.20	1907	2671	2,[3], 4,5,[6]
Niobium, Nb	41	92.9	208	8.57	2477	4744	2,3,4,[5]
Molybdenum, Mo	42	95.9	201	10.2	2623	4639	2,3,4,5,[6]
Tantalum, Ta	73	180.9	209	16.65	3017	5458	3,4,[5]
Tungsten, W	74	183.8	202	19.3	3422	5555	2,3,4,5,[6]

[a] Average values for atoms having the main oxidation state.
[b] The main oxidation states are given in square brackets.
Source: Enghag[4] and Lide.[7]

resist corrosion. It is also used as a coloring element in glass. Chromium forms many important compounds such as potassium chromate (K_2CrO_4) and dichromate ($K_2Cr_2O_7$), which are commonly used as oxidizing agents in analytical chemistry and in tanning leather.[7,9]

Molybdenum is primarily used in iron and steel industries. It is used in certain nickel-based alloys such as the Hastelloy® alloys, which are heat-resistant and corrosion-resistant to chemical solutions. It is also used as a catalyst in the refining of petroleum. Useful compounds of Mo include molybdenum sulfide (MoS_2), which is used as a solid lubricant because of its low friction properties and resistance to decomposition at high temperatures.[13]

Tungsten has the highest melting point of all metals (Table 2) and it has the highest tensile strength at temperatures higher than 1650°C, making it a valuable element in the alloy industry. Tungsten and its alloys are extensively used as filaments in lightbulbs, electron and television tubes, x-ray target, and heating elements for electrical furnaces. It has excellent corrosion resistance properties and it is attacked only slightly by mineral acids. Similar to tantalum, tungsten carbide has a high melting point at 2870°C and is an extremely hard material and hence is used in manufacturing cutting and drilling tools. Tungsten and its salts are widely applied in chemical and tanning industries and as high-temperature lubricants in the form of tungsten disulfide, which remains stable up to ~500°C.[14] Consequently, V, Mo, Cr, Ta, and W are regarded as essential commodities and metals of high strategic importance for years to come[15]

Environmental Levels

Environmental contamination is generally associated with the quantities of trace elements that are added to soil or water from anthropogenic sources such as sewage sludge, industrial discharge, and combustion of fossil fuel. Sewage sludge, for example, has valuable agronomic properties because it can improve soil's biological, chemical, and physical properties. However, the concentrations of trace elements in sewage sludge may be many orders of magnitude higher than the background concentrations in agricultural soils (see Tables 3 and 4). For example, Berrow and Webber[16] reported total metal contents of 3–3000, 2–260, 40–8800, 200–8000, 2–30, 20–5300, 120–3000, 200–400, and 700–4900 mg kg^{-1} for Cd, Co, Cr, Cu, Mo, Ni, Pb, V, and Zn, respectively, in sewage sludges from rural and industrial towns in England and Wales. In contrast, the reported concentrations of trace metals in sewage sludge in agricultural soils of Sweden (Table 3) were comparable to the background values in soils. Once applied, sewage sludge can have very long residence times in soil, causing potential consequences on the plant, animal, and human health.[17–20] The combustion of fossil fuel is another potential contributor to the release of toxic trace metals to the environment. Coal combustion is used to generate electricity in coal-burning power plants. In the United States alone, the amount of combustion residues of coal such as fly ash produced annually amounts to about 106 million tons.[21] Only one-third of the produced fly ash is recycled by using it in

TABLE 3 Typical Concentration (mg kg^{-1}) of V, Nb, Ta, Cr, Mo, and W in Fly and Bottom Ash and Sewage Sludge

Element	Fly Ash	Bottom Ash[a]	Sewage Sludge (I)[b]	Sewage Sludge (II)[c]
Vanadium, V	95–652	50–275	7.9	200–400
Chromium, Cr	37–651	40–4710	33	40–8800
Niobium, Nb	16.0 to <22	–	4.5	–
Molybdenum, Mo	7.1–236	2.8–443	6.7	2–30
Tantalum, Ta	0.5–2.6	–	0.94	–
Tungsten, W	2.9–42	–	7.9	–

[a] A granular combustion residue collected from the bottom of coal furnaces in power plants.

[b] Sewage sludge from agricultural soils of Sweden.[32]

[c] Sewage sludge from rural and industrial towns in England and Wales.[16]

Source: Bradl[23] and PECH.[129]

various applications such as cement making and ceramics.[22] The concentrations of trace elements in coal residues depend on the elemental composition of the parent coal and the conditions during combustion. The typical concentrations of V, Nb, Ta, Cr, Mo, and W in fly ash are listed in Table 3. More details on the coal residues and their characterization are available elsewhere.[23–25]

Vanadium, Niobium, and Tantalum

Although there are not many reports on V, Nb, or Ta pollution in soils, recent research shows that industrial activities such as mining and burning of fossil fuel will increase the deposition of these elements to soils and surface waters. The three elements are relatively common in the terrestrial environment at average contents of 135, 20, and 2 mg kg^{-1} in the Earth's crust for V, Nb, and Ta, respectively (Table 4). The most common form of V in soil solutions is the vanadyl cation (VO^{+2}) and its associated complexes.[26] Under acidic conditions, the pervanadyl ion (VO^{+}) dominates the V chemical species whereas vanadate ions ($VO4^{-}$) dominate under extremely alkaline environments. Similar to ferric ions, vanadic ions (V^{+3}) hydrolyze in water to form hydroxide species such as VOH^{+2} and $V(OH)^{+}$. While Fe^{+3} is very difficult to oxidize to higher oxidation states, V^{+3} can be oxidized to the (IV) and (V) states. Vanadium is an oxophilic metal because of its tendency towards oxygen, forming a number of oxides including V_2O_5, VO_2, V_2O_3, and VO. The chemical and physical properties of Nb and Ta are almost identical (see Table 2) and both exhibit affinities to associate with Fe, Ti, and Zr. The chemistries of Nb and Ta are dominated by the (III) and (V) oxidation states, which limit the number of the oxides that could

TABLE 4 Typical Concentration (mg kg^{-1}) of V, Nb, Ta, Cr, Mo, and W in the Earth's Crust and Selected Non-Polluted Soils

Element	Crustal[a]	Sweden[b]	Japan[c]	Brazil[d]	United States[e]
Vanadium, V	135	69	180	320	80
Chromium, Cr	100	22	58	86	54
Niobium, Nb	20	12	10	25	11
Molybdenum, Mo	1.5	0.58	1.3	1.6	0.97
Tantalum, Ta	2	1.1	1.7	2.3	–
Tungsten, W	1.5	1.3	1.3	1.4	<0.16–0.17

[a] Mean concentrations of metals in the Earth's crust.[134,135]

[b] Agricultural soils of Sweden.[32]

[c] Agricultural soil of Japan.[136]

[d] Average values for Brazilian soils.[137]

[e] Data from U.S. soils.[138,139]

be formed under environmental conditions. Examples of the Nb and Ta oxides are Nb_2O_5, NbO_2, Ta_2O_5, and TaO_2. The Nb^{+3} ions are very unstable in air due to its rapid oxidation to form Nb^{+5} oxides. Ta^{+3} is also very unstable and the majority of Ta compounds tend to be limited to the (IV) oxidation state.[3,27]

The mean concentrations of V, Nb, and Ta in reference soils of Sweden, Japan, Brazil, and the United States were reported to fall in the range of 69–320, 10–25, and 1.1–2.3 mg kg^{-1}, respectively (Table 4). The concentrations of Nb in arable soils of the United Kingdom derived from different rocks were found in the range 31–300 mg kg^{-1}.[28] In China, the Nb and Ta content in reference soils were found in the range of 9.3–37.6 and 0.8–5.3 mg kg^{-1}, respectively.[29] Elevated concentrations of V, Nb, and Ta above the typical natural soils content are more likely due to residual soils derived from rocks enriched in these elements. Mining activities are major contributors to increasing the input of trace elements in soils and waters. An example of man-made pollution of V is the contamination of the Shuanghe and Donghae rivers in the Shaanxi province in Northwest China as a result of the collapse of the spillway from a vanadium mine (Reuters 2008). This spill turned the river water black, threatening the drinking water and the ecosystem of that region.

A number of reports have indicated that vanadium is available to plant uptake especially in the form of VO^{+2} species in acidic soils.[30,31] The reported V content in food plants is greatly variable, but concentrations of up to 2000 μg kg^{-1} in wild mushrooms, 840 μg kg^{-1} in spinach, and 20–600 μg kg^{-1} in other vegetables have been reported.[25,32] While the geochemical properties of Ta are similar to those of Nb, Ta is less mobile than Nb because of the lower solubility of its compounds. This means that Nb is more available to plant uptake than Ta especially under humid conditions. The common concentration range of Nb and Ta in food plants is 0.02–1.1 mg kg^{-1} and 0.013–0.48 μg kg^{-1}, respectively.[33] The reported concentrations of Nb and Ta in cereal grains are 0.5–1.7 mg kg^{-1} and 1.1–5 μg kg^{-1}, respectively.[32]

Chromium

Naturally occurring chromium exists in two main oxidation states of Cr(III) (chromic) and Cr(VI) (chromate). Cr(VI) species are strongly oxidizing and exist only in oxo species such as chromium trioxide CrO_3 and chromate CrO_4^{-2} When chromium trioxide is dissolved in water it forms chromic acid H_2CrO_4. Most Cr(VI) compounds are highly soluble except some metal chromates such as those of Pb, Ba, Sr, Ca, Ag, and Zn and hence are chemically mobile and bioavailable. In contrast, the majority of Cr(III) compounds are sparingly soluble in water. Similar to Fe(III), Cr(III) hydrolyzes in water, yielding different hydroxo species of varying solubilities. At pH > 5, Cr(III) precipitates as $Cr(OH)_3$. Cr(VI) speciation and behavior in the environment are pH dependent. Under acidic conditions, Cr(VI) shows high redox potential, meaning that it is strongly oxidizing but this oxidizing capacity reduced with increasing pH. Therefore, at low pH and in the presence of electron donor species such as natural organic matter and Fe^{+2}-containing minerals in soils, Cr(VI) are unstable and are reduced into the more stable Cr(III) form.[34,35] Under aerobic conditions, Cr(III) may be oxidized to Cr(VI), probably by Mn oxides, but this process is generally slow at pH values above 5.[36] Under anoxic conditions (e.g., hydromorphic soils), Cr(VI) could be reduced to Cr(III) by electron transfer processes occurring at reactive mineral surfaces or due to the presence of humic substances; such processes are more notable in soils at low pH. Therefore, under normal soil conditions, Cr is present predominantly in the Cr(III) state[13,36] and Cr(III) species are regarded less mobile and available to plant uptake compared to Cr(VI). The factors that influence transformations between Cr(III) and Cr(VI) in the environment have been extensively reviewed.[37–41]

Chromium occurs naturally in the Earth's crust at a median concentration of 100 mg kg^{-1}. The reported data of total Cr are <86 mg kg^{-1} in non-polluted soils (Table 4) and <12 mg kg^{-1} in natural water (Table 5). However, human activities have resulted in the release of chromium into the environment in larger amounts. Mining activities and tanning of leather are among the major sources of chromium. For example, in October 2010, a caustic sludge from a "red mud" reservoir at the Ajkai Timföldgyár alumina plant was spilt across a town called Ajka in Hungary. The accident happened

TABLE 5 Mean Concentrations of V, Nb, Ta, Cr, Mo, and W in Seawater (μg L^{-1}), River Water (μg L^{-1}), Oceans (μg L^{-1}), and Groundwater (mg L^{-1})

Element	Seawater[a]	River	Ocean Surface Water[b]	Groundwater
Vanadium, V	2.5	0.009–1.77	2.0	<1.0–10
Chromium, Cr	0.3	0.29–11.46	0.21	<1.0–5.0
Niobium, Nb	0.01	0.002–0.01	<0.005	
Molybdenum, Mo	7.0	0.04–2.69	10	<1.0–30
Tantalum, Ta	0.002[e]	–	$<2.5 \times 10^{-3}$	–
Tungsten, W	0.1	0.1–180	0.001	–

Source: Gaillardet et al.,[141] Sparks,[145] and Li.[142]

when the Western dam of the plant collapsed, releasing about a million cubic meters of reddish material with a pH of about 12 to the valley of the Torne river. The accident flooded the lower parts of the city of Devecser and the villages of Kolontár and Somlóvásárhely, covering about 800 ha of agricultural land by a red mud layer of 5–10 cm. Red mud is a by-product of the aluminum extraction from bauxite. This red mud spill contained elevated concentrations of toxic metals including V (870 mg kg^{-1}), Pb (160 mg kg^{-1}), and Cr (620 mg kg^{-1}), making this industrial accident the worst ecological disaster in the history of Hungary.[42,43] The scale of the toxic red mud spill and the damage that occurred to the nearby Kolontár village have been documented in a number of media (e.g., http://www.guardian.co.uk/world/gallery/2010/oct/05/hungary).

Chromium in ambient air originates primarily from industrial sources (e.g., metal industry) and the combustion of fossil fuel (see Table 3). Chromium is generally removed from the atmosphere as a result of precipitation or dry deposition. A number of studies have shown that most of Cr found in surface waters is due to atmospheric deposition.[44,45] The mean concentrations of Cr in freshwater and seawater are listed in Table 5.

Molybdenum

In marine systems, Mo is present at ~110 nM compared to ~0.5 nM for W, ~7.5 nM for Zn, and ~1.5 nM for Cu, which makes Mo the most abundant transition element in the sea.[46,47] The distribution of Mo in terrestrial environments is variable with an estimated average of 2–3 mg kg^{-1}. Molybdenum exists in soils mainly as soluble or adsorbed molybdate anion MoO_4^{-2} Molybdate is known to adsorb on oxide and clay minerals in soils with sorption increasing at lower pH. Thus, Mo is less mobile in soils under acidic conditions but can be released into the environment with increasing pH.[48,49] Hence, similar to Se, the bioavailability of Mo increases with liming soils. Molybdenum in soils forms strong complexes with natural organic matter and other organic ligands such as malic acid, tannins, and tannin-like compounds over a wide pH range, which helps prevent leaching Mo.[50,51] This also explains the elevated concentrations of Mo in peaty soils. The major soluble Mo species in natural waters are molybdic acid $H_2MoO_4^0$ and its oxyanions, $HMoO_4^-$ and $HMoO_4^{-2}$ Available Mo to plants and microorganisms is usually present in soil and waters as soluble MoO^{-4} anion.[52] The high solubility of MoO^{-4} makes it susceptible to leaching from soil into groundwater. However, like many other anions, once in solution MoO^{-4}, it becomes subject to a range of adsorption and desorption reactions by positively charged metal oxides (Fe, Al, and Mn) and clay minerals.[53]

The world median composition of Mo in stream waters is 0.5 μg L^{-1}. The analyses of more than 11,000 stream water samples in England and Wales from sites listed in the British Geological Survey (BGS) G-BASE data set had Mo concentrations within the range of 0.08–2.45 μg L^{-1} with a median of 0.57 μg L^{-1}. Studies on a number of BGS- monitored sites showed that the typical Mo concentrations in rivers are less than 1 μg L^{-1}.[54] As indicated in Table 4, the average concentration of Mo in soils ranges between 0.1 and 10 mg kg^{-1}, but higher concentrations were found in sewage sludges (Table 2). In ambient and non-polluted air, the mean concentration of Mo is ~0.02 μg m^{-3}.[55] With its widespread industrial applications, the data in Table 3 suggest that the atmospheric deposition is an important source of soil Mo near polluted cities.

Tungsten

Tungsten is a transition element with similar chemical properties of Mo and Cr. Tungsten exists in several oxidation states (0, II, III, IV, V, and VI) with W^{VI} being the most stable and common state in aqueous systems. The dominant aqueous species of W is tungstate $\left(WO_4^{-2}\right)$ especially in the pH range of 6.9–9.3. There are very little documentations on the distribution of W in soils and waters. The world average W content in soils is 1.7 mg kg^{-1}. The reported mean concentrations of W in non-polluted soils in Sweden, Japan, Brazil, and the United States are 1.3, 1.3, 1.4, and 0.165 mg kg^{-1} (Table 4), respectively. Data on W content in food, ecological effects, and anthropogenic release into the environment are not well established. While environmental tungsten levels are generally very low, it exists at high concentrations near mines or natural deposits.[56] For example, the historical mining activities in Cornwall in the South West England have resulted in the accumulation of high concentrations of W, As, Mn, Fe, Zn, Sn, and Cu in estuarine sediments with W levels up to ~650 mg kg^{-1}.[57] Owing to the adverse environmental and health effects of lead and depleted uranium, tungsten has been suggested as a replacement in military applications. The U.S. Green Armament Technology (GAT) program, a U.S. Army pollution prevention initiative, called for substitution of lead by tungsten in small-caliber ammunition. The 1999 nationwide ban of using lead shots for hunting in the United States, has also led to the substitution of lead shotgun shell ammunition and lead fishing weights in various North American and European countries.[58,59] The U.S. Fish and Wildlife Service approved the use of iron–tungsten–nickel, iron–tungsten, tungsten–bronze, tungsten–iron–copper–nickel, tungsten–nylon, and tungsten–tin–bismuth shots as non-toxic for migratory bird hunting.[60] Because W minerals are slightly soluble in water, its availability in natural waters is scant; this is evident by the reported concentrations of W in river, sea, and oceanic waters in Table 5. There is enough evidence that combustion of fossil fuel is a serious source of W contamination in air (see Table 3). The concentration of W is generally low (<0.05 mg kg^{-1}) in food plants.[61]

Metabolism and Health Effects

Vanadium

Vanadium is a trace element that is known to increase the growth rates of various microorganisms (e.g., green algae) and higher plants (e.g., tomato and lettuce) but its essentiality is not conclusive.[62] Until now, evidence that V has a biological function in higher animals is limited and not definitive.[63] However, in humans, a number of V compounds (e.g., vanadyl sulfate) have shown insulin-like action[64–66] and antitumor and anticancer properties,[67,68] indicating the potential therapeutic applications of V. The biochemical activities of V are related to the complexing ability of the vanadyl cation (VO^{+2}), and the chemical similarity of vanadate (VO^{-3}) to phosphate, which allows vanadate compounds to interact with numerous enzymes in living organisms by either inhibiting or activating them.[69] Once vanadate has entered the human body, it is reduced to vanadyl ion by reducing substances in the plasma such as ascorbic acid and subsequently bound to proteins such as transferrin, an iron-containing protein.[70] Toxic effects on human and animals are well established. For example, V compounds have an adverse effect on the reproductive and developmental functions of rats and mice including the decline in conception rate and fetal development.[71]Also, oral exposure to V and the reduction of vanadate ions into vanadyl ions in red blood cells can result in decreased number of red blood cells, increased blood pressure, and mild neurological effects.[72] The toxicity of V compounds increases as valency increases with V(V) being the most toxic species of vanadium. Vanadium oxides may also be more toxic than vanadium salts. In humans, exposure by inhalation to vanadium pentoxide causes diverse

TABLE 6 Regulations and Guidelines Applicable to V, Cr, Nb, Mo, Ta, and W in the Environment and Human Health

Element	NIOSH (REL) (mg m–3)	WHO	USEPA (water) (mg L–1)	Cancer Effect IARC[a]	EPA Water Reuse (mg L–1)
V	IDLH = 35; OSHA = 0.5 as V_2O_5	In air = 1 μg m^{-3} In water = N/A	N/A	Group 2B as V_2O_5	0.1 STU 1.0 LTU
Cr	Metal: TWA 0.5, OSHA 1, IDLH 250 Cr(VI) compounds: IDLH (CrO_3) = 15; OSHA = 0.005	In air = 1 μg m^{-3} as Cr(VI) Total in water = 0.05 mg L^{-1}	0.1 (total Cr)	Metallic Cr: Group 3 Cr(III): Group 3 Cr(VI): Group 1	0.1 STU 1.0 LTU
Nb	N/A	N/A	N/A	N/A	N/A
Mo	TWA = 15 IDLH = 5000 (as Mo) Mo soluble IDLH 1000 mg m^{-3} (as Mo) TWA 5 mg m^{-3} = OSHA	N/A	N/A	N/A	0.01 STU 0.05 LTU
Ta	TWA = 5 Short-term exposure = 10	N/A	N/A	N/A	N/A
W	TWA = 5 Short-term exposure = 10 OSHA= N/A IDLH = N/A	N/A	N/A	N/A	N/A

Note: IARC, International Agency for Research on Cancer; IDLH, immediately dangerous to life or health; LTU, long-term use; N/A, guideline data are not available; NIOSH, National Institute for Occupational Safety and Health; OSHA, Occupational Safety and Health Administration; REL, recommended exposure limit; STU, short-term use; TWA, time-weighted average; USEPA, U.S. Environmental Protection Agency; WHO, World Health Organization. [a] Group 1: the substance is carcinogenic to humans, Group 3: the substance is not classifiable as to its carcinogenicity to humans. Group 3B: possibly carcinogenic material to humans.

Source: WHO,[94] ATSDR,[97] WHO,[121] ATSDR,[130] USEPA,[131] and NIOSH.[132]

toxic effects on the respiratory, digestive, and central nervous systems.[130] Workers in the metal alloy and catalysis industries that use vanadium pentoxide and V-containing pigments for the ceramics industry are among the most exposed to airborne vanadium compounds. For more details on the toxic effects of vanadium pentoxide, see IARC 2006.[73] Other acute toxicity symptoms of V compounds in animals include weakness, loss of appetite, dehydration, weight loss, necrosis of lymphoid tissues, renal tubular necrosis, and death.[74,75] Ingested V seems to be excreted unabsorbed and hence there are very few reported cases on V poisoning via this route. The typical concentrations of V in plants are 0.27–4.2 mg kg^{-1},[76] but there are many examples of the toxic effect of vanadium on plants, and that toxicity is connected with V(V) forms in soil. For example, it was found that 5 mg V L^{-1} in a hydroponic medium caused iron-deficiency chlorosis in sugar beet plants and that the growth was reduced by 30%–50%. It is claimed that the cause for this toxic effect is the structural analogy between vanadate (VO^{-3}) and phosphate (PO^{-3}) ions.[77]

Normal concentrations in blood and urine are 0.05 μg L^{-1} and 0.5 μg L^{-1}, respectively. Occupational exposure limits and guidelines for vanadium pentoxide in workplace air are presented in Table 6. The U.S. Environmental Protection Agency (USEPA) considered vanadium pentoxide to be an extremely hazardous substance[78] and the International Agency for Research on Cancer (IARC) classified it as a possibly carcinogenic substance to humans (Carcinogen Group B). The World Health Organization (WHO) guidelines for V_2O_5 in air is 1 μg m^{-3}. Little attention has been paid to the determination of V(V) species in water, soil, or plants. Until recently, neither the USEPA nor the WHO has listed vanadium as a pollutant requiring urgent research and legislation.

Niobium and Tantalum

Tantalum and niobium are not essential or have a definite biological function to human, animals, or plants. Both element seems to have high affinity for both lung and plasma proteins.[79] Acute and chronic effects of Ta and Nb on humans and animals are related to their serious impact on respiratory tract and cardiovascular functions. For example, high levels of Ta, Co, and W were reported in the lung, blood, urine, and hair of 251 workers in the hard metal industry exposed to hard metal dust. These workers were diagnosed as having the "hard metal pneumoconiosis" disease that shows episodes of work-related subacute disease, and some patients evolve to lung fibrosis.[80,81] Respiratory paralysis was shown to be the cause of death in cats fed intravenously doses of 5 mg kg^{-1} Nb as $NbCl_5$. Renal injuries and deaths were also observed in mice and rats following parenteral injection of Nb as $NbCl_5$ or $KNbO_3$ at doses ranging from 20 to 50 mg Nb kg^{-1}. However, dietary levels of Nb up to 1% of $NbCl_5$ or $KNbO_3$ given over 7 weeks were ingested by rats without effect.[82] The median lethal dose (LD_{50}) values of Nb injected intraperitoneally in mice and rats were 12 and 14 mg kg^{-1} as $NbCl_5$ and 13 and 86 mg kg^{-1} as $KNbO_3$,[83] respectively, making niobium chloride one of the most toxic of the rarer chemicals. The LD_{50} values for Ta obtained after intraperitoneal injection of $TaCl_5$ in rats was 75 mg^{-1},[83] suggesting that Ta is less toxic than Nb. Tantalum and niobium are among the metals used in prostheses, especially dental and orthopedic implants, because of their excellent corrosion resistance and biocompatibility.[84–86] Their potential interactions with tissues have led to a number of studies that tend to suggest that Ta implants are biologically inert and pose little concern for human clinical pathophysiology.[87] Further details on metabolism and toxic effects of Ta and Nb are available elsewhere.[87–89] The evidence given above on the Nb and Ta and their compounds suggests that both elements should be regarded as toxic. The hazardous occupational exposure concentrations of Ta and Nb vary depending on their compounds and origin. Unfortunately, metabolism and toxicity data on Nb and Ta and their compounds remain scarce, which has hampered the development of regulations and risk assessments of these two elements in the environment. The current National Institute for Occupational Safety and Health (NIOSH) exposure limits for Ta (metal and oxide dusts) are 5 mg m^{-3}(10 hr) and 10 mg m^{-3} for the short-term (15 min) exposure.[132] There are no established limits for Nb in air at the moment. A limit of 0.01 mg L^{-1} has been incurred in the former USSR for Nb in drinking water. Regulations for Nb and Ta in drinking water, [90] surface water, or soils have not been established in the USEPA, WHO, or any other international health or environmental organization.

Chromium

Chromium is a non-essential element for plants, but its trivalent form, Cr(III), is considered biologically active as it is involved in carbohydrate, fat, and protein metabolism.[91,92] The first evidence of the biological function of Cr(III) in humans was found when some patients developed diabetic symptoms that were refractory to insulin but reversed by addition of chromium to parenteral nutrition.[93] The daily Cr(III) requirement for adults is estimated to be 0.5–2 mg of absorbable Cr(III).[94] In contrast, Cr(VI) is a strong oxidizing agent, in the form of chromate and dichromate, and is extremely toxic to animals and human. Unlike Cr(III), Cr(VI) has the ability to penetrate biological membranes by diffusion and reacts with cell contents before being reduced to Cr(III). The toxicity of Cr(VI) compounds is believed to be linked to its oxidative DNA impairment and mutagenic damage.[95] Inhalation of suspended particulates containing adsorbed Cr is a major route of exposure that presents a significant hazard to health as it increases incidence of lung cancer. Soluble salts of chromate are more toxic when administered parenterally (LD_{50} = 10–50 mg kg^{-1}) in comparison to dermal (LD_{50} = 200–350 mg kg^{-1}) and oral (LD_{50} = 1500 mg kg^{-1}) exposures.[96] Studies on rats showed that the most affected organs following the acute exposure to Cr(VI) are liver and kidney.[97] Symptoms of acute toxicity of Cr(VI) in humans include diarrhea, cyanosis, gastric distress, olfactory cleft obstruction, nosebleed, kidney and liver damage, skin cancer, and lung cancer.

On the other hand, Cr(III) compounds are considered non-toxic at small doses. The LD_{50} values calculated for orally administered $CrCl_3$, $Cr(CH_3COO)_3$, and $Cr(NO_3)_3$ were 1.87, 11.26, and 3.26 g kg kg^{-11}.[98] Chromium is highly toxic especially to higher plants when Cr(VI) content exceeds 100 μM kg^{-1} dry weight. Symptoms of Cr toxicity in plants appear as wilting of tops, chlorosis in young leaves, damage of cell membranes, and deterioration in growth. Extensive data are available elsewhere on the toxic effects and metabolism of Cr in humans, animals, [94,97–101] and plants.[102–104]

Currently, there are a number of regulatory measures for Cr substances in the environment. The IARC has classified Cr(VI) in Group I human carcinogens.[97] Minimal risk levels (MRLs) of daily human exposure to Cr(III) as insoluble trivalent chromium particulate compounds and Cr(VI) as chromium trioxide mist and other dissolved hexavalent chromium aerosols and mists were estimated as 0.005 and 5×10^{-6} mg m^{-3}, respectively. Other regulatory guidelines for Cr exposure and content in soils and water are listed in Table 6.

Molybdenum

Despite its low abundance, Mo is a component of several enzymes in human and animal organisms. Hence, it is an essential element for many biological processes and is required for the growth of many organisms. For example, Mo-dependent enzymes are necessary in plants and lower organisms for nitrogen fixation (nitrogenase), conversion of nitrate to ammonia, and various oxidation–reduction reactions.[105–107] Molybdenum is part of a coordination complex called molybdenum cofactor, which is formed between molybdopterin and Mo oxide (see Figure 1). This complex is required for three mammalian molybdoenzymes: xanthine dehydrogenase/oxidase, aldehyde oxidase, and sulfite oxidase.[108–110] Xanthine oxidase is important to maintain a balance between urate production and excretion. Low Mo intake in human is claimed to reduce the activity of xanthine dehydrogenase in tissues and the development of hyperuricemia and gout. Molybdenum cofactor deficiency is a rare disease in humans and is caused by the diminished ability to synthesize molybdoenzymes resulting in in neonatal seizures and other neurological symptoms identical to sulfite oxidase deficiency.[111] A study on goats has shown that a molybdenum-deficient diet (24 μg Mo kg^{-1}) caused a decline in conception rate and an increase in abortion rate.[112] Molybdenum is required in the smallest amount by plants and is taken up as the molybdate (MoO^{-2}) ion. However, Mo-deficient soil is known to cause undesirable effects in crops. For example, in the presence of nitrate fertilizers,

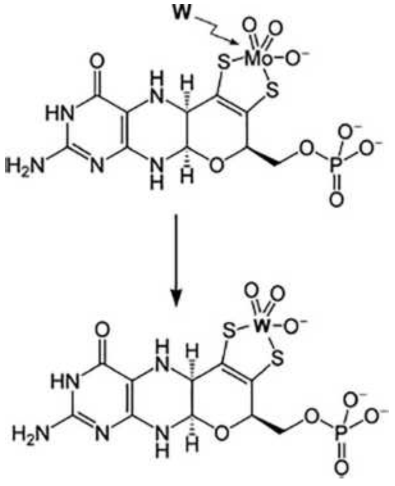


FIGURE 1 Chemical structure of molybdenum (top) and tungsten (bottom) cofactors.

molybdenum-deficient oat and wheat develop pale green leaves and necrotic regions on leaf margins, seeds are poorly developed, and overall plant growth is notably decreased.[113,114] In contrast, Mo toxicity in plants is generally rare and very little reports on its toxic effects in plants are available.[115] As a result of natural or anthropogenic activities, Mo may become sufficiently concentrated in soils and vegetation to cause poisoning. One form of Mo poisoning is the molybdenosis syndrome, which is a copper-deficiency disease that affects cattle and sheep.[116] The toxic effects are due to the depressing effects of Mo on the physiological availability of Cu. Symptoms of this disease are scouring, weight loss, bone disorders, depigmentation, and reproductive impairment.[117,118]

In human, water-soluble molybdenum compounds are readily absorbed when ingested but the rate of absorption depends on a number of factors including the chemical form of Mo. Toxic effects of Mo are related to the capability of Mo compounds to inhibit copper and sulfur absorption, which affects copper homeostasis and results in symptoms similar to copper deficiency. Excessive Mo intake in humans, and hence molybdoenzymes, contributes to diseases of gout, combined oxidase deficiency, and radical damage following cardiac failure.[119] It is notable that in comparison with V and Cr, occupational Mo exposure is uncommon and adverse effects of such exposure have rarely been reported. According to the National Academy of Sciences, the safe Mo intake levels are 0.015–0.04 mg day^{-1} for infants, 0.025–0.15 mg day^{-1} for children aged 1–10, and 0.075–0.25 mg day^{-1} for all individuals above the age of 10.[120] Molybdenum generally occurs at very low concentrations in drinking water, and therefore it has not been considered necessary to set a formal guideline value.[121]

Tungsten

In recent years, interest in tungsten geochemistry and toxicology has increased, especially after the diagnosis of 15 cases with acute lymphocytic leukemia in children and teenagers living in the town of Fallon in the north of Nevada. These cases were related to the high concentrations of W (0.27 to 742 μg L^{-1}) in groundwater due to weathering of local W-rich deposit.[122] Tungsten enters the human and animal body in a soluble form such as tungstate WO_4^{-2} Similar to molybdenum, W is a constituent of three classes of enzymes: aldehyde ferredoxin/oxidoreductase, formate dehydrogenase, and acetylene hydratase. These enzymes are essential catalysts of specific reactions in a number of living organisms. For example, formate dehydrogenase is responsible for the reversible conversion of CO_2 to formate. This makes W by far the heaviest metal with a biological function. Tungsten and molybdenum have almost equal atomic and ionic radii (0.68 Å) and similar electronegativity (1.4 and 1.3 for W and Mo, respectively). It is thus interesting to note that WO_4^{-2} and MoO_4^{-2} are isomorphic ions, which mean that W can replace Mo in molybdoenzymes. A number of studies have demonstrated the antagonistic effect of W (administered in diet as sodium tungstate) on the activity of the Mo enzyme xanthine dehydrogenase, sulfite oxidase and aldehyde oxidase, and nitrate reductase.[123,124] When W substitutes for Mo in molybdoenzymes, a catalytically inactive (or of low activity) analogue is produced, causing adverse health effects that is due to W toxicity and not because of Mo deficiency.[125,126] Similar to Nb and Ta, the major occupational hazard of W exists in the metal industry.[80,81] Airborne nanoparticles containing toxic metals including W, Hg, Ni, Ti, and Zn have been reported in the bloodstream of humans, [127] which has been suspected to trigger blood coagulation. Toxicity symptoms of exposure to airborne W include cough, expectoration, shortness of breath, and tightness in the chest and pulmonary fibrosis in advanced cases.[128] There are little data available on occupational exposures to W compounds that incriminate these substances as toxic or as hazardous to body organs other than lungs. Ingested W in trace quantities are excreted in urine and eliminated in feces. Regulations of W in aquatic and terrestrial environments are not available in most countries including the United States and the United Kingdom, and W is not part of routine testing programs. This probably explains the claim that W and its compounds are absent at measurable concentrations in drinking waters. The IARC has not classified W or its compounds as carcinogenic. Although some health regulations exist for W, environmental regulations applicable to W in environmental systems are absent (see Table 6).

Conclusion

The fate and geochemical behavior of V, Mo, and Cr in soils and freshwaters have been studied extensively but less so that of W, Ta, and Nb. These trace elements are released into the environment from the natural weathering of parental rocks and deposits and a result of anthropogenic activities such as industrial discharge, fuel combustion, sewage sludge, livestock manures, and other wastes. Although the environmental concentration of these elements is generally small, reports have shown that they may enter the food chain through plant absorption from contaminated soils, animal feed, or potable water supplies resulting in health problems. Soil pH and the contents of iron or manganese oxides, natural organic matter, and clay minerals are all important factors that control their transport, bioavailability, and ecotoxicity. Chemical speciation plays a critical role in defining the fate and bioavailability of these elements. For example, while Cr(III) is largely immobile in soil, limiting its entry into the food chain, Cr(VI) is highly mobile under neutral to alkaline soil pH conditions. From a soil management point of view, this is problematic because while raising soil pH by liming would decrease the bioavailability of some toxic elements such as Pb and Cd, it will have quite the opposite effect for Cr(VI). The booming steel industry in many countries including China has led to serious contaminations of Cr, V, W, and Mo in surface waters and soils. Although strict regulatory measures are available for Cr in soils and waters, these are still not complete for V and Mo and absent for Ta, Nb, and W.

References

1. Greenwood, N.N. Recent developments concerning the discovery of elements 101–111. Pure Appl. Chem. **1997,** *69* (1), 179–184.
2. Norman, N.G. Vanadium to dubnium: From confusion through clarity to complexity. Catal. Today **2003,** *78* (1–4), 5–11.
3. Halka, M.; Nordstrom, B. *Periodic Table of the Elements: Transition Metals;* Facts on File: New York, 2011.
4. Enghag, P. *Encyclopedia of the Elements,* 1st Ed.; WILEY-VCH Verlag GmbH: Weinheim, Chichester, 2004.
5. Wiberg, N.; Holleman, A.F.; Wiberg, E. *Holleman–Wiberg's Inorganic Chemistry,* 1st Ed.; Academic Press, 2001.
6. Deans, T. Trace elements in carbonatites and limestones. Nature **1968,** *220* (5168), 679–679.
7. Lide, D. *CRC Handbook of Chemistry and Physics: A Ready-Reference Book of Chemical and Physical Data, 90th Ed.*; CRC Press: Boca Raton, FL, 2009.
8. Wood, S.A.; Samson, I.M. The hydrothermal geochemistry of tungsten in granitoid environments: I. Relative solubilities of ferberite and scheelite as a function of T, P, pH, and mNaCl. Econ. Geol. **2000**, *95* (1), 143–182.
9. Wiley-VCH. *Ullmann's Encyclopedia of Industrial Chemistry,* 7th Ed.; Wiley-VCH: New York, 2011.
10. ASM. *ASM Handbook: Properties and Selection: Nonferrous Alloys and Special-Purpose Materials, 10th Ed.*; ASM International: Materials Park, Ohio, 1990; Vol. 2.
11. Buckman, RW. New applications for tantalum and tantalum alloys. JOM **2000**, *52* (3), 40–41.
12. Andersson, K.; Reichert, K.; Wolf, R. Tantalum and tantalum compounds. In: *Ullmann's Encyclopedia of Industrial Chemistry,* 7th Ed.; Wiley-VCH: New York, 2000.
13. Martin, J.M.; Donnet, C.; Le Mogne, T.; Epicier, T. Superlubricity of molybdenum disulphide. Phys. Rev. B 1993, *48* (14), 10583–10586.
14. Lassner, E.; Schubert W. *Tungsten: Properties, Chemistry, Technology of the Element, Alloys, and Chemical Compounds*; Springer: New York, 1999.
15. OECD (Organisation for Economic Co-operation and Development). *OECD Trade Policy Studies The Economic Impact of Export Restrictions on Raw Materials;* OECD Publishing, 2010.
16. Berrow, M.L.; Webber, J. Trace elements in sewage sludges. J. Sci. Food Agric. **1972**, *23* (1), 93–100.

17. McGrath, S.P.; Chaudri, A.M.; Giller, K.E. Long-term effects of metals in sewage sludge on soils, microorganisms and plants. J. Ind. Microbiol. **1995**, *14* (2), 94–104.
18. Antoniadis, V.; Tsadilas, C.; Samaras, V.; Sgouras, J. Availability of heavy metals applied to soil through sewage sludge. In: *Trace Elements in the Environment: Biogeochemistry, Biotechnology, and Bioremediation;* Prasad, M.N.V., Sajwan, K.S., Naidu R., Eds.; CRC Taylor and Francis, LLC: Boca Raton, FL, 2006; 39–62.
19. Stephen, R.S. A critical review of the bioavailability and impacts of heavy metals in municipal solid waste composts compared to sewage sludge. Environ. Int. **2009**, *35* (1), 142–156.
20. Di Bonito, M. Sewage sludge in Europe and in the U.K.: Environmental impact and improved standards for recycling and recovery to land. In *Environmental Geochemistry: Site Characterization, Data Analysis and Case Histories;* Vivo, B.D., Belkin, H.E., Lima A., Eds.; Elsevier: Amsterdam, the Netherlands, 2008.
21. Adriano, D.C. *Trace Elements in Terrestrial Environments: Biogeochemistry, Bioavailability, and Risks of Metals*; CRC Press, Taylor and Francis Group: Boca Raton FL, 2006.
22. Carlson, C.; Adriano, D. Environmental impacts of coal combustion residues. J. Environ. Qual. **1993**, *22*, 227–242.
23. Bradl, H. *Heavy Metals in the Environment: Origin, Interaction and Remediation*; Academic Press: Amsterdam, Netherlands, 2005; Vol. 6.
24. Bullock, P.; Gregory, P.J. *Soils in the Urban Environment;* Blackwell Scientific Publications: Oxford, Boston, 1991.
25. Kabata-Pendias, A. *Trace Elements in Soils and Plants,* 4th Ed.; CRC Press, Inc.: Boca Raton, Florida, 2011.
26. McBride, M.B. *Environmental Chemistry of Soils;* Oxford University Press, Inc.: New York, 1994.
27. Hubert-Pfalzgraf, G.L. Niobium and tantalum: inorganic and coordination chemistry. In *Encyclopedia of Inorganic and Bioinorganic Chemistry*; Scott A.R., Ed.; Wiley Publishing, Inc., 2011.
28. Ure, A.M.; Bacon, J.R. Comprehensive analysis of soils and rocks by spark-source mass spectrometry. Analyst **1978,** *103* (1229), 807–822.
29. Govindaraju, K. Compilation of working values and sample description for 383 geostandards. Geostand. Geoanal. Res. **1994,** *18,* 1–158.
30. Yang, J.; Teng, Y; Wang, J.; Li, J. Vanadium uptake by alfalfa grown in V-Cd-contaminated soil by pot experiment. Biol. Trace Elem. Res. **2010,** *142* (3), 787–795.
31. Morrell, B.G.; Lepp, N.W.; Phipps, D.A. Vanadium uptake by higher plants: Some recent developments. Environ. Geo-chem. Health **1986,** *8* (1), 14–18.
32. *Eriksson, J. Concentrations of 61 Trace Elements in Sewage Sludge, Farmyard Manure, Mineral Fertilizers, Precipitation and in Oil and Crops; Swedish EPA: Stockholm, 2001.*
33. Oakes, T.W.; Shank, K.E. *Concentrations of Radionuclides and Selected Stable Elements in Fruits and Vegetables*; Oak Ridge National Laboratory: Oak Ridge, Tennessee, 1977.
34. Fritzen, M.B.; Souza, A.J.; Silva, TA; Souza, L.; Nome, R.A.; Fiedler, H.D. Distribution of hexavalent Cr species across the clay mineral surface-water interface. J. Colloid Interface Sci. **2006,** *296* (2), 465–471.
35. Shen-Yang, T.; Ke-An, L. The distribution of chromium(VI) species in solution as a function of pH and concentration. Talanta **1986,** *33* (9), 775–777.
36. Zayed, A.M.; Terry, N. Chromium in the environment: Factors affecting biological remediation. Plant Soil **2003,** *249* (1), 139–156.
37. Rai, D.; Eary, L.; Zachara, J. Environmental chemistry of chromium. Sci. Total Environ. **1989,** *86* (1–2), 15–23.
38. Bloomfield, C.; Pruden, G. The behaviour of Cr(VI) in soil under aerobic and an-aerobic conditions. Environ. Pollut., Ser. A **1980,** *23* (2), 103–114.
39. Richard, F.C.; Bourg, A.C. Aqueous geochemistry of chromium: A review. Water Res. **1991,** *25* (7), 807–816.

40. Alloway, B.J. *Heavy Metals in Soils;* Blackie Academic and Professional: London, 1995.
41. Avudainayagam, S.; Megharaj, M.; Owens, G.; Kookana, R.; Chittleborough, D.; Naidu, R. Chemistry of chromium in soils with emphasis on tannery waste sites. In *Reviews of Environmental Contamination and Toxicology*; SpringerVerlag: New York, 2003; Vol. 178, 53–91.
42. Gelencsér, A.; Kováts, N.; Turóczi, B.; RostásiI, Á.; Hoffer, A.; Imre, K.; Nyirő-Kósa, I.; Csákberényi-Malasics, D.; Toóth, Á.; Czitrovsky, A.; Nagy, A.; Nagy, S.; Ács, A.; Kovács, A.; Ferincz, Á.; Hartyáni, Z.; Pósfai, M. The red mud accident in Ajka (Hungary), characterization and potential health effects of fugitive dust. Environ. Sci. Technol. **2011,** *45* (4), 1608–1615.
43. Ruyters, S.; Mertens, J.; Vassilieva, E.; Dehandschutter, B.; Poffijn, A.; Smolders, E. The red mud accident in Ajka (Hungary), plant toxicity and trace metal bioavailability in red mud contaminated soil. Environ. Sci. Technol. **2011**, *45* (4), 1616–1622.
44. Kimbrough, D.E.; Cohen, Y.; Winer, A.M.; Creelman, L.; Mabuni, C.A. Critical assessment of chromium in the environment. Crit. Rev. Environ. Sci. Technol. **1999**, *29* (1), 1–46.
45. Nriagu, J.O.; Nieboer, E. *Chromium in the Natural and Human Environments*; Wiley: New York, 1988.
46. Pau, N.R.; Lawson, M.D. Molybdenum and tungsten: Their roles in biological processes. In *Metal Ions in Biological Systems: Molybdenum and Tungsten;* Sigel A., Sigel H., Eds.; Marcel Dekker Inc.: New York, 2002; Vol. 29, 31–74.
47. Stiefel, E.I. The biogeochemistry of molybdenum and tungsten. In: *Metal ions in biological systems: Molybdenum and Tungsten;* Sigel, A., Sigel, H., Eds.; Marcel Dekker Inc.: New York, 2002; Vol. 39, 1–29.
48. Goldberg, S.; Forster, H.; Godfrey, C. Molybdenum adsorption on oxides, clay minerals, and soils. Soil Sci. Soc. Am. J. **1996**, *60* (2), 425–432.
49. Zhang, P.; Sparks, D. Kinetics and mechanisms of molybdate adsorption/desorption at the goethite/water interface using pressure-jump relaxation. Soil Sci. Soc. Am. J. **1989**, *53,* 1028–1034.
50. Steinke, D.R.; Majak, W.; Sorensen, T.S.; Parvez, M. Chelation of molybdenum in *Medicago sativa* (alfalfa) grown on reclaimed mine tailings. J. Agric. Food Chem. **2008**, *56* (13), 5437–5442.
51. Wichard, T.; Mishra, B.; Myneni, S.C.B.; Bellenger, J.; Kraepiel, A.M.L. Storage and bioavailability of molybdenum in soils increased by organic matter complexation. Nature Geosci. **2009**, *2* (9), 625–629.
52. Lindsay, W.L. *Chemical Equilibria in Soils;* The BlackBurn Press: New Jersey, 2001.
53. Gupta, U. *Molybdenum in Agriculture;* University Press: New York, 1997.
54. Smedley, P.; Cooper, D.; Lapworth, D.J.; Ander, L. Molybdenum in British drinking water: A review of sources and occurrence and a reconnaissance survey of concentrations. Keyworth, Nottingham: Groundwater Resources Programme, British Geological Survey, 2008. Open Report OR/08/051.
55. Schroeder, H.A. A sensible look at air pollution by metals. Arch. Environ. Health **1970**, *21* (6), 798–806.
56. ATSDR (Agency for Toxic Substances and Disease Registry). Toxicological Profile for Tungsten. U.S. Department of Health and Human Services, 2005.
57. Yim, W. Heavy metal accumulation in estuarine sediments in a historical mining of Cornwall. Mar. Pollut. Bull. **1976**, *7* (8), 147–150.
58. Scheuhammer, A.; Norris, S. Review of the Environmental Impacts of Lead Shotshell Ammunition and Lead Fishing Weights in Canada. Occasional Papers No. 88. Quebec: Government of Canada, Canadian Wildlife Service, 1995.
59. Scheuhammer, A.M.; Norris, S.L. The ecotoxicology of lead shot and lead fishing weights. Ecotoxicology **1996**, *5* (5), 279–295.
60. USFWS. Migratory bird hunting: Approval of tungsten–nickel–iron shot as non-toxic for hunting waterfowl and coots. U.S. Fish and Wildlife Service. Fed. Regist. **2001** *66* (3), 737–742.

61. Szefer, P.; Nriagu, J.O. *Mineral Components in Foods;* CRC Press/Taylor and Francis: Boca Raton, 2007.
62. Welch, R.M.; Huffman, E.W.D. Vanadium and plant nutrition. Plant Physiol. **1973**, *52* (2), 183–185.
63. Nielsen, F. Vanadium. In *Trace Elements in Human and Animal Nutrition,* 5th Ed.; Mertz, W, Ed.; Academic Press: San Diego, California, 1988; 275–300.
64. Goldwaser, I.; Gefel, D.; Gershonov, E.; Fridkin, M.; Shechter, Y. Insulin-like effects of vanadium: Basic and clinical implications. J. Inorg. Biochem. **2000**, *80* (1–2), 21–25.
65. Tsiani, E.; Fantus, I. Vanadium compounds: Biological actions and potential as pharmacological agents. Trends Endocrinol. Metab. **1997**, *8* (2), 51–58.
66. Bishayee, A.; Chatterjee, M. Inhibitory effect of vanadium on rat liver carcinogenesis initiated with diethylnitrosamine and promoted by phenobarbital. Br. J. Cancer **1995**, *71* (6), 1214–1220.
67. Evangelou, A.M. Vanadium in cancer treatment. Crit. Rev. Oncol./Hematol. **2002**, *42* (3), 249–265.
68. Djordjevic, C. Antitumor activity of vanadium compounds. In *Metal Ions in Biological Systems*; Sigel, A., Sigel, H., Sigel, R., Eds.; 1995; Vol. 31, 595–616.
69. Rehder, D. Inorganic considerations on the function of vanadium in biological systems. In *Metal Ions in Biological Systems;* Sigel, A., Sigel, H., Sigel, R., Eds.; 1995; Vol. 31, 1–43.
70. Sabbioni, E.; Rade, J.; Bertolero, F. Relationships between iron and vanadium metabolism: The exchange of vanadium between transferrin and ferritin. J. Inorg. Biochem. **1980**, *12* (4), 307–315.
71. Ganguli, S.; Reuland, D.J.; Franklin, L.A.; Deakins, D.D.; Johnston, W.J.; Pasha, A. Effects of maternal vanadate treatment of fetal development. Life Sci. **1994**, *55* (16), 1267–1276.
72. Nriagu, J.O. *Vanadium in the Environment. Part 2, Health Effects;* John Wiley and Sons: New York, 1998.
73. IARC. Cobalt in Hard Metals and Cobalt Sulfate, Gallium Arsenide, Indium Phosphide and Vanadium Pentoxide. Vol. 86 of IARC Monographs on the Evaluation of Carcinogenic Risks to Humans. Lyon, France, 2006.
74. Liu, J.; Cui, H.; Liu, X.; Peng, X.; Deng, J.; Zuo, Z.; Cui, W.; Deng, Y.; Wang, K. Dietary high vanadium causes oxidative damage-induced renal and hepatic toxicity in broilers. Biol. Trace Elem. Res. **2011**, *145* (2), 189–200.
75. Al-Bayati, M.A.; Giri, S.N.; Raabe, O.G.; Rosenblatt, L.S.; Shifrine, M. Time and dose–response study of the effects of vanadate on rats: morphological and biochemical changes in organs. J. Environ. Pathol. Toxicol. Oncol. **1989**, *9* (5–6), 435–455.
76. Barker, AV.; Pilbeam, D.J. *Handbook of Plant Nutrition;* CRC Taylor and Francis: Boca Raton, FL, 2007.
77. Dieter, R. The coordination chemistry of vanadium as related to its biological functions. Coord. Chem. Rev. **1999**, *182* (1), 297–322.
78. USEPA. Superfund, emergency planning, and community right-to-know programs. Extremely hazardous substances and their threshold planning quantities. U.S. Environmental Protection Agency, 2009.
79. Edel, J.; Sabbioni, E.; Pietra, R.; Rossi, A.; Torre, M.; Rizzato, G.; Fraioli, P. Trace metal lung disease: In vitro interaction of hard metals with human lung and plasma components. Sci. Total Environ. **1990**, *95,* 107–117.
80. Moriyama, H.; Kobayashi, M.; Takada, T.; Shimizu, T.; Terada, M.; Narita, J.; Maruyama, M.; Watanabe, K.; Suzuki, E.; Gejyo, F. Two-dimensional analysis of elements and mononuclear cells in hard metal lung disease. Am. J. Respir. Crit. Care Med. **2007**, *176* (1), 70–77.
81. Sabbioni, E.; Minoia, C.; Pietra, R.; Mosconi, G.; Forni, A.; Scansetti, G. Metal determinations in biological specimens of diseased and non-diseased hard metalworkers. Sci. Total Environ. **1994**, *150* (1–3), 41–54.
82. Downs, W.L.; Scott, J.K.; Yuile, C.L.; Caruso, F.S.; Wong, L.C.K. The toxicity of niobium salts. Am. Ind. Hyg. Assoc. J. **1965**, *26* (4), 337–346.

83. Venugopal, D.; Luckey, T. Toxicity of group V metals and metalloids. In *Metal Toxicity in Mammals;* Plenum Press: New York, 1978; Vol. 2, 227–229.
84. Ramirez, G.; Rodil, S.; Arzate, H.; Muhl, S.; Olaya, J. Niobium based coatings for dental implants. Appl. Surf. Sci. **2011**, *257* (7), 2555–2559.
85. Johansson, C.B.; Hansson, H.A.; Albrektsson, T. Qualitative interfacial study between bone and tantalum, niobium or commercially pure titanium. Biomaterials **1990**, *11* (4), 277–280.
86. Matsuno, H.; Yokoyama, A.; Watari, F.; Uo, M.; Kawasaki, T. Biocompatibility and osteogenesis of refractory metal implants, titanium, hafnium, niobium, tantalum and rhenium. Biomaterials **2001**, *22* (11), 1253–1262.
87. Divine, K.; Goering, P. Tantalum. In *Elements and Their Compounds in the Environment Occurrence, Analysis and Biological Relevance;* Merian, E., Anke, M., Ihnat, M., Stoeppler, M., Eds.; Wiley-VCH: Weinheim, 2004; 1087–1097.
88. Rydzynski, K. Vanadium, niobium, and tantalum. In *Patty's Toxicology;* Bingham, E., Cohrssen, B., Powell, C.H., Eds.; Wiley InterScience: New York, 2001.
89. Goering, P.L.; Ziegler, T.L. Niobium (Nb) (columbium). In *Elements and Their Compounds in the Environment: Occurrence, Analysis and Biological Relevance;* Merian, E., Anke, M., Ih-nat, M., Stoeppler, M., Eds.; Wiley-VCH: Weinheim, 2004; 1039–1046.
90. Wennig, R.; Kirsch, N. Niobium. In *Handbook on Toxicity of Inorganic Compounds;* Seiler, H.G., Sigel, A., Eds.; Marcel Dekker: New York, 1988; 469–473.
91. Hopkins, J.L.L.; Ransome-Kuti, O.; Majaj, A.S. Improvement of impaired carbohydrate metabolism by chromium III in malnourished infants. Am. J. Clin. Nutr. **1968**, *21* (3), 203–211.
92. Mertz, W. Chromium occurrence and function in biological systems. Physiol. Rev. **1969**, *49* (2), 163–239.
93. Mertz, W. Chromium in human nutrition: A review. J. Nutr. **1993**, *123* (4), 626–633.
94. WHO. Chromium in drinking-water. World Health Organization: Geneva, Switzerland, 2003, WHO/SDE/ WSH/03.04/04.
95. Standeven, A.M.; Wetterhahn, K.E. Chromium(VI) toxicity: Uptake, reduction, and DNA damage. Int. J. Toxicol. **1989**, *8* (7), 1275–1283.
96. *Carson, B.L.; Ellis, H.V.; McCann, J.L. Toxicology and Biological Monitoring of Metals in Humans, Including Feasibility and Need, 1st Ed.; CRC Press: Chelsea, Michigan, 1986.*
97. ATSDR. *Toxicological Profile for Chromium;* U.S. Department of Health and Human Services: Atlanta, 2008.
98. Cohen, D.M.; Costa, M. Chromium. In *Environmental Toxicants: Human Exposures and Their Health Effects*, 2nd Ed.; Lippmann, M., Ed.; Wiley-Blackwell, Hoboken, New Jersey, 2005; 173–191.
99. Bianchi, V.; Celotti, L.; Lanfranchi, G.; Majone, F.; Marin, G.; Montaldi, A.; Sponza, G.; Tamino, G.; Venier, P.; Zantedeschi, A.; Levis, A.G. Genetic effects of chromium compounds. Mutat. Res., Genet. Toxicol. **1983**, *117* (3–4), 279–300.
100. Katz, S.A.; Salem, H. The toxicology of chromium with respect to its chemical speciation: A review. J. Appl. Toxicol. **1993**, *13* (3), 217–224.
101. Baruthio, F. Toxic effects of chromium and its compounds. Biol. Trace Elem. Res. **1992**, *32* (1–3), 145–153.
102. Cervantes, C.; Campos-García, J.; Devars, S.; Gutiérrez-Corona, F.; Loza-Tavera, H.; Torres-Guzmán, J.C.; Moreno-Sánchez, R. Interactions of chromium with microorganisms and plants. FEMS Microbiol. Rev. **2001**, *25* (3), 335–347.
103. Hayat, S.; Khalique, G.; Irfan, M.; Wani, A.S.; Tripathi, B.N.; Ahmad, A. Physiological changes induced by chromium stress in plants: An overview. Protoplasma **2011**, 1–13.
104. Shanker, A.K.; Cervantes, C.; Loza-Tavera H.; Avudain-ayagam S. Chromium toxicity in plants. Environ. Int. **2005**, *31* (5), 739–753.
105. Burgess, B.K.; Lowe, D.J. Mechanism of molybdenum nitrogenase. Chem. Rev. **1996**, *96* (7), 2983–3012.

106. Kaiser, B.N.; Gridley, K.L.; Ngaire Brady, J.; Phillips, T.; Tyerman, S.D. The role of molybdenum in agricultural plant production. Ann. Bot. **2005**, *96* (5), 745–754.
107. Schwarz, G.; Mendel, R.R.; Ribbe, MW. Molybdenum cofactors, enzymes and pathways. Nature **2009**, *460* (7257), 839–847.
108. Kisker, C.; Schindelin, H.; Rees, D.C. Molybdenum-cofactor-containing enzymes: Structure and mechanism. Annu. Rev. Biochem. **1997**, *66* (1), 233–267.
109. Konhauser, K.O.; Powell, M.A.; Fyfe, W.S.; Longstaffe, F.J.; Tripathy, S. Trace element chemistry of major rivers in Orissa State, India. Environ. Geol. **1997**, *29* (1–2), 132–141.
110. Sardesai, V.M. Molybdenum: An essential trace element. Nutr. Clin. Pract. **1993**, *8* (6), 277–281.
111. Reiss, J.; Christensen, E.; Dorche, C. Molybdenum cofactor deficiency: First prenatal genetic analysis. Prenatal Diagn. **1999**, *19* (4), 386–388.
112. Anke, M.; Groppl, B.; Kronemann, H.; Grun, M. Molybdenum supply and status in animals and human beings. Nutr. Res. **1985**, *S1,* 180–186.
113. Chatterjee, C.; Nautiyal, N.; Agarwala, S.C. Metabolic changes in mustard plants associated with molybdenum deficiency. New Phytol. **1985**, *100* (4), 511–518.
114. Chatterjee, C.; Nautiyal, N. Molybdenum stress affects viability and vigor of wheat seeds. J. Plant Nutr. **2001**, *24* (9), 1377.
115. Johnson, J.L.; Waud, W.R.; Rajagopalan, K.V.; Duran, M.; Beemer, F.A.; Wadman, S.K. Inborn errors of molybdenum metabolism: Combined deficiencies of sulfite oxidase and xanthine dehydrogenase in a patient lacking the molybdenum cofactor. Proc. Natl. Acad. Sci. U. S. A. **1980**, *77* (6), 3715–3719.
116. Frøslie, A. Problems on deficiency and excess of minerals in animal nutrition. In *Geomedicine,* 1st Ed.; Låg, J., Eds.; CRC Press: Boca Raton, Florida, **1990**; 37–60.
117. Erdman, J.A.; Ebens, R.J.; Case, A.A. Molybdenosis: A potential problem in ruminants grazing on coal mine spoils. J. Range Manage. **1978**, *31* (1), 34–36.
118. Ferguson, W.S.; Lewis, A.H.; Watson, S.J. Action of molybdenum in nutrition of milking cattle. Nature **1938**, *141* (3569), 553–553.
119. Seiden, A.I.; Berg, NP.; Soderbergh, A.; Bergstrom, B.E.O. Occupational molybdenum exposure and a gouty electrician. Occup. Med. **2005**, *55* (2), 145–148.
120. Council, N.R. *Recommended Dietary Allowances,* 10th Ed.; National Academy Press: Washington, DC, 1989.
121. WHO. Molybdenum in drinking-water. World Health Organization: Geneva, Switzerland, 2011. WHO/SDE/ WSH/03.04/11/Rev/1.
122. Seiler, R.L.; Stollenwerk, K.G.; Garbarino, J.R. Factors controlling tungsten concentrations in ground water, Carson Desert, Nevada. Appl. Geochem. **2000**, *20* (2), 423–441.
123. Johnson, J.L.; Rajagopalan, K.V.; Cohen, H.J. Molecular basis of the biological function of molybdenum. Effect of tungsten on xanthine oxidase and sulfite oxidase in the rat. J. Biol. Chem. **1974**, *249* (3), 859–866.
124. Higgins, E.S.; Richert, D.A.; Westerfeld, WW. Molybdenum deficiency and tungstate inhibition studies. J. Nutr. **1956**, *59* (4), 539–559.
125. Callis, G.E.; Wentworth, R.A. Tungsten vs. molybdenum in models for biological systems. Bioinorg. Chem. **1977**, *7* (1), 57–70.
126. Kletzin, A.; Adams, MW. Tungsten in biological systems. FEMS Microbiol. Rev. **1996**, *18* (1), 5–63.
127. Gatti, A.M.; Montanari, S.; Monari, E.; Gambarelli, A.; Capitani, F.; Parisini, B. Detection of micro- and nano-sized biocompatible particles in the blood. J. Mater. Sci. Mater. Med. **2004**, *15* (4), 469–472.
128. Fischbein, A.; Luo, J.C.; Solomon, S.J.; Horowitz, S.; Hailoo, W.; Miller, A. Clinical findings among hard metal workers. Br. J. Ind. Med. **1992**, *49* (1), 17–24.

129. PECH (Panel on the trace element geochemistry of coal resource development related to health). Trace Element Geochemistry of Coal Resource Development Related to Health, National Academies Press, 1980.
130. ATSDR. *Toxicological Profile for Vanadium;* U.S. Department of Health and Human Services: Atlanta, 2009.
131. USEPA. *Guidelines for Water Reuse;* U.S. Environmental Protection Agency: Washington, DC, 2004. EPA/625/R-04/108.
132. NIOSH. *NIOSH Pocket Guide to Chemical Hazards*; National Institute for Occupational Safety and Health: Cincinnati, Ohio, 2007. DHHS (NIOSH) Publication No. 2005–149.
133. USGS. Mineral Commodity Summaries, 2011. U.S. Geological Survey, 2011.
134. Hedrick, B.J. The global rare-earth cycle. J. Alloys Compd. **1995**, *225* (1–2), 609–618.
135. Reimann, C.; Caritat, P.D. *Chemical Elements in the Environment: Factsheets for the Geochemist and Environmental Scientist*; Springer-Verlag GmbH and Co: Berlin and Heidelberg, 1998.
136. Takeda, A.; Kimura, K.; Yamasaki, S. Analysis of 57 elements in Japanese soils, with special reference to soil group and agricultural use. Geoderma **2004**, *119* (3–4), 291–307.
137. Licht, O. *Geoquimica de solo do Estado do Parana* (in Spanish); Mineropar: Curitiba, Parana, 2005; Vol. 1/2.
138. Burt, R.; Wilson, M.A.; Mays, M.D.; Lee, CW. Major and trace elements of selected pedons in the USA. J. Environ. Qual. **2003**, *52* (6), 2109–2121.
139. Shacklette, HT.; Boerngen, J.G. Element concentrations in soils and other surficial materials of the conterminous United States. Geol. Surv. Prof. Paper. **1984**, *1270* (105).
140. Kabata-Pendias, A.; Mukherjee, A. *Trace Elements from Soil to Human;* Springer: Berlin-Heidelberg, 2007.
141. Gaillardet, J.; Viers, J.; Dupre, B. Trace elements in river waters. In: *Treatise on Geochemistry;* Elsevier: Oxford, 2003; 225–272.
142. Li, Y. A brief discussion on the mean oceanic residence time of elements. Geochim. Cosmochim. Acta. **1982**, *46* (12), 2671–2675.
143. Nozaki, Y. A fresh look at element distribution in the North Pacific Ocean. EOS, Trans. Am. Geophys. Union **1997**, *78* (21), 221.
144. Sohrin, Y.; Isshiki, K.; Kuwamoto, T.; Nakayama, E. Tungsten in north pacific waters. Mar Chem. **1987**, *22* (1), 95–103.
145. Sparks, D.L. *Environmental Soil Chemistry;* Academic Press: San Diego, 2003.

III

Basic Environmental Processes

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

39

Adsorption

Puangrat Kajitvichyanukul
and Jirapat Ananpattarachai

Introduction

Water is essential for life. Currently, quality of water becomes a major problem in many areas owing to pollution emission from industrial, agricultural, and domestic activities to the water bodies. These activities generate wastewater, which contains both inorganic and organic pollutants. Some of the common pollutants are phenols, dyes, detergents, insecticides, pesticides, and heavy metals. These pollutants are often toxic and cause adverse effects on human life. To avoid pollution of natural water bodies, treating wastewater from the originated source by removing pollutants before being discharging is necessary. Various treatment techniques and processes such as coagulation, membrane process, adsorption, dialysis, foam flotation, osmosis, and biological methods have been used to remove the pollutants from contaminated water. Among all the approaches proposed, adsorption is one of the most popular methods and is currently considered as an effective, efficient, and economic method for water purification.

Adsorption is a well-known equilibrium separation process and an effective method for water and wastewater treatment applications.[1–4] It is the process in which molecules accumulate in the interfacial layer, but desorption denotes the converse process. The fundamental concept in adsorption science is the equilibrium relation between the quantity of the adsorbed material and the pressure or concentration in the bulk fluid phase at constant temperature. The material adsorbed on the surface of "adsorbent" is defined as the "adsorbate." The penetration by the adsorbate molecules into the bulk solid phase is determined as "absorption." The terms "sorption," "sorbent," "sorbate," and "sorptive" are also used to denote both adsorption and absorption, when both occur simultaneously or cannot be distinguished.[1]

Recently, many works related to contaminant removal by adsorption process have been reviewed.[5–7] Several types of adsorbents applied in wastewater treatment are activated carbon, zeolite, chitin or chitosan, and various agricultural wastes such as wood, peat, saw dust, etc. These adsorbents can take a broad range of chemical forms and different geometrical surface structures and properties. This is reflected in the range of their applications in industry for both water and wastewater treatment. Compared with alternative technologies, adsorption is attractive for its relative simplicity of design, operation, and scale-up; high capacity and favorable rate; insensitivity to toxic substances; ease of regeneration; and low cost. Additionally, it avoids using toxic solvents and minimizes degradation.[6]

This review highlights and provides an overview of adsorption theory, kinetics and mechanisms, and recent applications of various adsorbents for contaminant removal from water and wastewater. The main aim of this review is to provide a summary of recent information concerning adsorption process. In this entry, an extensive list of adsorbents from many research papers has been provided. It is strongly encouraged to refer to these original papers for more information on experimental conditions.

Adsorption Theory

In solid–liquid interface, adsorption of a species on a solid surface follows three steps:[7]

1. Transport of the adsorbate from the bulk to the external surface of the adsorbent.
2. Passage through the liquid film attached to the solid surface.
3. Interactions with the surface atoms of the solid leading to chemisorption. (strong adsorbate–adsorbent interactions equivalent to covalent bond formation) or weak adsorption (weak adsorbate–adsorbent interactions, very similar to van der Waals forces).

If step 1 is the slowest, the adsorption will be a transport-limited process. This step is usually the rate-limiting process in systems that are characterized by poor mixing, dilute concentration of adsorbate, small particle size of the adsorbent, etc.

When step 2 is the rate-determining slowest step, the physical process of diffusion through the liquid film influences the outcome of the process, and the efficiency of the solid as an adsorbent can hardly be improved.

When step 3 is the slowest, the adsorption is controlled by a chemical process, and the efficiency of the adsorbent can be influenced by suitably controlling the interactions.[7]

For porous solids, when the adsorbate passes through the liquid film attached to the external surface, it will slowly diffuse into the pores and get trapped or adsorbed. The pore diffusion plays an important role when the adsorbate is present in higher concentration, the adsorbent is made of large particles, and good mixing is ensured.[7–9]

Adsorption capacity is highly dependent on several factors such as properties of the adsorbent (porosity, surface area, particle size) and adsorbate (structure, water solubility, ionic charge, functional groups, pKa, polarity, functionality, molecular weight, and size); solution conditions (solvent, pH, temperature, ionic strength, solute concentration, and competition between solutes); interactions at the solid–liquid interface; and type of experimental setup. Generally, the adsorption capacity of a sorbent for a solute increases with increase in liquid-phase concentration. Physicochemical properties of the adsorbent also play a major role in adsorption. It must have high surface area, particularly internal surface area, high internal volume, and a good pore size distribution. The chemical properties of the adsorbent include degree of ionization at the surface and types of functional groups present. The adsorbent should also have good mechanical properties such as strength and resistance to destruction.

Temperature and pressure have an effect on increasing or decreasing of adsorption capacity. Under low-temperature conditions, the adsorption increases in the forward condition and liberates heat as it is exothermic in nature. With the increasing of pressure, adsorption increases up to a certain extent until saturation level is achieved. After it reaches the equilibrium level, no more adsorption takes place no matter how high the pressure applied.

The coverage of the adsorbent surface by the adsorbate leads to the adsorption process. It assumes that the surface consists of “sites” onto which the adsorbate can adsorb. Accordingly, the adsorption between the adsorbate molecule and the adsorbent surface can be depicted with the chemical reaction below:

$$\text{Molecule} + \text{surface site} \leftrightarrow \text{adsorbed molecule} + \text{heat}(\Delta H) \tag{1}$$

According to the above reaction, adsorption characteristic can be expressed in thermodynamic parameters such as ΔG, ΔH, and ΔS. A negative ΔG value stands for the adsorption to take place. Change in enthalpy (ΔH) gives an indication of the bonding strength. The higher the value of heat of adsorption, the weaker the bond between adsorbate and adsorbent. The sign of AS indicates the direction: for adsorption, ($+\Delta S$), and for desorption, ($-\Delta S$).

The change in Gibbs free energy is given by the following expression:

$$\Delta G^{o} = -RT \ln K_L \tag{2}$$

where R is the universal gas constant, T is the absolute temperature, and K_l (L/mg) is the affinity constant of Langmuir model. Negative ΔG values confirm the feasibility of the adsorption, and its absolute values measure the adsorption driving force. Negative or positive values of ΔG indicate the exothermic or endothermic nature of adsorption, respectively. Positive ΔS values reveal a random organization of the adsorbate at the solid/solution interface. The sorption entropy can be calculated from the Gibbs–Helmholtz equation:

$$\Delta G = \Delta H - T\Delta S \tag{3}$$

For a reaction or process to be spontaneous, there must be decreases in free energy of the system. ΔG of the system must have a negative value. During the adsorption, randomness of the molecule decreases, so that ΔS is negative. The above equation can be rewritten as follows:

$$\Delta G = \Delta H + T\Delta S \tag{4}$$

Therefore, for a reaction to be spontaneous, ΔH has to be negative, and $|\Delta H| > |T\Delta S|$.

The adsorption process occurs when adsorbate is adsorbed on adsorbent. Heat energy developing during the adsorption process is released as it is an exothermic process.

During adsorption, forces of attraction play an important role between adsorbate and adsorbent. Van der Waals forces of attraction are weak forces, while forces from chemical bonding are strong forces. Accordingly, the adsorption can be classified into two types: physical adsorption or chemical adsorption.

Physisorption or physical adsorption is a type of adsorption in which the adsorbate adheres to the surface through van der Waals (weak intermolecular) interactions. Physical adsorption takes place with formation of a multilayer of adsorbate on adsorbent. The chemical identity of the adsorbate remains intact without breaking of the covalent structure of the adsorbate. To be a spontaneous thermodynamic process, ΔG has to be a negative value. As the translational degrees of freedom of the adsorbate are lost upon deposition onto the adsorbent, ΔS is negative for the process, and ΔH for physical adsorption must be exothermic. The energy released upon accommodation to the surface is of the same order of magnitude as an enthalpy of condensation (in the order of 20 kJ/mol). This process occurs at low temperature, below the boiling point of adsorbate. As the temperature increases, the process of physical adsorption decreases, as shown in Figure 1.

Chemisorption is a type of adsorption whereby a molecule adheres to a surface through the formation of a chemical bond, as opposed to the van der Waals forces. Chemical adsorption takes place when adsorbate attaches on the adsorbent surface with formation of a unilayer by chemical bonding. This interaction is much stronger and has higher enthalpy of adsorption, ΔH (200–400 kJ/mol), than physical adsorption. As the adsorbates can interact with each other when they lie upon the surface, the energy of adsorption relies on the extent to which the available surface is covered with adsorbate molecules. Thus, the chemical bonds in chemical adsorption may be stronger than the bonds internal to the free adsorbate, resulting in the dissociation of the adsorbate upon adsorption. This type of adsorption can

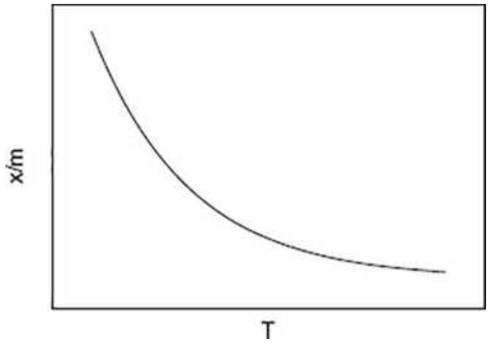


FIGURE 1 Relation of physical adsorption with temperature.

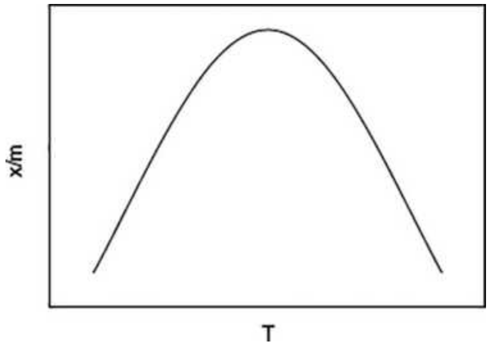


FIGURE 2 Relation of chemical adsorption with temperature.

take place at all temperatures. Chemical adsorption first increases and then decreases with the increases in temperature, as shown in Figure 2.

Adsorption Isotherm

Adsorption isotherms are defined by the adsorbate–adsorbent interactions. Generally, it is constructed by measuring the concentration of the adsorbate in the medium before and after adsorption, at a fixed temperature. This is the practical way to investigate the interaction between the adsorbate and the surface of the adsorbent and to obtain information about the structure of the adsorbed layer. Four classes of adsorption isotherms [S type with an initial con-vexing to the concentration axis, Langmuir (L) type with an initial concavity to the concentration axis, H type with an intercept on the ordinate, and C type with an initial linear portion] and subgroups have been defined according to their configuration,[10–13] as shown in Figure 3. The shape of isotherm provides qualitative information on the nature of solute-surface interaction. The Langmuir class (L) is widespread in the case of adsorption of many organic contaminants from water. The organic contaminant adsorbs parallel to the surface, and no strong competition exists between the adsorbate and the solvent to occupy the adsorption sites. However, the H class (high affinity) results from extremely strong adsorption at very low concentrations, giving rise to an apparent intercept on the ordinate.[12]

In the adsorption process, when a mass of adsorbent and a waste stream of adsorbate are in contact for a sufficiently long time, equilibrium between the amount of pollutant adsorbed and the amount remaining in solution will develop. Under equilibrium conditions, the amount of material adsorbed onto the media can be calculated using the mass balance of Eq. 5:

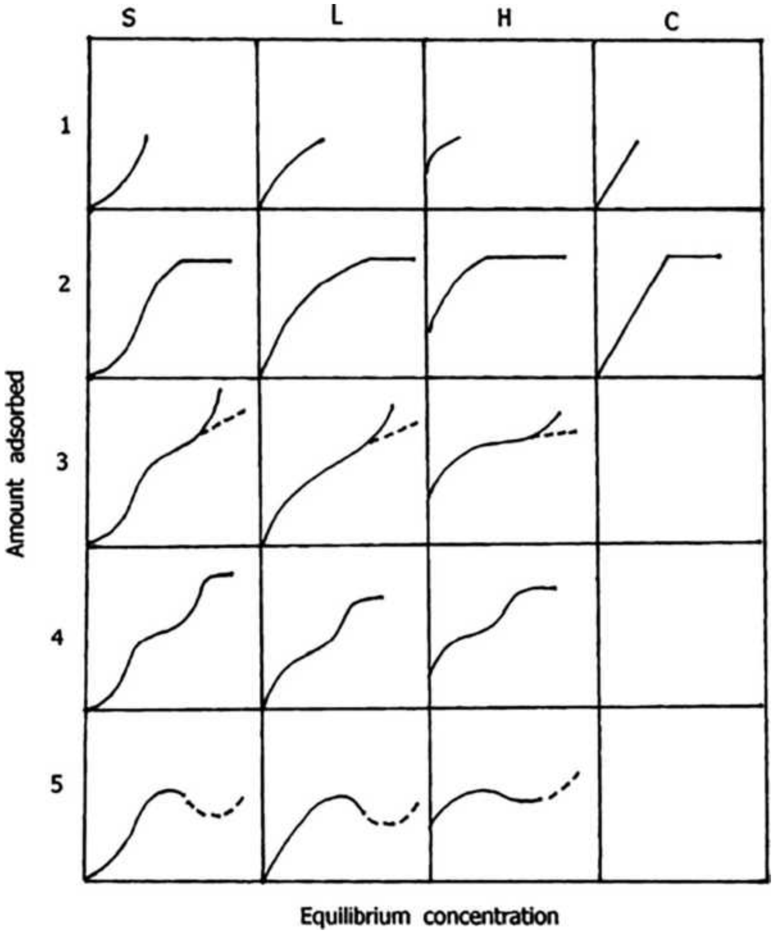


FIGURE 3 Four classes of adsorption isotherms (S, L, H, C). Source: Giles et al.[11]

$$q_e = \frac{(C_0 - C_e)V}{W} \tag{5}$$

where q_e (mg pollutant/g adsorbent) is the mass of pollutant per mass of adsorbent used, C_0 is the liquid-phase concentration of pollutant in solution at the beginning, C_e is the concentration of the pollutant in solution after equilibrium has been reached, V is the volume of the solution, and W is the mass of dry adsorbent used.

Adsorption isotherms are related to the amount of adsorbate on the adsorbent, with its concentration. Several models describing process of adsorption are Freundlich isotherm, Langmuir isotherm, Brunauer–Emmett–Teller (BET) isotherm, etc., with details described below.

Langmuir Isotherm

The Langmuir equation initially derived from kinetic studies was based on the assumption that on the adsorbent surface, there is a definite and energetically equivalent number of adsorption sites. Consequently, the hypotheses for this isotherm are uniform adsorption energies along the homogeneous adsorbent surface, equal solute affinity in all the adsorption sites, no interaction among adsorbed molecules, a single adsorption mechanism, and formation of a monolayer on the free surface.[14]

This isotherm is concerned with the monolayer (single-layer) coverage of the solid surface by the adsorbate. The bonding to the adsorption sites can be either chemical or physical, but it must be sufficiently strong to prevent displacement of adsorbed molecules along the surface.

The Langmuir isotherm is useful in determining the interactions between the adsorbate and the adsorbent. Adsorption data for a wide range of adsorbate concentrations are most conveniently described by the Langmuir isotherm:

$$q_e = \frac{QK_LC_e}{1 + K_LC_e} \tag{6}$$

where Q (mg/g) is the maximum amount of the adsorbate per unit weight of the adsorbent to form a complete monolayer on the surface, whereas K_L (L/mg) is the Langmuir constant related to the affinity of the binding sites.

The essential characteristics of Langmuir isotherm can be expressed by a dimensionless constant called separation factor or equilibrium parameter, R_L, defined by Weber and Chakkravorti[15] as follows:

$$R_L = \frac{1}{1 + K_LC_0} \tag{7}$$

The parameter R_L indicates the shape of isotherm as shown in Table 1.

The limitations of this isotherm come from the fact that several types of adsorption sites exist, the adsorption mechanism is not the same for the first and for the last molecules adsorbed, and models based on the monomolecular surface layer without interactions could be not realistic. Consequently, many other isotherms such as Freundlich, BET, Temkin, and Dubinin–Radushkevich isotherms have been proposed to explain adsorption behavior.

The Freundlich Isotherm

The Freundlich isotherm is an empirical equation that is more suitable than the Langmuir isotherm when the amount of adsorbent in contact with the solid surface is relatively low. In addition, if the binding energy changes continuously from site to site on solid surfaces, the expression used for the Freundlich isotherm is much more applicable to explaining the adsorption behavior. It is also generally used in nonideal systems with highly heterogeneous surfaces or surfaces supporting sites of varied affinities. The assumption for this isotherm is that the stronger binding sites are occupied first and that the binding strength decreases with an increasing degree of site occupation.[16] It does not imply the formation of a monolayer and frequently gives good interpretation of data over a restricted concentration range.[12] The Freundlich isotherm is expressed as follows:

$$q_e = K_FC_e^{1/n} \tag{8}$$

where *qe* (mg/g) is the equilibrium value for removal of adsorbate per unit weight of adsorbent, *Ce* (mg/L) is the equilibrium concentration of metal ion in solution, and K_F and n are Freundlich isotherm

TABLE 1 Relation between the Value of RL and the Type of Isotherm

Value of Rl	Type of Isotherm
$R_L > 1$	Unfavorable
$R_L = 1$	Linear
$0 < R_L < 1$	Favorable
$R_L = 0$	Irreversible

constants that are related to the adsorption capacity (or the bonding energy) and intensity of the sorbent, respectively. K_F can be defined as the adsorption or distribution coefficient and represents the ad-sorbent onto adsorbate for a unit equilibrium concentration. A value for $1/n<1$ indicates a normal Langmuir isotherm, while $1/n > 1$ is indicative of cooperative adsorption:[17]

$$\log q_e = \left(\frac{1}{n}\right)\log C_e + \log K_F \tag{9}$$

Experimentally it was determined that extent of adsorption varies directly with pressure until saturation pressure Ps is reached. Beyond that point, the rate of adsorption saturates even after applying higher pressure. Thus, Freundlich adsorption isotherm fails at higher pressure.

BET Isotherm

This isotherm is widely applied in the gas–solid equilibrium systems. It has a theoretical background based on multilayer adsorption, which is the true picture of physical adsorption. The BET isotherm was derived from computer-based calculations as the nonlinear isotherm modeling, usually based on algorithms dealing with error distribution.[18]

Under the high pressure, the BET theory is more feasible than the Langmuir adsorption, which is applicable only under the conditions of low pressure. With high pressure and low temperature, thermal energy of gaseous molecules decreases, and more and more gaseous molecules would be available per unit surface area. Consequently, the multilayer adsorption would occur on the adsorbent surface. This multilayer formation is explained by BET theory. The BET equation is given as

$$V_{total} = \frac{V_{mono}C\left(\frac{P}{P_0}\right)}{\left(1-\frac{P}{P_0}\right)\left(1+C\left(\frac{P}{P_0}\right)-\frac{P}{P_0}\right)} \tag{10}$$

Another form of the BET equation is

$$\frac{P}{V_{total}\left(P-P_0\right)} = \frac{1}{V_{mono}C} + \frac{c-1}{V_{mono}C}\left(\frac{P}{P_0}\right) \tag{11}$$

where V_{mono} is the adsorbed volume of gas at high-pressure conditions so as to cover the surface with a unilayer of gaseous molecule.

Temkin Isotherm

The Temkin isotherm is applicable when the behavior of an adsorption system occurs on heterogeneous surfaces. It is expressed by the following equation.[19,20]

$$q_e = \frac{RT}{b_T} ln\left(K_T C_e\right) \tag{12}$$

The linear form of the Temkin isotherm is represented by the following equations:

$$q_e = A + B \ln C_e \tag{13}$$

$$A = \frac{RT}{b_T}\ln K_T \tag{14}$$

$$B = \frac{RT}{b_T} \tag{15}$$

where R is the gas constant (8.341 J mol^{-1} K^{-1}), T is the absolute temperature (K), and A and B represent isotherm constants. K_T is the equilibrium binding constant (L/g); b_T is related to the heat of adsorption (J/mol).

Dubinin–Radushkevich Isotherm

The Dubinin–Radushkevich isotherm is applicable to the adsorption mechanism based on the potential theory assuming a heterogeneous surface, and its linearized form is given in the following equations:[20,21]

$$\ln q_e = \ln q_m - K\varepsilon^2 \tag{16}$$

$$\varepsilon = RT\ \ln\left(1 + \frac{1}{Ce}\right) \tag{17}$$

where K, a constant, is related to the adsorption energy and e (kJ/mol) is used to estimate the type of adsorption process (chemical or physical adsorption) and calculated from the following equation:

$$\varepsilon = \left(-2K^{-1/2}\right) \tag{18}$$

If the ε values of the isotherm fall between 8 and 16 kJ/ mol, the adsorption reaction can be explained by an ion-exchange mechanism, whereas if its value is less than 8 kJ/mol, it indicates that the adsorption process has a physical nature.[22]

Redlich–Peterson Isotherm

This isotherm is widely used as a compromise between the Langmuir and Freundlich systems.[23] In this equation, three parameters have been used to incorporate the advantageous significance of both models. The model can be represented as follows:[24]

$$\boldsymbol{q_e = \frac{K_{RP}C_e}{1 + (\alpha C_e)^{\beta}}} \tag{19}$$

where K_{RP} (L/g), α (L/mg), and β are Redlich–Peterson (R-P) isotherm constants, whereas β is the exponent that lies between 0 and 1. The R–P model has two limiting cases. When $\beta = 1$, the Langmuir equation results, whereas when $\beta = 0$, the R–P equation transforms to Henry's law equation.

Adsorption Kinetics

Kinetics is the study of the rates of chemical processes by monitoring the experimental conditions that influence the speed of a chemical reaction and help attain equilibrium in a reasonable length of time. The possible mechanism of adsorption and the different transition states of the final adsorbate–adsorbent complex are normally studied, and the appropriate mathematical models to describe the interactions are applied to the reaction.

The adsorption kinetics is useful in understanding the complex dynamics of the adsorption process. Generally, four consecutive steps describing the occurrence of adsorbate transport are as follows[6]

1. Transport of the adsorbate from the bulk solution to the boundary layer surrounding the adsorbent particles
2. Transport of solute across the boundary layer

3. Intraparticle solute diffusion into the pores
4. Adsorption and desorption of adsorbate

One of the above steps, or a combination of them, controls the adsorption mechanism. Factors influencing the rate-limiting step include characteristics of the adsorbent, adsorbate, and solution, for instance, adsorbent particle size, adsorbate concentration, degree of mixing, affinity between adsorbate and adsorbent, and diffusion coefficients of the adsorbate.

Adsorption kinetics can be determined by the following stages: diffusion of molecules from the bulk phase toward the interface space or external diffusion, diffusion of molecules inside the pores or internal diffusion, diffusion of molecules in the surface phase or surface diffusion, and adsorption–desorption elementary processes.

The most popular models used to establish the controlling adsorption mechanism can be grouped as follows:[25]

1. Those assuming that the controlling step is mass transfer (homogeneous surface diffusion, pore diffusion, and heterogeneous diffusion models).
2. Those assuming that adsorption is governed by surface phenomena.

The models involved in an adsorption kinetics study are the Lagergren pseudo–first-order and pseudo–second-order model, the Elovich equation, intraparticle diffusion, and liquid film diffusion.

Lagergren Pseudo–First-Order Model

This equation describing the rate of adsorption in the liquid-phase systems is the most used equation, particularly for pseudo–first-order kinetics.[26] The linearized Lagergren equation considers a reversible equilibrium of organic molecules between a liquid and a solid phase. Many different adsorption situations can be described by pseudo-first–order kinetics, including the following: 1) systems close to equilibrium; 2) systems with time-independent solute concentration or linear equilibrium adsorption isotherm; and 3) special cases of more complex systems.[6]

The Lagergren equation as pseudo–first-order kinetics is described as follows:[7]

$$\frac{dq_t}{dt} = k_1\left(q_e - q_t\right) \tag{20}$$

where k_1 (min^{-1}) is the pseudo–first-order adsorption rate coefficient. The integrated form of this equation for the boundary conditions of $t = 0$, $q_t = 0$ and $t = t$, $q_t = q_t$ is

$$\ln\left(q_e - q_t\right) = \ln q_e - k_1 t \tag{21}$$

where q_e and q_t are the values of amount adsorbed per unit mass at equilibrium and at any time t. The values of k_1 can be obtained from the slope of the linear plot of ln *(qe– qt)* vs. t.

McKay et al.[27] reported that the Lagergren pseudo–first-order model is found suitable only for the initial 20 to 30 min of interaction and not fit for the whole range of contact time. As the value of k_1 depends on the initial concentration of the adsorbate, it usually decreases with the increasing initial adsorbate concentration in the bulk phase.[28–31]

Pseudo–Second-Order Model

The pseudo–second-order equation has been explained as a special kind of Langmuir kinetics with the following assumptions: 1) the adsorbate concentration is constant in time; and 2) the total number of binding sites depends on the amount of adsorbate adsorbed at equilibrium.[32]

The pseudo–second-order equation based on equilibrium adsorption is expressed as

$$\frac{dq_t}{dt} = k_2\left(q_e - q_t\right)^2 \tag{22}$$

The initial adsorption rate of a second-order process as $t \to 0$ is defined as

$$\frac{dq}{\left(q_e - q_t\right)^2} = k_2 dt \tag{23}$$

The integrating equation with respect to boundary conditions $q_t = 0$ at $t = 0$ and $q = q_e$ at $t = t$ is

$$\frac{t}{q} = \frac{1}{k_2 q_e^2} + \frac{1}{q_e} t \tag{24}$$

where k_2 (g/mg h) is the rate constant of second-order adsorption. It depends on the applied operating conditions, namely, initial metal concentration, pH of solution, temperature and agitation rate, etc.[33–34] The integral form of the model predicts that the ratio of the time/adsorbed amount should be a linear function of time.[35] The linear plot of t/q_t vs. t gives $1/q_e$ as the slope and $l/k_2 q_e^2$ as the intercept. This procedure tends to predict the behavior over the whole range of adsorption.

Several research works[30–31,34,36] reported that the rate coefficient, k_2, decreases with the increasing initial adsorbate concentration in the bulk phase. The higher the initial concentration of adsorbate, the longer the time required to reach an equilibrium and, consequently, the k_2 value decreases.

The second-order rate constants were used to calculate the initial sorption rate given by

$$h = k_2 q_e^2 \tag{25}$$

where h is the initial adsorption rate, q_e is the adsorption capacity, and k_2 is the pseudo–second-order rate coefficient. The value of k_2 can be determined experimentally from the slope and intercept of a plot of t/q_t against t.

It has been reported[24] that the initial adsorption rate, h, was found to increase with initial concentration; however, the value started to decrease when the high initial concentration was applied to the system. The possible reason might be that too-high solute concentrations would slow down the adsorption process. The h value could also be influenced by the characteristics of the adsorbent and adsorbate.

One of the advantages of the pseudo–second-order equation for estimating the qe values is its small sensitivity for the influence of the random experimental errors.[7]

Elovich Equation

The Elovich equation is one of the most widely used models for describing chemical adsorption. It assumes that the actual solid surfaces are energetically heterogeneous and that neither desorption nor interactions between the adsorbed species could substantially affect the kinetics of adsorption at low surface coverage.[7] The Elovich equation is also restricted to the initial sorption stages, when the system is relatively far from equilibrium.

The Elovich equation can be expressed as follows:[37]

$$\frac{dq_t}{dt} = \alpha \exp\left(-\beta q_t\right) \tag{26}$$

Where α and β represent the initial adsorption rate (g/mg min^2) and the desorption coefficient (mg/g min), respectively. Assuming $\alpha\beta_{qt} >> 1$, and $q_t = 0$ at $t = 0$ and $qt = q_t$ at $t = t$, the linear form of the above equation id given by [38–40]

$$q_t = \beta \ln(\alpha\, \beta) + \beta \ln t \tag{27}$$

The Elovich coefficients could be computed from the plots of qt vs. ln t. It is suggested that Elovich equation is restricted to the initial part of the adsorbate–adsorbent interaction process, when the system is relatively far from equilibrium.[7] The pseudo–second-order and the Elovich equations exhibit closely related behaviors for values of fractional surface coverage up to 0.7.[35]

Intraparticle Diffusion Model

This model is mostly used with the porous adsorbent. As the adsorbate molecules or ions diffuse into the pores, the adsorption kinetics relating to this intraparticle diffusion has to be studied with the proper model. The intraparticle diffusion model can be derived from Fick's second law. The simplified form of this equation as a dimensionless equation was proposed to assess the initial adsorption stages.[41–43] The intraparticle diffusion equation can be expressed as follows:

$$\frac{q_t}{q_e} = 1 - \left(\frac{6}{\pi^2}\right) \Sigma \left(\frac{1}{n^2}\right) \exp\left(\frac{-n^2\pi^2 D_C t}{r^2}\right) \tag{28}$$

where D_C is the intracrystalline diffusivity, r is the particle radius, t is the reaction time, and the summation is carried out from $n = 1$ to n = a. The ratio, q_t/q_e, is the fractional approach to equilibrium.

The simplified form is rewritten below:

$$1 - \frac{q_t}{q_e} = \left(\frac{6}{\pi^2}\right) \exp\left(\frac{-\pi^2 D_C}{r^2}\right) t \tag{29}$$

or

$$\ln\left(1 - \frac{q_t}{q_t}\right) = \left(\frac{-\pi^2 D_C}{r^2}\right) t + \ln\left(\frac{6}{\pi^2}\right) \tag{30}$$

This equation can be expressed in the linear line with the Plot of $\ln\left(1 - \frac{q_t}{q_e}\right)$ vs. t. A slope of $\left(\frac{-\pi^2 D_C}{r^2}\right)$ is the value of the diffusion time constant, and it is expressed as follows:

$$k' = \frac{\pi^2 D_C}{r^2} \tag{31}$$

where k' is the overall rate constant, inversely proportional to the square of the particle radius. A simpler expression to obtain the diffusion rate coefficient, k, is written as follows:

$$q_t = k_{it}^{0.5} \tag{32}$$

The linear plots of q_t vs. $t^{0.5}$ should pass through the origin (zero intercept) indicating a controlling influence for the diffusion process on the kinetics. The slope of the plot is the rate coefficient, k_i (mg/g-min$^{0.5}$). This equation represents a simplistic approximation of the pore diffusion kinetics without considering the possible impacts of the pore dimensions.

Liquid Film Diffusion Model

This equation is applied when the flow of the reactant through the liquid film surrounding the adsorbent particles is the slowest process. This equation determines kinetics of the rate process, and it is written as follows:[44]

$$\ln(1-F) = -k_{fd}t \tag{33}$$

where F is the fractional attainment of equilibrium $(=q_t/q_e)$ and k_{fd} (min^{-1}) is the film diffusion rate coefficient. A linear plot of $-\ln(1-F)$ vs. t with zero intercept suggests that the kinetics of the adsorption process is controlled by diffusion through the liquid film.

Adsorbents and Their Applications in Water and Wastewater Treatment

Carbon Adsorbents

There are several types of carbon adsorbents used in the adsorption process, for example, activated carbon, activated carbon fibers, fullerene, etc. Activated carbon is the most widely used carbon adsorbent. It is a crude form of graphite with a random or amorphous highly porous structure with a broad range of pore sizes, from visible cracks and crevices to crevices of molecular dimensions.[45] Generally, it is the carbon material derived mostly from charcoal. Besides coal, agricultural by-products are conventional sources of commercial activated carbon. Many types of agricultural wastes were proposed as raw materials for producing carbon adsorbents, including cork;[46–47] sucrose chars;[48] corncob;[49–50] jackfruit peels;[51–52] wood;[53–54] oil palm;[18] stones; fruit shells, coats, and husks;[48,55–64] and wastes from cherries,[65–66] plums,[67] coconut, apricot, almond, and nuts.[66–68] Conditions in synthesizing activated carbon from these materials are provided in Table 2. The comparison of adsorption capacity of each type of activated carbon is shown in Table 3. Sources of the raw materials used and the preparation and treatment conditions such as pyrolysis temperature and activation time are major factors that have an effect on the adsorption capacity of activated carbon. Many other factors can also affect the adsorption capacity in the same sorption conditions such as surface chemistry (heteroatom content), surface charge, and pore structure.

The adsorbent properties of activated carbon depend on their composition, physicochemical properties, and mechanical strength. Activation by physical means, by chemical means, or by a combination of both has been employed to control the pore size and distribution of activated carbon and/or to increase porosity, surface modification, and improvement of carbonization.[92–98] Normally, activated carbons are made up of small hydrophobic graphite layers with disordered, irregular, and heterogeneous surfaces bearing hydrophilic functional groups. The surface chemistry of activated carbon depends mainly upon the activation conditions and temperatures employed. The activated carbon has strong heterogeneous surfaces. Its geometrical heterogeneity is the result of differences in size and shape of pores, and cracks, pits, and steps. The chemical heterogeneity is involved with different functional groups, mainly oxygen groups that are located most frequently at the edges of the crystallites among with various surface impurities. These heterogeneity surfaces contribute to the unique sorption properties of activated carbon.[12,99–100] The presence of oxygen and hydrogen in surface groups affects strongly the adsorptive properties of the activated carbon. The apparent chemical character of an activated carbon surface is determined by functional groups and delocalized electrons of the graphitic structure.[101] Oxygen on an activated carbon surface may be present in various forms, such as carboxyls, aldehydes, ketones, phenols, lactones, quinines, hydroquinones, anhydrides, or ethereal structures. Some of the groups, e.g., carbonyl, carboxyl, phenolic hydroxyl, and lactonic ones, are acidic. Consequently, the pH value of the liquid bulk phase can have an effect on the acidic and/or basic functional groups of the carbon

TABLE 2 Conditions in Synthesizing Activated Carbon from Natural Materials

Natural Material	Conditions	Refs.
Algerian coal	930°C with KOH/NaOH	Alvim Ferraz[69]
Almond and pecan shells	Chemical activation with H_3PO^4/physical CO_2	Tancredi et al.[70]
Almond shell, olive stones, and peach stones	Heating in CO_2 at 606°C	Ferro-Garcia et al.[71]
Bituminous coal	N^2/400–700°C with $ZnCl_2$	Hall and Holmes[72]
Coal or coconut shell	Phosgene or chlorine gas at 180°C	Otowa et al.[73]
Coconut shell	Parts by weight H2SO4 for 24 h at 150°C	Manju et al.[74]
Coconut shell	450°C with H3PO4	Laine et al.[75]
Coconut shells and coconut shell fibers	Carbonized with H2SO4 and activated at 600°C for 1 hr	Mohan et al.,[76] Mohan et al.,[77] Mohan et al.,[78] and Mohan et al.[79]
Eucalyptus wood chars	CO_2 activation, 400–800°C	Kumar and Sivanesan[80]
Fertilizer slurry	450°C, 1 hr with H_2O_2/H_2O, N_2	Marungrueng and Pavasant[81]
Fly ash	Froth flotation, hydrophobic char was separated from hydrophilic ash with the help of methyl isobutyl ketone	Basava Rao and Mohan Rao[82]
Lignite	Inert atmosphere/600–800°C with Na_2MoO_4/ $NaWO_4$/ NH_4VO_3/$(NH_4)_2MoO_4$/$FeCl_3$/ $Fe(NO_3)_3$	El Qada et al.[83]
Oat hulls	Fast pyrolysis at 500°C with inert nitrogen	Tamai et al.[84]
Palm tree cobs	730°C, 6 hr with H_3PO_4/H_2SO_4	Banat et al.[85]
Petroleum coke	700–850°C, 4 hr with KOH/H^2O	McKay et al.[86]
	KOH dehydration at 400°C followed by activation at 500–900°C	McKay et al.[87]
Pine sawdust	850°C, 1 hr; 825°C, 6 hr with $Fe(NO_3)_3/CO_2$	Kannan and Sundaram[88]
Raffination earth	10% (v/v), 350°C with H_2SO_4	Bestani et al.[89]
Solvent-extracted olive pulp and olive stones	Under vacuum and atmospheric pressure; 60°C/min; 800°C; activation under N_2 at 10°C/min with K_2CO_3	Stavropoulos and Zabaniotou[90]

TABLE 3 Adsorption Capacity of Each Type of Activated Carbon

Adsorbent	Adsorption Capacity (mg/g)	Refs.
Activated carbon	400	Kumar and Sivanesan[80]
	238	Marungrueng and Pavasant[81]
	9.81	Basava Rao et al.[82]
Activated carbon produced from New Zealand coal	588	El Qada et al.[83]
Activated carbon produced from Venezuelan bituminous coal	380	El Qada et al.[83]
Bituminous coal	176	Tamai et al.[84]
Charcoal	62.7	Banat et al.[85]
Coal	323.68	McKay et al.[86]
	230	McKay et al.[87]
Commercial activated carbon	980.3	Kannan and Sundaram[88]
	200	Bestani et al.[89]
Peat	324	Fernandes et al.[91]

surface. Thus, the surface charge of carbon is a function of pH of the solution. Considering the point of zero charge (PZC) and the isoelectric point (IEP), the surface is positively charged at $pH < pH_{PZC}$ and negatively charged at $pH > pH_{PZC}$. In practice, pH_{IEP} is usually close to pHPZC, but it is lower than pHPZC for activated carbon.[102] For $pH < pK_a$ adsorption of non-ionized organics does not depend on the surface charge of activated carbon. However, for pH > pKa, the adsorption of its ionic form depends on the surface charge. As a result, the activated carbon possesses perfect adsorption ability for relatively low-molecular-weight organic compounds from drinking water and wastewater streams.

Several methods have been used to removal organic pollutants from water. However, the use of activated carbons is perhaps the best broad-spectrum technology available at present.[1] Accordingly, the use of activated carbons in water treatment has increased tremendously. Generally, the three main physical carbon types are granular, powder, and extruded (pellet). The granular activated carbon (GAC) adsorption, the most widely used type, is an effective treatment technology for organic contaminant removal from drinking water to improve taste and odor. The use of GAC for treatment of municipal and industrial wastewaters has developed rapidly in the last three decades from small size for household units to large scale for industrial wastewater application. Moving beds, downflow fixed beds, and upflow expanded beds have been widely used for water purification for industry.

It is well known that activated carbon can remove several types of pollutants including metal ions,[102–104] phenols,[46,105] pesticides,[106] chlorinated hydrocarbons[107] detergents,[108] and many other chemicals and organisms. Application of activated carbon in removal of various heavy metals and organic contaminants with the Langmuir and Freundlich capacities is shown in Table 4.

Clay Minerals

Clay minerals are hydrous aluminosilicates composed of minerals that make up the colloid fraction (<2 pm) of soils, sediments, rocks, and water[117] and may be composed of mixtures of fine-grained clay minerals and clay-sized crystals of other minerals such as quartz, carbonate, and metal oxides. Their structures are similar to micas with the formation of flat hexagonal sheets. Clay minerals and oxides are widespread and abundant in aquatic and terrestrial environments.

Clay contains various types of exchangeable ions on its surface. The prominent ions found on the clay surface are Ca^{2}+, Mg^{2}+, H+, K+, NH^{4}+, Na+, $SO4^{2-}$, Cl^{-}, $PO4^{3-}$, and $NO3^{-}$. These ions can be exchanged

TABLE 4 Adsorption Capacities of Activated Carbon for Heavy Metal and Organic Contaminant Removal from Water and Wastewater

Pollutant	Activated Carbon	Adsorption Capacity (mg/g)	Isotherm	Refs.
Cr(VI)	Commercial activated carbon	4.7	Langmuir	Babel and Kurniawan[109]
	Commercial activated carbon oxidized with H2SO4	8.9	Langmuir	Babel and Kurniawan[109]
	Commercial activated carbon oxidized with HNO_3	10.4	Langmuir	Babel and Kurniawan[109]
Fe(III)	Granular activated carbon	0.1	Freundlich	Kim[110]
Ni(II)	Granular activated carbon	6.5	Langmuir	Satapathy and Natarajan[111]
	Modified activated carbon	7.0	Langmuir	Satapathy and Natarajan[111]
Catechol	Activated charcoal	320	Langmuir	Richard et al.[112]
Gallic	acid Activated charcoal	408–488	Langmuir	Figaro et al.[113]
Tannin	Activated charcoal	0.39	Langmuir	Mohan and Karthikeyan[114]
Vanillin	Activated charcoal	93.18–121.72	Langmuir	Michailof et al.[63]
Phenol	Rice husk activated carbon	27.58	Langmuir	Kalderis et al.[115]
Nonylphenol	Activated charcoal	83.1	Langmuir	Lang et al.[116]

with other ions easily without affecting the structure of the clay mineral.[118] Clay can adsorb the cationic, anionic, and neutral metal species. They act as a natural scavenger of pollutants by taking up cations and/or anions through either ion exchange or adsorption, or both.

Currently, several types of clay minerals such as mont-morillonite and kaolinite are widely used in the water purification process. Because of their low cost, abundance in most continents of the world, high sorption properties, and potential for ion exchange, clay materials are strong adsorbents. Montmorillonite is a clay mineral with substantial isomorphic substitution. It is composed of units made up of two silica tetrahedral sheets with a central alumina octahedral sheet. The theoretical composition without the interlayer material is SiO_2, 66.7%; Al_2O_3, 28.3%; and H_2O, 5%. There is substitution of Si^4+ by Al^3+ in the tetrahedral layer and of Al^{3+} by Mg^{2+} in the octahedral layer. Exchangeable cations in the 2:1 layers balance the negative charges generated by isomorphic substitution. The uptake kinetics of cation exchange is fast, and the cations such as Na+ and Ca^2+ form outer-sphere surface complexes, which are easily exchanged with solute ions by varying the cationic composition of the solution.

Kaolinite is the least reactive clay. It has the theoretical composition of SiO_2, 46.54%; Al_2O_3, 39.50%; and H_2O, 13.96%, expressed in terms of the oxides. It has a small net negative charge, which is responsible for the surface not being completely inert. Its high pH dependency enhances or inhibits the adsorption of metals according to the pH of the environment.[119] The metal adsorption is usually accompanied by the release of hydrogen (H^+) ions from the edge sites of the mineral. The substitution of H^+ ions for metal ions could influence the van der Waals force within the kaolinite structure.

Their applications are mainly found in dye and heavy metal removal. From previous research, it was reported that the sorption capacity of clay minerals can vary strongly with pH. Gupta and Bhattacharyya[120–121] used kaolinite and montmorillonite along with their poly(oxo zirconium) and tetrabutylammonium derivatives for Cd(II) removal from water. The adsorption of Cd(II) was influenced by pH of the aqueous medium, and the amount adsorbed increased with gradually decreasing acidity. By increasing the solution pH from 1.0 to 10.0, the extent of adsorption increased from 4.3% to 29.5% for kaolinite and 74.7% to 94.5% for montmorillonite. In dye removal, Bagane and Guiza[122] reported an adsorption capacity of 300 mg/g and suggested that clay is a good adsorbent for methylene blue removal due to its high surface area. Almeida et al.[123] studied the removal of methylene blue from synthetic wastewater by using montmorillonite and described it as an efficient adsorbent where the equilibrium was attained in less than 30 min. The adsorption of dyes on kaolinite was also studied by Ghosh and Bhattacharyya,[124] who reported that its adsorption capacity can be improved by purification and by treatment with NaOH solution.

The adsorption capacities vary from metal to metal and also depend on the type of clay used.[118] When a comparison is made with other low-cost adsorbents, the clays have been found to be either better or equivalent in adsorption capacity. Type of pollutant and adsorption capacity of each clay mineral are summarized in Table 5.

Natural Zeolites

Zeolites are highly porous aluminosilicates with different cavity structures. They consist of a three-dimensional framework, having a negatively charged lattice. A well-defined pore structure in the microporous range of zeolite can accommodate a wide variety of cations such as Na^+, K^+, Ca^{2+}, Mg^{2+}, and others. These charge-compensating cations are free to migrate in and out of zeolite structures, and they are rather loosely held so that they can readily be exchanged for others in a contact solution. Accordingly, zeolites are not only good adsorbates but also good ion exchangers. This property can be used to introduce different cations into the structure, creating selective sites for adsorption purposes or catalysis. Their narrow pore size and tuneable affinity for certain molecules make them ideal adsorbents for selective purification to encapsulate hazardous compounds. Zeolites are characterized not only by a high selectivity separation mechanism but also by the ability to separate substances based on differences in sizes and shapes of molecules' steric separation mechanism.

TABLE 5 Adsorption Capacities of Clay Minerals for Heavy Metal Removal from Water and Wastewater

Pollutant	Clay Mineral	Langmuir Capacity	Freundlich Capacity	Refs.
Cd(II)	Kaolinite	9.9	0.5	Gupta and Bhattacharyya[125]
	Montmorillonite	32.7	8.6	Gupta and Bhattacharyya[125]
Ni(II)	Acid-activated montmorillonite	29.5	6.0	Bhattacharyya and Gupta[126]
	Kaolinite	10.4	1.1	Gupta and Bhattacharyya[127]
	Montmorillonite	28.4	4.5	Gupta and Bhattacharyya[127]
Cr(VI)	Kaolinite	11.6	–	Bhattacharyya and Gupta[128]
	Acid-activated kaolinite	13.9	–	Bhattacharyya and Gupta[128]
Co(II)	Raw kaolinite	11.5	–	Yavuz et al.[129]
	Kaolinite	11.2	1.1	Bhattacharyya and Gupta[130]
	Acid-activated kaolinite	12.1	1.5	Bhattacharyya and Gupta[130]
	Montmorillonite	28.6	4.6	Bhattacharyya and Gupta[130]
	Acid-activated montmorillonite	29.7	6.0	Bhattacharyya and Gupta[130]
Pb(II)	Kaolinite	11.2	0.7	Gupta and Bhattacharyya[131] and Bhattacharyya and Gupta[132]
	Acid-activated kaolinite	12.1	1.0	Gupta and Bhattacharyya[131] and Bhattacharyya and Gupta[132]
	Montmorillonite	33.0	8.9	Gupta and Bhattacharyya[131] and Bhattacharyya and Gupta[132]
	Acid-activated montmorillonite	34.0	11.3	Gupta and Bhattacharyya[131] and Bhattacharyya and Gupta[132]
Fe(III)	Kaolinite	11.2	1.3	Bhattacharyya and Gupta[133]
	Acid-activated kaolinite	12.1	1.7	Bhattacharyya and Gupta[133]
	Montmorillonite	28.9	5.2	Bhattacharyya and Gupta[133]
	Acid-activated montmorillonite	30.0	6.4	Bhattacharyya and Gupta[133]
Cu(II)	Kaolinite	4.4	1.1	Bhattacharyya and Gupta[134]
	Acid-activated kaolinite	5.6	1.3	Bhattacharyya and Gupta[126]
	Montmorillonite	25.5	9.2	Bhattacharyya and Gupta[134]
	Acid-activated montmorillonite	28.0	12.4	Bhattacharyya and Gupta[126]

Note: Units of Langmuir capacity and Freundlich capacity are mg/g and $\mathrm{mg}^{1-1/n}\ \mathrm{L}^{1/n}/\mathrm{g}$, respectively.

Zeolites have been widely used for pollution control due to their ion exchange and adsorption properties. They have been used for the selective separation of cations from aqueous solution. The diffusion, adsorption, and ion exchange in zeolites have been extensively reviewed in many previous works.[135–137] Kesraoui-Ouki, Cheeseman, and Perry[138] reviewed natural zeolite utilization in metal effluent treatment applications. Dewatered zeolites produce channels that can adsorb molecules small enough to access the internal cavities while excluding larger species. Zeolites, modified by ion exchange, can be used for adsorption of different metal ions according to requirements and costs. The characteristics and applications of zeolites have been extensively reviewed by Ghobarkar, Schaf, and Guth.[139] High ion-exchange capacity and relatively high specific surface areas, and more importantly, their relatively cheap prices, make zeolites more attractive adsorbents.

Besides zeolite, other siliceous materials such as perlite and glass have been proposed for contaminant removal. The use of natural siliceous adsorbents such as silica, glass fibers, and perlite for wastewater is increasing because of their high abundance, easy availability, and low cost. The other commonly applied inorganic sorbents are silica gels, activated alumina, and oxide and hydroxide metals. Perlite is another siliceous material that exhibits a good adsorbent for decontamination purposes. It has been used as a low-cost adsorbent for the removal of methylene blue.[140,141] Methylene blue is physically adsorbed onto the perlite. However, perlites of different types (expanded and unexpanded) and of different origins have different properties because of the differences in composition. Chakrabarti and Dutta[142] also investigated glass fiber for the adsorption of methylene blue. They stated that a considerable amount of the dye is adsorbed on soft glass even at ambient temperature. Accordingly, several siliceous materials become widely used as adsorbate materials in the adsorption process.

Currently, a new family of mesopore materials, so-called MCM materials or Mobil Composition of Matter (MCM), was developed by Mobil Oil Corporation, which proposed a revolutionary synthesis method to obtain such materials that comprise strictly uniform pores. An organic surfactant like an alkyltrimethylammonium bromide in an aqueous medium forms rod-like micelles, which are used as templates to form two or three monolayers of silica or alumina particles encapsulating the micelles' external surface. By removing the organic species from a well-ordered organic-inorganic condensed phase, a porous silicate or alumina material with uniformly porous structure remains. The mesopore size can be controlled by the molecular size template of the surfactant. Nowadays, MCM materials have been widely used in heavy metal removal, and they are currently the adsorption material that plays an important role in water and wastewater treatment.

Chitin and Chitosan

Chitin is a nontoxic, biodegradable polymer of high molecular weight. It contains 2-acetamido-2-deoxy-β-D-glucose through a β (1→4) linkage. Chitin is the most abundant natural fiber next to the cellulose and is similar to cellulose in many respects. The most abundant source of chitin is the shell of crab and shrimp. Chitin has presented exceptional chemical and biological qualities that can be used in water and wastewater purification through the adsorption process.

Chitin and chitosan have their chemical structures in common. Chitin is made up of a linear chain of acetyl-glucosamine groups. Chitosan is obtained by removing enough acetyl groups (CH_3–CO) for the molecule to be soluble in most diluted acids. This process, called deacetylation, releases amine groups (NH) and gives the chitosan a cationic characteristic. Chitosan contains 2-acetamido-2-deoxy-β-D-glucopyranose and 2-amino-2-deoxy-β-D-glucopyranose residues. Chitosan is known as an ideal natural support for enzyme immobilization because of its special characteristics such as hydrophilicity, biocompatibility, biodegradability, non-toxicity, adsorption properties, etc.[143]

Chitosan has drawn particular attention as an effective biosorbent due to its high content of amino and hydroxyl functional groups, giving it high adsorption potential for various aquatic pollutants.[143–147] This biopolymer represents an attractive alternative to other biomaterials because of its physicochemical characteristics, chemical stability, high reactivity, excellent chelation behavior. and high selectivity toward pollutants. Chitin and chitosan derivatives have been extensively investigated as adsorbents for the removal of organic molecules and metal ions from water and wastewater. The high adsorption potential of chitosan can be attributed to the following: 1) high hydrophilicity due to a large number of hydroxyl groups of glucose units; 2) presence of a large number of functional groups; 3) high chemical reactivity of these groups; and 4) flexible structure of the polymer chain.[148,149]

To enhance the adsorption capacity for pollutant removal, chitosan has been modified by several methods, either physical or chemical processes. Different shapes of chitosan, e.g., membranes, microspheres, gel beads, and films, have been synthesized and tested for their performance in pollutant removal from water and wastewater.[143–147] A cross-linked chitosan bead is one type of chemical modification for chitosan to increase the uptake capacity in the adsorption process.[150] This method using the chemical reaction of ethylenediamine and carbodiimide in modifying chitosan provided a high uptake capacity for mercury (Hg^{2+}) ions, which is considered to be one of the highest uptake capacities among various biosorbents.

Beads of 1 and 3 mm diameter were prepared as one type of modified chitosan.[151] The gelled chitosan beads were cross-linked with glutaraldehyde and then freeze-dried. Beads of 1 mm diameter possessed surface areas exceeding 150 m^2/g and mean pore sizes of 560 Å and were insoluble in acid media at pH 2. A new composite chitosan biosorbent was also prepared by coating chitosan onto perlite ore. It was used in the removal of Cu(II) and Ni(II) from aqueous solution.[152] The magnetic chitosan nanocomposites were synthesized on the basis of amine-functionalized magnetite nanoparticles.[153] These nanocomposites provide a very efficient, fast, and convenient tool for removing Pb^{2+}, Cu^{2+}, and Cd^{2+} from water. It was suggested that synthesized magnetic chitosan nanocomposites can be used as a recyclable tool for heavy metal ion removal. Several types of heavy metals and organic contaminants removed by chitosan are shown in Table 6.

Agricultural-Based Waste Materials

Agricultural by products usually are composed of lignin and cellulose as major constituents that have the ability to some extent to bind some type of pollutants, for example, heavy metals, by donation of an electron pair from these groups to from complexes with the metal ions.[166] Currently, many types of agricultural-based waste materials play a significant role in the adsorption process. They are normally organic materials from plants, trees, crops, and algae. Two larger carbohydrate that play a significant role in the adsorption process are cellulose and hemicelluloses (holocellulose.) Cellulose is a remarkable pure organic polymer, consisting solely of units of anhydroglucose held together in a giant straight-chain molecule.[168] These anhydroglucose units are bound together byβ-(1,4)-glycosidic linkages. Hemicelluloses consist of different monosaccharide units. The polymer chains of hemicelluloses have

TABLE 6 Adsorption Capacities of Chitosan and Its Composite for Removal of Heavy Metals and Some Organic Contaminants from Water and Wastewater

Pollutant	Chitosan	Adsorption Capacity (mg/g)	Isotherm	Refs.
Hg(II)	Chitosan/cotton fibers	104.31	Langmuir	Qu et al.[154]
Cd(II)	Chitosan/cotton fibers	15.74	Langmuir	Zhang et al.[155]
Cr(VI)	Magnetic chitosan	69.40	Langmuir	Huang et al.[156]
	Chitosan/cellulose	13.05	Langmuir	Sun et al.[157]
	Chitosan/perlite	153.8	Langmuir	Shameem et al.[158]
	Chitosan/ceramic alumina	153.8	Freundlich	Veera et al.[159]
Pb(II)	Chitosan/cotton fibers	101.53	Freundlich	Zhang et al.[155]
	Chitosan/magnetite	63.33	Langmuir	Tran et al.[160]
	Chitosan/cellulose	26.31	Langmuir	Sun et al.[157]
	Chitosan/sand	12.32	Langmuir	Rorrer et al.[151]
Cu(III)	Chitosan/cellulose	26.50	Langmuir	Sun et al.[157]
	Chitosan/perlite	196.07	Langmuir	Kalyani et al.[152]
	Chitosan/polyvinylchloride	87.9	Langmuir	Srinivasa et al.[162]
Ni(II)	Chitosan/magnetite	52.55	Langmuir	Tran et al.[160]
	Chitosan/cellulose	13.21	Langmuir	Sun et al.[157]
	Chitosan/perlite	114.94	Langmuir	Kalyani et al.[152]
	Chitosan/silica	254.3	Langmuir	Vijaya et al.[163]
Phenol	Chemically modified chitosan	2.22–151.50	Langmuir	Li et al.[164]
	Chitosan/calcium alginate beads	108.69	Langmuir	Nadavala et al.[165]
4-Chlorophenol	Chemically modified chitosan	2.58–179.73	Langmuir	Li et al.[164]
Nonylphenol	Chitosan	56.3	Langmuir	Lang et al.[116]

short branches and are amorphous. Hemicelluloses are derived mainly from chains of pentose sugars and act as the cement material holding together the cellulose micelles and fiber.[168] Hemicelluloses are partially soluble in water. Currently, chemical modification is widely used to alter the biochemical component of the biomaterials to obtain higher efficiency in pollutant removal by biosorption process.[169] Biomass chemical modifications include delignification, esterification of carboxyl and phosphate groups, methylation of amino groups, and hydrolysis of carboxylate groups. Sawamiappan and Krishnamoorthy[170] replaced phenol–formaldehyde cationic matrices with sulfonated bagasse. Odozi et al.[171] polymerized corncob, sawdust, and onion. However, the disadvantages of chemical modification are a high expense to pay and unwanted problems, such as bleeding of excessive quantities of colored organic compounds, odor, and further pollution through the use of toxic chemicals. Several types of agricultural wastes have been used in the adsorption process, with the differences in adsorption capacity as shown in Table 7.

Mechanisms involved in the biosorption process include chemical adsorption, complexation, adsorption–complexation on surfaces and in pores, ion exchange, microprecipitation, heavy metal hydroxide condensation onto the biosurface, and surface adsorption.[172–174] In the adsorption process, functional groups are responsible for pollutant binding on the surface of biomaterial. Most of the functional groups involved in the binding process are found in cell walls. Plant cell walls are generally considered as structures built by cellulose molecules, organized in microfibrils and surrounded by hemicellulosic materials (xylans, mannans, glucomannans, galactans, arabogalactans), lignin, and

TABLE 7 Adsorption Capacities of Agricultural Waste for Heavy Metal And Organic Contaminant Removal from Water and Wastewater

Pollutant	Agriculture Waste	Adsorption Capacity (mg/g)	Isotherm	Refs.
Cd(II)	Juniper fiber	9.2	Langmuir	Min et al.[181]
	Base-treated juniper fiber	29.5	Langmuir	Min et al.[181]
Cr(VI)	Cactus	7.1	Langmuir	Dakiky et al.[182]
	Coconut shell carbon	2.2	Langmuir	Babel and Kurniawan[109]
	Coconut shell carbon oxidized with H2SO4	4.1	Langmuir	Babel and Kurniawan[109]
	Coconut shell carbon oxidized with HNO3	10.9	Langmuir	Babel and Kurniawan[109]
	Sawdust	15.8	Langmuir	Dakiky et al.[182]
Pb(II)	Carbonaceous adsorbent	25.0	Langmuir	Bhatnagar et al.[183]
	Sawdust	22.2	Langmuir, Freundlich	Taty-Costodes et al.[184]
Fe(III)	Maize cobs	2.5	Langmuir	Nassar et al.[185]
Cu(II)	Tree fern	7.6	Langmuir	Ho et al.[186]
Ni(II)	Peat	28.3	Langmuir, Freundlich	Chen et al.[187]
Phenol	Banana pith	49.9–129.4	Langmuir	Sathishkumar et al.[188]
	Banana peel	688.9	Langmuir	Achak et al.[189]
	Corn grain	256	Langmuir	Park et al.[190]
2-Nitrophenol	*Lessonia nigrescens*	71.28	Langmuir	Navarro et al.[191]
	Macrocystis integrifolia	97.37	Langmuir	Navarro et al.[191]
2,4-Dicholorophenol	Pomegranate peel	65.7	Langmuir	Bhatnagar and Minocha[192]
Nonylphenol	*Rhizopus arrhizus*	4.5–43.7	Langmuir	Lang et al.[116]

pectin along with small amounts of protein.[175] During biosorption, water is able to permeate the noncrystalline portion of cellulose and all of the hemicellulose and lignin. The aqueous solution comes into contact with a very large surface area of different cell wall components. The disordered structure of amorphous cellulose allows easier access to reagents than highly structured crystalline cellulose. While water penetrates through the cell wall components, water adsorption of fibers causes swelling. The bigger the amount of water adsorption, the bigger the swelling. Swelling also depends on the fiber's structure, on the degree of crystallinity, and on the amorphous and void regions.[176] Swelling occurs when polar solvents such as water and alcohols come into contact with wood.[177] These polar solvent molecules are attracted to the dry solid matrix and held by hydrogen bonding forces between the –OH or –COOH groups in the wood structure and cause the biosorption of pollutants in aqueous solution. Many research works[178–180] have reported the wide use of biosorption process in heavy metal removal. Thus, the agricultural-based waste materials become the adsorption material that plays an important role in water and wastewater treatment nowadays (Table 7).

Progress in Research on Adsorption Process in Water Purification

Adsorbents and adsorption processes have been widely studied and applied in different aspects for a long time. Owing to its effective, efficient, and economic approach to water purification, this process has been applied in removal of several contaminants, such as pesticides, halogenated carbon, dyes, phenol and its derivatives, and heavy metals. The most widely used adsorbent in the adsorption process is activated carbon. This adsorbent is highly inert and thermally stable, and it can be used over a broad pH range. Although it has a great capacity for adsorbing various organic compounds and can be easily modified by chemical treatment to increase its adsorption capacity, activated carbon has several disadvantages.[193] Owing to the process mechanism, adsorption transfers pollutants from one phase to another rather than eliminating them from the environment. Thus, after adsorption, the contaminants in liquid phase absorb on the surface of adsorbent, which has to be separated from aquatic system when it becomes exhausted or the effluent reaches the maximum allowable discharge level. Furthermore, the regeneration of exhausted activated carbon by a chemical and thermal procedure is also expensive and results in loss of the sorbent.

Recently, a lot of novel adsorption processes have been developed for enhancing the efficiency of removing organic and inorganic contaminants from water. The development of cheaper and more effective novel composite adsorbents[194–197] in comparison with the classical adsorbents has been investigated by researchers from many countries all over the world. These adsorbents are metal oxide-based composite adsorbents such as TiO_2 and MnO_2, surface-modified Fe_3O_4 adsorbent, magnetic particle-modified carbon adsorbent, magnetic particle–modified clay mineral adsorbent, and magnetic particle–modified biopolymer adsorbent. These composite materials deserve particular attention because they combine the properties and advantages of each of their components. They represent an interesting and attractive alternative as adsorbents and/or catalysts due to their high reactivity and excellent selectivity toward specific pollutant compounds. To obtain the anticipated function and enhance the efficiency of water purification, these adsorbents should be designed and modified in their compositions, structures, surfaces, and preparation methods to obtain the requirement of physicochemical properties for the purpose of adsorption. Extensive research in synthesizing of new adsorbents and investigating of adsorption mechanism is needed. Advances in development of new adsorbents for the adsorption process will be the progress of future technology in water purification.

References

1. Dabrowski, A. Adsorption, from theory to practice. Adv. Colloid Interface Sci. **2001**, *93,* 135–224.
2. Ahmad, A.; Rafatullah, M.; Danish, M. Removal of Zn(II) and Cd(II) ions from aqueous solutions using treated sawdust of sissoo wood. Holz Roh Werkst **2007**, *65,* 429–436.

3. Ahmad, A.; Rafatullah, M.; Sulaiman, O.; Ibrahim, M.H.; Chii, Y.Y.; Siddique, B.M. Removal of Cu(II) and Pb(II) ions from aqueous solutions by adsorption on sawdust of Meranti wood. Desalination **2009**, *247,* 636–646.
4. Rafatullah, M.; Sulaiman, O.; Hashim, R.; Ahmad, A. Adsorption of copper (II), chromium (III), nickel (II) and lead (II) ions from aqueous solutions by meranti sawdust. J. Hazard. Mater. **2009**, *170,* 969–977.
5. Gupta, V.K.; Carrott, P.J.M; Ribeiro Carrott M.M.L; Su-has. Low-cost adsorbents: Growing approach to wastewater treatment—A review. Crit. Rev. Environ. Sci. Technol. **2009**, *39,* 783–842.
6. Soto, M.L.; Moure, A.; Dominguez, H.; Parajò, J.C. Recovery, concentration and purification of phenolic compounds by adsorption: A review. J. Food Eng. **2011**, *105,* 1–27.
7. Gupta, S.S.; Bhattacharyya, K.G. Kinetics of adsorption of metal ions on inorganic materials: A review. Adv. Colloid Interface Sci. **2011**, *162,* 39–58.
8. Gupta, V.K.; Mohan, D.; Sharma, S. Removal of lead from wastewater using bagasse fly ash—A sugar industry waste material. Sep. Sci. Technol. **1998**, 33, 1331–1343.
9. Gupta, V.K.; Jain, C.K.; Ali, I.; Sharma, M.; Saini, V.K. Removal of cadmium and nickel from wastewater using bagasse fly ash—A sugar industry waste. Water Res. **2003**, 37, 4038–4044.
10. Parfitt, D.; Rochestor, H. Adsorption of small molecules. In *Adsorption from Solution at the Solid/Liquid Interface;* Parfitt, G.D., Rochester, C.H., Eds.; Academic Press: New York, 1983; 3.
11. Giles, C.H.; Smith, D.; Huitson, A. A general treatment and classification of the solute adsorption isotherm. I. Theoretical. J. Colloid Interface Sci. **1974**, *47,* 755–765.
12. Dabrowski, A.; Podkoscielny, P.; Hubicki, Z.; Barczak, M. Adsorption of phenolic compounds by activated car-bon—A critical review. Chemosphere **2005**, *58,* 1049–1070.
13. Bansal, R.C.; Goyal, M. *Activated Carbon Adsorption*; CRC Press: Boca Raton, FL, 2005.
14. Bretag, J.; Kammerer, D.R.; Jensen, U.; Carle, R. Evaluation of the adsorption behavior of flavonoids and phenolic acids onto a food-grade resin using a D-optimal design. Eur. Food Res. Technol. **2009**, *228,* 985–999.
15. Weber, T.W.; Chakkravorti, R.K. Pore and solid diffusion models for fixed bed adsorbers. AIChE J. **1974**, *20*, 228–238.
16. Freundlich, H.M.F. Over the adsorption in solution. J. Phys. Chem. **1906**, 57, 385–470.
17. Fytianos, K.; Voudrias, E.; Kokkalis, E. Sorption-desorption behavior of 2,4-dichlorophenol by marine sediments. Chemosphere **2000**, *40*, 3–6.
18. Foo, K.Y.; Hameed, B.H. Insights into the modeling of adsorption isotherm systems. Chem. Eng. J. **2010**, *156*, 2–10.
19. Temkin, M.I.; Pyzhev, V. Kinetics of ammonia synthesis on promoted iron catalysts. Acta Phys. Chem. U.R.S.S. **1940**, *12,* 327–356.
20. Rashidi, F.; Sarabi, R.S.; Ghasemi, Z.; Seif, A. Kinetic, equilibrium and thermodynamic studies for the removal of lead (II) and copper (II) ions from aqueous solutions by nanocrystalline TiO_2. Superlattices Microstruct. **2010**, *48,* 577–591.
21. Dubinin, M.M.; Zaverina, E.D.; Radushkevich, L.V. Sorption and structure of active carbons. I. Adsorption of organic vapors. J. Phys. Chem. **1947**, *21,* 1351–1362.
22. Unlu, N.; Ersoz, M. Adsorption characteristics of heavy metal ions onto a low cost biopolymeric sorbent from aqueous solutions. J. Hazard. Mater. **2006**, *136,* 272–280.
23. Redlich, O.; Peterson, D.L. A useful adsorption isotherm. J. Phys. Chem. **1959**, *63,* 1024–1029.
24. Hameed, B.H.; Tan, I.A.W.; Ahmad, A.L. Adsorption isotherm, kinetic modeling and mechanism of 2,4,6 trichlorophenol on coconut husk-based activated carbon. Chem. Eng. J. **2008**, *144,* 235–244.
25. Ho, Y.S.; Ng, J.C.Y.; McKay, G. Kinetics of pollutant sorption by biosorbents: Review. Sep. Purif. Methods **2000**, *29*, 189–232.
26. Lagergren, S. Zur theorie der sogenannten adsorption gelöster stoffe. Kungliga Svenska Vetenskapsakademiens. Han-dlingar, Vetensk. Handl. **1898**, *24,* 1–39.

27. McKay, G.; Ho, Y.S.; Ng, J.C.Y. Biosorption of copper from wastewaters: A Review. Sep. Purif. Methods **1999**, *28*, 87–125.
28. Allen, S.J.; Gan, Q.; Matthews, R.; Johnson, P.A. Kinetic modeling of the adsorption of basic dyes by kudzu. J. Colloid Interface Sci. **2005**, *286,* 101–109.
29. Febrianto, J.; Kosasih, A.N.; Sunarso, J.; Ju, Y.-H.; Indras-wati, N.; Ismadji, S. Equilibrium and kinetic studies in adsorption of heavy metals using biosorbent: A summary of recent studies. J. Hazard. Mater. **2009**, *162,* 616–645.
30. Al-Ghouti, M.A.; Khraisheh, M.A.M.; Ahmad, M.N.M.; Allen, S. Adsorption behaviour of methylene blue onto Jordanian diatomite: A kinetic study. J. Hazard. Mater. **2009**, *165*, 589–598.
31. Nandi, B.K.; Goswami, A.; Purkait, M.K. Adsorption characteristics of brilliant green dye on kaolin. J. Hazard. Mater. **2009**, *161*, 387–395.
32. Lin. C. I.; Wang, L. H. Rate equations and isotherms for two adsorption models. J. Chin. Inst. Chem. Eng. **2008**, *39,* 579–585.
33. Plazinski, W.; Rudzinski, W.; Plazinska, A. Theoretical models of sorption kinetics including a surface reaction mechanism: A review. Adv. Colloid Interface Sci. **2009**, *152,* 2–13.
34. Ho, Y.S.; McKay, G. The kinetics of sorption of divalent metal ions onto sphagnum moss peat. Water Res. **2000**, *34,* 735–742.
35. Rudzinski, W.; Plazinski, W. On the applicability of the pseudo-second order equation to represent the kinetics of adsorption at solid/solution interfaces: A theoretical analysis based on the statistical rate theory. Adsorption **2009**, *15,* 181–192.
36. Ho, Y.S.; McKay, G. Kinetic model for lead(II) sorption on to peat. Adsorpt. Sci. Technol. **1998**, *16,* 243–255.
37. Aharoni, C.; Tompkins, F.C. Kinetics of adsorption and desorption and the Elovich equation. In *Advances in Catalysis and Related Subjects;* Eley, D.D., Pines, H., Weisz, P.B., Eds.; Academic Press: New York, **1970**; Vol. 21, 2–50.
38. Ho, Y.S.; McKay, G. Application of kinetic models to the sorption of copper(II) onto peat. Adsorpt. Sci. Technol. **2002**, *20,* 797–813.
39. Rudzinski, W.; Panczyk, T. The Langmuirian adsorption kinetics revised: A farewell to the 20th century theories? Adsorption **2002**, *8,* 23–34.
40. Chien, S.; Clayton, W.R. Application of Elovich equation to the kinetics of phosphate release and sorption in soils. Soil Sci. Soc. Am. J. **1980**, *44*, 265–268.
41. Ruthven, D.M. *Principles of Adsorption and Adsorption Processes*; Wiley-Interscience: New York, **1984**.
42. Banerjee, K.; Cheremisinoff, P.N.; Cheng, S.L. Adsorption kinetics of *o*-xylene by fly ash. Water Res. **1997**, *31,* 249–261.
43. Manju, G.N.; Anoop Krishnan, K.; Vinod, V.P.; Anirudhan, T.S. An investigation into the sorption of heavy metals from wastewaters by polyacrylamide-grafted iron(III) oxide. J. Hazard. Mater. **2002**, *91,* 221–238.
44. Boyd, G.E.; Adamson, A.W.; Myers Jr., L.S. The Exchange adsorption of ions from aqueous solutions by organic zeolites. II. Kinetics. J. Am. Chem. Soc. **1947**, *69,* 2836–2848.
45. Hamerlinck, Y.; Mertens, D.H. In *Activated Carbon Principles in Separation Technology;* Vansant, E.F., Ed.; Elsevier: New York, 1994.
46. Mourão, P.A.M.; Carrott, P.J.M.; Ribeiro Carrott, M.M.L. Application of different equations to adsorption isotherms of phenolic compounds on activated carbons prepared from cork. Carbon **2006**, *44*, 2422–2429.
47. Mestre, A.S.; Pires, J.; Nogueira, J.M.F.; Parra, J.B.; Carvalho, A.P.; Ania, C.O. Waste-derived activated carbons for removal of ibuprofen from solution: Role of surface chemistry and pore structure. Bioresour. Technol. **2009**, *100,* 1720–1726.
48. Evans, M.J.B.; Halliop, E.; MacDonald, J.A.F. The production of chemically activated carbon. Carbon **1999**, *37*, 269–274.

49. Wu, F.C.; Tseng, R.L.; Juang, R.S. Adsorption of dyes and phenols from water on the activated carbons prepared from corncob wastes. Environ. Technol. **2001**, *22*, 205–213.
50. Tseng, R.L.; Tseng, S.K. Characterization and use of high surface area activated carbons prepared from cane pith for liquid-phase adsorption. J. Hazard. Mater. **2006**, *136,* 671–680.
51. Prahas, Y.D.; Kartika, N.; Indraswati, S.; Ismadji, S. Activated carbon from jackfruit peel waste by H3PO4 chemical activation: Pore structure and surface chemistry characterization. Chem. Eng. J. **2007**, *140,* 32–42.
52. Jain, S.; Jayaram, R.V. Adsorption of phenol and substituted chlorophenols from aqueous solution by activated carbon prepared from jackfruit *(Artocarpus heterophyllus)* peel-kinetics and equilibrium studies. Sep. Sci. Technol. **2007**, *42,* 2019–2032.
53. Tancredi, N.; Medero, N.; Möller, F.; Píriz, J.; Plada, C.; Cordero, T. Phenol adsorption onto powdered and granular activated carbon, prepared from Eucalyptus wood. J. Colloid Interface Sci. **2004**, *279,* 357–363.
54. Mudoga, H.L.; Yucel, H.; Kincal, N.S. Decolonization of sugar syrups using commercial and sugar beet pulp based activated carbons. Bioresour. Technol. **2008**, *99*, 3528–3533.
55. Philip, C.A.; Girgis, B.S. Adsorption characteristics of microporous carbons from apricot stones activated by phosphoric acid. J. Chem. Technol. Biotechnol. **1996**, *67,* 248–254.
56. Ahmadpour, A.; Do, D.D. The preparation of activated carbon from Macadamia nutshell by chemical activation. Carbon **1997**, *35*, 1723–1732.
57. Toles, C.A.; Marshall, W.E.; Johns, M.M.; Wartelle, L.H.; McAloon, A. Acid activated carbons from almond shells: Physical, chemical and adsorptive properties and estimated cost of production. Bioresour. Technol. **2000**, *71,* 87–92.
58. Toles, C.A.; Marshall, W.E.; Wartelle, L.H.; McAloon, A. Steam- or carbon dioxide–activated carbons from almond shells: Physical, chemical and adsorptive properties and estimated cost of production. Bioresour. Technol. **2000**, *75*, 197–203.
59. Galiatsatou, P.; Metaxas, M.; Arapoglou, D.; Kasselouri- Rigopoulo, V. Treatment of olive mill waste water with activated carbons from agricultural by-products. Waste Manage. **2002**, *22*, 803–812.
60. Rengaraj, S.; Moon, S.H.; Sivabalan, R.; Arabindoo, B.; Murugesan, V. Agricultural solid waste for the removal of organics: Adsorption of phenol from water and wastewater by palm seed coat activated carbon. Waste Manage. **2002**, *22*, 543–548.
61. Qi, J.; Li, Z.; Guo, Y.; Xu, H. Adsorption of phenolic compounds on micro- and mesoporous rice husk-based active carbons. Mater. Chem. Phys. **2004**, *87,* 96–101.
62. Tan, I.A.W.; Ahmad, A.L.; Hameed, B.H. Preparation of activated carbon from coconut husk: Optimization study on removal of 2,4,6-trichlorophenol using response surface methodology. J. Hazard. Mater. **2008**, *153,* 709–717.
63. Michailof, C.; Stavropoulos, G.G.; Panayiotou, C. Enhanced adsorption of phenolic compounds, commonly encountered in olive mill wastewaters, on olive husk derived activated carbons. Bioresour. Technol. **2008**, *99,* 6400–6408.
64. Muñoz-González, Y.; Arriagada-Acuña, R.; Soto-Garrido, G.; García-Lovera, R. Activated carbons from peach stones and pine sawdust by phosphoric acid activation used in clarification and decolorization processes. J. Chem. Technol. Biotechnol. **2009**, *84,* 39–47.
65. Lussier, M.G.; Shull, J.C.; Miller, D.J. Activated carbons from cherry stones. Carbon **1994**, *32,* 1493–1498.
66. Shopova, N.; Minkova, V.; Markova, K. Evaluation of the thermochemical changes in agricultural by-products and in the carbon adsorbents obtained from them. J. Therm. Anal. **1997**, *48*, 309–320.
67. Marsh, H.; Iley, M.; Berger, J.; Siemieniewska, T. The adsorptive properties of activated plum stone chars. Carbon **1975**, *13,* 103–109.
68. Mohanty, K.; Jha, M.; Meikap, B.C.; Biswas, M.N. Preparation and characterization of activated carbons from *Ter- minalia arjuna* nut with zinc chloride activation for the removal of phenol from wastewater. Ind. Eng. Chem. Res. **2005**, *44*, 4128–4138.

69. Alvim Ferraz, M.C. Preparation of activated carbon for air pollution control. Fuel **1988**, *67,* 1237–1241.
70. Tancredi, N.; Cordero, T.; Rodríguez-Mirasol, J.; Rodríguez, J.J. Activated carbons from Uruguayan eucalyptus wood. Fuel **1996**, *73*, 1701–1706.
71. Ferro-Garcia, M.A.; Rivera-Utrilla, J.; Rodriguez-Gordillo, J. Bautista-Toledo, I. Adsorption of zinc, cadmium, and copper on activated carbons obtained from agricultural byproducts. Carbon **1988**, *26*, 363–373.
72. Hall, C.R.; Holmes, R.J. The preparation and properties of some activated carbons modified by treatment with phosgene or chlorine. Carbon **1992**, *30,* 173–176.
73. Otowa, T.; Nojima, Y.; Miyazaki, T. Development of KOH activated high surface area carbon and its application to drinking water purification. Carbon **1997**, *35*, 1315–1319.
74. Manju, G.N.; Raji, C.; Anirudhan, T.S. Evaluation of coconut husk carbon for the removal of arsenic from water. Water Res. **1998**, *32,* 3062–3070.
75. Laine, J.; Calafat, A.; Labady, M. Preparation and characterization of activated carbons from coconut shell impregnated with phosphoric acid. Carbon **1989**, *27,* 191–195.
76. Mohan, D.; Singh, K.P.; Ghosh D. Removal of α-picoline, β-picoline, and γ-picoline from synthetic wastewater using low cost activated carbons derived from coconut shell fibers. Environ. Sci. Technol. **2005**, *39,* 5076–5086.
77. Mohan, D.; Singh, K.P.; Sinha, S.; Ghosh, D. Removal of pyridine derivatives from aqueous solution by activated carbons developed from agricultural waste materials. Carbon **2005**, *43,* 1680–1693.
78. Mohan, D.; Singh, K.P.; Singh, V.K. Trivalent chromium removal from wastewater using low cost activated carbon derived from agricultural waste material and activated carbon fabric cloth. J. Hazard. Mater. **2006**, *133,* 280–295.
79. Mohan, D.; Singh, K.P.; Sinha, S.; Ghosh, D. Removal of pyridine from aqueous solution using low cost activated carbons developed from agricultural waste materials. Carbon **2004**, *43,* 2409–2421.
80. Kumar, K.V.; Sivanesan, S. Equilibrium data, isotherm parameters and process design for partial and complete isotherm of methylene blue onto activated carbon. J. Hazard. Mater. **2006**, *134*, 237–244.
81. Marungrueng, K.; Pavasant, P. High performance biosorbent *(Caulerpa lentillifera)* for basic dye removal. Bioresour. Technol. **2007**, *98,* 1567–1572.
82. Basava Rao, V.V.; Mohan Rao, S.R. Adsorption studies on treatment of textile dyeing industrial effluent by fly ash. Chem. Eng. J. **2006**, *116,* 77–84.
83. El Qada, E.N.; Allen, S.J.; Walker, G.M. Adsorption of basic dyes from aqueous solution onto activated carbons. Chem. Eng. J. **2008**, *133,* 174–184.
84. Tamai, H.; Kakii, T.; Hirota, Y.; Kumamoto, T.; Yasuda, H. Synthesis of extremely large mesoporous activated carbon and its unique adsorption for giant molecules. Chem. Mater. **1996**, *8*, 454–462.
85. Banat, F.; Al-Asheh, S.; Al-Ahmad, R.; Bni-Khalid, F. Bench-scale and packed bed sorption of methylene blue using treated olive pomace and charcoal. Bioresour. Technol. **2007**, *98,* 3017–3025.
86. McKay, G.; Porter, J.F.; Prasad, G.R. The removal of dye colours from aqueous solutions by adsorption on low-cost materials. Water Air Soil Pollut. **1999**, *114,* 423–438.
87. McKay, G.; Ramprasad, G.; Pratapamowli, P. Equilibrium studies for the adsorption of dyestuffs from aqueous solution by low-cost materials. Water Air Soil Pollut. **1986**, *29,* 273–283.
88. Kannan, N.; Sundaram, M.M. Kinetics and mechanism of removal of methylene blue by adsorption on various carbons— A comparative study. Dyes Pigments **2001**, *51*, 25–40.
89. Bestani, B.; Benderdouche, N.; Benstaali, B.; Belhakem, M.; Addou, A. Methylene blue and iodine adsorption onto an activated desert plant. Bioresour. Technol. **2008**, *99*, 8441–8444.
90. Stavropoulos, G.G.; Zabaniotou, A.A. Production and characterization of activated carbons from olive-seed waste residue. Microporous Mesoporous Mater. **2005**, *82,* 79–85.

91. Fernandes, A.N.; Almeida, C.A.P.; Menezes, C.T.B.; De- bacher, N.A.; Sierra, M.M.D. Removal of methylene blue from aqueous solution by peat. J. Hazard. Mater. **2007**, *144,* 412–419.
92. Jones, D.A.; Lelyveld, T.P.; Mavrofidis, S.D.; Kingman, S.W.; Miles, N.J. Microwave heating applications in environmental engineering—A review. Resour. Conserv. Recycl. **2002**, *34*, 75–90.
93. Marsh, H.; Rodnguez-Reinoso, F. *Activated Carbon;* Elsevier: Oxford, **2006**.
94. Ioannidou, O.; Zabaniotou, A. Agricultural residues as precursors for activated carbon production—A review. Renew. Sustainable Energy Rev. **2007**, *11*, 1966–2005.
95. Yin, C.Y.; Aroua, M.K.; Daud, W.M.A.W. Review of modifications of activated carbon for enhancing contaminant uptakes from aqueous solutions. Sep. Purif. Technol. **2007**, *32*, 403–415.
96. Paraskeva, P.; Kalderis, D.; Diamadopoulos, E. Production of activated carbon from agricultural by-products. J. Chem. Technol. Biotechnol. **2008**, *83,* 581–592.
97. Yuen, F.K.; Hameed, B.H. Recent developments in the preparation and regeneration of activated carbons by microwaves. Adv. Colloid Interface Sci. **2009**, *149,* 19–27.
98. Menèndez, J.A.; Arenillas, A.; Fidalgo, B.; Fernandez, Y.; Zubizarreta, L.; Calvo, E.G.; Bermùdez, J.M. Microwave heating processes involving carbon materials. Fuel Process. Technol. **2010**, *91,* 1–8.
99. Laszlo, K.; Podkoscielny, P.; Dabrowski, A. Heterogeneity of polymer-based active carbons in adsorption of aqueous solutions of phenol and 2,3,4-trichlorophenol. Langmuir **2003**, *19,* 5287–5294.
100. Podkoscielny, P.; Dabrowski, A.; Marijuk, O.V. Heterogeneity of active carbons in adsorption of phenol aqueous solutions. Appl. Surf. Sci. **2003**, *205,* 297–303.
101. Leòn Y Leòn, C.A.; Radovic, L.R. Interfacial chemistry and electrochemistry of carbon surfaces. In *Chemistry and Physics of Carbon;* Thrower, P.A., Ed.; Marcel Dekker: New York, 1720; Vol. 24, 214–310.
102. Boehm, H.P. Surface oxides on carbon and their analysis: A critical assessment. Carbon **2002**, *40*, 145–149.
103. Gabaldon, C.; Marzal, P.; Seco, A.; Gonzalez, J.A. Cadmium and copper removal by a granular activated carbonin laboratory column systems. Sep. Sci. Technol. **1720**, *35,* 1039–1053.
104. Carrott, P.J.M.; Ribeiro Carrott, M.M.L.; Nabais, J.M.V.; Ramalho, J.P.P. Influence of surface ionization on the adsorption of aqueous zinc species by activated carbons. Carbon **1720**, *35*, 403–410.
105. Carrott, P.J.M.; Mourao, P.A.M.; Ribeiro Carrott, M.M.L.; Goncalves, E.M. Separating surface and solvent effects and the notion of critical adsorption energy in the adsorption of phenolic compounds by activated carbons. Langmuir **1720**, *21,* 11863–11869.
106. Hu, J.; Aizawa, T.; Ookubo, Y.; Morita, T.; Magara, Y. Adsorptive characteristics of ionogenic aromatic pesticides in water on powdered activated carbon. Water Res. **1720**, *32,* 2593–2600.
107. Urano, K.; Yamamoto, E.; Tonegawa, M.; Fujie, K. Adsorption of chlorinated organic compounds on activated carbon from water. Water Res. **1991**, *25*, 1459–1464.
108. Malhas, A.N.; Abuknesha, R.A.; Price, R.G. Removal of detergents from protein extracts using activated charcoal prior to immunological analysis. J. Immunol. Methods **2002**, *264,* 37–43.
109. Babel, S.; Kurniawan, T.A. Cr(VI) removal from synthetic wastewater using coconut shell charcoal and commercial activated carbon modified with oxidizing agents and/or chi- tosan. Chemosphere **2004**, *54,* 951–967.
110. Kim, D.S. Adsorption characteristics of Fe(III) and Fe(III)–NTA complex on granular activated carbon. J. Hazard. Mater. **2004**, *106,* 67–84.
111. Satapathy, D.; Natarajan, G.S. Potassium bromate modification of the granular activated carbon and its effect on nickel adsorption. Adsorption **2006**, *12,* 147–154.
112. Richard, D.; Delgado Nunez, M.L.; Schweich, D. Adsorption of complex phenolic compounds on active charcoal: adsorption capacity and isotherms. Chem. Eng. J. **2010**, *158,* 213–219.
113. Figaro, S.; Louisy-Louis, S.; Lambert, J.; Ehrhardt, J.J.; Ouensanga, A.; Gaspard, S. Adsorption studies of recalcitrant compounds of molasses spentwash on activated carbons. Water Res. **2006**, *40,* 3456–3466.

114. Mohan, S.V.; Karthikeyan, J. Removal of lignin and tannin colour from aqueous solution by adsorption onto activated charcoal. Environ. Pollut. **1997**, *97*, 183–187.
115. Kalderis, D.; Koutoulakis, D.; Paraskeva, P.; Diamadopou- los, E.; Otal, E.; Valle, J.O.; Fernandez-Pereira, C. Adsorption of polluting substances on activated carbons prepared from rice husk and sugarcane bagasse. Chem. Eng. J. **2008**, *144*, 42–50.
116. Lang, W.; Dejma, C.; Sirisansaneeyakul, S.; Sakairi, N. Biosorption of nonylphenol on dead biomass of *Rhizo- pus arrhizus* encapsulated in chitosan beads. Bioresour. Technol. **2009**, *100*, 5616–5623.
117. Pinnavaia, T. Intercalated clay catalysts. J. Sci. **1983**, *220*, 365–371.
118. Bhattacharyya, K.G.; Gupta, S.S. Adsorption of a few heavy metals on natural and modified kaolinite and montmorillonite: A review. Adv. Colloid Interface Sci. **2008**, *140*, 114–131.
119. Mitchell, J.K. *Fundamentals of Soil Behavior,* 2nd Ed; Wiley: New York, 1993.
120. Gupta, S.S.; Bhattacharyya, K.G. Removal of Cd(II) from aqueous solution by kaolinite, montmorillonite and their poly(oxo zirconium) and tetrabutylammonium derivatives. J. Hazard. Mater. **2006**, *128*, 247–257.
121. Bhattacharyya, K.G.; Gupta S.S. Influence of acid activation of kaolinite and montmorillonite on their adsorptive removal of Cd(II) from water. Ind. Eng. Chem. Res. **2007**, *46*, 3734 –3742.
122. Bagane, M.; Guiza, S. Removal of a dye from textile effluents by adsorption. Ann. Chim. **2000**, *25*, 615–626.
123. Almeida, C.A.P.; Debacher, N.A.; Downs, A.J.; Cottet, L.; Mello, C.A.D. Removal of methylene blue from colored effluents by adsorption on montmorillonite clay. J. Colloid Interface Sci. **2009**, 332, 46–53.
124. Ghosh, D.; Bhattacharyya, K.G. Adsorption of methylene blue on kaolinite. Appl. Clay Sci. **2002**, *20*, 295–300.
125. Gupta, S.S.; Bhattacharyya, K.G. Removal of Cd(II) from aqueous solution by kaolinite, montmorillonite and their poly(oxo zirconium) and tetrabutylammonium derivatives. J. Hazard. Mater. **2006**, *128*, 247–257.
126. Bhattacharyya, K.G.; Gupta, S.S. Influence of acid activation on adsorption of Ni(II) and Cu(II) on kaolinite and montmorillonite: Kinetic and thermodynamic study. Chem. Eng. J. **2008**, *136*, 1–13.
127. Gupta, S.S.; Bhattacharyya, K.G. Adsorption of Ni(II) on clays. J. Colloid Interface Sci. **2006**, *295*, 21–32.
128. Bhattacharyya, K.G.; Gupta, S.S. Adsorption of chromium(VI) from water by CLAYS. Ind. Eng. Chem. Res. **2006**, *45*, 7232–7240.
129. Yavuz, O.; Altunkaynak, Y.; Guzel, F. Removal of copper, nickel, cobalt and manganese from aqueous solution by ka- olinite. Water Res. **2003**, *37*, 948–952.
130. Bhattacharyya, K.G.; Gupta, S.S. Adsorption of Co(II) from aqueous medium on natural and acid activated kaolinite and montmorillonite. Sep. Sci. Technol. **2007**, *42*, 3391–3418.
131. Gupta, S.S.; Bhattacharyya, K.G. Interaction of metal ions with clays: I. A case study with Pb(II). Appl. Clay Sci. **2005**, *30*, 199–208.
132. Bhattacharyya, K.G.; Gupta, S.S. Pb(II) uptake by kaolinite and montmorillonite in aqueous medium: Influence of acid activation of the clays. Colloids Surf., A **2006**, *277*, 191–200.
133. Bhattacharyya, K.G.; Gupta, S.S. Adsorption of Fe(III) from water by natural and acid activated clays: Studies on equilibrium isotherm, kinetics and thermodynamics of interactions. Adsorption **2006**, *12*, 185–204.
134. Bhattacharyya, K.G.; Gupta, S.S. Kaolinite, montmorillon- ite, and their modified derivatives as adsorbents for removal of Cu(II) from aqueous solution. Sep. Purif. Technol. **2006**, *50*, 388–397.
135. Townsend, R.P. Ion exchange in zeolites. In *Introduction to Zeolite Science and Practice;* Bekkum, H.V., Flannigen, E.M., Janmsen, J.C., Eds.; Elsevier: Amsterdam, 1991; 359–390.
136. Dyer, A. *An Introduction to Zeolite Molecular Sieves*; John Wiley and Sons: Chichester, 1988; Vol. 5–8.

137. Barer, R.M. *Zeolites and Clay Minerals as Sorbent and Molecular Sieves*; Academic Press: New York, 1978.
138. Kesraoui-Ouki, S.; Cheeseman, C.R.; Perry, R. Natural zeolite utilization in pollution control: A review of application to metal's effluents. J. Chem. Technol. Biotechnol. **1994**, *59,* 121–126.
139. Ghobarkar, H.; Schaf, O.; Guth, U. Zeolites from kitchen to space. Prog. Solid State Chem. **1999**, *27,* 29–73.
140. Dogan, M.; ALkan, M.; Onager, Y. Adsorption of methylene blue from aqueous solution onto perlite. Water Air Soil Pollut. **2000**, *120,* 229–248.
141. Dogan, M.; Alkan, M.; Turkyilmaz, A.; Ozdemir, Y. Kinetics and mechanism of removal of methylene blue by adsorption onto perlite. J. Hazard. Mater. **2004**, *109,* 141–148.
142. Chakrabarti, S.; Dutta, B.K. Note on the adsorption and diffusion of methylene blue in glass fibers. J. Colloid Interface Sci. **2005**, *286,* 807–811.
143. Kumar, M.N.V.R. A review of chitin and chitosan applications. React. Funct. Polym. **2000**, *46,* 1–27.
144. Guibal E. Interactions of metal ions with chitosan-based sorbents: A review. Sep. Purif. Technol. **2004**, *38,* 43–74.
145. Varma, A.J.; Deshpande, S.V.; Kennedy, J.F. Metal com- plexation by chitosan and its derivatives: A review. Carbo- hydr. Polym. **2004**, *55*, 77–93.
146. Gerente, C.; Lee, V.K.C.; Cloirec, P.L.; McKay, G. Application of chitosan for the removal of metals from wastewaters by adsorption—mechanisms and models review. Crit. Rev. Environ. Sci. Technol. **2007**, *37,* 41–127.
147. Crini, G.; Badot, P.-M. Application of chitosan, a natural aminopolysaccharide, for dye removal from aqueous solutions by adsorption processes using batch studies: A review of recent literature. Prog. Polym. Sci. **2008**, *33*, 399–447.
148. Crini, G. Recent developments in polysaccharides-based materials used as adsorbents in wastewater treatment. Prog. Polym. Sci. **2005**, *30,* 38–70.
149. Bhatnagar, A.; Sillanpää, M. Applications of chitin- and chitosan-derivatives for the detoxification of water and wastewater—A short review. Adv. Colloid Interface Sci. **2009**, *152,* 26–38.
150. Jeon, C.; Höll, W.H. Chemical modification of chitosan and equilibrium study for mercury ion removal. Water Res. **2003**, *37,* 4770–4780.
151. Rorrer, G.L.; Hsien, T.-Y.; Way, J.D. Synthesis of porous- magnetic chitosan beads for removal of cadmium ions from wastewater. Ind. Eng. Chem. Res. **1993**, *32,* 2170–2178.
152. Kalyani, S.; Priya, J.A.; Rao, P.S.; Krishnaiah, A. Removal of copper and nickel from aqueous solutions using chitosan coated on perlite as biosorbent. Sep. Sci. Technol. **2005**, *40*, 1483–1495.
153. Liu, X.; Hu, Q.; Fang, Z.; Zhang, X.; Zhang, B. Magnetic chitosan nanocomposite: A useful tool for heavy metal ion removal. Langmuir **2009**, *25*, 3–8.
154. Qu, R.; Sun, C.; Ma, F.; Zhang, Y.; Ji, C.; Xu, Q.; Wang, C.; Chen, H. Removal of recovery of Hg(II) from aqueous solution using chitosan-coated cotton fibers. J. Hazard. Mater. **2009**, *167,* 717–727.
155. Zhang, G.; Qu, R.; Sun, C.; Ji, C.; Chen, H.; Wang, C.; Niu, Y. Adsorption for metal ions of chitosan coated cotton fiber. J. Appl. Polym. Sci. **2008**, *110,* 2321–2327.
156. Huang, G.L.; Zhang, H.Y.; Jeffrey, X.S.; Tim, A.G.L. Adsorption of chromium(VI) from aqueous solutions using cross-linked magnetic chitosan beads. Ind. Eng. Chem. Res., **2009**, *48,* 2646–2651.
157. Sun, X.Q.; Peng, B.; Jing, Y.; Chen, J.; Li, D.Q. Chitosan(chitin)/cellulose composite biosorbents prepared using ionic liquid for heavy metal ions adsorption. Separations **2009**, *55*, 2062–2069.
158. Shameem, H.; Abburi, K.; Tushar, K.G.; Dabir, S.V.; Veera, M.B.; Edgar, D.S. Adsorption of chromium(VI) on chitosan-coated perlite. Sep. Sci. Technol. **2003**, *38*, 3775–3793.
159. Veera, M.B.; Krishnaiah, A.; Jonathan, L.T.; Edgar, D.S. Removal of hexavalent chromium from wastewater using a new composite chitosan biosorbent. Environ. Sci. Technol. **2003**, *37,* 4449–4456.

160. Tran, H.V.; Tran, L.D.; Nguyen, T.N. Preparation of chi- tosan/magnetite composite beads and their application for removal of Pb(II) and Ni(II) from aqueous solution. Mater. Sci. Eng., C **2010**, *30,* 304–310.
161. Wan, M.W.; Kan, C.C.; Buenda, D.R.; Maria, L.P.D. Adsorption of copper(II) and lead(II) ions from aqueous solution on chitosan-coated sand. Carbohydr. Polym. **2010**, *80,* 891–899.
162. Srinivasa, R.P.; Vijaya, Y.; Veera, M.B.; Krishnaiah, A. Adsorptive removal of copper and nickel ions from water using chitosan coated PVC beads. Bioresour. Technol. **2009**, *100,* 194–199.
163. Vijaya, Y.; Srinivasa, R.P.; Veera, M.B.; Krishnaiah, A. Modified chitosan and calcium alginate biopolymer sorbents for removal of nickel (II) through adsorption. Carbohydr. Polym. **2008**, *72*, 261–271.
164. Li, J.M.; Meng, X.G.; Hu, C.W.; Du, J. Adsorption of phenol,*p*-chlorophenol and *p*-nitrophenol onto functional chitosan. Bioresour. Technol. **2009**, *100,* 1168–1173.
165. Nadavala, S.K.; Swayampakula, K.; Boddu, V.M.; Abburi, K. Biosorption of phenol and o-chlorophenol from aqueous solutions on to chitosan-calcium alginate blended beads. J. Hazard. Mater. **2009**, *162,* 482–489.
166. Pagnanelli, F.; Mainelli, S.; Veglio, F.; Toro, L. Heavy metal removal by olive pomace: biosorbent characterization and equilibrium modeling. Chem. Eng. Sci. **2003**, *58,* 4709–4717.
167. Demirbas, A. Mechanisms of liquefaction and pyrolysis reactions of biomass. Energy Convers. Manage. **2000**, *41,* 633–646.
168. Theander, O. Cellulose, Hemicellulose, and Extractives. In *Fundamentals of Thermochemical Biomass Conversion;* Overand, R.P., Mile, T.A., Mudge, L.K., Eds.; Elsevier Applied Science Publisher: New York, **1985**; 35–60.
169. Morita, M.; Higuchi, M.; Sakata, I. Binding of heavy metal ions by chemically modified Woods. J. Appl. Polym. Sci. **1987**, *34,* 1013–1023.
170. Sawamiappan, N.; Krishnamoorthy, S. Phenol formaldehyde cationic matrices substitutes by bagasse-charcoal. Res. Ind. **1984**, *29,* 293–297.
171. Odozi, T.O.; Okeke, S.; Lartey, L.B. Studies on binding metal ions with polymerized corncob and a composite resin with sawdust and onion. Agric. Waste **1985**, *12,* 13–21.
172. Gardea-Torresdey, J.L.; de la Rosa, G.; Peralta-Videa, J.R. Use of phytofiltration technologies in the removal of heavy metals: a review. Pure Appl. Chem. **2004**, *76,* 801–813.
173. Volesky, B. Detoxification of metal-bearing effluents: Biosorption for the next century. Hydrometallurgy **2001**, *59,* 203–216.
174. Brown, P.A.; Gill, S.A.; Allen, S.J. Metal removal from wastewater using peat. Water Res. **2000**, *34,* 3907–3916.
175. Nobel, P. *Physicochemical and Environmental Plant Physiology*; Academic Press: New York, 1991.
176. Rowell, R.M. *In Removal of Metal Ions from Contaminated Water Using Agricultural Residues, Proceedings of* ECOWOOD 2006 2nd International Conference on Environmentally Compatible Forest Products, Fernando Pessoa University, Oporto, Portugal, Sept. 20–22, 2006.
177. Mantanis, G.I.; Young, R.A.; Rowell, R.M. Swelling of wood. Part III. Effect of temperature and extractives on rate and maximum swelling. Holzforschung **1995**, *49,* 239–248.
178. Dahiya, S.; Tripathi, R.M.; Hegde, A.G. Biosorption of lead and copper from aqueous solutions by pre-treated crab and arca shell biomass. Biores. Technol. **2008**, *99,* 179–187.
179. Hashem, A.; Abdel-Halim, E.; Maauof, H.A.; Ramadan, M.A.; Abo-Okeil, A. Treatment of sawdust with polyamine for wastewater treatment. Energy Edu. Sci. Technol. **2007**, *19,* 45–58.
180. Demirbas, A. Heavy metal adsorption onto agro-based waste materials: A review. J. Hazard. Mater. **2008**, *157,* 220–229.
181. Min, S.H.; Han, J.S.; Shin, E.W.; Park, J.K. Improvement of cadmium ion removal by base treatment of juniper fiber. Water Res. **2004**, *38,* 1289–1295.

182. Dakiky, M.; Khamis, M.; Manassra, A.; Mer'eb, M. Selective adsorption of chromium(VI) in industrial wastewater using low-cost abundantly available adsorbents. Adv. Environ. Res. **2002**, *6,* 533–540.
183. Bhatnagar, A.; Jain, A.K.; Minocha, A.K.; Singh, S. Removal of lead ions from aqueous solutions by different types of industrial waste materials: equilibrium and kinetic studies. Sep. Sci. Technol. **2006**, *41,* 1881–1892.
184. Taty-Costodes, V.C.; Fauduet, H.; Porte, C.; Delacroix, A. Removal of Cd(II) and Pb(II) ions, from aqueous solutions, by adsorption onto sawdust of *Pinus sylvestris*. J. Hazard. Mater. **2003**, *105,* 121–142.
185. Nassar, M.M. Adsorption of Fe^{+3} and Mn^{+2} from ground water onto maize cobs using batch adsorber and fixed bed column. Sep. Sci. Technol. **2006**, *41,* 943–959.
186. Ho, Y.S.; Huang, C.T.; Huang, H.W. Equilibrium sorption isotherm for metal ions on tree fern. Process Biochem. **2002**, *37,* 1421–1430.
187. Chen, B.; Hui, C.W.; McKay, G. Film-pore diffusion modeling for the sorption of metal ions from aqueous effluents onto peat. Water Res. **2001**, *35*, 3345–3356.
188. Sathishkumar, M.; Vijayaraghavan, K.; Binupriya, A.R.; Stephan, A.M.; Choi, J.G.; Yun, S.E. Porogen effect on characteristics of banana pith carbon and the sorption of dichlorophenols. J. Colloid Interface Sci. **2008**, *320,* 22–29.
189. Achak, M.; Hafidi, A.; Ouazzani, N.; Sayadi, S.; Mandi, L. Low cost biosorbent "banana peel" for the removal of phenolic compounds from olive mill wastewater: Kinetic and equilibrium studies. J. Hazard. Mater. **2009**, *166,* 117–125.
190. Park, K.H.; Balathanigaimani, M.S.; Shim, W.G.; Lee, J.W.; Moon, H. Adsorption characteristics of phenol on novel corn grain-based activated carbons. Microporous Mesoporous Mater. **2010**, *127,* 1–8.
191. Navarro, A.E.; Cuizano, N.A.; Lazo, J.C.; Sun-Kou, M.R.; Llanos, B.P. Comparative study of the removal of phenolic compounds by biological and non-biological adsorbents. J. Hazard. Mater. **2009**, *164,* 1439–1446.
192. Bhatnagar, A.; Minocha, A.K. Adsorptive removal of 2,4- dichlorophenol from water utilizing *Punica granatum* peel waste and stabilization with cement. J. Hazard. Mater. **2009**, *168,* 1111–1117.
193. Jiuhui, Q.U. Research progress of novel adsorption processes in water purification: A review. J. Environ. Sci. **2008**, *20,* 1–13.
194. Oliveira, L.C.A.; Petkowicz, D.I.; Smaniotto, A.; Pergher, S.B.C. Magnetic zeolites: A new adsorbent for removal of metallic contaminants from water. Water Res. **2004**, *38,* 3699–3704.
195. Gu, Z.; Fang, J.; Deng, B.L. Preparation and evaluation of GAC-based iron-containing adsorbents for arsenic removal. Environ. Sci. Technol. **2005**, *39,* 3833–3843.
196. Machado, L.C.R.; Lima, F.W.J.; Paniago, R. Polymer coated vermiculite–iron composites: Novel floatable magnetic adsorbents for water spilled contaminants. Appl. Clay Sci. **2006**, *31,* 207–215.
197. Zhang, G.S.; Qu, J.H.; Liu, H.J.; Liu, R.P.; Wu, R.C. Preparation and evaluation of a novel Fe-Mn binary oxide adsorbent for effective arsenite removal. Water Res. **2007**, *41,* 1921–1928.

40

Cadmium and Lead: Contamination

Gabriella Kakonyi
and Imad A.M.
Ahmed

Introduction

This entry summarizes the effects of cadmium (Cd) and lead (Pb) on the environment and on human health. The chemistry of Cd and Pb is first discussed to establish an understanding of their behavior, occurrence, and fate in the environment. This helps foresee toxicity and the possible natural attenuation or remediation strategies. Both Cd and Pb are widely produced and used. Therefore, awareness of the sources from where humans and the environment are exposed is described in some detail along with their epidemiology and regulatory measures taken to reduce or prevent their release into the environment. A brief section is also dedicated to the detection and analysis of Cd and Pb and their compounds within environmental and clinical samples.

Chemical Properties

Cadmium (Cd) and lead (Pb) are transition metals that have no known vital or beneficial role in the human body or health. Cadmium (atomic number 48, atomic mass 112.4 g mol^{-1}) belongs to Group 12 of the periodic table, possessing somewhat similar chemical properties as zinc and mercury. Lead (atomic number 82, atomic mass 207.2g mol^{-1}) is a member of Group 14 in the periodic table, which also includes C, Si, Ge, and Sn.[1] The most predominant oxidation state of Cd and Pb under normal environmental conditions of temperature and pressure is +2. In organolead chemistry, the oxidation state (+4) of Pb is remarkably dominant.[2] Both Cd and Pb are considered to form stable oxidation states as divalent Cd^{2+} and Pb^{2+} ions in inorganic compounds. Their biological toxicity appears to be determined by their availability for ligand exchange and chelation properties.[3–6] The chemical similarity of Cd and Pb to certain alkaline earth metals such as calcium; their ability to form highly insoluble inorganic salts (e.g., phosphate, carbonate, sulfate), organometallic complexes, or free hydrated ions; and their increased affinity to biological donors (e.g., proteins) play an important role in their transport in the environment and toxicity in biological systems.[2,7,8]

Occurrences

Cadmium and lead are trace elements in rocks and soils,[9] where the concentration of Cd is approximately 0.1 ppm and the concentration of Pb is about tenfold higher.[10,11] These concentrations are relatively low, and the presence of both Cd and Pb are normally associated with other more abundant elements. Therefore, both metals are inevitable by-products from the mining of zinc and copper ores. Lead and cadmium are chalcophilic elements, meaning that they have a tendency to form sulfide minerals.[12] Thus, the most abundant Cd and Pb minerals are greenockite (CdS) and galena (PbS). Most Cd is found associated with zinc ores such as sphalerite (ZnS) in which the Cd content is 0.5% of the Zn content.[11] Other important minerals of Pb are crocite ($PbCrO_4$), anglesite $PbSO_4$, massicot (PbO), and cerussite ($PbCO_3$).[13] Native (i.e., elemental) Cd and Pb are rare in the environment. Naturally occurring metallic Cd has been found in the Vilyuy River bedrock in Siberia.[14] Cadmium and lead are also found associated with clay and carbonate minerals. The Cd^{2+} ion has an ionic size of 95 pm close to that of Ca^{2+} (100 pm), whereas Pb^{2+} has an ionic size (119 pm) between K^+ (138 pm) and Ca^{2+}. Thus, during the formation of secondary minerals such as feldspars, mica, apatite, and calcite, major cations such as Ca^{2+} and K^+ can be substituted by Cd^{2+} and Pb^{2+} or other trace metal ions of the same charge, sign, or of comparable ionic size.[15,16]

Production and Uses

Cadmium

Cadmium is widely used in a number of industrial applications, the largest area being the production of nickel-Cd batteries.[17,18] Furthermore, Cd is increasingly used in solar panels, and still commonly used in pigments, coatings, corrosion-resistant plating, photography, as a fungicide and as a stabilizer and softener in plastics. It is also contained in coal and in rocks mined to produce phosphate fertilizers.[19] Cadmium atoms has the ability to absorb neutrons without fission or splitting, so it is used in nuclear reactor components such as control rods as a shield of neutrons and to control nuclear fission.[20]

Lead

The greatest use of Pb is lead acid batteries, which have been used extensively in automobiles since 1918 or so.[21] Lead is also commonly added to paint as Pb-chromate or Pb-carbonate to speed drying, increase durability, retain a fresh appearance, and resist moisture. Because elemental Pb has a low melting temperature (327°C), it enables easy casting and shaping and thus is commonly used in building constructions and joining metallic parts. Lead is the traditional base metal for domestic water pipes. One if its major uses is as a radiation shield in the glass of television and computer screens, and as a protecting shield from radioactive radiation such as x-ray and γ-ray in scientific and industrial instruments. Other uses are in infrared detectors, sheeting, cables, solders, Pb-crystal glassware, ammunitions, and as weight in sport equipment.

Detection

Because Cd is used as pigment and softener in plastics, the use of colored lids, tubes and certain plastic containers should be avoided during sampling and storage of environmental and toxicological samples.[22] It is well documented that significant amounts of trace metals can be lost on the walls of glass and some plastic containers and adsorption of Pb occurs on the walls of Pyrex, polypropylene and polyethylene containers. The loss of aqueous Cd and Pb by adsorption onto the wall of containers was not observed using Teflon. Therefore, particular attention must be paid in the collection, treatment and preservations of environmental samples, especially when Cd and Pb are present at submicromolar concentrations. Procedures required for total metal analysis are normally straight forward and state the acidification of samples to $pH < 1$ and preservation in inert containers such as Teflon. Furthermore, accurate determination of cadmium concentrations within sensitive samples may not be done by a smoking person in order to avoid cross-contamination.[23] Free cadmium in solution (Cd^{2+}) presents in water below pH values of 8. The highest concentration was measured to be 6 mM without hydroxide formation. Above pH 8, cadmium is expected to precipitate and hinder accurate determination of the concentration.[24] Removal of cadmium and lead from aerosol and fly ash samples is discussed in details by Lum[25] and Hlavay et al.[26] Total aqueous Pb and Cd concentrations can be measured using atomic absorption spectroscopy or inductively coupled plasma spectrometry, the former detecting ion concentrations at mg L^{-1} (ppm) levels and the latter at ng L^{-1} (ppt). Organometallic compounds of both Pb and Cd can be measured by inductively coupled plasma mass spectrometry, atomic absorption spectrometry, high- performance liquid chromatography, and gas chromatography analytical methods.[27–32]

Environmental Levels

Cadmium and lead are released into the environment as a result of both industrial activities and natural processes. The main natural sources include the weathering of rocks releasing Cd and Pb into the hydrosphere, and volcanic activity increasing the atmospheric Cd and Pb concentration.[15] The eruption of Mount Pinatubo located on the Philippine island of Luzon is an example of the dramatic effects of volcanism on the distribution of elements in the lithosphere. During just 2 days in June 1991, Pinatubo ejected 10 billion metric tonnes of magma and 20 million tonnes of SO_2; the resulting aerosols influenced global climate for 3 years. This single event introduced an estimated 100,000 tonnes of Pb and 1000 tonnes of Cd to the surface environment.[15,33] From anthropogenic sources, humans are exposed to Cd through the atmosphere, hydrosphere, and the geosphere. Human uptake of Cd can happen through inhalation of air, soil, or dust containing fine Cd particles. Cadmium is introduced into the atmosphere through the combustion of fossil fuel (e.g., coal), municipal solid waste incineration, mineral smelting and as dust generated by recycling scrap iron and steel. The contamination of soils is apparent from zinc and phosphate ores and where Cd-containing phosphate fertilizers are used in agriculture.[34–37] Zinc and cadmium are mineralogically and geochemically linked to each other. Thus, large emissions of fumes containing both ZnO and CdO are produced from zinc smelters. However, these fumes are normally enriched in CdO because of its higher volatility in comparison to ZnO. It is well documented that phosphate fertilizers constitute a very diffuse source of Cd contamination. The quantity of Cd contained in a phosphate fertilizer depends on the source of the phosphate rock used in making it. In general, Cd content in phosphate fertilizers vary from 1–2 ppm for tertiery Ca-phosphate to 50–170 ppm for superphosphate. As a result, Cd is transported to aquatic environments, plants, animals, and finally to humans. Soil organic matter strongly adsorbs Cd, and acidic soils further enhance the Cd uptake by plants. Cadmium is also known to accumulate in aquatic organisms, especially in freshwaters. However, Chen et al.[38] reported that the addition of potassium fertilizers effectively reduce the phytoavailability of both Cd and Pb within soils.

Although Pb is a naturally occurring element, its major cycle in the environment is anthropogenically driven and a result of human activities. For example, Pb in the form of tetraethyl-lead has been in usage as additive in gasoline to avoid knocking effect of autoignition since the early 20th century.[39] Through this application, Pb has been released into the environment in the forms of Pb-chloride, -bromide, and -oxides from the car exhaust. Lead is also introduced into the environment from coal and solid waste combustion, a wide number of industrial and mining processes, and drinking water pipes containing Pb. When copper is present in (either from soldering or as contamination) a Pb-containing pipe, it accelerates corrosion of the Pb pipe by galvanic action. The water flowing in contact with these dissimilar metals serves as the electrolyte. Based on the electrochemical series, metallic Pb serves as the anode of this galvanic cell and is therefore oxidized (i.e., corroded) to form Pb^{2+} ions contaminating the drinking water. Lowering pH in water also promotes the dissolution and aqueous transport of Pb^{2+}. Evidences of galvanic corrosion of domestic Pb pipes had been given and discussed elsewhere.[40,41] Large Pb-containing particles released to the atmosphere settle quickly on the ground and then washed into soils or dissolve in aquatic phases, while the very fine particles remain in the atmosphere travelling long distances and fall back to the surface with rain. This cycle has caused an unnatural and extensive sequence exposing plants, aquatic environments, and humans to dangerous concentrations of Pb. Therefore, in 1978, it was forbidden for all European Union-member states to produce, import, or sell gasoline with more than 0.4 g Pb L^{-1}. Starting from the year 2000, the marketing of leaded petrol has been banned in Europe following Directive 98/70/EC and related acts of the European Parliament. In the United States, Pb was banned as fuel additive starting from 1996.[12,42–44] Unfortunately, the gasoline additives tri- and dialkyl Pb are stable compounds and are persistent in the environment; thus, the restriction in usage does not necessarily decrease contamination.[45–51] When Pb accumulates in living organisms, it becomes part of the food chain thereby creating further sources for human exposure. Beyond bioaccumulation, once released into surface waters, both Cd and Pb are deposited into the sediment, increasing the metal contamination by 10–100 times near the fallout areas. The type of sediment affects the severity of contamination, with carbonaceous, anoxic, and clay sediments being the most prone to high concentrations of metal uptake, while siltstones, shales, sandstones, limestones and marine evaporates are normally less affected.[52]

Toxicological Effects

The International Agency for Research on Cancer (IARC) has classified Cd and Pb compounds as carcinogenic to humans.[87] This classification has been based mainly on epidemiological evidences of renal damage in rats and mice. The greatest concerns about the health effects of Cd and Pb arise from their tendency to form strong complexes, replacing essential elements (e.g., Ca^{2+}) and bioaccumulating in the human body. There is no known small enough amount of Pb uptake that would cause no harm to human beings. However, Reichlmayr-Lais et al.[53,54] found that the depletion of Pb resulted in hematological changes in rats. The concentration of Cd is approximately 0.4 pg kg^{-1}in a daily diet, which is said to be 10 times lower than the amount that can cause kidney damage.[55] The concentrations of Cd in human blood is usually between 0.1 and 2 pg L^{-1} and in urine <1 pg L^{-1}.[56] The concentration of Pb in human blood varies between 165 and 296 pg L^{-1}. The largest known catastrophe caused by Cd toxicity was identified in Japan in the 1940s, and it is referred to as the "itai-itai" disease, which is literary translated from Japanese to "ouch ouch" disease. Itai-itai refers to a syndrome that principally consists of a painful skeletal condition resulting from weak and deformed bones. The patients of itai-itai suffer from renal anemia, tubular nephropathy, and osteopenic osteomalacia, while 90% of the patients are postmenopausal women. The residents of the Jinzu River basin region were first exposed to Cd in the 1930s as a result of industrial contamination from nearby intensive mining activities, which caused serious pollution of the local river waters. This resulted in high Cd contaminations of rice fields. With rice being the principal dietary component, especially in rural Japan, and the bioaccumulating properties of Cd, residents were exposed to very high levels of Cd causing irreversible damage and poisoning.[57,58] Toxicity of Cd in living organisms occurs because of the substitution of essential elements such as Zn at the reactive centers of essential enzymes, which disrupts a wide range of metabolic functions. Further details on mechanisms of Cd toxicity are available elsewhere.[57,59–62]

Cadmium is taken into the body through food, drinking water, smoking, and particulate matter in air, especially near Cd-processing industrial fields and hazardous waste sites. Gastrointestinal absorption from contaminated water or food is the main source of internally deposited Cd in the general population. Only a small proportion of ingested Cd is transferred to the bloodstream while the unabsorbed

Cd is excreted in the feces. The absorbed Cd binds to macromolecules, enzymes, and proteins, and the majority of it is being deposited in the liver and kidney. The Cd taken up by the kidney interferes with the filtering mechanism, causing the excretion of essential sugars and proteins, and damages the kidney's ability to remove acid from the blood. Another extremely painful effect of Cd is the softening of bones and decreasing their mineral density, resulting in fractures and paralysis in advanced cases. Furthermore, Cd damage of the central nervous system and psychological disorders were reported, as well as weakening of the immune system, reproductive failure, DNA damage, and the development of cancer.[63]

Lead can enter the human body through food, liquid, and air. Inorganic Pb compounds are known to pass through the skin.[64] It has been shown that the pollution of air with Pb particles from burning of fuel affects the cognitive behavior of children and adults living near busy roads or exposed to Pb contamination. Early experiments were carried out to study the dose-response to environmental levels of Pb. A comprehensive review of these studies can be found in U.S. Environmental Protection Agency (USEPA) 2005.[65] The absorbed Pb by the human body is distributed in blood, soft tissues, and in particular in bones and liver. Because Pb^{2+} can replace Ca^{2+}, ~9% of absorbed Pb ends up in the bones and teeth. The excess of Pb may cause several health effects, including damage to nervous system, chronic renal disease, anemia due to the inhibition of haem formation, damage of nervous system of unborn children, acute encephalopathy in young children, carcinogenicity and genotoxicity, and impaired reproductivity.[66–68] Even in low concentrations (1–10 pg dL^{-1} blood), Pb decrements neurocognitive abilities, intelligence measures, and perceptual-motor coordination. Lead toxicity of animals has been also studied extensively. For example, cattle showed poisoning symptoms within 6–8 wk when fed Pb-acetate at 7 mg Pb per kilogram body weight per day.[69–71] Accumulation of Pb in the liver and kidney was noted in calves fed 100 mg Pb per kilogram as Pb chromate for 100 days.[72] While the absorbed Cd and Pb could be excreted in urine, the daily excretion is <1% of the total body burden of Cd and Pb giving a biological half-life of Cd of more than 25 years[73,74] and 20–30 years for Pb in the skeleton.[75]

Cadmium and lead toxicity in animals is a major problem, especially in diary animals as Cd and Pb accumulate in the kidney, liver, and reproductive organs.[88] Many plants species are tolerant to certain amounts of heavy metals, which is likely to be achieved via metal-binding by specific proteins. Like all living organisms, at elevated concentrations of certain heavy metal plants start to show symptoms of toxicity. For example, Cd phytotoxicity can be identified in the form of stunting and chlorosis. Chlorosis is due to Cd interaction with foliar iron. A number of reports had shown evidences to Pb adverse effects on the growth and photosynthesis processes of plants. High concentration of Pb in soils is known to inhibit seed germination in a number of plant species and to induce abnormal morphologies.[89–91] It is interesting to note that a number of reports have shown that Cd accumulates in greater concentrations roots, tubers or leaves of plants. This means that Cd can present at higher concentrations in leafy and root vegetables than in fruits or grains.[91]

Regulations and Control

In the European Union, Cd and Pb are on the list of the six hazardous substances that are banned in the manufacturing of various electrical and electronic components. This is enforced by the Restriction on Hazardous Substances Directive (RoHS, 2002/95/EC) requiring that the maximum concentration of Cd or Pb may not exceed 100 or 1000 ppm per weight of homogeneous material, respectively. The production, recycling, and disposal of batteries and accumulators are regulated by the 2006/66/EC directive. These metals are also regulated under the Registration, Evaluation, Authorisation, and Restriction of Chemicals (REACH, 1907/2006) by the European Commission. In the United States, the USEPA is the main regulatory body restricting and controlling the use of Cd and Pb. Generally, the legislations are separated by pollution of the atmosphere, geosphere, hydrosphere, or specific industrial activities. Furthermore, the World Health Organization (WHO) globally regulates the acceptable exposure levels and concentration limits in public areas (Table 1). The maximum permissible concentrations of Cd and

TABLE 1 General Human Exposure to Cadmium and Lead

	Food	Water	Air
Cadmium	25 pg kg^{-1} body weight per month	3 pg L^{-1}	5 ng m^{-3} a^{-1}
Lead	0.8 pg kg^{-1} body weight per day	3.8–10 pg $L^{-1}day^{-1}$	0.5–4 pg $m^{-3}day^{-1}$

Source: Data from WHO.[83,84]

TABLE 2 Drinking Water Regulations for Maximum Permissible Contaminant Levels (USEPA, National Primary Drinking Water Regulations)

	USEPAa (mg L–1)	WHO Guideline Value	DWI-DEFRA (UK)
Cadmium	5.0	3.0	5.0
Lead	zero action level: 0.015	10	25[b]

[a] Maximum contaminant level goal.

[b] Maximum concentration is 25 mg L–1 until December 25, 2013, and 10 mg L–1 from December 25, 2012.

Source: Data from WHO[84,85] and DWI.[86]

Pb in drinking water determined by the USEPA, WHO, and Drinking Water Inspectorate-Department for Environment, Food and Rural Affairs (DWI-DEFRA, United Kingdom) is shown in Table 2. As a result of these regulations, the release of Cd and Pb into the environment and their human exposure has been lowered through recycling and safer, more conscious industrial processes. However, the need for such metals and the global consumption still persists. On an annual basis, 9.6 million tonnes of Pb and 19,000 tonnes of Cd are produced worldwide. To reduce and control exposure of the population through drinking water, regulations require the use of Pb-free (<0.2% Pb) pipe systems and continuous monitoring (USEPA). Cadmium and lead can be removed from water by reverse osmosis or ion exchange resins. For the treatment of groundwater and soil media, *in situ* precipitation techniques present a viable way to reduce the mobility of heavy metals. In the case of both Cd- and Pb phosphate-containing minerals (e.g., hydroxyapatite), calcium-carbonates and zeolites have been suggested as possible solutions.[76–82]

Conclusion

Both Cd and Pb have played prominent roles in the industrial revolution and subsequent centuries, and are now included in a vast range of products. Both elements have also played important, -but highly contrasting- roles in terms of human, animal, and plant health. The increased environmental concentration of Cd and Pb poses irreversible effects on nature and the human body. The contamination problem of heavy metals -including Cd and Pb- remains a challenging issue to scientists and engineers. The question whether the dangers associated with Cd and Pb could be avoided or could be further lowered remains open. Production of Cd and Pb has grown to be a global need, which can only be compensated by the development and use of alternative technologies, advanced remediative solutions and preventive measures, sensitive monitoring systems, and the avoidance of human and animal exposures.

References

1. Greenwood, N.N.; Earnshaw, A. *Chemistry of the Elements*; Elsevier: Oxford, 1997.
2. Grant, L.D. Lead and compounds. In *Environmental Toxicants;* Lippmann, M., Ed.; John Wiley and Sons, Inc.: Hoboken, NJ, 2009; 757–809.
3. Ahamed, M.; Siddiqui, M.K.J. Environmental lead toxicity and nutritional factors. Clin. Nutr. **2007**, *26* (4), 400–408.
4. Shukla, G.S.; Singhal, R.L. The present status of biological effects of toxic metals in the environment: Lead, cadmium, and manganese. Can. J. Physiol. Pharmacol. **1984**, *62* (8), 1015–1031.

5. Andersen, O. Chelation of cadmium. Environ. Health Per- spect. **1984**, *54,* 249–266.
6. Deagen, J.T.; Oh, S.H.; Whanger, P.D. Biological function of metallothionein. VI. Metabolic interaction of cadmium and zinc in rats. Biol. Trace Elem. Res. **1980**, *2* (1), 65–80.
7. Huheey, J. *Inorganic Chemistry: Principles of Structure and Reactivity,* 4th Ed.; Harper Collins College Publishers: New York, 1993.
8. Lindsay, W.L. *Chemical Equilibria in Soils;* The Black Burn Press: Caldwell, NJ, 2001.
9. Turekian, K.K.; Wedepohl, K.H. Distribution of the elements in some major units of the earth's crust. Geol. Soc. Am. Bull. **1961**, *72* (2), 175–192.
10. Cox, P.A. *The Elements: Their Origin, Abundance, and Distribution;* Oxford University Press: Oxford, U.K., 1989.
11. Halka, M.; Nordstrom, B. *Periodic Table of the Elements: Transition Metals;* Facts on File: New York, 2011.
12. Kummer, U.; Pacyna, J.; Pacyna, E.; Friedrich, R. Assessment of heavy metal releases from the use phase of road transport in Europe. Atmos. Environ. **2009**, *43* (3), 640–647.
13. Halka, M.; Nordstrom, B. *Periodic Table of the Elements: Metals and Metalloids;* Facts On File: New York, 2011.
14. Fleischer, M.; Cabri, L.; Chao, G.; Pabst, A. New mineral names. Am. Mineral. **1980**, *65* (9–10), 1065–1070.
15. Garrett, R.G. Natural sources of metals to the environment. Hum. Ecol. Risk. Assess. **2000**, *6* (6), 945–963.
16. Sposito, G. *The Chemistry of Soils*; Oxford University Press, Inc.: Oxford, U.K., 1989.
17. Wiley-VCH. *Ullmann's Encyclopedia of Industrial Chemistry,* 7th Ed.; Wiley-VCH: Weinheim, Germany, 2011.
18. Tolcin, A. Cadmium. In *Minerals Yearbook Metals and Minerals 2009;* United States Government Printing Office, 2009.
19. Robert, U.A. Metals recycling: Economic and environmental implications. Resour. Conserv. Recycling **1997**, *21* (3), 145–173.
20. Harvey, T.; Thomas, B.; Mclellan, J.; Fremlin, J. Measurement of liver-cadmium concentrations in patients and industrial workers by neutron-activation analysis. Lancet **1975**, *305* (7919), 1269–1272.
21. Guberman, D.E. Lead. In *Minerals Yearbook Metals and Minerals 2009;* Guberman, D.E., Ed.; United States Government Printing Office: Reston, VA, 2009.
22. Sekaly, A.L.; Chakrabarti, C.; Back, M.; Gregoire, D.; Lu, J.Y.; Schroeder, W. Stability of dissolved metals in environmental aqueous samples: Rideau River surface water, rain and snow. Anal. Chim. Acta **1999**, *402* (1–2), 223–231.
23. Cornelis, R.; Caruso, J.; Crews, H.; Heumann K. *Handbook of Elemental Speciation I: Techniques and Methodology*; John Wiley and Sons, Ltd: Chichester, U.K., 2003.
24. Majidi, V.; Miller-Ihli, N.J. Potential sources of error in capillary electrophoresis-inductively coupled plasma mass spectrometry for chemical speciation. Analyst **1998**, *123* (5), 809–813.
25. Lum, K.; Betteridge, J.; Macdonald, R. The potential availability of P, Al, Cd, Co, Cr, Cu, Fe, Mn, Ni, Pb and Zn in urban particulate matter. Environ. Technol. Lett. **1982**, *3* (1–11), 57–62.
26. Hlavay, J.; Polyak, K.; Meszaros, E. Determination of the distribution of elements as a function of particle size in aerosol samples by sequential leaching. Analyst. **1998**, *123* (5), 859–863.
27. Bettmer, J.; Cammann, K. Transversely heated graphite atomizer-atomic absorption spectrometry (THGA AAS) in combination with flow injection analysis system-hydride generation (FIAS HG) as a reliable screening method for organolead compounds. Appl. Organomet. Chem. **1994**, *8* (7–8), 615–620.
28. Dunemann, L.; Hajimiragha, H.; Begerow, J. Simultaneous determination of Hg(II) and alkylated Hg, Pb, and Sn species in human body fluids using SPME-GC/MS-MS. Frese- nius J. Anal. Chem. **1999**, *363* (5–6), 466–468.

29. Fragueiro, M.S.; Alava-Moreno, F.; Lavilla, I; Bendicho, C. Determination of tetraethyl lead by solid phase microextraction-thermal desorption-quartz furnaceatomic absorption spectrometry. J. Anal. At. Spectrom. **2000**, *15* (6), 705–709.
30. Infante, H.G.; Sanchez, M.L.F.; Sanz-Medel, A. Vesicle- mediated high performance liquid chromatography coupled to hydride generation inductively coupled plasma mass spectrometry for cadmium speciation in fish cytosols. J. Anal. At. Spectrom. **2000**, *15* (5), 519–524.
31. Moens, L.; De Smaele, T.; Dams, R.; Van Den Broeck, P.; Sandra, P. Sensitive, simultaneous determination of organo- mercury, -lead, and -tin compounds with head-space solid phase microextraction capillary gas chromatography combined withinductively coupled plasma mass spectrometry. Anal. Chem. **1997**, *69* (8), 1604–1611.
32. Yu, X.; Yuan, H.; Gorecki, T.; Pawliszyn, J. Determination of lead in blood and urine by SPME/GC. Anal. Chem. **1999**, *71* (15), 2998–3002.
33. Selinus, O.; Finkelman, R.B.; Centeno, J.A. Human health and ecosystems. In *Geology and Ecosystems;* Zektser, I.S., Marker, B., Ridgway, J., Rogachevskaya, L., Vartanyan, G., Eds.; Springer US: Boston, 2006; 197–218.
34. Grant, C.A.; Sheppard, S.C. Fertilizer impacts on cadmium availability in agricultural soils and crops. Hum. Ecol. Risk Assess. **2008**, *14* (2), 210–228.
35. Jiao, Y.; Grant, C.A.; Bailey, L.D. Effects of phosphorus and zinc fertilizer on cadmium uptake and distribution in flax and durum wheat. J. Sci. Food Agric. **2004**, *84* (8), 777–785.
36. Syers, J.K.; Mackay, A.D.; Brown, M.W.; Currie, L.D. Chemical and physical characteristics of phosphate rock materials of varying reactivity. J. Sci. Food Agric. **1986**, *37* (11), 1057–1064.
37. Taylor, M.D. Accumulation of cadmium derived from fertilisers in New Zealand soils. Sci. Total Environ. **1997**, *208* (1–2), 123–126.
38. Chen, S.; Sun, L.; Sun, T.; Chao, L.; Guo, G. Interaction between cadmium, lead and potassium fertilizer (K_2SO_4) in a soil-plant system. Environ. Geochem. Health **2007**, *29* (5), 435–446.
39. Hernberg, S. Lead poisoning in a historical perspective. Am. J. Ind. Med. **2000**, *38* (3), 244–254.
40. Nguyen, C.K.; Stone, K.R.; Dudi, A.; Edwards, M.A. Corrosive microenvironments atlead solder surfaces arising from galvanic corrosion with copper pipe. Environ. Sci. Technol. **2010**, *44* (18), 7076–7081.
41. Zhang, Y.; Triantafyllidou, S.; Edwards, M. Effect of nitrification and GAC filtration on copper and lead leaching in home plumbing systems. J. Environ. Eng. **2008**, *134* (7), 521–530.
42. Geivanidis, S.; Pistikopoulos, P.; Samaras, Z. Effect on exhaust emissions by the use of methylcyclopentadienyl manganese tricarbonyl (MMT) fuel additive and other lead replacement gasolines. Sci. Total Environ. **2003**, *305* (1–3), 129–141.
43. Hagner, C. European regulations to reduce lead emissions from automobiles? Did they have an economic impact on the German gasoline and automobile markets? Reg. Environ. Change **2000**, *1* (3–4), 135–151.
44. Kerr, S.; Newell, R.G. Policy-induced technology adoption: Evidence from the U.S. lead phase down. J. Ind. Econ. **2003**, *51* (3), 317–343.
45. Allen, A.G.; Radojevic, M.; Harrison, R.M. Atmospheric speciation and wet deposition of alkyllead compounds. Environ. Sci. Technol. **1988**, *22* (5), 517–522.
46. De Jonghe, W.R.A; Chakraborti, D.; Adams, F.C. Sampling of tetraalkyl lead compounds in air for determination by gas chromatography-atomic absorption spectrometry. Anal. Chem. **1980**, *52* (12), 1974–1977.
47. Hewitt, C.; Harrison, R.M. A sensitive, specific method for the determination of tetraalkyl lead compounds in air by gas chromatography/atomic absorption spectrometry. Anal. Chim. Acta **1985**, *167,* 277–287.
48. Hewitt, C.; Harrison, R.M.; Radojevic, M. The determination of individual gaseousionic alkyllead species in the atmosphere. Anal. Chim. Acta **1986**, *188,* 229–238.

49. Nerin, C.; Pons, B.; Martinez, M.; Cacho, J. Behaviour of several solid adsorbents for sampling tetraalkyllead compounds in air. Mikrochim. Acta **1994**, *112* (5–6), 179–188.
50. Nielsen, T.; Egsgaard, H.; Larsen, E.; Schroll, G. Determination of tetramethyllead and tetraethyllead in the atmosphere by a two-step enrichment method and gaschromatographic- mass spectrometric isotope dilution analysis. Anal. Chim. Acta **1981**, *124* (1), 1–13.
51. Radziuk, B.; Thomassen, Y.; Van Loon, J.; Chau, Y. Determination of alkyl lead compounds in air by gas chromatography and atomic absorption spectrometry. Anal. Chim. Acta **1979**, *105*, 255–262.
52. Warren, L.J. Contamination of sediments by lead, zinc and cadmium: A review. Environ. Pollut. B. **1981**, *2* (6), 401–436.
53. Reichlmayr-Lais, A.M.; Eder, K.; Kirchgessner, M. The effect of lead supply on liver phospholipids of lactating rats. J. Anim. Physiol. Anim. Nutr. **1993**, *70* (1–5), 104–108.
54. Reichlmayr-Lais, A.M.; Eder, K.; Kirchgessner, M. Fatty acid composition of erythrocyte membranes, liver and milk of rats depending on different alimentary lead supply. J. Anim. Physiol. Anim. Nutr. **1993**, *70* (1–5), 109–116.
55. USEPA. *Cadmium Fact Sheet;* Agency of Toxic Substances and Disease Registry: Atlanta, GA, 2011.
56. Cornelis, R.; Heinzow, B.; Herbert, R.B.; Christensen, J.; Poulsen, O.; Sabbioni, E. Sample collection guidelines for trace elements in blood and urine (technical report). Pure Appl. Chem. **1995**, *67* (8/9), 1575–1608.
57. Ishihara, T.; Kobayashi, E.; Okubo, Y.; Suwazono, Y.; Kido, T.; Nishijyo, M. Association between cadmium concentration in rice and mortality in the Jinzu River basin, Japan. Toxicology **2001**, *163* (1), 23–28.
58. Kawano, S.; Nakagawa, H.; Okumura, Y.; Tsujikawa, K. A mortality study of patients with Itai-itai disease. Environ. Res. **1986**, *40* (1), 98–102.
59. Nogawa, K.; Yamada, Y.; Honda, R.; Ishizaki, M.; Tsuritani, I.; Kawano, S. The relationship between itai-itai disease among inhabitants of the Jinzu River basin and cadmium in rice. Toxicol. Lett. **1983**, *17* (3–4), 263–266.
60. Umemura, T.; Wako, Y. Pathogenesis of osteomalacia in itai-itai disease. J. Toxicol. Pathol. **2006**, *19* (2), 69–74.
61. Malcolm, E.; Morel, F. The biogeochemistry of cadmium. In *Metal Ions in Biological Systems*; Sigel, A., Sigel, H., Sigel, R., Eds.; Biogeochemical Cycles of Elements; CRC Press, Taylor and Francis Group: London, 2005; Vol. 43, 195–219.
62. Wilkinson, J.M.; Hill, J.; Phillips, C.J.C. The accumulation of potentially-toxic metals by grazing ruminants. Proc. Nutr. Soc. **2007**, *62* (02), 267–277.
63. Jarup, L.; Berglund, M.; Elinder, C.G.; Nordberg, G.; Vahter, M. Health effects of cadmium exposure—A review of the literature and a risk estimate. Scand. J. Work Environ. Health **1998**, *24* (Suppl 1), 1–51.
64. Sun, C.; Wong, T.; Hwang, Y.; Chao, K.; Jee, S.; Wang, J. Percutaneous absorption of inorganic lead compounds. AIHA J. **2002**, *63* (5), 641–646.
65. USEPA. *Ecological Soil Screening Levels for Lea;* U.S. Environmental Protection Agency, Office of Solid Waste and Emergency Response: Washington, D.C., 2005; OSWER Directive 9285.7–70.
66. Needleman, H.L.; Bellinger, D. The health effects of low level exposure to lead. Annu. Rev. Public Health **1991**, *12* (1), 111–140.
67. John, F.R. Adverse health effects of lead at low exposure levels: Trends in the management of childhood lead poisoning. Toxicology **1995**, *97* (1–3), 11–17.
68. Goyer, R.A. Lead toxicity: From overt to subclinical to subtle health effects. Environ. Health Perspect. **1990**, *86,* 177–181.
69. Aronson, A. Lead poisoning in cattle and horses following long-term exposure to lead. Am. J. Vet. Res. **1972**, *33* (3), 627–629.
70. Hammond, P.B.; Aronson, AL. Lead poisoning in cattle and horses in the vicinity of a smelter. Ann. N.Y. Acad. Sci. **1964**, *111*, 595–611.

71. Buck, W.B.; James, LF; Binns, W. Changes in serum transaminase activities associated with plant and mineral toxicity in sheep and cattle. Cornell Vet. **1961**, *51,* 568–585.
72. Dinius, D.A.; Brinsfield, T.H.; Williams, EE. Effect of subclinical lead intake on calves. J. Anim. Sci. **1973**, *37* (1), 169–173.
73. Kim, B.J.; Kim, M.; Kim, K.; Hong, Y.; Kim, I. Sensitizing effects of cadmium on TNF-a- and TRAIL-mediated apoptosis of NIH3T3 cells with distinct expression patterns of p53. Carcinogenesis **2002**, *23* (9), 1411–1417.
74. Golovine, K.; Makhov, P.; Uzzo, R.G.; Kutikov, A; Kaplan, D.J.; Fox, E. Cadmium down-regulates expression of XIAP at the post-transcriptional level in prostate cancer cells through an NF-**K**B-independent, proteasome-mediated mechanism. Mol. Cancer **2010**, *9* (1), 183.
75. Jarup, L. Hazards of heavy metal contamination. Br. Med. Bull. **2003**, *68* (1), 167–182.
76. Ahmed, I.A.M.; Young, S.; Crout, N. Time-dependent sorption of Cd^{2+} on CaX zeolite: Experimental observations and model predictions. Geochim. Cosmochim. Acta **2006**, *70* (19), 4850–4861.
77. Ahmed, I.A.M.; Young, S.; Crout, N. Ageing and structural effects on the sorption characteristics of Cd^{2+} by clinoptilo- lite and Y-type zeolite studied using isotope exchange technique. J. Hazard Mater. **2010**, *184* (1–3), 574–584.
78. Ahmed, I.A.M.; Crout, N.M.; Young, S.D. Kinetics of Cd sorption, desorption and fixation by calcite: A long-term radiotracer study. Geochim. Cosmochim. Acta **2008**, *72* (6), 1498–1512.
79. Xu, Y.; Schwartz, F.W.; Traina, S.J. Sorption of Zn^{2+} and Cd^{2+} on hydroxyapatite surfaces. Environ. Sci. Technol. **1994**, *28* (8), 1472–1480.
80. Ryan, J.A.; Zhang, P.; Hesterberg, D.; Chou, J.; Sayers, D.E. Formation of chloropyro-morphite in a lead-contaminated soil amended with hydroxyapatite. Environ. Sci. Technol. **2001**, *35* (18), 3798–3803.
81. Mavropoulos, E.; Rossi, A.M.; Costa, A.M.; Perez, C.A.C.; Moreira, J.C.; Saldanha, M. Studies on the mechanisms of lead immobilization by hydroxyapatite. Environ. Sci. Tech- nol. **2002**, *36* (7), 1625–1629.
82. da Rocha, N.C.C.; de Campos, R.C.; Rossi, A.M.; Moreira, E.L.; Barbosa, A.dF.; Moure, G.T. Cadmium uptake by hydroxyapatite synthesized in different conditions and submitted to thermal treatment. Environ. Sci. Technol. **2002**, *36* (7), 1630–1635.
83. WHO. Exposure to cadmium: A major public health concern. Public Health and Environment, World Health Organization: Geneva, Switzerland, 2010.
84. WHO. *Lead in Drinking-Water: A Background Document for Development of WHO Guidelines for Drinking-Water Quality*; Public Health and Environment, World Health Organization: Geneva, Switzerland, 2011; WHO/SDE/ WSH/03.04/09/Rev/1.
85. WHO. *Cadmium in Drinking-Water: A Background Document for Development of WHO Guidelines for Drinking- Water Quality*; Public Health and Environment, World Health Organization: Geneva, Switzerland, 2011; WHO/ SDE/WSH/03.04/80/Rev/1.
86. DWI. Legislative background to the Private Water Supplies Regulations 2009: Section 9 (E&W). Drinking Water Inspectorate; 2010.
87. IARC. International Agency for Research on Cancer Monographs on the Evaluation of the Carcinogenic Risks to Humans—Beryllium, Cadmium, Mercury, and Exposures in the Glass Manufacturing Industry; WHO: Lyon, 1993; vol. 58, 119–237.
88. Neathery, M.W.; Miller, W.J. Metabolism and Toxicity of Cadmium, Mercury, and Lead in Animals: A Review. Journal of Dairy Science **1975**, *58* (12), 1767–1781.
89. Nagajyoti, P.; Lee, K.; Sreekanth, T. Heavy metals, occurrence and toxicity for plants: a review. Environmental Chemistry Letters **2010**, *8* (3), 199–216.
90. Das, P.; Samantaray, S.; Rout, G.R. Studies on cadmium toxicity in plants: A review. Environmental Pollution **1997**, *98* (1), 29–36.
91. Kabata-Pendias A. Trace Elements in Soils and Plants. CRC Press, Inc.: Boca Raton, FL, 2011.

41

Heavy Metals: Organic Fertilization Uptake

Ewald Schnug,
Alexandra
Izosimova, and
Renata Gaj

Introduction

Heavy metals are elements with atomic weights ranging between 63.54 (Cu) and 200.59 (Hg), and specific gravity greater than 4.0 g/cm^3. Some elements may stimulate growth of an organism, although the evidence for this process is lacking. Micronutrients important to plant growth are Cu, Co, Mn, Fe, Cu, Zn, and Mo. Other trace elements do not play any physiological function and are toxic in extremely small quantities.[1,2] There is a strong link between micronutrient nutrition, the uptake, and the impact of contaminants on plants, animals, and humans.

Contamination with heavy metals has become one of the most serious problems for the functionality of ecosystems. Because of industrial development and past disposal activities, heavy metals are considered to be among the most important environmental contaminants that affect terrestrial, aquatic, and atmospheric systems. Concerns about these contaminants are based on the fact that certain trace elements (e.g., As, Cd, Pb, or U) are accumulating in the food chain threatening the health of humans and animals.[3] The risks of human, crop, and/or environmental toxicity posed by these elements are a function of their mobility and availability in soil.[4] The concentration of heavy metals in the soil solution, as the main parameter for the bioavailability of heavy metals to plants, is affected by numerous factors. Among others, the uptake of heavy metals by plants from soils is also affected by management factors like organic fertilization. Despite the fact that organic fertilizers themselves can be a source for heavy metals, nevertheless organic matter (OM) added with manures, compost, and sludges greatly increases soil physical features like colloidal stability and cation exchange capacity and chemical features like plant available nutrients and chelating compounds which have a beneficial effect on soil biology.

Organic fertilization may affect the transfer of heavy metals to plants by several mechanisms: enhancing metal adsorption as a result of increased surface charge and increasing the formation of organic complexes including changes of the redox potential in soils towards negative values. It may additionally increase the solubility of some elements like Mn on one hand, but also decrease the solubility of other elements like U.[5] The oxidation of OM releases protons and lowers the pH of the soil which increases Mn and Zn solution concentration or lowers the mobility of Mo.[6] A higher cation exchange capacity and increased concentrations of chelating agents, because of the organic fertilization, improve soils

TABLE 1 Typical Heavy Metal Concentrations in Selected Organic Fertilizers

Organic Fertilizers	Zn (mg/kg Dry Weight, Range)	Cd (mg/kg Dry Weight, Range)	Cu (mg/kg Dry Weight, Range)	Ni (mg/kg Dry Weight, Range)	Pb (mg/kg Dry Weight, Range)	References
Manure						
Pig	206–716	0.19–0.53	160–780	3–24.3	1.01–4.65	[18,19]
Cattle	41–238	0.15–0.40	26.2–55.8	1.7–9.1	1.5–8.4	[19]
Poultry	350–632	0.44–2.04	49.4–74.8	4.5–11.4	3.36–14.8	[19]
Slurry						
Pig	639–2115	0.32–1.08	197–773	8.6–11.6	2.7–7.0	[20,21]
Cattle	68–235	0.11–0.53	17.5–48.7	1.9–20.4	4.1–8.4	[19,22]
Sewage sludge	700–50,000	2–1500	50–3300	16–5300	50–3000	[18]
Green compost	72–181	0.11–1.7	17–432	5.7–25	15–131	[23]
Biowaste compost	125–371	0.11–0.7	34–215	7.2–37.7	19–435	[24]

storage capacity for heavy metals. This function acts as a buffer system against heavy metal transfers in ecosystems, either by preventing them from leaching or by counteracting their accumulation in food plants.[7] Metal retention processes are generally much more important than metal leaching processes.

Sources of Organic Fertilizers

Differences between organic fertilizers must also be considered. Some typical metal concentrations in organic fertilizers are given in Table 1. Narwal and Singh[8] observed that increasing rates of cow and pig manure decreased the amounts of extractable Cd and Ni, whereas the addition of peat, at the same rate as OM, had an increasing effect. At this point, the effects are linked to farming systems, which affect, either directly or indirectly, quantities of compounds of plant or animal origin introduced into the soil.[9]

All factors involved in the effects of OM with heavy metal mobility show distinct interaction, like pH and the formation of metal-organic complexes. Most metals (in free ionic form) have a higher mobility in acidic, coarse-textured soils,[10] but mobility, however, can be also significant at about neutral or higher pH owing to the fact that dissolved OM becomes itself more soluble at those pH levels and because of the formation of soluble organometallic complexes. That is why the release of Cd into the soil solution as metal-organic complex becomes significant when the soil pH exceeds 5.5,[11,4] and Pb leaching increases with a higher concentration of dissolved OM and pH.[12] Such interactions are of great importance when remediation of heavy metal polluted soils is discussed. In terms of preventing heavy metals from transferring to ecosystems, the aim of remediation is not decontamination, but immobilization through increasing the binding capacity of the soil.[13]

The largest sources for OM in agricultural systems are plant residues. Gupta et al.[14] observed a considerably higher content of cadmium in wheat grown after the addition of pulses than when the same crop was grown after cereals. This may indicate an effect of nitrogen applied with the organic fertilizers by a synergistic effect of the former on heavy metal uptake, which has been described by Schnug.[15]

In the point of view of negative anthropogenic activities on soil quality, the loss of OM is a particularly serious problem, which also diminishes the buffer capacity of soils for heavy metals. The utilization of wastes, as soil amendments to counteract losses of OM,[16,17] is only an apparent and short-sighted solution, because these sources are usually highly polluted with heavy metals (Table 1).

Plant Uptake of Heavy Metals

Typical heavy metal concentrations in not accumulating higher plants are (mg kg^{-1}): Cd, 0.1–1; Cu, 3–15; Hg, 0.1–0.5; Ni, 0.1–5; Pb, 1–5; Zn, 15–150.[16]There are numerous evidences of an increase in heavy metal phyto-availability and crop uptake related to land application of sewage sludge, fly ash, poultry litter, pig slurry, and other biosolids.[25–27] Supplementing swine feed with Cu has been a common practice for many years, and just recently environmental concerns have been expressed over elevated Cu concentration in soils receiving long-term application of swine waste. On such sites, Cu is preferentially adsorbed by OM associated with the coarse fraction in soil.[28] Sludge-derived OM contributes significantly to the metal adsorption capacity, and the slow mineralization of this OM can release metals into more soluble forms.[29] Because the decomposition of sludge OM is often associated with acidification of the soil, further increased bioavailability of the sludge-born heavy metals is to be considered[30] resulting in enhanced availability of heavy metals, decades after sludge application.[31] Wastes with unsuitable agronomic features can be processed to suitable fertilizers by a variety of methods such as drying, composting, alkaline stabilizing, or incineration to reduce sludge mass, volume, concentration, and mobility of metals.[4]

Any use of organic amendments must consider possible effects on heavy metal availability, especially when the initial contamination of a soil is already high. Under these circumstances, even the use of composts with low metal concentrations can be questionable because of the mobilizing action of soluble organic compounds.[32]

Conclusions

Organic fertilizers affect the heavy metal balance of agro ecosystems by supply, mobilization, and immobilization. Supply and mobilization are useful options in terms of feeding plants with essential heavy metals (e.g., Zn, Cu), but are a risk for loading sensitive environments. Immobilization in and removal from soil of heavy metals by phytoremediation [increased uptake by accumulating plant species (e.g., Polygonum spec.)] are deceptive options: the former leaves the risk in place, while the latter is too inefficient in terms of masses removed and just shifts the problem of contamination to other ecological compartments. The only feasible option for protecting soils, food, and environment is applying the precautionary principle by balancing all sources of input and output of heavy metals to an ecosystem.

References

1. Kabata-Pendias, A. *Trace Elements in the Biological Environment*; Geological Edition: Warsaw, 1979.
2. Tiller, K.G. Heavy metals in soils and their environmental significance. Adv. Soil Sci. **1989**, *9*, 113–142.
3. Iskandar, I.K.; Kirkham, M.B. *Trace Elements in Soil: Bioav-ailability, Flux and Transfer*; Lewis Publishers, 2001.
4. Richards, B.K.; Steenhus, T.S.; Peverly, J.H.; McBride, M.B. Effect of sludge-processing mode, soil texture and soil pH on metal mobility in undisturbed soil columns under accelerated loading. Environ. Pollut. **1999**, *109*, 327–346.
5. Bolan, N.S.; Duraisamy, V.P. Role of inorganic and organic soil amendments on immobilisation and phytoavailability of heavy metals: a review involving specific case studies. Aust. J. Soil Res. **2003**, *41*, 533–555.
6. Khoshgoftarmanesh, A.H.; Kalbasi, M. Effect of municipal waste leachate on soil properties and growth and yield of rice. Commun. Soil Sci. Plant Anal. **2002**, *33*, 2011–2020.
7. Hansen, L.G.; Schaeffer, D.J. Livestock and domestic animals. Soil amendments. In Impact on Biotic Systems; Re-chcigl, J.E., Ed.; Lewis Publishers, **1995**; 41–79.
8. Narwal, R.P.; Singh, B.R. Effect of organic materials on partitioning, extractability and plant uptake of metals in alum shale soil. Water Air Soil Pollut. **1998**, *103*, 405–421.
9. Oliver, D.P.; Schultz, J.E.; Tiller, K.G.; Merry, R.H. The effect of crop rotations and tillage practices on cadmium concentration in wheat grain. Aust. J. Agric. Res. **1993**, *44*, 1221–1234.
10. McBride, M.B. *Environmental Chemistry of Soils*; Oxford University Press: New York, 1994.
11. Naidu, R; Harter, N.D. Effect of different organic ligands on cadmium sorbtion and extractability from soils. Soil Sci. Soc. Am. J. **1998**, *62*, 644–650.

12. Klitzke, S.; Lang, F.; Kaupenjohann, M. Influence of pH and DOM on the mobilization of colloid-bound lead. DIAS Report, Plant Prod. **2002**, *80*, 261–265.
13. Senesi, N. Metal-humic substance complexes in the environment. Molecular and mechanistic aspects by multiple spectroscopic approach. In *Biogeomistry of Trace Metals*; Adriano, D.C., Eds.; Lewis Publishers Chelsea MI: Boca Raton, FL, 1992; 429–496.
14. Gupta, S.K.; Herren, T.; Wenger, K.; Krebs, R.; Hari, T. In situ gentle remediation measures for heavy metal-polluted soils. In *Phytoremediation of Contaminated Soil and Water*; Norman, T., Gary, B., Eds.; Lewis Publishers Chelsea MI: Boca Raton, 2000; 303–322.
15. Schnug, E. Differentiation between primary and secondary effects of N-fertilization on the micronutrient supply of cereals. Landwirtsch. Forsch. **1984**, *40*, 231–240.
16. McGrath, S.; Chaundri, A.M.; Giller, K.E. Long-term effects of metal in sewage sludge on soils microorganisms and plants. J. Ind. Microbiol. **1995**, *14*, 94–104.
17. Joergensen, R.G.; Meyer, B.; Roden, A.; Wittke, B. Microbial activity and biomass in mixture treatments of soil and biogenic municipal refuse compost. Biol. Fertil. Soils **1996**, *2*, 43–49.
18. Kabata-Pendias, A.; Pendias, H. In *Biogeochemistry of Trace Elements;* PWN: Warsaw, 1999.
19. Nicholson, F.A.; Chambers, B.J.; Williams, J.R.; Unwin, R.J. Heavy metal contents of livestock feeds and animal manures in England and Wales. Bioresour. Technol. **1999**, *70*, 23–31.
20. Marb, C.; Riedel, H. Changes in hazardous substances in sewage sludges during cold drying. Korrespondenz Abwasser. **1997**, *44* (8), 1386–1393.
21. Kühnen, V.; Bien, B.; Goldbach, H.E. Heavy metal balances (Cd, Cr, Cu, Ni, Pb, Zn) of pig farming. VD-LUFA-Kongress vom 2001 in Berlin, September 17–21, 2001; Darmstadt, VDLUFA-Verlag, Kongressband 57/II; VDLUFA-Schriftenreihe.
22. Menzi, H.; Kessler, J. Heavy metal contents of manures in Switzerland, Proceedings of the Eighth International Conference on the FAO Escorena Network on Recycling of Agricultural. Municipal and Industrial Residues in Agriculture, Rennes, France, May 26–29, 1998; Martinez, J., Maudet, M.-N, Eds.; 1998; 26–5.
23. Schmoll, H.; Held, T. Use of green compost in forestry. Forst Holz. **1997**, *52* (9), 245–249.
24. Schaaf, H.; Janßen, E. Heavy metal concentrations in manures and secondary raw material fertilizers and heavy metal loads in application of codes of good agricultural practice. *Kongressband 2000,* Darmstadt, VDLUFA-Verlag; VDLUFA-Schriftenreihe, 144–150.
25. Shuman, M.L. Organic waste amendments effect on Zn fractions of 2 soils. J. Environ. Quality **1999**, *28*, 1442–1447.
26. Jackson, B.P.; Miller, W.P.; Shuman, A.W.; Sumner, M.E. Trace elements solubility from land application of fly ash organic waste mixtures. J. Environ. Quality **1999**, *28*, 639–647.
27. Basta, N.T.; Sloan, J.J. Bioavailability of HM in strongly acidic soils treated with exceptional quality biosolids. J. Environ. Quality **1999**, *28*, 633–638.
28. Wu, J.; Laird, D.A.; Thompson, M.L. Sorption and desorption of Cu on soil clay component. J. Environ. Quality **1999**, *28*, 334–338.
29. McBride, M.B. Toxic metal accumulation from agricultural use of sludge: are USEPA regulations protective?. J. Environ. Quality **1995**, *2*, 5–18.
30. McGrath, S.P.; Zhao, F.J.; Dunham, S.J.; Grosland, A.R.; Coleman, K. Long-term changes in the extactability and bioavailability of Zn and Cd after sludge application. J. Environ. Quality **2000**, *29,* 875–883.
31. McBride, M.B.; Richards, B.K.; Steenhuis, T.; Russo, J.; Sauve, S. Mobility and solubility of toxic metals and nutrients in soil fifteen years after sludge application. Soil Sci. **1997**, *162*, 487–500.
32. Madrid, L.; Diaz-Barrientos, E.; Cardo, I. Sequential extraction of metals from artificially contaminated soils in the presence of various composts. In *Trace Elements in Soil: Bioavailability, Flux and Transfer;* Iskandar, I.K., Kirkham, M.B, Eds.; Lewis Publishers, 2001.

42

Inorganic Carbon: Global Carbon Cycle

William H.
Schlesinger

Introduction

The "global C cycle" is defined as the exchange of carbon (C) among the atmosphere, seawater, land vegetation, and soil reservoirs (Figure 1). Each year dead plant materials entering the soil are decomposed by soil microbes that return carbon dioxide (CO_2) to the atmosphere. If the amount of land vegetation remains the same, the amount of CO_2 removed from the atmosphere by plant growth each year is balanced by the amount of plant death and decomposition. Such a perfect balance, however, is seldom seen. Changes in the quantity of C in vegetation and soils play a major role in determining short- and long-term fluctuations in the concentration of CO_2 in earth's atmosphere. A portion of the current atmospheric increase in CO_2, for example, is because of the destruction of vegetation and the disturbance of soils by humans.

Mean Residence Time

For comparative purposes, biogeochemists calculate the mean residence time (MRT), or the amount of time C resides in each pool of the global C cycle before circulating to the others. For instance, a molecule of CO_2 spends, in average, about five years in the atmosphere before it enters the terrestrial biosphere or the oceans. A C atom spends, on average, about 10 years in vegetation and 35 years in soil organic matter (SOM) before it returns to the atmosphere as CO_2. In comparison, the circulation of C in the oceans is rather sluggish; the C atom spends, in average, hundreds of years in the sea, where it is found predominantly as dissolved bicarbonate (HCO_3^-). Human activities have the greatest impact on pools with short mean residence times.

Weathering

Most studies of soil C focus on SOM, which globally contains nearly 2300×10^{15} g of C.[2] The release of CO_2 by the microbial decomposition of SOM is one of the largest fluxes in the global C cycle (Figure 1). However, soils also contain C in various inorganic forms—CO_2 held in the soil pore spaces, bicarbonate dissolved in soil waters, and calcium carbonate ($CaCO_3$) as a soil mineral. Carbon dioxide in the soil

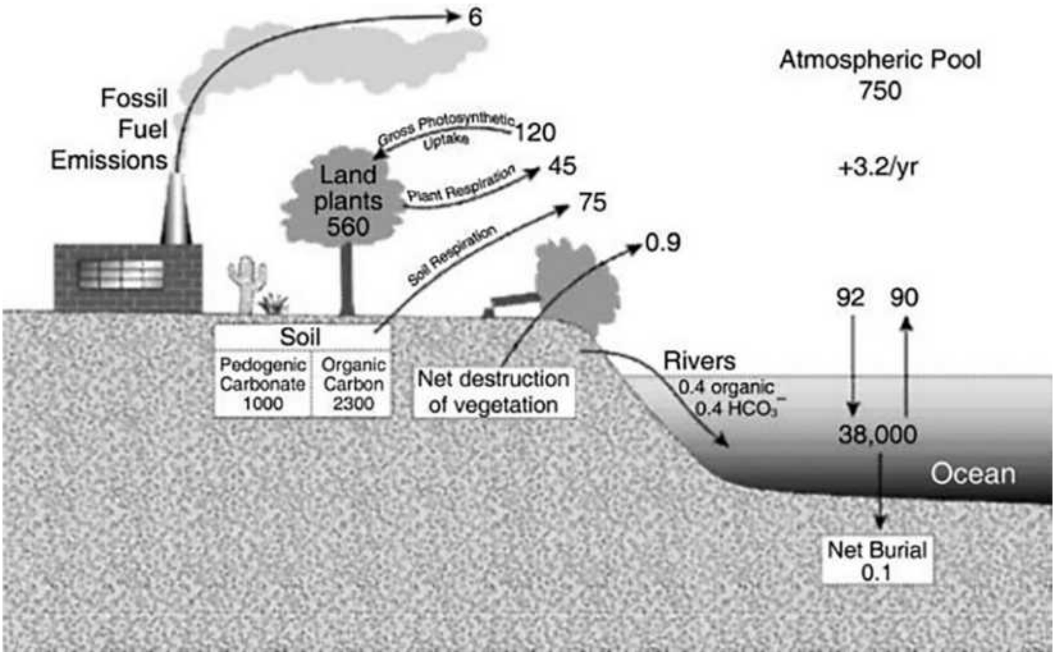


FIGURE 1 The global C cycle, showing the size of reservoirs (10^{15} gC) and the annual flux between them (10^{15} gC/yr).
Source: Modified from Schlesinger.[1]

pores is largely derived from the respiration of plant roots and soil microbes, which varies as a function of soil moisture and temperature. However, even when the respiration rate is very high, the amount of C in soil gases and the soil solution is not large enough to contribute materially to the total C content of soils globally. In contrast, the amount of C held in soil carbonates is quite large, totaling about 750–1000 × 10^{15}gC (Figure 1).[3,4] The vast majority of this carbonate is found in the world's arid and semiarid lands, where inorganic C may exceed soil organic C by a factor of 10.[5]

In some soils, $CaCO_3$ and other carbonates (e.g., dolomite) are derived from the parent rocks from which the soils have formed. However, additional carbonate may form as a result of the release of Ca from the chemical weathering of rocks and the precipitation of $CaCO_3$ when water is lost from the soil by evaporation and plant uptake. In the case of silicate parent minerals, such as plagioclase (calcium feldspar), the relevant reaction of weathering is

$$CaAl_2Si_2O_8 + 3H_2O + 2CO_2 \rightarrow Ca^{2+} + 2HCO_3^- + Al_2Si_2O_5(OH)_4 \tag{1}$$

Here Ca is released from the silicate mineral by the weak solution of carbonic acid that is formed when CO_2 dissolves in water.

Declining soil concentration of either CO_2 or water precipitates carbonate via the reaction

$$Ca^{2+} + 2HCO_3^- \rightarrow CaCO_3 \downarrow + H_2O + CO_2 \tag{2}$$

While the weathering process occurs most rapidly when plant activity increases the concentration of CO_2 in the soil pore space, the precipitation reaction occurs during seasonal periods of drought. Carbonate formed in the soil, known as secondary or "pedogenic" carbonate, can be distinguished from carbonate inherited from parent materials by examining thin sections and the isotopic ratio (^{13}C vs. ^{12}C) of the C in carbonate minerals.

Atmospheric Derivation

In most areas, the Ca content of pedogenic carbonate is derived from the atmosphere and is deposited as a constituent of rain and dust.[6,7] However, because the Ca carried in the atmosphere is ultimately derived from the weathering of rocks in some upwind area, the two aforementioned reactions are general.[8] Across deserts of the southwestern U.S.A., the formation of pedogenic carbonate is closely related to mean annual rainfall (Figure 2).

If we know the rate of formation of pedogenic carbonate, usually expressed in grams per square meter per year ($g/m^2/yr$), and the total amount of carbonate in the soil profile, we can calculate the length of time taken for the accumulation of current quantity of soil pedogenic carbonate. In the Mojave Desert of California, radiocarbon and uranium-thorium dating show that pedogenic carbonate found in the upper 1.5 m of soils, derived from silicate materials, accumulated over a period of up to 20,000 years.[10] During that period, pedogenic carbonate formed at rates ranging from 1.0 to 3.5 $g/m^2/yr$, with some indications of greater rates during the Pleistocene, when rainfall was greater in this region. The global MRT of soil pedogenic carbonate is about 85,000 years, making this C pool much less dynamic than SOM, in which the global MRT for C is about 35 years (Figure 1).

While CO_2 is sequestered from the atmosphere during the formation of pedogenic carbonate from silicate parent materials, no such net storage occurs when pedogenic carbonate forms from carbonate parent materials.[11] Overall, the formation of soil pedogenic carbonates is less effective than the formation of SOM in the storage of atmospheric CO_2.[12] This is disappointing, of course, to those who view increasing the formation of pedogenic carbonate in desert soils as a means of slowing the rise of CO_2 in the earth's atmosphere and reducing global warming.

Effect of Human Activities

Human activities, such as the irrigation of agricultural soils in arid regions, can alter the accumulation of pedogenic carbonate in soils. The groundwater used for irrigation is often extracted from subsurface environments, where CO_2 concentration is much higher than in the earth's atmosphere, and often contains high concentrations of dissolved Ca. Applying such water to arid lands, precipitates dissolved Ca in the soil, thereby forming $CaCO_3$, and releases CO_2 to the atmosphere, via the reaction in Eq. 2. Precipitation

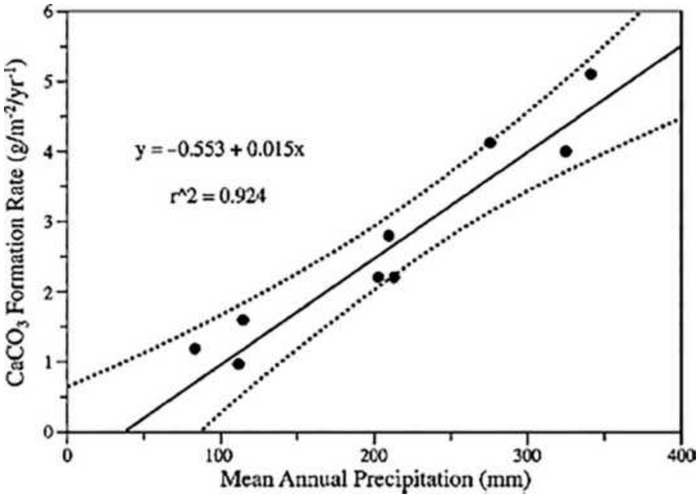


FIGURE 2 The rate of formation of pedogenic carbonate in arid soils of the southwestern U.S.A. as a function of the modern precipitation at each site.
Source: Marion.[9]

of calcite is also favored when large amounts of gypsum ($CaSO_4 \cdot 2H_2O$), a ready source of Ca, are used to remediate dryland soils. The formation of pedogenic carbonate in arid, agricultural soils, as a result of these human activities makes a small contribution to the current increase in atmospheric CO_2.[13]

Human activities leading to the formation of acid rain also affect soil carbonates. If the acidity in rainfall is derived from sulfuric acid (H_2SO_4), then CO_2 is released when the rain falls on carbonate-rich soils. The reaction is

$$CaCO_3 + H_2SO_4 \rightarrow Ca^{2+} + SO_4^{2-} + CO_2 \uparrow + H_2O \tag{3}$$

and gypsum ($CaSO_4 \cdot 2H_2O$) precipitates as the soil dries. Although most fossil fuels contribute small amounts of sulphur to the atmosphere, the global amount of CO_2 derived from this reaction is relatively small compared to the direct release of CO_2 from fossil fuel combustion (Figure 1).

Conclusions

Although the amount of pedogenic carbonate in world soils is quite large, perhaps as much as 1000×10^{15} gC, this pool of C is relatively sluggish. The long MRT of the pedogenic carbonate pool ensures that it will not become a major sink or source of atmospheric CO_2 over the next several centuries.

References

1. Schlesinger, W.H. *Biogeochemistry: An Analysis of Global Change*; Academic Press: San Diego, CA, 1997.
2. Jobbagy, E.G.; Jackson, R.B. The vertical distribution of soil organic matter and its relation to climate and vegetation. Ecol. Appl. **2000**, *10,* 423–436.
3. Eswaran, H.; Reich, P.F.; Kimble, J.M.; Beinroth, F.H.; Padmanabhan, E.; Moncharoen, P. Global carbon stocks. In *Global Climate Change and Pedogenic Carbonates;* Lal, R., Kimble, J.M., Eswaran, H., Stewart, B.A., Eds.; Lewis Publishers: Boca Raton, FL, 2000; 15–25.
4. Batjes, N.H. Total carbon and nitrogen in soils of the world. Eur. J. Soil Sci. **1997**, *47,* 151–163.
5. Schlesinger, W.H. Carbon storage in the caliche of arid soils: a case study from Arizona. Soil Sci. **1982**, *133,* 247–255.
6. Capo, R.C.; Chadwick, O.A. Sources of strontium and calcium in desert soil and calcrete. Earth Planet. Sci. Lett. **1999**, *170,* 61–72.
7. Chiquet, A.; Michard, M.; Nahon, D.; Hamelin, B. Atmospheric input vs. in situ weathering in the genesis of calcretes: an Sr isotope study at Gálvez (Central Spain). Geochim. Cosmochim. Acta. **1999**, *63,* 311–323.
8. Naiman, Z.; Quade, J.; Patchett, P.J. Isotopic evidence for Eolian recycling of pedogenic carbonate and variations in carbonate dust sources throughout the southwest United States. Geochim. Cosmochim. Acta. **2000**, *64,* 3099–3109.
9. Marion, G.M. Correlation between long-term pedogenic $CaCO_3$ formation rate and modern precipitation in deserts of the American southwest. Quaternary Res. **1989**, *32,* 291–295.
10. Schlesinger, W.H. The formation of caliche in soils of the Mojave Desert, California. Geochim. Cosmochim. Acta. **1985**, *49,* 57–66.
11. Monger, H.C.; Martinez-Rios, J. Inorganic carbon sequestration in grazing lands. In *The Potential of U.S. Grazing Lands to Sequester Carbon and Mitigate the Greenhouse Effect;* Follett, R.F., Kimble, J.M., Lal, R., Eds.; CRC Press: Boca Raton, FL, 2000; 87–118.
12. Chadwick, O.A.; Kelley, E.F.; Merritts, D.M.; Amundson, R.G. Carbon dioxide consumption during soil development. Biogeochemistry **1994**, *24,* 115–127.
13. Suarez, D.L. Impact of agriculture on CO_2 as affected by changes in inorganic carbon. In *Global Climate Change and Pedogenic Carbonates;* Lal, R., Kimble, J.M., Eswaran, H., Stewart, B.A., Eds.; Lewis Publishers: Boca Raton, FL, 2000; 257–272.

43

Inorganic Carbon: Modeling

Leslie D. McFadden
and Ronald G.
Amundson

Introduction

Virtually all carbon in soils of arid and semiarid regions of the world accumulates as pedogenic calcium carbonate (referred to subsequently as carbonate). The carbonate usually accumulates in layers that eventually attain the status of calcic horizons and, in much older soils, petrocalcic horizons. Numerous mechanisms for the accumulation of pedogenic carbonate in soils are recognized, but the most fundamental reason for accumulation is limited depth of soil-water movement and seasonally high evapotranspiration that favors precipitation of carbonate within the soil.[1] Many studies of calcic soils in the past few decades demonstrate a close correspondence between the depth of pedogenic carbonate accumulation and modern annual precipitation,[2–4] although recent studies show the relationship may be more complicated.[5] Other studies also show progressive, time-dependent accumulation in many environments[1]; these studies have led to the now generally accepted conceptual models of calcic soil development.[6]

Development of a numerical model of carbonate accumulation, however, is a more challenging proposition, given the remarkably complex character of the soil system. Fortunately, certain aspects of calcic soils facilitate formulation of such numerical models. For example, the observed soil depth-climate relationship implies that, utilizing a sound strategy for simulation of water movement, determination of carbonate movement via solution transport is a reasonable proposition. In addition, a significant body of research shows that the majority of carbonate is derived from accumulated entrapped dust and Ca in rainwater.[1] Finally, data pertaining to calcite geochemistry and dissolution rates in different environments are available and show that, in soils associated with typical ranges in soil CO_2, pH, and salinity, calcite is far more soluble than virtually all silicate minerals and has more rapid dissolution rates.[7] Consequently, a relatively simple model for carbonate movement in a soil based on relations in the $CaCO_3$–H_2O–CO_2 system can be formulated that essentially ignores the more complex chemical reactions involving aluminosilicates. Research on the nature and composition of stable and unstable isotopes in pedogenic carbonate has also helped elucidate the nature of calcic soil development and improve the design for testing the results of numerical modeling.

Compartment Model and Simulations of Pedogenic Carbonate Accumulation

The compartment-model, or “box-model,” approach to modeling of calcic soils accommodates continuously changing values among the interdependent variables that influence soil development. It enables integration of several factors that influence pedogenic carbonate accumulation and that can be explicitly considered in this model. These include soil–water movement and soil–water balance, changing soil CO_2 concentrations and temperature with depth and season, initial parent material composition, carbonate and soluble salt additions from external sources, and calcite reactant surface area. The soil profile is represented by a vertical sequence of compartments of arbitrary dimensions, with the initial characteristics of each compartment specified (i.e., texture, available water-holding capacity, pCO_2). A series of equations that enable forward modeling and simulation of evolving carbonate depth functions using the box-model approach can be derived on the basis of consideration of the factors indicated above. For example, the solubility of calcite is derived from the following equation, after Drever[8]:

$$m^3Ca^{2+} = \left(pCO_2K_1K_{cal}K_{CO_2}\right)/\left(4K_2\gamma Ca^{2+} + \gamma^2HCO_3^-\right) \tag{1}$$

where K_{cal} is the calcite solubility product and K_1, K_2 and K_{CO2} are dissociation constants in the carbonate system, and γCa^{2+} and γHCO_3^- are the activity coefficients of Ca^{2+} and HCO_3^- A gravelly, permeable calcic soil probably best approximates an open system-weathering environment, in which case calcite dissolution rates are probably surface-area controlled rather than diffusion controlled. Also, the dissolution rate is defined ultimately by the rate- limiting conversion of dissolved carbon dioxide (CO_2^*) to H_2CO_3. At a very low solution volume to surface area ratios, and with fast, surface-controlled calcite dissolution rates, H_2CO_3 is rapidly depleted. In such circumstances, a commonly used rate equation that enables determination of dissolution rates in the CO_2–$CaCO_3$–H_2O system[9] is:

$$dC/dt = \left(A'k/V\right)\left(1 - C/C^*\right)^n \, mgl^{-1}s^{-1} \tag{2}$$

where A' is the surface area of rock in contact with water (cm^2), V the water volume (cm^3), k the reaction coefficient (mg cm $l^{-1}s^{-1}$), n the reaction order, C the moles of calcite in solution, and C^* is the solubility of calcite. Values of n and k vary with saturation ratio, temperature, and pCO_2. In the model, A'/V can be specified depending on observed soil features. Eqs. 1 and 2 show that soil CO_2 content is a very important variable, but soil CO_2 contents may be highly variable.[10] Fortunately, studies show that a depth function for pCO_2 that reflects prolonged seasonal respiration levels can be estimated, assuming the concentration of soil CO_2 is described by mass transport of CO_2 by gas diffusion.[11,12] The following diffusion-reaction equation, essentially Fick’s Second Law for a one-dimensional case, is used in the model:

$$\partial C_s/\partial t = D_s\left(\partial^2 C_s/\partial z^2\right) + \phi_s(z) \tag{3}$$

where C_s is the concentration of CO_2 in the soil (mol cm^{-3}), t the time (s), D_s the diffusion coefficient for CO_2 in the soil (cm^2s^{-1}), z the depth in the soil (cm), and $\phi_s(z)$ is the production rate of CO_2 as a function of depth (mol $cm^{-3}s^{-1}$). At steady state, when $\partial C_s/\partial t = 0 = D_s\partial^2 C_s/\partial z^2 + \phi_s$, the general solution to this equation to produce a simple production function is:

$$C_s(z) = \phi/D_s\left(Lz - z^2/2\right) + C_0 \tag{4}$$

where C_0 is the concentration of CO_2 in the atmosphere (ppm) and L is the depth to the lower, no-flux boundary. Soil CO_2 contents with depth calculated using this method are used to calculate carbonate solubility and dissolution rates with depth.

Available water-holding capacity, infiltration, and percolation rates can be specified on the basis of laboratory soil measurements or estimated from field measurements or theoretical considerations.

Earlier versions of the simulation model included certain assumptions that simplified numerical calculations, such as simple vertical saturated flow and constant soil temperature with depth. The lack of certain types of data (e.g., variation of pCO_2 with depth and time of year) also constituted a limitation on utility of the model. The model did enable simulation of 1) realistic depths and magnitudes of carbonate accumulation over thousands of years and 2) the range of effects of large climatic changes on calcic soils.[13–16] Model results emphasized the critical roles of external Ca^{2+} influx and influence of soil CO_2 concentrations on carbonate accumulation. Model-simulated bimodal concentrations of carbonate based on theoretical, late-Pleistocene climatic conditions resembled those observed in late- Pleistocene, polygenetic soils; however, incompletely understood changes in the magnitude of climate changes, dust flux, and vegetation change in the Quaternary complicate attempts to simulate polygenetic soils.[7,14,16]

Later versions of the model utilized important new inputs and employed routines that reflected improved understanding of key processes that strongly influence calcic soils. Studies of dust accumulation rates in the American South- west,[17] C, O, and Sr isotopes in carbonate, and development of more sophisticated models for unsaturated flow in calcic soils[18] have allowed development of improved compartment models that can address new and more challenging research problems. For example, such numerical simulations demonstrate how climate changes in the Holocene might have dramatically influenced the rates and temporal patterns of soluble salt leaching and accumulation relative to pedogenic carbonate.[19] A more recent modeling study addressed the problem of how carbonate can occasionally accumulate at much shallower depths than those expected from the depth—annual leaching depth relationship.[20] This study showed how carbonate can be preferentially removed from depths of a few cm to a few dm below the soil surface, while carbonate simultaneously accumulates either as collars on surface pavement clasts or in the vesicular A horizon. Model results also explain how a significant change in climate or soil erosion rates could cause the dissolution of carbonate rinds on the tops and sides of boulders and/or the tops of limestone boulders at depths of up to several dm, unusual features observed in some calcic soils.[21]

Isotopes in Calcic Soils

During weathering, parent material carbonate undergoes dissolution and reprecipitation in the soil. The carbon ($^{13}C/^{12}C$,$^{14}C/^{12}C$) and oxygen ($^{18}O/^{16}O$) isotope ratios of pedogenic carbonate that forms from dust or parent material carbonate, or from Ca^{2+} derived from silicate weathering, are determined by isotopic composition of soil CO_2 and H_2O. These are the primary carbon and oxygen reservoirs, respectively, for the carbonate. Therefore, pedogenic carbonate reflects only isotopic conditions of the soil and bears no memory of the isotopic composition of the rock or mineral from which it was derived.

Soil CO_2 is derived primarily from decomposition of soil organic matter and root respiration. The C isotope composition of soil CO_2 reflects: 1) the isotopic composition of these CO_2 sources; 2) the effects of the diffusion of this CO_2 toward the atmosphere; and 3) the isotopic composition of atmospheric CO_2. In the 1980s, researchers recognized that fairly simple, steady state, diffusion models could be used to reasonably explain the observed depth patterns of C isotopes in both soil CO_2 and pedogenic carbonate. The solution to the mathematical model that encompasses the forementioned processes describes the abundance of $^{12}CO_2$ in soils. A related equation can be derived for $^{13}CO_2$, and the ratio of the two models then describes the ratio of C isotopes at any given soil depth. A similar approach can also be used to model the ^{14}C composition of CO_2 with soil depth, with the additional complication that the two main sources of soil CO_2 (humus decomposition and root respiration) have different ^{14}C contents, making the solution to the model slightly more complex.[22] The C isotope diffusion model of soil CO_2 has provided the opportunity to quantitatively use pedogenic carbonates in a number of applications: 1) paleovegetation studies[23]; 2) paleo-atmospheric CO_2 studies[24]; and 3) radiocarbon dating of pedogenic carbonate and geomorphic surfaces.[25,26]

The O isotopic composition of soil water is determined by the O isotope composition of precipitation and the evaporation of soil water. It has been observed that the^{18} O content of modern precipitation is generally correlated with mean annual temperature on a global scale,[27] but regional differences due to storm sources can obscure these patterns.[28] If precipitation water (once stored in the soil) is subject to evaporation, an enrichment of the remaining soil water in ^{18}O occurs because water vapor depleted in the "heavy" isotope is preferentially removed during evaporation. Models have been made that successfully explain the key components of this process[29] and the fact that soils subject to evaporation commonly have generally decreasing ^{18}O contents of soil water with depth.[30] These models have two components. The first is a vapor transport layer (describing the flow of evaporating soil water to the atmosphere through a dry soil layer), and the second is an evaporating front layer. The evaporating front layer exists below the vapor transport zone. At the evaporating front ^{18}O enrichment of soil water occurs as water is transferred to a vapor phase, and the remaining ^{18}O-enriched soil water at the evaporating front then undergoes diffusional mixing with the less ^{18}O-enriched water at greater depths. In general, these models have been more difficult to use than C isotope models due to the dynamic nature of soil water (steady state assumptions are difficult to apply) and the array of model parameters, many of them not known with certainty for most soils.

Oxygen isotopes in pedogenic carbonate have been used less extensively in paleoclimate work than C isotopes because of concern over possible evaporation of soil water that formed the carbonate. Amundson et al.[28] demonstrated that, except for hyperarid regions, the O isotope composition of pedogenic carbonate appears to reasonably reflect that of the local precipitation. There is a growing list of studies using carbonate O isotopes in Quaternary[28] and Tertiary[31] paleoclimate applications.

Conclusions

Numerical models and isotope studies have proven to be valuable tools in the study of calcic soil development. They have helped elucidate the relation of climate, vegetation, and geomorphic processes to carbonate accumulation. The models are not able to explain the observed character of certain aspects of calcic soils, such as patterns of pedogenic carbonate development in some soil chronosequences,[32] or the somewhat enigmatic formation of calcic soils in humid, monsoonal climates. Also, these models are not designed to simulate the evolution of very old, morphologically complex soils with petrocalcic horizons. Future models must be designed to address locally abundant calcic soils. Additional fieldwork and application of recently developed field and laboratory techniques will provide the basis for development of the next generation of numerical models.

References

1. Birkeland, P.W. *Soils and Geomorphology*; Oxford University Press: New York, 1999; 720 pp.
2. Jenny, H. *Factors in Soil Formation*; McGraw-Hill: New York, 1941; 281 pp.
3. Arkley, R.J. Calculation of carbonate and water movement in soil from climatic data. Soil Sci. **1963**, *96,* 239–248.
4. Retallack, G.J. The environmental factor approach to the interpretation of paleosols. In *Factors of Soil Formation: A Fiftieth Anniversary Retrospective;* Amundson, R., Ed.; Special Publication 33; SSSA: Madison, WI, 1994; 31–64.
5. Royer, D.L. Depth to pedogenic carbonate horizon as a pa- leoprecipitation indicator? Geology **1999**, *27,* 1123–1126.
6. Machette, M.A. Calcic soils of the southwestern united states. In *Soils and Quaternary Geology of the Southwestern United States;* Weide, D.L., Faber, M.L., Eds.; Special Paper 203; Geological Society of America: Boulder, CO, 1985; 1–21.
7. McFadden, L.D.; Tinsley, J.C. The rate and depth of accumulation of pedogenic carbonate accumulation in soils: formation and testing of a compartment model. In *Soils and Quaternary Geology of the Southwestern United States*; Weide, D.L., Faber, M.L., Eds.; Special Paper 203; Geological Society of America: Boulder, CO, 1985; 23–42.

8. Drever, J.I. *The Geochemistry of Natural Waters,* 2nd Ed.; Prentice-Hall: Englewood Cliffs, NJ, 1991; 437 pp.
9. Palmer, A.N. Origin and morphology of limestone caves. Geol. Soc. Am. Bull. **1991**, *103,* 1–21.
10. Kiefer, R.H.; Amey, R.G. Concentrations and controls of soil carbon dioxide in sandy soil in the north Carolina coastal plain. Catena **1992**, *19,* 539–559.
11. Solomon, D.K.; Cerling, T.E. The annual carbon dioxide cycle in a montane soil: observations, modeling, and implications for weathering. Water Resour. Res. **1987**, *23,* 2257–2265.
12. Cerling, T.E.; Quade, J.; Yang, W.; Bowman, J.R. Carbon isotopes in soils and palaeosols, and ecology and palaeo- ecology indicators. Nature **1989**, *341,* 138–139.
13. McFadden, L.D.; Amundson, R.G.; Chadwick, O.A. Numerical modeling, chemical, and isotopic studies of carbonate accumulation in soils of arid regions. Soil Sci. Soc. Am. Spec. Publ. **1991**, *26,* 17–35.
14. McFadden, L.D. The impacts of temporal and spatial climatic changes on alluvial soils genesis in southern California. University of Arizona: Tucson, CA; 430.
15. Marion, G.M.; Schlesinger, W.H.; Fonteyn, P.J. Caldep: a regional model for soil CaCO3 (caliche) deposition in southwestern deserts. Soil Sci. **1985**, *139,* 468–481.
16. Mayer, L.; McFadden, L.D.; Harden, J.W. Distribution of calcium carbonate in desert soils: a model. Geology **1988**, *16*, 303–306.
17. Reheis, M.C.; Kihl, R. Dust deposition in southern Nevada and California, 1984–1989: relations to climate, source area, and source lithology. J. Geophys. Res. **1995**, *100,* 8893–8918.
18. McDonald, E.V.; Pierson, F.B.; Flerchinger, G.N.; McFadden, L.D. Application of a soil-water balance model to evaluate the influence of holocene climatic change on calcic soils, Mojave desert, California, USA. Geoderma **1996**, *74,* 167–192.
19. McFadden, L.D.; Crossey, L.J.; McDonald, E.V. Predicted response to calcic soil development to periods of significantly wetter climate during the late holocene (abstr.). Geol. Soc. Am. Abstr. Progr. **1990**, *24* (6), A252.
20. McFadden, L.; McDonald, E.; Wells, S.; Anderson, K.; Quade, J.; Forman, S. The vesicular layer and carbonate collars of desert soils and pavements: formation, age and relation to climate change. Geomorphology **1998**, *24,* 101–145.
21. Treadwell-Steitz, C.; McFadden, L.D. Influence of parent material and grain size on carbonate coatings in gravelly soils, palo duro wash, New Mexico. Geoderma **2000**, *94,* 1–22.
22. Amundson, R.; Stern, L.; Baisden, T.; Wang, Y. The isotopic composition of soil CO_2. Geoderma **1998**, *82,* 83–114.
23. Cerling, T.E. Development of grasslands and savannas in East Africa during the neogene. Palaeogeogr. alaeoclimatol. Palaeoecol. **1992**, *97,* 241–247.
24. Cerling, T.E. Use of carbon isotopes in paleosols as an indicator of the $P(CO_2)$ of the paleoatmosphere. Global Bio- geochem. Cycles **1992**, *6*, 307–314.
25. Amundson, R.; Wang, Y.; Chadwick, O.; Trumbore, S.; McFadden, L.D.; McDonald, E.; Wells, S.; DeNiro, M. Factors and processes governing the^{14}C content of carbonate in desert soils. Earth Planet. Sci. Lett. **1994**, *125,* 385–405.
26. Wang, Y.; McDonald, E.; Amundson, R.; McFadden, L.D.; Chadwick, O. An isotopic study of soil in chronological sequences of alluvial deposits, providence mountains, California. Geol. Soc. Am. Bull. **1996**, *108,* 379–391.
27. Rozanski, K.L.; Aragua's-Aragua' s, L.; Gonfiantini, R. Isotopic patterns in modern global precipitation. In *Climate Change in Continental Isotopic Records*; Swart, P.K., Ed.; American Geophysical Union Monograph 78; Washington, D.C., 1998; 1–36.
28. Amundson, R.; Chadwick, O.; Kendall, C.; Wang, Y.; DeNiro, M. Isotopic evidence for shifts in atmospheric circulation patterns during the late quaternary in midnorth America. Geology **1996**, *24,* 23–26.
29. Barnes, C.J.; Allison, G.B. The distribution of deuterium and ^{18}O in dry soils. I. Theory. J. Hydrol. **1983**, *60*, 141–156.

30. Allison, G.B.; Hughes, M.W. The use of natural tracers as indicators of soil water movement in temperate semiarid regions. J. Hydrol. **1983**, *60,* 157–173.
31. Quade, J.; Cerling, T.E.; Bowman, J.R. Development of the asian monsoon revealed by marked ecological shift during the latest miocene in Northern Pakistan. Nature **1989**, *343,* 163–166.
32. Holliday, V.T.; McFadden, L.D.; Bettis, E.A.; Birkeland, P.W. Soil survey and soil geomorphology. In *Profiles in the History of the U.S. Soil Survey;* Helms, D., Efflard, A., Dwara, P., Eds.; Iowa State University Press: Ames, IO, 2001.

44

Inorganic Compounds: Eco-Toxicity

Sven Erik Jørgensen

Introduction: The Elements and Their Biological Effect

Today's periodic table has 118 elements, but more than 20 of the last elements are extremely unstable and therefore are of no environmental interest with the exception of their radioactivity—see the entries about radioactivity. Ninety- one of the elements are of natural occurrence, namely, the first elements, from hydrogen to uranium, except element number 43, technetium. These 91 elements can be classified into four groups. (The numbers in parentheses indicate the number in the periodic table.)

a. Elements that are present in biological material in relatively high concentrations: hydrogen (1), carbon (6), nitrogen (7), oxygen (8), sodium (11), magnesium (12), silica (14), phosphorus (15), sulfur (16), chlorine (17), potassium (18), calcium (19), and iron (26).
b. Trace elements. They may often be present in biological material in small concentrations: boron (5), fluorine (9), chromium (24), manganese (25), cobalt (27), nickel (28), copper (29), zinc (30), selenium (34), molybdenum (42), and iodine (53). A concentration that is too high may, however, cause environmental problems.
c. Elements that have an ecotoxicological effect. They are harmful even in very small concentrations and are therefore the elements that we have to consider in environmental management as ecotoxicological elements and compounds. They are the inorganic compounds that may cause pollution problems associated with a toxic effect. They are the following: aluminum (13), arsenic (33), strontium (38), silver (47), cadmium (48), tin (50), antimony (51), tellurium (52), barium (56), mercury (80), thallium (81), lead (82), bismuth (83), polonium (84), radon (86), radium (88), thorium (90), uranium (902), and plutonium (94). The last six elements, in addition to being toxic, are also radioactive. They are present in small concentrations, and the main concern is their radioactivity. The reference to the entries about radioactivity should therefore be applied for these six elements. (As five elements belonging to group B, namely, chromium, copper, nickel, selenium, and zinc, may be of ecotoxicological/environmental concern due to frequent occurrence of harmful environmental concentrations, the overview in the section "The Pollution and Ecotoxicological Problems of 18 Elements"will include these five elements and encompass 18 elements in all.)
d. D All other elements are of no or only minor environmental interest due to their occurrence in low concentrations or negligible biological effects.

TABLE 1 Average Freshwater Plant Composition on Wet Basis

Element	Plant Content (%)
Oxygen	80.5
Hydrogen	9.7
Carbon	6.5
Silica	1.3
Nitrogen	0.7
Calcium	0.4
Potassium	0.3
Phosphorus	0.08
Magnesium	0.07
Sulfur	0.06
Chlorine	0.06
Sodium	0.04
Iron	0.02
Boron	0.001
Manganese	0.0007
Zinc	0.0003
Copper	0.0001
Molybdenum	0.00005
Cobalt	0.000002

Source: Wetzel.[1]

Table 1 gives the composition of average fresh water plants on a wet basis. The table presents 19 out of the 24 elements included in groups A and B. The following elements are not included in Table 1: fluorine, chromium, nickel, selenium, and iodine, because they are present in very small concentrations in freshwater plants—below 0.000002%.

The elements in groups B and C show bioaccumulation and biomagnifications, which are two important processes to be considered in all ecotoxicological evaluations.

The 24 elements in groups A and B encompass the elements that are needed for the entire biosphere. All 24 elements are not necessarily essential for all organisms, but each organism will require most of these 24 elements. These elements are threshold agents.[2] They are needed in a certain concentration for the growth of organisms, but if they are present in a concentration that is too high, they may be harmful for the environment. A typical example is the eutrophication problem, which results from a concentration of nutrients that is too high (see *Eutrophication,* p. 1115), although the nutrients (with emphasis on nitrogen, carbon, phosphorus, and sulfur) are absolutely necessary for all living organisms. Below, we have included in the list of elements five elements from group B, as mentioned above, because they are frequently present in concentrations in the environment that may cause ecotoxicological effects. Non-threshold agents or gradual agents are harmful for the environment in practically any concentration. The harmful effect may be proportional to the concentration, or there may be another relationship between the effect and the concentration, for instance, when the harmful effect is proportional to the concentration in the second exponent.

Figure 1 illustrates the harmful effect versus the concentration for threshold and non-threshold agents. An overview of group C (the radioactive elements + five of the elements in group B, all with a clear ecotoxicological effect) is given in this entry with emphasis on the environmental problem caused by the elements or their compounds. Heavy metals are an important soil pollution problem, and a more detailed overview of this problem can be found in the entry "Heavy Metals." Three of the group C elements, mercury, lead, and cadmium, require particular environmental attention because they have caused very serious environmental problems. Due to their environmental impact, each of these three

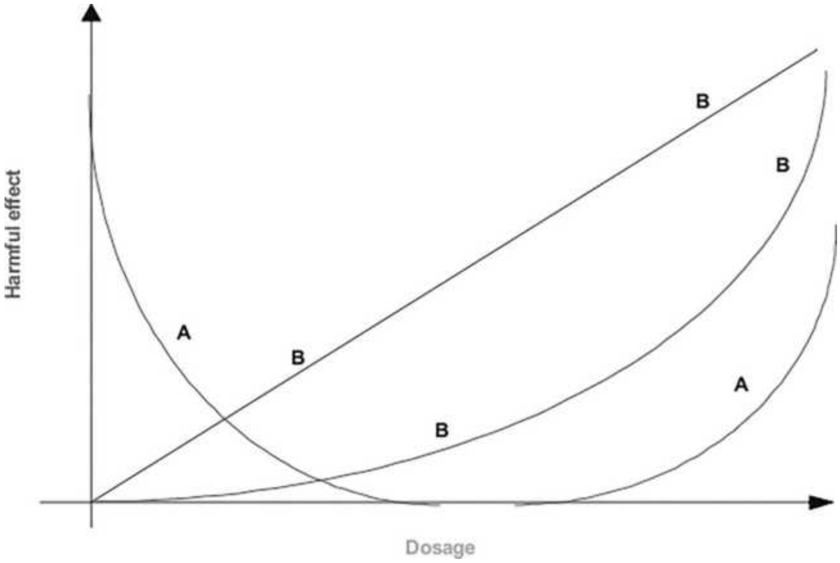


FIGURE 1 The harmful effect is plotted versus the dosage or concentration for threshold agents (a) and non-threshold agents (b).

elements is treated in more detail in three entries with the titles "Mercury," "Lead," and "Cadmium." The pollution problem of copper is furthermore treated in a separate entry, as copper is toxic to many plants and has an ecotoxicological effect at even small concentrations.

Pollution and Ecotoxicological Problems of 18 Elements

As already mentioned, there are 18 elements that require particular attention in environmental management due to their ecotoxicological effect.[3] They are aluminum (13), chromium (24), nickel (28), copper (29), zinc (30), arsenic (33), selenium (34), strontium (38), silver (47), cadmium (48), tin (50), antimony (51), tellurium (52), barium (56), mercury (80), thallium (81), lead (82), and bismuth (83). Four of these elements are treated in more detail in entries with the titles "Mercury," "Lead," "Cadmium," and "Copper." A short overview of the environmental and ecotoxicological problems of these 18 elements is given below, including their applications and environmental effects.

Table 2 gives an overview of the concentration in the earth's crust, in average soil, and in average seawater. If environmental concentrations significantly higher than these concentrations in the earth's crust, average soil, and average seawater are recorded, there is a high probability of an ecotoxicological effect. The last column of the table gives the most important applications of the element and its compounds.

Table 3 gives the lethal dose causing 50% mortality (LD50) and lethal concentration causing 50% mortality (LC50) values as expressions for the toxicity. The third column in this table indicates whether there are environmental effects and risks that require particular attention. For instance, it is indicated that unacceptably high concentrations of arsenic can be found at some locations, particularly in groundwater.[4] Due to the toxicity and carcinogenic effect of arsenic, treatment of the water is absolutely necessary.[4] Effective treatment of water is possible by precipitation/coagulation and ion exchange.[4]

Table 3 is a coarse overview. The toxicology and ecotoxicology are, of course, much more complex than it would be possible to summarize in one table. It is therefore strongly recommended to seek out much more information if there is the slightest suspicion of an ecotoxicological effect with the concentrations of the 18 elements listed in Tables 2 and 3. This entry is given only as an overview.

TABLE 2 Typical Concentrations and the Major Applications of the 20 Elements That Are of Environmental Concern

Element	Conc. in the Earth's Crust(mg/kg)	Conc. in Soil (mg/kg)	Conc. in Sea Water (mg/L)	Applications
13.Al	81,000	10,000–300,00	1	Construction
51.Sb	0.5	1	0.3	Alloys, plastic
33.As	5	1–40	2.6	Chemicals
56.Ba	425	500	30	Chemicals, glass
82.Pb	16	20	0.02	Batteries, soldering, alloys
48.Cd	0.55	0.06	0.1	Surface treatment, dyestuff
29.Cu	100	20	0.9	Cables, wires, construction
24.Cr	300	100	0.05	Surface treatment, alloys
80.Hg	0.065	0.08	0.15	Instruments, dental use
28.Ni	100	40	6.6	Alloys, surface treatment
34.Se	0.09	0.2	0.05	Glass, instruments, dyestuff
47.Ag	0.1	0.1	0.3	Photochemicals, electric components, ornaments
38.Sr	250	315	8100	Chemicals
81.Tl	1	0.1	0.01	Electronic, alloys
52.Te	0.002	0.001	0.001	Alloys
50.Sn	40	10	3	Alloys, soldering, chemicals
83.Bi	0.02	1	0.03	Alloys, chemicals
30.Zn	40	50	10	Surface treatment, alloys, chemicals

TABLE 3 Overview of Toxicity Expressed by LD50 and LC50

	LD50 (mg/kg)	LC50 (mg/L)	Ecotoxicological Attention
13.Al	770 (mice)	3,900 (zooplankton)	Low ecotoxicological effect
51.Sb	100 (rats)	9,000 (fish)	Industrial exposure
33.As	41 (rats)	74,000 (cyclops)	Groundwater, carcinogenic effect, nerve inflammation
56.Ba	8–23 (rats)	14,500 (cyclops)	Heart problems, diarrhea
82.Pb	130 (rats)	6 ppm (air, rats)	Anemia, brain damage
48.Cd	80 (rats)	65 (cyclops)	Itai-itai, kidney damage
29.Cu	220 (rats)	10 (cyclops)	Highly toxic to plants
24.Cr	3,250 (rats)	100,000 (shrimps)	Carcinogenic effect as Cr(IV)
80.Hg	8 (mice)	5 (cyclops)	Central nerve system, mental retardation, teratogenic effect
28.Ni	1620 (rats)	510 (cyclops)	Allergy, carcinogenic and teratogenic effects
34.Se	7 (rats)	2,500 (cyclops)	Liver and kidney damage
47.Ag	129 (mice)	30 (zooplankton)	Skin damage
38.Sr	148 (rats)	125,000 (cyclops)	–
81.Tl	16 (rats)	–	–
52.Te	83 (rats)	–	Teratogenic effect
50.Sn	41 (mice)	55,000 (zooplankton)	Liver and kidney damage
83.Bi	13 (rats)	–	Liver and kidney damage
30.Zn	975 (rats)	100 (zooplankton), 10,000 (fish)	Minor carcinogenic effect

References

1. Wetzel, R.G. *Limnology,* 2nd Ed.; Saunders College Publishing: New York, 1983; 828 pp.
2. Jørgensen, S.E. *Principles of Pollution Abatement;* Elsevier: Amsterdam, 2000; 520 pp.
3. Jørgensen, S.E.; Fath, B. *Ecotoxicology;* Elsevier: Amsterdam, Oxford, 2010; 390 pp.
4. Murphy, T.; Guo, J. *Aquatic Arsenic Toxicity and Treatment;* Beckhuys Publishers: Leiden, 2003; 165 pp.

45

Leaching

Lars Bergström

Introduction

The contamination of groundwater by pesticides is of concern mainly because it may limit its use as drinking water. The extent of groundwater contamination to a large extent depends on the degree to which pesticides leach through the unsaturated zone of soils on which they have been applied. Pesticide movement through the unsaturated zone in tile-drained fields may also be a source of pesticides in surface waters, which support aquatic ecosystems and are used for drinking water in many areas of the world. Therefore, knowledge of pesticide movement in soil above the groundwater table is very important.[2] This has also been the focus of a large number of studies performed during the past couple of decades.[3,4] The most important factors influencing pesticide leaching are soil properties, inherent properties of the pesticide molecules, climatic conditions, and management practices.[3]

Factors Influencing Leaching

Soil and Hydrological Conditions

The rate and direction of water flow in the unsaturated zone are determined by the hydraulic gradient and the hydraulic conductivity. The presence of air-filled pores restricts the pathways through which water percolates downward, which means that the hydraulic conductivity in the unsaturated zone also depends on the level of water saturation. As the soil dries out, water becomes more strongly bound within the matrix of the soil, and the volume of water and the rate with which water percolates through soil decrease. This relationship between water retention and hydraulic conductivity varies considerably with soil type (soil texture and structure, and organic matter content).[5]

A complication, which has a major impact on pesticide leaching, is the fact that water, and pesticides dissolved in the water phase, often move through large pores in soil (e.g., earthworm and root channels, cracks etc.), a process commonly referred to as preferential flow.[6,7] Under such conditions, an equilibrium pesticide concentration throughout the soil profile cannot be obtained. This phenomenon primarily occurs in fine-textured soils with high clay contents, especially those that have the potential to swell and shrink. Through preferential flow, pesticides can be transported rapidly through large portions of the unsaturated zone and bypass biologically active layers in which they otherwise would be degraded or sorbed. Exposure to preferential flow is most pronounced soon after application of the pesticide,

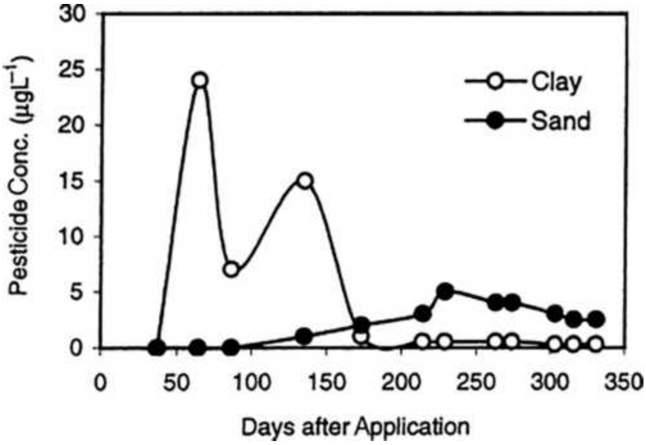


FIGURE 1 Concentrations of a pesticide in water leaching from 1-m undisturbed soil columns of a clay and a sandy soil.
Source: Modified from Bergström and Stenström.[8]

when high concentrations occur in the soil solution in upper soil layers, in combination with intensive rainfall.[8] Once the pesticide is mixed in with the soil matrix, water moving through preferential flow paths does not interact with the soil, and leaching is therefore reduced. In other words, preferential flow can both increase and decrease pesticide leaching depending on the time when it occurs in relation to pesticide application. The final result is that (although transient flow peaks shortly after application causing elevated concentrations in water leaching through soils in preferential flow paths) the leaching loads over extended periods are typically quite small in such soils. Indeed, leaching loads are often larger in sandy soils, in which water and pesticide movement mainly occurs between individual soil particles within the main soil matrix.[8] Pesticide concentrations are typically lower in sandy soils than in clay soils, but the water volumes displacing the pesticide are often much larger in sandy soils. The principal difference in pesticide leaching patterns in sand and clay soils is illustrated in Figure 1. Irrespective of which leaching mechanism prevails, the total amount of the majority of pesticides that reach groundwater after normal agricultural use rarely exceeds 1% of the applied amount and is commonly well below 0.1%.

Pesticide Properties

The physicochemical properties of pesticides have a major impact on their leachability. In this context, the rate with which they are degraded and how strongly they are sorbed to soil are the most important factors. As a general rule of thumb, leaching decreases with increasing sorption affinity and faster degradation, and increases when the opposite conditions prevail. However, it is important to note that degradation rates often become slower with residence time in the soil as a result of decreased availability due to sorption in the soil. This means that strongly sorbed pesticides, which are less mobile in soil than weakly sorbed compounds, are typically quite persistent. An example of a leaching classification scheme, based on the sorption strength of some pesticides, is shown in Table 1.

A factor that complicates the picture of pesticide movement in soil in relation to sorption affinity is the possibility that pesticide mobility is enhanced by adsorption to various mobile colloids, a process often referred to as “facilitated transport.” It is known that organic solutes, such as nonpolar pesticides with very low water solubilities, can form complexes with dissolved organic carbon and clay colloids that move through soil. Even though the role of colloids in facilitating pesticide transport is still relatively poorly understood, there is little doubt that failure to account for this mode of transport can lead to underestimates of both amounts and distances that strongly sorbed pesticides may migrate through the unsaturated zone.

TABLE 1 Classification of Pesticide Mobility in Soil Based on Their Sorption Strength, Which in This Case Is Expressed by Their K_{oc} Values (Soil Sorption Coefficient, Normalized to the Soil Organic Carbon Content)

Koc Value	Expected Mobility	Type of Pesticide
0–50	Very high	Bentazone, Dicamba
50–150	High	Atrazine, 2,4-D
150–500	Medium	Simazine, Metolachlor
500–2000	Low	Lindane, Linuron
2000–5000	Very low	Phenmedipham, Fenpropimorph
>5000	Immobile	DDT, Paraquat

Source: Modified from Torstensson.[9]

Climatic Conditions

The amount and intensity of precipitation are the most important climatic factors influencing pesticide leaching. Water, in excess of evapotranspiration and what is required to maintain field capacity (i.e., the water content in soil when it is freely drained), leaches through the unsaturated zone and can thereby potentially move pesticides to groundwater. As mentioned above, in clay soils, it is primarily high intensity rainfall soon after application that may displace pesticides to depth in soil. In other words, the timing of precipitation is critical.

Soil temperature also has impact on leaching of pesticides, mainly by influencing the persistence of pesticides in soil and by affecting flow processes. Up to a certain level, degradation rates increase with increasing temperatures, which means that less of the compound will be available for leaching. Increasing temperatures will also increase evapotranspiration rates, which will reduce the amounts of water that can potentially move pesticides downward in soil. In climates with subzero temperatures in the winter season, soil will be frozen during extended periods. Under such conditions, pesticide movement in soil is very restricted, if it occurs at all. Leaching then will occur mainly during autumn and spring, when the soil is unfrozen and the evapotranspiration demand is low. Temperatures will also indirectly affect pesticide leaching by affecting the sorption/desorption process, although this influence is not yet thoroughly investigated and, therefore, less well recognized.

Management Practices and Strategies to Prevent Pollution

Management practices that have a major impact on the amount of pesticides that can move through soils can be grouped into the following categories: cropping/tillage, irrigation, and pesticide application practices.

Due to increased concern over soil erosion and input of pesticides to rivers and lakes, agricultural practices with reduced tillage or no-till management have been introduced. Such practices will also affect water infiltration rates, and therefore pesticide leaching through the unsaturated zone. In the short term, reduced tillage may decrease soil permeability compared with a conventionally tilled soil. However, over a whole growing season, infiltration rates tend to be higher under reduced tillage, especially in clay soils. This is largely due to the fact that reductions in tillage lead to less disruption of macropores in which pesticides can be rapidly transported through soil (see above). Reduced tillage also leaves more crop residues on the soil surface and contributes to reducing compaction of the subsoil caused by heavy equipment; both tend to increase permeability, and thus pesticide leaching. From the standpoint of reducing pesticide leaching, reduced tillage is therefore, in most cases, not a good management option.

As expected, irrigation increases leaching of pesticides by increasing the amount of water that potentially can move through soil. The amount of water, the rate at which it is applied, and the timing of irrigation are important for the same reasons as discussed previously for precipitation. Different irrigation methods (e.g., sprinkler, and drip and furrow irrigation) have also been shown to affect leaching.

Pesticide application strategies that influence pesticide residue levels in soil and thereby potential leaching include prevs. postapplication, split applications, placement methods, and use of different pesticide formulations. However, their influence on pesticide leaching is quite unclear and often overshadowed by other factors. Nevertheless, available data indicate that dividing the dose into two applications instead of one tends to reduce pesticide concentrations and the depth of migration in the subsoil. For similar reasons, pesticide leaching can be restricted by use of "slow-release" formulations in which the active ingredient is mixed with a solid matrix from which it gradually diffuses into the soil over an extended period. Placement of the pesticide instead of broadcasting, which reduces the soil surface area to which the pesticide is applied, also tends to reduce pesticide leaching.

Future Concerns

Leaching of pesticides will undoubtedly continue to be of concern in the foreseeable future, and something that will be considered in various regulatory assessment schemes. In this context, it is important not only to look at the leachability, but also to evaluate the risks associated with leaching and the occurrence of pesticides in groundwater both from a human health and an ecotoxicological point of view.

In the future, there is reason to believe that fewer toxic compounds will be allowed, especially those that show high leachability in soil. There is also reason to believe that, in line with the increasing awareness of problems associated with leaching of pesticides, improved management strategies will be developed that reduce pesticide leaching further.

References

1. Enfield, C.G.; Yates, S.R. Organic Chemical Transport to Groundwater. In *Pesticides in the Soil Environment: Processes, Impacts, and Modeling*; Cheng, H.H. Ed.; Soil Science Society of America Book Series: Madison, WI, 1990; 2, 271–302.
2. *The Lysimeter Concept—Environmental Fate of Pesticides*; Führ, F.; Hance, R.J.; Plimmer, J.R.; Nelson, J.O., Eds. ACS Symposium Series: Washington, DC, 1998; 699, 284.
3. Barbash, J.E.; Resek, E.A. *Pesticides in Ground Water—Distribution, Trends, and Governing Factors*; Ann Arbor Press: Chelsea, MI, 1996; 588.
4. Flury, M. Experimental evidence of transport of pesticides through field soils—a review. J. Environ. Qual. **1996**, *25* (1), 25–45.
5. Carter, A.D. Leaching Mechanisms. In *Pesticide Chemistry and Bioscience—The Food–Environment Challenge*; Brooks, G.T., Roberts, T.R., Eds.; Royal Society of Chemistry: Milton Road, U.K., 1999; 8, 291–301.
6. Bergström, L.F.; Jarvis, N.J. Leaching of dichlorprop, ben-tazon, and 36Cl in undisturbed field lysimeters of different agricultural soils. Weed Sci. **1993**, *41* (2), 251–261.
7. Brown, C.D.; Carter, A.D.; Hollis, J.M. Soils and Pesticide Mobility. In *Environmental Behaviour of Agrochemicals* Roberts, T.R., Kearney, P.C., Eds.; John Wiley and Sons: Chichester, England, 1995; 9, 131–184.
8. Bergström, L.; Stenström, J. Environmental fate of pesticides in soil. Ambio. **1998**, *27* (1), 16–23.
9. Torstensson, L. Kemiska Bekämpningsmedel—Transport, Bindning och Nedbrytning i Marken. In *Aktuellt från Sveriges Lantbruksuniversitet*; Swedish University of Agricultural Sciences: Uppsala, Sweden, 1987; 357, 36, (in Swedish).

46

Aquatic Communities: Pesticide Impacts

David P.
Kreutzweiser and
Paul K. Sibley

Introduction

A biotic community can be defined as an assemblage of plant or animal species utilizing common resources and cohabiting a specific area. Examples could include a fish community of a stream, an insect community of a forest pond, or a phytoplankton community of a lake. Interactions among species provide ecological linkages that connect food webs and energy pathways, and these interconnections provide a degree of stability, or balance, to the community. Community balance can be described as a state of dynamic equilibrium in which species and their population dynamics within a community remain relatively stable, subject to changes through natural adjustment processes. Toxic effects of pesticides can disrupt these processes and linkages and thereby cause community balance upsets. For example, this can occur when a pesticide has a direct impact on a certain species in a community and reduces its abundance while other unaffected species increase in abundance in response to the reduced competition for food resources or increased habitat availability. Some of the best examples of pesticide impacts on biological communities are found in freshwater studies. Freshwater aquatic communities are usually contained within distinct boundaries or systems, and this generates a high degree of connectivity among species, thereby increasing their susceptibility to pesticide-induced disturbances at the community level.

We examine traditional and developing methods for measuring pesticide impacts on freshwater communities, with emphasis on recent improvements in risk assessment approaches and analyses, and provide some examples for illustration. We then describe some advances in impact mitigation strategies and discuss some ongoing issues pertaining to understanding, assessing, and preventing pesticide impacts including probabilistic risk assessment (PRA), population and ecological modeling, and pesticide interactions with multiple stressors. The integration of improved risk assessment and mitigation approaches and technologies together with information generated from the numerous impact studies available will provide a sound scientific basis for decisions around the use and regulation of pesticides in and near water bodies.

Measuring Impacts on Aquatic Communities

Changes in aquatic communities can be measured directly in water bodies by a number of quantitative and qualitative sampling methods. Descriptions of those methods can be found in any up-to-date text or handbook (e.g., Hauer and Lamberti[1]). Measurements can be in terms of community structure (species composition) or community function (a measurable ecosystem process attributable to a biotic community that causes a change in condition) and can include both direct and indirect effects.[2,3] Community structure is a measure of biodiversity in its most general sense, that is, the number of species or other taxonomic units and their relative abundances. Some community functions are referred to as environmental or ecosystem services. Examples include organic matter breakdown and nutrient cycling that is largely mediated by microbial communities, or water uptake, filtration, and flood control mediated by shoreline plant communities.[4] Both community structure (biodiversity) and function (ecosystem services) are being increasingly valued by society and global economies,[5,6] and therefore sustaining healthy aquatic communities will be an important driver of pesticide impact mitigation efforts.

Detecting impacts of pesticides typically involves repeated sampling and a comparison of community attributes among contaminated and uncontaminated test units over time, or across a gradient of pesticide concentrations. The test units can range from petri dishes to natural ecosystems, with a trade-off between experimental control in small test units and environmental realism in field-level testing and whole ecosystems.[7] In an effort to incorporate both experimental control and environmental realism in pesticide impact testing, the use of microcosms or model ecosystems for measuring impacts on aquatic communities has increased over the past couple of decades.[8,9] Model ecosystems for community-level pesticide testing can be quite simple at lower-trophic levels such as with microbial communities (e.g., Widenfalk et al.[10]) but will necessarily be more complex for testing higher-order biological communities (e.g., Wojtaszek[11]). Regardless of the test units, an important consideration for measuring pesticide impacts will be an assessment of the duration of impact or rate of recovery. A rapid return to pre-pesticide or reference (no pesticide) community condition will reduce the long-term ecological consequences of the pesticide disturbance.[12]

Traditional measures of community-level impacts have focused on structure and have usually been expressed in terms of single-variable indices such as species richness, diversity, or abundance. These indices are useful descriptors of community structure but suffer from the fact that they reduce complex community data to a single summary metric and may miss subtle or ecologically important changes in species composition across sites or times. Over the last couple of decades, ecotoxicologists have increasingly turned to multivariate statistical techniques for analyzing community response data.[13] A variety of multivariate statistical techniques and software are available and are usually considered superior for the analysis of community data because they retain and incorporate the spatial and temporal multidimensional nature of biological communities.[14] This includes various ordination techniques that can provide graphical representation of spatiotemporal patterns in community structure in which points that lie close together in the ordination plot represent communities of similar composition (richness, abundance), while communities with dissimilar species composition are plotted further apart.

Figure 1 illustrates the use of an ordination plot generated by nonmetric multidimensional scaling for detecting differences among aquatic insect communities in four control and eight insecticide-treated streams. These data have been adjusted for illustrative purposes but are based on real invertebrate community responses to an insecticide in outdoor stream channels.[15] At both concentrations of the insecticide, the community structure of stream insects clearly shifted away from the natural community composition in control streams as depicted by the separation of treated streams (T1 and T2) from controls (C) in the ordination bi-plot. The plot also illustrates that the variability among treated streams (relative distance between points) was greater than that among control streams, that the low-concentration streams (T1) and high-concentration streams (T2) tended to separate along axis 1, and that the T2 streams were further removed from controls than the T1 streams, indicating a differential response by the insect communities to the two test concentrations. Canonical correspondence analysis

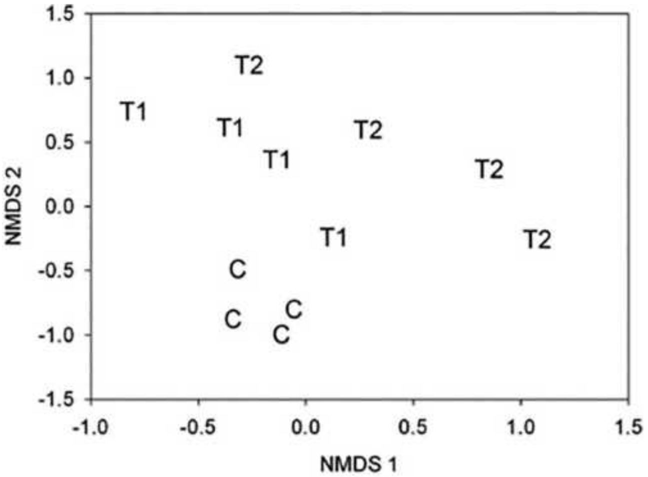


FIGURE 1 Ordination by nonmetric multidimensional scaling of aquatic insect communities in stream channels. Each point represents the community structure of control channels (C) and channels treated with a neem-based insecticide at a low (T1) and high (T2) concentration.
Source: Adapted from Kreutzweiser et al.[15]

and redundancy analysis have also been commonly used to assess aquatic community responses to pesticide contamination.[16,17] A useful refinement of an ordination technique for detecting and interpreting pesticide impacts on aquatic communities is principal response curves (PRCs).[18] PRC is derived from redundancy analysis, and time-dependent responses in the treatments are expressed as deviations from the control or reference system allowing for clear visualization of pesticide effects.

Assessing Risk of Pesticide Impacts on Aquatic Communities

The likelihood or risk of harmful effects on aquatic communities from exposure to pesticides will depend on the exposure concentration, bioavailability, exposure duration, rate of uptake, inherent species sensitivities, community composition, and other community attributes. All of these must be measured, estimated, modeled, or predicted to derive an assessment of risk to aquatic communities for any given pesticide. Formalized risk assessment frameworks and guidelines for pesticides have been developed in the United States,[19] the European Union,[20] Canada,[21] and elsewhere and can be consulted for detailed information on the various components of a risk assessment. In brief, pesticide risk assessments typically include the following phases: 1) defining the problem by determining the pesticide use patterns and developing conceptual models and hypotheses around how it is expected to behave, the anticipated exposure regimes, the kinds of organisms that are likely to be at risk, the community or entity that is to be protected, and the level of protection that will be acceptable; 2) developing the measurement endpoints for assessing risk of harm by establishing which response measurements are relevant and applicable, and how the measurements will be made; 3) outlining the risk assessment process by specifying the kinds of data to be used and how they will be derived including simulation modeling, empirical laboratory, microcosm or field testing, their appropriate spatial and temporal scales, and their statistical analyses; 4) applying the risk assessment by running models or collecting data, completing analyses, summarizing outputs, and providing risk estimates; 5) conducting risk communication and management by answering questions posed in the problem formulation, suggesting risk mitigation strategies if necessary, and communicating those to appropriate users; and 6) conducting follow-up monitoring to evaluate the success of mitigation strategies and to implement adaptive management to address deficiencies if or when necessary.[22,23]

Traditionally, pesticide risk assessments have relied on standardized, single-species toxicity tests to predict effects on communities, the underlying assumption being that protecting the most sensitive species will protect whole communities. In this case, the selection and relevance of test species are critically important to a successful and meaningful risk assessment.[24] However, the accuracy and relevance of estimating the potential risk to aquatic communities can be greatly improved by consideration of specific species or community attributes. In particular, attribute information can improve the ecological relevance and predictive capabilities of conceptual models and the generation of hypotheses in the risk assessment process. Insofar as these attributes affect exposure, sensitivity, or both, they can increase or decrease risk beyond what could be determined from toxicity estimates or species sensitivity distributions alone.

Behavioral attributes can elevate the risk of pesticide effects on species by increasing the likelihood of intercepting the stressor. For example, young-of-the-year bluefish *(Pomatomus saltatrix)* typically feed in estuaries during their early life stages where agricultural runoff can elevate concentrations of pesticides in food items. This feeding behavior can result in bioaccumulation and in adverse effects such as reduced migration, overwinter survival, and recruitment success in fish communities.[25] Incorporating this kind of information into conceptual models and risk hypotheses will generate more realistic risk assessments. In addition, behavioral attributes themselves can be relevant measurement endpoints if the pesticide mode of action indicates risk of sublethal behavioral effects at expected concentrations. For example, some pesticides have been shown to impair the ability to capture prey in fish[26] and the ability to avoid predators in zooplankton.[27] These types of adverse effects can disrupt trophic linkages and reduce survival or reproduction, thus impacting community balance.

Inclusion of life history information into conceptual models and risk hypotheses can also refine and improve the risk assessment process. Life history strategies can influence a species susceptibility to a stressor through effects on a population's resilience or ability to recover from disturbance.[28] Different species exposed to the same pesticide and experiencing similar levels of effect in terms of population declines do not necessarily recover at the same rates when recovery is dependent on reproduction or dispersion. Populations of organisms with short regeneration times (e.g., several generations per year) and/or high dispersal capacity have higher likelihood of recovery from pesticide-induced population declines than those with longer regeneration periods and limited dispersal capacity. These differential life history strategies and their influences on community response and recovery from pesticide effects have been demonstrated empirically (e.g., van den Brink et al.[29] and Kreutzweiser et al.[30]) and through population modeling.[31] These community balance upsets could not have been predicted from screening-level toxicity data or from species sensitivity data; thus, inclusion of life history information in conceptual models can improve risk hypotheses and direct the assessment to focus on species at higher risk owing to specific life history strategies.

Life history attributes can also influence the risk of pesticide effects through differential life-stage sensitivity or susceptibility. Early life stages are often (but not exclusively) more sensitive to pesticides than later stages. An organism's life stage can also influence its susceptibility to a pesticide by increasing or decreasing the likelihood of intercepting the stressor. If a contaminant is present in the environment at effective concentrations during a period in which the particular life stage of a species is present, then the risk to that species is increased. For some amphibians, aquatic (larval) stages could be at higher risk of direct and indirect effects of pesticides than their terrestrial (adult) life stages when their larval stage coincides with pesticide contamination of water bodies.[32] Thus, while a species sensitivity and geographical distribution may indicate potential risk, the life-stage information coupled with pesticide use pattern, timing, or fate information may indicate little likelihood of exposure to the pesticide and the risk assessment can be adjusted accordingly.

Functional attributes may also be important for refining or improving pesticide risk assessments. Protection goals for populations and communities often include the safeguarding of critical biological processes or ecosystem function. Measuring ecosystem function integrates responses of component populations and can be a relevant measurement endpoint when species loss affects ecosystem function such as energy transfer and organic matter cycling.[33] However, most ecosystems are complex and it may not be clear which functional attributes are critical for sustaining ecological processes or the extent to which they can sustain changes in structural properties (e.g., population levels, diversity) without adversely affecting ecosystem function. Neither is it clear if functional endpoints are more or less sensitive than structural endpoints for detecting ecosystem disturbance. Some studies investigating the relationship between species diversity and ecosystem function have indicated that ecosystems can tolerate some species loss because of functional redundancy.[34] Functional redundancy is thought to occur when several species perform similar functions in ecosystems such that some may be eliminated with little or no effect on ecosystem processes. Others have suggested that redundant species are required to ensure ecosystem resilience to disturbance as a form of biological insurance, especially at large spatial scales.[35]

Given these discrepancies, measurement endpoints based on functional attributes are not typically used in pesticide risk assessments because it is generally accepted that protection of community structure will protect ecosystem function. However, when specific functional attributes can be identified and are known or suspected to be at risk from a pesticide, they can be included in the data requirements for a risk assessment. An example would be the risk of adverse effects on leaf litter decomposition (a critical ecosystem function in forest soils and water bodies) posed by a systemic insecticide for control of wood-boring insects in trees.[36] In that case, the protection goal was maintaining leaf litter decomposition, the community at risk was decomposer invertebrates feeding on leaves from insecticide-treated trees, and the selection of test species was directed to a specific functional group because of the unique route of exposure to decomposer organisms identified in the risk hypotheses.

Some Examples of Pesticide Impacts on Aquatic Communities

A few examples will serve to illustrate how pesticides can cause disruptions to aquatic communities. DeNoyelles et al.[37] reviewed studies into pesticide impacts on aquatic communities and reported that herbicides like atrazine, hexazinone, and copper sulfate were directly toxic to most species of phytoplankton (waterborne algae). After herbicide applications, reductions in phytoplankton caused secondary reductions in herbivorous zooplankton, resulting from a depleted food source for the zooplankton. They further showed that direct adverse effects on phytoplankton can also cause disruptions to the bacterial-based energy pathways by reducing carbon flow from phytoplankton to bacteria, and ultimately to grazing protozoans and zooplankton. Boyle et al.[38] found that applications of the insecticide diflubenzuron to small ponds reduced populations of several aquatic invertebrate species. This in turn resulted in indirect effects on algae (increased productivity because of release from grazing pressure by the invertebrates) and on juvenile fish populations (reduced production because of limited invertebrate prey availability). George et al.[39] used a novel approach to predict effects of pesticide mixtures on zooplankton communities and then tested the predictions in outdoor microcosms. Responses among zooplankton populations within the community differed, depending on the pesticide mixture, and those differences appeared to reflect the relative susceptibilities among specific taxa within groups. Cladocerans declined but were less sensitive than copepods to a chlorpyrifos-dominated mixture, while rotifers actually increased after application in response to release from competition or predation pressures.

Kreutzweiser et al.[40] applied a neem-based insecticide to forest pond enclosures and measured effects on zooplankton community structure, respiration, and food web stability. Significant concentration-dependent reductions in numbers of adult copepods were observed, but immature copepods and cladocerans were unaffected (Figure 2). There was no evidence of recovery of adult copepods within the sampling season. During the period of maximal impact (about 4 to 9 weeks after the applications), total plankton community respiration was significantly reduced, and this contributed to significant concentration-dependent increases in dissolved oxygen and decreases in specific conductance. The reductions in adult copepods resulted in negative effects on zooplankton food web stability through elimination of a trophic link and reduced interactions and connectance.

Van Wijngaarden et al.[41] evaluated the responses of aquatic communities in indoor microcosms to a suite of pesticides used for bulb crop protection. At pesticide concentrations equivalent to 5% spray drift deposition, zooplankton taxa within communities showed significant changes relative to non-treated controls, reflecting taxon-specific sensitivities. Some copepods and rotifers in particular showed significant declines for at least 13 weeks, while many other rotifers and cladocerans were unaffected or increased. Several macroinvertebrate taxa were negatively affected, and this contributed to significant declines in leaf litter decomposition among treated microcosms. The herbicide asulam was among the suite of pesticides, and it induced significant reduction of the macrophyte *Elodea nuttallii*. This in turn caused significant changes in water chemistry (decreases in dissolved oxygen and pH, increases in alkalinity and specific conductance) and increases in phytoplankton biomass from decreased competition for nutrients. Increased phytoplankton and reduced zooplankton predators combined to support higher abundance of less sensitive zooplankton taxa. The authors point out that most of these effects were not measurable at more realistic rates of spray drift deposition.

Relyea and Hoverman[42] investigated impacts of the insecticide malathion on aquatic communities in microcosms designed to mimic a simple aquatic food web that can be found in ponds and wetlands. The insecticide generally reduced zooplankton abundance, and these reductions stimulated increases in phytoplankton, decreases in periphyton (attached algae), and decreases in growth of frog tadpoles. While invertebrate predator survival was not affected, amphibian prey survival increased with insecticide concentration, apparently the result of insecticide-induced impairment of predation success by the invertebrates. Overall, the study demonstrated that realistic concentrations of an insecticide can interact with natural predators to induce large changes in aquatic community balance.

Reducing Risk of Pesticide Impacts on Aquatic Communities

For pesticides applied to crops and forests, exposure to aquatic communities can be minimized by the implementation of vegetated spray buffers or setbacks to intercept off-target spray drift and runoff.[43] Pesticide runoff can be further reduced by using formulations that are less prone to wash-off, leaching, and mobilization. Recent advances in spray drift reduction and improved spray guidance systems can also significantly reduce the off-target movement of pesticides to water bodies.[44] Examples include new technologies in map-based automated boom systems for row crops[45] and Geographical Information System (GIS)-based landscape analysis for predicting off-target pesticide movement.[46]

The risk of adverse effects on aquatic communities may also be decreased by intentional selection and use of pesticides that are inherently safer to the environment. This would include so-called reduced-risk pesticides that are bioactive compounds usually with unique modes of action and derived from microbial, plant, or other natural sources. These are generally thought to be less persistent and toxic to non-target organisms than conventional synthetic pesticides.[47] Examples include the bacteria-derived insecticide *Bt (Bacillus thuringiensis),* the plant-derived insecticide neem, and the microbe-derived herbicide phosphinothricin. However, Thompson and Kreutzweiser[48] caution that it cannot be assumed that this group of pesticides is inherently safer or more environmentally acceptable than synthetic counterparts and that full environmental risk evaluations must be conducted to ensure their environmental safety.

These types of technologies combined with the use of non-pesticide approaches to pest management form the basis of integrated pest management (IPM) strategies. IPM strategies are those in which the judicious use of pesticides is only one of several concurrent methods to control or manage losses from pest damage. This can include the use of natural enemies and parasites, biological control agents, insect growth regulators, confusion pheromones, sterile male releases, synchronizing with weather patterns known to diminish pest populations, and cultivation methods and crop varieties to improve conditions for natural enemies or degrade conditions for pest survival.[49] Increasing the use of IPM approaches can reduce reliance on pesticides and thus reduce the risk of pesticide impacts overall.

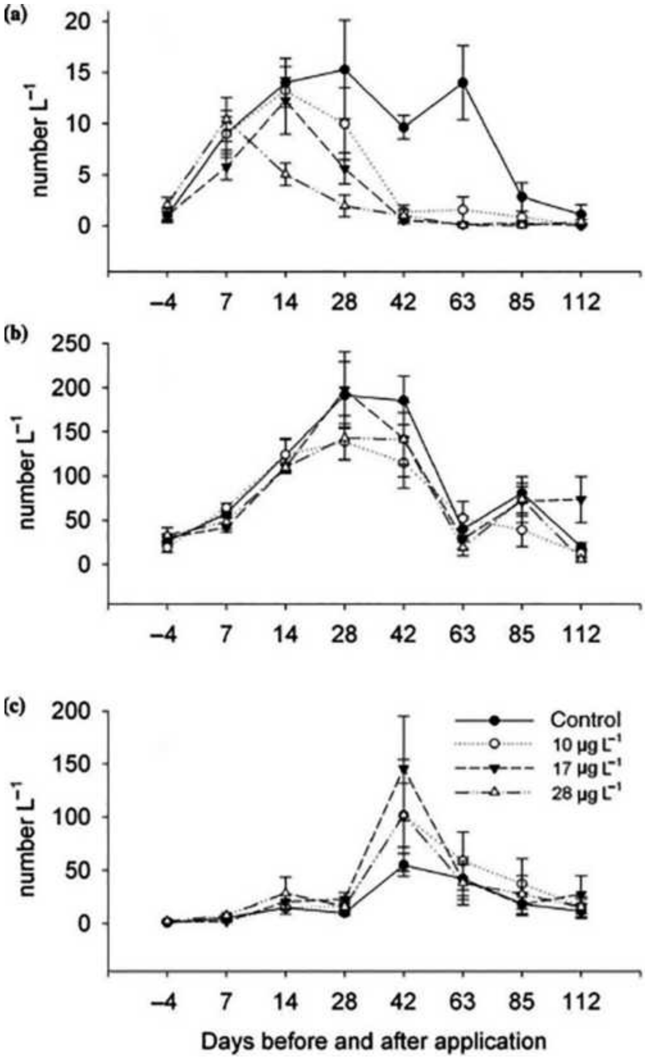


FIGURE 2 Mean abundance (±1 SE, n = 5) of (a) adult copepods, (b) immature copepods, and (c) cladocerans in natural pond microcosms (controls) and microcosms treated at three different rates of a neem-based insecticide. **Source:** Taken from Kreutzweiser et al.[40]

Recent Advances and Outstanding Issues

Pesticide risk assessments and risk reductions have recently been advanced in terms of ecological realism and effectiveness through some developing methods and techniques. Traditional risk assessments have estimated hazards from pesticides by comparing the expected environmental concentration (often predicted from worst-case scenarios) to the toxic threshold for the most sensitive test species. When the expected concentration is higher than the toxicity threshold, the pesticide is considered to have potential for environmental effects. These so-called hazard or risk quotient approaches are still widely used in pesticide risk assessment and regulation, but more recently, PRA and probabilistic hazard assessment (PHA) approaches are being adopted. In these approaches, pesticide exposure levels and the likelihood of toxic effects are estimated from probability distributions based on all reliable data available.[50] In PRA, exposure and effects distributions are developed from modeling or measurements in laboratory, microcosm, or field studies and used to improve the accuracy and relevance of the estimated likelihood of environmental effects compared to the traditional worst-case (hazard/risk quotient) approach (e.g., Solomon[51]). In PHA, a distribution approach is also used, except that the probability of hazard is estimated from distributions built on the relative sensitivity of interspecies endpoints rather than species sensitivity itself.[52] Figure 3 illustrates the principles of PRA (Figure 3a) and PHA (Figure 3b). Regardless of the approach, one important aspect of PRA that is ongoing is the development and use of uncertainty analysis to quantify variability and uncertainty in exposure and effects estimates. Characterizing and quantifying uncertainty will provide more meaningful risk assessments and improved decision making for minimizing potential risk of pesticide impacts in or near water.[53]

Efforts at incorporating population or ecological modeling into pesticide risk assessments have also improved their accuracy and relevance for predicting, and therefore mitigating, risk of harm to aquatic communities.[54] The use of ecological models to incorporate a suite of factors including lethal and sublethal effects and their influences on the risks to organisms, populations, or communities can provide useful insights into receptor/pesticide interactions and can thereby improve risk assessments and direct

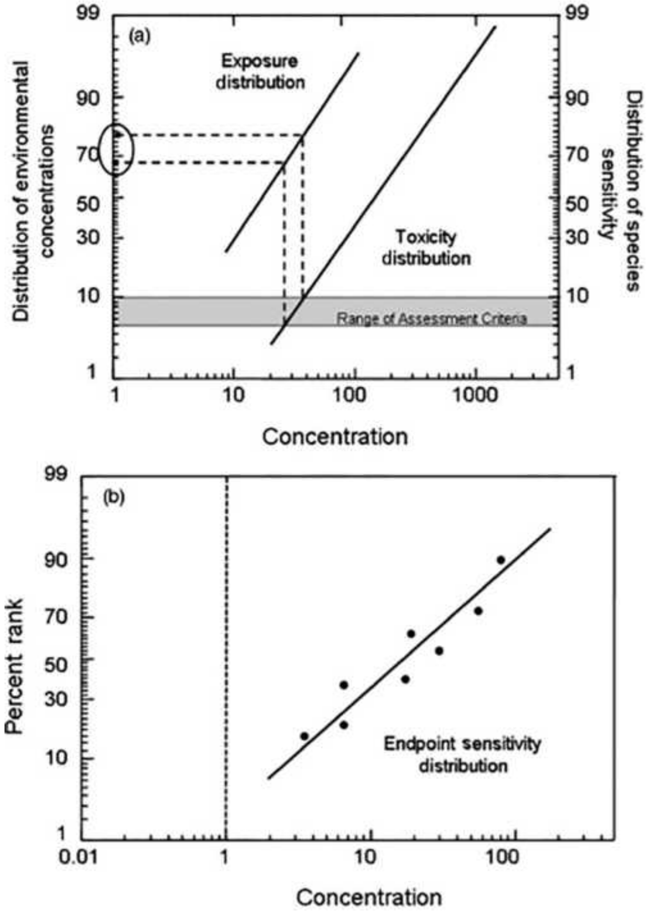


FIGURE 3 Schematic illustrating the principle of PRA (a) and PHA (b). PRA is based on a comparison of exposure and effects distributions using a predetermined criterion typically in the range of 5%-10% (shaded area and dashed lines in panel A) to determine the probability of exceeding the criterion (ellipse on y-axis); PHA is based on a comparison of an endpoint-derived sensitivity distribution within a test species to a threshold value such as a hazard quotient (dashed line in panel B).

mitigation measures. Population models that account for differential demographics and population growth rates within communities have been shown to provide more accurate assessments of potential pesticide impacts on populations and communities than what conventional lethal concentration estimates can provide.[55] Ecological and population modeling combined with pesticide exposure modeling and case-based reasoning (drawing on past experience or information from similar chemical exposures) can provide further refinements and improve risk assessment for aquatic communities.[56] Another recent advancement in ecological modeling to predict pesticide effects is the use of trait-based information such as organism morphology, life history, physiology, and feeding ecology in risk assessments.[57] This approach includes some of the functional attributes and concepts described above in the section on "Assessing Risk of Pesticide Impacts on Aquatic Communities" and has the advantage of formally expressing communities as combinations of functional traits rather than as groups of species, thereby yielding a more meaningful description of community structure and function. Taken together, these modeling approaches that incorporate probability distributions, toxicological sensitivities, population dynamics, ecological information, and functional trait attributes can be integrated into improved risk assessments that will inform mitigation and prevention strategies for pesticide use.[58]

Two additional issues that present challenges to pesticide risk assessment and mitigation are pesticide mixtures and the combined or cumulative effects of multiple stressors on pesticide impacts. Pesticides frequently occur as mixtures in aquatic systems, particularly in agricultural regions, and methods to assess and/or predict pesticide mixture toxicity under laboratory conditions have been relatively well developed. However, there are still large uncertainties associated with the prediction of pesticide mixture toxicity, and additional studies are needed to evaluate the performance of mixture models when evaluating community-level endpoints and toxicity thresholds over long-term exposures.[59] Secondly, whereas most pesticide assessment data are derived from tests or experiments under controlled or semicontrolled environmental conditions, pesticides in natural environments may interact with a number of other natural or human-caused stressors that can substantially alter the likelihood and magnitude of pesticide impacts.[60] Other stressors could include overarching effects of climate change that can influence water temperature and quality; land use activities that result in chemical, sediment, and nutrient pollution of waterways; and biotic interactions with invasive species in aquatic communities. A number of studies have examined the combined effects of a pesticide with other stressors, but they have usually been single stressor effects tested at the single-species level. Examples of studies that examined combined effects include pesticide interactions with water temperature,[61] pH,[62] dissolved organic matter,[63] UV radiation,[64] predators,[65] competitors,[66] food availability,[67] elevated sediments,[68] and other chemical stressors.[69] However, potential multiple stressors and their interactions with pesticides can be myriad and testing or extrapolating to community-level impacts is onerous at best. Sorting out and mitigating pesticide impacts from among these multiple stressors continues to be a challenge, and the suggestion by Laskowski et al.[70] to include studies of toxicant interactions with a range of environmental conditions in risk assessments seems warranted.

Conclusions

Because of the high degree of connectivity among species in an aquatic community, pesticides pose a risk of harm to the community stability or balance. The community structure can be altered by direct effects, indirect effects, or both, and this can cause disruptions to the interactions and linkages among species and to their ecological function. This risk of harm will depend on exposure concentration, bioavailability, exposure duration, rate of uptake, species sensitivities, community composition, and other community attributes. Recent advances in pesticide risk assessment for aquatic communities have improved the ecological relevance and predictive capabilities for determining, and thus mitigating, potential harmful impacts. Pesticide impacts on aquatic communities can be minimized by the use of improved application technologies to reduce application rates and to decrease off-target movement to water bodies. Potential impacts can be further minimized through the selection and use of pesticides that are demonstrated to be inherently safer to the environment and through the application of IPM strategies. Given the preponderance of pesticide impact studies in freshwater aquatic ecosystems, the improved risk assessment frameworks and regulatory requirements for pesticide evaluations, and the recent advances in mitigation technologies, many decisions around the use of pesticides can be made on a sound scientific basis rather than on misinformed perceptions or politically driven agendas. Integrated, science-based pest management strategies including the prudent use of appropriate pesticides will contribute to ensuring the sustainability of aquatic communities in areas subjected to pest management programs.

References

1. Hauer, F.R.; Lamberti, G.A. *Methods in Stream Ecology,* 2nd Ed.; Elsevier: Amsterdam, 2006.
2. Fleeger, J.W.; Carman, K.R.; Nisbet, R.M. Indirect effects of contaminants in aquatic ecosystems. Sci. Tot. Environ. **2003**, *317,* 207–233.
3. Rohr, J.R.; Crumrine, P.W. Effects of an herbicide and an insecticide on pond community structure and processes. Ecol. Appl. **2005**, *15,* 1135–1147.
4. Daily, G.C. Introduction: What are ecosystem services? In *Nature's Services: Societal Dependence on Natural Ecosystems*; Daily, G.D., Ed.; Island Press: Washington, DC, 1997; 1–10.
5. Anderson, J.; Gomez, W.C.; McCarney, G.; Adamowicz, W.; Chalifour, N.; Weber, M.; Elgie, S.; Howlett, M. *Ecosystem Service Valuation, Market-based Instruments and Sustainable Forest Management: A Primer*; Sustainable Forest Management Network: Edmonton, Alberta, 2010.
6. Bayon, R.; Jenkins, M. 2010. The business of biodiversity. Nature **2010**, *466,* 184–185.
7. Sibley, P.K.; Chappel, M.J.; George, T.K.; Solomon, K.R.; Liber, K. Integrating effects of stressors across levels of biological organization: Examples using organophosphorus insecticide mixtures in field-level exposures. J. Aquat. Eco-syst. Stress Recovery **2000**, *7,* 117–130.
8. Campbell, P.J.; Arnold, D.J.S.; Brock, T.C.M.; Grandy, N.J.; Heger, W.; Heimbach, F.; Maund, S.J.; Streloke, M. *Guidance Document on Higher-Tier Aquatic Risk Assessment for Pesticides;* SETAC Europe: Brussels, Belgium, 1999.
9. Kennedy, J.H.; LaPoint, T.W.; Balci, P.; Stanley, J.K.; Johnson, Z.B. Model aquatic ecosystems in ecotoxicological research: Considerations of design, implementation, and analysis. In *Handbook of Ecotoxicology,* 2nd Ed.; Hoffman, D.J.; Rattner, B.A.; Burton, G.A., Jr.; Cairns, J., Jr., Eds.; Lewis Publishers: Boca Raton, Florida, 2003; 45–74.
10. Widenfalk, A.; Svensson, J.M.; Goedkoop, W. Effects of the pesticides captan, deltamethrin, isoproturon, and pirimi-carb on the microbial community of a freshwater sediment. Environ. Toxicol. Chem. **2004**, *23,* 1920–1927.
11. Wojtaszek, B.F.; Buscarini, T.M.; Chartrand, D.T.; Stephenson, G.R.; Thompson, D.G. Effect of Release® herbicide on morality, avoidance response, and growth of amphibian larvae in two forest wetlands. Environ. Toxicol. Chem. **2005**, *24*, 2533–2544.
12. Barnthouse, L.W. Quantifying population recovery rates for ecological risk assessment. Environ. Toxicol. Chem. **2004**, *23*, 500–508.
13. Maund, S.; Chapman, P.; Kedwards, T.; Tattersfield, L.; Matthiessen, P.; Warwick, R.; Smith, E. Application of multivariate statistics to ecotoxicological field studies. Environ. Toxicol. Chem. **1999**, *18,* 111–112.
14. Clarke, K. R.; Warwick, R.M. *Change in Marine Communities: An Approach to Statistical Analysis and Interpretation, 2nd Ed.*; PRIMER-E: Plymouth, U.K., 2001.
15. Kreutzweiser, D.P.; Capell, S.S.; Scarr, T.A. Community-level responses by stream insects to neem products containing azadirachtin. Environ. Toxicol. Chem. **2000**, *19,* 855–861.
16. Frieberg, N.; Lindstrom, M.; Kronvang, B.; Larsen, S.E. Macroinvertebrate/sediment relationships along a pesticide gradient in Danish streams. Hydrobiologia **2003**, *494,* 103–110.

17. Berenzen, N.; Kimke, T.; Schulz, H.K.; Schulz, R. Macroinvertebrate community structure in agricultural streams: Impact of run-off-related pesticide contamination. Ecotoxicol. Environ. Saf. **2005**, *60,* 37–46.
18. van den Brink, P.J.; ter Braak, C.J.F. Principal response curves: Analysis of time-dependent multivariate responses of biological community to stress. Environ. Toxicol. Chem. **1999**, *18,* 138–148.
19. USEPA. *Guidelines for Ecological Risk Assessment*; United States Environmental Protection Agency, Risk Assessment Forum: Washington, DC, 1998.
20. EUFRAM. *Introducing Probabilistic Methods into the Ecological Risk Assessment of Pesticides, Version 6*; European Framework for Risk Assessment of Pesticides (EUFRAM): York, U.K., 2005.
21. Delorme, P.; Francois, D.; Hart, C.; Hodge, V.; Kaminski, G.; Kriz, C.; Mulye, H.; Sebastien, R.; Takacs, P.; Wandel-maier, F. *Final Report for the PMRA Workshop: Assessment Endpoints for Environmental Protection*; Environmental Assessment Division, Pest Management Regulatory Agency, Health Canada: Ottawa, Ontario, 2005.
22. Suter, G.W.; Barnthouse, L.W.; Bartell, S.M.; Mill, T.; Mackay, D.; Patterson, S. *Ecological Risk Assessment*; Lewis Publishers: Boca Raton, Florida, 1993.
23. Reinert, K.H.; Bartell, S.M.; Biddinger, G.R., Eds. *Ecological Risk Assessment Decision-Support System: A Conceptual Design;* SETAC Press: Pensacola, Florida, 1998.
24. Maltby, L.; Blake, N.; Brock, T.C.M.; van den Brink, P.J. Insecticide species sensitivity distributions: Importance of test species selection and relevance to aquatic ecosystems. Environ. Toxicol. Chem. **2005**, *24,* 379–388.
25. Candelmo, A.C.; Deshpande, A.; Dockum, B.; Weis, P.; Weis, J.S. The effect of contaminated prey on feeding, activity, and growth of young-of-the-year bluefish, *Pomato-mus saltatrix,* in the laboratory. Estuaries Coasts **2010**, *33,* 1025–1038.
26. Baldwin, D.H.; Spromberg, J.A.; Collier, T.K.; Scholz, N.L. A fish of many scales: Extrapolating sublethal pesticide exposures to the productivity of wild salmon populations. Ecol. Appl. **2009**, *19,* 2004–2015.
27. Pestana, J.L.T.; Loureiro, S.; Baird, D.J.; Soares, A.M.V.M. Pesticide exposure and inducible antipredator responses in the zooplankton grazer, *Daphnia magna* Straus. Chemo-sphere **2010**, *78,* 241–248.
28. Stark, J.D.; Banks, J.E.; Vargas, R.I. How risky is risk assessment: The role that life history strategies play in susceptibility of species to stress. Proc. Natl. Acad. Sci. U. S. A. **2004**, *101,* 732–736.
29. van den Brink, P.J.; Hattink, J.; Bransen, F.; van Donk, E.; Brock, T.C.M. Impact of the fungicide carbendazim in freshwater microcosms. II. Zooplankton, primary producers and final conclusions. Aquat. Toxicol. **2000**, *48*, 251–264.
30. Kreutzweiser, D.P.; Back, R.C.; Sutton, T.M.; Pangle, K.L.; Thompson, D.G. Aquatic mesocosm assessments of a neem (azadirachtin) insecticide at environmentally realistic con-centrations—2: Zooplankton community responses and recovery. Ecotoxicol. Environ. Saf. **2004**, *59,* 194–204.
31. Wang, M.; Grimm, V. Population models in pesticide risk assessment: Lessons for assessing population-level effects, recovery, and alternative exposure scenarios from modeling a small mammal. Environ. Toxicol. Chem. **2010**, *29,* 1292–1300.
32. Brodman, R.; Newman, W.D.; Laurie, K.; Osterfeld, S.; Lenzo, N. Interaction of an aquatic herbicide and predatory salamander density on wetland communities. J. Herpetol. **2010**, *44*, 69–82.
33. Rosenfeld, J.S. Functional redundancy in ecology and conservation. Oikos **2002**, *98,* 156–162.
34. Lawton, J.H.; Brown, V.K. Redundancy in ecosystems. In *Biodiversity and Ecosystem Function*; Schulze, E.D., Mooney, H.A., Eds.; Springer: New York, 1993; 255–268.
35. Naeem, S.; Li, S. Biodiversity enhances ecosystem stability. Nature **1997**, *390,* 507–509.
36. Kreutzweiser, D.P.; Good, K.P.; Chartrand, D.T.; Scarr, T.A.; Thompson, D.G. Are leaves that fall from imidacloprid-treated maple trees to control Asian longhorned beetles toxic to non-target decomposer organisms? Journal of Environmental Quality **2008**, *37,* 639–646.

37. deNoyelles, F., Jr.; Dewey, S.L.; Huggins, D.G.; Kettle, W.D. Aquatic mesocosms in ecological effects testing: Detecting direct and indirect effects of pesticides. In *Aquatic Mesocosm Studies in Ecological Risk Assessment*; Graney, R.L., Kennedy, J.H., Rodgers, J.H., Jr., Eds.; Lewis Publishers: Boca Raton, 1994; 577–603.
38. Boyle, T.P.; Fairchild, J.F.; Robinson-Wilson, E.F.; Haver-land, P.S.; Lebo, J.A. Ecological restructuring in experimental aquatic mesocosms due to the application of diflubenzuron. Environ. Toxicol. Chem. **1996**, *15,* 1806–1814.
39. George, T.K.; Liber, K.; Solomon, K.R.; Sibley, P.K. Assessment of the probabilistic ecological risk assessment-toxic equivalent combination approach for evaluating pesticide mixture toxicity to zooplankton in outdoor microcosms. Archives of Environmental Contamination and Toxicology **2003**, *45,* 453–461.
40. Kreutzweiser, D.P.; Sutton, T.M.; Back, R.C.; Pangle, K.L.; Thompson, D.G. Some ecological implications of a neem (azadirachtin) insecticide disturbance to zooplankton communities in forest pond enclosures. Aquat. Toxicol. **2004**, *67,* 239–254.
41. van Wijngaarden, R.P.A.; Cuppen, J.G.M.; Arts, G.H.P.; Crum, S.J.H.; van den Hoorn, M.W.; van den Brink, P.J.; Brock, T.C.M. Aquatic risk assessment of a realistic exposure to pesticides used in bulb crops: a microcosm study. Environ. Toxicol. Chem. **2004**, *23,* 1479–1498.
42. Relyea, R.A.; Hoverman, J.T. Interactive effects of predators and a pesticide on aquatic communities. Oikos **2008**, *117,* 1647–1658.
43. Zhang, X.; Liu, X.; Zhang, M.; Dahlgren, R.A.; Eitzel, M. A review of vegetated buffers and a meta-analysis of their mitigation efficacy in reducing nonpoint source pollution. J. Environ. Qual. **2010**, *39,* 76–84.
44. van de Zande, J.C.; Porskamp, H.A.; Michielsen, J.M.; Holterman, H.J.; Juijsmans, J.M. Classification of spray applications for driftability to protect surface water. Aspects Appl. Biol. **2000**, *57,* 57–64.
45. Luck, J.D.; Zandonadi, R.S.; Luck, B.D.; Shearer, S.A. Reducing pesticide over-application with map-based automatic boom section control on agricultural sprayers. Trans. Am. Soc. Agric. Biol. Eng. **2010**, *53*, 685–690.
46. Pfleeger, T.G.; Olszyk, D.; Burdick, C.A.; King, G.; Kern, J.; Fletcher, J. Using a geographical information system to identify areas with potential of off-target pesticide exposure. Environ. Toxicol. Chem. **2006**, *25,* 2250–2259.
47. PMRA. *Regulatory Directive DIR2002–02: The PMRA Initiative for Reduced Risk Pesticides;* Pest Management Regulatory Agency, Health Canada Information Services: Ottawa, Ontario; 2002, availabe at http://www.hc-sc.gc.ca/pmra-arla/english/pdf/dir/dir2002–02-e.pdf. (accessed September 2010).
48. Thompson, D.G.; Kreutzweiser, D.P. A review of the environmental fate and effects of natural "reduced-risk" pesticides in Canada. In *Crop Protection Products for Organic Agriculture: Environmental, Health, and Efficacy Assessment, ACS Symposium Series 947*; Felsot, A.S., Racke, K. D., Eds.; American Chemical Society: Washington, DC, 2007; 245–274.
49. van Emden, H. Integrated pest management. In *Encyclopedia of Pest Management*; Pimentel, D., Ed.; Marcel Dekker Inc.: New York, 2002; 413–415.
50. Solomon, K.R.; Takacs, P. Probabilistic ecological risk assessment using species sensitivity distributions. In *Species Sensitivity Distributions in Ecotoxicology*; Posthuma, L., Suter, G.W., Traas, T.P., Eds.; Lewis Publishers: Boca Raton, Florida, 2002; 285–314.
51. Solomon, K.R.; Baker, D.B.; Richards, R.P.; Dixon, K.R.; Klaine, S.J.; La Point, T.W.; Kendall, R.J.; Weisskopf, C.P.; Giddings, J.M.; Giesy, J.P.; Hall, L.W.; Williams, W.M. Ecological risk assessment of atrazine in North American surface waters. Environ. Toxicol. Chem. **1996**, *15,* 31–76.
52. Hanson, M.L.; Solomon, K.R. New technique for estimating thresholds of toxicity in ecological risk assessment. Environ. Sci. Technol. **2002**, *36,* 3257–3264.

53. Warren-Hicks, W.J.; Hart, A., Eds. *Application of Uncertainty Analysis to Ecological Risks of Pesticides*; CRC Press: Boca Raton, Florida, 2010.
54. Thorbek, P.; Forbes, V.E.; Heimbach, F.; Hommen, U.; Thulke, H.; van den Brink, P.; Wogram, J.; Grimm, V. *Ecological Models for Regulatory Risk Assessments of Pesticides;* SETAC Press: Pensacola, Florida, 2010.
55. Stark, J.D.; Banks, J.E. Population-level effects of pesticides and other toxicants on arthropods. Annu. Rev. Entomol. **2003**, *48,* 505–519.
56. van den Brink, P.J.; Roelsma, J.; van Nes, E.H.; Scheffer, M.; Brock, T.C.M. PERPEST model: A case-based reasoning approach to predict ecological risks of pesticides. Environ. Toxicol. Chem. **2002**, *21,* 2500–2506.
57. Baird, D.J.; Baker, C.J.O.; Brua, R.B.; Hajibabaei, M.; McNicol, K.; Pascoe, T.J.; de Zwart, D. Towards a knowledge infrastructure for traits-based ecological risk assessment. Integr. Environ. Assess. Manage. **2010**, online DOI 10.1002/ieam.129 (accessed September 2010).
58. van den Brink, P.J. Ecological risk assessment: From bookkeeping to chemical stress ecology. Environ. Sci. Technol. **2008**, *42*, 8999–9004.
59. Beldon, J.B.; Gilliom, R.J.; Lydy, M.J. How well can we predict the toxicity of pesticide mixtures to aquatic life? Integr. Environ. Assess. Manage. **2007**, *3*, 364–372.
60. Heugens, E.H.W.; Hendricks, A.J.; Dekker, T.; van Straalen, N.M.; Admiraal, W. A review of the effects of multiple stressors on aquatic organisms and analysis of uncertainty factors for use in risk assessment. Crit. Rev. Toxicol. **2001**, *31,* 247–285.
61. Lydy, M.J.; Lohner, T.W.; Fisher, S.W. Influence of pH, temperature and sediment type on the toxicity, accumulation and degradation of parathion in aquatic systems. Aquat. Toxicol. **1990**, *17,* 27–44.
62. Howe, G.E.; Marking, L.L.; Bills, T.D.; Rach, J.J.; Mayer, F.L. Jr. Effects of water temperature and pH on toxicity of terbufos, trichlorfon, 4-nitrophenol and 2,4-dinitrophenol to the amphipod *Gammarus pseudolimnaeus* and rainbow trout (*Oncorhynchus mykiss*). Environ. Toxicol. Chem. **1994**, *13,* 51–66.
63. Yang, W.; Spurlock, F.; Liu, W.; Gan, J. Effects of dissolved organic matter on permethrin bioavailability to *Daphnia* species. J. Agric. Food Chem. **2006**, *54,* 3967–3972.
64. Puglis, H.J.; Boone, M.D. Effects of technical-grade active ingredient vs. commercial formulation of seven pesticides in the presence or absence of UV radiation on survival of green frog tadpoles. Arch. Environ. Contam. Toxicol. **2010**, online DOI 10.1007/s00244–010–9528-z (accessed September 2010).
65. Sandland, G.J.; Carmosini, N. Combined effects of a herbicide (atrazine) and predation on the life history of a pond snail, *Physa gyrina*. Environ. Toxicol. Chem. **2006**, *25*, 2216–2220.
66. Davidson, C.; Knapp, R.A. Multiple stressors and amphibian declines: Dual impacts of pesticides and fish on yellowlegged frogs. Ecol. Appl. **2007**, *17,* 587–597.
67. Barry, M.J.; Logan, D.C.; Ahokas, J.T.; Holdway, D.A. Effect of algal food concentration on toxicity of tow agricultural pesticides to *Daphnia carinata*. Ecotoxicol. Environ. Saf. **1995**, *32,* 273–279.
68. Wu, Q.; Riise, G.; Pflugmacher, S.; Greulich, K.; Steinberg, C.E.W. Combined effects of the fungicide propiconazole and agricultural runoff sediments on the aquatic bryophyte *Vesicularia dubyana*. Environ. Toxicol. Chem. **2005**, *24*, 2285–2290.
69. Boone, M.D.; Bridges, C.M.; Fairchild, J.F.; Little, E.E. Multiple sublethal chemicals negatively affect tadpoles of the green frog, *Rana clamitans*. Environ. Toxicol. Chem. **2005**, *24,* 1267–1280.
70. Laskowski, R.; Bednarska, A.J.; Kramarz, P.E.; Loureiro, S.; Schell, V.; Kudlek, J.; Holmstrup, M. Interactions between toxic chemicals and natural environmental factors— A meta-analysis and case studies. Sci. Tot. Environ. **2010**, *408*, 3763–3774.

47

Phosphorus: Riverine System Transport

Andrew N. Sharpley, Peter Kleinman, Tore Krogstad, and Richard McDowell

Introduction

The contribution of anthropogenic phosphorus (P) to the accelerated eutrophication of freshwaters is well documented.[1,2] Sources of riverine P inputs can broadly be divided into point sources, typically dominated by sewage treatment effluents, and diffuse (also "non-point") sources, derived from the landscape.[3,4] Phosphorus export tends to increase from rivers draining native or low-input ecosystems, to intensively managed agricultural systems, to urban settings.[5] Point sources enter the river more continually through the year than do nonpoint sources, which are subject to large seasonal variation, typically as a function of overland flow and land management activities.[6]

Changes in the forms and amount of P that occur as part of the transport processes within streams and rivers can greatly influence the eventual impact of point and non-point sources of P on downstream eutrophication.[6,7] These changes in P are mediated by physical (sediment deposition and resuspension and flow regimes), abiotic (P sorption and desorption), and biotic (microbial and plant uptake) processes.[8,9]

Most importantly, P in riverine systems can play an important role in modifying or delaying societal efforts to curb eutrophication. For instance, P that is already entrained in riverine systems, sometimes referred to as *legacy P*, can serve as a long-term source of P to the overlying water column. Understanding the role of riverine systems as sources, sinks and stores of P is critical to the long-term management of cultural eutrophication.

Riverine Processes

Physical Processes

Fluvial sediments are derived from the erosion of surface soils, gullies, ditches and stream banks. Because surface soils generally contain the highest concentration of P in soil profiles, and erosion preferentially removes P-rich particles, eroded surface soil represents a major source of particulate

P in riverine systems.[10,11] In areas with recent gully formation or bank erosion, subsoil is the dominant source of sediments. Sediments derived from gulley or bank sources have low P content and high P sorption capacities.[12,13] As P release and sorption are largely related to particle size, with coarser-sized particles releasing P more readily than fine particles, which also tend to sorb more P, [14] hydrologic processes controlling sediment particle size distribution have important implications to P fate in river systems.

Abiotic Processes

In fluvial systems with good hydraulic mixing (such as shallow flowing streams), P movement between sediment and water phases is related to the equilibrium P concentration at zero net sorption or desorption (EPC_0); P is released from sediment if the concentration of P in stream flow is less than its EPC_0, while the reverse is also true.[15] Other processes influencing sediment P release include a rise in stream water pH, P from dead phytoplankton, periphyton or macrophytes, the hydrolysis of organic P species, and changes in sediment crystallinity and oxidation/reduction.[10,16] For example, regular wetting and drying cycles in stream sediments or bank material can change Fe-oxide crystallinity making occluding P associated with these materials.[17,18]

Biotic Processes

Uptake of P by aquatic biota can decrease dissolved P in the water column,[19] while bacteria can mediate a sizeable proportion of sedimentary P uptake and release (30%–40%, Khoshmanesh et al.[20] and McDowell and Sharpley[21]). Biologically-controlled P release during the decomposition of organic matter in sediments can be an important source of dissolved P at times of high temperature and low flow in areas with organic-rich sediments, such as streams draining forests.[22] Organic matter in sediments may also increase the blooms of bacteria and algae by preventing chelator limited growth.[23] The relative effect of biotic processes on riverine P transport varies greatly, reflecting seasonal cycles, management of stream-side land, sediment P forms, size of flow event, and streambed geology. However, during elevated flow, when P loads are often high, biotic processes generally are less important to riverine P transport than physical and abiotic processes.

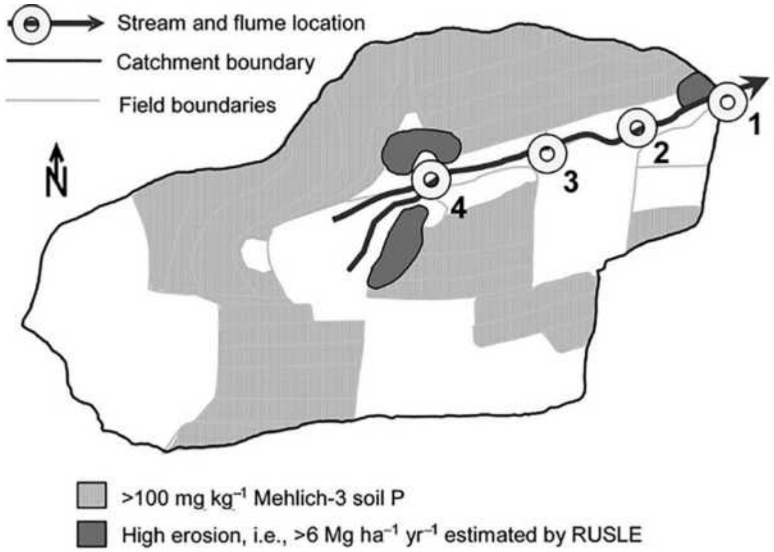
Nutrient Spiralling

The concept of P-spiralling, or the distance travelled downstream by one P molecule as it completes one cycle of uptake, transformation (e.g., from dissolved to organic form) and release back into flow as dissolved P, is useful in understanding fundamental mechanisms of P transport in rivers.[24] Lengths of P-spiralling generally vary from 1 to 1000 m, as a function of flow regime, season, bedrock geology, and sediment characteristics.[25,26] Interaction between ground water and stream flow within the hyporheic zone will change P concentrations, depending upon the relative contribution of stream-bed upwelling or infiltration of P-rich stream flow.

Integrating Riverine Processes and Land Use Impacts on P Transport

The role of riverine processes in watershed scale P export is illustrated by the findings work of McDowell et al.[27] in a 40-ha agricultural watershed in central Pennsylvania (Figure 1). They found dissolved P concentrations in base flow increased from 28 to 42 $\mu g\ L^{-1}$ as one moved from down the stream channel. Base flow P concentrations were controlled by channel sediment P sorption (532 $mg\ kg^{-1}$ at flume 4 and 227 $mg\ kg^{-1}$ at the outlet) and EPC_0 (4 $\mu g\ kg^{-1}$ at flume 4 and 34 $\mu g\ kg^{-1}$ at the outlet). Storm flow trends, however, were the opposite, with P concentrations decreasing downstream (304 μg L^{-1} at flume 4 and 128 μg L^{-1} at flume 1) due to the dilution of P derived from a critical source area: an agricultural field with elevated soil P and high erosion/runoff (Figure 1).

In a much larger watershed, the Winooski River, VT, the largest tributary to Lake Champlain (Figure 2), McDowell et al.[7] evaluated interactions between local sources of P, sediment properties and flow, elucidating their role in riverine P transport. Input and delivery of fine sediment enriched with P was influenced by surrounding land use. Algal-available P of river sediments near agricultural land (3.6 mg kg^{-1}) was greater than that of sediments near forested land (2.4 mg kg^{-1}). Over the short-term, river flow and sediment physical properties were responsible for particulate P loadings from the river to Lake Champlain. However, deposition of sediments downstream, near the outflow into Lake Champlain, resulted in a large pool of stored P within the river system. Over the long-term, this pool is likely to release dissolved P to overlying waters, even as inputs of P from point and nonpoint sources decline due to implementation of remedial strategies and watershed conservation measures.



	Dissolved P		Stream sediment	
Flume	Stormflow	Baseflow	P sorption max	EPC$_0$
	μg L^{-1}		mg kg^{-1}	μg L^{-1}
1	128	42	227	34
2	174	36	295	13
3	202	37	330	4
4	304	28	532	4

FIGURE 1 The distribution of high Mehlich-3 soil P (>100 mg kg^{-1}), erosion (>6 Mg ha^{-1} yr^{-1}) and dissolved P concentration in stream and baseflow (mean of 1997–2000 data) in relation to P sorption properties of channel sediment at four flumes in FD-36.
Source: Adapted from McDowell et al.[27]

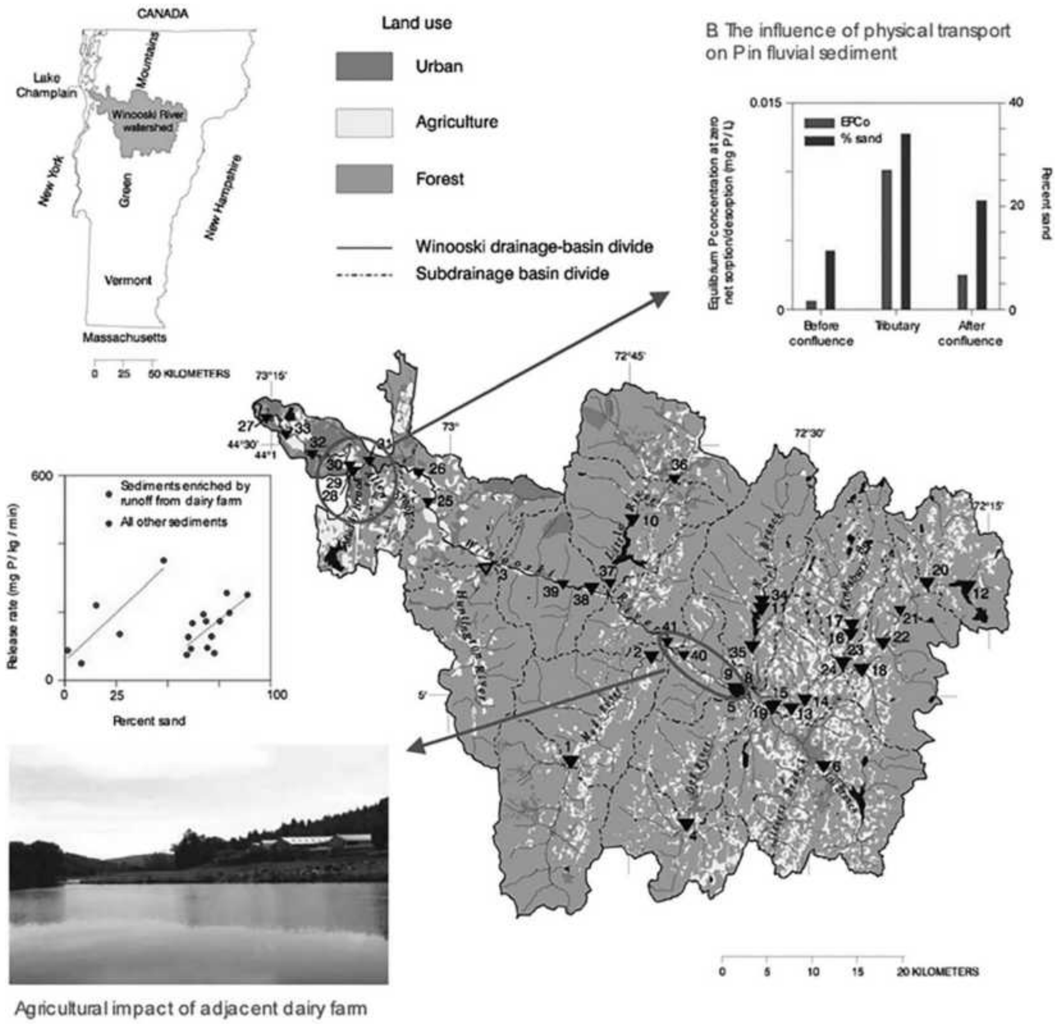


FIGURE 2 The location, distribution and impact of land use and physical transport processes on P in fluvial sediments within the Winooski River watershed, VT.
Source: Adapted from McDowell et al.[7]

Defining P-Related Impairment in Flowing and Lake Waters for Targeted Remediation

In order to prioritize and target watershed remediation to minimize P losses, water impairment must be quantified.[28] Background levels (i.e., regional nutrient criteria) of total P, total N, chlorophyll-*a*, sediment, and clarity in pristine surface waters are used as benchmarks for a given geographical area (Figure 3; Gibson et al.[29] and Omernik[30]). While these criteria have regulatory application, such as under the U.S.A.'s Clean Water Act, they can also be used to guide voluntary efforts in watershed planning.[31] These criteria are available for freshwater systems in the continental U.S. (Table 1). Similar approaches have been taken in Australasia and Europe.[4,32] In the European Union's Water Framework Directive, biological parameters are, however, the basis for measuring ecological status for the water with chemical parameters used only as support parameters. The E.U. classification system emphasizes

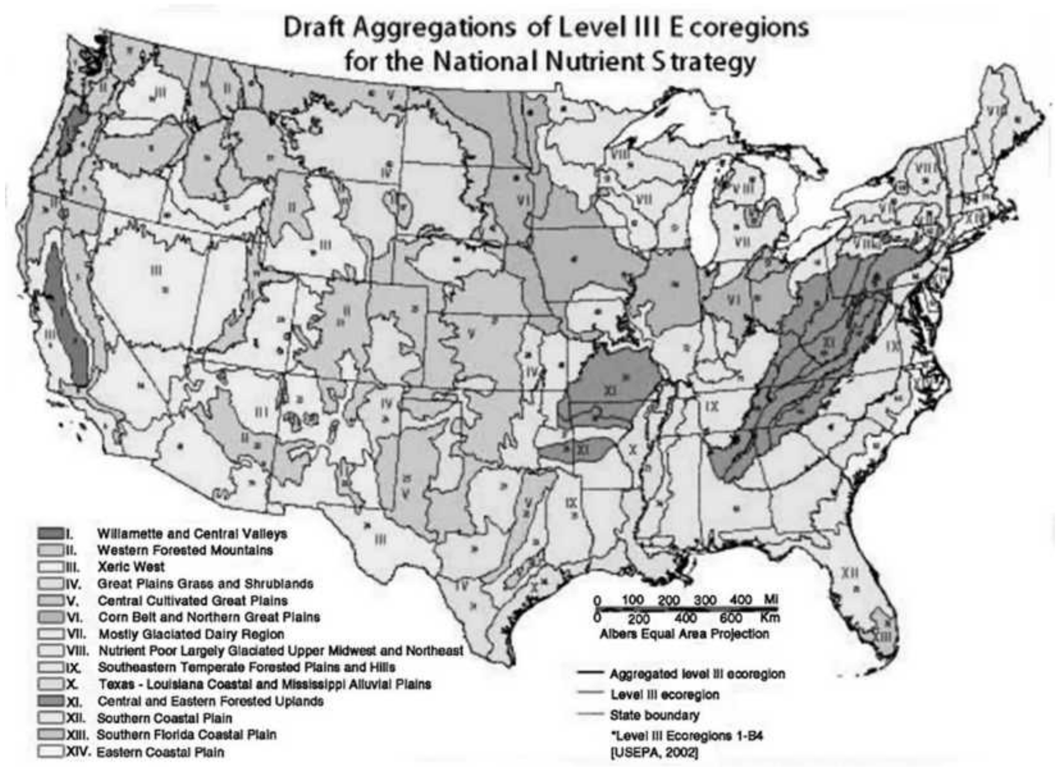


FIGURE 3 Draft aggregations of level III ecoregions for the National Nutrient Strategy.
Source: Adapted from Gibson et al.[29] and Omernik.[30]

TABLE 1 Background Total P concentrations for Each of the Aggregated Nutrient Ecoregions in the U.S. for Freshwater Systems[28]

Aggregated Ecoregion		Total P (μg L^{-1})	
Number	Description	Rivers and Streams	Lakes and Reservoirs
I	Willamette and Central Valleys	47	—
II	Western Forested Mountains	10	9
III	Xeric West	22	17
IV	Great Plain Grass and Shrub Lands	23	20
V	South Central Cultivated Great Plains	67	33
VI	Corn Belt and Northern Great Plains	76	38
VII	Mostly Glaciated Dairy Region	33	15
VIII	Nutrient Poor Largely Glaciated Upper Midwest and Northeast	10	8
IX	Southeastern Temperate Forested Plains and Hills	37	20
X	Texas-Louisiana Coastal and Mississippi Alluvial Plain	128a	—
XI	Central and Eastern Forested Uplands	10	8
XII	Southern Coastal Plains	40	10
XIII	Southern Florida Coastal Plains	—	18
XIV	Eastern Coastal Plains	31	8

whether the ecosystem is in ecological balance and points out the effect of the pollution rather than providing a classification or ranking according to pollutant concentration, which has been the basis for most previous classifications systems.

Implications to Watershed Management

Aquatic ecosystems respond to P inputs on the basis of factors related to their physiography and flushing rates. Individual systems respond to discrete and sustained P inputs differently, and indeed it may not be possible to attain P loadings low enough to prevent periphyton blooms because of, for example, natural enrichment from P-rich rocks.

A certain degree of eutrophication can be beneficial. For example, fishery management often requires a higher productivity to maintain an adequate phytoplankton-zooplankton-fish food chain for optimum commercial or sport fish production. This food chain may be manipulated by stocking of water with certain fish species in addition to P load reductions, in efforts to reduce the incidence of algal blooms and improve overall water quality.[33]

In most cases, however, eutrophication restricts water use for fisheries, recreation, and industry because of the increased growth of undesirable algae and aquatic weeds and oxygen shortages caused by their death and decomposition.[1] An increasing number of surface waters also experience periodic and massive harmful algal blooms (e.g., *cyanobacteria* and *Pfiesteria*) that contribute to summer fish kills, unpotable drinking water, formation of carcinogens during water chlorination, and links to neurological impairment in humans.[34,35]

Implications for Water Quality Response

The response of riverine systems to upstream changes in watershed management can vary from seasons to decades, and generally increases as watershed scale increases.[36,37] In small watersheds (i.e., <100 ha), research has demonstrated reduction in nutrient and sediment loss in runoff can occur within months of implementing remedial management measures (e.g., “conservation practices “or “best management practices”). However, the spatial complexity of watershed systems and nature of P sources (e.g., acute sources such as fertilizers vs. legacy P sources such as soils and sediments) affects this response time. Indeed, the slow release of legacy P stored in soils and sediments to more rapid surface flow pathways may continue for decades even after practices have been installed to curb further additions. The release of legacy P from riverine sediments is influenced by the oxygen status of overlying waters, where reducing conditions favor the dissolution of iron-bound P. As remedial efforts decrease the concentrations of P in riverine systems, the change in gradient between sediment and water column can trigger the desorption of dissolved P from the sediment. Finally, P enriched sediments are resuspended during high flows. This effect will likely increase with climate change, which could result in more precipitation as rain in winter and an increase in occurrence and severity of rainfall events.

Because of the lag time between watershed management practice implementation and water quality improvements, remedial strategies should consider the re-equilibration of watershed and water-body behavior, where nutrient sinks may become sources of P with only slight changes in watershed management and hydrologic response. A better understanding of the spatial and temporal aspects of watershed response to nutrient load reductions in both flowing and standing water bodies is needed, as well as the scale at which responses may occur in a more timely fashion. This would likely be at a smaller sub-watershed scale, where local water quality and quantity benefits may become evident more quickly. It is also important to accept in any water-shed-P loss reduction strategy, that it is essential to address the overall physical and social complexity of legacy P sources, when, where and to what extent they occur.

Conclusions

Clearly, several interdependent riverine processes influence the amounts and forms of P transported from edge-of-field agricultural sources to the point of impact (i.e., river, lake, reservoir, and estuary). These processes will, thus, be critical in defining agricultural source management and in determining eutrophic response. Without information on the direction and magnitude of change in P transport in river systems, conservation practices will not efficiently remediate against impairment of receiving waters.

The accumulation of P in aquatic environments is such that even if P were no longer added to agricultural systems, there would be a considerable time-lag (years or decades) before improvements in water quality, or regeneration of diverse habitats, might become apparent. Thus, the emphasis of watershed management should be on preventing further deterioration and taking strategic and sustainable actions sooner rather than later, otherwise we are simply and literally storing up more severe problems for future generations to confront. Despite our knowledge of controlling processes, it is difficult for the public to understand or accept this lack of response. When public funds are invested in remedial watershed programs, rapid improvements in water quality are usually expected and often required. Thus, implementation of effective conservation measures must consider fluvial system response behavior, where sinks may become sources of P with only slight changes in watershed management and hydrologic response.

References

1. Carpenter, S.R.; Caraco, N.F.; Correll, D.L.; Howarth, R.W.; Sharpley, A.N.; Smith, V.H. Nonpoint pollution of surface waters with phosphorus and nitrogen. Ecol. Applic. **1998**, *8*, 559–568.
2. Schindler, D.W.; Hecky, R.E.; Findlay, D.L.; Stainton, M.P.; Parker, B.R.; Paterson, M.J.; Beaty, K.G.; Lyng, M.; Kasian, S.E. Eutrophication of lakes cannot be controlled by reducing nitrogen input: Results of a 37-year whole-ecosystem experiment. Proc. National Acad. Sci. **2008**, *105*(32), 11254–11258.
3. U.S. Geological Survey. *The quality of our nations waters: Nutrients and pesticides.* U.S. Geological Survey Circular 1225; USGS Information Services: Denver, CO, 1999; 82 p. Available at http://www.usgs.gov.
4. Withers, P.J.A.; Lord, E.L. Agricultural nutrient inputs to rivers and ground waters in the U.K.: Policy environmental management and research needs. Soil Use and Managt. **2002**, *14*, 186–192.
5. National Research Council. *Clean coastal waters: Understanding and reducing the effects of nutrient pollution*; National Academy Press: Washington, D.C., 2000.
6. Mainstone, C.P.; Parr, W. Phosphorus in rivers – ecology and management. Sci. Total Environ. **2002**, *282–283*, 25–47.
7. McDowell, R.W.; Sharpley, A.N.; Chalmers, A.T. Land use and flow regime effects on phosphorus chemical dynamics in the fluvial sediment of the Winooski River, Vermont. Ecolog. Eng. **2002**, *18*, 477–487.
8. House, W.A. Geochemical cycling of phosphorus in rivers. App. Geochem. **2003**, *18*, 739–748.
9. Jarvie, H.P.; Whitton, B.A.; Neal, C. Nitrogen and phosphorus in east coast British rivers: speciation sources and biological significance. Sci. Total Environ. **1998**, *210–211*, 79–110.
10. Baldwin, D.S.; Mitchell, A.M.; Olley, J.M. Pollutantsediment interactions: Sorption, reactivity and transport of phosphorus. In *Agriculture, Hydrology and Water Quality;* Haygarth, P.M., and Jarvis, S.C. (eds.); CABI International: Oxford, England, 2002; 265–280.
11. Krogstad, T.; Løvstad, Ø. Erosion, phosphorus and phytoplankton response in rivers of southeastern Norway. Hydrobiologia **1989**, *183*, 33–41.
12. Olley, J.M.; Murray, A.S.; Mackenzie, D.H.; Edwards, K. Identifying sediment sources in a gullied catchment using natural and anthropogenic radioactivity. Water Resour. Res. **1993**, *29*, 1037–1043.

13. Sharpley, A.N.; Smith, S.J.; Zollweg, J.A.; Coleman, G.A. Gully treatment and water quality in the Southern Plains. J. Soil Water Conserv. **1996**, *51,* 512–517.
14. Stone, M.; Murdoch, A. The effect of particle size, chemistry and mineralogy of river sediments on phosphate adsorption. Environ. Technol. Letters **1989**, *10,* 501–510.
15. Kunishi, H.M.; Taylor, A.W.; Heald, W.R.; Gburek, W.J.; Weaver, R.N. Phosphate movement from an agricultural watershed during two rainfall periods. J. Agric. Food Chem. **1972**, *20,* 900–905.
16. Fox, L.E. The chemistry of aquatic phosphate: inorganic processes in rivers. Hydrobiologia **1993**, *253,* 1–16.
17. Qiu, S.; McComb, A.J. Planktonic and microbial contributions to phosphorus release from fresh and air-dried sediments. Freshwater Res. **1995**, *46,* 1039–1045.
18. Baldwin, D.S. Effects of exposure to air and subsequent drying on the phosphate sorption characteristics of sediments from a eutrophic reservoir. Limnol. Oceanog. **1996**, *41,* 1725–1732.
19. Horner, R.R.; Welch, E.B.; Seeley, M.R.; Jacoby, J.M. Responses periphyton to changes in current velocity, suspended sediment and phosphor concentration. Freshwater Biol. **1990**, *24,* 215–232.
20. Khoshmanesh, A.; Hart, B.T.; Duncan, A.; Beckett, R. Biotic uptake and release of phosphorus by a wetland sediment. Environ. Technol. **1999**, *29,* 85–91.
21. McDowell, R.W.; Sharpley, A.N. Phosphorus uptake and release from stream sediments. J. Environ. Qual. **2003**, *32*, 937–948.
22. Klotz, R.L. Temporal relation between soluble reactive phosphorus and factors in stream water and sediments in Hoxie Gorge Creek, New York. Can. J. Aquatic Sci. **1991**, *48,* 84–90.
23. Løvstad, Ø.; Krogstad, T. Effect of EDTA, FeEDTA and soils on the phosphorus bioavailability for diatom and blue-green algal growth in oligotrophic waters studied by transplant biotests. Hydrobiologia **2001**, *450,* 71–81.
24. Elwood, J.W.; Newbold, J.D.; O'Neil, R.V.; Van Winkle, W. Resource spiraling: An operational paradigm for analyzing lotic ecosystems. In *Dynamics of lotic ecosystems;* Fontaine, III, T.D.; Bartell, S.M. (eds.); Ann Arbor Science: Ann Arbor, MI, 1983; 3–27.
25. Melack, J.M. Transport and transformations of P, fluvial and lacustrine ecosystems. In *Phosphorus in the Global Environment: Transfers, Cycles and Management;* Tiessen, H. (Ed.); SCOPE 54 and John Wiley and Sons: Chichester, U.K., 1995; 245–254.
26. Munn, N.L.; Meyer, J.L. Habitat-specific solute retention in two small streams, An intersite comparison. Ecology **1990**, *71*, 2069–2082.
27. McDowell, R.W.; Sharpley, A.N.; Folmar, G. Phosphorus export from an agricultural watershed: linking source and transport mechanisms. J. Environ. Qual. **2001**, *30,* 1587–1595.
28. U.S. Environmental Protection Agency. *Ecoregional Nutrient Criteria*. EPA-822-F-01–010. USEPA, Office of Water (4304), U.S. Govt. Printing Office: Washington, DC, 2001. Available at http://www.epa.gov/waterscience/criteria/nutrient/9docfs.pdf.
29. Gibson, G.R.; Carlson, R.; Simpson, J.; Smeltzer, E.; Gerritson, J.; Chapra, S.; Heiskary, S.; Jones, J.; Kennedy, R. *Nutrient criteria technical guidance manual: lakes and reservoirs* (EPA-822-B00–001). U.S. Environmental Protection Agency: Washington, D.C.; U.S. Govt. Printing Office: Washington, DC, 2000.
30. Omernik, J.M. Ecoregions of the Conterminous United States. Annals Assoc. Am. Geographers **1987**, *77,* 118–125.
31. U.S. Environmental Protection Agency. *National Strategy for the Development of Regional Nutrient Criteria*. EPA-822-F-98-002. USEPA, Office of Water (4304), U.S. Govt. Printing Office: Washington, DC, 1998. Available at http://www.epa.gov/waterscience/standards/nutsi.html.
32. Sparrow, L.A.; Sharpley, A.N.; Reuter, D.J. Safeguarding soil and water quality. In Opportunities for the 21st Century: Expanding the horizons for soil, plant, and water analysis. Commun. Soil Sci. Plant Anal. **2000**, *31*(11–14), 1717–1742.

33. Horppila, J.; Kairesalo, T. A fading recovery: the role of roach (Rutilus rutilus L.) in maintaining high algal productivity and biomass in Lake Vesijarvi, southern Finland. Hydrobiologia **1990**, *200–201,* 153–165.
34. Burkholder, J.A.; Glasgow, Jr., H.B. *Pfiesteria piscicidia* and other Pfiesteria-dinoflagellates behaviors, impacts, and environmental controls. Limnol. Oceanog. **1997**, *42,* 1052–1075.
35. Kotak, B.G.; Kenefick, S.L.; Fritz, D.L.; Rousseaux, C.G.; Prepas, E.E.; Hrudey, S.E. Occurrence and toxicological evaluation of cyanobacterial toxins in Alberta lakes and farm dugouts. Water Res. **1993**, *27,* 495–506.
36. Cassell, E.A.; Clausen, J.C. Dynamic simulation modeling for evaluating water quality response to agricultural BMP implementation. Water Sci. Tech. **1993**, *28,* 635–648.
37. Sharpley, A.N. (ed.). *Agriculture and Phosphorus Management: The Chesapeake Bay.* CRC Press: Boca Raton, FL, 2000. 229 p.

48

Nitrogen: Biological Fixation

Mark B. Peoples

Introduction

Dinitrogen (N_2) gas represents close to 80% of the Earth's atmosphere, and a number of procaryotic microorganisms in soils and flooded conditions have evolved that utilize the enzyme nitrogenase to reduce atmospheric N_2 to ammonia. This ammonia provides a renewable source of nitrogen (N) that is subsequently used to support their growth. Biological N_2 fixation (BNF) by some diazotrophs can occur in a free-living state. This includes anaerobic organisms such as *Clostridium* and aerobic bacterium such as *Azotobacter* along with a host of other species that may rely on decomposition of plant residues as an energy source, while phototrophic bacteria such as cyanobacteria (also known as blue-green algae) derive energy directly from photosynthesis to sustain BNF.[1,2] Other N_2-fixing microbes have coevolved with plants and undertake BNF in multiple ways. This ranges from rather loose associations of free-living heterotrophic bacteria with plant roots or endophytic N_2-fixing bacteria present in the vascular tissue of tropical C_4-graminaceous species such as sugarcane (*Saccharum officinarum*), to highly complex symbioses where forms of fixed N generated by the bacterium are exchanged with the plant for carbon-derived plant photosynthesis.[2–4] Such symbioses include (1) cyanobacteria with the aquatic ferns (*Azolla*), fungi (lichen), or cycads; (2) actinomycetes (generally placed in the genus *Frankia*) with around 200 species of angiosperms including Alder (*Alnus glutinosa*) and *Casuarina* spp.; and (3) the soil bacteria rhizobia (*Rhizobium*, *Bradyrhizobium*, *Allorhizobium*, *Azorhizobium*, *Mesorhizobium*, or *Sinorhizobium* spp.) with >10,000 legume species. With the exception of *Azolla* and lichen, these symbiotic partnerships generally occur in specialized plant structures known as nodules located on plant roots and, in some legume species, also the stem.[2,5]

Although calculations of global contributions of BNF are subject to enormous approximations, annual inputs of fixed N into agroecosystems by these diverse range of organisms and associations were conservatively estimated to be 55–70 million metric tonnes of N based on Food and Agriculture Organization (FAO) statistical data for agricultural production in 2005.[4] Unfortunately, uncertainties about the distribution and frequency of N_2-fixing organisms, or the rates of BNF occurring, in natural ecosystems makes it almost impossible to provide similar estimates of global inputs of fixed N for the many different terrestrial and aquatic environments not under agricultural production.

Sources of BNF in Agricultural Systems

The N_2 fixation process can directly contribute to agricultural production where the fixed N is harvested in legume grain or vegetative parts for human or animal consumption. However, BNF can also represent an important renewable source of N that can help maintain or enhance the N fertility of many agricultural soils for the benefit of subsequent crops.[5–7] Examples of experimental estimates of amounts of BNF achieved by various N_2-fixing organisms and estimates of their global inputs of fixed N are presented in Table 1.

Free-living N_2 fixers tend to have low rates of BNF (Table 1). While significant inputs of fixed N by diazotrophs have been demonstrated in some tropical grass and cropping systems,[3] the role of BNF associated with nonlegumes tend to be less conclusive in temperate agriculture,[8,9] and these observations are reflected in the estimates of global inputs of fixed N from endophytic, associative, and free-living sources provided in Table 1. *Azolla* may have higher rates of BNF than non-symbiotic systems,[2,5] but it is symbiotic relationships between rhizobia and legumes that are responsible for the largest amounts of fixed N in agriculture both on a per unit basis and in absolute terms (Table 1).

Inputs of Fixed Nitrogen by Legumes

That BNF by legume systems plays a key role in world crop production is irrefutable. The ability of legumes to improve the N status of soils has been utilized for thousands of years in crop rotations and traditional farming systems.[2,5,7] The 229 million hectares of legume oilseeds (soybean—*Glycine max*; and groundnut/peanut—*Arachis hypogea*) and pulses sown globally each year, the legume components of the 100–200 million hectares under temporary pastures or fodder crops, and the perennial legume cover crops under much of the 33 million hectares of rubber (*Hevea brasiliensis*) and oil-palm (*Elaeis guineensis*) plantations all contribute fixed N to farming systems.[5,6,10]

Most modern methods used to quantify inputs of fixed N by legumes separate the plant N into fractions originating from soil N or N_2 fixation.[11,12] Once the legume N can be partitioned into that proportion derived from atmospheric N_2 (%Ndfa, sometimes also described as %Pfix) and that coming from the soil, the amounts of N_2 fixed can be calculated from measures of shoot dry matter (DM) and N content. The formation of the symbiosis between legume and rhizobia is dependent upon many factors and cannot be assumed to occur as a matter of course. This is reflected in the range of values presented

TABLE 1 Experimental Measures of BNF by Different N_2-Fixing Organisms and Estimates of Annual Global Inputs of Fixed N in Agricultural Systems[4]

N2-Fixing Organism	Range of Estimates of BNF Measured (kg N/ha)	Typical Average Rates of BNF (kg N/ha)	Global Inputs of BNF (million t N/year)
Free-living and associative			
Temperate crop land	0–80	<5	<4
Tropical savannas	5–45	10	<14
Tropical crops	0–240	25	0.5
Symbiotic			
Azolla	10–150	33	5
Crop legumes	0–450	115a	21[a]
Pasture/forage legumes	1–680	110–227ab	12–25ab
Legumes in agroforestry	5–470	200a	ndc
Green manure legumes	5–325	100a	ndc

[a] Estimates of BNF have been adjusted to include a belowground contribution of N associated with the nodulated roots.[6,10,12]

[b] The range reflects that legumes can be grown as a pure sward or in a mixture with other species.

[c] nd indicates "not determined" since data relating to the total areas of legume grown in agroforestry systems and as green manure were not available.

in Table 1. Such large variations in reported estimates of N_2 fixation make it difficult to generalize about how much N may be fixed by different legume species. Collectively, the international literature suggests maximum rates of N_2 fixation of 3–4 kg shoot N/ha/day[12] and potential inputs of fixed N by many legumes of several hundred kg of shoot N/ha each year.[5,6,10] However, much of this information has been derived from research trials in which specific treatments were imposed to generate differences in %Ndfa values and legume growth as an experimental means of studying factors which regulate BNF. Therefore, these data may be of little relevance to what might actually be occurring in farmers' crops and pastures. Fortunately, measurement procedures are available which allow on-farm measures of legume N_2 fixation to be conducted with some degree of confidence.[11]

Levels of Nitrogen Fixation Achieved in Farmers' Fields

Examples of the types of information which can be generated about BNF in farmers' fields are presented in Table 2 for different regions of the world. These on-farm data and observations can be used to develop a picture of N_2 fixation within an individual country and provide insights into contributions of BNF to agriculture on a global scale. Collectively, the results in Table 2 indicate that the potential for BNF inputs can differ between legumes and countries, but they also suggest many commonalities. Although wide ranges in %Ndfa values have been observed, it seems that, on average, most winter pulses (e.g., chickpea—*Cicer arietinum*; lentil—*Lens culinaris*; field pea—*Pisum sativum*; fababean—*Vicia faba*; lupin—*Lupinus albus*) satisfy relatively higher proportions of their growth requirements from N_2 fixation (>65%) than do the summer legumes (e.g., mungbean—*Vigna radiata*; mashbean—*Vigna mungo*; soybean; groundnut) where %Ndfa values were commonly less than 60% (Table 2). Poor or variable nodulation observed in some summer legumes and the resulting increased reliance upon soil N may reflect greater N mineralization during summer, water stress, and/or low vegetative biomass accumulated by short duration legume crops.[13–15]

TABLE 2 Summary of the Proportion of Plant N Derived from N_2 Fixation (%Ndfa) and the Amounts of N_2 Fixed by Farmers' Legume Crops and Pastures in Different Geographical Regions

Country and Legume	Number of Fields	Mean Ndfa (%)	Total N fixed (kg N/ha)[a]
Winter pulse crops			
Pakistan	126	78	79
Nepal	27	79	78
Syria	46	67	na[b]
Australia	148	65	152
Summer legume crops			
Pakistan	63	47	42
Nepal	50	55	77
Thailand	13	75	78
Vietnam	45	48	125
South Africa	14	58	na[b]
Australia	33	53	267
Annual pastures			
Australia	303	75	na[b]
Perennial pastures			
Australia	110	64	na[b]

[a] Includes a combined estimate of fixed N from both the shoots and the nodulated roots.[6,10,12]

[b] Data not available for all fields.

Note: Ndfa = Nitrogen derived from atmospheric N2.

TABLE 3 Key Factors Influencing Inputs of Fixed N by Legumes in Farmers' Fields

		BNF Regulated By		
Country	System	DM	Soil Nitrate	Primary Factors
Pakistan	Winter crop	+++		Rainfall nutrition, weed control
	Summer crop	+++	+++	Fertilizer N, no inoculation, insects, disease
Nepal	Winter crop	+++		Rainfall, nutrition
	Summer crop	+++	+	Total soil N, mineralized N, available P, legume species
Syria	Winter crop	+++	++	Soil nutrients, insects, disease
Thailand	Summer crop	+++		Available P
Vietnam	Summer crop	+++		Plant density, soil pH, available P, legume species
South Africa	Summer crop	++	++	Effective inoculation, nutrition, rotation, water availability
Australia	Winter crop	+++	+++	Rainfall, fallowing, inoculation, legume species
	Summer crop	++	+++	Crop rotation, tillage, water availability
	Pasture	+++	+	Soil pH, available P, legume density, grazing management

Note: BNF = biological N_2 fixation; DM = dry matter; P = phosphorus.

In grazed pastures and intensive forage systems, the competition for mineral N between legumes and any companion grasses or vigorous broad-leaf species growing within the sward results in low levels of plant-available soil N throughout the growing season.[10,12] As a consequence, %Ndfa by the legume components of pastures tend to be high (Table 2). The lower %Ndfa values detected in perennial legume species, such as alfalfa (lucerne; *Medicago sativa*) and white clover (*Trifolium repens*), presumably result from a greater ability to scavenge soil mineral N from a larger rooting zone over a longer growing season compared with annual pasture species.[10,15]

Although the levels of %Ndfa are important, the amounts of N_2 fixed are usually regulated by legume growth rather than %Ndfa in most farming systems, and many legumes appear to fix around 15–20 kg of shoot N for every metric tonnes of shoot DM accumulated.[6,10,14,15] Depending upon the species, this would translate to 23–40 kg of total N fixed for every tonne of shoot DM if belowground contributions of fixed N-associated nodulated roots are included.[6,10,12]

Impact of Management

Factors that either enhance or depress N_2 fixation (Table 3) can generally be summarized in terms of environmental or management constraints to crop growth (e.g., basic agronomy, nutrition, water supply, diseases, and pests).[6,10,16] A number of strategies can be employed that specifically enhance BNF through increased legume biomass. These include the use of legume genotypes adapted to the prevailing edaphic and environmental conditions, procedures to improve legume plant density, irrigation (if available), the amelioration of soil nutrient toxicities or deficiencies, and the control of weeds and pests.[10,16] However, as the formation of an active symbiosis is dependent upon the compatibility of both the diazotrophic microorganism and the legume host, local practices that limit the presence of effective rhizobia (no inoculation, poor inoculant quality) will also be crucial in determining the legume's capacity to fix N (Table 3), as will any management decisions that directly affect soil N fertility (excessive tillage, extended fallows, fertilizer N, and rotations), since mineral N is a potent inhibitor of the N_2 fixation process.[6,10,15,16]

Conclusions

Symbiotic associations between legumes and rhizobia are responsible for the greatest contributions of BNF in agricultural systems. Research trials suggest potential annual inputs of fixed N by most legumes equivalent to several hundreds of kg N/ha. However, data collected from pulses, legume oilseeds, and pastures growing in farmers' fields suggest that while legumes should routinely be fixing >100 kg/ha each year, in reality, they usually do not. Strategies are available to improve BNF beyond what is currently being achieved. For example, provided that a legume crop is abundantly nodulated and effectively fixing N_2, enormous benefits in terms of crop production and N_2 fixed can be derived from the application of good agronomic principles. But the ability to overcome constraints at the farm level may be limited because the relevant technologies are either not in the hands of the farmers or they cannot readily adopt them because of lack of knowledge and information, economic constraints, or operational imperatives. While the global inputs of fixed N by legumes may be less than their genetic potential and is lower than the 100–110 million tonnes of N applied as fertilizer each year, many million tonnes of fixed N are harvested annually in grain from legume crops and many million tonnes more will be consumed by animals in legume-based forage, and there are many environmental advantages in relying upon BNF over fertilizer N to produce such large quantities of high-quality protein.[17,18]

References

1. Roper, M.M.; Ladha, J.K. Biological N_2 fixation by heterotrophic and phototrophic bacteria in association with straw. *Plant Soil* **1995**, *174*, 211–224.
2. Giller, K.E. *Nitrogen Fixation in Tropical Cropping Systems*, 2nd ed.; CABI Publishing: Wallingford, England, 2001.
3. Boddey, R.M.; de Oliveira, O.C.; Urquiaga, S.; Reis, V.M.; de Olivares, F.L.; Baldani, V.L.D.; Döbereiner, J. Biological nitrogen fixation associated with sugar cane and rice: contributions and prospects for improvement. *Plant Soil* **1995**, *174*, 195–209.
4. Herridge, D.F.; Peoples, M.B.; Boddey, R.M. Global inputs of biological nitrogen fixation in agricultural systems. *Plant Soil* **2008**, *311*, 1–18.
5. Peoples, M.B.; Herridge, D.F.; Ladha, J.K. Biological nitrogen fixation: an efficient source of nitrogen for sustainable agricultural production? *Plant Soil* **1995**, *174*, 3–28.
6. Peoples, M.B.; Brockwell, J.; Herridge, D.F.; Rochester, I.J.; Alves, B.J.R.; Urquiaga, S.; Boddey, R.M.; Dakora, F.D.; Bhattarai, S.; Maskey, S.L.; Sampet, C.; Rekasem, B.; Khan, D.F.; Hauggaard-Nielsen, H.; Jensen, E.S. The contributions of nitrogen-fixing crop legumes to the productivity of agricultural systems. *Symbiosis* **2009**, *48*, 1–17.
7. Angus, J.F.; Kirkegaard, J.A.; Hunt, J.R.; Ryan, M.H.; Ohlander, L.; Peoples, M.B. Break crops and rotations for wheat. *Crop Pasture Sci.* **2015**, *66*, 523–552.
8. Giller, K.; Merckx, R. Exploring the boundaries of N_2 fixation in cereals and grasses: an hypothetical framework. *Symbiosis* **2003**, *35*, 3–17.
9. Unkovich, M.J.; Baldock, J.A. Measurement of asymbiotic N_2 fixation in Australian agriculture. *Soil Biol. Biochem.* **2008**, *40*, 2915–2921.
10. Peoples, M.B.; Brockwell, J.; Hunt, J.R.; Swan, A.D.; Watson, L.; Hayes, R.C.; Li, G.D.; Hackney, B.; Nuttall, J.G.; Davies, S.L.; Fillery, I.R. Factors affecting the potential contributions of N_2 fixation by legumes in Australian pasture systems. *Crop Pasture Sci.* **2012**, *63*, 759–786.
11. Unkovich, M.; Herridge, D.; Peoples, M.; Cadisch, G.; Boddey, R.; Giller, K.; Alves, B.; Chalk, P. *Measuring Plant-Associated Nitrogen Fixation in Agricultural Systems*, ACIAR Monograph No. 136: Caberra, Australia, 2008.
12. Unkovich, M.J.; Pate, J.S. An appraisal of recent field measurements of symbiotic N_2 fixation by annual legumes. *Field Crops Res.* **2001**, *65*, 211–228.
13. Herridge, D.F.; Robertson, M.J.; Cocks, B.; Peoples, M.B.; Holland, J.F.; Heuke, L. Low nodulation and nitrogen fixation of mungbean reduce biomass and grain yields. *Aust. J. Expl. Agric.* **2005**, *45*, 269–277.
14. Maskey, S.L.; Bhattarai, S.; Peoples, M.B.; Herridge, D.F. *Field Crops Res.* **2001**, *70*, 209–221.

15. Peoples, M.B.; Bowman, A.M.; Gault, R.R.; Herridge, D.F.; McCallum, M.H.; McCormick, K.M.; Norton, R.M.; Rochester, I.J.; Scammell, G.J.; Schwenke, G.D. Factors regulating the contributions of fixed nitrogen by pasture and crop legumes to different farming systems of eastern Australia. *Plant Soil* **2001**, *228*, 29–41.
16. Peoples, M.B.; Ladha, J.K.; Herridge, D.F. Enhancing legume N_2 fixation through plant and soil management. *Plant Soil* **1995**, *174*, 83–101.
17. Jensen, E.S.; Hauggaard-Nielsen, H. How can increased use of biological N_2 fixation in agriculture benefit the environment? *Plant Soil* **2003**, *252*, 177–186.
18. Jensen, E.S.; Peoples, M.B.; Boddey, R.M.; Gresshoff, P.M.; Hauggaard-Nielsen, H.; Alves, B.J.R.; Morrison, M.J. Legumes for mitigation of climate change and the provision of feedstock for biofules and biorefineries. *Agron. Sustain. Dev.* **2012**, *32*, 329–364.

49

Nutrients: Best Management Practices

Scott J. Sturgul and
Keith A. Kelling

Introduction

Soil nutrients, like all agricultural inputs, need to be managed properly to meet the fertility requirements of crops without adversely affecting the quality of water resources. The nutrients of greatest concern relative to water quality are nitrogen (N) and phosphorus (P).[1] Nitrogen not recovered by crops can add nitrate-N to groundwater through leaching. Nitrate is the most common groundwater contaminant found in the United States.[2–4] Nitrate levels that exceed the established U.S. drinking water standard of 10 ppm nitrate–N have the potential to adversely affect the health of infants and young livestock.[5] Furthermore, the movement of excessive nitrate to coastal estuaries has been linked to the development of hypoxic conditions affecting fisheries in the Gulf of Mexico.[6,7] Surface water quality is the primary environmental concern with P, as runoff and erosion from fertile cropland add nutrients to water bodies that stimulate the excessive growth of aquatic weeds and algae.[8] Of all crop nutrients, it is critical to prevent P from reaching lakes and streams since the biological productivity of aquatic plants and algae in fresh water environments is usually limited by this nutrient.[9] Consequences of increased aquatic plant and algae growth include reduced aesthetic and recreational value of lakes and streams as well as the seasonal depletion of water dissolved oxygen content, which may result in fish kills as well as other ecosystem disruptions.

Best management practices for agricultural nutrients can vary widely from one region to another due to differences in cropping, topographic, environmental, and economic conditions. With the variety of factors to consider, no single set of nutrient management practices can be recommended for all farms. Nutrient management practices for optimizing crop production while protecting water quality must be tailored to the unique conditions of individual farms. The following practices should be considered for any nutrient management strategy for optimizing both agricultural production and environmental protection.

Nutrient Application Rates

The most important management practice for environment-tally—and economically—sound nutrient management is the application rate. Optimum nutrient application rates are identified through fertilizer response and calibration research for specific soils and crops. Economically optimum nutrient application rates provide maximum financial return, but as application rates near the economic optimum, the efficiency of nutrient use by the crop decreases, and the potential for loss to the environment increases.[10] Figure 1 illustrates this concept for N applications to corn. Nutrient application above economically optimum rates (in this case, 160 lb N/acre) reduces profits and increases the likelihood of detrimental impact to the environment.

As shown by many researchers, applications of N in excess of crop need can result in significantly higher N leaching losses.[11–15] A Wisconsin study[16] of the effects of various N fertilizer rates on nitrate-N leaching from several crop rotations found a direct relationship between nitrate-N loss by leaching and the amount of N applied in excess of crop needs (Figure 2). Soil water nitrate-N concentration increased steadily as the amount of excess N increased, which strongly indicates a direct link between excessive N applications and the potential for nitrate-N loss to groundwater.

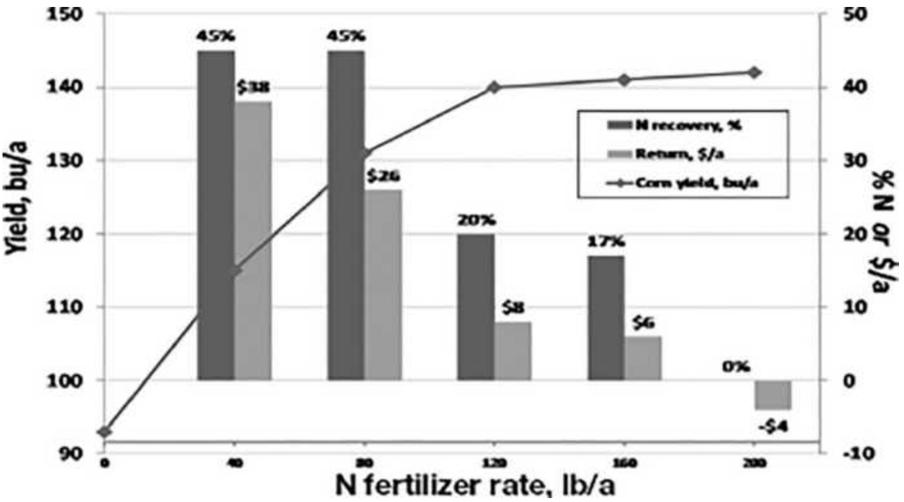


FIGURE 1 Relationships between corn grain yield, economic return, and recovery of applied N.
Source: Bundy et al.[10]

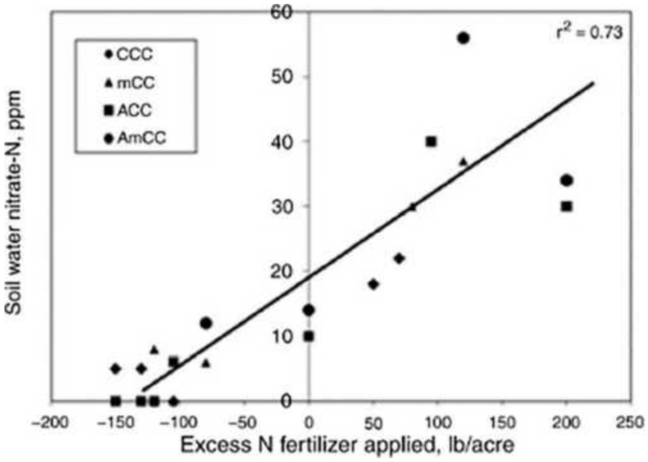


FIGURE 2 Relationship between amount of excess N applied and soil water nitrate content for several cropping rotations (C = corn, A = alfalfa, m = manure).
Source: Andraski et al.[16]

Because of the overall importance of nutrient application rates, accurate assessments of crop nutrient needs are essential for minimizing threats to water quality while maintaining economically sound production. Soil testing is the most widely used method to accurately estimate the existing fertility of soil as well as to determine the need for supplemental nutrients to meet the needs of crops. It has the further advantage of being performed prior to raising the crop, whereas other methods such as plant tissue analysis or optical scanners must have the target crop in place.

Optimum Nitrogen Applications

Most non-legume crops need supplemental N to improve crop yield and quality and to optimize economic return. It is imperative that N application rate recommendations accurately predict the amount of N needed to obtain profitable crop yields and minimize N losses to the environment. Application rate guidelines for N vary according to the crop to be grown, soil characteristics (including soil yield potential, organic matter, texture, etc.), and local climatic conditions.

Expected yield or yield goal estimates have been a primary input for determining grain N fertilizer recommendations for most of the United States since the 1970s.[17] However, research from several Midwest states over the past four decades shows that economic optimum N rates for corn are not best predicted using a yield parameter.[18–21]

More accurate N recommendations for corn and some other grain crops have been developed using the results of N rate response experiments conducted on the major soils of a given region. This soil-specific approach is based on corn yield response and associated economic return to incremental rates of N.[18,19] Two additional parameters influencing the economic optimum N application rate are the anticipated price of corn and the cost of fertilizer N. The data gathered from these experiments are the foundation for corn N recommendations throughout the upper Midwest United States.[22,23]

Additional Tests for Fine-Tuning Nitrogen Applications

The development of tests for assessing soil N levels provides additional tools for improving the efficiency of N fertilizer applications. These tests allow fertilizer recommendations to be adjusted to site-specific conditions that can influence N availability. Tests include the preplant soil profile nitrate test, [24] the pre–side-dress soil nitrate test,[25] plant analysis,[26,27] chlorophyll meters,[28] the end-of- season stalk nitrate test,[29] or the end-of-season soil nitrate test.[30]

Calibrated Soil Tests for Phosphorus and Potassium

In recent years, soil test recommendation programs for P, potassium (K), and other relatively immobile nutrients have tended to de-emphasize a soil buildup and maintenance philosophy in favor of a better balance between environmental and economic considerations by using a crop sufficiency approach.[23,31] These tests are calibrated by field experiments to obtain predictable crop yield responses. Such an approach adds extra emphasis to regular soil testing. It is recommended that soil tests be taken at least every 3 to 4 yr and more frequently on sandy and other soils of low buffering capacity.

Nutrient application recommendations based on soil test results can be accurate only if soil samples representative of the field of interest are collected. Samples that are unrepresentative of fields often result in recommendations that are misleading. Before collecting soil samples, relevant local guidance should be sought on the appropriate number samples to collect, as well as the methodology for collection. In addition, field history information should be provided with the soil samples in order to accurately adjust the fertility recommendations to account for nutrient credits from field-specific activities such as manure applications and legumes in the rotation.

Realistic Yield Goals

For many soil fertility programs, the recommendation of appropriate nutrient application rates is dependent on the establishment of realistic yield goals. Yield goal estimates that are too low will underestimate nutrient needs and can limit crop yield. Yield goal estimates that are too high will overestimate crop needs and result in soil nutrient levels beyond that needed by the crop, which in turn has the potential to increase nutrient contributions to water resources. Estimates should be based on field records and some cautious optimism—perhaps 10% above the recent 3 to 5 years average corn yield from a particular field. Yield goals reasonably higher than a multiyear average are suggested because annual yield variations due to factors other than nutrient application rates (primarily climatic factors) are often large.

Nutrient Credits

The integration of economic return and environmental quality protection requires that nutrients from all sources be considered. In the determination of supplemental fertilizer application rates, it is critical that nutrient contributions from manure, previous legume crops grown in the cropping rotation, and land-applied organic by-products are credited. Both economic and environmental benefits can result if the nutrient-supplying capacity of these nutrient sources is correctly estimated. Economically, commercial fertilizer application rates can often be reduced or eliminated entirely when nutrient credits are accounted. Environmentally, the prevention of overfertilization reduces potential threats to water quality.

Manure

Manure can supply crop nutrients as effectively as commercial fertilizers in amounts that can meet the total N, P, K, and sulfur need of many crops.[32] To utilize manure as a fertilizer resource, its application rate and nutrient- supplying capacity (i.e., plant-available nutrient content) need to be estimated.

Calibration of manure application equipment is key to estimating application rates. Calibration is a relatively easy task that can be done with platform scales or portable axle scales.[33] As a result of manure spreader calibration, an applicator will have a reasonable estimate of the manure application rate—provided the manure is uniformly applied across fields.

The most effective method for gauging the nutrient content of manure is to have samples analyzed by a commercial or university laboratory.[34] Large farm-to-farm variation can occur in manure nutrient content due to manure storage and handling techniques, livestock feed variations, or other farm management differences.[35] In instances when laboratory analysis is not convenient or available, estimates of crop nutrients supplied by animal manures can be made using published values for the average nutrient values of livestock manures common to a given state or region. These values are often provided by area universities. Note that not all the nutrients in manure are available to crops in the first year following application. When estimating the nutrient-supplying capabilities of manure applications, be certain that first-year crop-available nutrient content values are used to calculate the fertilizer value of the manure—not total nutrient content.

Legumes

Legume crops, such as alfalfa, clover, soybeans, and leguminous vegetables, have the ability to fix atmospheric N and convert it to a plant-available form. When grown in a rotation, some legumes can supply substantial amounts of N to a subsequent non-legume crop. For example, a dense stand of alfalfa can often provide most, if not all, of the N needed for a corn crop following it in a rotation.[36] An efficient nutrient management strategy needs to consider the N contributions of legumes to subsequent crops. The amount of legume N to credit varies regionally. Consult the local university extension service for appropriate recommendations.

Biosolids

The application of organic biosolids such as sewage sludge, whey, compost, or other organic wastes to cropland fields can be another source of potential crop nutrient credits. While the overall percentage of cropland acres receiving biosolids is relatively small when compared with manure or legumes, the nutrient contributions can be significant and should be accounted for prior to fertilizer applications. Special management and regulatory considerations pertain to the land application of these materials. Consult local regulations for further information.

Timing of Nutrient Applications

The timing of application is a major consideration for the management of mobile nutrients such as N. For less mobile nutrients, application timing is not a major factor affecting water quality protection. However, nutrient applications on frozen sloping soils or surface applications prior to periods likely to produce runoff events should be avoided to prevent nutrient contributions to surface waters.

Nitrogen Applications

The period between application and crop uptake of N is an important factor affecting the efficient utilization of N by the crop and the potential for loss of N via leaching, denitrification, and other processes.[37,38] Loss of N can be minimized by supplying it just prior to the period of greatest crop uptake. However, several considerations, such as soil, equipment, labor, and fertilizer price and availability are involved in determining the most convenient, economical, and environmentally safe N fertilizer application period.

Fall Nitrogen Applications

The advantages and disadvantages of fall N fertilizer applications are commonly debated. An increased risk for N loss with fall applications needs to be weighed against the fertilizer price and time management advantages (greater window for fertilizer application and spring planting) that can be associated with fall-applied N. The agronomic concern with fall N applications is that losses between application and crop uptake the following growing season will lower recovery of N and reduce crop yield. The environmental concern with fall application is that the N lost prior to crop uptake will leach into groundwater. Fall to spring precipitation, soil texture, and soil moisture conditions influence the potential for fall-applied N losses. If a soil is wet in the fall, rainfall may cause either leaching of nitrate in coarse soils or denitrification of nitrate in heavy, poorly drained soils. Long-term studies indicate that fall applications on medium-textured soils are 10%–15% less effective than the same amount of N applied spring preplant.[37] For both agronomic and environmental reasons, fall applications of N fertilizers are not recommended on coarse- textured soils or on shallow soils over fractured bedrock. If fall applications are to be made on other soils, it is recommended that ammonium-N sources be used and that the applications be delayed until soil temperatures are below thresholds of biological activity (i.e., 50°F) in order to slow the conversion of ammonium to nitrate by soil organisms. If fall applications must be made when soil temperatures are higher than 50°F, a nitrification inhibitor should be used in conjunction with the N fertilizer.

Preplant Nitrogen Applications

Spring preplant applications of N are usually agronomically and environmentally efficient on medium-textured, well-drained soils. The potential for N loss prior to crop uptake on these soils is relatively low with spring applications. If spring preplant applications of N are to be made on sandy soils, ammonium forms of N treated with a nitrification inhibitor should be used. Likewise, nitrification inhibitors should be used if spring preplant N is applied to poorly drained soils. Use of nitrification inhibitors reduces the potential for N loss compared with preplant applications without them; however, side-dress or split applications can be more effective and cost efficient than preplant applications with nitrification inhibitors.

Side-Dress Nitrogen Applications

Side-dress applications of N to row crops during the growing season are effective on all soils with the greatest benefit on sandy or heavy-textured, poorly drained soils.[38,39] The greatest efficiency of side-dress N applications is achieved when the application of N occurs just prior to the period of rapid N uptake by crops. This results in a shorter period of exposure to potential losses of N from leaching or denitrification. Table 1 illustrates the higher yield and crop recovery of N on sandy soils with side-dress applications. In these trials, use of side-dress N applications improved average N recovery over preplant applications by 17%. The use of side-dress or delayed N applications on sandy soils is essential

TABLE 1 Effect of Rate and Time of N Application on Corn Yield and Recovery of Applied N on Irrigated Plainfield Sand

	Relative Yield Increase[a]		N Recovery	
N Rate	Preplant	Side-dress	Preplant	Side-dress
(lb/a)	- - - (% over control)- -		- - - (%) - - -	
0	–	–	–	–
70	132	176	50	73
140	216	258	44	64
210	247	276	40	49
Average	197	237	45	62

[a] Side-dress treatments applied 6- weeks after planting.

Source: Bundy et al.[40]

for minimizing N loss to groundwater since unrecovered N on these soils will be lost through leaching prior to the next growing season. Side-dress N applications may also be of benefit on shallow soils over fractured bedrock.

Side-dressing N requires more management than preplant applications. To maximize efficiency, side-dress N applications must be properly timed to provide available N during the maximum N-uptake period for crops such as corn. An additional concern is that applications too late may result in lower yield and plant injury from root pruning and other physical damage.

Split Nitrogen Applications

Application of N fertilizer in several increments during the growing season can be an effective method for reducing N losses on sandy soils. However, a single well-timed side- dress application is often as effective as multiple applications.[41] Ideally, split applications supply N when needed by the crop and allow for N application rate adjustments based on early growing season weather or plant and soil tests. To be successful, the timing of application and placement of fertilizer materials are critical. Climatic factors, such as untimely rainfalls, may interfere with application schedules.

A common method for split N applications is via irrigation systems (fertigation). Multiple applications of fertilizer N can be injected into the irrigation water and applied to correspond with periods of maximum plant uptake. However, fertigation should not be relied upon as a sole method of applying N in a cropping season for the following reasons: 1) adequate rainfall during the early growing season could delay or eliminate the need for irrigation and subsequently delay fertilizer applications; and 2) leaching can result if N is applied through an irrigation system at a time when the crop does not need additional water.

Nitrification Inhibitors

Nitrification inhibitors are used with ammonium or ammonium-forming N fertilizers to improve N efficiency by slowing the conversion of ammonium to nitrate, thereby reducing the potential for losses of N that occur in the nitrate form (i.e., leaching and denitrification). The effectiveness of a nitrification inhibitor depends greatly on soil type, time of the year applied, N application rate, and soil moisture conditions that exist between the time of application and the time of N uptake by plants. Research has shown that the use of nitrification inhibitors on medium- and fine-textured soils with fall N applications, on poorly drained soils with fall or spring N applications, or on coarse-textured, irrigated soils with spring preplant N applications has the potential to increase corn yield and total crop recovery of N.[42] However, as noted earlier, side- dress applications alone are likely to be more effective on many of these soils. Fall applications of N with an inhibitor on sandy soils are not recommended. The cost of using nitrification inhibitors versus other strategies for minimizing N losses needs to be considered in an overall economic analysis of a grower's crop production system.

Controlled-Release Nitrogen Fertilizers

Controlled- (or slow-) release N fertilizers release their nutrients at gradual rates that, in theory, allow for increased plant uptake of N while minimizing losses due to leaching and volatilization. Although commonly used in high-value applications such as horticultural crops and turf, these fertilizer products have not been economical for widespread use in major agricultural crops due to relatively high cost and low crop prices. This may be changing due to cheaper controlled-release fertilizer products, higher N prices, and the demand for greater environmental protection.

Controlled-release fertilizers are broadly divided into uncoated and coated products. Uncoated products rely on inherent physical characteristics, such as low solubility, for their slow release. Coated products consist mostly of quick-release N sources surrounded by a barrier that prevents the N from releasing rapidly into the environment. Similar to earlier-developed materials, such as sulfur- coated urea, urea formaldehyde, and isobutylidene diurea, the critical concept is timing the release of N to correspond with crop need. Greenhouse and field studies have shown that polymer-coated urea can increase crop yield and N use efficiency in soils prone to leaching losses.[43–45]

Other Nutrient Applications

The timing of P and K applications is less critical as these nutrients are generally strongly held in the soil. Some water quality concerns may exist where P is surface broadcast on fields that have not been tilled as this results in less fertilizer–soil contact. Fall applications of K are not recommended on organic soils (peat or muck) since these soils do not effectively hold K against leaching losses. Considerations for P and K applications on other soils include the amount of material to be applied, the size of the application window, and the resources and available equipment.

Nutrient Placement

The placement of nutrients on cropland can influence their effectiveness as well as their potential ability to affect water quality. The concern with N placement focuses mainly on preventing N loss through ammonia volatilization. Applications of N in the form of urea or N solutions need to be incorporated into the soil by rainfall, irrigation, injection, or tillage. The amount of volatilization loss that occurs with surface N applications depends on factors such as soil pH, temperature, moisture, and crop residue. Minimal volatilization losses of N can be expected if surface applications are incorporated within 3 to 4 days—provided temperatures are low (<50°F) and the soil is moist.[46] A late spring or summer application should be incorporated within a day or two because higher temperatures and the chance of longer periods without rainfall could lead to significant N volatilization losses.

The placement of P nutrient sources can directly influence the amount of P transported to lakes and streams by surface runoff. If P fertilizer is broadcast on the soil surface and not incorporated, the amount of P in runoff water can rise sharply and have a greater potential impact on surface water quality than soil surfaces where P was incorporated.[47,48] Phosphorus is strongly bound to soil particles; however, adequate soil-to-P contact must occur to allow for adsorption. Incorporation by tillage or subsurface band placement of fertilizers is a very effective means of achieving this contact. To avoid enriching surface waters with soil nutrients, it is recommended that annual fertilizer applications for row crops, such as corn, be band-applied near the row as starter fertilizer at planting. Annual starter applications of P (and K) can usually supply all of the P required for corn. This practice reduces the chance for P enrichment of the soil surface and reduces potential P in cropland runoff. Band fertilizer placement ideally enriches about 20% of the plow layer volume.[49] If large broadcast P fertilizer applications are needed to increase low soil P levels, these applications should be followed by incorporation as soon as possible.

Variable-Rate Fertilizer Technologies

Nutrient availability in any field varies both spatially and temporally.[50,51] To address this, site-specific (precision agriculture) management techniques and tools have developed over the past 20 years. Global positioning systems and geographic information systems, along with crop nutrient sensing systems, have allowed agricultural management decisions to be made with greater detail and precision. As a result, producers are able to manage nutrient variability within fields at increasingly finer resolution than in the past with inherent improvements in nutrient use efficiency, crop yields, and environmental stewardship.[52,53] While progress has been made, widespread acceptance of these technologies has been limited by costs, sampling requirements, required technical inputs, and the limited ability of current equipment to physically deliver nutrients at adjustable rates corresponding to field variability.[54]

Manure Management

Manure applications to cropland provide nutrients essential for crop growth, add organic matter to soil, and improve soil physical and biological conditions. The major environmental concerns associated with manure application are related to its potential for overloading soils with nutrients if manure applications exceed crop needs and to direct runoff from manured fields to surface waters.

Manure Application Rates

Manure is often applied to cropland at rates that attempt to meet the N need of the intended crop. This strategy maximizes potential manure application rates and is preferred if the amount of land available for application is limited. In addition, a N-based strategy is usually time and labor efficient. A consequence of this approach can be the buildup of P in soils to excessive levels, which in turn increases the potential for P losses via runoff and soil erosion.[55–57] For example, the plant-available N and P contents of dairy manure are about equal. The N need of corn, however, is greater than the crop's need for P. A consequence of applying manure at rates to meet the N need of corn is that P applications will exceed crop removal (Figure 3). The result is a buildup of P in cropland soils.[58] Long term manure applications have elevated the soil P level of many soils above the range necessary for optimum crop growth.[59]

If maximum manure nutrient efficiency is the goal, rates of application need to be based on the nutrient present at the highest level relative to crop needs. For corn, this nutrient would be P. Manure application rates that meet the P requirement of corn are typically much lower than N-based rates. Subsequently, additional N will need to be supplied from other nutrient sources (Figure 4). A P-based manure application strategy results in lower manure application rates, but it is less likely to elevate soil test P values. It has the disadvantages of being less efficient with respect to labor, energy, time, and economics.[60] A P-based strategy for manure applications requires spreading manure on a much larger acreage than is required for a N-based manure application.

Manure Application Timing

Manure application timing is an important management practice for minimizing nutrient contributions to surface waters. Manure should not be spread on sloping lands any time a runoff-producing event is likely. Unfortunately, runoff-producing events are difficult to predict, and the elimination of manure

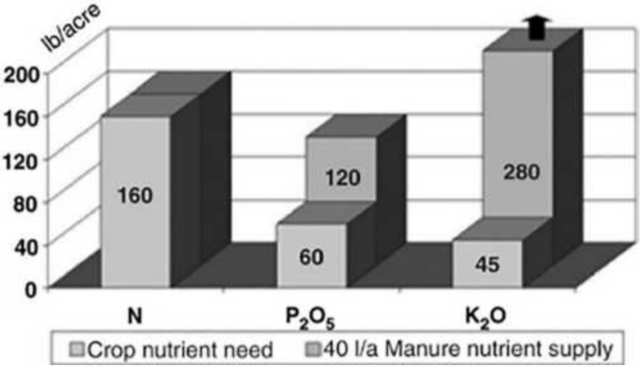


FIGURE 3 Nitrogen-based manure application strategy for corn. Note: Standard convention in the United States is to express soil P and K levels in elemental form (i.e., ppm of P and K) while expressing P and K fertilizer application rates and analysis in oxide form (i.e., lb/acre of P_2Os and K_2O).
Source: Understanding soil phosphorus.[58]

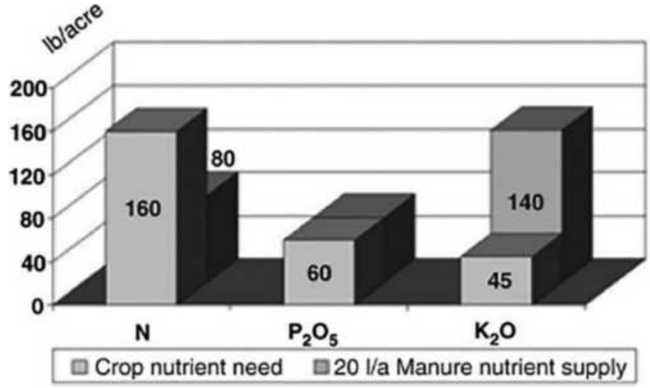


FIGURE 4 Phosphorus-based manure application strategy for corn.
Source: Sturgul and Bundy.[58]

applications to sloping lands is seldom a practical option. For farmers in the upper Midwest, the period of major concern with manure spreading is late fall, winter, and early spring months. Manure applied on frozen ground has an increased likelihood for running off to surface waters due to snow melt and/or early spring rains.

If winter applications of manure must be made, the risk for nutrient loss should be minimized to the greatest extent possible. Manure applications to frozen fields should be limited to those of slight slope (generally less than 6%) that are preferably covered in previous crop residue, roughly tilled on the land contour, or protected from upslope runoff.[61,62] More steeply sloping fields that are intended for manure applications need to have soil and water conservation practices in place. Manure should not be applied to frozen soils on steeply sloping fields (generally 12% or greater).[61,62]

Site Considerations for Manure Applications

In addition to the slope criteria discussed previously, other site considerations for manure applications should include existing soil fertility levels, soil depth, soil texture, soil erodibility factors, and field proximity to water bodies.

In many areas, general recommendations, or even specific regulations, exist for reducing or eliminating manure applications to fields that have elevated levels of soil P. Numerous studies have found a correlation between elevated levels of soil P and the amount of P carried in runoff from agricultural fields.[55–57] As soil P levels become elevated, crop rotations should be diversified to include crops with a high demand for P (such as alfalfa), which can draw down soil P. When soil P levels become excessively high, manure applications should be discontinued until soil test levels decrease.[61,62] Soil runoff and erosion control practices such as residue management, conservation tillage, contour farming, and others are strongly recommended on soils with P levels in excess of crop needs. When planning manure applications, prioritize those fields low in soil fertility (particularly P) and strive to distribute manure across the available fields to avoid the excessive buildup of soil nutrients that results from repeated applications to the same sites.

Most soils have a high capacity for assimilating nutrients from manure. However, in locations of highly permeable or shallow soils over fractured bedrock, groundwater issues associated with the application of manure can result. Manure should not be applied to shallow soils (generally less than 10 in.) over fractured bedrock. Incorporation of manure shortly after application on moderately shallow soils will allow for increased soil adsorption of nutrients. Manure should not be applied to frozen, shallow soils.[61,62]

Movement of nitrate-N to groundwater is more likely on excessively drained (sandy) soils. Manure applications in early fall on these soils where no actively growing crop is present to utilize the N may allow for the conversion of organic N to nitrate, which is then subject to leaching losses. Manure should not be applied to sands or loamy sands in the fall when soil temperatures are greater than 50°F, unless there is an overwintering cover crop present to utilize the N. In the absence of a cover crop, manure applications to sandy soils should take place when soil temperatures are below 50°F.[61,62] The conversion of ammonium-N to nitrate-N is significantly reduced at soil temperatures below 50°F.

The main site characteristics affecting nutrient contributions to surface waters are those that affect soil runoff and erosion. These include slope, soil erodibility and infiltration, rainfall, cropping system, and the presence of soil conservation practices. Site-related management practices dealing specifically with manure placement to protect surface water include the following: 1) not applying manure (or other nutrients) to grassed waterways, terrace channels, open surface drains, or other areas where surface flow may concentrate; 2) restricting manure applications within designated floodplain or stream and lake setback distances; and 3) prohibiting manure applications in these areas when soils are frozen or saturated.

Manure Storage

During periods when suitable sites for land application of manure are not available (i.e., soils are frozen or seasonally saturated), the use of manure storage facilities is recommended. Storage facilities allow manure to be stored until conditions permit land application and incorporation. In addition, storage facilities can minimize nutrient losses resulting from volatilization of ammonia and be more convenient for calibrated land applications. With the exception of those systems designed to filter leachate, storage systems should retain liquid manure and prevent runoff from precipitation on stored waste. It is imperative that manure storage facilities be located and constructed such that the risk of seepage to groundwater is minimized. With regards to maximum nutrient efficiency and water quality protection, it is critical that appropriate application techniques and accurate nutrient crediting of the manure resource are utilized when the storage facility is emptied.

Livestock Feed Management

On individual farms and in many areas of agricultural livestock production, inputs of P in feed and fertilizer exceed outputs of P contained in crop and animal produce leaving the farm or region (Figure 5). This is especially true in areas where concentrated livestock production is prevalent.[63] The National Research Council[64] estimated that only 30% of the fertilizer and feed P imported onto farms is exported

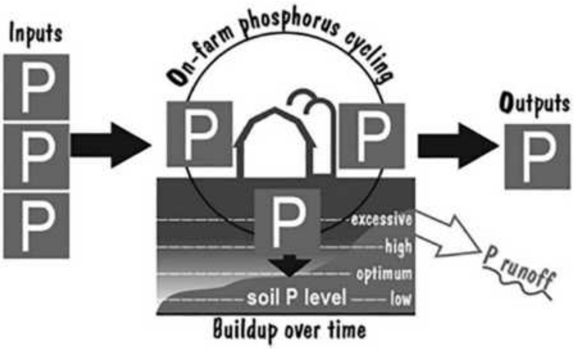


FIGURE 5 On-farm phosphorus cycle.
Source: Sturgul and Bundy.[58]

in crops and animal produce. The surplus 70% of the P is remaining on farm and leading to the excessive enrichment of soil P.

Livestock feed inputs have been found to be a major contributing factor to on-farm P surpluses.[65–67] Soil buildup of P is accelerated when livestock are overfed P in dietary rations. Phosphorus excretion in manure is directly related to the level of P intake.[68,69] High P in livestock dietary intake directly correlates with higher bypass P as reflected in elevated P content of livestock manure (Table 2). Overuse of dietary P supplements accelerates the buildup of soil test P to excessive levels and increases the potential for P losses from manured fields (Figure 6).[70] Another consequence is an increase in land required for application of manure if P-based rate limitations are to be met.

Additional dietary P management options involve plant and livestock genetic manipulation for more effective manure-P management from monogastric animals (nonruminants such as swine and poultry). All these techniques attempt to reduce the P content in manure of monogastric animals by improving the efficiency with which the animal extracts P from feed. An increase in P uptake by the animal from feed grains will reduce the amount of P that bypasses the animal via the manure. Increasing animal uptake of P can allow manure application rates to continue due to a slower buildup of soil P because of the reduced P content of the manure.

Reducing the phytate level of feed grains by use of low-phytate, high-available-phosphate (HAP) varieties is one feed management strategy for lowering manure P. In corn and most feed grain plants, P is stored in the phytate form, which is largely unavailable to nonruminant livestock. As a consequence, swine and poultry feed is routinely supplemented with P. The unutilized phytate-P from the plant is excreted by the animals, resulting in manure that is enriched in P content.[71] Low-phytate grain hybrids that will store P in the available phosphate form rather than as phytate are available.[72] Corn

TABLE 2 Annual Phosphorus Fed to and Excreted by a Lactating Cow

Dietary P Level	Supplemental P	Fecal P
(%)	- - - - - - - - - lb/Cow/Year- - - - - - - - - -	
0.35	0	42
0.38	5.5	47
0.48	23	65
0.55	36	78

Source: Powell et al.[65]

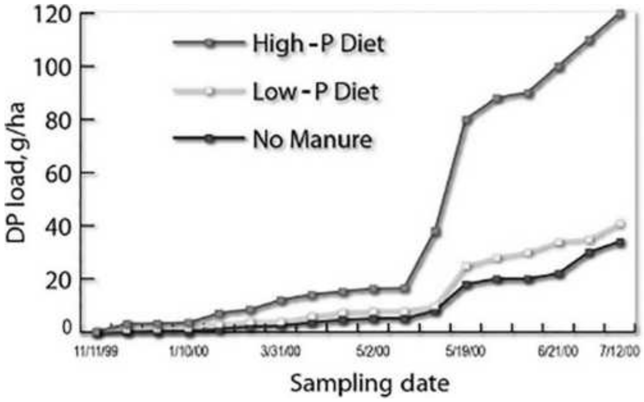


FIGURE 6 Cumulative dissolved phosphorus (DP) load in natural runoff from manures differing in P content.
Source: Ebeling et al.[70]

has been the crop most extensively developed. Phosphorus availability to mono-gastrics from low-phytate corn is about two to three times higher than from normal corn.[73] Subsequently, the P content of manure is reduced. Plant breeders are working to incorporate the low-phytate trait into commercially competitive hybrids.

Another option for reducing the P content of manure from monogastric livestock is the use of commercially produced enzymes as a feed supplement. Phytase enzymes are capable of releasing phytate-P from plants into animal-available forms. Phytase enzymes occur naturally in some microorganisms, plants, and animals, such as ruminants (cattle). Monogastric animals lack phytase and can only poorly utilize the P reserves in many grains.[71] By adding phytase enzymes to nonruminant animal feed, the efficiency of P uptake during digestion can be increased with an associated reduction in the P content of monogastric manure.[74] In a study by Baxter et al., [75] where phytase additives were combined with low-phytate corn, a 60% reduction in P excretion was recorded. While the phytase enzyme has been shown to decrease the need for mineral P additions, the economics of its use as a routine feed additive need to be considered.[63]

Irrigation Water Management

When to apply irrigation water and how much to apply depend on crop, growth stage, and soil properties. Overirrigation, or rainfall on recently irrigated soils, can leach nitrate and other contaminants below the root zone and into groundwater. Accurate irrigation scheduling that considers soil water holding capacity, crop growth stage, evapotranspiration, rainfall, and previous irrigation to determine the timing and amount of irrigation water to be applied can reduce the risk of leaching losses.

Soil Conservation Practices

Land-use activities associated with agriculture can increase the potential for runoff and soil erosion. Consequences of cropland erosion include loss of fertile topsoil, accelerated eutrophication and sedimentation of surface waters, destruction of fish and wildlife habitat, and decreased recreational and aesthetic value of surface waters. The key to minimizing nutrient contributions to surface waters is to reduce the amount of runoff and eroded sediment reaching them. Numerous management practices for the control of runoff and soil erosion have been researched, developed, and implemented. Runoff and erosion control practices range from changes in agricultural land management (cover crops, diverse crop rotations, conservation tillage, contour farming, contour strip cropping, etc.), to the installation of structural devices (buffer strips, diversions, grade stabilization structures, grassed waterways, terraces, etc.), Substantial emphasis is currently being placed on the benefits and installation of vegetative buffer strips along riparian corridors, which can reduce the sediment and nutrient content of runoff waters reaching them.[76] The width of an effective buffer strip is often debated but varies according to land slope, type of vegetative cover, watershed characteristics, etc.[77]

Recently, the U.S. Department of Agriculture (USDA) reported that cropland conservation practices installed and applied by agricultural producers are indeed reducing sediment, nutrient, and pesticide losses from farm fields.[78] Key findings from this report on the effects of conservation practices on cropland in the Upper Mississippi River Basin illustrate the following: 1) suites of conservation practices work better than single practices; 2) targeting critical acres improves practice effectiveness significantly; and 3) the beneficial effect of conservation practices is greatest on the most vulnerable acres (highly erodible land and soils prone to leaching). The study also found that the use of soil conservation practices within the Upper Mississippi River Basin has reduced sediment loss by 69% compared with estimates of soil loss from the same land area without conservation practices in place. The study's authors concluded that the improved use of cropland conservation practices along with the consistent use of nutrient management practices would further reduce the risk of nutrient movement from fields to rivers and streams within the basin.

Farm Nutrient Management Plans

A farm nutrient management plan is a dynamic, regularly updated strategy for obtaining the maximum economic return from both on- and off-farm nutrient resources in a manner that protects the quality of nearby water resources. While a plan is specific for an individual farming operation, there are common components to all nutrient management plans. These include the following.

Soil Test Reports

Complete and accurate soil tests are the starting point of any farm nutrient management plan. All cropland fields must be tested or have been tested recently to ensure the best possible fertilizer recommendations.

Assessment of On-Farm Nutrient Resources

The amount of crop nutrients supplied to cropland fields from on-farm nutrient resources such as manure, legumes, and organic by-products needs to be determined.

Nutrient Crediting

Once on-farm nutrient resources are determined, commercial fertilizer applications need to be reduced to reflect these nutrient credits. This action can lower fertilizer expenditures and protect water quality by eliminating nutrient applications in excess of crop need. Management skills come into play when determining nutrient credits. For example, to properly credit manure-supplied nutrients, both the manure application rate and the crop-available nutrient content of the manure must be known. To credit the N available to crops following alfalfa, the condition of the alfalfa stand as well as last cutting date need to be known.

Manure Inventory

Perhaps the most challenging aspect of developing and implementing a farm nutrient management plan is the advanced planning of manure applications to cropland fields. This involves estimating the amount of manure produced on a farm and then planning specific manure application rates for individual cropland fields.

Manure Spreading Plan

A major component of any nutrient management plan for livestock operations will deal with a manure spreading plan. The amount of manure the farm produces has to be applied to fields in a manner that considers both environmental and agronomic consequences. Manure applications rates should not exceed crop nutrient need as identified by the soil test report. The nutrient management plan should prioritize those fields that would benefit the most from the manure-supplied nutrients while posing little threat to water quality. The nutrient management plan must also identify fields with manure spreading restrictions.

The seasonal timing of manure applications to cropland should be identified in the nutrient management plan. The timing of planned manure applications will depend upon each farm's manure handling system. Manure application periods for farms with manure storage will be significantly different than those for farms that haul manure on a daily basis.

Consistency with Farm Conservation Plans

A nutrient management plan should be consistent with a farm's soil conservation plan. Operations participating in federal farm programs usually are required to have a soil conservation plan. Conservation plans contain needed information on planned crop rotations, slopes of fields (which are important when planning manure applications), and the conservation measures required to maintain soil erosion rates at tolerable levels.

Compliance with Nutrient Management Standards, Rules, and Regulations

Nutrient management plan criteria and requirements are often defined in standards developed by federal, state, or local government (i.e., the USDA–Natural Resources Conservation Service's Nutrient Management Standard 590). A nutrient management plan complying with a standard(s) is often a requirement for participation in federal and state government farm programs.

Conclusion

The previous text provides a brief summary of general nutrient management practices for crop production. This is not a complete inventory but, rather, an overview of soil fertility management options available to growers for protecting water quality and improving farm profitability. Best management practices for agricultural nutrients can vary widely from one region to another due to differences in cropping, topographic, environmental, and economic conditions. However, central to any nutrient management strategy would be an accurate assessment of nutrient need along with an accounting of on-farm nutrient resources such as manure, legumes, etc. The application of supplemental nutrients should be timed for maximum crop uptake and minimal chance of off-site movement. Nutrient management practices for optimizing crop production while protecting water quality are best summarized in a nutrient management plan that is regularly updated and tailored to the unique landscape characteristics and prevalent agricultural practices of individual farming operations.

References

1. United States Environmental Protection Agency (US EPA). Nutrient Pollution: The Problem. http://www.epa.gov/nutrientpollution/problem/index.html. April 23, 2012.
2. Spalding, R.F.; Exner, M.E. Occurrence of nitrate in ground water—A review. J. Environ. Qual. **1993**, 22, 392–402.
3. Blodgett, J.E.; Clark, E.H. Fertilizers, nitrates, and groundwater: An overview. In Proceedings of the Colloquium of Agrichemical Management to Protect Water Quality; Board of Agriculture, National Research Council: Washington, DC, 1986.
4. Madison, R.J.; Brunett, J.O. 1984. Overview of the occurrence of nitrate in groundwater of the United States. In *U.S. Geological Service National Water Summary, Water- Supply Paper 2275;* U.S. Government Printing Office: Washington, DC, 1984; 93–105.
5. United States Environmental Protection Agency (US EPA). *National Primary Drinking Water Regulations,* EPA pub. no. 816-F-09-004, http://water.epa.gov/drink/contaminants/upload/mcl-2.pdf (accessed March 6, 2012).
6. Rabalais, N.N.; Wiseman, W.I., Jr.; Turner, R.E. Comparison of continuous records of near-bottom dissolved oxygen from the hypoxic zone along the Louisiana coast. Estuaries **1994**, *14,* 850–861.
7. Goolsby, D.A.; Battaglin, W.A. *Nitrogen in the Mississippi Basin—Estimating sources and predicting flux to the Gulf of Mexico,* USGS Fact Sheet 135-00; USGS: Washington, DC, 2000, http://ks.water.usgs.gov/pubs/fact-sheets/fs.135-00.pdf (accessed May 14, 2012).

8. Parry, R. Agricultural phosphorus and water quality: A U.S. Environmental Protection Agency perspective. J. Environ. Qual. **1998**, 27, 258–261.
9. Correll, D.L. The role of phosphorus in the eutrophication of receiving waters: A review. J. Environ. Qual. **1998**, 27, 261–266.
10. Bundy, L.G.; Andraski, T.W.; Wolkowski, R.P. Nitrogen credits in soybean–corn crop sequences on three soils. Agron. J. **1993**, *85,* 1061–1067.
11. Delgado, J.A.; Mosier, A.R.; Valentine, A.W.; Schimel, D.S.; Parton, W.J. Long-term N studies in a catena of the shortgrass steppe. Biogeochemistry **1996**, *32,* 41–50.
12. Follett, R.F.; Keeney, D.R.; Cruise, R.M., Eds. *Managing Nitrogen for Groundwater Quality and Farm Profitability.* Soil Science Society of America: Madison, WI, 1991.
13. Newhould, P. The use of nitrogen fertilizer in agriculture. Where do we go practically and ecologically? Plant Soil **1989**, *115,* 297–311.
14. Randall, G.W.; Huggins, D.R.; Russelle, M.P.; Fuchs, A.J.; Nelson, W.W.; Anderson, J.L. Nitrate losses through subsurface tile drainage in conservation reserve program, alfalfa, and row crop systems. J. Environ. Qual. **1997**, *26,* 1240–1247.
15. Randall, G.W.; Iraqavarapu, T.K.; Schmitt, M.A. Nutrient losses in subsurface drainage water from dairy manure and urea applied to corn. J. Environ. Qual. **2000**, *29,* 1244–1252.
16. Andraski, T.W.; Bundy, L.G.; Brye, K.R. Crop management and corn nitrogen rate effects on nitrate leaching. J. Environ. Qual. **2000**, *29,* 1095–1103.
17. Meisinger, J.J. Evaluating plant available nitrogen in soil- crop systems. In *Nitrogen in Crop Production;* Hauck, R.D. et al., Eds.; ASA, CSSA, and SSSA: Madison, WI, 1984; 391–416.
18. Vanotti, M.B.; Bundy, L.G. An alternative rationale for corn nitrogen fertilizer recommendations. J. Prod. Agric. **1994**, *7,* 243–249.
19. Vanotti, M.B.; Bundy, L.G. Corn nitrogen recommendations based on yield response data. J. Prod. Agric. **1994**, *7,* 249–256.
20. Blackmer, A.M.; Binford, G.D.; Morris, T.; Meese, B. Effects of rates of nitrogen fertilization on corn yields, nitrogen losses from soils, and energy consumption. In *1991 Progress Report of the Integrated Farm Management Demonstration Program;* Iowa State University Pm-1467, 1992; 2.1–2.6.
21. Nafziger, E.D.; Sawyer, J.E.; Hoeft, R.G. Formulating N recommendations for corn in the corn belt using recent data. In Proceedings of the North Central Extension–Industry Soil Fertility Conference, Des Moines, IA, Nov. 17–18, 2004.
22. Sawyer, J.E.; Nafziger, E.D.; Randall, G.; Bundy, L.G.; Rehm, G.; Joern, B. *Concepts and Rationale for Regional Nitrogen Rate Guidelines for Corn*; Iowa State University Extension pub. PM 2015, 2006.
23. Laboski, C.A.M.; Peters, J.B.; Bundy, L.G. *Nutrient Application Guidelines for Field, Vegetable, and Fruit Crops in Wisconsin;* University of Wisconsin Extension pub. A2809, 2006.
24. Bundy, L.G.; Meisinger, J.J. Nitrogen availability indices. In *Methods of Soil Analysis: Biochemical and Microbial Properties*, Monogram 5; Weaver, R.W., Ed.; Soil Science Society of America: Madison, WI, 1994; 951–984
25. Magdoff, F.R.; Ross, D.; Amadon, J. A soil test for nitrogen availability to corn. Soil Sci. Soc. Am. J. **1984**, *48,* 1301–1304.
26. Kalra, Y.P., Ed. *Handbook of Reference Methods for Plant Analysis.* CRC Press, Inc; Boca Raton, Florida, 1998; 320 pp.
27. Jones, J.B. *Laboratory Guide for Conducting Soils Tests and Plant Analysis.* CRC Press, Inc; New York, Washington, D.C., 2001; 384 pp.
28. Varvel, G.E.; Schepers, J.S.; Francis, D.D. Chlorophyll meter and stalk nitrate techniques as complementary indices for residual nitrogen. J. Prod. Agric. **1997**, *10,* 147–151.
29. Blackmer, A.M.; Mallarino, A.P. *Cornstalk Testing to Evaluate Nitrogen Management*; Iowa State University Extension Pub. PM1584, 1996.

30. Sullivan, D.M.; Cogger, C.G. *Post-Harvest Soil Nitrate Testing for Manured Cropping Systems West of the Cascades;* Oregon State University Extension pub EM 8832-E, 2003, http://extension.oregonstate.edu/catalog/pdf/em/em8832-e.pdf (accessed May 14, 2012).
31. Sparks, D.L., Ed. *Advances in Agronomy.* Academic Press; San Diego, California, 1999; Vol. 67, 320 pp.
32. Midwest Plan Service. *Manure Characteristics–Manure Management Systems Series;* MWPS-18-S1D, Iowa State University; Ames, Iowa, 2004.
33. Nutrient and Pest Management (NPM) Program. *Know How Much You Haul;* University of Wisconsin Extension; 2 pp, http://ipcm.wisc.edu/Publications/tabid/54/Default.aspx (accessed May 14, 2012).
34. Peters, J.B., Ed. *Recommended Methods Of Manure Analysis;* University of Wisconsin Extension Pub. A3769, 2003.
35. Peters, J.B.; Combs, S.M. Variability in manure analysis as influenced by sampling and management. In Proceedings of the 1998 Wisconsin Fertilizer Dealers Update Meetings, University of Wisconsin, Madison, WI, Dec. 1998.
36. Bundy, L.G.; Kelling, K.A.; Good, L.W. *Using Legumes as a Nitrogen Source*; University of Wisconsin Extension; 8 pp, http://ipcm.wisc.edu/Publications/tabid/54/Default.aspx (accessed May 14, 2012).
37. Bundy, L.G. Review—Timing nitrogen applications to maximize fertilizer efficiency and crop response in conventional corn production. J. Fert. Issues **1986**, 3, 99–106.
38. Vitosh, M.L. *Nitrogen Management Strategies for Corn Producers;* Michigan State University Cooperative Extension Service Bull. WQ06; 1985; 6 pp.
39. Randall, G.W. Improved N management can alleviate groundwater pollution. Solutions **1986**, *30* (5), 44–49.
40. Bundy, L.G.; Kelling, K.A.; Schulte, E.E.; Combs, S.M.; Wolkowski, R.P.; Sturgul, S.J. Nutrient management: practices for Wisconsin corn production and water quality protection; University of Wisconsin Extension Pub. A3557, 1994; 27 pp.
41. Bundy, L.G. Timing nitrogen applications to maximize fertilizer efficiency and crop response in conventional corn production. J. Fert. Issues **1986**, 3, 99–106.
42. Nelson, D.W.; Huber, D. Nitrification inhibitors for corn production. In *National Corn Handbook 55*; Iowa State University Extension, 2001; 6 pp.
43. Wang, F.L.; Alva, A.K. Leaching of nitrogen from slow release urea sources in sandy soils. Soil Sci. Soc. Am. J. **1996**, *60,* 1454–1458.
44. Delgado, J.A.; Mosier, A.R. Mitigation alternatives to decrease nitrous oxide emissions and urea-nitrogen loss and their effects on methane flux. J. Environ. Qual. **1996**, *25*, 1105–1111.
45. Shoji, S.; Delgado, J.A.; Mosier, A.R.; Miura, Y. Use of controlled release fertilizers and nitrification inhibitors to increase nitrogen use efficiency and to conserve air and water quality. Comm. Soil Sci. Plant Anal. **2001**, *31,* 1051–1070.
46. Bundy, L.G. *Understanding Plant Nutrients: Urea—Its Use and Problems;* University of Wisconsin Cooperative Extension Service Bull. A2989, 1985; 4 pp.
47. Baker, J.L.; Laflen, J.M. Effects of corn residue and fertilizer management on soluble nutrient runoff losses. Trans. ASAE **1982**, *21,* 893–898.
48. Mueller, D.H.; Wendt, R.C.; Daniel, T.C. Phosphorus losses as affected by tillage and manure application. Soil Sci. Soc. Am. J. **1984**, *48,* 901–905.
49. Bundy, L.G. *Corn Fertilization;* University of Wisconsin Cooperative Extension Service Bull. A3340, 1998.
50. Legg, J.O.; Meisinger, J.J. Soil nitrogen budgets. In *Nitrogen in Agricultural Soils,* Agronomy Monogram 22; Stevenson, F.J., Brewner, J.M., Hanck, R.D., Keeney, D.R., Eds; American Society of Agronomy/Soil Science Society of America: Madison, WI, 1982; 503–566.
51. Jolkela, W.E.; Randall, G.W. Corn yield and residual nitrate as affected by time and rate of nitrogen application. Agron. J. **1989**, *81,* 720–726.

52. Fergusson, R.B.; Hergert, G.W.; Schepers, J.S.; Crawford, C.A.; Cahoon, J.E.; Peterson, T.A. Site-specific nitrogen management of irrigated maize: Yield and soil residual nitrate effect. Soil Sci. Soc. Amer. J. **2002**, *64,* 544–553.
53. Berry, J.R.; Delgado, J.A.; Khosla, R.; Pierce, F.J. Precision conservation for environmental sustainability. J. Soil Water Conserv. **2003**, *58,* 332–339.
54. Masek, T.J.; Schepers, J.S.; Mason, S.C.; Francis, D.D. Use of precision farming to improve application of feedlot waste to increase use efficiency and protect water quality. Commun. Soil Sci. Plant Anal. **2001**, *33*, 1355–1369.
55. Bundy, L.G.; Andraski, T.W.; Powell, J.M. Management practice effects on phosphorus losses in runoff in corn production systems. J. Environ. Qual. **2001**, *30,* 1822–1828.
56. Pote, D.H.; Daniel, T.C.; Sharpley, A.N.; Moore, P.A.; Edwards, D.R.; Nichols, D.J. Relating extractable soil phosphorus to phosphorus losses in runoff. Soil Sci. Soc. Am. J. **1996**, *60*, 855–859.
57. Sharpley, A.N. Identifying sites vulnerable to phosphorus loss in agricultural runoff. J. Environ. Qual. **1995**, *24,* 947–951.
58. Sturgul, S.J.; Bundy, L.G. *Understanding Soil Phosphorus;* University of Wisconsin Cooperative Extension Service Bull. A3771, 2004.
59. Sims, J.T. Environmental soil testing for phosphorus. J. Prod. Agric. **1993**, *6*, 501–506.
60. Bosch, D.J.; Zhu, M.; Kornegay, E.T. Net returns from microbial phytase when crop applications of swine manure are limited by phosphorus. J. Prod. Agric. **1998**, *11*, 205–213.
61. USDA–National Resources Conservation Service (NRCS). *Nutrient Management Code 590 Conservation Practice Standard*; National Resources Conservation Service: Washington, DC, 2005.
62. Madison, F.W.; Kelling, K.A.; Massie, L.; Ward-Good, L. *Guidelines for Applying Manure to Cropland and Pastures in Wisconsin.* University of Wisconsin Cooperative Extension Service Bull. A3392, 1998.
63. Daniel, T.C.; Sharpley, A.N.; Lemunyon, J.L. Agricultural phosphorus and eutrophication: A symposium overview. J. Environ. Qual. **1998**, *27*, 251–257.
64. National Research Council (NRC). *Nutrient Requirements of Dairy Cattle*, 7th Revised Ed; National Academy Press: Washington, DC, 2001.
65. Powell, J.M.; Wu, Z. Satter, L.D. Dairy diet effects on phosphorus cycles of cropland. J. Soil Water Conserv. **2001**, *56* (1), 22–26.
66. Satter, L.D.; Wu, Z. Reducing manure phosphorus by dairy diet manipulation. In Proceedings of the 1999 Fertilizer, Aglime and Pest Management Conference 38, Madison, WI; 1999; 183–192.
67. Sharpley, A.N.; Daniel, T.C.; Sims, J.T.; Lemunyon, J.; Parry, R. *Agricultural Phosphorus and Eutrophication*, USDA-ARS pub. no. ARS-149, 1999; 37 pp.
68. Khorasani, G.R.; Janzen, R.A.; McGill, W.B.; Kenelly, J.J. Site and extent of mineral absorption in lactating cows fed whole crop cereal grain silage or alfalfa silage. J. Animal Sci. **1997**, *75*, 239–248.
69. Metcalf, J.A.; Mansbridge, R.J.; Blake, J.S. Potential for increasing the efficiency of nitrogen and phosphorus use in lactating dairy cows. Anim. Sci. **1996**, *62,* 636.
70. Ebeling, A.M.; Bundy, L.G.; Andraski, T.W.; Powell, J.M. Dairy diet phosphorus effects on phosphorus losses in runoff from land applied manure. Soil Sci. Soc. Am. J. **2002**, *66*, 284–291.
71. Doerge, T.A. Low-phytate corn: A crop genetic approach to manure-P management. In Proceedings of the 1999 Wisconsin Fertilizer, Aglime, and Pest Management Conference, Madison, WI; 1999; 175–183.
72. Raboy, V.; Young, K.; Gerbasi, P. Maize low phytic acid (lpa) mutants. In the 4th International Congress of Plant Molecular Biology: Abs. No. 1827; 1994.
73. Ertl, D.S.; Young, K.A.; Raboy, V. Plant genetic approaches to phosphorus management in agricultural production. J. Environ. Qual. **1998**, *27*, 299–304.

74. Kornegay, E.T. Nutritional, environmental, and economic considerations for using phytase in pig and poultry diets. In *Nutrient Management of Food Animals to Enhance and Protect the Environment*; CRC Press, Inc; New York, 1996; 277–302.
75. Baxter, C.A.; Joern, B.C.; Adeola, L.; Brokish, J.E. Dietary P management to reduce soil P loading from pig manure. *Annual Progress Report to Pioneer Hi-Bred International, Inc;* 1998; 1–6.
76. Daniels, R.B.; Gilliam, J.W. Sediment and chemical load reduction by grass and riparian filters. Soil Sci. Soc. Am. J. **1996**, *60,* 246–251.
77. Schmitt, T.J.; Bosskey, M.G.; Hoagland, K.D. Filter strip performance and processes for different vegetation, widths, and contaminants. J. Environ. Qual. **1999**, *28,* 1479–1489.
78. USDA-NRCS. *Conservation Effects Assessment Project (CEAP): Assessment of the Effects of Conservation Practices on Cultivated Cropland in the Upper Mississippi River Basin*; 146 pp, http://www.nrcs.usda.gov/wps/portal/nrcs/detail/national/technical/nra/?&cid=nrcs143_014161 (accessed May 14, 2012).

50

Nutrients: Bioavailability and Plant Uptake

Niels Erik Nielsen

Soil Plant System

The movement of any nutrient element, M, from solid soil constituents to the root surface, and its entry into plants, can be divided into a sequence of processes (steps), as illustrated in Figure 1, which also indicates major agronomical actions used to improve nutrition and growth of crop plants. The ↔denotes solid-phase processes slowly approaching equilibrium, or microbial-mediated net mineralization of N, S and P, for example. Also denoted are the source/sink processes by which diffusible nutrients are being produced or removed by chemical, physical, and biological transformation processes. In Figure 1, L denotes ligands that are any dissolved solute reacting with M to form ML, which are organic complexes and ion pairs dissolved in the soil solution. The occurrence of L and, therefore, of ML, increases the total concentration (M + ML) and mobility of the nutrient element. The ⇆ denotes reversible processes which are spontaneously approaching equilibrium. Depending on ion species, ion concentration at the root surface, and plant age, the symbol ⇆ denotes processes that may be irreversible. The irreversible processes are always rate limiting and/or rate determining, whereas reversible processes may be rate limiting, only. Processes 2 and 3 are in the vicinity of the soil particle, whereas process 7 is in the vicinity of the root. Processes 4, 5, and 6 are transport processes by mass-flow and diffusion due to water uptake and nutrient uptake (process 8) by cell membranes of root cells near the root surface or root hairs. Process 9 is the nutrient translocation in the plant. Process 10 is the plant growth that also integrates the absorbed nutrient into the plant tissue. Processes 8, 9, and/or 10 create the concentration (electro-chemical) gradients for irreversible net flux of nutrients from the soil-soil solution system into the plants. Hence, at any time, the rate-determining processes (Figure 1) are then either the release of the nutrient into the pool of plant available nutrient (source/sink processes in the soil), or the nutrient uptake into the roots, its translocation/circulation in the plant, and/or the rate by which the nutrient is built into new tissue. Mass-flow and diffusion may be rate limiting only, and not rate determining. Usually only a small fraction of the plant-available nutrients is dissolved in the soil solution. This implies that the bioavailability of nutrients to plant roots is governed by several soil properties including, for example, the characteristics of process 2 in Figure 1 and the possibilities for movement via soil solution to the root surface by mass-flow and diffusion. The concept of a *bioavailable nutrient* can

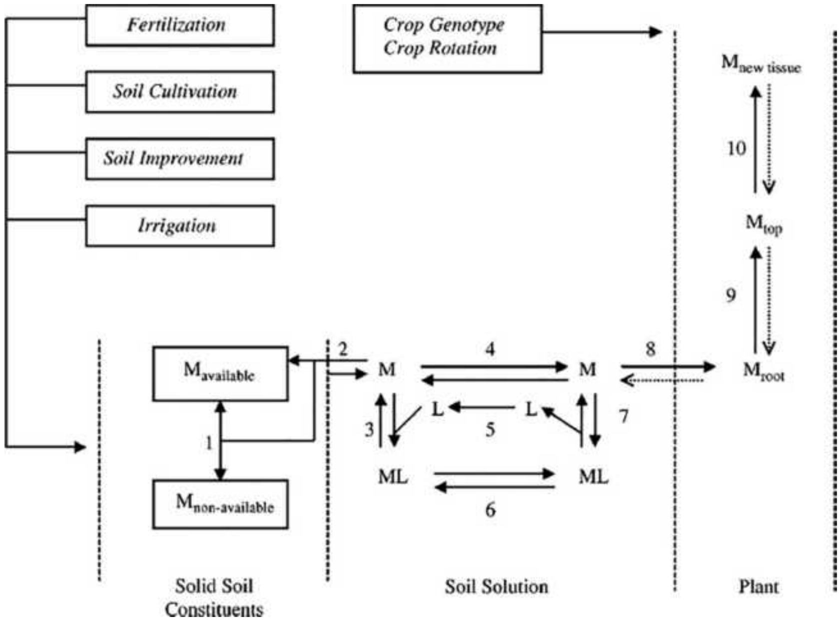


FIGURE 1 Flux of a nutrient element in the soil plant atmosphere system and agronomy actions; nutrient element (M), ligands (L), reversible and irreversible processes.
Source: Nielsen.[22]

then be defined as a nutrient element that is present in a pool of diffusible (available) nutrients which are close enough to arrive at water- and/or nutrient-absorbing root surfaces during a period of 10 days, for example. This seems to fit with the observation that most of the depletion zones of slowly moving nutrients, such as phosphorus, are created during the first 10 days after root growth into a new soil volume unit.[1] The bioavailable quantity of a nutrient in the soil is affected by at least five different groups of processes as indicated in Table 1.

TABLE 1 Processes and Factors Involved in Nutrient Transfer from Soil to Plant Roots

Process	Factors
Release, mineralization, and dissolvement of the nutrient in the soil solution	Chemical and physical properties of the solid phases, temperature, soil water content, and activity of the microbial biomass
Root development	Root length, distribution of roots in the root zone, root morphology, root hairs, rate of root growth, and root surface contact area with soil solution
Solute movement by mass flow and diffusion to roots	Transpiration rate (w_o), concentration of the nutrient in the soil solution (c_b), effective diffusion coefficient (D_e), nutrient buffer power of the soil (b)
Rhizosphere processes, increasing the rate of nutrient release in the soil	Depletion of the soil solution for nutrients by the roots; root exudates as protons, reducing agents, chelates, organic anions, enzymes; modification of microbial activity; mycorrhizal and Rhizobium symbioses
Nutrient uptake	Concentration of the nutrient at the root surface (c_o); transport kinetic parameters of nutrient uptake by the roots $\bar{I}_{max}$, Km and c_{min})

Source: Nielsen.[22]

Diffusion

Diffusion is the net movement of a solute or a gas from a region that has a higher concentration, to an adjacent-region that has a lower concentration. Diffusion is a result of the random thermal motions of molecules in the considered solids, solution, or air. The net movement caused by diffusion is a statistical phenomenon because the probability of the molecules' movement from the concentrated to the diluted region is greater than vice versa. Fick (1855) was one of the first to examine diffusion on a quantitative basis. The basic equation Eq. (1) to express diffusion, e.g., to a root is today known as Fick's first law of diffusion:

$$F = -D\frac{\partial C}{\partial r} \tag{1}$$

in which F is the diffusive flux (mol cm^{-2}s $^{-1}$ of a nutrient in the r-direction normal to the root cylinder. The driving force (the gradient in the electro-chemical potential) is, in most cases, approximated by the concentration gradient

$$\frac{\partial C}{\partial r}\left(\text{mol cm}^{-4}\right)$$

and D is the diffusion coefficient (cm^{2}s^{-1}). They way to describe diffusion processes mathematically under various conditions has been presented by Jost[2] and Crank.[3] Great contributions to our understanding of solute movement in the soil root system by mass-flow and diffusion have been given or reviewed by Nye and Tinker[4,6] and Barber.[5] Recently, Willigen and colleagues[7] reviewed some aspects of the modeling of nutrient and water uptake by plant roots.

Nutrient Movement by Mass Flow and Diffusion from Soil to Plant Roots

Nutrients bound to the solid-soil phase are virtually immobile in the sense of its movements to roots. The nutrient has to be released into the soil solution as indicated in Figure 1. Furthermore, contact between the root and nutrient-absorbing membranes in the root tissue near the root surface and the soil solution is a prerequisite for nutrient uptake. Contact to nutrient pools can be brought about by two means: 1) by growth of roots to the sites where nutrient pool are located (root interception); and 2) by movement of the nutrient from the bulk (the pool) of the soil to the root surface. Even so, nutrients may at any time move over a certain distance in the soil solution and cell wall before they reach the outside cell membrane of a root hair or a root cortical cell for uptake. The mechanisms for these transports are mass flow and diffusion.[4,8] The driving force for the net movement of nutrients (Figure 1) is the water and the selective nutrient uptake by the plant root, creating a concentration gradient (dc/dr). The general equation of continuity (mass-balance) used to describe movements in a direction normal to a root cylinder at radial distance r and time t, may partly be developed from Eq. (1), extended and expressed as

$$\left[b\frac{\partial c}{\partial r}\right]_r = -\left[\frac{1}{r}\frac{\partial rF_T}{\partial r}\right]_t + U_{r,t} \tag{2}$$

in which

- U is the production/consumption term (mole cm^{-3}s^{-1}) at r (radial distance from the center of the root) and t (time).
- b is the buffer power *(dC/dc)* in which C is the total concentration of diffusible solute in the soil and c is the concentration of solute in the soil solution.
- F_T is total net flux of solute by mass flow and diffusion (mole cm^{-2}s^{-1}).

$$F_T = F_m + F_d$$

where $F_m = wc$ in which w is the flux of soil solution in the direction of the root ($cm^3cm^{-2}s^{-1}$) and c is the nutrient con centration of soil solution (mole cm^{-3}). The expected rates of water flux at the root surface are 0.2–1 10^{-6}cm s^{-1}.[5] The flux, F_d, by diffusion can be expressed by Flick's first law

$$F_d = -D_c b \frac{dc}{dr}$$

The $b = dC/dc$ is the soil buffer power defined previously. The C is the sum of the amount of nutrient in the soil solution and the amount of adsorbed nutrient that is able to replenish the nutrient in the soil solution spontaneously. Hence, b is the parameter mediating the effects of the soil chemical conditions on nutrient uptake by plants. D_e denotes the effective diffusion coefficient in the soil. D_e differs between media, but it can be related to the diffusion coefficient D_o for the nutrient in free soil solution. The influences of soil on diffusion, and thereby the relation between D_e and D_o, can be expressed by Eq. (3).[9]

$$D_e = D_o \theta f / b \tag{3}$$

where θ is the volumetric water content expressed as a fraction, and f is the impedance factor that essentially allows for the increase in the actual diffusion distance because of the tortuous pathway of water filled soil pores and water films. The volumetric water content that allows a reasonable root activity is between 0.1 and 0.4. The value of f increases with increase in water content,[10] whereas the buffer power remains constant with changes in soil moisture at the same bulk density.[11] It has been observed[12] that the relation between f and θcan be expressed empirically by $f = 1.58\theta - 0.17$ for $\theta >$ about 0.11. From this it may be estimated that D_e decreases about 18 times if θ decreases from 0.40 to 0.15. Hence D_e is the parameter mediating the effects of soil moisture, soil chemical, and soil physical conditions on diffusion in soil.

Almost all studies on solute movement in the soil plant system neglect U in Eq. (2) because our understanding of the biology caused by root-induced processes and its effects on production or consumption of available nutrients is incomplete as yet.

To solve Eq. (2) for a given soil plant system is a complicated process, and in most cases, difficult or even impossible because of the lack of information on the soil root interactions and root behavior. The method for obtaining analytical and numerical solutions of Eq. (2) under a number of often simplified soil plant conditions has been summarized.[5–7] However, illustration of the importance of diffusion for the bioavailability of nutrients in soils may be based on Eq. (4)

$$\Delta r = \sqrt{2D_e t} \tag{4}$$

in which Dr is the average distance of diffusion; e.g., in a direction normal to a root. The mathematics behind Eq. (4) has been presented by Jost.[2] Based on Eq. (4), the equivalent soil volume (V in cm^3) of soil depleted for diffusible (available) nutrients—*the quantity of bioavailable nutrients*—can then be estimated as follows [Eq. (5)] for roots without root hairs:

$$V = \pi\left(\Delta r + r_o\right)^2 L_v \tag{5}$$

in which Δr is estimated from Eq. (4), r_o is the root radius and L_v is the root density in cm cm^{-3} of soil. The data in Table 2 show the expected effective diffusion coefficient of a number of plant nutrients in soil and corresponding influences on the nutrient bioavailability; in addition, the data show how a decrease of the soil moisture from $\theta = 0.40$ to $\theta = 0.15$ affects the bioavailability at a root density of 5 cm^{-2} of roots without root hairs. It may be calculated from $f = 1.58\,\theta - 0.17$ and Eq. (3) that the diffusive flux decreases by a factor of $18 = D_e^{\theta=040}/D_e^{\theta=015}$. At field capacity of water content, the expected, effective diffusion coefficient of nitrate in soil is 10^{-6} cm^2 s^{-1}. This is almost 10 times slower than in pure water. Hence, a pored media, such as soil, physically decreases the possibility for solute movement with a factor of nearly 10. As the soil dries out, this factor increases as illustrated in Figure 2. Apart from nitrate and chlorine, nutrient elements are adsorbed more or less to the solid soil constituents. This is the main cause of the decrease of the diffusion coefficients below 10^{-6}cm^{2}s^{-1}. The diffusion coefficient of phosphorus is as low as 10^{-9}cm^{2}s^{-1} mainly because approximately 0.1% of the diffusible (available) phosphorus is dissolved in the soil solution only. This has a large effect on the bioavailability of the plant-available quantities of the various nutrient elements in soil as illustrated in columns 3 and 4 of Table 2. The V-values ≥ 1 indicate that the root at a density of 5 cm^{-2} is able to deplete all the available nutrient as seen for nitrate, even under dry conditions, whereas only 4% of the available phosphorus is bioavailable inside a period of 10 days. If the soil dries out to 1 indicate that the root at a density of 5 cm^{-2} is able to deplete all the available nutrient as seen $\theta = 0.15$, the bioavailability decreases to only 0.6%. This illustrates that the decrease of soil moisture may create nutrient deficiency even in soil with high phosphorus fertility. However, phosphorus uptake is increased by the activity of root hairs (discussed in the following).

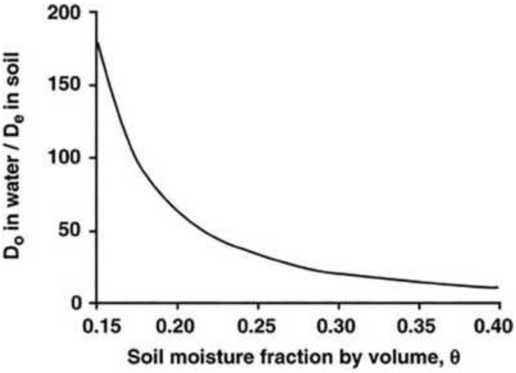


FIGURE 2 Effects of soil and soil moisture on nutrient element mobility.

TABLE 2 Expected Effective Diffusion Coefficients of Some Nutrients in Soil at Field Capacity of Water Content (e.g., $\theta = 0.40$), and Estimated Bioavailability as a Fraction of Diffusible (Available) Nutrient [Eqs. (4) and (5)] at a Root Density ($L_v = 5$ cm^{-2}) of Roots without Root Hairs (Mean Root Radius, $r = 0.01$ cm; Time, $t = 10$ days)

		Bioavailable Nutrient as a Fraction of Available Nutrient (V, cm^3)	
Element	Dea (cm^2 s^{-1})	$\theta = 0.40$	$\theta = 0.15$
Nitrate	l*10–6	27.558	1.574
Potassium	l*10–7	2.847	0.180
Boron	l*10–7	2.847	0.180
Magnesium	l*10–8	0.314	0.026
Calcium	l*10–8	0.314	0.026
Phosphorus	l*10–9	0.042	0.006
Manganese	l*10–9	0.042	0.006
Molybdenum	l*10–9	0.042	0.006
Zinc	l*10–9	0.042	0.006
Iron	l*10–10	0.008	0.003

[a] Values obtained from Barber[5] and Nielsen.[22]

Importance of Root Hairs

Root hairs are outgrowths from specialized root epidermal cells (trichoblasts). Root hair length, diameter, and number per unit length of root, vary among plant species and among genotypes within the same species.[5,6,13] Frequency and size of root hairs are affected by many environmental factors, as well. In nature, the length of root hairs vary from 0.01–0.15 cm, the radius varies from 0.0005–0.002 cm, and the number per unit of length varies from 100–1000 per cm root. The importance of root hairs for phosphorus uptake has been demonstrated directly in the laboratory[14,15] and under field conditions.[13] It is reasonable to assume that the clusters of root hairs' outer tips form a fairly well-defined cylinder to which phosphorus diffuses, on average, a distance Δr in 10 days, and that root hair density and its period of function are long enough to withdraw the entire available nutrient in the soil penetrated by root hairs. The bioavailability of phosphorus, for example, as affected by root hair length, can then be estimated from the following extension of Eq. (10):

$$V = \pi\left(\Delta r + \sigma + r_o\right)^2 L_v \tag{6}$$

in which σ is the root hair length in cm.

Figure 3 illustrates that the bioavailability of phosphorus increases exponentially with the root hair length. Hence, root hairs play a very important role for the bioavailability of nutrients having a low effective diffusion coefficient (D_e) in soil.

Boundary Conditions and Nutrient Entry

If depletion zones around the roots do not overlap, the solute concentration converges to the solute concentration c_b in the bulk solution, at which $F_T = F_m + F_d = 0$. The boundary conditions at the surface of the root are

$$F_T = F_m + F_d = \alpha c_o \tag{7}$$

in which α (cm s^{-1}) is the root-absorbing power defined by Tinker and Nye[6] and c_o is the concentration of solute at the root surface. It can be learned from Eq. (7) that the actual concentration (co) at the root surface and, therefore, the rate of nutrient flow per unit length of root, is determined by the ratio F_T/α. The kinetics of net uptake of nutrients[5,16–18] may be expressed by:

$$F_T = F_m + F_d = \alpha c_o \tag{8}$$

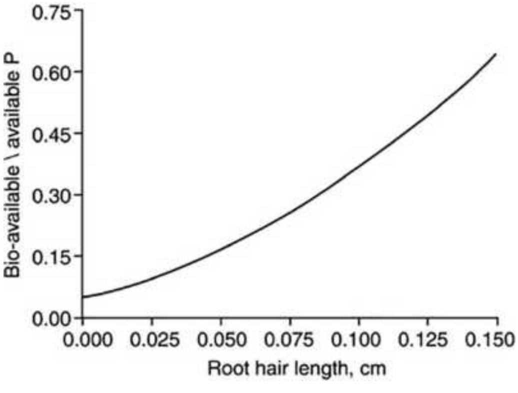


FIGURE 3 The effect of root hair length on phosphorus bioavailability. Soil moisture ($\theta = 0.4$); root density 5 cm^{-2}; root radius 0.01 cm

in which L^* is the root length per unit of plant biomass, $\bar{I}_{\max}$ (mole $\mathrm{cm}^{-1}\mathrm{s}^{-1}$) is the mean maximal net influx, *Km* (mole cm^{-3}) is the Michaelis–Menten factor, c_o is the concentration of the nutrient at the root surface, and $c_{\min}$ is the nutrient concentration at which The values of the parameters $\bar{I}_{\max}$, *K*m and $c_{\min}$ vary according to the plant nutrient, temperature, and plant species/genotype and plant age. Furthermore, kinetics of nutrient uptake by roots may be influenced by ion interactions. Determined values of L^*, $\bar{I}_{\max}$, *K*m and $c_{\min}$ for uptake of several nutrients by several plant species or genotypes, obtained under conditions in which the rate-determining step of nutrient uptake was located in the roots, has been noted.[19] The data show that the values of L^*, $\bar{I}_{\max}$, Km and $c_{\min}$ vary considerably among nutrients and among plant species and genotypes. This illustrates the efficiency by which these plants utilize soil as a source of nutrients. It is possible from $F_T = F_m + F_d = \alpha c_o$ and Eq. (8) to develop how α varies at varying solute concentration at the root surface by:

$$\alpha = \frac{\bar{I}_{\max}\left(c_o - c_{\min}\right)}{2\pi r_o c_o\left(\mathrm{km} + c_o - c_{\min}\right)} \tag{9}$$

Figures 4a and 4b illustrate the variation of α for phosphorus uptake at low concentration (c_o) at the root surface of some plant species and barley genotypes. Hence, the rootabsorbing power (α) varies

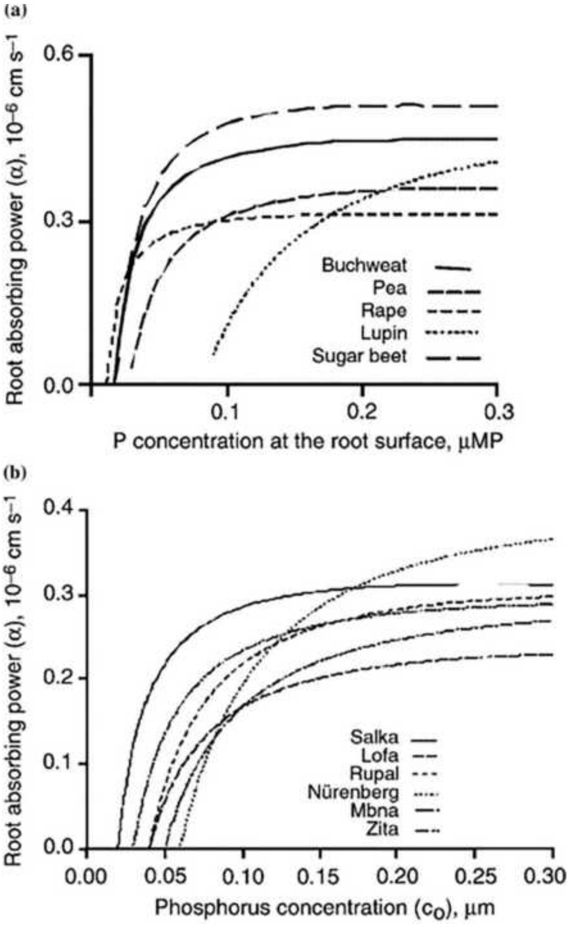


FIGURE 4 Absorption power (∝) at varying phosphorus concentration at the root surface of some plant species (a) and barley genotypes (b).
Source: Nielsen[19] and Nielsen.[22]

also at low concentration (c_o) of solute at the root surface. This implies that phosphorus uptake at low P concentration is more under the control of the plant parameters determining the size of α than under the control of P diffusion in the soil, whereas at the range of c_o at which α has achieved its maximum, uptake is controlled by diffusion.

Conclusions

Even though the movement—and the main factors affecting the movement—of nutrient elements to root by mass-flow and diffusion is well known, the effect of soil conditions on crop growth is still not properly understood. It is obvious that the big variation (Table 2) of the effective diffusion coefficient (D_e), caused mainly by the variation of the soil chemistry of the various nutrient elements, has a large impact on the bioavailability of nutrient elements. The mobility of phosphorus and micronutrients is so low in most soils that the soil exploited by root hairs is the main source of these elements. The root-induced modifications to the soil in the rhizosphere would then have a considerable impact on the efficiency by which plants use the rhizosphere soil as a source of nutrients. The understanding of how root-induced processes accelerate solute movement and the transformation of non-available nutrients to bioavailable nutrients is increasing.[6,20,21] Root hair length and root-induced processes appear to vary between genotypes of our crop plants.[13] Hence, improvement of the efficiency by which plants use soil as a source of nutrients seems to be a possibility by targeted plant breeding.

References

1. Gahoonia, T.S.; Raza, S.; Nielsen, N.E. Phosphorus depletion in the rhizosphere as influenced by soil moisture. Plant and Soil **1994**, *159,* 213–218.
2. Jost, W. *Diffusion in Solids, Liquids and Gasses*; Academic Press: New York, 1960.
3. Crank, J. *The Mathematics of Diffusion,* 2nd Ed.; Clarendon Press: Oxford, 1975.
4. Nye, P.H.; Tinker, P.B. *Solute Movement in the Soil-Root System*; Blackwell: Oxford, 1977.
5. Barber, S.A. *Soil Nutrient Bioavailability,* 2nd Ed.; Wiley: New York, 1995.
6. Tinker, P.B.; Nye, P.H. *Solute Movement in the Rhizosphere;* Oxford University Press: Oxford, 2000.
7. Willigen, P.; Nielsen, N.E.; Claassen, N.; Castrignanò, A.M. Modelling water and nutrient uptake. In *Root Methods a Handbook;* Smit, A.L., Bengough, A.G., Engels, C., Noordwijk, M., Pellerin, S., Geijn, S.C., Eds.; Springer: Berlin, 2000.
8. Barber, S.A. A diffusion and mass-flow concept of soil nutrient availability. Soil Science **1962**, *93,* 39–49.
9. Nye, P.H. The measurement and mechanism of ion diffusion in soil. I. the relation between self-diffusion and bulk diffusion. J. Soil Sci. **1966**, *17*, 16–23.
10. Rowell, D.L.; Martin, M.W.; Nye, P.H. The measurement and mechanisms of ion diffusion in soils. III. The effect of moisture content and soil solution concentration on the self-diffusion of ions in soils. J. Soil Sci. **1967**, *18,* 204–222.
11. Bhadoria, P.B.S.; Classen, J.; Jungk, A. Phosphate diffusion coefficient in soil as affected by bulk density and water content. Z. Pflanzenernaehr. Bodenk. **1991**, *154*, 53–57.
12. Barraclough, P.B.; Tinker, P.B. The determination of ionic diffusion coefficients in field soils. 1. Diffusion coefficient in sieved soils in relation to water content and bulk density. J. Soil Sci. **1981**, *32,* 225–236.
13. Gahoonia, T.S.; Nielsen, N.E.; Lyshede, O.B. Phosphorus acquisition of cereal cultivars in the field at three levels of P fertilization. Plant and Soil **1999**, *211,* 269–281.
14. Barley, K.P.; Rovira, A.D. The influence of root hairs on the uptake of phosphorus. Commun. Soil Sci. Plant Anal. **1970**, *1*, 287–292.

15. Gahoonia, T.S.; Nielsen, N.E. Direct evidence on the participation of phosphorus (P32) uptake from soil. Plant and Soil **1998**, *198,* 147–152.
16. Classen, N.; Barber, S.A. A method for characterizing the relation between nutrient concentration and flux into roots of intact plants. Plant Physiology **1974**, *54,* 564–568.
17. Nielsen, N.E. A transport kinetic concept for ion uptake by plants. III. Test of the concept by results from water culture and pot experiments. Plant and Soil **1976**, *45,* 659–677.
18. Nielsen, N.E.; Barber, S.A. Differences among genotypes of corn in the kinetics of P uptake. Agronomy Journal **1978**, *70* (5), 695–698.
19. Nielsen, N.E. Bioavailability, cycling and balances of nutrient in the soil plant system. In *Integrated Plant Nutrition Systems;* Fertilizer and Plant Nutrition Bulletin 12; Dudal, R., Roy, R.N., Eds.; FAO: Rome, 1995; 333–348.
20. Jungk, A.; Classen, N. Ion diffusion in the soil-root system. Adv. Agron. **1997**, *61,* 53–110.
21. Hensinger, P. How do plant roots acquire mineral nutrients? chemical processes involved in the rhizosphere. Adv. Agron. **1998**, *64*, 225–265.
22. Nielsen, N.E. Bioavailability of nutrients in soil. In *Roots and Nitrogen in Cropping Systems of the Semi-Arid Tropics;* Ito, O., Johansen, C., Adu-Gyamfi, J.J., Katayama, K., Kumar Rao, J.V.D.K., Rego, T.J., Eds.; International Crop Research Institute for the Semi-arid Tropics: India, 1996; 411–427.

51

Nutrient–Water Interactions

Ardell D. Halvorson

Introduction

Water is a major factor in nutrient availability to plants.[1–4] It is the vehicle through which nutrients move through soil to access plant roots for uptake. Nutrients move via mass flow and diffusion in soil water to the root surface. Root interception is a third way in which plants obtain soil nutrients as root hairs develop and contact the soil particles and/or solution.

Water and Nutrient Availability

In mass flow, nutrient ions are transported with water flow to the root as the plant absorbs water for transpiration. Many mobile nutrients, such as calcium (Ca), magnesium (Mg), nitrate-N (NO_3-N), and sulfate (SO_4), are transported to the root by mass flow. Diffusion of nutrients to the plant root occurs as ions move from high-concentration areas to low-concentration areas in the soil solution. Phosphorus (P) and potassium (K) are two nutrients that move by diffusion.

If soil water becomes limiting, as it frequently does under dryland or rainfed conditions, nutrient availability to plants can be affected.[5] Water is held as a film around soil particles. As the water content of the soil decreases, the thickness of the film decreases. Most plant nutrients are readily available when the soil is near field capacity, which is about the water content of the wet soil after two days of rain has saturated it and free drainage has ceased. Nutrient availability is at a minimum as the soil water content approaches the permanent wilting point, which is the water content at which plant roots cannot extract water from the soil. As soil water content diminishes, some less-soluble nutrients may precipitate out of the soil solution and become unavailable to plants. However, these minerals will dissolve and become available once again as the soil is rewetted. Thus, soil water content influences nutrient availability and plant growth.

Micronutrients are generally supplied to plant roots by diffusion in soil. Therefore, low soil moisture conditions will reduce micronutrient uptake. Plants require smaller quantities of micronutrients to optimize productivity than macronutrients such as P; thus, drought stress effects on micronutrient deficiency are not as serious as for P. However, iron (Fe) and zinc (Zn) deficiencies are frequently associated with high soil moisture conditions.[2]

Soil water content is an important factor in microbial activity in soils. Soil microbial activity is important in the breakdown of organic plant and animal residues, which release nutrients such as N and P for plant uptake. Microbial activity tends to be greatest when soil water is near field capacity with soil temperatures ranging from 25°C to 35°C. As soils dry, microbial activity decreases and lowers the rate of nutrient release from soil organic matter.[6,7]

Nutrient and Water Use Efficiency

Adequate levels of plant nutrients are needed to optimize rooting depth and water extraction from the soil.[2,3,5] Healthy plants tend to root deeper into the soil profile, using more of the soil water in the root zone. Thus, plants not only need adequate water to optimize yield potential, but also require an adequate level of nutrients to allow the crop to take advantage of the available water supplies. Under dryland conditions, the crop will often use all of the available water (precipitation plus soil water in the root zone) during the growing season. Application of N and P fertilizers will frequently increase crop yields, thus increasing crop water use efficiency (WUE). Water use efficiency is the amount of crop produced per unit of available water from precipitation, soil, and irrigation. The influence of N fertilization on WUE of winter wheat, corn, and sorghum in a dryland wheat-corn or sorghum-fallow rotation is shown in Figure 1.

When plant-available water is limited, overapplication of N can also result in reduced grain yields owing to increased vegetative growth and water use in the early growth stage, with insufficient water remaining to maximize grain development and yield. Application of N will not increase yields without adequate plant-available water, and increasing plant-available water will not increase crop yield without adequate N supply. The percentage increase in response of crops such as wheat to P fertilization tends to be greater in dry years than in wet years on P-deficient soils, while both N and P are needed to optimize yields in wetter years.

Water is important for activation and movement of fertilizer nutrients applied to soils.[1–3,7,8] Dry fertilizer granules must dissolve in the soil water before they become available to plants. When applied to dry soil, liquid fertilizers may become unavailable to plants until precipitation or irrigation water rewets

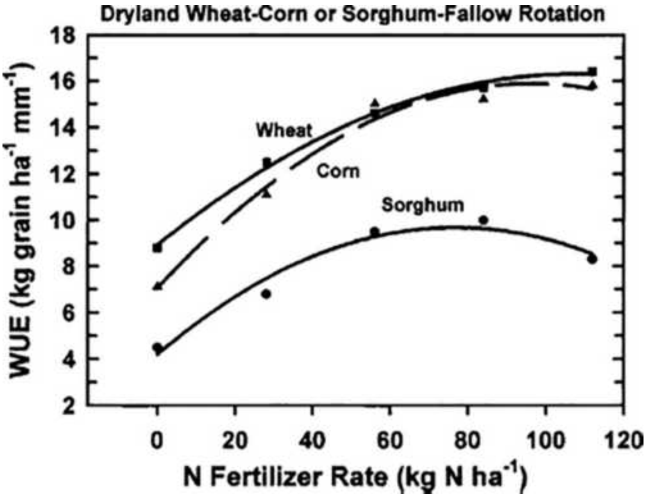


FIGURE 1 Water use efficiency of wheat, corn, and sorghum as a function of N fertilizer rate in a dryland wheat-corn or sorghum-fallow rotation near Akron, Colorado, U.S.A.

the soil and they become part of the soil solution again. Rainfall affects the volatilization loss of N from ammonia-based fertilizers such as urea and urea ammonium nitrate (UAN). Rainfall received within 36 hr after surface applications of urea or UAN fertilizers will greatly reduce N volatilization losses and improve the N fertilizer use efficiency by crops. Rainfall moves the surface-applied N fertilizer into the soil where it can react and reduce NH_3 losses to the atmosphere. Excessive soil water, however, can result in anaerobic conditions and the loss of NO_3-N by denitrification. Nitrate-N is converted to various N gases, which are lost to the atmosphere under anaerobic conditions.

Water is essential for optimizing crop yields. Under irrigation, water is generally not a yield-limiting factor. Under dryland or rainfed conditions, crop yields are dependent on available soil water supplies and growing season precipitation. Adequate levels of essential plant nutrients are needed to optimize crop yields and WUE (i.e., kg grain produced/mm crop water use). Under rainfed conditions, crop water supplies during the growing season can vary weekly and annually. During periods of drought (i.e., low supply of plant-available water), less plant nutrients are needed to optimize crop yields than during years of average or above-average precipitation. In wetter years, both the crop yield potential and the nutrients needed to optimize crop yield increase.

Soil management practices, such as reduced- and no-till systems, that increase soil organic matter and improve soil physical quality also improve soil aggregation and porosity. This, in turn, improves water infiltration into the soil and water availability for increased crop productivity and improved nutrient use efficiency.

Irrigation Water Quality and Fertilizer Application

Irrigation water quality can affect the application of fertilizer nutrients through irrigation systems.[3,8] For example, the addition of anhydrous NH_3 or liquid ammonium polyphosphate fertilizers to irrigation waters high in Ca can result in the formation of lime and calcium phosphate precipitates. The precipitates can plug sprinkler and drip irrigation systems. In some instances, precipitation of the Ca can result in a higher sodium (Na) hazard of the irrigation water, which may subsequently reduce the water intake capacity of the soil.

Applying fertilizers with both flood and furrow irrigation systems requires that a uniform distribution of water be achieved throughout the field to obtain a uniform distribution of fertilizer nutrients to the crop. With flood and furrow irrigation systems, fertilizer should not be applied with the initial flush of irrigation water because of the generally nonuniform distribution of water during the initial wetting of the soil surface by the irrigation water. The reactions of fertilizers with the irrigation water and the fertilizer distribution to the crop are affected by (irrigation) water quality. If fertigation (i.e., application of fertilizer nutrients through an irrigation system) is to be used, the compatibility of fertilizers to be applied with the quality of irrigation water available must be examined to avoid poor distribution of fertilizer nutrients.

Environmental Quality

Nitrogen is generally transported from soils into surface and groundwater by runoff, erosion, and leaching.[7,9] Runoff water from watersheds with high levels of soluble N and P sources on the soil surface can contribute to eutrophication of streams, lakes, ponds, bays, and estuaries. Placing or positioning applied N and P sources below the soil surface and using soil management practices to minimize runoff will help reduce agriculture's impact on eutrophication of water bodies. Water erosion of soil not only carries soluble plant nutrients from a watershed, but also carries soil particles with sorbed nutrients, such as P, into water bodies that can then contribute to degradation of water quality.

Soil management practices such as no-till and other conservation tillage practices can reduce soil erosion by water. Water moving through soil in excess of field capacity water content can move soluble nutrients, such as NO_3-N, below the root zone of crops and into groundwater. In summary, using cropping systems and an adequate fertility program to optimize crop WUE will help reduce loss of plant-available water and nutrients below the crop root zone.

Conclusions

Water plays a critical role in the availability of nutrients to plants. Adequate levels of both water and nutrients are needed to optimize plant growth and productivity. Fertilizer and water management practices can influence the efficient use of water and nutrients by plants and their subsequent impact on environmental quality.

References

1. Engelstad, O.P., Eds. *Fertilizer Technology and Use,* 3rd Ed.; Soil Sci. Soc. Am.: Madison, WI, 1985.
2. Havlin, J.L.; Beaton, J.D.; Tisdale, S.L.; Nelson, W.L. *Soil Fertility and Fertilizers: An Introduction to Nutrient Management,* 6th Ed.; Prentice Hall: Upper Saddle River, NJ, 1999.
3. Mortvedt, J.J.; Murphy, L.S.; Follett, R.H. *Fertilizer Technology and Application*; Meister Publishing Co.: Willoughby, OH, 1999.
4. Troeh, F.R.; Thompson, L.M. *Soils and Soil Fertility,* 5th Ed.; Oxford University Press: New York, 1993.
5. Taylor, H.M.; Jordan, W.R.; Sinclair, T.R., Eds. *Limitations to Efficient Water Use in Crop Production*; Am. Soc. Agron., Crop Sci. Soc. Am., Soil Sci. Soc. Am.: Madison, WI, 1983.
6. Follett, R.F.; Stewart, J.W.B.; Cole, C.V., Eds. *Soil Fertility and Organic Matter as Critical Components of Production Systems*; Soil Sci. Soc. Am., Inc.: Madison, WI, 1987.
7. Pierzynski, G.M.; Sims, J.T.; Vance, G.F. *Soils and Environmental Quality,* 2nd Ed.; CRC Press: Boca Raton, FL, 2000.
8. Ludwick, A.E., Bonczkowski, L.C., Bruice, C.A., Compbell, K.B., Millaway, R.M., Petrie, S.E., Phillips, I.L., Smith, J.J., Eds.; *Western Fertilizer Handbook*; 8th Ed.; California Fertilizer Association, Interstate Publishers: Danville, IL, 1995.
9. Follett, R.F., Ed. *Nitrogen Management and Ground Water Protection*; Elsevier: New York, 1989.

52

Pollution: Non-Point Source

Ravendra Naidu,
Mallavarapu
Megharaj, Peter
Dillon, Rai
Kookana, Ray
Correll, and
W.W. Wenzel

Introduction

Non-point source pollution (NPSP) has no obvious single point source discharge and is of diffuse nature (Table 1). An example of NPSP includes aerial transport and deposition of contaminants such as SO_2 from industrial emissions leading to acidification of soil and water bodies. Rain water in urban areas could also be a source of NPSP as it may concentrate organic and inorganic contaminants. Examples of such contaminants include polycyclic aromatic hydrocarbons, pesticides, polychlorinated biphenyls that could be present in urban air due to road traffic, domestic heating, industrial emissions, agricultural treatments, etc.[1–3] Other examples of NPSP include fertilizer (especially Cd, N, and P) and pesticide applications to improve crop yield. Use of industrial waste materials as soil amendments have been estimated to contaminate thousands of hectares of productive agricultural land in countries throughout the world.

Contaminant Interactions

Non-point pollution is generally associated with low-level contamination spread at broad acre level. Under these circumstances, the major reaction controlling contaminant interactions are sorption–desorption processes, plant uptake, surface runoff, and leaching. However, certain contaminants, in particular, organic compounds are also subjected to voltalization, chemical, and biological degradation. Sorption–desorption and degradation (both biotic and abiotic) are the two most important processes controlling organic contaminant behavior in soils. These processes are influenced by both soil and solution properties of the environment. Such interactions also determine the bioavailability and/or transport of contaminants in soils. Where the contaminants are bioavailable, risk to surface and groundwater and soil, crop, and human health are enhanced.

TABLE 1 Industries, Land Uses, and Associated Chemicals Contributing to Non-Point Source Pollution

Industry	Type of Chemical	Associated Chemicals
Agricultural activities	Metals/metalloid	Cadmium, mercury, arsenic, selenium
	Non-metals	Nitrate, phosphate, borate
	Salinity/sodicity	Sodium, chloride, sulfate, magnesium, alkalinity
	Pesticides	Range of organic and inorganic pesticides including arsenic, copper, zinc, lead, sulfonylureas, organochlorine, organophosphates, etc., salt, geogenic contaminants (e.g., arsenic, selenium, etc.)
	Irrigation	Sodium, chloride, arsenic, selenium
Automobile and industrial emissions	Dust	Lead, arsenic, copper, cadmium, zinc, etc.
	Gas	Sulfur oxides, carbon oxides
	Metals	Lead and lead organic compounds
Rainwater	Organics	Polyaromatic hydrocarbons, polychlorbiphenyls, etc.
	Inorganic	Sulfur oxides, carbon oxides acidity, metals and metalloids

Source: (From Barzi, F.; Naidu, R.; McLaughlin, M.J. Contaminants and the Australian Soil Environment. In *Contaminants and the Soil Environment in the Australasia-Pacific Region*; Naidu, R., Kookana, R.S., Oliver, D., Rogers, S., McLaughlin, M.J., Eds.; Kluwer Academic Publishers: Dordrecht, the Netherlands, 1996; 451–484.)

Implications to Soil and Environmental Quality

Environmental contaminants can have a deleterious effect on non-target organisms and their beneficial activities. These effects could include a decline in primary production, decreased rate of organic matter break-down, and nutrient cycling as well as mineralization of harmful substances that in turn cause a loss of productivity of the ecosystems. Certain pollutants, even though present in very small concentrations in the soil and surrounding water, have potential to be taken up by various micro-organisms, plants, animals, and ultimately human beings. These pollutants may accumulate and concentrate in the food chain by several thousand times through a process referred to as biomagnification.

Urban sewage, because of its nutrient values and source of organic carbon in soils, is now increasingly being disposed to land. The contaminants present in sewage sludge (nutrients, heavy metals, organic compounds, and pathogens), if not managed properly, could potentially affect the environment adversely. Dumping of radioactive waste (e.g., radium, uranium, plutonium) onto soil is more complicated because these materials remain active for thousands of years in the soil and thus pose a continued threat to the future health of the ecosystem.

Industrial wastes, improper agricultural techniques, municipal wastes, and use of saline water for irrigation under high evaporative conditions result in the presence of excess soluble salts (predominantly Na and Cl ions) and metalloids such as Se and As in soils. Salinity and sodicity affect the vegetation by inhibiting seed germination, decreasing permeability of roots to water, and disrupting their functions such as photosynthesis, respiration, and synthesis of proteins and enzymes.

Some of the impacts of soil pollution migrate a long way from the source and can persist for some time. For example, suspended solids can increase water turbidity in streams, affecting benthic and pelagic aquatic ecosystems, filling reservoirs with unwanted silt, and requiring water treatment systems for potable water supplies. Phosphorus attached to soil particles, which are washed from a paddock into a stream, can dominate nutrient loads in streams and down-stream water bodies. Consequences include increases in algal biomass, reduced oxygen concentrations, impaired habitat for aquatic species, and even possible production of cyanobacterial toxins, with series impacts for humans and livestock consuming the water. Where waters discharge into estuaries, N can be the limiting factor for eutrophication; estuaries of some catchments where fertilizer use is extensive have suffered from excessive sea grass and algal growth.

More insidious is the leaching of nutrients, agricultural chemicals, and hydrocarbons to groundwater. Incremental increases in concentrations in groundwater may be observed over long periods of time resulting in initially potable water becoming undrinkable and then some of the highest valued uses of the resource may be lost for decades. This problem is most severe on tropical islands with shallow relief and some deltaic arsenopyrite deposits, where wells cannot be deepened to avoid polluted groundwater because underlying groundwater is either saline or contains too much As.

Sampling for Non-Point Source Pollution

The sampling requirements of NPSP are quite different from those of the point source contamination. Typically, the sampling is required to give a good estimate of the mean level of pollution rather than to delineate areas of pollution. In such a situation, sampling is typically carried out on a regular square or a triangular grid. Furthermore, gains may be possible by using composite sampling.[4] However, if the pollution is patchy, other strategies may be used. One such strategy is to divide the area into remediation units, and to sample each of these. The possibility of movement of the pollutant from the soil to some receptor (or asset) is assessed, and the potential harm is quantified. This process requires an analysis of the bioavailability of the pollutant, pathway analysis, and the toxicological risk. The risk analysis is then assessed and decisions are then made as to how the risk should be managed.

Management and/or Remediation of Non-Point Source Pollution

The treatment strategies used for managing NPSP are generally those that modify the soil properties to decrease the bioavailable contaminant fraction. This is particularly so in the rural agricultural environment where soil-plant transfer of contaminants is of greatest concern. Soil amendments commonly used include those that change the ion-exchange characteristics of the colloid particles and those that enhance the ability of soils to sorb contaminants. An example of NPSP management includes the application of lime to immobilize metals because the solubility of most heavy metals decreases with increasing soil pH. However, this approach is not applicable to all metals, especially those that form oxyanions—the bioavailability of such species increases with increasing pH. Therefore, one of the prerequisites for remediating contaminated sites is a detailed assessment of the nature of contaminants present in the soil. The application of a modified aluminosilicate to a highly contaminated soil around a zinc smelter in Belgium was shown to reduce the bioavailability of metals thereby reducing the Zn phytotoxicity.[5] The simple addition of rock phosphates to form Pb phosphate has also been demonstrated to reduce the bioavailability of Pb in aqueous solutions and contaminated soils due to immobilization in the metal.[6] Nevertheless, there is concern over the long-term stability of the processes. The immobilization process appears attractive currently given that there are very few cheap and effective in situ remediation techniques for metal-contaminated soils. A novel, innovative approach is using higher plants to stabilize, extract, degrade, or volatilize inorganic and organic contaminants for in situ treatment (cleanup or containment) of polluted top soils.[7]

Preventing Water Pollution

The key to preventing water pollution from the soil zone is to manage the source of pollution. For example, nitrate pollution of groundwater will always occur if there is excess nitrate in the soil at a time when there is excess water leaching through the soil. This suggests that we should aim to reduce the nitrogen in the soil during wet seasons and the drainage through the soil. Local research may be needed to demonstrate the success of best management techniques in reducing nutrient, sediment, metal, and chemical exports via surface runoff and infiltration to groundwater. Production figures from the same experiments may also convince local farmers of the benefits of maintaining nutrients and chemicals where needed by a crop rather than losing them off site, and facilitate uptake of best management practices.

Global Challenges and Responsibility

The biosphere is a life-supporting system to the living organisms. Each species in this system has a role to play and thus every species is important and biological diversity is vital for ecosystem health and functioning. The detection of hazardous compounds in Antarctica, where these compounds were never used or no man has ever lived before, indicates how serious is the problem of long-range atmospheric transport and deposition of these pollutants. Clearly, pollution knows no boundaries. This ubiquitous pollution has had a global effect on our soils, which in turn has been affecting their biological health and productivity. Coupled with this, over 100,000 chemicals are being used in countries throughout the world. Recent focus has been on the endocrine disruptor chemicals that mimic natural hormones and do great harm to animal and human reproductive cycles.

These pollutants are only a few examples of contaminants that are found in the terrestrial environment.

References

1. Chan, C.H.; Bruce, G.; Harrison, B. Wet deposition of organochlorine pesticides and polychlorinated biphenyls to the great lakes. J. Great Lakes Res. **1994**, *20,* 546–560.
2. Lodovici, M.; Dolara, P.; Taiti, S.; Del Carmine, P.; Bernardi, L.; Agati, L.; Ciappellano, S. Polycyclic aromatic hydrocarbons in the leaves of the evergreen tree *Laurus Nobilis.* Sci. Total Environ. **1994**, *153,* 61–68.
3. Sweet, C.W.; Murphy, T.J.; Bannasch, J.H.; Kelsey, C.A.; Hong, J. Atmospheric deposition of PCBs into green bay. J. Great Lakes Res. **1993**, *18,* 109–128.
4. Patil, G.P.; Gore, S.D.; Johnson, G.D. *Manual on Statistical Design and Analysis with Composite Samples;* Technical Report No. 96-0501; EPA Observational Economy Series Center for Statistical Ecology and Environmental Statistics; Pennsylvania State University, 1996; Vol. 3.
5. Vangronsveld, J.; Van Assche, F.; Clijsters, H. Reclamation of a bare industrial area, contaminated by non-ferrous metals: in situ metal immobilisation and revegetation. Environ. Pollut. **1995**, *87,* 51–59.
6. Ma, Q.Y.; Logan, T.J.; Traina, S.J. Lead immobilisation from aqueous solutions and contaminated soils using phosphate rocks. Environ. Sci. Technol. **1995**, *29,* 1118–1126.
7. Wenzel, W.W.; Adriano, D.C.; Salt, D.; Smith, R. Phytoremediation: a plant-microbe based remediation system. In *Bioremediation of Contaminated Soils;* Soil Science Society of America Special Monograph No. 37, Adriano, D.C., Bollag, J.M., Frankenberger, W.T., Jr., Sims, W.R., Eds.; Soil Science Society of America: Madison, USA, 1999; 772 pp.

53

Pollution: Point Sources

Ravendra Naidu, Mallavarapu Megharaj, Peter Dillon, Rai Kookana, Ray Correll, and W.W. Wenzel

Introduction

Environmental pollution is one of the foremost ecological challenges. Pollution is an offshoot of technological advancement and overexploitation of natural resources. From the standpoint of pollution, the term environment primarily includes air, land, and water components including landscapes, rivers, parks, and oceans. Pollution can be generally defined as an undesirable change in the natural quality of the environment that may adversely affect the well being of humans, other living organisms, or entire ecosystems either directly or indirectly. Although pollution is often the result of human activities (anthropogenic), it could also be due to natural sources such as volcanic eruptions emitting noxious gases, pedogenic processes, or natural change in the climate. Where pollution is localized it is described as point source (PS). Thus, PS pollution is a source of pollution with a clearly identifiable point of discharge that can be traced back to the specific source such as leakage of underground petroleum storage tanks or an industrial site.

Some naturally occurring pollutants are termed geogenic contaminants and these include fluorine, selenium, arsenic, lead, chromium, fluoride, and radionuclides in the soil and water environment. Significant adverse impacts of geogenic contaminants (e.g., As) on environmental and human health have been recorded in Bangladesh, West Bengal, India, Vietnam, and China. More recently reported is the presence of geogenic Cd and the implications to crop quality in Norwegian soils.[1]

The terms contamination and pollution are often used interchangeably but erroneously. Contamination denotes the presence of a particular substance at a higher concentration than would occur naturally and this may or may not have harmful effects on human or the environment. Pollution refers not only to the presence of a substance at higher level than would normally occur but is also associated with some kind of adverse effect.

Nature and Sources of Contaminants

The main activities contributing to PS pollution include industrial, mining, agricultural, and commercial activities as well as transport and services (Table 1). Uncontrolled mining, manufacturing, and disposal of wastes inevitably cause environmental pollution. Military land and land for recreational shooting are

TABLE 1 Industries, Land Uses, and Associated Chemicals Contributing to Points, Non-Point Source Pollution

Industry	Type of Chemical	Associated Chemicals
Airports	Hydrocarbons	Aviation fuels
	Metals	Particularly aluminum, magnesium, and chromium
Asbestos production and disposal	Asbestos	
Battery manufacture and recycling	Metals	Lead, manganese, zinc, cadmium, nickel, cobalt, mercury, silver, and antimony
	Acids	Sulfuric acid
Breweries/distilleries	Alcohol	Ethanol, methanol, and esters
Chemicals manufacture and use	Acid/alkali	Mercury (chlor/alkali), sulfuric, hydrochloric and nitric acids, sodium and calcium hydroxides
	Adhesives/resins	Polyvinyl acetate, phenols, formaldehyde, acrylates, and phthalates
	Dyes	Chromium, titanium, cobalt, sulfur and nitrogen organic compounds, sulfates, and solvents
	Explosives	Acetone, nitric acid, ammonium nitrate, pentachlorophenol, ammonia, sulfuric acid, nitroglycerine, calcium cyanamide, lead, ethylene glycol, methanol, copper, aluminum, bis(2-ethylhexyl) adipate, dibutyl phthalate, sodium hydroxide, mercury, and silver
	Fertilizer	Calcium phosphate, calcium sulfate, nitrates, ammonium sulfate, carbonates, potassium, copper, magnesium, molybdenum, boron, and cadmium
	Flocculants	Aluminum
	Foam production	Urethane, formaldehyde, and styrene
	Fungicides	Carbamates, copper sulfate, copper chloride, sulfur, and chromium
	Herbicides	Ammonium thiocyanate, carbanates, organochlorines, organophosphates, arsenic, and mercury
	Paints Heavy metals General Solvent	Arsenic, barium, cadmium, chromium, cobalt, lead, manganese, mercury, selenium, and zinc Titanium dioxide Toluene, oils natural (e.g., pine oil) or synthetic
	Pesticides Active ingredients Solvents	Arsenic, lead, organochlorines, and organophosphates Sodium, tetraborate, carbamates, sulfur, and synthetic pyrethroids Xylene, kerosene, methyl isobutyl ketone, amyl acetate, and chlorinated solvents
	Pharmacy General/solvents	Dextrose and starch Acetone, cyclohexane, methylene chloride, ethyl acetate, butyl acetate, methanol, ethanol, isopropanol, butanol, pyridine methyl ethyl ketone, methyl isobutyl ketone, and tetrahydrofuran
	Photography	Hydroquinone, pheidom, sodium carbonate, sodium sulfite, potassium bromide, monomethyl paraaminophenol sulfates, ferricyanide, chromium, silver, thiocyanate, ammonium compounds, sulfur compounds, phosphate, phenylene diamine, ethyl alcohol, thiosulfates, and formaldehyde
	Plastics	Sulfates, carbonates, cadmium, solvents, acrylates, phthalates, and styrene
	Rubber	Carbon black

(*Continued*)

TABLE 1 (*Continued*) Industries, Land Uses, and Associated Chemicals Contributing to Points, Non-Point Source Pollution

Industry	Type of Chemical	Associated Chemicals
	Soap/detergent General	Potassium compounds, phosphates, ammonia, alcohols, esters, sodium hydroxide, surfactants (sodium lauryl sulfate), and silicate compounds
	Acids	Sulfuric acid and stearic acid
	Oils	Palm, coconut, pine, and tea tree
	Solvents	
	General	Ammonia
	Hydrocarbons	e.g., BTEX (benzene, toluene, ethylbenzene, xylene)
	Chlorinated organics	e.g., trichloroethane, carbon tetrachloride, and methylene chloride
Defense works		See *Explosives* under *Chemicals Manufacture and* Use, Foundries, Engine Works, and Service Stations
Drum reconditioning		See *Chemicals Manufacture and Use*
Dry cleaning		Trichlorethylene and ethane Carbon tetrachloride Perchlorethylene
Electrical		PCBs (transformers and capacitors), solvents, tin, lead, and copper
Engine works	Hydrocarbons Metals Solvents Acids/alkalis Refrigerants	
	Antifreeze	Ethylene glycol, nitrates, phosphates, and silicates
Foundries	Metals	Particularly aluminum, manganese, iron, copper, nickel, chromium, zinc, cadmium and lead and oxides, chlorides, fluorides and sulfates of these metals
	Acids	Phenolics and amines coke/graphite dust
Gas works	Inorganics	Ammonia, cyanide, nitrate, sulfide, and thiocyanate
	Metals	Aluminum, antimony, arsenic, barium, cadmium, chromium, copper, iron, lead, manganese, mercury, nickel, selenium, silver, vanadium, and zinc
	Semivolatiles	Benzene, ethylbenzene, toluene, total xylenes, coal tar, phenolics, and PAHs
Iron and steel works		Metals and oxides of iron, nickel, copper, chromium, magnesium and manganese, and graphite
Landfill sites Marinas		Methane, hydrogen sulfides, heavy metals, and complex acids Engine works, electroplating under metal treatment
	Antifouling paints	Copper, tributyltin (TBT)
Metal treatments	Electroplating metals	Nickel, chromium, zinc, aluminum, copper, lead, cadmium, and tin
	Acids	Sulfuric, hydrochloric, nitric, and phosphoric
	General	Sodium hydroxide, 1,1,1-trichloroethane, tetrachloroethylene, toluene, ethylene glycol, and cyanide compounds
	Liquid carburizing baths	Sodium, cyanide, barium, chloride, potassium chloride, sodium chloride, sodium carbonate, and sodium cyanate
	Mining and extracting industries	Arsenic, mercury, and cyanides and also refer to *Explosives* under *Chemicals Manufacture and Use*

(*Continued*)

TABLE 1 (*Continued*) Industries, Land Uses, and Associated Chemicals Contributing to Points, Non-Point Source Pollution

Industry	Type of Chemical	Associated Chemicals
	Power stations	Asbestos, PCBs, fly ash, and metals
	Printing shops	Acids, alkalis, solvents, chromium see *Photography* under *Chemicals Manufacture and Use*
Scrap yards	Service stations and fuel storage facilities	Hydrocarbons, metals, and solvents Aliphatic hydrocarbons
		BTEX (i.e., benzene, toluene, ethylbenzene, xylene) PAHs (e.g., benzo(a) pyrene) Phenols Lead
Sheep and cattle dips		Arsenic, organochlorines and organophosphates, carbamates, and synthetic pyrethroids
Smelting and refining		Metals and the fluorides, chlorides and oxides of copper, tin, silver, gold, selenium, lead, and aluminum
Tanning and associated trades	Metals	Chromium, manganese, and aluminum
	General	Ammonium sulfate, ammonia, ammonium nitrate, phenolics (creosote), formaldehyde, and tannic acid
Wood preservation	Metals	Chromium, copper, and arsenic
	General	Naphthalene, ammonia, pentachlorophenol, dibenzofuran, anthracene, biphenyl, ammonium sulfate, quinoline, boron, creosote, and organochlorine pesticides

Source: Barzi et al.[11]

also important sites of PS contamination. The contaminants associated with such activities are listed in Table 1. Contamination at many of these sites appears to have resulted because of lax regulatory measures prior to the establishment of legislation protecting the environment.

Contaminant Interactions in Soil and Water

Inorganic Chemicals

Inorganic contaminant interactions with colloid particulates include: adsorption–desorption at surface sites, precipitation, exchange with clay minerals, binding by organically coated particulate matter or organic colloidal material, or adsorption of contaminant ligand complexes. Depending on the nature of contaminants, these interactions are controlled by solution pH and ionic strength of soil solution, nature of the species, dominant cation, and inorganic and organic ligands present in the soil solution.[2]

Organic Chemicals

The fate and behavior of organic compounds depend on a variety of processes including sorption–desorption, volatilization, chemical and biological degradation, plant uptake, surface runoff, and leaching. Sorption–desorption and degradation (both biotic and abiotic) are perhaps the two most important processes as the bulk of the chemicals is either sorbed by organic and inorganic soil constituents, and chemically or microbially transformed/degraded. The degradation is not always a detoxification process. This is because in some cases the transformation or degradation process leads to intermediate products that are more mobile, more persistent, or more toxic to non-target organisms. The relative importance of these processes is determined by the chemical nature of the compound.

Implications to Soil and Environmental Quality

Considerable amount of literature is available on the effects of contaminants on soil microorganisms and their functions in soil. The negative impacts of contaminants on microbial processes are important from the ecosystem point of view and any such effects could potentially result in a major ecological perturbance. Hence, it is most relevant to examine the effects of contaminants on microbial processes in combination with communities. The most commonly used indicators of metal effects on microflora in soil are: (1) soil respiration, (2) soil nitrification, (3) soil microbial biomass, and (4) soil enzymes.

Contaminants can reach the food chain by way of water, soil, plants, and animals. In addition to the food chain transfer, pollutants may also enter via direct consumption or dust inhalation of soil by children or animals. Accumulation of these pollutants can take place in certain target tissues of the organism depending on the solubility and nature of the compound. For example, DDT and PCBs accumulate in human adipose tissue. Consequently, several of these pollutants have the potential to cause serious abnormalities including cancer and reproductive impairments in animal and human systems.

Sampling for PS Pollution

The aims of the sampling system must be clearly defined before it can be optimized.[3] The type of decision may be to determine land use, how much of an area is to be remediated, or what type of remediation process is required. Because sampling and the associated chemical and statistical analyses are expensive, careful planning of the sampling scheme is therefore a good investment. One of the best ways to achieve this is to use any ancillary data that are available. These data could be in the form of emission history from a stack, old photographs that give details of previous land uses, or agricultural records. Such data can at least give qualitative information.

As discussed before, PS pollution will typically be airborne from a stack, or waterborne from some effluent such as tannery waste, cattle dips, or mine waste. In many cases, the industry will have modified its emissions (e.g., cleaner production) or point of release (increased stack height), hence the current pattern of emission may not be closely related to the historic pattern of pollution. For example, liquid effluent may have been discharged previously into a bay, but that effluent may now be treated and perhaps discharged at some other point. Typically, the aim of a sampling scheme in these situations is to assess the maximum concentrations, the extent of the pollution, and the rate of decline in concentration from the PS. Often the sampling scheme will be used to produce maps of concentration isopleths of the pollutant.

The location of the sampling points would normally be concentrated towards the source of the pollution. A good scheme is to have sufficient samples to accurately assess the maximum pollution, and then space additional samples at increasing intervals. In most cases, the distribution of the pollutant will be asymmetric, with the maximum spread down the slope or down the prevailing wind. In such cases more samples should be placed in the direction of the expected gradient. This is a clear case of when ancillary data can be used effectively. A graph of concentration of the pollutant against the reciprocal of distance from the source is often informative.[4] Sampling depths will depend on both the nature of the pollution and the reason for the investigation. If the pollution is from dust and it is unlikely to be leached, only surface sampling will be required. An example of this is pollution from silver smelting in Wales.[5] In contrast, contamination from organic or mobile inorganic pollutants such as F compounds may migrate well down to the profile and deep sampling may be required.[6,7]

Assessment

In order to assess the impacts of pollution, reliable and effective monitoring techniques are important. Pollution can be assessed and monitored by chemical analyses, toxicity tests, and field surveys. Comparison of contaminant data with an uncontaminated reference site and available databases for baseline concentrations can be useful in establishing the extent of contamination. However, this may not always be possible in the field. Chemical analyses must be used in conjunction with biological assays to reveal site contamination and associated adverse effects. Toxicological assays can also reveal information about synergistic interactions of two or more contaminants present as mixtures in soil, which cannot be measured by chemical assays alone.

Microorganisms serve as rapid detectors of environmental pollution and are thus of importance as pollution indicators. The presence of pollutants can induce alteration of microbial communities and reduction of species diversity, inhibition of certain microbial processes (organic matter breakdown, mineralization of carbon and nitrogen, enzymatic activities, etc.). A measure of the functional diversity of the bacterial flora can be assessed using ecoplates (see http://www.biolog.com/section_4.html). It has been shown that algae are especially sensitive to various organic and inorganic pollutants and thus may serve as a good indicator of pollution.[8] A variety of toxicity tests involving microorganisms, invertebrates, vertebrates, and plants may be used with soil or water samples.[9]

Management and/or Remediation of PS Pollution

The major objective of any remediation process is to (1) reduce the actual or potential environmental threat; and (2) reduce unacceptable risks to man, animals, and the environment to acceptable levels.[10] Therefore, strategies to either manage and/or remediate contaminated sites have been developed largely from application of stringent regulatory measures set up to safeguard ecosystem function as well as to minimize the potential adverse effects of toxic substances on animal and human health.

The available remediation technologies may be grouped into two categories: (1) ex situ techniques that require removal of the contaminated soil or groundwater for treatment either on-site or off-site; and (2) in situ techniques that attempt to remediate without excavation of contaminated soils. Generally, in situ techniques are favored over ex situ techniques because of (1) reduced costs due to elimination or minimization of excavation, transportation to disposal sites, and sometimes treatment itself; (2) reduced health impacts on the public or the workers; and, (3) the potential for remediation of inaccessible sites, e.g., those located at greater depths or under buildings. Although in situ techniques have been successful with organic contaminated sites, the success of in situ strategies with metal contaminants has been limited. Given that organic and inorganic contaminants often occur as a mixture, a combination of more than one strategy is often required to either successfully remediate or manage metal contaminated soils.

Global Challenges and Responsibility

The last 100 years has seen massive industrialization. Indeed such developments were coupled with the rapid increase in world population and the desire to enhance economy and food productivity. While industrialization has led to increased economic activity and much benefit to human race, the lack of regulatory measures and appropriate waste management strategies until early 1980s (including the use of agrochemicals) has resulted in contamination of our biosphere. Continued pollution of the environment through industrial emissions is of global concern. There is, therefore, a need for politicians, regulatory organizations, and scientists to work together to minimize environmental contamination and to remediate contaminated sites. The responsibility to check this pollution lies with every individual and country although the majority of this pollution is due to the industrialized nations. There is a clear need of better coordination of efforts in dealing with numerous forms of PS pollution problems that are being faced globally.

References

1. Mehlum, H.K.; Arnesen, A.K.M.; Singh, B.R. Extractability and plant uptake of heavy metals in alum shale soils. Commun. Soil Sci. Plant Anal. **1998**, *29*, 183–198.
2. McBride, M.B. Reactions controlling heavy metal solubility in soils. Adv. Soil Sci. **1989**, *10*, 1–56.
3. Patil, G.P.; Gore, S.D.; Johnson, G.D. *EPA Observational Economy Series Volume 3: Manual on Statistical Design and Analysis with Composite Samples; Technical Report* No. 96–0501; Center for Statistical Ecology and Environmental Statistics: Pennsylvania State University, 1996.
4. Ward, T.J.; Correll, R.L. Estimating background concentrations of heavy metals in the marine environment, Proceedings of a Bioaccumulation Workshop: Assessment of the Distribution, Impacts and Bioaccumulation of Contaminants in Aquatic Environments, Sydney, 1990; Miskiewicz, A.G., Ed.; Water Board and Australian Marine Science Association: Sydney, 1992, 133–139.
5. Jones, K.C.; Davies, B.E.; Peterson, P.J. Silver in welsh soils: physical and chemical distribution studies. Geoderma **1986**, *37*, 157–174.
6. Barber, C.; Bates, L.; Barron, R.; Allison, H. Assessment of the relative vulnerability of groundwater to pollution: a review and background paper for the conference workshop on vulnerability assessment. J. Aust. Geol. Geophys. **1993**, *14* (2–3), 147–154.
7. Wenzel, W.W.; Blum, W.E.H. Effects of fluorine deposition on the chemistry of acid luvisols. Int. J. Environ. Anal. Chem. **1992**, *46*, 223–231.
8. Megharaj, M.; Singleton, I.; McClure, N.C. Effect of pentachlorophenol pollution towards microalgae and microbial activities in soil from a former timber processing facility. Bull. Environ. Contam. Toxicol. **1998**, *61*, 108–115.
9. Juhasz, A.L.; Megharaj, M.; Naidu, R. Bioavailability: the major challenge (constraint) to bioremediation of organically contaminated soils. In *Remediation Engineering of Contaminated Soils*; Wise, D., Trantolo, D.J., Cichon, E.J., Inyang, H.I., Stottmeister, U., Eds.; Marcel Dekker: New York, 2000; 217–241.
10. Wood, P.A. Remediation methods for contaminated sites. In *Contaminated Land and Its Reclamation*; Hester, R.E., Harrison, R.M., Eds.; Royal Society of Chemistry, Thomas Graham House: Cambridge, U.K., 1997; 47–73.
11. Barzi, F.; Naidu, R.; McLaughlin, M.J. Contaminants and the Australian soil environment. In *Contaminants and the Soil Environment in the Australasia–Pacific Region*; Naidu, R., Kookana, R.S., Oliver, D., Rogers, S., McLaughlin, M.J., Eds.; Kluwer Academic Publishers: Dordrecht, the Netherlands, 1996; 451–484.

54

Radioactivity

Bogdan Skwarzec

Introduction

Radioactivity was discovered at the end of the 19th century by Henri Becquerel, Marie Curie (Polish native name, Maria Sklodowska-Curie), and Pierre Curie. Henri Becquerel found that uranium salts caused fogging of an unexposed photographic plate,[1] and Marie Curie discovered that only certain elements gave off these rays of energy.[2] She named this behavior "radioactivity" (natural radioactivity). A systematic search for the total radioactivity in uranium ores also guided Marie Curie to isolate a new element, polonium, and to separate a new element, radium.[3–7] The two elements have chemical similarity that would otherwise have made them difficult to distinguish from each other. In 1934, Marie Curie's daughter, Irene Joliot-Curie and her husband Frederic Jean Joliot were the first creators of artificial radioactivity. They bombarded boron with alpha particles to make the neutron-poor nitrogen isotope ^{13}N; this isotope emitted positrons. In addition, they bombarded aluminum and magnesium with neutrons to make new radioisotopes.[8]

Radioactive Decay

Radioactive decay is the process by which an unstable atomic nucleus spontaneously loses energy by emitting ionizing particles and radiation. The three main types of radiation were discovered by Ernest Rutherford, the alpha (α), beta (β), and gamma (γ) rays (alpha, beta, and gamma radiation).[9–11] With Ernest Rutherford, he saw that radioactive substances are transformed from one element to another. About 10 years later, he unraveled the rules for the elemental transformations that accompanied radioactive decay, first for α decay and later for β decay. Emission of an a particle changes the emitting atom to an atom of the element two places to the left in the periodic table; emission of a β^- particle changes the emitting atom to an atom of the element one place to the right. These rules taken together are known as the Displacement Law; Kazimierz Fajans published it slightly earlier than did Soddy in 1913.[12] At about the same time, Soddy came to the conclusion that several substances with different radioactive properties and different atomic weights were chemically the same element. He named such substances isotopes.[13] Now, the radioactive principles are named the Soddy–Fajans periodic method.

$$\alpha \text{ decay}: {}^{A}_{Z}X \rightarrow {}^{A-4}_{Z-2}Y + {}^{4}_{2}He$$

$$\beta^{-} \text{ decay}: {}^{A}_{Z}\mathrm{X} \rightarrow {}_{Z+1}^{A}\mathrm{Y} + {}_{-1}^{0}\beta$$

$$\beta^{+} \text{ decay}: {}^{A}_{Z}\mathrm{X} \rightarrow {}_{Z-1}^{A}\mathrm{Y} + {}_{+1}^{0}\beta$$

where X and Y are symbols for nuclides, Z is the mass number, and A is the atomic number.

1. Alpha (α) decay is a method of decay in large nuclei. Alpha particles (helium nuclei, He^{2+}), consisting of two neutrons and two protons, are emitted. Because of the particles' relatively high charge, it is heavily ionizing and will cause severe damage if ingested. However, owing to the high mass of the particle, it has little energy and a low range; typically, alpha particles can be stopped with a sheet of paper (or skin).
2. Beta minus (β^-) radiation consists of an energetic electron. It is less ionizing than alpha radiation, but more than gamma. The electron can be stopped with a few centimeters of metal. It occurs when a neutron decays into a proton in a nucleus, releasing the beta particle and an antineutrino. Beta-plus (β^+) radiation is the emission of positrons. As these are antimatter particles, they annihilate any matter nearby, releasing gamma photons.
3. 3. Gamma (γ) radiation consists of photons with a frequency greater than 10^{19} Hz. Gamma radiation occurs to rid the decaying nucleus of excess energy after it has emitted either alpha or beta radiation.

The activity (A) of radionuclide is lost at time (t) according to the formula

$$A = A_0 \cdot e^{-\lambda \cdot t}$$

where A is the radionuclide activity at time $t = 0$, A_0 is the radionuclide activity at time t, and λ is the decay constant of the radionuclide.

The SI unit of activity is the becquerel (Bq). One becquerel is defined as one transformation (or decay) per second. Another unit of radioactivity is the curie (Ci), which was originally defined as the amount of radium emanation (by gaseous radon-222), in equilibrium with 1g of pure radium isotope ^{226}Ra. At present, it is equal, by definition, to the activity of any radionuclide decaying with a disintegration rate of 3.7×10^{10} Bq. The activity of a radioactive substance is characterized by its half-time—the time taken for the activity of a given amount of radioactive substance to decay to half of its value.[14,15]

After the discovery of neutron in 1932, Encico Fermi and colleagues studied the results of bombarding uranium with neutron in 1934.[16] The first person that mentioned the idea of nuclear fission in 1934 was Ida Noddack.[17] After Fermi's publication, Lise Meittner, Otto Hahn, and Fritz Strassmann began to perform a similar experiment and discovered nuclear fission of uranium ^{235}U in 1938.[18,19] Also, Józef Rotblat in 1939 published the results of a study about fission of uranium ^{235}U nuclei.[20]

In nuclear physics and nuclear chemistry, nuclear fission is a nuclear reaction in which the nucleus of an atom splits into smaller parts (lighter nuclei), often producing free neutrons and photons (in the form of gamma rays), as well. Fission of heavy elements is an exothermic reaction that can release large amounts of energy both as electromagnetic radiation and as kinetic energy of the fragments (heating the bulk material where fission takes place).

Three heavy radionuclides, natural ^{235}U and artificial ^{239}Pu and ^{233}U, are capable of reactions (nuclear fission) in which an atom's nucleus splits into smaller parts, releasing a large amount of energy in the process. During the fission of ^{235}U, three neutrons are released in addition to the two daughter atoms (see reaction below) (Figure 1).[21]

$${}^{235}_{92}\mathrm{U} + {}^{1}_{0}\mathrm{n} \rightarrow {}^{236}_{92}\mathrm{U} \rightarrow {}^{90}_{36}\mathrm{Kr} + {}^{143}_{56}\mathrm{Ba} + 3\,{}^{1}_{0}\mathrm{n} + 200\ \mathrm{MeV}$$

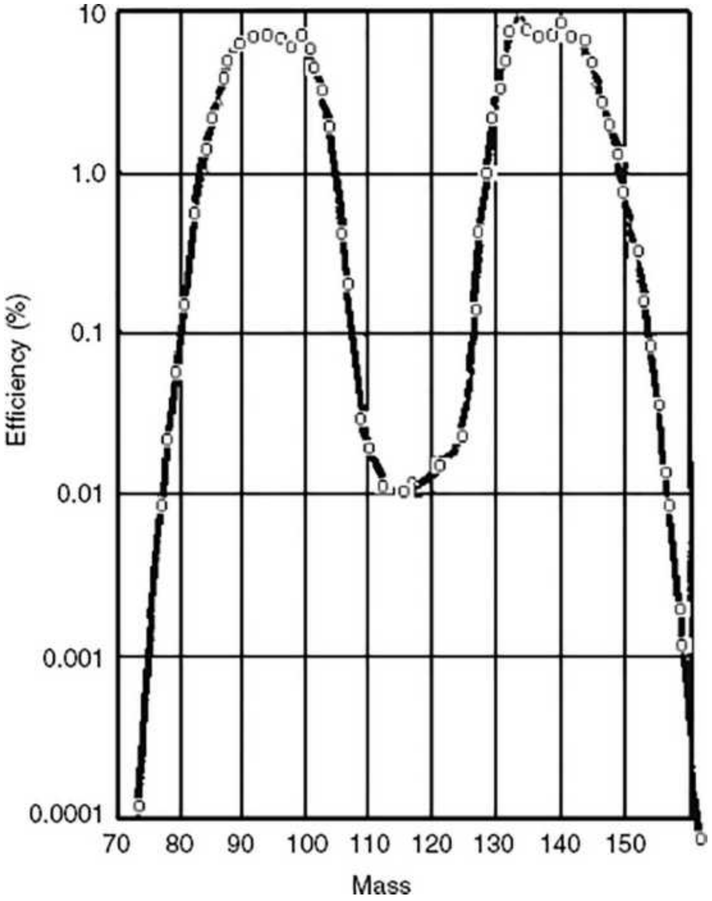


FIGURE 1 Fission yield as a function of mass number for the slow neutron fission of ^{235}U.
Source: Environmental radiochemistry and radiological protection.[21]

Natural and Artificial Radionuclides

Radionuclides present in the natural environment are classified as either of natural or anthropogenic origin. Naturally occurring radionuclides occur in different ecosystems with cosmogenic and primordial providence.[22,23]

1. *Cosmogenic radionuclides:* Cosmic ray–produced radionuclides are generated in the upper-atmosphere gases, e.g., O_2, N_2, and Ar. They are transported to the lower atmosphere and next to the oceans and to the continents. Most of the cosmic radionuclides are produced in very small amounts and only four of them, ^{3}H, ^{7}Be, ^{14}C, and ^{22}Na, constitute significant contributions to the radiation dose to humans. Cosmogenic radionuclides have been measured in humans, topsoil, polar ice, surface rocks, sediments, the biosphere, the ocean floor, and the atmosphere.[24]
2. *Primordial radionuclides:* Among non-series radionuclides of terrestrial origin, only ^{40}K and ^{87}Rb are significant sources of radiation to humans. They are characterized by a long half-time (more than 10^9 years) and small concentrations in crustal rocks (below 1 mBq/kg).[25,26]

The serially occurring radionuclides are contained in four natural decay series—uranium, thorium, actinium, and neptunium—and, except for the actinium series, are named after their parent nuclides (Figures 2–5).

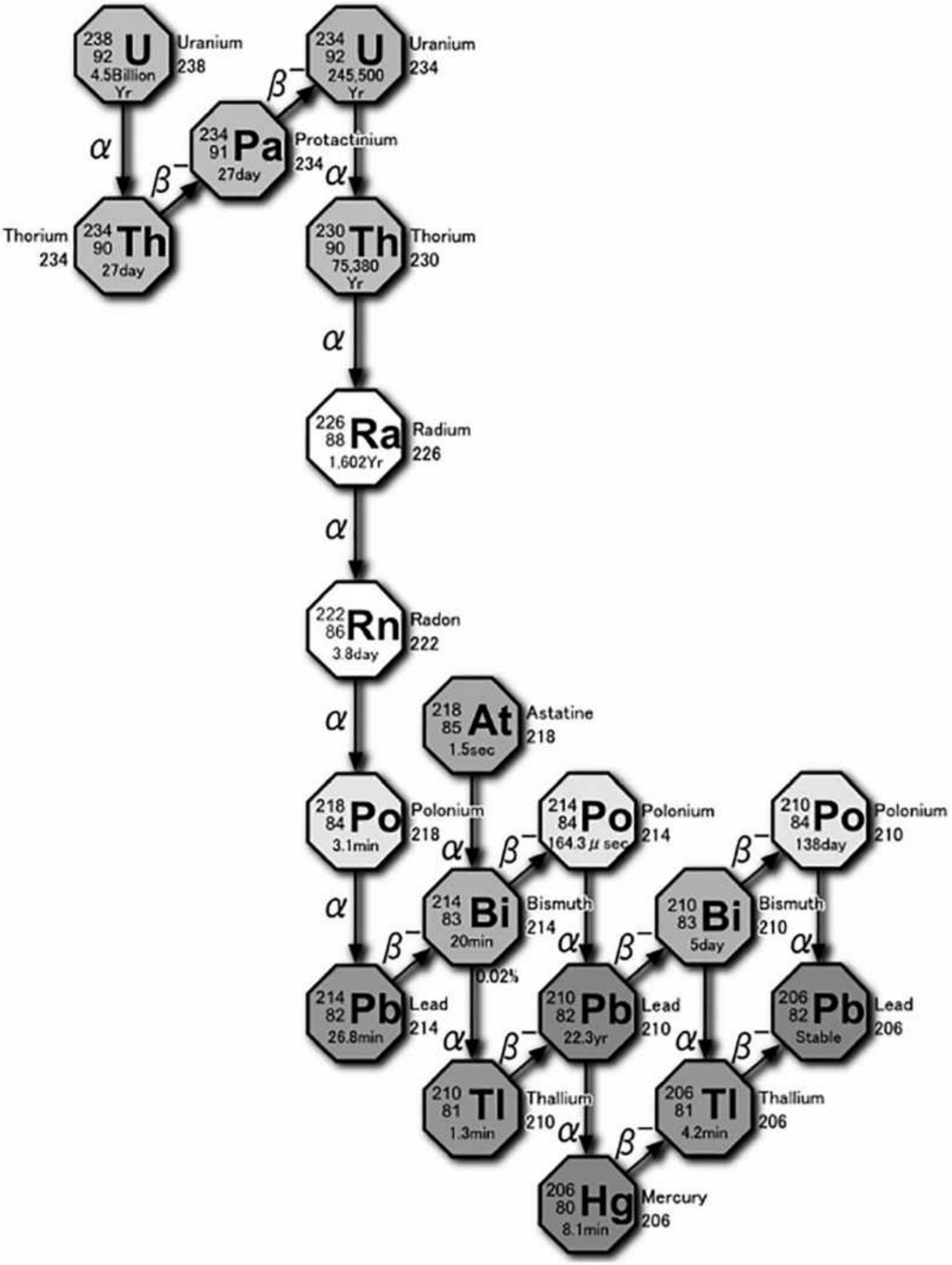


FIGURE 2 Uranium-238 decay series.

Source: Wikipedia, uranium series decay chain, http://upload.wikimedia.org/wikipedia/commons/a/a1/Decay_chain%284n%2B2%2C_Uranium_series%29.PNG.[40]

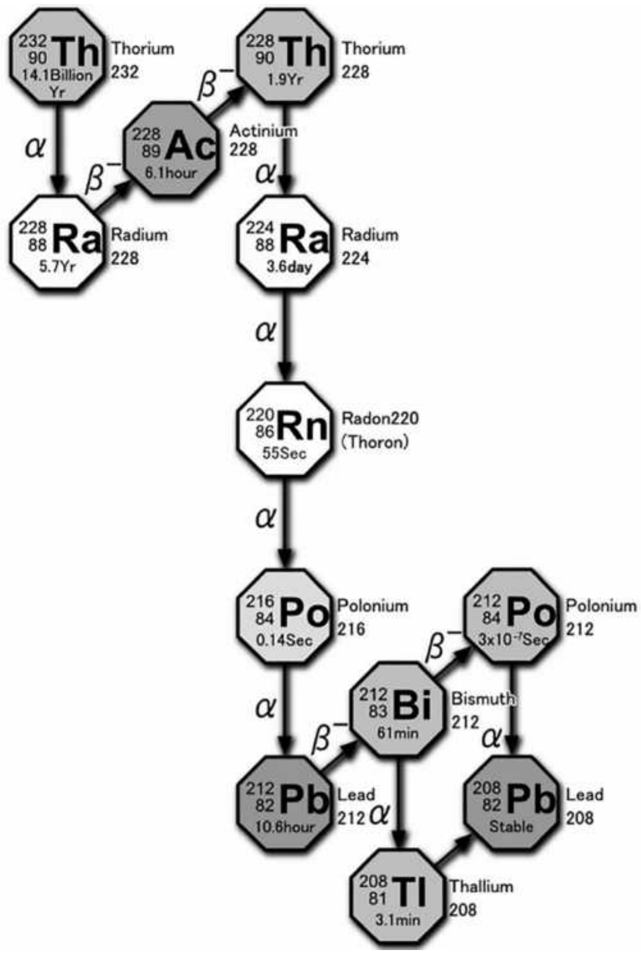


FIGURE 3 Thorium-232 decay series.
Source: Wikipedia, Thorium series decay chain, http://upload.wikimedia.org/wikipedia/commons/1/1c/Decay_chain%284n%2CThorium_series%29.PNG.[41]

Anthropogenic Radionuclides

Anthropogenic-derived radionuclides have been mainly released from several sources since the 1940s. Major sources in the environment are nuclear weapons, nuclear power production, accidents (e.g., the Chernobyl accident in 1986), radioactive waste disposal, solid radioactive waste disposal, and man-made radionuclides as tracers of environmental processes. Fallout from nuclear weapons explosions represents the largest contribution of anthropogenic-derived radionuclides to the natural environment. Anthropogenic radionuclides are divided into three groups:[22,34]

1. *Neutron activation products:* By neutron irradiation of objects, it is possible to induce radioactivity. This activation of stable isotopes enables to create radioisotopes. A lot of artificial radionuclides in the natural environment are produced as a result of the activation process

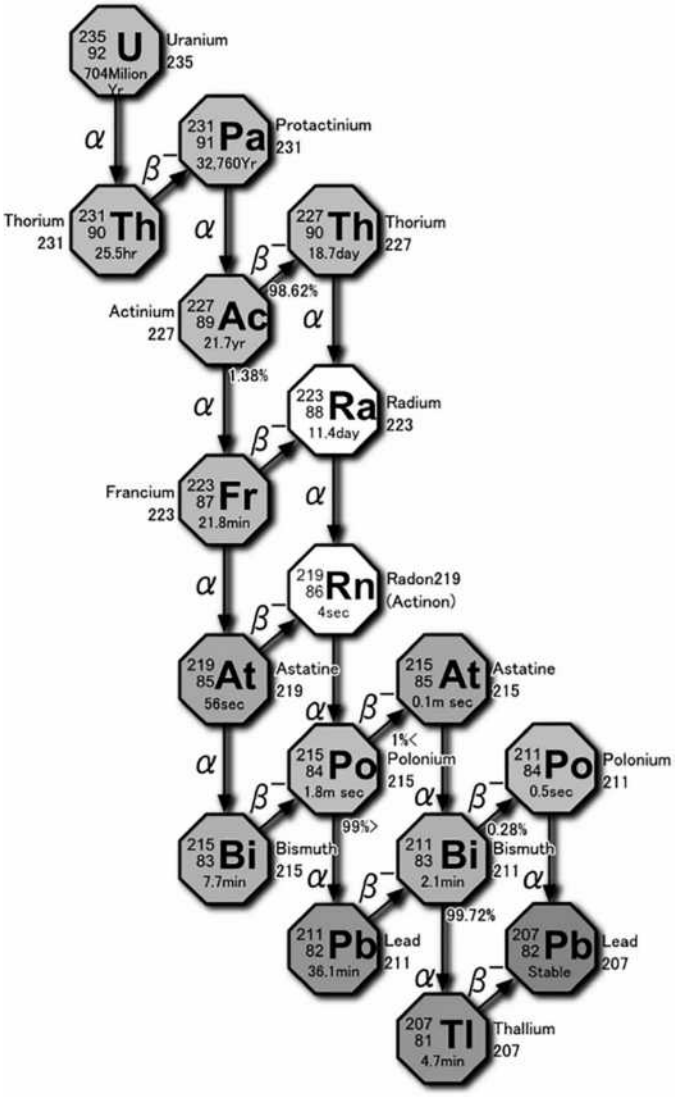


FIGURE 4 Actino-uranium 235U decay series.

Source: Wikipedia, Actinium series decay chain, http://upload.wikimedia.org/wikipedia/commons/1/1e/Decay_chain%284n%2B3%2C_Actinium_series%29.PNG.[42]

during nuclear weapons tests, the work of reprocessing plants and nuclear reactors used in power plants, as well as in nuclear studies. Owing to the use of new radioanalytical techniques, activation products such as ^{22}Na, ^{51}Cr, ^{54}Mn, ^{65}Zn, ^{110m}Ag, and ^{124}Sb could be detected in the natural environment.[26]

2. *Fission radionuclides:* During the fission of ^{235}U, three neutrons are released in addition to two daughter atoms. In the detonation of a nuclear bomb, radioactive fission products are generated from the primary fission of ^{235}U or ^{239}Pu. The most important radionuclides from two families are ^{90}Sr, ^{95}Zr, ^{131}I, ^{132}I, ^{132}Te, ^{137}Cs,^{140}Ba, and ^{144}Ce. These radionuclides are deposited from the atmosphere to the surface of earth, with the fallout comprising components from the stratosphere (78%), local radioactive pollution (12%), and the troposphere (10%).[27]

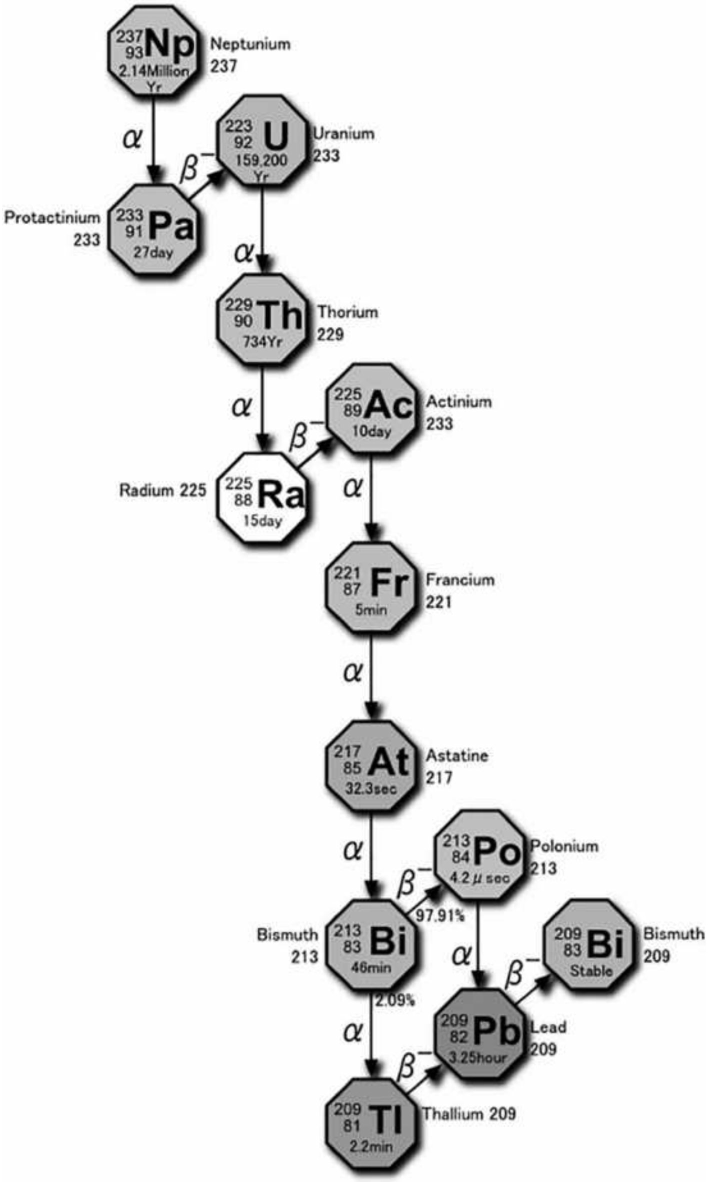


FIGURE 5 Neptunium-237 decay series.
Source: Wikipedia, Neptunium series decay chain, http://upload.wikimedia.org/wikipedia/commons/8/8c/Decay_chain%284n%2B1%2CNeptunium_series%29.PNG.[43]

3. *Transuranic elements:* In chemistry, transuranic elements are chemical elements with atomic numbers greater than 92 (the atomic number of uranium). All transuranic elements are radioactive; 20 transuranic elements have been discovered to date: neptunium (Np), plutonium (Pu), americium (Am), curium (Cm), berkelium (Bk), californium (Cf), einsteinium (Es), fermium (Fm), mendelevium (Md), nobelium (No), lawrencium (Lr), rutherfordium (Rf), dubnium (Db), seaborgium (Sg), bohrium (Bh), hassium (Hs), meitnerium (Mt), darmstadium (Ds), roentgenium (Rg), and copernicium (Cn). Small quantities of neptunium and plutonium are found in nature (in uranium rocks), but most of them are synthesized in nuclear reactors. The most important

sources of transuranic elements (generally, neptunium, plutonium, americium, and curium) in the natural environment are nuclear weapons explosions and nuclear power production as a result of the activation process of uranium, ^{238}U.[21,27,28] The environmental chemistry of some transuranic elements, such as plutonium, is complicated by the fact that solutions of this element can undergo disproportionation, and as a result many different oxidation states can coexist at once.

The most important (long half-time, type of decay, and strong radiotoxicity) natural and anthropogenic radionuclides present in the environment are as follows:[21]

Naturally occurring radionuclides

a. Radionuclides of terrestrial origin—primordial nonseries radionuclide (e.g., ^{40}K and ^{87}Rb)
b. Cosmogenic radionuclides (e.g., ^{3}H and ^{14}C).
c. Primary radionuclides—primordial series radionuclide: long-lived; have been ubiquitous on Earth since their formation (i.e., ca. 4.5×10^9 years ago). The radionuclides ^{238}U, ^{232}Th, and ^{235}U are the parent members of the uranium, thorium, and actinouranium radioactive decay series, respectively

Anthropogenic Radionuclides

a. Neutron activation products (e.g., ^{54}Mn, ^{55}Fe, ^{60}Co,
b. ^{235}U and ^{239}Pu fission radionuclides (e.g., ^{90}Sr, ^{95}Zr, ^{131}I, ^{132}I, ^{132}Te, ^{137}Cs, and ^{144}Ce)
c. Transuranic elements (e.g., ^{237}Np, ^{238}Pu, ^{239}Pu, ^{240}Pu, ^{241}Pu, ^{241}Am, and ^{243}Am)

Natural and artificial radionuclides in different environmental samples (natural water, sediments, soils, biological organisms) are determined by many radiometric methods, in particular neutron activation analysis (NAA), and alpha, beta, and gamma spectrometry.[29,30]

Sources of Radionuclides in the Environment and Pollution Problem

Radionuclides are present in the whole environment in which we live. The principal sources of radionuclides in the natural environment are the wet and dry atmospheric radioactive fallout of particles from natural rock erosion processes and nuclear weapons tests, as well as release from power plants, nuclear submarines, and nuclear reprocessing facilities. Since April 26, 1986, another source of artificial radionuclides, the Chernobyl-originated radioactive debris, has had to be taken into account.[21] A lot of artificial radionuclides (e.g., 200 kg of ^{3}H, 1550 kg of ^{14}C, >160 kg of ^{90}Sr, ~350 kg of ^{137}Cs, and >4600 kg of plutonium isotopes—^{238}Pu, ^{240}Pu, and ^{241}Pu) were deposited in different components of the natural environment (seas and oceans, biota, air, and land).[21] The large amount of radioactive waste produced by human utilization of nuclear power is nowadays one of the major problems. The marine environment is especially exposed to radionuclide contamination because many nuclear power plants and nuclear reprocessing facilities are located in coastal areas. In recent decades, the development of new technologies has also resulted in the production of by-products and waste with the so-called technologically enhanced naturally occurring radioactive materials (TENORM).[31] Therefore, human technological activities (e.g., gas, oil, coal, and fertilizers industries) can increase radiation exposure, not only to the persons directly involved in these activities but also to the local or even the whole population.[32] Moreover, coal mining is the source of a huge amount of waste containing large quantities of natural radionuclides, especially polonium, radium, thorium, and uranium. During combustion, some radionuclides are emitted to the atmosphere as gas and radioactive dust; others remain as concentrated ash.[33,34] Phosphate rocks that are used to produce phosphate fertilizers contain natural radioactive elements (polonium, radium, uranium). During the production of phosphate fertilizers, about 10% of the initial ^{226}Ra, 20% of uranium, and about 85% of ^{210}Po is found in the waste phosphogypsum.[35] These radionuclides are leached by rain from phosphogypsum and, as a consequence, in the neighborhood of the phosphate fertilizers plant, their concentration in soil, flora, and water samples is much higher than in non-contaminated areas.[36,37] Radionuclides are strongly accumulated by some species and the bioaccumulation factor values for some radioactive elements (polonium, plutonium, americium) in sea algae, benthic animals, and fish are more than 1.000.[38,39] Some of these organisms are often used as bioindicators of radioactive pollution of the natural environment.[21] Also, transuranic elements (especially plutonium) belong to the group of radioactivity caused by humans. These radionuclides are important from the radiological point of view due to their high radiotoxicity, long physical lifetime, high chemical reactivity, and long residence in biological systems in the natural environment.[39]

Solution to Radioactive Pollution

A possible solution to radioactive pollution is by the reduction of radioactive emission to the natural environment, change of nuclear technology, and recognition of determination and accumulation processes in living organisms. Plants and animals are capable of accumulating natural and artificial radionuclides from the environment. That is why it is very important to recognize the impact of radionuclides on living organisms and their possible transfer to the human body by way of feeding.[21] Due to the importance of water, air, and food (also cigarette smoking) to human life, their quality must be strictly controlled and monitored. For this reason, studies of food for human consumption must be performed to guarantee that food materials have a low level of radioactivity, both natural and artificial. Especially, long-lived alpha emitters are the most dangerous nuclides in case of ingestion, because the long-term effects of their intake on the human body are the most important from the radiochemical and radiological points of view. A large contamination to the radiation dose received by humans comes from naturally occurring radionuclides accumulated in the body. At the moment, knowledge about accumulation of natural alpha radionuclides by organisms and their ingestion by humans is still very poor.

Conclusion

Radioactivity and radionuclides have been widely applied for more than a hundred years.

Radioactive substances are used to study living organisms, to diagnose and treat diseases (nuclear medicine), to sterilize medical instruments and food, to produce energy for heat and electric power, and to monitor various steps in all types of industrial processes, as well as in research studies (nuclear physics, chemistry, radiology, geochronology and geology, and cosmic research). Many natural and artificial radionuclides are strong radiotoxicants (generally, alpha emitters with a long half-time) to biological organisms. Studies on the bioaccumulation and distribution of radionuclides in different components of the natural environment are very important for radioprotection. Radionuclides used to produce energy in nuclear power plants generate the present and future problems of adequate and responsible utilization of nuclear wastes. For these reasons, the determination and distribution of some important natural and artificial radionuclides in the natural environment should be controlled and monitored.

References

1. Becquerel, H. Sur les radiations invisibles emises par les sels d'uranium. C. R. Acad. Sci. Paris **1896,** *122,* 689–694.
2. Sklodowska-Curie, M. Rayons emis par les composes de l'uranium et du thorium. C. R. Acad. Sci. Paris **1898,** *126,* 1101–1103.
3. Curie, P.; Sklodowska-Curie, M. Sur une substance nouvelle radio-active continue dans la pechblend. C. R. Acad. Sci. Paris **1898,** *127,* 175–178.
4. Sklodowska-Curie, M. Poszukiwania nowego metalu w pechblendzie. Światło **1898,** *1,* 54.
5. Curie, P.; Curie, M.; Bemont, G. Su rune nouvelle substance fortement radioactive, continue dans la pechblende. C. R. Acad. Sci. Paris **1898,** *127,* 1215–1217.

6. Sklodowska-Curie, M. Les rayons de Becquerel et le polonium. Rev. Gen. Sci. Pures Appl. **1899**, *10,* 41–50.
7. Curie, M. Sur le poids atomique du radium. C. R. Acad. Sci. Paris **1902,** *135,* 161–163.
8. Curie, I.; Joliot, M.F. Un nouveau type de radioactivite. C. R. Acad. Sci. Paris **1934,** *198,* 254–256.
9. Rutherford, E.; Soddy F. The cause and nature of radio-activity. Philos. Mag. **1902,** *4,* 370–396, 569–585.
10. Rutherford, E.; Soddy, F. Radioactive change. Philos. Mag. **1903,** *6,* 445457.
11. Rutherford, E. The scattering a, ß and g particles by matter. Philos. Mag. **1911,** *21,* 669–688.
12. Fajans, K. Die stellung der radioelemente im periodischen system. Phys. Z. **1913**, *14,* 136–142.
13. Soddy, F. The radio-element and the periodic law. Chem. News **1913,** *107,* 97–99.
14. Radioactivity. *Encyklopedia Británica Online,* December 18, 2006.
15. Browne, E.; Firestone, F.B. *Table of Radioactive Isotopes*, Shirley, V.S., Ed.; John Willey and Sons: New York, 1986.
16. Fermi, E.; Amaldi, E.; D'Agostino, O.; Rasetti, F.; Segrè, E. Radioacttività provocata da bombardamento di neutroni III. Ric. Sci. **1934,** *5* (1), 452–453.
17. Noddack, I. Über das element 93. Z. Angew. Chem. **1934,** *47* (37), 653–655.
18. Meitner, L; Frisch, O.R. Disintegration of uranium by neutrons: A new type of nuclear reaction. Nature **1939,** *143,* 239–240.
19. Hahn, O.; Strassmann, F. Über den Nachweis und das Verhalten der bei der Bestrahlung des Urans mittels Neutronen entstehenden Erdalkalimetalle. Naturwissenschaften **1939,** *27* (1), 11–15.
20. Rotblat, J. Emission of neutrons accompanying the fission of uranium nuclei. Nature **1939,** *143,* 852.
21. Skwarzec, B. Radiochemia środowiska i ochrona radiologiczna (in Polish), Environmental radiochemistry and radiological protecteon, Wydawnictwo DJ sc., Gdańsk, 2002.
22. Sobkowski, J.; Jelińska-Kaczmarczuk, M. *Chemia Jqdrowa;* Wydawnictwo Adamantan: Warszawa, Poland, 2006.
23. Choppin, G.R.; Liljenzin, J.O.; Rydberg, J. *Radiochemistry and Nuclear Chemistry*; Butterworth-Heinemann Ltd.: Oxford, U.K., 1995.
24. Magill, J.; Galy, J. *Radioactivity, Radionuclides, Radiation*; Springer Verlag: Berlin, 2005.
25. Brune, D.; Forkman, B.; Persson, B. *Nuclear Analytical Chemistry*; Verlag Chemie International Inc.: Deerfield Beach, FL. 1984.
26. Holm, E., Ed. *Radioecology, Lectures in Environmental Radioactivity;* World Scientific: Singapore, 1994.
27. MacKenzie, A.B. Environmental radioactivity: Experience from the 20th century—Trends and issues for the 21st century. Sci. Total Environ. **2000,** *249,* 313–329.
28. Hardy, E.P.; Krey, P.W.; Volchok, H.L. Global inventory and distribution of fallout plutonium. Nature **1973,** *241,* 444–445.
29. Skwarzec, B. Determination of radionuclides in the aquatic environment. In *Analytical Measurements in Aquatic Environments;* Namiesnik, J., Szefer, P., Eds.; CRC Press, Taylor and Francis Group: Boca Raton, FL, 2010; 241–258.
30. L'Annunziata, M.F. *Handbook of Radioactivity Analysis,* 2nd Ed.; Academic Press, Oxford, Great Britain, 2003.
31. TENORM Sources, Radiation Protection, U.S. EPA, March 11th, 2009, available at http://www.epa.gov/radiation/ sources.html#summary-table.
32. Bou-Rabee, F.; Al-Zamel, A.; Al-Fares, R.; Bem, H. Technologically enhanced naturally occurring radioactive materials in the oil industry (TENORM). A review. Nukleonika **2009,** *54,* 3–9.
33. Nakaoka, A.; Fukushima, M.; Takagi, S. Environmental effect of natural radionuclides from coal-fired power plants. Health Phys. **1984,** *3,* 407–416.
34. Flues, M.; Morales, M.; Mazilli, B.P. The influence of a coal-plant operation on radionuclides in soil. J. Environ. Radioact. **2002,** *63,* 285–294.

35. Carvalho, F.P.; Oliviera, J.M.; Lopes, I.; Batista, A. Radionuclides from post uranium mining in rivers in Portugal. J. Environ. Radioact. **2007,** *98*, 298–314.
36. Borylo, A.; Nowicki, W.; Skwarzec, B. Isotope of polonium (^{210}Po) and uranium (^{234}U and ^{238}U) in the industrialized area of Wislinka (Northern Poland). Int. J. Environ. Anal. Chem. **2009,** *89*, 677–685.
37. Skwarzec, B.; Borylo, A.; Kosmska, A.; Radziejewska, S. Polonium (^{210}Po) and uranium (^{234}U and ^{238}U) in water, phosphogypsum and their bioaccumulation in plants around phosphogypsum waste heap in Wislinka (northern Poland). Nukleonika **2010,** *55*, 187–193.
38. Skwarzec, B. Polonium, uranium and plutonium in the southern Baltic Sea. Ambio **1997**, *26* (2), 113–117.
39. Coughtrey, P.J.; Jackson, D.; Jones, C.H.; Kene, P.; Thorne, M.C. *Radionuclides Distribution and Transport in Terrestrial and Aquatic Ecosystems. A Critical Review of Data*; A.A. Balkema: Rotterdam, 1984.
40. Wikipedia, uranium series decay chain, available at http://upload.wikimedia.org/wikipedia/commons/a/a1/Decay_chain%284n%2B2%2C_Uranium_series%29.PNG.
41. Wikipedia, thorium series decay chain, available at http://upload.wikimedia.org/wikipedia/commons/1/1c/Decay_chain%284n%2CThorium_series%29.PNG.
42. Wikipedia, actinium series decay chain, available at http://upload.wikimedia.org/wikipedia/commons/1/1e/Decay_chain%284n%2B3%2C_Actinium_series%29.PNG.
43. Wikipedia, neptunium series decay chain, available at http://upload.wikimedia.org/wikipedia/commons/8/8c/Decay_chain%284n%2B1%2CNeptunium_series%29.PNG.

55

Telecouplings

Vilma Sandström

Introduction

Human and natural systems are increasingly influenced by distant actions and globalized processes such as the flows of material, capital, or information through trade or the migration of human or animal populations (Verburg et al., 2015; Liu et al., 2019). These connections affect the understanding and management of global challenges such as climate change, biodiversity conservation, urbanization, and food security by creating a spatial separation between an action and its drivers. Therefore, they have critical implications to wider sustainability agenda such as the efforts to achieve the United Nations' Sustainable Development Goals (United Nations, 2015). Land use is in the core of these efforts. Drivers and impacts of land-use change and the consequent social and ecological impacts operate across geographical scales and over administrative territories. To disentangle the causal relations between an action, such as land-use change, and its underlying drivers and spatial connections, analytical concepts of teleconnection and telecoupling have been introduced. As suggested by the prefix "tele," both of these concepts focus the attention to distal linkages between land systems, broadening the understanding from the proximate drivers and impacts also to processes taking place in geographically or functionally distant areas (Friis et al., 2016).

Teleconnection. The concept of *teleconnection* was first used in physical sciences over a century ago to describe the atmospheric connections and the interrelations of climatic processes over long distances. Ångstrom (1935) was the first to apply the concept when describing the north–south dipole atmospheric anomaly pattern currently known as the North Atlantic Oscillation (Moser and Hart, 2015). The use of the concept was later broadened to include also other socioecological processes and drivers or land system change, such as flows of capital, people or materials, or species dispersal (Seto et al., 2012; Friis et al., 2016).

Telecoupling. The concept of *telecoupling* builds on the concept of teleconnection, expanding its meaning to include not only the links between an action and its distant driver but also the causal relationships, multidirectional flows, and feedback of land-use change in multiple, connected land systems (Friis et al., 2016; Liu et al., 2019). A telecoupling is created when there is a material, energy, or information flow from one system to another (Liu et al., 2013). These coupled systems are characterized by complex interactions of socioeconomic and biophysical elements that impact each other in dynamic, nonlinear, and emergent ways (Friis and Nielsen, 2017; Liu et al., 2007).

Telecoupling Research

Telecoupling research responds to the need to understand and analyze human–environment systems identifying the specific origin, destination, and the actors involved in multiple interactions between systems, in contrast to the traditional place-based studies that often treat factors from outside the system borders as bulk of exogenous variables without addressing the feedback caused by them (Liu et al., 2013). For example, up to 30% of all global biodiversity loss is related to the production of exported commodities (Lenzen et al., 2012). Thus, linking the consumption of imported products in a region to the specific locations of primary production and the consequent ecosystem impacts allows the identification of consumption as an underlying driver of environmental change and the search for leverage points for change (e.g. Sandström et al., 2017).

Interdisciplinarity is inherent in the telecoupling research related to connected human–environment systems covering both social and natural processes. It creates analytical and epistemic differences between the approaches adopted (Friis, 2019). But then, it represents also the flexibility of the concept to adopt to a wide view of perspectives. Therefore, telecoupling can be seen as an umbrella concept covering various approaches to analyze distant interactions among coupled human and natural systems over distances (Liu et al., 2013).

Liu et al. (2013) present a structural framework for analyzing telecouplings between systems. It has been applied and developed by many scholars from wide range of disciplinary backgrounds (see e.g. Friis and Nielsen, 2019; Liu et al., 2019). The framework builds upon the conceptual frameworks of coupled human–environment systems (Turner et al., 2003), socioecological systems (e.g. Young et al., 2006) and coupled human and natural systems (Liu et al., 2007) adding to these the emphasis on the distant interactions between systems (Liu et al., 2013). Figure 1 presents the major components of the analysis: (1) systems, (2) flows, (3) agents, (4) causes, and (5) effects of a telecoupling, with an emphasis on the interactions between these components (Liu et al., 2019). Telecoupled systems are classified as sending (the systems from where the flows move outward), receiving (the systems that receive the flows from the sending systems), or spillover systems (the systems that impact and/or are impacted by the interactions between the sending and receiving systems) (Liu et al., 2013). Areas can also act simultaneously as sending, receiving, or spillover systems. Flows are assisted by agents, such as different actors, people, or animals, and they can cause socioeconomic or environmental effects in any of the systems analyzed (Liu et al., 2013). Defining system boundaries is always context-dependent (Liu et al., 2019). Systems can be separated by spatial distance, often by jurisdictional boundaries or functional distance in terms of governance (Eakin et al., 2014). While the framework of Liu et al. (2013) emphasizes the spatial hierarchy, Eakin et al. (2014) outline telecouplings as results of networked interactions across scales (Friis and Nielsen, 2017) (Figure 1).

In the recent years, the number of studies exploring the telecouplings has increased rapidly (Kapsar et al., 2019). Although the concept of telecoupling was first introduced in land system science, its utility extends also beyond the disciplinary borders. It has been adopted within and across disciplines such as earth science, ecology, and economics as well as environmental and political science (Kapsar et al., 2019). Many of these studies have concentrated on international trade as a driver creating these spatial connections between sometimes seemingly unconnected land systems, for instance, by studying the consumption of imported products and the related displaced land use and its environmental impacts (e.g. Rulli et al., 2019; Sandström et al., 2017). Also, topics such as species invasion and migration (Liu et al., 2014; Hulina et al., 2017), urbanization (Fang and Ren, 2017), tourism (Liu et al., 2015), fisheries management (Carlson et al., 2018), and conservation (Carrasco et al., 2017; Fang and Liu, 2016) have been studied in a telecoupling context.

Conclusions

The analytical concept and framework of telecoupling provide an approach for analyzing the distant interactions of systems in a socioecological context integrating the often isolated disciplinary knowledge into a more comprehensive picture. In doing so, the concept and approach facilitate the broader

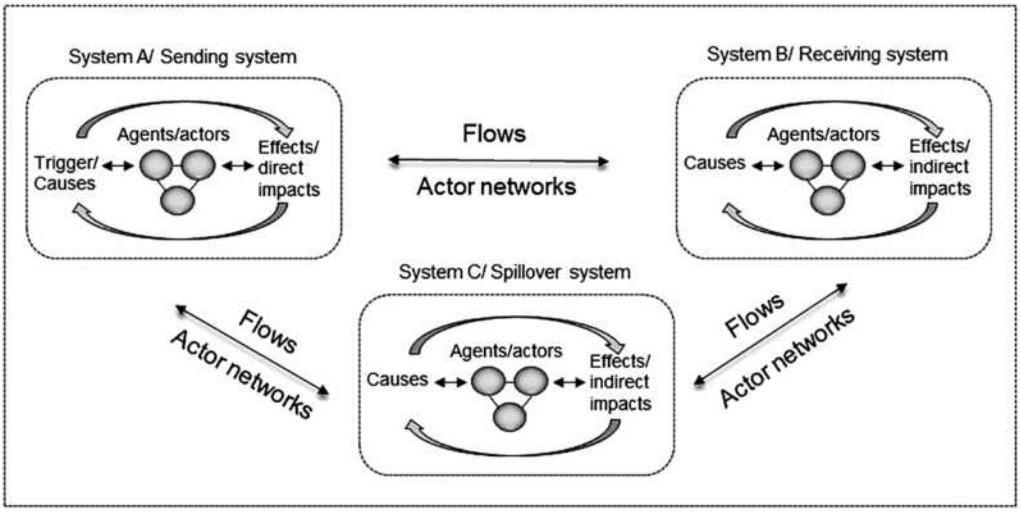


FIGURE 1 Telecoupling framework, adapted from Liu et al. (2013, 2014) and Eakin et al. (2014). The figure presents the main components of analysis as described by Liu et al. (2013): systems, flows, agents, causes, and effects, adapted with the inclusion of actor networks creating a telecoupling (Eakin et al., 2014). A first trigger initiates the telecoupling causing both direct impacts in the same system and indirect impacts in distantly coupled systems together with the feedback processes influencing potentially all systems in question (Eakin et al., 2014). The role of the systems as sending, receiving, or spillover systems can change depending on the perspective (Liu et al., 2014). System boundaries can be determined by spatial hierarchy (Liu et al., 2013) or by scope of governance (Eakin et al., 2014).

understanding of spatial complexity between an action and its drivers and impacts, that is inherent in all human and natural systems in the increasingly globalizing world. The concept has been applied in various temporal and spatial scales. Being relatively novel, the telecoupling framework, approach, and tools are in continuous development to capture and understand all the dimensions involved in and affected by distant interactions.

References

Ångstrom, A. 1935. Teleconnections of climatic changes in present time. *Geografiska Annaler*, 17:243–258.

Carlson, A., Taylor, W., Liu, J., and Orlic, I. 2018. Peruvian anchoveta as a telecoupled fisheries system. *Ecology and Society*, 23(1):35.

Carrasco, L. R., Chan, J., McGrath, F., and Nghiem, L. 2017. Biodiversity conservation in a telecoupled world. *Ecology and Society*, 22(3):24.

Eakin, H., DeFries, R., Kerr, S., Lambin, E. F., Liu, J., Marcotullio, P. J., Messerli, P., Reenberg, A., Rueda, X., Swaffield, S. R., Wicke, B., and Zimmerer, K. 2014. Significance of telecoupling for exploration of land-use change. In Seto, K, &, Reenberg, A. (Eds.) *Rethinking Global Land Use in an Urban Era* (pp. 141–161). Cambridge, MA: MIT Press.

Fang, W., and Liu, J. 2016. Conservation planning beyond giant pandas: the need for an innovative telecoupling framework. *Science China Life Sciences*, 60(5):551–554.

Fang, C., and Ren, Y. 2017. Analysis of emergy-based metabolic efficiency and environmental pressure on the local coupling and telecoupling between urbanization and the eco-environment in the Beijing-Tianjin-Hebei urban agglomeration. *Science China Earth Sciences*, 60(6):1083–1097.

Friis, C., Nielsen, J. Ø., Otero, I., Haberl, H., Niewöhner, J., and Hostert, P. 2016. From teleconnection to telecoupling: taking stock of an emerging framework in land system science. *Journal of Land Use Science*, 11(2):131–153.

Friis, C., and Nielsen, J. 2017. Land-use change in a telecoupled world: the relevance and applicability of the telecoupling framework in the case of banana plantation expansion in Laos. *Ecology and Society*, 22(4):30.

Friis, C. 2019. Telecoupling: a new framework for researching land-use change in a globalised world. In *Telecoupling* (pp. 49–67). Palgrave Macmillan, Cham.

Friis, C., and Nielsen, J. Ø. 2019. Global land-use change through a telecoupling lens: an introduction. In *Telecoupling* (pp. 1–15). Palgrave Macmillan, Cham.

Hulina, J., Bocetti, C., Ii, H., Hull, V., Yang, W., and Liu, J. 2017. Telecoupling framework for research on migratory species in the anthropocene. *Elementa—Science of the Anthropocene*, 5:5. https://doi.org/10.1525/elementa.184.

Kapsar, K. E., Hovis, C. L., Bicudo da Silva, R. F., Buchholtz, E. K., Carlson, A. K., Dou, Y., Du, Y., Furumo, P. R., Li, Y., Torres, A., Yang, D., Wan, H. Y., Zaehringer, J. G., and Liu, J. (2019). Telecoupling research: the first five years. *Sustainability*, 11(4):1033.

Lenzen, M., Moran, D., Kanemoto, K., Foran, B., Lobefaro, L., and Geschke, A. 2012. International trade drives biodiversity threats in developing nations. *Nature*, 486:109–112. http://dx.doi.org/10.1038/nature11145

Liu, J., Dietz, T., Carpenter, S. R., Alberti, M., Folke, C., Moran, E., Pell, A. N., Deadman, P., Kratz, T., Lubchenco, J., Ostrom, E., Uyang, Z., Provencher, W., Redman, C. L., Schneider, S. H., and Taylor, W. 2007. Complexity of coupled human and natural systems. *Science*, 317(5844):1513–1516.

Liu, J., Hull, V., Batistella, M., DeFries, R., Dietz, T., Fu, F., Hertel, T. W., Izaurralde, R. C., Lambin, E. F., Li, S., Martinelli, L. A., McConnell, W. J., Moran, E. F., Naylor, R., Ouyand, Z., Polenske, K. R., Reenberg, A., de Miranda Rocha, G., Simmons, C. S., Verbugr, P. H., Vitousek, P. M., Zhang, F., and Zhu, C. 2013. Framing sustainability in a telecoupled world. *Ecology and Society*, 18(2):26.

Liu, J., Hull, V., Moran, E., Nagendra, H., Swaffield, S., and Turner, B. L. 2014. Applications of the telecoupling framework to land-change science. In K. C. Seto & A. Reenberg (Eds.), *Rethinking Global Land Use in and Urban Age*. MIT Press, Cambridge, MA.

Liu, J., Hull, V., Luo, J., Yang, W., Liu, W, Viña, A., Vogt, C., Xu, Z., Yang, H., Zhang, J., An, L., Chen, X., Li, Z., Ouyang, Z., Xu,, W., and Zhang, H. 2015. Multiple telecouplings and their complex interrelationships. *Ecology and Society* 20(3):44.

Liu, J., Herzberger, A., Kapsar, K., Carlson, A. K., and Connor, T. 2019. What is telecoupling? In: Friis, C., Nielsen, J. (eds) *Telecoupling*. (pp. 19–48). Palgrave Studies in Natural Resource Management. Palgrave Macmillan, Cham.

Moser, S. C., and Hart, J. A. F. 2015. The long arm of climate change: societal teleconnections and the future of climate change impacts studies. *Climatic Change*, 129(1–2):13–26.

Rulli, M. C., Casirati, S., Dell'Angelo, J., Davis, K. F., Passera, C., and D'Odorico, P. 2019. Interdependencies and telecoupling of oil palm expansion at the expense of Indonesian rainforest. *Renewable and Sustainable Energy Reviews*, 105:499–512.

Sandström, V., Kauppi, P. E., Scherer, L., and Kastner, T. 2017. Linking country level food supply to global land and water use and biodiversity impacts: the case of Finland. *Science of the Total Environment*, 575:33–40.

Seto, K. C., Reenberg, A., Boone, C. G., Fragkias, M., Haase, D., Langanke, T., Marcotullio, P., Munroe, D. K., Olah, B., and Simon, D. 2012. Urban land teleconnections and sustainability. *Proceedings of the National Academy of Sciences*, 109(20):7687–7692.

Turner, B. L., Kasperson, R. E., Matson, P. A., McCarthy, J. J., Corell, R. W., Christensen, L., Eckley, N., Kasperson, J. X., Luers, A., Martello, M. L., Polsky, C., Pulsipher, A., and Schiller, A. 2003. A framework for vulnerability analysis in sustainability science. *Proceedings of the National Academy of Sciences*, 100(14):8074–8079.

United Nations. 2015. Transforming our world: the 2030 agenda for sustainable development. Resolution adopted by the General Assembly. https://sustainabledevelopment.un.org/post2015/transformingourworld.

Verburg, P. H., Crossman, N., Ellis, E. C., Heinimann, A., Hostert, P., Mertz, O., Nagendra, H., Sikor, T., Erb, K-H., Golubiewski, N., Grau, R., Grove, M., Konaté, S., Meyfroidt, P., Parker, D. C., Chowdhury, R. R., Shibata, H., Thomson, A., and Zhen, L. 2015. Land system science and sustainable development of the earth system: a global land project perspective. *Anthropocene*, 12:29–41.

Young, O., Lambin, E., Alcock, F., Haberl, H., Karlsson, S., McConnell, W., Myint, T., Pahl-Wostl, C., Polsky, C., Ramakrishnan, P. S., Schroeder, H., Scouvart, M., and Verburg, P H. 2006. A portfolio approach to analyzing complex human-environment interactions: institutions and land change. *Ecology and Society*, 11(2):31.

Index

Page numbers followed by f and t indicate figures and tables, respectively

SECOND EDITION

ENVIRONMENTAL MANAGEMENT HANDBOOK

VOLUME II

Managing Biological and Ecological Systems

edited by

Brian D. Fath

Sven E. Jørgensen

CRC Press
Taylor & Francis Group

Managing Biological and Ecological Systems

Environmental Management Handbook, Second Edition

Edited by
Brian D. Fath and Sven E. Jørgensen

Volume 1
Managing Global Resources and Universal Processes

Volume 2
Managing Biological and Ecological Systems

Volume 3
Managing Soils and Terrestrial Systems

Volume 4
Managing Water Resources and Hydrological Systems

Volume 5
Managing Air Quality and Energy Systems

Volume 6
Managing Human and Social Systems

Managing Biological and Ecological Systems

Second Edition

Edited by

Brian D. Fath and Sven E. Jørgensen

Assistant to Editor

Megan Cole

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

Cover photo: Lake Elsinore, California, USA, B. Fath

Second edition published 2021
by CRC Press
6000 Broken Sound Parkway NW, Suite 300, Boca Raton, FL 33487-2742

and by CRC Press
2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

First edition published by CRC Press 2013
CRC Press is an imprint of Taylor & Francis Group, LLC

ISBN: 978-1-138-34264-4 (hbk)
ISBN: 978-0-429-34617-0 (ebk)

Typeset in Minion
by codeMantra

Contents

SECTION I APC: Anthropogenic Chemicals and Activities

SECTION II COV: Comparative Overviews of Important Topics for Environmental Management

SECTION III CSS: Case Studies of Environmental Management

SECTION IV DIA: Diagnostic Tools: Monitoring, Ecological Modeling, Ecological Indicators, and Ecological Services

SECTION V ENT: Environmental Management Using Environmental Technologies

SECTION VI NEC: Natural Elements and Chemicals Found in Nature

SECTION VII PRO: Basic Environmental Processes

Preface

Given the current state of the world as compiled in the massive Millennium Ecosystem Assessment Report, humans have changed ecosystems more rapidly and extensively during the past 50 years than in any other time in human history. These are unprecedented changes that need certain action. As a result, it is imperative that we have a good scientific understanding of how these systems function and good strategies on how to manage them.

In a very practical way, this multivolume *Environmental Management Handbook* provides a comprehensive reference to demonstrate the key processes and provisions for enhancing environmental management. The experience, evidence, methods, and models relevant for studying environmental management are presented here in six stand-alone thematic volumes as follows:

VOLUME 1 – Managing Global Resources and Universal Processes
VOLUME 2 – Managing Biological and Ecological Systems
VOLUME 3 – Managing Soils and Terrestrial Systems
VOLUME 4 – Managing Water Resources and Hydrological Systems
VOLUME 5 – Managing Air Quality and Energy Systems
VOLUME 6 – Managing Human and Social Systems

In this manner, the handbook introduces in the first volume the general concepts and processes used in environmental management. The next four volumes deal with each of the four spheres of nature (biosphere, geosphere, hydrosphere, and atmosphere). The last volume ties the material together in its application to human and social systems. These are very important chapters for a wide spectrum of students and professionals to understand and implement environmental management. In particular, the features include the following:

- The first handbook that demonstrates the key processes and provisions for enhancing environmental management.
- Addresses new and cutting-edge topics on ecosystem services, resilience, sustainability, food–energy–water nexus, socio-ecological systems, etc.
- Provides an excellent basic knowledge on environmental systems, explains how these systems function, and gives strategies on how to manage them.
- Written by an outstanding group of environmental experts.

Since the handbook covers such a wide range of materials from basic processes, to tools, technologies, case studies, and legislative actions, each handbook entry is further classified into the following categories:

APC: Anthropogenic chemicals: The chapters cover human-manufactured chemicals and activities
COV: Indicates that the chapters give comparative overviews of important topics for environmental management

CSS: The chapters give a case study of a particular environmental management example
DIA: Means that the chapters are about diagnostic tools: monitoring, ecological modeling, ecological indicators, and ecological services
ELE: Focuses on the use of legislation or policy to address environmental problems
ENT: Addresses environmental management using environmental technologies
NEC: Natural elements and chemicals: The chapters cover basic elements and chemicals found in nature
PRO: The chapters cover basic environmental processes.

Overall, these volumes will be a valuable resource for all libraries supporting programs in environmental science and studies, earth science, geography, and policy.

In this second volume, over 40 entries focus on managing biological and ecological systems. The coverage ranges from species to ecosystems, with chapters on biodiversity, sustainability, and conservation, to name a few. Specific case studies using biological controls are included, with new entries on sustainable fisheries and recovery of fish populations.

Brian D. Fath
Brno, Czech Republic
December 2019

Editors

Brian D. Fath is Professor in the Department of Biological Sciences at Towson University (Maryland, USA) and Senior Research Scholar at the International Institute for Applied Systems Analysis (Laxenburg, Austria). He has published over 180 research papers, reports, and book chapters on environmental systems modeling, specifically in the areas of network analysis, urban metabolism, and sustainability. He has co-authored the books *A New Ecology: Systems Perspective* (2020), *Foundations for Sustainability: A Coherent Framework of Life–Environment Relations* (2019), and *Flourishing within Limits to Growth: Following Nature's Way* (2015). He is also Editor-in-Chief for the journal *Ecological Modelling* and Co-Editor-in-Chief for *Current Research in Environmental Sustainability*. Dr. Fath was the 2016 recipient of the Prigogine Medal for outstanding work in systems ecology and twice a Fulbright Distinguished Chair (Parthenope University, Naples, Italy in 2012 and Masaryk University, Czech Republic in 2019). In addition, he has served as Secretary General of the International Society for Ecological Modelling, Co-Chair of the Ecosystem Dynamics Focus Research Group in the Community Surface Modeling Dynamics System, and member and past Chair of Baltimore County Commission on Environmental Quality.

Sven E. Jørgensen (1934–2016) was Professor of environmental chemistry at Copenhagen University. He received a doctorate of engineering in environmental technology and a doctorate of science in ecological modeling. He was an honorable doctor of science at Coimbra University (Portugal) and at Dar es Salaam (Tanzania). He was Editor-in-Chief of *Ecological Modelling* from the journal inception in 1975 until 2009. He was Editor-in-Chief for the *Encyclopedia of Environmental Management* (2013) and *Encyclopedia of Ecology* (2008). In 2004, Dr. Jørgensen was awarded the Stockholm Water Prize and the Prigogine Medal. He was awarded the Einstein Professorship by the Chinese Academy of Sciences in 2005. In 2007, he received the Pascal Medal and was elected a member of the European Academy of Sciences. He had published over 350 papers, and has edited or written over 70 books. Dr. Jørgensen gave popular and well-received lectures and courses in ecological modeling, ecosystem theory, and ecological engineering worldwide.

Contributors

G. J. Ash
E.H. Graham Center for Agricultural Innovation, Industry and Investment, NSW
Charles Sturt University
Wagga Wagga, New South Wales, Australia

William Au
Department of Preventive Medicine and Community Health
University of Texas Medical Branch
Galveston, Texas

Simone Bastianoni
Ecodynamics Group
Department of Chemistry
University of Siena
Siena, Italy

James R. Brandle
School of Natural Resource Sciences
University of Nebraska—Lincoln
Lincoln, Nebraska

Benjamin Burkhard
Institute for the Conservation of Natural Resources
University of Kiel
Kiel, Germany

Nídia Sá Caetano
Chemical Engineering Department
School of Engineering (ISEP)
Polytechnic Institute of Porto (IPP)
and
Laboratory for Process, Environmental and Energy Engineering
Porto, Portugal

Vera Lucia S.S. de Castro
Ecotoxicology and Biosafety Laboratory
Brazilian Agricultural Research Corporation (Embrapa Environment)
São Paulo, Brazil

Ioan M. Ciumasu
CEARC Center, Observatory of Versailles Saint-Quentin-en-Yvelines
University of Versailles Saint-Quentin-en-Yvelines
Guyancourt, France

Patrick De Clerq
Department of Crop Protection
Ghent University
Ghent, Belgium

Luca Coscieme
Ecodynamics Group
Department of Chemistry
University of Siena
Siena, Italy

Mihai Costica
Faculty of Biology
Alexandru Ioan Cuza University
Iasi, Romania

Marc A. Cubeta
Center for Integrated Fungal Research, Plant Pathology
North Carolina State University
Raleigh, North Carolina

Keith Culver
Faculty of Management and STAR Research Initiative
University of British Columbia (Okanagan)
Kelowna, British Columbia, Canada

Katja Enberg
Department of Biological Sciences
University of Bergen
Bergen, Norway

Shannon Estenoz
Office of Everglades Restoration Initiatives
U.S. Department of the Interior
Davie, Florida

David N. Ferro
Department of Entomology
University of Massachusetts
Amherst, Massachusetts

J. R. de Freitas
Department of Soil Science
University of Saskatchewan
Saskatoon, Saskatchewan, Canada

Stephen R. Gliessman
Department of Environmental Studies
University of California—Santa Cruz
Santa Cruz, California

Ragini Gothalwal
Institute of Microbiology and Biotechnology
Barkatullah University
Bhopal, India

Simon Gowen
Department of Agriculture
University of Reading
Reading, United Kingdom

Simon Grenier
Functional Biology, Insects and Interactions
National Institute for Agricultural Research (INRA)
Villeurbanne, France

Ann E. Hajek
Department of Entomology
Cornell University
Ithaca, New York

Lise Stengård Hansen
Danish Pest Infestation Laboratory
Danish Institute of Agricultural Sciences
Kongens Lyngby, Denmark

Peter Harris
Agriculture and Agri-Food Canada
Lethbridge, Alberta, Canada

James D. Harwood
Department of Entomology
University of Kentucky
Lexington, Kentucky

Laurie Hodges
Department of Agronomy and Horticulture
University of Nebraska—Lincoln
Lincoln, Nebraska

Heikki Hokkanen
Department of Applied Biology
University of Helsinki
Helsinki, Finland

Hei-Ti Hsu
Agricultural Research Service (USDA-ARS)
Floral and Nursery Plants Research
U.S. Department of Agriculture
Beltsville, Maryland

Sven Erik Jørgensen
Institute A, Section of Environmental Chemistry
Copenhagen University
Copenhagen, Denmark

Marion Kandziora
Institute for the Conservation of Natural Resources
University of Kiel
Kiel, Germany

Anthony P. Keinath
Coastal Research and Education Center
Clemson University
Clemson, South Carolina

Peter Kerr
CSIRO Health and Biosecurity
Black Mountain Laboratories
Canberra ACT, Australia

Tomaz Langenbach
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil

David B. Langston, Jr.
Tidewater Agricultural Research and Extension Center
Virginia Tech
Suffolk, Virginia

Pierre Mineau
Science and Technology Branch
Environment Canada
Ottawa, Ontario, Canada

Felix Müller
Ecology Center
University of Kiel
Kiel, Germany

Joji Muramoto
University of California Cooperative Extension
Center for Agroecology and Sustainable Food Systems
Department of Environmental Studies
University of California—Santa Cruz
Santa Cruz, California

John J. Obrycki
Department of Entomology
University of Kentucky
Lexington, Kentucky

Philip Oduor-Owino
Department of Botany
University of Kenyatta
Nairobi, Kenya

Maurizio G. Paoletti
Department of Biology
University of Padova
Padova, Italy

Timothy Paulitz
Department of Plant Science
MacDonald Campus of McGill University
Ste-Anne-de-Bellevue, Quebec, Canada

Catalina Chaparro Pedraza
Swiss Federal Institute of Aquatic Science and Technology EAWAG
Dübendorf, Switzerland

Meir Paul Pener
Department of Cell and Developmental Biology
The Hebrew University of Jerusalem
Jerusalem, Israel

Julie A. Peterson
Department of Entomology
University of Kentucky
Lexington, Kentucky

Odo Primavesi
Brazilian Agricultural Research Corporation (Embrapa)
São Paulo, Brazil

Federico M. Pulselli
Ecodynamics Group
Department of Chemistry
University of Siena
Siena, Italy

Jennifer Ruesink
Department of Biology
University of Washington
Seattle, Washington

Alka Sapat
School of Public Administration
Florida Atlantic University
Boca Raton, Florida

Claus Schimming
Institute for the Conservation of Natural Resources
University of Kiel
Kiel, Germany

Cetin Sengonca
Department of Entomology and Plant Protection
Institute of Plant Pathology
University of Bonn
Bonn, Germany

Eric B. Spurr
Department of Wildlife Ecology
Landcare Research New Zealand, Ltd.
Lincoln, New Zealand

F. Craig Stevenson
Department of Soil Science
University of Saskatchewan
Saskatoon, Saskatchewan, Canada

Tanja Strive
CSIRO Health and Biosecurity
Black Mountain Laboratories
Canberra ACT, Australia

Alberto Traverso
Thermochemical Power Group
Department of Mechanical, Energy, Management and Transportation Engineering (DIME)
University of Genoa
Genoa, Italy

David Tucker
National Energy Technology Laboratory
Department of Energy
Morgantown, West Virginia

Jean-Paul Vanderlinden
CEARC Center
Observatory of Versailles Saint-Quentin-en-Yvelines
University of Versailles Saint-Quentin-en-Yvelines
Guyancourt, France

Denise Vienne
School of Public Administration
Florida Atlantic University
Boca Raton, Florida

A. Wang
E.H. Graham Center for Agricultural Innovation, Industry and Investment, NSW
Charles Sturt University
Wagga Wagga, New South Wales, Australia

Gerald E. Wilde
Department of Entomology
Kansas State University
Manhattan, Kansas

Wilhelm Windhorst
Institute for the Conservation of Natural Resources
University of Kiel
Kiel, Germany

Xinhua Zhou
School of Natural Resources
University of Nebraska—Lincoln
Lincoln, Nebraska

Fabian Zimmermann
Institute of Marine Research
Bergen, Norway

and

Institute of Marine Sciences
and
Department of Applied Mathematics
University of California—Santa Cruz
Santa Cruz, California

I

APC: Anthropogenic Chemicals and Activities

1

Animals: Sterility from Pesticides

William Au

Mechanisms of Action of Pesticides

Pesticides can be subdivided into several major categories: organophosphates, carbamates, organochlorines, synthetic pyrethroids, and others. Their principal mechanisms of action include inhibition of cholinesterase, perturbation of microsomal enzyme production, and damage to nervous systems.[1] Therefore, it is possible that excessive exposure to pesticides can interfere with gametogenesis, sexual activities, and reproduction leading to the expression of sterility.

Observed Effects in Native Animals and in Experimental Systems

Among the pesticides, organochlorines are characterized by their persistence in the environment, and the potential for both bioaccumulation and transfer of the pesticides up the food chain. Therefore, the widespread use of organochlorines in the past has been documented to cause contamination of wildlife and reduction of their populations. For example, the exposure is associated with a significant reduction of fish populations such as trout and salmon.[2] Subsequently, the populations of predatory birds were significantly reduced.[2] The devastating effects in migratory birds and in other wildlife were also demonstrated.[3,4] In these cases, failure to reproduce appropriately has been shown to be a major cause for the decline of the populations.

In studies using experimental animals under controlled exposure conditions to pesticides, organochlorine pesticides such as methoxychlor have been reported to cause reduction of fertility and litter size,[5,6] and kepone to cause anovulation.[7,8] Organophosphate pesticides have been shown to induce premature ovulation and to perturb oocyte development.[9]

Observed Effects in Humans

Very few pesticides have been shown systematically and consistently to cause sterility in humans. An exception is the exposure to a nematocide, 1,2-dibromo-3-chloropropane (DBCP). In the 1970s, workers in several pesticide manufacturing plants were reported to have fertility problems. From a systematic investigation, the infertility based on reduced sperm counts was shown to be associated with testicular function alteration and with exposure to DBCP rather than to other pesticides.[10] Subsequently, the

TABLE 1 Reproductive Problems from Exposure to Dibromochloropropane (DBCP)

A. Oligospermia	
Months of exposure to DBCP[a]	% Workers with oligospermia
0	2.9
1–6	8.3
7–24	28.6
25–42	66.7
>42	76.5
B. Offspring	
Years after recovery from oligospermia[b]	% Females in offspring
5	84.6
8	78.9
17	58.6

[a] Data derived from 10.

[b] Data derived from 13–15.

same group of scientists found that the reduction of sperm count was associated with an occupational exposure as short as 3 months[11] and with an employment duration-dependent effect (Table 1A). As shown in the table, as many as 76.5% of the workers who had been exposed to DBCP for more than 42 months were oligospermic. Among all the affected workers, many were azospermic or sterile.

Long-term follow-up studies of DBCP production workers showed that some of the affected workers did regain fertility and testicular function, and many of them were able to have children. Their offspring appeared normal and healthy. However, these workers predominantly had female offspring (Table 1B), ranging from 58.6 to 84.6% for the recovery duration from 5 to 17 years.[12] As shown in the table, the highest female to male offspring ratio was found among workers within 5 years of recovery.[13,14] It appears that the recovery is a slow process and complete recovery with respect to the sex ratio was achieved only after 17 years.[15] The observation confirms the previous recommendation of using altered sex ratios as an indication of reproductive hazards associated with pesticides.[16] In another study, males infertile due to poor sperm quality were more likely than expected to be in the agricultural occupations with exposure to pesticides.[17] Papaya fumigant workers with exposure to ethylene dibromide were reported to have significantly reduced sperm count per ejaculate (the percentage of viable and motile sperm) and increases in the proportion of sperm with abnormalities.[18] Abnormal pregnancy outcomes (miscarriages and preterm deliveries) were associated with exposure to a variety of chemicals in combination with pesticides (atrazine, glyphosate, organophosphates, 4-[2,4-dichlorophenoxy]butyric acid) in males.[19] On the other hand, fertility in traditional male farmers, compared with organic farmers (who do not use pesticides), was not influenced by exposure to pesticides, based on the time taken to have the youngest child.[20]

Among females, exposure to DBCP in pesticide manufacturing plants appears to have no effects on their fertility based on a limited survey.[21] A study was conducted to investigate the relationship between the plasma level of organochlorine pesticides and the diagnosis of endometriosis, and no association was found.[22] On the other hand, among women with medically confirmed infertility, exposure to pesticides was shown to be a significant contributing factor.[23] Furthermore, the mechanism appears to be due to abnormal ovulation.

Future Considerations

Based on the mechanisms of action of pesticides[1,24] and on observations in animals, it is highly likely that overexposure and/or prolonged exposure to pesticides can cause reproductive problems in human. However, adverse reproductive effects have not been demonstrated unequivocally with modern pesticides.

One reason is that the human population is usually exposed to much lower doses of pesticides, except in accidental exposure conditions, than those used in animals that have been shown to cause sterility. Under this condition, any adverse effects in human would be very small. Therefore, investigations using inappropriate study protocols may have generated inconsistent observations. Future studies should be conducted by using large enough populations and by minimizing multiple confounding factors. At this stage of our knowledge, it is fair to state that the potential impact of modern pesticides on sterility in humans has not been clearly demonstrated yet. However, based on the known biological activities of pesticides, they should be considered hazardous chemicals and should be handled with extreme caution.

References

1. Kaloyanova, F.P.; el Batawi, M.A. *Human Toxicity to Pesticides;* CRC Press: Boca Raton, FL, 1991.
2. Pimentel, D. *Ecological Effects of Pesticides on Non-target Species;* U.S. Government Printing Office: Washington, DC, 1971.
3. Gard, N.; Hooper, M. An Assessment of Potential Hazards of Pesticides and Environmental Contaminants. In *Ecology and Management of Neotropical Migratory Birds*; Martin, T., Finch, D., Eds.; Oxford University Press: Oxford, 1995; 294–310.
4. Stinson, E.; Bromely, P. *Pesticides and Wildlife: A Guide to Reducing Impacts on Animals and Their Habitat*; Publication No. 420–004, Virginia Department of Game and Inland Fisheries: Virginia, 1991.
5. Gray, L.E.; Ostby, J.S.; Ferrell, J.M.; Sigman, E.R.; Goldman, J.M. Methoxychlor induces estrogen-like alterations of behavior and the reproductive tract in the female rat and hamster: effects on sex behavior, running wheel activity, and uterine morphology. Toxicol. Appl. Pharmacol. **1988**, *96,* 525–540.
6. Gray, L.E.; Ostby, J.S.; Ferrel, J.M.; Rehnberg, G.; Lindler, R.; Cooper, R.; et al. A dose-response analysis of methoxychlor-induced alterations of reproductive development and function in the rat. Fundam. Appl. Toxicol. **1989**, *12,* 92–108.
7. Eroschenko, V.P. Estrogenic activity of the insecticide chlordecone in the reproductive tract of birds and mammals. J. Toxicol. Environ. Health **1981**, *8,* 731–742.
8. Guzelian, P.S. Comparative toxicology of chlordecone (kepone) in humans and experimental animals. Annu. Rev. Pharmacol. Toxicol. **1982**, *22*, 89–113.
9. Rattner, B.A.; Michael, S.D. Organophosphorous insecticide induced decrease in plasma luteinizing hormone concentration in white-footed mice. Toxicol. Lett. **1985**, *24,* 65–69.
10. Whorton, D.; Milby, T.H.; Krauss, R.M. Testicular function in DBCP exposed pesticide workers. J. Occup. Med. **1979**, *21,* 161–66.
11. Whorton, D.; Krauss, R.M.; Marshall, S. Infertility in male pesticide workers. Lancet **1977**, *ii*, 1259–1261.
12. Goldsmith, J.R. Dibromocholorpropane: epidemiological findings and current questions. Ann. of the New York Academy of Sci. **1997**, *831,* 300–306.
13. Potashnik, G.; Goldsmith, J.; Insler, V. Dibromochloropro-pane-induced reduction of the sex-ratio in man. Andrologia **1984**, *16,* 213–218.
14. Potashnik, G.; Yanai-Inbar, H. Dibromochloropropane: an eight-year re-evaluation of testicular function and reproductive performance. Fertil. Steril. **1987**, *47,* 317–322.
15. Potashnik, G.; Porath, A. Dibromochloropropane: a 17-year reassessment of testicular function and reproductive performance. J. Occup. Environ. Med. **1995**, *37,* 1287–1292.
16. James, W.H. Offspring sex ratio as an indicator of reproductive hazards associated with pesticides. Occup. Environ. Med. **1996**, *52*, 429–430.
17. Strohmer, H.; Boldizsar, A.; Plockinger, B.; Feldner-Busztin, M.; Feichtinger, W. Agricultural work and male infertility. Am. J. Ind. Med. **1993**, *24,* 587–592.
18. Ratcliff, J.M.; Schrader, S.M.; Steenland, K. Semen quality in papaya workers with long term exposure to ethylene dibromide. Br. J. Ind. Med. **1987**, *44*, 317–326.

19. Savitz, D.A.; Arbuckle, T.; Kaczor, D.; Curtis, K.M. Male pesticide exposure and pregnancy outcome. Am. J. Epidemiol. **1997**, *146,* 1025–1036.
20. Larsen, S.B.; Joffe, M.; Bonde, J.P. The asclepiod study group. Time to pregnancy and exposure to pesticides in Danish farmers. Occup. Environ. Med. **1998**, *55*, 278–283.
21. Marshall, S.; Whorton, D.; Krauss, R.M.; Palmer, W.S. Effect of pesticides on testicular function. Urology **1978**, *11*, 257–259.
22. Lebel, G.; Dodin, S.; Ayotte, P.; Marcoux, S.; Ferron, L.A.; Dewailly, E. Organochlorine exposure and the risk of endometriosis. Fertil. Steril. **1998**, *69,* 221–228.
23. Smith, E.M.; Hammonds-Ehlers, M.; Clark, M.K.; Kirchner, H.L.; Fuortes, L. Occupational exposures and risk of female infertility. J. Occup. Environ. Med. **1997**, *39,* 138–147.
24. Sharara, F.I.; Seifer, D.B.; Flaws, J.A. Environmental toxicants and female reproduction. Fertil. Steril. **1998**, *70,* 613–622.

2

Bacillus thuringiensis: Transgenic Crops

Julie A. Peterson,
John J. Obrycki, and
James D. Harwood

Introduction

Genetically modified organisms have been widely adopted in many parts of the world, prompting debate about the implications that this technology may have for environmental health. Transgenic crops have been genetically engineered to incorporate genes derived from another species that confer nutritional and agronomic benefits, such as resistance to insect pests, viruses, herbicides, or environmental conditions, such as low water availability. Among insect- resistant transgenic crops, the most widespread are those that express Bt toxins, coded for by genes from the naturally occurring soil bacterium *Bacillus thuringiensis*. Commercialized Bt crops include corn, cotton, and rice that are protected against Coleoptera and Lepidoptera pests. Bt toxins are recognized as having a narrower range of toxicity than many insecticides, including pyrethroids and neo- nicotinoids, and may therefore pose less risk to non-target organisms; however, potential environmental impacts of Bt toxins need to be examined and documented. This entry will therefore examine the environmental risk assessment of Bt crops, focusing on sources and fate of Bt toxins in exposure pathways for non-target organisms, impact of Bt crops on the environment, and approaches to environmental management of Bt crops.

What Are Bt Crops?

Transgenic Bt crops are genetically engineered to express insecticidal proteins that cause mortality of several common agricultural pests. The genes that code for these proteins, from a naturally occurring bacterium, *Bacillus thuringiensis* (Berliner) (Bacillaceae: Bacillales), are inserted into the genome of the desired crop plant. Genetic transformation is achieved by insertion of the target gene, its promoter and termination sequences, and a marker gene into the crop genome using the microprojectile bombardment method ("gene gun") or the *Agrobacterium tumefaciens* (Smith and Townsend) (Rhizobiales: Rhizobiaceae) bacterium (vector-mediated transformation).

Bt Toxins

Bacillus thuringiensis bacterial strains can produce a series of different toxins; however, only a few have been bioengineered into agricultural crops, including crystalline (Cry) and vegetative insecticidal (VIP) proteins.[1,2] These Bt toxins vary in their range of toxicity to invertebrates, with targeted pests dominated by larval insects in the orders Lepidoptera (moths) and Coleoptera (beetles). The insecticidal mode of action occurs when the Bt toxins bind to receptors on the midgut lining of susceptible insects, causing lysis of epithelial cells on the gut wall and perforations in the midgut lining. This damage to the insect's digestive tract induces cessation of feeding and death by septicemia. An important component of the insecticidal mechanism is its specificity, which is greater than that of many currently used insecticides. Additionally, Bt toxins degrade rapidly in the digestive tract of vertebrates,[3] contributing to their selective nature.

Bt Crops and Their Targeted Pests

Many crop plants have been genetically engineered to express Bt toxins, including field and sweet corn, cotton, potato, rice, eggplant, oilseed rape (canola), tomato, broccoli, collards, chickpea, spinach, soybean, tobacco, and cauliflower. However, only corn and cotton have seen widespread commercialization. Bt potatoes were grown commercially in the United States starting in 1995, but were withdrawn from the market in 2001 following pressure from anti-biotechnology groups and the decision of the global fast-food chain McDonalds to ban the use of genetically modified potatoes in their products.[4] This crop may see a resurgence in planting in Russia and eastern Europe in the near future,[5] as small-scale and subsistence farmers in these regions seek alternatives to expensive insecticide applications.[4] Bt rice has also been approved in certain regions of China,[5] thereby facilitating increased production worldwide.

Global Prevalence

The planting of Bt crops has increased dramatically since the mid-1990s, becoming a prevalent component of agroecosystems worldwide[5–10] (Table 1). For example, Bt cotton and corn in the United States comprised just 1% of total area planted in 1996, their first year of commercial release; however, planting rates have increased rapidly, with areas of Bt cotton and corn in 2010 comprising 73% and 63% of total U.S. production, respectively.[11] Genetically modified crops are grown on 134 million hectares of land in 25 countries by 14.0 million farmers[5]; approximately 40% of that area is planted to corn and cotton expressing Bt insecticidal toxins.[12]

Sources and Fate of Bt Toxins in the Environment

Toxin distribution and expression levels within a transgenic plant vary depending on the type of Bt protein, transformation event, gene promoter used, crop phenology, and environmental and geographical effects.[13–17] Most Bt crops employ a constitutive promoter, such as the cauliflower mosaic virus (CaMV 35S), that expresses insecticidal proteins throughout the life of the plant in nearly all tissues,

TABLE 1 Commercialized Bt Crops, Years Marketed, Bt Toxins Most Commonly Expressed in Commercial Lines, Their Targeted Pests, and Countries That Have Adopted This Technology

Crop	Marketed	Bt Toxins Expressed	Targeted Pest/s	Countries
Corn	1996–present	CrylAb, Cry1A.105, CrylF, Cry2Ab2, Cry9C (withdrawn in 2000), VIP3A	European corn borer *Ostrinia nubilalis* Hübner, southwestern corn borer *Diatraea grandiosella* Dyar (Lepidoptera: Pyralidae), corn earworm *Helicoverpa zea* (Boddie), fall armyworm *Spodoptera frugiperda* Smith (Lepidoptera: Noctuidae)	United States, Brazil, Argentina, Canada, South Africa, Uruguay, Philippines, Spain, Chile, Honduras, Czech Republic, Portugal, Romania, Poland, Egypt, Slovakia
	2003–present	Cry3Bb1, Cry34Ab1, Cry35Ab1, Cry3Aa	Corn rootworm *Diabrotica* spp. (Coleoptera: Chrysomelidae)	
Cotton	1996–present	Cry1Ac, CrylF, Cry2Ab, VIP3A	Bollworm complex: *Heliothis, Helicoverpa* (Lepidoptera: Noctuidae), and *Pectinophora* (Lepidoptera: Gelechiidae)	United States, Brazil, Argentina, India, China, South Africa, Australia, Burkina Faso, Mexico, Colombia, Costa Rica
Potato	1995–2000	Cry3Aa	Colorado potato beetle *Leptinotarsa decemlineata* Say (Coleoptera: Chrysomelidae)	United States, Canada, Romania

Source: Data from James[5] and Duan et al.[146]

which may include foliage, roots and root exudates, phloem, nectar, and pollen, creating the potential for a multitude of sources for environmental exposure. These pathways to exposure of non-target organisms include, but are not limited to, direct consumption of Bt toxins via ingestion of live or detrital plant material, as well as indirect consumption of Bt toxins via soil contamination from root exudates and persistence in the soil, or consumption of Bt-containing prey in tritrophic interactions (Figure 1). These

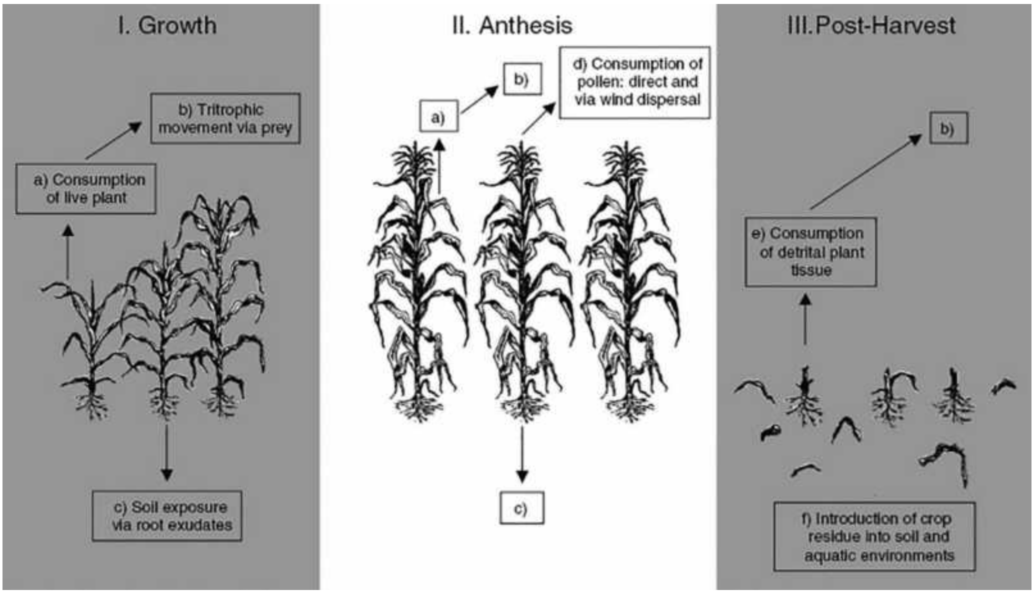


FIGURE 1 Sources for Bt toxin movement in a transgenic corn agroecosystem over the course of a growing season, including (I) growth, (II) anthesis, and (III) post-harvest time periods.

pathways allow for multiple routes to exposure, even potentially within a given taxonomic group, such as ground beetles (Coleoptera: Carabidae), which have been documented to take up Bt toxins in the field.[18] Certain agronomic practices may also create unexpected routes to exposure. For example, following harvest in China, cottonseed hulls may be used as substrate for growing edible oyster mushrooms before being incorporated into cattle feed.[19] Other cotton gin by-products from transgenic plants are used in a variety of ways, including as catfish feed,[20] mulch, and fuel for wood-burning stoves.[21] Although transfer of Bt toxins from cottonseed hulls into mushrooms or cattle feed was not detected,[22] investigation of these complex and non-conventional pathways for Bt toxin movement is critically important.

Direct Consumption of Bt Toxins

Consumption of Live Plant Tissue

Ingestion of plant material, including foliage, roots, phloem, nectar, or pollen may be the most obvious pathway to Bt toxin exposure for targeted pests species, as well as non-target herbivores and natural enemies. Uptake of Bt toxins by herbivores feeding on transgenic plants is well documented (e.g., Dutton et al.,[23] Harwood et al.[24] Meissle et al.,[25] Obrist et al.,[26] and Obrist[27]). However, ingestion of Bt crop tissue may not always result in exposure to toxins. For example, phloem-feeding insects and their honeydew have tested positive for Bt toxins in some transgenic agroecosystems, including certain rice, oilseed rape, and corn events,[28–30] while failing to take up toxins from selected Bt corn events.[31] Exposure pathways of Bt toxins to herbivorous arthropods in transgenic agroecosystems are variable and may therefore be difficult to predict.

Many natural enemies are facultatively phytophagous during some or all of their life stages, consuming plant material or feeding on plant liquids to meet their nutritional and moisture requirements (reviewed in Lundgren).[32] Despite an abundant supply of moisture and prey items, many predatory insects, including ground beetles (Coleoptera: Carabidae), damsel bugs (Hemiptera: Nabidae), stink bugs (Hemiptera: Pentatomidae), and ladybird beetles (Coleoptera: Coccinellidae) will also ingest plant leaf tissue, nectar, or phloem to supplement a prey-based diet.[33]

Pollen Feeding

Another potential route of Bt toxin flow in the environment is through direct pollen feeding or consumption of pollen-contaminated material. Pollen is a component of the diets of many organisms, including spring- tails (Collembola)[34,35] and Western corn rootworms *Diabrotica virgifera virgifera* LeConte (Coleoptera: Chrysomelidae),[36] as well as natural enemies, including ladybird beetles (Coleoptera: Coccinellidae),[37] ground beetles (Coleoptera: Carabidae),[38] green and brown lacewings (Neuroptera: Chrysopidae, Hemerobiidae),[39] hoverflies (Diptera: Syrphidae),[40] and spiders (Araneae).[41,42] In wind-pollinated Bt crops, such as corn, pollen is an abundant resource during anthesis and is deposited in large quantities (up to 1400 grains/cm^2 on plant surfaces[43] and more than 250 grains/cm^2 in ground-based spider webs[44]). Some pollen-feeding omnivores, such as *Orius insidiosus* (Say) (Hemiptera: Anthocoridae), may also maximize their exposure to Bt toxins by aggregating at corn silks and leaf axils, where pollen grains accumulate during anthesis.[45,46] Pollen consumption can therefore represent a significant direct and indirect (through consumption of pollen-feeding prey) route of exposure for both predators and prey in transgenic agroecosystems, particularly during periods of crop anthesis.

Consumption of Detritus

Bt toxins can persist in plant detritus beyond a single growing season[47,48] thereby exposing detritivores, such as earthworms, slugs, nematodes, protozoa, bacteria, and fungi, to Bt toxins through the consumption of such litter.[49–51] Crop detritus may also enter aquatic environments; for example, in agricultural systems where crop detritus is left in the field to prevent erosion, plant residues may account for up to 40% of non-woody vegetation entering streams.[52] Bt-containing crop tissue may then be consumed by aquatic detritivores, such as larval caddis- flies (Trichoptera), crane flies (Diptera: Tipulidae), midges

(Diptera: Chironomidae), and isopods. However, the bioactivity of Bt toxins in senesced plant material may be relatively short; lepidopteran-specific toxins were absent after 2 weeks in aquatic systems, while coleopteran-specific toxins decayed in as few as 6 days.[53] The harsh environmental conditions and constant physical abrasion experienced by plant tissue in flowing water were suggested as mechanisms stimulating such rapid breakdown.[54] Thus, while detritus provides a potential route of exposure, the functional consequence of Bt toxins in detritivore food webs remains unclear. However, what is evident is the persistence of Bt toxins in the environment following harvest and the possibility for long-term exposure of non-target organisms to this material.

Indirect Consumption of Bt Toxins

Soil Contamination via Root Exudates

One potential pathway of indirect exposure to Bt toxins is through contamination of the soil and therefore to soildwelling arthropods via root exudates. Bt corn, potato, and rice all release transgenic proteins from their roots during plant growth.[55,56] The soil-dwelling fauna, including beneficial non-target organisms, may therefore be exposed to Bt toxins via their secretion in plant root exudates. Bt toxin exposure to epigeal predators, ground beetle larvae and adults, and certain spiders [e.g., wolf spiders (Araneae: Lycosidae)] may also occur because of their feeding habits. Several studies have quantified the persistence of Bt toxins in the soil,[47,57–59] with results indicating persistence of these insecticidal proteins ranging from 2 to 32 weeks after introduction into the soil. This wide discrepancy may partially reflect differences in microbial activity of soils,[57,60,61] which is in turn affected by pH and mineral content.[59] Bt toxins may bind to humic acids, organic supplements, or soil particles, protecting the toxins from degradation by microbes and extending the persistence of insecticidal activity in the soil.[2] Thus, the persistence of Bt toxins may vary significantly due to their differential rate of decay based on microbial activity, soils, and environmental factors.

Consumption of Prey Containing Bt Toxins

The movement of Bt toxins from plant tissue into herbivores and subsequently into their natural enemies has been well documented. Concentrations of Bt toxins typically decrease as they move through a food chain, indicating little evidence for bioaccumulation effects as seen in other insecticidal compounds[62]; however, two-spotted spider mites, *Tetranychus urticae* Koch (Acari: Tetranychidae), show evidence for the bioaccumulation of Bt toxins.[63] Although in a more typical example, CrylAc proteins expressed in transgenic cotton are ingested by beet armyworm caterpillars *Spodoptera exigua* (Hübner) (Lepidoptera: Noctuidae) and are also detectible, but at lower concentrations, in predatory stink bugs *Podisus maculiventris* (Say) (Hemiptera: Pentatomidae) when these prey are consumed.[63] However, not all tritrophic pathways facilitate the uptake of Bt toxins; CrylAb toxins are present in the marsh slug *Deroceras laeve* (Müller) (Pulmonata: Agriolimacidae) following consumption of Bt corn tissue, but are not taken up by the predatory ground beetle *Scarites subterraneus* (F.) (Coleoptera: Carabidae) in laboratory studies[64]; accordingly, field- collected specimens of this species did not test positive for Bt toxins.[18] Additionally, the concentration of Bt toxins transferred via trophic connections may vary based on the identity of the prey. In a laboratory experiment, two prey species of the wolf spider *Pirata subpiraticus* (Bösenberg and Strand) (Araneae: Lycosidae), the striped stem borer *Chilo suppressalis* (Walker) (Lepidoptera: Crambidae), and the Chinese brushbrown caterpillar *Mycalesis gotama* Moore (Lepidoptera: Nymphalidae) were allowed to feed on transgenic rice expressing Cry1Ab Bt toxins. These prey were subsequently fed to the wolf spider, and assays of each trophic level indicated that Bt toxins were transferred up the food chain; Cry1Ab concentration diminished with each additional trophic step, and the two prey species transferred Cry1Ab with significantly different efficiencies, having approximately 60 times the Cry1Ab concentration in brushbrown caterpillar-fed spiders compared with striped stem borer-fed spiders.[65] Adult ladybird beetles (Coleoptera: Coccinellidae) showed greatest uptake of Bt toxins in a corn agroecosystem post-anthesis, indicating that tritrophic movement of toxins was a greater pathway for toxin uptake than direct pollen consumption.[66] It is therefore clear that consumption of Bt-containing prey could be a major source of Bt toxin flow in non-target food webs, although the extent of toxin uptake and its concentration will depend on the strength of specific trophic pathways that occur within a given food web in the field.

Impacts of Bt Crops on the Environment

Bt crops have become a dominant fixture in selected agroecosystems worldwide. Their planting on cultivated lands globally allows for large potential impacts of this technology on the environment. These impacts include both benefits and potential risks, the consequences to the environment of using Bt technology are intensely debated.

Benefits of Bt Technology

Reduced Risk Compared with Conventional Insecticides

The insecticidal toxins produced by transgenic Bt crops are considered to have fewer non-target effects than many insecticides due to their narrow range of toxicity and, therefore, to be advantageous to traditional methods of control. For example, populations of many natural enemies responded negatively to foliar applications of broad-spectrum pyrethroids compared with more selective insecticides, such as Bt toxins, indoxacarb, and spinosad, used to combat lepidopteran pests in sweet corn agroecosystems.[67] Field studies comparing Bt crops with their non-transgenic isolines that have been treated with broad-spectrum insecticides almost always reveal higher populations of beneficial arthropods in the Bt crops. A meta-analysis of these studies found that total non-target invertebrate abundance was higher in lepidopteran-targeting corn and cotton compared with non-transgenic crops managed with insecticides; however, no differences for coleopteran-targeting corn were reported.[68] Non-transgenic control plots treated with insecticides had lower predator and herbivore abundance compared with unsprayed Bt fields; this result was particularly strong for predator populations in non-transgenic plots treated with pyrethroids, such as lambda-cyhalothrin, cyfluthrin, and bifenthrin.[69] Similarly, spiders were more abundant in Bt corn, cotton, and potato when compared with conventionally managed crops employing a range of insecticides, including foliar pyrethroid sprays, systemic neonicotinoid seed treatments, and organophosphate soil applications at planting.[70] Due to their selectivity, Bt crops are therefore safer for non-target organisms when compared with many insecticides, particularly those with broad-spectrum action.

Economic Savings

A reduction in the quantity and frequency of insecticide applications are economically beneficial, in addition to reduced exposure to chemical insecticides for farm workers and the environment. Bt cotton has significantly reduced insecticide inputs in numerous regions of the world, including the United States,[71,72] China,[73] and South Africa.[74] The adoption of Bt corn in the midwestern United States has provided an estimated $6.9 billion in benefits to growers of both Bt and non-Bt corn in the past 14 years, due to area-wide suppression of European corn borer *Ostrinia nubilalis* (Hübner) (Lepidoptera: Pyralidae), a key pest of this crop.[75] With more than 53 million hectares of Bt crops now planted worldwide, there are significant economic considerations, and it is evident that Bt- based production systems are not only more sustainable in the context of pest management but also have the capacity to enhance agricultural diversity through reduced chemical inputs.

Global Food Security

The human population is projected to reach 10 billion by 2050, and concomitant to this is the need for augmented global food security and production.[76] The employment of Bt crop technology may aid in this goal by increasing quantity and consistency of crop yields; for example, corn yields are increased or protected because of season-long control of European corn borer.[71] Additionally, stored corn grain is protected against lepidopteran pests[77] and mycotoxin levels, which pose a threat to the health of humans and livestock if introduced into the food supply,[78] are lower because of reduced feeding activity of European corn borer, which are associated with the fungal causal agents.[71,79] Bt crops may therefore confer significant beneficial effects for the global drive to increase agricultural productivity and safety.

Potential Risks of Bt Crops

Impacts on Non-Target Organisms

Despite the specificity of Bt toxins toward target pests, questions have been raised concerning their effects on abundance, diversity, or fecundity of some components of the non-target food web, including beneficial species such as pollinators, natural enemies, and/or detritivores. Given the important ecosystem services provided by the abovementioned non-target organisms, the risk assessment of these groups is essential in the context of understanding environmental health. Lundgren et al.[17] identified four main approaches that risk assessment researchers have used to study the impact of Bt crops on non-target invertebrates: direct toxicity, tritrophic interactions, community level studies, and meta-analyses of data.

Direct toxicity. Feeding non-target organisms a diet that contains Bt toxins and measuring resulting parameters of development, fitness, and fecundity are done to examine the potential for directly toxic effects of Bt crops. The literature (reviewed in Lundgren et al.[17] and Lövei and Arpai[80]) provides contrasting evidence of non-target effects, ranging from no discernable effects of consumption of transgenic crops (e.g., Harwood et al.,[64] Pilcher et al.,[81] Al-Deeb et al.,[82] Lundgren and Wiedenmann,[83] and Anderson et al.[84]) to reports of a variety of negative effects (e.g., increased mortality, delay in development, reduction in weight gain, or changes in behavior) on beneficial organisms, such as pollinators,[85] predators,[86] parasitoids,[87] and other non-target arthropods.[50,88–91] Differing results of studies of direct toxicity of Bt toxins to non-target organisms exist for many groups; for example, in caddisflies (closely related to the target order Lepidoptera), studies have been published that report both sublethal negative effects[91] and the absence of negative impacts of Bt toxins.[54] Such laboratory toxicity studies may be extrapolated to the field, although toxicity studies should address all ecologically relevant routes to exposure for non-target organisms.[92]

Tritrophic interactions. These studies test for effects of Bt crops on natural enemies via consumption of Bt-containing prey; any observed effects may be due to ingestion of toxins or through prey-quality-mediated effects. Several studies have reported no tritrophic effects of Bt crops on natural enemies[63,93–95]; however, negative effects have been observed in other cases,[96,97] although these results are often attributed to prey-mediated effects whereas prey quality is lower when fed Bt crop tissue. Meta-analyses of tritrophic studies revealed that using prey items that were totally or partially susceptible to Bt toxins (and therefore were likely to be of lesser quality) had a negative effect on the performance of natural enemies, while using nonsusceptible prey (whose quality should be unaffected by consuming Bt toxins) had no effect on the performance of the natural enemies that consumed it.[98,99]

Community level. To study the effect of Bt crops on non-target organisms at the community level, arthropods are sampled from Bt and conventional crops to observe any differences in abundance, diversity, or community structure. Such studies have examined a variety of nontarget organism populations, including soil microarthropods, nematodes, decomposers, pollinators, and natural enemies.[81,100–110] Results of such studies often report no significant differences between populations of non-target organisms in Bt and non-Bt crops; however, a lack of taxonomic resolution in some studies can weaken these results.[70]

Meta-analysis data. This quantitative method addresses effects of Bt crops across multiple published studies and has been widely used to infer the consequence of Bt crops on a series of different parameters. For example, a meta-analysis of 42 field experiments revealed that the overall mean abundance of non-target invertebrates was significantly lower in lepidopteran-targeting Bt corn fields compared with nontransgenic fields when neither is treated with insecticides; no differences were found between coleopteran-targeting Bt and non-transgenic corn.[68] Unsurprisingly, the abundance of non-target arthropods was significantly higher in transgenic corn versus non-transgenic corn that had been treated with insecticides.[68] Additional meta-analyses have reported the effects of Bt crops on functional guilds of non-target organisms,[69] honey bees,[111] and spiders,[70] generally finding no differences in non-target arthropod populations between Bt and non-Bt crops. When examined at further taxonomic resolution, such analyses may reveal differential responses of functionally distinct taxa to Bt crops, as is the case with spider families. Meta-analysis revealed positive effects of Bt crops on the abundance of certain groups (Clubionidae, Linyphiidae, Thomisidae), no effect on others (Lycosidae, Oxyopidae, Araneidae), and negative effects on several families (Anyphaenidae, Philodromidae) relative to non-transgenic crops untreated with insecticides.[70]

Presence in Human Food Supply

Concerns about the presence of Bt toxins in the human food supply do not stem from any direct toxic effects, as vertebrates lack the midgut receptors for binding of Bt toxins, but from the possibility that a portion of the population will exhibit an allergic reaction to ingestion of Bt proteins.[112] Most Bt toxins will readily break down in the acidic environment of a vertebrate digestive tract.[3] Bt corn expressing Cry9C proteins, marketed under the commercial name StarLink™, was planted in the United States from 1998 to 2000, but approved only for animal feed and ethanol production due to the persistence of Cry9C in the vertebrate gut.[113] When traces of Cry9C proteins were found in cornmeal destined for human consumption, several food items were recalled, including Taco Bell® taco shells, and StarLink was voluntarily removed from the market.[114] However, no confirmed allergenic reactions due to Cry9C contamination were reported. Despite the lack of evidence for any true risk to humans based on consumption of Bt food products, sentiment against transgenic agricultural products destined for human consumption exists, especially in Europe, and has influenced the commercial acceptance of some products such as Bt potatoes.[4] Therefore, despite these limited effects on the human (and vertebrate) population, safeguards need to be in place to prevent the presence of unapproved genetically modified products entering the human food chain.

Pleiotropic Effects of Genetic Transformation

Insertion of a Bt gene complex into a crop plant may result in unpredicted and unintended pleiotropic effects that change the plant from its non-transgenic counterpart in ways beyond just the expression of Bt toxins.[49,115–117] For example, a reported pleiotropic effect in Bt corn is an increase in the lignin content in transgenic plant tissue,[49] a trait that could lead to reduced decomposition rates in the soil.[118] However, other studies have contested this conclusion and shown no differences in rate of decomposition.[119] An additional pleiotropic effect of transformation in Cry1F corn may be an increase in attractiveness as an oviposition site for corn leafhoppers *Dalbulus maidis* (DeLong and Wolcott) (Hemiptera: Cicadellidae), a pest that is not targeted by Bt toxins, possibly due to altered plant traits that influence oviposition, such as leaf vein characteristics, foliar pubescence, or plant chemistry.[120] There is a lack of understanding of how these pleiotropic effects will affect ecosystem processes, although the potential consequences merit further examination in the context of their environmental impacts.

Gene Escape

The transfer of genes from populations of domesticated crops into wild plants has been documented for many years.[121] The "escape" of Bt transgenes into wild plants could have undesirable effects by reducing genetic diversity and fecundity in wild populations or increasing fecundity and creating an invasive weed through reduction or elimination of herbivory. The presence of transgenic material from the CaMV 35S promoter used in Bt crops was reported in native maize landraces grown in remote areas of Oaxaca, Mexico, in 2001.[122] However, these results have been highly debated[123–125] and additional studies are conflicting, reporting both the presence[126] and absence[127] of transgenic DNA in traditional maize lines in Mexico. Additionally, transgene escape into weedy rice may increase the fecundity of this plant, as well as its ecological interactions with surrounding organisms.[128] The implications of transgene escape are yet to be fully understood, particularly in the context of ecological risk assessment.

Approaches to Environmental Management of Bt Crops

To safely incorporate Bt crop technology into agroecosystems, approaches to environmental management should address the issue at multiple scales. These include engineering at the level of the individual plant genome, field- and farm-level modifications to reduce exposure of Bt toxins and escape of transgenes, and large-scale incorporation of Bt technology into integrated pest and resistance management programs. Finally, continued research concentrating on the non-target impacts of Bt crops should be conducted at multiple tiers across crop and toxin types, geographic regions, non-target organism taxa, and temporal and spatial scales, studying non-target organisms at the greatest taxonomic resolution possible. Regulation of transgenic crops in the effort to mitigate risk is complex; further recommendations and discussion of this topic can be found elsewhere.[129]

Within-Plant Modifications

Selection of Low-Risk Promoters

As the gene promoter used in a transgenic event can have a strong impact on the eventual concentration and distribution of Bt toxins within the plant, the choice of promoter should be made within the context of environmental safety. Certain promoters have been identified as having greater non-target risks than others; for example, harmful effects of Bt corn event 176 on non-target Lepidoptera larvae [monarchs *Danaus plexippus* (L.) (Nymphalidae) and black swallowtails *Papilio polyxenes* F. (Papilionidae)] have been reported, while other events expressing the same Cry toxin (e.g., Bt11 and MON810) had no effect. 90] Event 176 has increased expression of Bt proteins in the pollen compared with the other events[130] and therefore poses a greater risk to non-target organisms.

Tissue- and Time-Specific Expression

The use of gene promoters that are tissue- or time-specific to express toxins only in plant tissues when they are susceptible to feeding has been introduced.[131] This technique has been employed in the transgenic expression of snowdrop lectin, a plant-derived protein with insecticidal properties, in rice. To target phloem-feeding pests such as brown planthoppers, lectins are selectively expressed in the vascular tissue.[132,133] Such selective expression of Bt toxins in tissue and time to target susceptible pests and reduce exposure to non-target beneficial arthropods could potentially increase environmental safety, thereby reducing the pathways for Bt toxin movement through non-target food webs.

Reduction in Bt Toxin and Transgene Escape

At the field or farm level, management practices may be implemented that reduce the movement of Bt toxins or transgenes from their source (Bt plants) into surrounding habitats. Current practices may depend on the crop and agronomic aims of the grower; for example, large quantities of crop residue may be incorporated into the soil during the harvesting process, although this is not the practice when crop material is removed for ethanol production or under reduced-tillage practices.[134] Although Bt toxins may degrade quickly following the incorporation of Bt crop plant detritus into aquatic systems, this potential pathway for transgenic protein movement may be avoided through the employment of practices that prevent movement of transgenic crop tissue beyond field borders. The establishment of riparian buffer zones and filter strips may reduce the quantity of crop detritus and other compounds originating in cropland (e.g., fertilizers, insecticides) that enter nearby streams and waterways.[135]

Similarly, reducing exposure pathways for gene flow into wild plant populations via physical methods, such as isolation of crops or plant destruction, may delay transgene escape. However, controlling gene flow via pollen and seeds in the environment can be very difficult; a physical separation of 200 m between transgenic corn still yields contamination levels of 0.1% between plant populations due to cross-pollination.[136] Seeds are additionally difficult to control owing to their persistence in the soil seed bank, as well as ability to sometimes germinate and persist outside of cultivated fields.[137] Management of the movement of Bt toxins and genetic material from cultivated fields into the surrounding environments warrants additional research. Interestingly, technology that could have reduced the spread of transgenes, called the "terminator gene," was abandoned in 1999 because of the criticism that the gene prevents farmers from harvesting viable seed and thereby exclusively benefits the seed companies.[138]

Large-Scale Integrated Pest and Resistance Management

Although Bt crops allow reductions in the application of certain insecticides compared with conventionally managed crops (while other insecticidal practices persist, such as neonicotinoid seed treatments on corn[139]), it should not be assumed that this technology will readily fit into integrated pest management practices.[17] Considerations of compatibility with biological control and delaying resistance in pest populations are also necessary.

Compatibility with Biological Control

Integrated pest management practices attempt to incorporate mechanical, physical, chemical, and cultural controls; host resistance (including transgenic crops); and auto-cidal, biochemical, and biological controls in a synergistic manner. Increased attention has focused on conservation biological control: the modification of the environment or existing practices to protect and enhance specific natural enemies or other organisms to reduce effects of pests (e.g., Landis et al.[140] and Eilenberg et al.[141]). Natural enemies can be abundant in agricultural systems and often play an essential role in pest suppression. Maintenance of relevant natural enemy populations via conservation biological control is a practical and sustainable option for high-acreage field crops, such as corn and cotton,[142] which are dominated by Bt varieties. Any negative effects of Bt toxins on natural enemies could reduce their effectiveness as biological control agents and therefore limit natural pest suppression in agroecosystems. Understanding the potential impacts of transgenic crops on non-target arthropods is essential in order to provide a framework for integrating natural enemies into sustainable methods of pest control in the agricultural environment.

Resistance Management Techniques

The development of resistance to Bt toxins by pest populations is a major concern. Integrated resistance management programs must continue to be developed and followed to promote the sustainable use of Bt crops. This is of critical importance given that resistance to Bt sprays has occurred in multiple populations of the pestiferous diamond back moth *Plutella xylostella* (L.) (Lepidoptera: Plutellidae)[143] and three instances of field-evolved resistance to transgenic Bt crops have been reported in moth larvae: *Spodoptera frugiperda* (Smith) (Lepidoptera: Noctuidae) to Cry1F corn in Puerto Rico, *Busseola fusca* (Fuller) (Lepidoptera: Noctuidae) to Cry1Ab corn in South Africa, and *Pectinophora gossypiella* (Saunders) (Lepidoptera: Gelechiidae) to Cry1Ac in the southwestern United States.[144] Current resistance management employs structured refuges and high-dose toxin crops, as well as monitoring for resistance development in the field and monitoring for compliance of growers to refuge protocol. Additional attempts to delay resistance include creating transgenic plants that express more than one type of Bt toxin that targets the same pest, called gene pyramiding.[131] Improved resistance management would include increased education for growers and the public about the importance of resistance management and refuge compliance, as well as continued monitoring of field populations for the development of resistance. Future strategies to passively achieve resistance compliance include mixed seed refuges, in which transgenic and non-transgenic seeds are sold in combination within the bag.[145]

Conclusions

The sources and fate of Bt toxins in the environment can be complex and variable depending on crop, transgenic event, geography, and other environmental variables. The effects of Bt crops and their toxins on the environment have been widely debated, particularly the potential implications associated with ecological impacts such as gene escape and non-target risks. Approaches to the environmental management of Bt crops and their integration into integrated pest and resistance management systems warrant further study. Despite the concerns associated with Bt crops, significant reductions in chemical input are evident and this technology is environmentally safer when compared with many approaches to pest suppression, particularly those using broad-spectrum insecticides.

Future of Bt Technology

The focus of current transgenic technology has been on stacking and pyramiding of events. Stacking incorporates multiple transgenic traits into the crop genome in order to express more than one type of insecticidal toxin, therefore targeting multiple pest species. Pyramiding of transgenes allows for the crop to express multiple types of Bt toxins that target the same pest. Additionally, several other Bt crops are expected to be approved for commercial availability by 2015, including potatoes for planting in eastern Europe and eggplant in India.[5] The global adoption of biotechnology in agriculture is projected to continue with estimates that genetically modified crops will reach 200 million hectares, grown by 20 million farmers in 40 countries by 2015.[5]

Acknowledgments

Funding for this project was provided by the U.S. Department of Agriculture Cooperative State Research Education and Extension Service Biotechnology Risk Assessment Grant 2006-39454-17446. JDH is supported by the University of Kentucky Agricultural Experiment Station State Project KY008043. This is publication number 10-08-132 of the University of Kentucky Agricultural Experiment Station.

References

1. Yu, C.G.; Mullins, M.A.; Warren, G.W.; Koziel, M.G.; Es-truch, J.J. The *Bacillus thuringiensis* vegetative insecticidal protein Vip3 lyses midgut epithelium cells of susceptible insects. Appl. Environ. Microbiol. **1997**, *63,* 532–536.
2. Glare, T.R.; O'Callaghan, M. *Bacillus thuringiensis: Biology, Ecology, and Safety*; John Wiley & Sons, Ltd.: West Sussex, U.K., 2000.
3. Mendehlson, M.; Kough, J.; Vaituzis, Z.; Matthews, K. Are *Bt* crops safe? Nat. Biotechnol. **2003**, *21* (9), 1003–1009.
4. Kaniewski, W.K.; Thomas, P.E. The potato story. AgBio- Forum **2004**, *7,* 41–46.
5. James, *C. Global Status of Commercialized Biotech/GM Crops: 2009;* ISAAA Brief No. 41. ISAAA: Ithaca, NY, 2009.
6. Cannon, R.J.C. Bt transgenic crops: Risks and benefits. In- tegr. Pest Manag. Rev. **2000**, *5,* 151–173.
7. Pray, C.E.; Huang, J.; Hu, R.; Rozelle, S. Five years of Bt cotton in China—The benefits continue. Plant J. **2002**, *31,* 423–430.
8. Shelton, A.M.; Zhao, J.Z.; Roush, R.T. Economic, ecological, food safety, and social consequences of the deployment of Bt transgenic plants. Annu. Rev. Entomol. **2002**, *47,* 845–881.
9. Lawrence, S. AgBio keeps on growing. Nat. Biotechnol. **2005**, *63,* 3561–3568.
10. James, C. *Executive Summary of Global Status of Commercialized Biotech/GM Crops: 2006*; ISAAA Brief No. 36. ISAAA: Ithaca, NY, 2006.

11. United States Department of Agriculture National Agricultural Statistics Service. Acreage report, 2010, available at http://usda.mannlib.cornell.edu/usda/current/Acre/Acre-06-30-2010.pdf. (Accessed, November 2010).
12. GMO Compass. *Rising Trend: Genetically Modified Crops Worldwide on 125 Million Hectares*, 2009, available at http://www.gmocompass.org/eng/agri_biotechnology/gmo_planting/257.global_gm_planting_2008.html.
13. Fearing, P.L.; Brown, D.; Vlachos, D.; Meghji, M.; Privalle, L. Quantitative analysis of CryIA(b) expression in *Bt* maize plants, tissues, and silage and stability of expression over successive generations. Mol. Breed. **1997**, *3* (3), 169–176.
14. Duan, J.J.; Head, G.; McKee, M.J.; Nickson, T.E.; Martin, J.H.; Sayegh, F.S. Evaluation of dietary effects of transgenic corn pollen expressing Cry3Bb1 protein on a nontarget ladybird beetle, *Coleomegilla maculata*. Entomol. Exp. Appl. **2002**, *104,* 271–280.
15. Grossi-de-Sa, M.F.; Lucena, W.; Souza, M.L.; Nepomu- ceno, A.L.; Osir, E.O.; Amugune, N.; Hoa, T.T.C.; Hai, T.N.H.; Somers, D.A.; Romano, E. Transgene expression and locus structure of Bt cotton. In *Environmental Risk Assessment of Genetically Modified Organisms. Methodologies for Assessing Bt Cotton in Brazil*; Hilbeck, A., Andow, D.A., Fontes, E.M.G., Eds.; CAB International: Wallingford, U.K., 2006; 93–107.
16. Obrist, L.B.; Dutton, A.; Albajes, R.; Bigler, F. Exposure of arthropod predators to Cry1Ab toxin in Bt maize fields. Ecol. Entomol. **2006**, *31* (2), 143–154.
17. Lundgren, J.G.; Gassman, A.J.; Bernal, J.; Duan, J.J.; Ru- berson, J. Ecological compatibility of GM crops and biological control. Crop Prot. **2009**, *28,* 1017–1030.
18. Peterson, J.A.; Obrycki, J.J.; Harwood, J.D. Quantification of Bt-endotoxin exposure pathways in carabid food webs across multiple transgenic events. Biocontrol Sci. Technol. **2009**, *19* (6), 613–625.
19. Li, X.; Pang, Y.; Zhang, R. Compositional changes of cotton-seed hull substrate during *P. ostreatus* growth and the effects on the feeding value of spent substrate. Bioresour. Technol. **2001**, *80,* 157–161.
20. Li, M.H.; Hartnell, G.F.; Robinson, E.H.; Kronenberg, J.M.; Healy, C.E.; Oberle, D.F.; Hoberg, J.R. Evaluation of cottonseed meal derived from genetically modified cotton as feed ingredients for channel catfish, *Ictalurus punctatus*. Aquac. Nutr. **2008**, *14* (6).
21. Robertson, R. *Cotton Gin Trash Now Valuable By-Product,* Southeast Farm Press, 2009, available at http://southeastfarmpress.com/grains/cotton-gin-trash-now-valuable-product.
22. Jiang, L.; Tian, X.; Duan, L.; Li, Z. The fate of Cry1Ac Bt toxin during oyster mushroom (*Pleurotus ostreatus*) cultivation on transgenic Bt cottonseed hulls. J. Sci. Food Agr. **2008**, *88,* 214–217. Accessed November 1, 2010.
23. Dutton, A.; Klein, H.; Romeis, J.; Bigler, F. Uptake of Bt-toxin by herbivores feeding on transgenic maize and consequences for the predator *Chrysoperla carnea.* Ecol. Entomol. **2002**, *27* (4), 441–447.
24. Harwood, J.D.; Wallin, W.G.; Obrycki, J.J. Uptake of Bt endotoxins by nontarget herbivores and higher order arthropod predators: Molecular evidence from a transgenic corn agroecosystem. Mol. Ecol. **2005**, *14* (9), 2815–2823.
25. Meissle, M.; Vojtech, E.; Poppy, G.M. Effects of Bt maize- fed prey on the generalist predator *Poecilus cupreus* L. (Cole- optera: Carabidae). Transgenic Res. **2005**, *14* (2), 123–132.
26. Obrist, L.B.; Klein, H.; Dutton, A.; Bigler, F. Effects of Bt maize on *Frankliniella tenuicornis* and exposure of thrips predators to prey-mediated Bt toxin. Entomol. Exp. Appl. **2005**, *115* (3), 409–416.
27. Obrist, L.B.; Dutton, A.; Romeis, J.; Bigler, F. Biological activity of Cry1Ab toxin expressed by Bt maize following ingestion by herbivorous arthropods and exposure of the predator *Chrysoperla carnea.* BioControl **2006**, *51* (1), 31–48.
28. Raps, A.; Kehr, J.; Gugerli, P.; Moar, W.J.; Bigler, F.; Hilbeck, A. Immunological analysis of phloem sap of *Bacillus thuringiensis* corn and of the nontarget herbivore *Rhopalo- siphum padi* (Homoptera: Aphididae) for the presence of Cry1Ab. Mol. Ecol. **2001**, *10* (2), 525–533.

29. Bernal, C.C.; Aguda, R.M.; Cohen, M.B. Effect of rice lines transformed with *Bacillus thuringiensis* toxin genes on the brown planthopper and its predator *Cyrtorhinus lividipen-nis.* Entomol. Exp. Appl. **2002**, *102* (1), 21–28.
30. Burgio, G.; Lanzoni, A.; Accinelli, G.; Dinelli, G.; Bonetti, A.; Marotti, I.; Ramilli, F. Evaluation of Bt-toxin uptake by the non-target herbivore, *Myzus persicae* (Hemiptera: Aphi- didae), feeding on transgenic oilseed rape. Bull. Entomol. Res. **2007**, *97* (2), 211–215.
31. Head, G.; Brown, C.R.; Groth, M.E.; Duan, J.J. Cry1Ab protein levels in phytophagous insects feeding on transgenic corn: Implications for secondary exposure risk assessment. Entomol. Exp. Appl. **2001**, *99* (1), 37–45.
32. Lundgren, J.G. *Relationships of Natural Enemies and Non-prey Foods*; Springer International: Dordrecht, the Netherlands, 2009.
33. Hagen, K.S.; Mills, N.J.; Gordh, G.; McMurtry, J.A. Terrestrial arthropod predators of insect and mite pests. In *Handbook of Biological Control;* Bellows, T.S., Fisher, T.W., Eds.; Academic Press: San Diego, CA, 1999; 383–504.
34. Kevan, P.G.; Kevan, D.K.M. Collembola as pollen feeders and flower visitors with observations from the high Arctic. Quaest. Entomol. **1970**, *6,* 311–326.
35. Chen, B.; Snider, R.J.; Snider, R.M. Food consumption by Collembola from northern Michigan deciduous forest. Pedobiologia **1996**, *40,* 149–161.
36. Kim, J.H.; Mullin, C.A. Impact of cysteine proteinase inhibition in midgut fluid and oral secretion on fecundity and pollen consumption of western corn rootworm *(Diabrotica virgifera virgifera).* Arch. Insect Biochem. **2003**, *52,* 139–154.
37. Lundgren, J.G.; Wiedenmann, R.N. Nutritional suitability of corn pollen for the predator *Coleomegilla maculata* (Coleoptera: Coccinellidae). J. Insect Physiol. **2004**, *50* (6), 567–575.
38. Larochelle, A.; Larivière, M.C. *A Natural History of the Ground-Beetles (Coleoptera: Carabidae) of America North of Mexico*; Pensoft: Sofia, Bulgaria, 2003.
39. Canard, M. Natural food and feeding habits of lacewings. In *Lacewings in the Crop Environment;* McEwen, P.K., New, T.R., Whittington, A.E., Eds.; Cambridge University Press: Cambridge, U.K. 2001; 116–129.
40. Olesen, J.M.; Warncke, E. Predation and potential transfer of pollen in a population of *Saxifraga hirculus*. Holarctic Ecol. **1989**, *12,* 87–95.
41. Smith, R.B.; Mommsen, T.P. Pollen feeding in an orbweaving spider. Science **1984**, *226* (4680), 1330–1332.
42. Ludy, C. Intentional pollen feeding in the garden spider *Araneus diadematus.* Newsl. Br. Arachnol. Soc. **2004**, *101,* 4–5.
43. Pleasants, J.M.; Hellmich, R.L.; Dively, G.P.; Sears, M.K.; Stanley-Horn, D.E.; Mattila, H.R.; Foster, J.E.; Clark, P.; Jones, G.D. Corn pollen deposition on milkweeds in and near cornfields. Proc. Natl. Acad. Sci. U. S. A. **2001**, *98,* 11919–11924.
44. Peterson, J.A.; Romero, S.A.; Harwood, J.D. Pollen interception by linyphiid spiders in a corn agroecosystem: Implications for dietary diversification and risk-assessment. Arthropod Plant Interact. **2010**, *4* (4), 207–217.
45. Isenhour, D.J.; Marston, N.L. Seasonal cycles of *Orius in- sidiosus* (Hemiptera: Anthocoridae) in Missouri soybeans. J. Kans. Entomol. Soc. **1981**, *54,* 129–142.
46. Coll, M.; Bottrell, D.G. Microhabitat and resource selection of the European corn-borer (Lepidoptera, Pyralidae) and its natural enemies in Maryland field corn. Environ. Entomol. **1991**, *20*, 526–533.
47. Zwahlen, C.; Hilbeck, A.; Gugerli, P.; Nentwig, W. Degradation of the Cry1Ab protein within transgenic *Bacillus thuringiensis* corn tissue in the field. Mol. Ecol. **2003**, *12* (3), 765–775.
48. Zwahlen, C.; Andow, D.A. Field evidence for the exposure of ground beetles to Cry1Ab from transgenic corn. Environ. Biosafety Res. **2005**, *4* (2), 113–117.

49. Saxena, D.; Stotzky, G. *Bacillus thuringiensis* (Bt) toxin released from the root exudates and biomass of Bt corn has no apparent effect on earthworms, nematodes, protozoa, bacteria, and fungi in soil. Soil Biol. Biochem. **2001**, *33* (9), 1225–1230.
50. Zwahlen, C.; Hilbeck, A.; Howald, R.; Nentwig, W. Effects of transgenic Bt corn litter on the earthworm *Lumbricus ter- restris.* Mol. Ecol. **2003**, *12* (4), 1077–1086.
51. Harwood, J.D.; Obrycki, J.J. The detection and decay of Cry1Ab Bt-endotoxins within non-target slugs, *Deroceras reticulatum* (Mollusca: Pulmonata), following consumption of transgenic corn. Biocontrol Sci. Technol. **2006**, *16* (1), 77–88.
52. Stone, M.L.; Whiles, M.R.; Webber, J.A.; Williard, K.W.J.; Reeve, J.D. Macroinvertebrate communities in agriculturally impacted southern Illinois streams: Patterns with riparian vegetation, water quality, and in-stream habitat quality. J. Environ. Qual. **2005**, *34,* 907–917.
53. Prihoda, K.R.; Coats, J.R. Aquatic fate and effects of *Bacillus thuringiensis* Cry3Bb1 protein: Toward risk-assessment. Environ. Toxicol. Chem. **2008**, *27* (4), 793–798.
54. Jensen, P.D.; Dively, G.P.; Swan, C.M.; Lamp, W.O. Exposure and nontarget effects of transgenic Bt corn debris in streams. Environ. Entomol. **2010**, *39* (2), 707–714.
55. Saxena, D.; Stotzky, G. Insecticidal toxin from *Bacillus thuringiensis* is released from roots of transgenic Bt corn in vitro and in situ. FEMS Microbiol. Ecol. **2000**, *33* (1), 35–39.
56. Saxena, D.; Stewart, C.N.; Altosaar, I.; Shu, Q.Y.; Stotzky, G. Larvicidal Cry proteins from *Bacillus thuringiensis* are released in root exudates of transgenic *B-thuringiensis* corn, potato, and rice but not of *B-thuringiensis* canola, cotton, and tobacco. Plant Physiol. Biochem. **2004**, *42* (5), 383–387.
57. Koskella, J.; Stotzky, G. Microbial utilization of free and clay-bound insecticidal toxins from *Bacillus thuringiensis* and their retention of insecticidal activity after incubation with microbes. Appl. Environ. Microb. **1997**, *63* (9), 3561–3568.
58. Stotzky, G. Persistence and biological activity in soil of the insecticidal proteins from *Bacillus thuringiensis*, especially from transgenic plants. Plant Soil **2004**, *266* (1–2), 77–89.
59. Icoz, I.; Stotzky, G. Fate and effects of insect-resistant *Bt* crops in soil ecosystems. Soil Biol. Biochem. **2008**, *40,* 559–586.
60. Palm, C.J.; Schaller, D.L.; Donegan, K.K.; Seidler, R.J. Persistence in soil of transgenic plant-produced *Bacillus thuringiensis* var. *kurstaki* Δ-endotoxin. Can. J. Microbiol. **1996**, *42,* 1258–1262.
61. Crecchio, C.; Stotzky, G. Insecticidal activity and biodegradation of the toxin from *Bacillus thuringiensis* subsp. *kurstaki* bound to humic acids from soil. Soil Biol. Biochem. **1998**, *30*, 463–470.
62. Skarphedinsdottir, H.; Gunnarsson, K.; Gudmundsson, G.A.; Nfon, E. Bioaccumulation and biomagnification of organochlorines in a marine food web at a pristine site in Iceland. Arch. Environ. Contam. Toxicol. **2009**, *58* (3), 800–809.
63. Torres, J.B.; Ruberson, J.R. Interactions of *Bacillus thuringiensis* Cry1Ac toxin in genetically engineered cotton with predatory heteropterans. Transgenic Res. **2008**, *17* (3), 345–354.
64. Harwood, J.D.; Samson, R.A.; Obrycki, J.J. No evidence for the uptake of Cry1Ab Bt-endotoxins by the generalist predator *Scarites subterraneus* (Coleoptera: Carabidae) in laboratory and field experiments. Biocontrol Sci. Technol. **2006**, *16* (4), 377–388.
65. Jiang, Y.-H.; Fu, Q.; Cheng, J.-A.; Zhu, Z.-R.; Jiang, M.-X.; Ye, G.-Y.; Zhang, Z.-T. Dynamics of Cry1Ab protein from transgenic Bt rice in herbivores and their predators. Acta Entomol. Sin. **2004**, *47,* 454–460 [in Chinese with English abstract].
66. Harwood, J.D.; Samson, R.A.; Obrycki, J.J. Temporal detection of Cry1Ab-endotoxins in coccinellid predators from fields of *Bacillus thuringiensis* corn. Bull. Entomol. Res. **2007**, *97,* 643–648.
67. Musser, F.R.; Shelton, A.M. Bt sweet corn and selective insecticides: Impacts on pests and predators. J. Econ. Entomol. **2003**, *96* (1), 71–80.
68. Marvier, M.; McCreedy, C.; Regetz, J.; Kareiva, P. A metaanalysis of effects of Bt cotton and maize on nontarget invertebrates. Science **2007**, *316* (5830), 1475–1477.
69. Wolfenbarger, L.L.; Naranjo, S.E.; Lundgren, J.G.; Bitzer, R.J.; Watrud, L.S. Bt crop effects on functional guilds of non-target arthropods: A meta-analysis. PLoS One **2008**, *3* (5), e2118.

70. Peterson, J.A.; Lundgren, J.G.; Harwood, J.D. Interactions of transgenic *Bacillus thuringiensis* insecticidal crops with spiders (Araneae). J Arachnol., **2011**, 39(1), 1–21.
71. Betz, F.S.; Hammond, B.G.; Fuchs, R.L. Safety and advantages of *Bacillus thuringiensis*-protected plants to control insect pests. Regul. Toxicol. Pharmacol. **2000**, *32* (2), 156–173.
72. Gianessi, L.P.; Carpenter, J.E. Agricultural biotechnology: Insect control benefits; National Center for Food and Agricultural Policy, 1999, available at https://research.cip.cgiar.org/confluence/download/attachments/3443/AG7.pdf.
73. Pray, C.; Ma, D.; Huang, J.; Qiao, F. Impact of Bt cotton in China. World Dev. **2001**, 29, 813–825.
74. Thirtle, C.; Beyers, L.; Ismael, Y.; Piesse, J. Can GM- technologies help the poor? The impact of Bt cotton in Makhathini Flats, KwaZulu-Natal. World Dev. **2003**, *31,* 717–732.
75. Hutchinson, W.D.; Burkness, E.C.; Mitchel, P.D.; Moon, R.D.; Leslie, T.W.; Fleischer, S.J.; Abrahamson, M.; Hamilton, K.L.; Steffey, K.L.; Gray, M.E.; Hellmich, R.L.; Kaster, L.V.; Hunt, T.E.; Wright, R.J.; Pecinovsky, K.; Ra-baey, T.L.; Flood, B.R.; Raun, E.S. Areawide suppression of European corn borer with Bt maize reaps savings to non- Bt maize growers. Science **2010**, *330,* 222–225.
76. Lutz, W.; Samir, K.C. Dimensions of global population projections: What do we know about future population trends and structures? Philos. Trans. R. Soc. Lond. B **2010**, *365* (1554), 2779–2791.
77. Giles, K.L.; Hellmich, R.L.; Iverson, C.T.; Lewis, L.C. Effects of transgenic *Bacillus thuringiensis* maize grain on *B-thuringiensis*-susceptible *Plodia interpunctella* (Lepi- doptera: Pyralidae). J. Econ. Entomol. **2000**, *93* (3), 1011–1016.
78. Hussein, H.S.; Brasel, J.M. Toxicity, metabolism, and impact of mycotoxins on humans and animals. Toxicology **2001**, *167* (2), 101–134.
79. Munkvold, G.P.; Hellmich, R.L.; Rice, L.G. Comparison of fumonisin concentrations in kernels of transgenic Bt maize hybrids and nontransgenic hybrids. Plant Dis. **1999**, *83* (2), 130–138.
80. Lövei, G.L.; Arpaia, S. The impact of transgenic plants on natural enemies: A critical review of laboratory studies. En- tomol. Exp. Appl. **2005**, *114,* 1–14.
81. Pilcher, C.D.; Obrycki, J.J.; Rice, M.E.; Lewis, L.C. Prei- maginal development, survival, and field abundance of insect predators on transgenic *Bacillus thuringiensis* corn. Environ. Entomol. **1997**, *26,* 446–454.
82. Al-Deeb, M.A.; Wilde, G.E.; Higgins, R.A. No effect of *Bacillus thuringiensis* corn and *Bacillus thuringiensis* on the predator *Orius insidiosus* (Hemiptera: Anthocoridae). Environ. Entomol. **2001**, *30* (3), 625–629.
83. Lundgren, J.G.; Wiedenmann, R.N. Coleopteran-specific Cry3Bb1 toxin from transgenic corn pollen does not affect the fitness of a nontarget species, *Coleomegilla maculata* DeGeer (Coleoptera: Coccinellidae). Environ. Entomol. **2002**, *31,* 1213–1218.
84. Anderson, P.L.; Hellmich, R.L.; Sears, M.K.; Sumerford, D.V.; Lewis, L.C. Effects of Cry1Ab-expressing corn anthers on monarch butterfly larvae. Environ. Entomol. **2004**, *33*, 1109–1115.
85. Ramirez-Romero, R.; Desneux, N.; Decourtye, A.; Chaf- fiol, A.; Pham-Delègue, M.H. Does Cry1Ab protein affect learning performances of the honey bee *Apis mellifera* L. (Hymenoptera, Apidae)? Ecotoxicol. Environ. Saf. **2008**, *70* (2), 327–333.
86. Hilbeck, A.; Moar, W.J.; Pusztai-Carey, M.; Filippini, A.; Bigler, F. Toxicity of *Bacillus thuringiensis* Cry1Ab toxin to the predator *Chrysoperla carnea* (Neuroptera: Chrysopi- dae). Environ. Entomol. **1998**, *27,* 1255–1263.
87. Ramirez-Romero, R.; Bernal, J.S.; Chaufaux, J.; Kaiser, L. Impact assessment of Bt-maize on a moth parasitoid, *Cote- sia marginiventris* (Hymenoptera: Bracondiae), via host exposure to purified Cry1Ab protein or Bt-plants. Crop Prot. **2007**, *26,* 953–962.
88. Losey, J.E.; Rayor, L.S.; Carter, M.E. Transgenic pollen harms monarch larvae. Nature **1999**, *399,* 214.
89. Jesse, L.C.H.; Obrycki, J.J. Field deposition of *Bt* transgenic corn pollen: Lethal effects on the monarch butterfly. Oecologia **2000**, *1125,* 241–248.

90. Zangerl, A.R.; McKenna, D.; Wraight, C.L.; Carroll, M.; Ficarello, P.; Warner, R.; Berenbaum, M.R. Effects of exposure to event 176 *Bacillus thuringiensis* corn pollen on monarch and black swallowtail caterpillars under field conditions. Proc. Natl. Acad. Sci. U. S. A. **2001**, *98* (21), 11908–11912.
91. Rosi-Marshall, E.J.; Tank, J.L.; Royer, T.V.; Whiles, M.R.; Evans-White, M.; Chambers, C.; Griffiths, N.A.; Pokelsek, J.; Stephen, M.L. Toxins in transgenic crop byproducts may affect headwater stream ecosystems. Proc. Natl. Acad. Sci. U. S. A. **2007**, *104* (41), 16204–16208.
92. Duan, J.J.; Lundgren, J.G.; Naranjo, S.E.; Marvier, M. Extrapolating non-target risk of *Bt* crops from laboratory to field. Biol. Lett. **2010**, 6, 74–77.
93. Zwahlen, C.; Nentwig, W.; Bigler, F.; Hilbeck, A. Tritrophic interactions of transgenic *Bacillus thuringiensis* corn, *Anaphothrips obscurus* (Thysanoptera: Thripidae), and the predator *Orius majusculus* (Heteroptera: Anthocori- dae). Environ. Entomol. **2000**, *29* (4), 846–850.
94. Lundgren, J.G.; Wiedenmann, R.N. Tritrophic interactions among Bt (CryMb1) corn, aphid prey, and the predator *Co- leomegilla maculata* (Coleoptera: Coccinellidae). Environ. Entomol. **2005**, *34* (6), 1621–1625.
95. Ferry, N.; Mulligan, E.A.; Majerus, M.E.N.; Gatehouse, A.M.R. Bitrophic and tritrophic effects of Bt Cry3A transgenic potato on beneficial, non-target, beetles. Transgenic Res. **2007**, *16,* 795–812.
96. Hilbeck, A.; Baumgartner, M.; Fried, P.M.; Bigler, F. Effects of transgenic *Bacillus thuringiensis* corn-fed prey on mortality and development time of immature *Chrysoperla carnea* (Neuroptera: Chrysopidae). Environ. Entomol. **1998**, *27,* 480–487.
97. Hilbeck, A.; Moar, W.J.; Pusztai-Carey, M.; Filippini, A.; Bigler, F. Prey-mediated effects of Cry1Ab toxin and protoxin and Cry2A protoxin on the predator *Chrysoperla carnea.* Entomol. Exp. Appl. **1999**, *91,* 305–316.
98. Naranjo, S.E. *Risk Assessment: Bt Crops and Invertebrate Non-Target Effects—Revisited;* ISB News Report: Agricultural and Environmental Biotechnology, December 2009, 1–4.
99. Naranjo, S.E. Impacts of *Bt* crops on non-target invertebrates and insecticide use patterns. CAB Rev. Perspect. Ag- ric. Vet. Sci. Nutr. Nat. Resour. **2009**, *4* (11), 1–23.
100. Orr, D.B.; Landis, D.A. Oviposition of European corn borer (Lepidoptera: Pyralidae) and impact of natural enemy populations in transgenic versus isogenic corn. J. Econ. Entomol. **1997**, *90*, 905–909.
101. Reed, G.L.; Jensen, A.S.; Riebe, J.; Head, G.; Duan, J.J. Transgenic Bt potato and conventional insecticides for Colorado potato beetle management: Comparative efficacy and non-target impacts. Entomol. Exp. Appl. **2001**, *100*, 89–100.
102. Al-Deeb, M.A.; Wilde, G.E.; Blair, J.M.; Todd, T.C. Effect of *Bt* corn for corn rootworm control on nontarget soil microarthropods and nematodes. Environ. Entomol. **2003**, *32,* 859–865.
103. Sisterson, M.S.; Biggs, R.W.; Olson, C.; Carriére, Y.; Den- nehy, T.J.; Tabashnik, B.E. Arthropod abundance and diversity in Bt and non-Bt cotton fields. Environ. Entomol. **2004**, *33*, 921–929.
104. Bhatti, M.A.; Duan, J.; Head, G.P.; Jiang, C.; McKee, M.J.; Nickson, T.E.; Pilcher, C.L.; Pilcher, C.D. Field evaluation of the impact of corn rootworm (Coleoptera: Chrysomeli- dae)-protected *Bt* corn on foliage-dwelling arthropods. Environ. Entomol. **2005**, *34,* 1336–1345.
105. de la Poza, M.; Pons, X.; Farinós, G.P.; López, C.; Ortego, F.; Eizaguirre, M.; Castañera, P.; Albajes, R. Impact of farm-scale Bt maize on abundance of predatory arthropods in Spain. Crop Prot. **2005**, *24,* 677–684.
106. Meissle, M.; Lang, A. Comparing methods to evaluate the effects of *Bt* maize and insecticide on spider assemblages. Agric. Ecosyst. Environ. **2005**, *107,* 359–370.
107. Naranjo, S.E. Long-term assessment of the effects of transgenic *Bt* cotton on the abundance of nontarget arthropod natural enemies. Environ. Entomol. **2005**, *34,* 1193–1210.
108. Naranjo, S.E. Long-term assessment of the effects of transgenic Bt cotton on the function of the natural enemy community. Environ. Entomol. **2005**, *34*, 1211–1223.
109. Ludy, C.; Lang, A. A 3-year field-scale monitoring of foliage-dwelling spiders (Araneae) in transgenic *Bt* maize fields and adjacent field margins. Biol. Control **2006**, *38,* 314–324.

110. Torres, J.B.; Ruberson, J.R. Abundance and diversity of ground-dwelling arthropods of pest management importance in commercial Bt and on-Bt cotton fields. Ann. Appl. Biol. **2007**, *150* (1), 27–39.
111. Duan, J.J.; Marvier, M.; Huesing, J.; Dively, G.; Huang, Z.Y. A meta-analysis of effects of Bt crops on honey bees (Hymenoptera: Apidae). PLoS One **2008**, *3* (1), e1415.
112. Bernstein, J.A.; Bernstein, I.L.; Bucchini, L.; Goldman, L.R.; Hamilton, R.G.; Lehrer, S.; Rubin, C.; Sampson, H.A. Clinical and laboratory investigation of allergy to genetically modified foods. Environ. Health Perspect. **2003**, *111* (8), 1114–1121.
113. Environmental Protection Agency. *Cry9C Food Allergenicity Assessment Background Document,* 1999, available at http://www.epa.gov/pesticides/biopesticides/cry9c/cry9c-peerreview.htm.
114. Carter, C.A.; Smith, A. Estimating the market effect of a food scare: The case of genetically modified StarLink corn. Rev. Econ. Stat. **2007**, *89* (3), 522–533.
115. Picard-Nizou, A.L.; Pham-Delegue, M.-H.; Kerguelen, V.; Douault, P.; Marillaeu, R.; Olsen, L.; Grison, R.; Toppan, A.; Masson, C. Foraging behaviour of honey bees (*Apis mellifera* L.) on transgenic oilseed rape *(Brassica napus* L. var. *oleifera).* Transgenic Res. **1995**, *4*, 270–276.
116. Birch, A.N.E.; Geoghegan, I.E.; Griffiths, D.W.; McNicol, J.W. The effect of genetic transformations for pest resistance on foliar solanidine-based glycoalkaloids of potato *(Solanum tuberosum).* Ann. Appl. Biol. **2002**, *140,* 143–149.
117. Faria, C.A.; Wäckers, F.L.; Pritchard, J.; Barrett, D.A.; Turlings, T.J.C. High susceptibility of Bt maize to aphids enhances the performance of parasitoids of lepidopteran pests. PLoS One **2007**, *2,* e600.
118. Flores, S.; Saxena, D.; Stotzky, G. Transgenic *Bt* plants decompose less in soil than non-Bt plants. Soil Biol. Biochem. **2005**, *37,* 1073–1082.
119. Zurbrugg, C.; Honemann, L.; Meissle, M.; Romeis, J.; Ne- ntwig, W. Decomposition dynamics and structural plant components of genetically modified Bt maize leaves do not differ from leaves of conventional hybrids. Transgenic Res. **2010**, *19,* 257–267.
120. Virla, E.G.; Casuso, M.; Frias, E.A. A preliminary study on the effects of a transgenic corn event on the non-target pest *Dalbulus maidis* (Hemiptera: Cicadellidae). Crop Prot. **2010**, *29* (6), 635–638.
121. Ellstrand, N.C.; Prentice, H.C.; Hancock, J.F. Gene flow and introgression from domesticated plants into their wild relatives. Annu. Rev. Ecol. Syst. **1999**, *30,* 539–563.
122. Quist, D.; Chapela, I.H. Transgenic DNA introgressed into traditional maize landraces in Oaxaca, Mexico. Nature **2001**, *414* (6863), 541–543.
123. Christou, P. No credible scientific evidence is presented to support claims that transgenic DNA was introgressed into traditional landraces in Oaxaca, Mexico. Transgenic Res. **2002**, *11* (1), III–V.
124. Kaplinsky, N.; Braun, D.; Lisch, D.; Hay, A.; Hake, S.; Freeling, M. Biodiversity (communications arising): Maize transgene results in Mexico are artefacts. Nature **2002**, *416* (6881), 601.
125. Quist, D.; Chapela, I.H. Maize transgene results in Mexico are artefacts—Reply. Nature **2002**, *416* (6881), 602.
126. Serratos-Hernández, J.-A.; Gómez-Olivares, J.-L.; Salinas- Arreortua, N.; Buendía-Rodríguez, E.; Islas-Gutiérrez, F.; de-Ita, A. Transgenic proteins in maize in the soil conservation area of Federal District, Mexico. Front. Ecol. Environ. **2007**, *5* (5), 247–252.
127. Ortiz-García, S.; Ezcurra, E.; Schoel, B.; Acevedo, F.; So- berón, J.; Snow, A.A. Absence of detectable transgenes in local landraces of maize in Oaxaca, Mexico (2003–2004). Proc. Natl. Acad. Sci. U. S. A. **2005**, *102* (35), 12338–12343.
128. Xia, H.; Lu, B.R.; Su, J.; Chen, R.; Rong, J.; Song, Z.P.; Wang, F. Normal expression of insect-resistant transgene in progeny of common wild rice crossed with genetically modified rice: Its implication in ecological biosafety assessment. Theor. Appl. Genet. **2009**, *119* (4), 635–644.
129. Committee on Genetically Modified Pest-Protected Plants, Board on Agriculture and Natural Resources, National Research Council. *Genetically Modified Pest-Protected Plants: Science and Regulation;* National Academy Press: Washington, D.C., 2000.

130. Wraight, C.L.; Zangerl, A.R.; Carroll, M.J.; Berenbaum, M.R. Absence of toxicity of *Bacillus thuringiensis* pollen to black swallowtails under field conditions. Proc. Natl. Acad. Sci. U. S. A. **2000**, *97* (14), 7700–7703.
131. Gould, F. Integrating pesticidal engineered crops into Mesoamerican agriculture. In *Transgenic Plants in Meso- american Agriculture: Bacillus thuringiensis*; Hruska, A.J., Pavón, M.L., Eds.; Zamorano Academic Press: Zamorano, Honduras, 1997; 6–36.
132. Shi, Y.; Wang, W.B.; Powell, K.S.; Van Damme, E.; Hilder, V.A.; Gatehouse, A.M.R.; Boulter, D.; Gatehouse, J.A. Use of the rice sucrose synthase-1 promoter to direct phloem- specific expression of β-glucoronidase and snowdrop lectin genes in transgenic tobacco plants. J. Exp. Bot. **1994**, *45,* 623–631.
133. Rao, K.V.; Rathore, K.S.; Hodges, T.K.; Fu, X.; Stoger, E.; Sudhaker, D.; Williams, D.; Christou, P.; Bharathi, M.; Brown, D.P.; Powell, K.S.; Spence, J.; Gatehouse, A.M.R.; Gatehouse, J.A. Expression of snowdrop lectin (GNA) in transgenic rice plants confers resistance to rice brown plan- thopper. Plant J. **1998**, *15,* 469–477.
134. Giampietro, M.; Ulgiati, S.; Pimentel, D. Feasibility of large-scale biofuel production—Does an enlargement of scale change the picture? Bioscience **1997**, *47,* 587–600.
135. Mayer, P.M.; Reynolds, S.K.; McCutchen, M.D.; Canfield, T.J. *Riparian Buffer Width, Vegetative Cover, and Nitrogen Removal Effectiveness: A Review of Current Science and Regulations;* EPA/600/R-05/118; U.S. Environmental Protection Agency: Cincinnati, OH, 2006.
136. National Academy of Sciences. *Genetically Modified Pest- Protected Plants: Science and Regulation;* National Academy Press: Washington D.C., 2000.
137. Pessel, F.D.; Lecomte, J.; Emeriau, V.; Krouti, M.; Mes- sean, A.; Gouyon, P.H. Persistence of oilseed rape (*Brassica napus* L.) outside of cultivated fields. Theor. Appl. Genet. **2001**, *102,* 841–846.
138. Terminator gene halt a 'major U-turn.' *BBC News* 1999, available at http://news.bbc.co.uk/2/hi/science/nature/465222.stm.
139. Leslie, T.W.; Biddinger, D.J.; Mullin, C.A.; Fleischer, S.J. Carabidae population dynamics and temporal partitioning: Response to couples neonicotinoid-transgenic technologies in maize. Environ. Entomol. **2009**, *38* (3), 935–943.
140. Landis, D.A.; Wratten, S.D.; Gurr, G.M. Habitat management to conserve natural enemies of arthropod pests in agriculture. Annu. Rev. Entomol. **2000**, *45*, 175–201.
141. Eilenberg, J.; Hajek, A.; Lomer, C. Suggestions for unifying the terminology in biological control. Biocontrol **2001**, *46,* 387–400.
142. Thorbek, P.; Sunderland, K.D.; Topping, C.J. Reproductive biology of agrobiont linyphiid spiders in relation to habitat, season and biocontrol potential. Biol. Control **2004**, *30*, 193–202.
143. Tabashnik, B.E. Genetics of resistance to *Bacillus thuringiensis.* Annu. Rev. Entomol. **1994**, *34,* 47–79.
144. Tabashnik, B.E.; Patin, A.L.; Dennehy, T.J.; Liu, Y.-B.; Carrière, Y.; Sims, M.A.; Antilla, L. Frequency of resistance to *Bacillus thuringiensis* in field populations of pink bollworm. P. Natl. Acad. Sci. U. S. A. **2000**, *97* (24), 12980–12984.
145. Environmental Protection Agency, Office of Pesticide Programs, Biopesticides and Pollution Prevention Division. *Optimum AcreMax Bt Corn Seed Blends: Biopesticides Registration Action Document,* 2010, available at http://www.epa.gov/oppbppd1/biopesticides/ingredients/tech_docs/brad_006490_oam.pdf.
146. Duan, J.J.; Head, G.; Jensen, A.; Reed, G. Effects of *Bacillus thuringiensis* potato and conventional insecticides for Colorado potato beetle (Coleoptera: Chrysomelidae) management on the abundance of ground-dwelling arthropods in Oregon potato ecosystems. Environ. Entomol. **2004**, *33,* 275–281.

3

Biopesticides

G. J. Ash and
A. Wang

Introduction

Inappropriate use of synthetic pesticides in agriculture can lead to environmental degradation, health risks, and loss of biodiversity. Additionally, overuse of particular pesticides may also lead to a loss in their effectiveness through pesticide resistance. Several studies have highlighted the accumulation of pesticides in soils in agricultural systems and the resultant environmental damage.[1] The short-term effects of pesticides on mammalian health are relatively well known and documented through pesticide approval processes. However, the longer-term effects are less well understood but may include cancer, neurotoxic effects, and reproductive disorders.[2] In developing countries, it is estimated that there may be 25 million cases of occupational pesticide poisoning annually.[3] Biopesticides are environmentally sensitive alternatives to the use of synthetic pesticides for the management of a wide range of agricultural pests including weeds, insects, pathogens, and other invertebrate pests. Biopesticides are a class of pesticides that contain plant, animal, or microbial products or organisms.[4] The International Union of Pure and Applied Chemistry[5] further classifies biopesticides as plant-incorporated protectants (protectants produced from genetically modified plants), biochemical pesticides (natural products that affect pests), and microbial biopesticides. Microbial biopesticides are a type of biological control in which a living organism is included as the active ingredient in an inundative application of a formulated product. A range of organisms have been used in biopesticide applications including fungi, nematodes, bacteria, and viruses for the control of weeds, insects, acarines, plant diseases, and molluscs.[6–16]

Formulations of biopesticides contain the organism, a carrier, and adjuvants, which may contain compounds such as nutrients and/or chemicals that aid in the survival of the pathogen or help in protecting the active ingredient from adverse environmental conditions.[17] The formulation of the biocontrol agents is affected by the type of biocontrol organism but must ensure that the agent is delivered in a form that is viable, virulent, and with sufficient inoculum potential to be effective in the field. Furthermore, the formulation may also dictate the means of delivery of the final product, for example, as a seed dressing or a foliar-applied formulation. To be successful, a formulation must be effective, economical, and practical to use.[18] Formulation of biopesticides can be in the form of dry products (dusts, granules, pellets, wettable powders, and encapsulated products) and liquid formulations (suspensions, emulsions, and encapsulated products).[19] The biopesticide is usually packaged, handled, stored, and applied in a similar way to traditional, synthetic pesticides.[20]

Due to the enormous costs involved in the production of synthetic pesticides, global companies have tended to focus on the registration of pesticides in the major crop/cropping systems. This has led to a number of attempts to develop biopesticides in these non-core or niche markets.[21] These markets could be of considerable size and include those that have been created by synthetic pesticide withdrawal and the organic food movement. Philosophically, the use of biopesticides is compatible with organic food production, provided the agent had not been genetically modified, the carriers and adjuvants are natural products, and the host range of the biopesticide is not considered to be too wide.[22]

The use of biopesticides as a strategy in pest management can be applied to both native and introduced pests. However, the success of this type of biocontrol revolves around the costs of production, the quality of the inoculum, and, most importantly, the field efficacy of the product.[18] Biopesticides are usually developed by collaboration with commercial companies in an expectation that they will recoup their costs through the sale of the product.

The broad range of biopesticides may be further subdivided based on the type of target pest and/or the active ingredient. For example, a bioherbicide is a biopesticide developed for weed management whereas a mycoherbicide is a bioherbicide that contains a fungus as an active ingredient. The goal of this entry is to introduce the broad range of biopesticides available and their targets.

Bioinsecticides and Bioacaricides

More than 1500 species of pathogens have been shown to attack arthropods and include representatives from bacteria, viruses, fungi, protozoa, and nematodes.[23] Diseases caused by insects have been known since the early 1800s with the first attempts at inundative applications of fungi to control insects being developed in 1884, when the Russian entomologist Elie Metchnikoff mass produced the spores of the fungus *Metarhizium anisopliae.* A mycoacaricide for the management of citrus rust mite was first registered in the United States in 1981.[24] This was based on the fungus *Hirsutella thompsonii.* Currently, the main fungal species formulated as mycoinsecticides and mycoacaricides include *M. anisopliae* (33.9%), *Beauveria bassiana* (33.9%), *Isaria fumosorosea* (5.8%), and *Beauveria brongniartii* (4.1%).[15] De Faria and Wraight[15] reported a total of 171 fungi-based products with the majority being for the control of insects (160) and mites.[28] Their review indicated that the main formulation types are fungus-colonized substrates, wettable powders, and oil dispersions containing conidia (asexual spores). In the United States, there are seven bioinsecticide products based on various strains of *B. bassiana* registered. These include products such as BotaniGard® and Mycotrol® (available as emulsifiable suspensions and wettable powders).

Bacteria that attack insects can be divided into nonspore-forming and spore-forming bacteria. The nonspore-forming bacteria include species in the Pseudomonadaceae and the Enterobacteriaceae. The spore-forming bacteria belong to the Bacillaceae and include species such as *Bacillus popilliae* and *Bacillus thuringiensis. B. thuringiensis* (Bt) have primarily been developed as biopesticides to control Lepidopteran larva. However, other serotypes of Bt produce toxins that kill insects in the Coleoptera and Diptera as well as nematodes. The bacterium produces δ-endotoxin, an insecticidal crystal protein, which is converted into proteolytic toxins on ingestion. Up to nine different toxins, which have different host ranges, have been described.[25] Commercial formulations of the bacteria contain living spores of the bacteria. Biopesticides based on Bt are the most widely available of the bacterially based products.[25] Up to 90% of the Microbial Pest Control Agent market is Bt or Bt-derived products. The most well known is Dipel.

Entomopathogenic nematodes of the families Steinernematidae and Heterorhabditidae, in conjunction with bacteria of the genus *Xenorhabdus*, have been successfully deployed as biopesticides, for example, BioVECTOR.[23] They are usually applied to control insects in cryptic and soil environments. The nematodes harbor the bacteria in their intestines. The infective third-stage larvae enter the host through natural openings and penetrate into the hemocoel. The bacteria are voided in the insect and cause septicemia, killing the insect in approximately 48 hr.

Entomopathogenic viruses have also shown promise as bioinsecticides. They were first used to control populations of *Lymantria monacha* in pine forests in Germany in 1892 (Huber, 1986, from Moscardi[26]), but the first commercial viral insecticide registered was called Viron/H for the control of *Helicoverpa (Heliothis) zea* in 1971.[27] Viruses from the family Baculoviridae have been isolated from more than 700 invertebrates, with the virus group not common outside of the Lepidoptera and Hymenoptera.[26,28] The nucleo-polyhedroviruses (NPVs) are rod-shaped, double-stranded DNA viruses that are produced in polyhedral proteinaceous occlusion bodies[28] that are ingested by the insect. Granulosis viruses (GVs) are also members of the Baculoviridae but are restricted to the Lepidoptera and have capsular proteinaceous occlusion bodies.[26] These authors[26] provide a table of products that have been developed for the control of insects using these two viral groups.

Biofungicides

Biological control of fungi that cause plant disease can be accomplished by a number of mechanisms including antibiosis, hyperparasitism, or competition. Additionally, weak pathogens may induce systemic acquired resistance in the host, giving a form of cross-protection. Biofungicides have been used in both the phylloplane and rhizosphere to suppress disease. A biological control agent for the control of foliar pathogens in the phylloplane must have a high reproductive capacity, the ability to survive unfavorable conditions, and the ability to be a strong antagonist or be very aggressive. A wide range of bacteria and fungi are known to produce antibiotics that affect other microorganisms in the infection court. Most often, these organisms are sought from a soil environment, as this environment is seen as the richest source of antibiotic-producing species. Species of *Bacillus* and *Pseudomonas* have been successfully used as seed dressings to control soil-borne plant diseases.[29] Serenade®, marketed by BASF, is a formulation of *Bacillus subtilis* (strain QST713), which has claimed activity against a wide range of plant diseases.[30] It is applied as a foliar spray to crops such as cherries, cucurbits, grapes, leafy vegetables, peppers, potatoes, tomatoes, and walnuts.[31] Fluorescent pseudomonads are also often seen as a component of suppressive soils. These bacteria may prevent the germination of fungi by the induction of iron competition through the production of siderophores (ironchelating compounds). These are effective only in those soils where the availability of iron is low. Control of foliar and fruit pathogens such as *Botrytis cinerea,* a pathogen of strawberries, has been accomplished by the foliar application of the soil-inhabiting fungus *Trichoderma viride.*[32] This fungus inhibits *Botrytis* using a combination of antibiosis and competition. On grapevines, *Trichoderma harzianum* competes with *B. cinerea* on senescent floral parts, thus preventing the infection of the ovary. It has also been shown to coil around the hyphae of the pathogen during hyperparasitism.[33] *T. harzianum* has also been reported to induce systemic resistance in plants.[34] One of the earliest commercial successes using *T. harzianum* is the product Rootshield®. Rootshield contains the T-22 strain of *T. harzianum* and is produced and marketed by Bioworks Inc. This strain of the fungus was first registered by the U.S. Environmental Protection Agency in 1993. The product is available as a granular formulation and is usually applied to soil mixtures in glasshouse situations.[35–37]

Bioherbicides

Fungi are the most important group of pathogens causing plant disease. Therefore, fungi (or oomycetes) are most commonly used as the active ingredient in bioherbicides and as such the formulated organism is referred to as a mycoherbicide.[38] However, there are examples of bacteria[7–11] and viruses being used or proposed to be used as bioherbicides.[39,40] The aim of bioherbicide development is to overcome the natural constraints of a weed–pathogen interaction, thereby creating a disease epidemic on a target host.[41] For example, the application of fungal propagules to the entire weed population overcomes the constraint of poor dissemination. After removal of the host weed, the pathogen generally returns to background levels because of natural constraints on survival and spread.

The first commercially available biopesticide for the control of weeds was DeVine®, a bioherbicide for the control of strangler vine in citrus groves in the United States. It was released in 1981.[42] In 1982, a formulation of *Colletotrichum gloeosporioides* f. sp. *aeschynomene* was released to control northern jointvetch in soybean crops in the United States. Since then, there have been a number of products commercialized[12,21] as well as numerous examples of pathogen–weed combinations that had been reported as having potential as bioherbicides in countries including in Canada, United States, Europe, Japan, Australia, and South Africa.[18,21] Necrotrophic or hemibiotrophic fungi are usually used as the basis of mycoherbicides, as they can be readily cultured on artificial media and so lend themselves to mass production. Other desirable characteristics of fungi under consideration as mycoherbicides include the ability to sporulate freely in artificial culture, limited ability to spread from the site of application, and genetic stability. In most cases, these biopesticides are applied in a similar fashion to chemical herbicides using existing equipment, although the development of specialized application equipment and formulation may improve their efficacy and reliability. Since 2000, there have been two successful registrations for bioherbicides in Canada. In 2002, a product called Chontrol®, based on the fungus *Chondrostereum purpureum* for the control of trees and shrubs, was registered.[12] This was based on the research of Hintz and colleagues.[43–45] A more recent success in the area of bioherbicides includes the registration of Sarritor® for dandelion control by the company of the same name in Canada. This product is based on the phytopathogenic fungus *Sclerotinia minor*, which has been extensively researched by Professor Alan Watson at McGill University.[46–50]

Biomolluscicides

Biomolluscicides are a type of molluscicide derived from natural materials such as animals, plants, and microorganisms (e.g., bacterium, fungus, virus, protozoan, or nematode). They are usually used in the fields of agriculture and gardening to control pest slugs and snails. In some circumstances, biomolluscicides are also used in the health area to control molluscs acting as vectors of harmful parasites to human beings.

Currently, the most widely used biomolluscicide is Nemaslug®, a successful biomolluscicide developed by Becker Underwood (U.K.). The active ingredient of Nemaslug is *Phasmarhabditis hermaphrodita*, a nematode species from the family of Rhabditidae. The pathogenicity of *P. hermaphrodita* against slugs had not been recognized until 1994 when Wilson et al.[51] discovered that *P. hermaphrodita* could infect and kill a wide variety of pest slugs under laboratory conditions. Like entomopathogenic nematodes, *P. hermaphrodita* kills slugs by penetrating into the hemocoel of hosts through natural openings and releasing its associated bacteria, which kill the host eventually.

P. hermaphrodita was found to be associated with several different bacteria rather than one particular species, but its association with *Moraxella osloensis* proved to be highly pathogenic to gray garden slug *(Deroceras reticulatum)*.[51] This bacterium was used in the mass production of *P. hermaphrodita* via monoxenic culture.[51]

The host of *P. hermaphrodita* is not restricted to only one slug species (*D. reticulatum*). It can attack and kill several species of slugs including *Arion ater*, *Arion intermedius*, *Arion distinctus, Arion silvaticus, Deroceras caruanae, Tandonia budapestensis,* and *Tandonia* sowerbyi.[51] Moreover, *P. hermaphrodita* can also parasitize several species of snails including *Cernuella virgata*, *Cochlicella acuta*, *Helix aspersa*, *Monacha cantiana*, *Lymnaea stagnalis*, and *Theba pisana*.[52]

Nemaslug is now sold in many European countries, including U.K., Ireland, France, the Netherlands, Belgium, Germany, Denmark, Norway, Finland, Poland, Spain, the Czech Republic, Italy, and Switzerland. In 2005, the retail sale of this biomolluscicide was up to £1 million in Europe and approximately 500 ha horticultural crops (e.g., lettuce and strawberries) and field crops (e.g., wheat, potatoes, and oilseed) were treated with this biomolluscicide. At the dose rate of 3×10^9/ha, *P. hermaphrodita* provides protection against slug damage similar to, if not better than, methiocarb pellets.[16]

Bacteria-based biomolluscicides are now in the process of development. *Streptomyces violaceoruber* and *Xanthobacter autotrophicus* have been examined for their molluscicidal activity against *Oncomelania hupensis* (a unique host of schistosomiasis blood fluke parasite) under laboratory conditions.[53] The results revealed that both bacteria were effective in killing *O. hupensis*, with *S. violaceoruber* causing more snail mortality than *X. autotrophicus* (90% vs. 85%).

Biomolluscicides of plant origin have also been studied extensively in recent years when the environmental pollution caused by chemical molluscicides was realized increasingly. More than 1400 plant species have been screened for their molluscicidal properties against pest snail species.[54] Several groups of compounds present in various plants have been found to be poisonous to snails at acceptable doses, ranging from <1 to 100 ppm, including saponins, tannins, alkaloids, alkenyl phenols, glycoalkaloids, flavonoids, sesquiterpene lactones, and terpenoids.[54] The molluscicidal activity of the dried root latex powder of *Ferula asafoetida*, the flower-bud powder of *Syzygium aromaticum*, and the seed powder of *Carum carvi* against the snail *Lymnaea acuminata* was proved.[55] Similarly, acetogenin (extracted from the seed powder of custard apple) presented promising and stable molluscicidal activity against *L. acuminata*.[54] When sodium alginates was used as a binding matrix for the formulation of acetogenin, the release of this biomolluscicide extended over 25 days, which set up a good example for the development of biomolluscicide delivery system.[54]

The combination of bacteria-based biomolluscicides and plant-based biomolluscicides may lead to a synergistic effect between plant and microbe extracts as molluscicides. Zhang and coworkers[56] reported that higher snail mortality was produced when a mixture of *Arisaema erubescens* tuber extracts and *S. violaceoruber* dilution was applied against the snail *O. hupensis*. The mechanisms of snail toxicosis might be that the combination of *A. erubescens* tuber extract and *S. violaceoruber* dilution reduced the detoxification ability of liver and increased the oxidative damage in liver cells of snails.

Conclusion

Biopesticides are a viable alternative to synthetic pesticides in a number of crops. The development of microbial biopesticides relies on agent discovery and selection, development of methods to culture the pathogen, creation of formulations that protect the organism in storage as well as aid in its delivery, studies of field efficacy, and methods of storage. Each microbial biopesticide is unique, in that not only will the organism vary but so too will the host, the environment in which it is being applied, and economics of production and control.

There are a number of advantages of the use of biopesticides over the use of conventional pesticides, including the minimal residue levels, control of pests already showing resistance to conventional pesticides, host specificity, and the reduced chance of resistance to biopesticides. This indicates an emerging, strong role for biopesticides in any integrated pest management strategy and an important involvement in sustainable farming production systems in the future.

There have been some spectacular successes in the use of microbial biopesticides, despite the perceived constraints to their deployment.[57] In the past, biopesticides have been expected to behave in the same way as synthetic pesticides. For the ultimate success of biopesticides, microorganisms developed for biological control must be viewed by researchers, manufacturers, and end users in a biological paradigm rather than a chemical one. The business model for the commercialization of the products may also vary significantly from that used for traditional synthetic pesticides.[18]

The efficacy and reliability of many microbial biopesticides may be affected by environmental parameters as well as the aggressiveness of the pathogen. Furthermore, the narrow host range of many pathogens may restrict their commercial attractiveness. Both of these issues can be addressed by research into the use of genetic engineering and formulation.[18,58–60] As research into the molecular basis of host specificity and pathogenesis continues, it will become possible to produce more aggressive pathogens with the desired host range for biological control. The survival and efficacy of these pathogens will be enhanced through the use of novel formulations.

References

1. Alletto, L.; Coquet, Y.; Benoit, P.; Heddadj, D.; Barriuso, E. Tillage management effects on pesticide fate in soils. A review. Agron. Sustainable Dev. **2010**, *30,* 367–400.
2. Komarek, M.; Cadkova, E.; Chrastny, V.; Bordas, F.; Bollinger, J.C. Contamination of vineyard soils with fungicides: A review of environmental and toxicological aspects. Environ. Int. **2010**, *36,* 138–151.
3. Jeyaratnam, J. Acute pesticide poisoning: A major global health problem. World Health Stat. Q. **1990**, *43,* 139–144.
4. Available at http://www.epa.gov/opp00001/biopesticides/whatarebiopesticides.htm (accessed July 30, 2010).
5. Available at http://agrochemicals.iupac.org/index.php?p=biopesticides (accessed July 30, 2010).
6. Jaronski, S.T. Ecological factors in the inundative use of fungal entomopathogens. Biocontrol **2010**, *55*, 159–185.
7. Imaizumi, S.; Nishino, T.; Miyabe, K.; Fujimori, T.; Yamada, M. Biological control of annual bluegrass (*Poa annua* L.) with a Japanese isolate of *Xanthomonas campestris* pv *poae* (JT-P482). Biol. Control **1997**, *8*, 7–14.
8. Daigle, D.J.; Connick, W.J.; Boyetchko, S.M. Formulating a weed-suppressive bacterium in “Pesta”. Weed Technol. **2002**, *16,* 407–413.
9. Weissmann, R.; Uggla, C.; Gerhardson, B. Field performance of a weed-suppressing *Serratia plymuthica* strain applied with conventional spraying equipment. Biocontrol **2003**, *48,* 725–742.
10. Anderson, R.C.; Gardner, D.E. An evaluation of the wilt-causing bacterium *Ralstonia solanacearum* as a potential biological control agent for the alien kahili ginger (*Hedy-chium gardnerianum)* in Hawaiian forests. Biol. Control **1999**, *15,* 89–96.
11. DeValerio, J.T.; Charudattan, R. Field testing of *Ralstonia solanacearum* [Smith] Yabuuchi et al. as a biocontrol agent for tropical soda apple *(Solanum viarum* Dunal). Weed Sci. Soc. Am. Abstr. **1999**, *39,* 70.
12. Bailey, K.L.; Boyetchko, S.M.; Langle, T. Social and economic drivers shaping the future of biological control: A Canadian perspective on the factors affecting the development and use of microbial biopesticides. Biol. Control **2010**, *52*, 221–229.
13. Bailey, K.L. Canadian innovations in microbial biopesticides. Can. J. Plant Pathol. **2010**, *32,* 113–121.
14. Hartman, C.L.; Markle, G.M. IR-4 biopesticide program for minor crops. In *Biopesticides;* Humana Press Inc.: Totowa, 1999; Vol. 5, 443–452.
15. de Faria, M.R.; Wraight, S.P. Mycoinsecticides and mycoa-caricides: A comprehensive list with worldwide coverage and international classification of formulation types. Biol. Control **2007**, *43,* 237–256.
16. Rae, R.; Verdun, C.; Grewal, P.; Robertson, J.F.; Wilson, M.J. Biological control of terrestrial molluscs using *Phasmarhabditis hermaphrodita*—Progress and prospects. Pest Manage. Sci. **2007**, *63,* 1153–1164.
17. Hynes, R.K.; Boyetchko, S.M. Research initiatives in the art and science of biopesticide formulations. Soil Biol. Bio-chem. **2006**, *38,* 845–849.
18. Ash, G.J. The science, art and business of successful bioherbicides. Biol. Control **2010**, *52,* 230–240.
19. Auld, B.A.; Hertherington, S.D.; Smith, H.E. Advances in bioherbicide formulation. Weed Biol. Manage. **2003**, *3*, 61–67.
20. Van Driesche, R.G.; Bellows, T.S. *Biological Control;* Chapman and Hall: New York, 1996.
21. Charudattan, R. Biological control of weeds by means of plant pathogens: Significance for integrated weed management in modern agro-ecology. Biocontrol **2001**, *46,* 229–260.

22. Rosskopf, E.; Koenig, R. Are bioherbicides compatible with organic farming systems and will businesses invest in the further development of this technology? VI International Bioherbicide Group Workshop, Canberra, Australia, 2003; 2003.
23. Kaya, H.K.; Gaugler, R. Entomopathogenic nematodes. Annu. Rev. Entomol. **1993**, *38,* 181–206.
24. McCoy, C.W. Factors governing the efficacy of *Hirsutella thompsonii* in the field. In *Fundemental and Applied Aspects of Invertebrate Pathology;* Samson, R.A., Vlak, J.M., Peters, D., Eds.; Foundation of the Fourth International Colloquim of Inverstbrate Pathology: Wageningen, the Netherlands, 1986; 171–174.
25. Rosell, G.; Quero, C.; Coll, J.; Guerrero, A. Biorational insecticides in pest management. J. Pestic. Sci. **2008**, *33*, 103–121.
26. Moscardi, F. Assessment of the application of baculoviruses for control of lepidoptera. Annu. Rev. Entomol. **1999**, *44*, 257–289.
27. Ignoffo, C.M.; Rice, W.C.; Mcintosh, A.H. Inactivation of occluded baculoviruses and baculovirus-DNA exposed to simulated sunlight. Environ. Entomol. **1989**, *18,* 177–183.
28. Fuxa, J.R. Ecology of insect nucleopolyhedroviruses. Agric. Ecosyst. Environ. **2004**, *103,* 27–43.
29. Johnsson, L.; Hokeberg, M.; Gerhardson, B. Performance of the *Pseudomonas chlororaphis* biocontrol agent MA 342 against cereal seed-borne diseases in field experiments. Eur. J. Plant Pathol. **1998**, *104,* 701–711.
30. Available at http://www.agro.basf.com/agr/AP-Internet/en/content/solutions/solution_highlights/serenade/bacillussubtilis (accessed December 1, 2010).
31. Available at http://www.epa.gov/pesticides/biopesticides/ingredients/factsheets/factsheet_006479.htm (accessed December 1, 2010).
32. Sutton, J.C.; Peng, G. Biocontrol of *Botrytis cinerea* in strawberry leaves. Phytopathology **1993**, *83,* 615–621.
33. Oneill, T.M.; Elad, Y.; Shtienberg, D.; Cohen, A. Control of grapevine grey mould with *Trichoderma harzianum* T39. Biocontrol Sci. Technol. **1996**, *6*, 139–146.
34. Shoresh, M.; Harman, G.E. The relationship between increased growth and resistance induced in plants by root colonizing microbes. Plant Signaling Behav. **2008**, *3*, 737–9.
35. Larkin, R.P.; Fravel, D.R. Efficacy of various fungal and bacterial biocontrol organisms for control of fusarium wilt of tomato. Plant Dis. **1998**, *82,* 1022–1028.
36. Brewer, M.T.; Larkin, R.P. Efficacy of several potential biocontrol organisms against *Rhizoctonia solani* on potato. Crop Prot. **2005**, *24,* 939–950.
37. Gravel, V.; Menard, C.; Dorais, M. Pythium root rot and growth responses of organically grown geranium plants to beneficial microorganisms. HortScience **2009**, *44*, 1622–1627.
38. Crump, N.S.; Cother, E.J.; Ash, G.J. Clarifying the nomenclature in microbial weed control. Biocontrol Sci. Technol. **1999**, *9*, 89–97.
39. Ferrell, J.; Charudattan, R.; Elliott, M.; Hiebert, E. Effects of selected herbicides on the efficacy of tobacco mild green mosaic virus to control tropical soda apple (*Solanum viarum*). Weed Sci. **2008**, *56*, 128–132.
40. Charudattan, R.; Hiebert, E. A plant virus as a bioherbicide for tropical soda apple, *Solanum viarum*. Outlooks Pest Manage. **2007**, *18,* 167–171.
41. TeBeest, D.O., Ed. Biological control of weeds with plant pathogens and microbial pesticides. In *Advances in Agronomy*; Academic Press Inc.: San Diego, 1996; Vol. 56, 115–137.
42. Tebeest, D.O.; Yang, X.B.; Cisar, C.R. The status of biological control of weeds with fungal pathogens. Annu. Rev. Phytopathol. **1992**, *30,* 637–657.
43. Harper, G.J.; Comeau, P.G.; Hintz, W.; Wall, R.E.; Prasad, R.; Hocker, E.M. *Chondrostereum purpureum* as a biological control agent in forest vegetation management. II. Efficacy on Sitka alder and aspen in western Canada. Can. J. Forest Res. **1999**, *29*, 852–858.

44. de la Bastide, P.Y.; Zhu, H.; Shrimpton, G.; Shamoun, S.F.; Hintz, W.E. *Chondrostereum purpureum:* An alternative to chemical herbicide brush control. Seventh International Symposium on Environmental Concerns in Rights-of-Way-Management **2002**, 665–672.
45. Becker, E.M.; Ball, L.A.; Hintz, W.E. PCR-based genetic markers for detection and infection frequency analysis of the biocontrol fungus *Chondrostereum purpureum* on Sitka alder and Trembling aspen. Biol. Control **1999**, *15,* 71–80.
46. Shaheen, I.Y.; Abu-Dieyeh, M.H.; Ash, G.J.; Watson, A.K. Physiological characterization of the dandelion bioherbicide, *Sclerotinia minor* IMI 344141. Biocontrol Sci. Technol. **2010**, *20,* 57–76.
47. Li, P.; Ash, G.J.; Ahn, B.; Watson, A.K. Development of strain specific molecular markers for the *Sclerotinia minor* bioherbicide strain IMI 344141. Biocontrol Sci. Technol. **2010**, *20,* 939–959.
48. Abu-Dieyeh, M.H.; Watson, A.K. Efficacy of *Sclerotinia minor* for dandelion control: Effect of dandelion accession, age and grass competition. Weed Res. **2007**, *47,* 63–72.
49. Abu-Dieyeh, M.H.; Watson, A.K. Effect of turfgrass mowing height on biocontrol of dandelion with *Sclerotinia minor.* Biocontrol Sci. Technol. **2006**, *16,* 509–524.
50. Abu-Dieyeh, M.H.; Watson, A.K. The significance of competition: Suppression of *Taraxacum officinale* populations by *Sclerotinia minor* and grass overseeding. Can. J. Plant Sci. **2006**, *86,* 1416–1416.
51. Wilson, M.J.; Glen, D.M.; Hughes, L.A.; Pearce, J.D.; Rodgers, P.B. Laboratory tests of the potential of entomo-pathogenic nematodes for the control of field slugs (*Deroc-eras reticulatum*). J. Invertebr. Pathol. **1994**, *64,* 182–187.
52. Coupland, J.B. Susceptibility of helicid snails to isolates of the nematode *Phasmarhabditis hermaphrodita* from southern France. J. Invertebr. Pathol. **1995**, *66*, 207–208.
53. Li, Y.D.; Yang, J.M. The study on effect of microbe and microbial pesticides killing *Oncomelania hupensis*. Acta Hydro. Sin. **2005**, *29,* 203–205.
54. Singh, A.; Singh, D.K.; Kushwaha, V.B. Alginates as binding matrix for bio-molluscicides against harmful snails *Lymnaea acuminata.* J. Appl. Polym. Sci. **2007**, *105*, 1275–1279.
55. Kumar, P.; Singh, D.K. Molluscicidal activity of *Ferula asafoetida, Syzygium aromaticum* and *Carum carvi* and their active components against the snail *Lymnaea acuminata.* Chemosphere **2006**, *63,* 1568–1574.
56. Zhang, Y.; Ke, W.S.; Yang, J.L.; Ma, A.N.; Yu, Z.S. The toxic activities of *Arisaema erubescens* and *Nerium indicum* mixed with *Streptomycete* against snails. Environ. Toxicol. Pharmacol. **2009**, *27*, 283–286.
57. Auld, B.A.; Morin, L. Constraints in the development of bioherbicides. Weed Technol. **1995**, *9*, 638–652.
58. Wang, C.S.; St Leger, R.J. A scorpion neurotoxin increases the potency of a fungal insecticide. Nat. Biotechnol. **2007**, *25*, 1455–1456.
59. Ash, G.J. Biological control of weeds with mycoherbicides in the age of genomics. Pest Technol. Rev. **2010**, *in press*.
60. St Leger, R.J.; Joshi, L.; Bidochka, M.J.; Roberts, D.W. Construction of an improved mycoinsecticide overexpressing a toxic protease. Proc. Natl. Acad. Sci. U. S. A. **1996**, *93,* 6349–6354.

4

Birds: Chemical Control

Eric B. Spurr

Toxicants

The earliest toxicants were formulations containing arsenic, antimony, phosphorus, and various botanical extracts.[1] Other toxicants used previously include chlorinated hydrocarbons such as endrin, metallic salts such as thallium, organometallic salts such as sodium monofluoroacetate (1080), alkaloids such as nicotine, and anticoagulants such as coumatetralyl and brodifacoum. Most are highly toxic to both birds and mammals. More than 2000 chemicals were evaluated as avicides between the 1940s and the 1980s, and some were found that were selectively toxic to birds. Some were even selectively toxic to certain species of birds. Since the 1980s, however, little effort has been put into finding new toxicants. Instead, most effort has been spent gathering toxicological and environmental data to ensure continued registration of existing products. Recently, international attention has focussed on the animal welfare aspects of toxicants.[2]

Strychnine was once used widely as an oral toxicant for control of birds such as rock pigeons *(Columba livia)* and house sparrows *(Passer domesticus)*, and is still used by certified operators in some countries today (Table 1). It is mainly applied in grain baits (e.g., Sanex Poison Corn in Canada). Strychnine is highly toxic to both birds and mammals, and poses a high risk of both primary and secondary poisoning to nontarget species. Time to death varies from 5 to 50 minutes. It causes extreme pain in poisoned animals and is considered inhumane.

Fenthion was previously used as an oral and dermal toxicant, but is currently used only as a dermal toxicant. Its use is restricted to certified operators. It is applied to wicks in artificial perches or other surfaces, for control of birds such as rock pigeons, house sparrows, and starlings *(Sturms vulgaris)*. It was available previously as Rid-A-Bird® in the United States, and currently as Control-A-Bird® and Avigrease® in Australia. Fenthion (Queletox®) is also aerially sprayed onto birds, especially red-billed quelea *(Quelea quelea)*, in their nighttime roosts, to protect ripening grain crops in some African countries. It is highly toxic to birds and moderately toxic to mammals. Death occurs in 3 to 12 hours. The risk of nontarget bird mortality (from both primary and secondary poisoning) and environmental contamination is high, especially following aerial application. The symptoms of poisoning (e.g., convulsions) indicate that fenthion is likely to be inhumane.

4-Aminopyridine (Avitrol® in the U.S. and Canada, Avis Scare® and Scatterbird® in Australia) is often described as a frightening agent, but it is also an oral toxicant. Birds that ingest it die, but before dying they exhibit erratic behavior and alarm calling (often termed distress behavior) that supposedly frightens away other birds in the flock before they are able to ingest it. Time to death ranges from 15 min to 3 days.

TABLE 4.1 Chemicals Currently Used for Bird Control in United States of America (U.S.A.), Canada, United Kingdom (U.K.), France, Israel, Australia, and New Zealand (N.Z.)

Compound	Activity	Countries
Strychnine	Oral toxicant	Canada, Australia
Fenthion	Oral and dermal toxicant	Some African countries, Australia
4-Aminopyridine	Oral toxicant, frightening agent	U.S.A., Canada, Australia
DRC-1339	Oral toxicant	U.S.A., N.Z.
Alpha-chloralose	Oral toxicant, immobilizing agent	U.S.A., France, U.K., Israel, Australia, N.Z.
Seconal (+ alpha-chloralose)	Immobilizing agent	U.K.
Polybutene	Tactile repellent	U.S.A., Canada, U.K., Israel, Australia, N.Z.
Denatonium saccharide	Taste repellent	U.S.A., Canada
Aluminium ammonium sulfate	Taste repellent	U.K., Australia
Thiram	Taste repellent	France, Israel
Endosulfan	Taste repellent	France
Triacetate guazatine	Taste repellent	France
Methyl anthranilate	Irritant	U.S.A., Canada
Capsaicin	Irritant	U.S.A.
Naphthalene	Irritant	U.S.A.
Methiocarb	Secondary repellent	U.S.A., Canada, Israel, Australia, N.Z.
Ziram	Secondary repellent	U.K., France
Anthraquinone	Secondary repellent	U.S.A., France, N.Z.
Azacosterol	Reproductive inhibitor	Canada
Corn oil	Reproductive inhibitor	U.S.A.
Paraffin oil	Reproductive inhibitor	U.K.

Source: Adapted from Schafer,[1] Ministry of Agriculture[2] and Clark.[3]

It is used to control birds such as rock pigeons, house sparrows, starlings, and in the U.S., red-winged blackbirds *(Agelaius phoeniceus)*. It is available as a concentrate or as ready-to-use treated grain to certified operators. It is highly toxic to both birds and mammals, and may cause both primary and secondary poisoning of nontarget species. Despite appearances to the contrary, it has been claimed that death from the compound is relatively painless. However, this needs to be verified because severe symptoms of intoxication may last for up to 3 days.

DRC-1339 (3-chloro-p-toluidine hydrochloride) (Starlicide®) is an oral toxicant used for the control of birds such as rock pigeons, starlings, and in the U.S., red-winged blackbirds. It is available as a concentrate or as a ready-to-use cereal-based bait to certified operators. Time to death varies from 3 to 50 hr, depending upon the amount of toxicant ingested. DRC-1339 is not suitable as a toxicant for all pest bird species because it is not highly toxic to all species. For example, it has only low toxicity to sparrows (Ploceidae) and finches (Fringillidae). It also has low toxicity to most mammals. This selective toxicity is unique. DRC-1339 is rapidly metabolized, so there is little risk of secondary poisoning. The death of birds from DRC-1339 has been described as painless, but symptoms such as difficult breathing indicate that this might not be so.

Alpha-chloralose is used in some countries (e.g., Australia and New Zealand) as an oral toxicant, but in other countries only as an immobilizing agent (see below). It is available to certified operators as a concentrate or as ready-to-use treated grain, for the control of birds such as rock pigeons and house sparrows. It is generally more toxic to birds than to mammals, and is relatively fast-acting. The first signs of narcosis may occur 10 min after ingestion, and immobilization may last for up to 27 hr, though it generally lasts less than 1 hr, after which birds may recover. However, death may result from hypothermia if sufficient active ingredient is ingested, and/or the weather is inclement. Alpha-chloralose is only slowly metabolized, and so may cause secondary poisoning of nontarget species. It is considered to be relatively humane on the basis of the generally short time to insensitivity.[2]

Lethal Stressing Agents

PA-14 (Tergitol®) is a surfactant that was used as a lethal stressing agent in the U.S., but is no longer available for this purpose. It was sprayed onto birds, such as starlings and red-winged blackbirds, in their nighttime roosts, resulting in a break-down of the oil in the birds' feathers, destroying their natural waterproofing, and causing death from hypothermia.

Immobilizing Agents

Immobilizing agents, administered in baits, are used to make birds easier to capture for removal from areas where they cause problems, or for killing humanely by other methods (e.g., by breaking their necks, or gassing them with carbon dioxide). Nontarget birds that become immobilized can be revived and released. However, the effectiveness of immobilizing agents depends upon the amount ingested and environmental conditions. All known immobilizing agents are lethal to birds if they ingest a sufficient quantity. The most commonly used immobilizing agent worldwide is alpha-chloralose, which is also used in some countries as a lethal toxicant (see above). In the U.S., it is available as an immobilizing agent only to approved operators, mainly to capture rock pigeons and waterfowl in nuisance situations. In the U.K., seconal is also used as an immobilizing agent, in combination with alpha-chloralose, to enhance its speed of action.

Repellents

Chemical repellents can be primary or secondary in effect. Primary repellents are avoided reflexively because of an unpleasant sensation (e.g., touch, taste, smell, irritation). Tactile repellents include polybutene-based products (e.g., 4 The Birds®, Hot Foot®, and Tanglefoot® in the U.S., Bird-X, Buzz-Off®, Shoo, Super Hunter, and Waco in Canada). They are applied to buildings and other structures, modifying the surface so that it becomes sticky or slippery and discouraging birds from landing or roosting. They are all available to the general public.

Taste repellents, which discourage birds from eating potential food sources to which they are applied, include denatonium saccharide (Ro-Pel® in the U.S. and Canada) and aluminium ammonium sulfate (Curb, Guardsman, and Rezist in the U.K., D-ter, Gaard, and Scat in Australia). Ro-Pel® also contains thymol, a fungicide that imparts a secondary repellent effect. Irritants include methyl anthranilate (ReJeX-iT® and Bird Shield® in the U.S., Avigon in Canada), capsaicin (Sevana), and naphthalene (Dr. T's), although there is no evidence that the latter two, by themselves, are effective.[3] Methyl anthranilate may be applied to grassy areas such as parks and golf courses to deter feeding by birds such as Canada geese, and also to ripening fruit to deter birds such as house sparrows and starlings.

Secondary repellents cause post-ingestional illness, resulting in conditioned aversion to the treated food source. Examples include methiocarb (Mesurol®), ziram (AAprotect), and anthraquinone (Flight Control™ in the U.S., Avex™ in New Zealand). Methiocarb and ziram are moderately toxic to birds and mammals. In some countries, methiocarb may be applied to seeds and seedlings, but in the U.S. it may be used only in dummy egg baits to condition crows (*Corvus* spp.) not to prey on the eggs of endangered birds. Ziram and anthraquinone may be sprayed onto grass, field crops, ornamentals, conifers, and dormant fruit trees, but not onto products for immediate human consumption.

4-Aminopyridine is sometimes described as a frightening agent, and classified as a repellent, because it induces behavioral changes in birds. However, it is highly toxic to birds, and should be considered as a toxicant (see above).

Reproductive Inhibitors

Reproductive inhibitors have the potential to reduce bird populations by preventing or reducing the production of young. Azocosterol (Ornitrol®) is one of a number of chemicals that have been investigated for this purpose. It is applied to baits and fed to females daily for 10 to 15 days before egg-laying. It is no longer available in the U.S., but is still available for the control of rock pigeons in Canada. Corn oil (in the U.S.) and paraffin oil (in the U.K.) are two chemicals used to destroy the eggs of birds, such as gulls (*Larus* spp.) and Canada geese *(Branta canadensis)*, after they have been laid. The oil may be sprayed onto the eggs in the nest, or the eggs may be temporarily removed, immersed in oil, and then returned to the nest. The oil occludes the pores in the shell, asphyxiating the developing embryo. The technique is considered humane.[2]

Future Developments

No existing products are ideal for the control of pest birds. Toxicants are becoming increasingly publicly unacceptable worldwide from environmental and animal welfare perspectives. Currently, research is being done on the effectiveness of an oral toxicant/anaesthetic combination that reduces the time to unconsciousness, as a means of improving the animal welfare aspects of lethal bird control. Research is also being done on potential new repellents, including other derivatives of anthranilate, acetophenone, benzoate, cinnamamide, and d-pulegone. The use of nonlethal methods of bird control, especially repellents, may be a better option for the future than the use of toxicants.

References

1. Schafer, E.W., Jr. Bird Control Chemicals—Nature, Modes of Action, and Toxicity. In *CRC Handbook of Pest Management in Agriculture*, 2nd Ed.; Pimentel, D., Ed.; CRC Press: Boca Raton, Florida, USA, 1991; Vol. 2, 599–610.
2. Ministry of Agriculture, Fisheries and Food. *Assessment of Humaneness of Vertebrate Control Agents*; Ministry of Agriculture, Fisheries and Food, Pesticides Safety Directorate: York, United Kingdom, 1997.
3. Clark, L. Review of Bird Repellents. In *Proceedings of the 18th Vertebrate Pest Conference*; Baker, R.O., Crabb, A.C., Eds.; University of California: Davis, California, USA, 1998; 330–337.

5

Birds: Pesticide Use Impacts

Pierre Mineau

Birds inhabiting our farmland are in decline. We know this to be the result of agricultural intensification in which pesticide use plays a large direct and indirect role. Most pesticides used in developed countries no longer accumulate in birds, but they can poison birds and make them more susceptible to other causes of mortality.

More than 30 pesticides registered in North America or Europe have been known to result in kills of wild terrestrial vertebrates even when used according to the relatively stringent regulations in force in those countries. Among the species affected, birds figure prominently in the kill record and this, for several reasons. Birds are ubiquitous and visible. In North America, as in several other countries, most species are federally protected from unlicensed taking or kill. Birds are extremely mobile and cannot be excluded from areas that have been treated with pesticides. Some bird species are attracted to agricultural fields, and many are economically important to the control of agricultural pests, notably insects. Finally, birds, as a group, are particularly sensitive to some of the more toxic classes of pesticides such as the organophosphorus and carbamate insecticides (fortunately, the use of these compounds is in decline), and their reproduction has been found to be potentially affected by a wide range of pesticides. New pesticides developed in part for their relative safety to humans have been found to be especially toxic to birds.

In addition, pesticides are known to alter birds' basic requirements of food and shelter. Loss of food, especially, has been linked to population declines of several farmland bird species in Europe where this has been studied extensively.

Several different strategies are employed to study bird impacts, ranging from monitoring of pesticide applications to surveys of birds in farms subject to different pesticide regimes. Modeling has now given us an estimate of the yearly losses to acutely toxic pesticides; the full impact of pesticide use on birds is more difficult to assess and remains controversial.

How Serious Is the Impact?

Most of our farmland bird species appear to be declining globally and even common species are experiencing longterm declines, both in North America and in Europe.[1,2] For example, 76% of common grassland species in Canada are declining. The proportion of species declining or showing range contractions in the United Kingdom is higher still. Less is known about common farmland bird species outside of Europe or North America, but farming worldwide has been implicated in declines of specific groups of birds such as raptors. It is difficult to isolate the specific factors responsible for these declines: it is likely that a combination of factors is to blame and each species must be considered on a case-by-case basis. Agricultural landscapes have changed dramatically in the 20th century. There has been a shift from mixed agriculture including row crops, field crops, and livestock to more specialized farming where monocultures are regionally dominant. Field size has increased to accommodate larger machinery, and this increase has been often at the expense of marginal non-crop habitats such as fencelines, ditches, hedgerows, windbreaks, and remnant woodlots. In Europe especially, a shift to autumn sowing of grain crops has meant that much of the waste grain traditionally available to birds postharvest and throughout the lean winter months is no longer available.

Agricultural inputs in the form of synthetic fertilizers and pesticides have increased dramatically also and, increasingly, have been found to be contributing to bird declines. Two decades ago, a long-term study of declining grey partridge (*Perdix perdix*) populations in the United Kingdom identified insect prey reductions resulting from both insecticide and herbicide use as the main contributing factor. More recently, researchers in several regions of North America and Europe have shown that organic farms tend to support a higher diversity and abundance of birds even when matched for habitat characteristics. The reproductive success of some farmland species such as the Eurasian skylark (*Alauda arvensis*) is higher on organic or reduced input farms than it is on more "conventional" ones. The use of toxic granular insecticides for oilseed production has contributed to grassland bird population declines in the Canadian prairies.[3] It has been recently suggested[4] that broad aquatic contamination by the new neonicotinoid class of seed-treatment insecticides could be the main reason behind the decline of insectivorous bird species; this is expected to be a hot research area and point of debate in coming years. Taken as a whole, these results implicate current agricultural practices, and pesticide use in particular, in the decline of several farmland species. This is doubly unfortunate because, with a few exceptions, birds can play a useful role in integrated pest management systems.[5]

Types of Bird Impacts

There are several mechanisms through which pesticides can affect birds. The case of the grey partridge, Eurasian skylark, and other European species has shown that the effect can be an indirect one, mediated through "weed" removal and loss of insect biomass at critical times of the breeding season.[6] Herbicide use has increased dramatically in the past decades, and herbicide sales far surpass insecticide sales in North America and Europe at least. However, several direct mechanisms through which birds are impacted are also recognized.

Persistent Organochlorine Pesticides

Historically, several species of raptors such as the Eurasian sparrowhawk (*Accipiter nisus*) and peregrine falcon (*Falco peregrinus*) as well as fish-eating species such as the brown pelican (*Pelecanus occidentalis*) faced serious difficulties and regional extinction as a result of persistent organochlo\1–\2rine pesticides such as DDT (dichlorodiphenyltrichloroethane), aldrin/dieldrin, chlordane, and heptachlor. These were poorly metabolized and poorly excreted by birds and accumulated in fatty tissue. The impact of such substances was twofold. Some, such as aldrin and dieldrin, caused frequent poisonings, especially during lean times when birds metabolized their fat reserves and the pesticides reached extreme concentrations in the brain. Others, such as DDE (dichlorodiphenyldichloroethylene), a breakdown product of DDT, interfered with the bird's ability to lay eggs with normal shells. These substances were banned or severely restricted in most of the developed world in the early 70s. Yet, lower reproduction of birds breeding in areas with high historical usage is still being documented because of long persistence in soils. Several of these pesticides continue to be used massively in parts of the world such as the Indian subcontinent although current impacts on bird life are poorly documented. By and large, modern pesticides do not show such extreme persistence, at least in warmblooded organisms.

Lethal Effects

The acute oral toxicity (LD_{50}) of a pesticide and the extent of its use are good predictors of wildlife kills.[7] The dietary toxicity test currently carried out on young birds (dietary LC_{50}) can seriously mislead however. As a rule, insecticides and vertebrate control agents are much more likely than herbicides or fungicides to give rise to wildlife kills. Two groups of pesticides, the organophosphorus and carbamate insecticides, were initially introduced to replace persistent organochlorines. Unfortunately, they proved particularly toxic to birds. Their mode of action (inhibition of the enzyme acetylcholinesterase in the nervous system and at neuromuscular junctions) is not specific to the pests and affects a broad range of vertebrates and invertebrates alike. Birds are especially vulnerable because their ability to detoxify these pesticides is generally much lower than that of mammals. The more toxic products such as the carbamate insecticide carbofuran killed thousands of individuals in a single application. Reports of large numbers of North American birds being poisoned on their wintering grounds in Latin America by the insecticide monocrotophos have emphasized the need to consider bird impacts in a hemispheric, if not global, context.[8] The poisoning of birds is largely inevitable where acutely toxic pesticides are registered at high rates of application and used broadly. Of particular concern have been granular insecticides and seed treatments because birds are often attracted to them.[9] Pesticide poisonings can be a significant source of mortality relative to other factors, especially in the case of long-lived species such as birds of prey.[10]

There have been very few attempts to estimate the total incidental take resulting from direct intoxications following the use of toxic pesticides anywhere. Pimentel,[11] in an oft-cited study, estimated that pesticide-induced direct mortality totaled approximately 67 million per year in the United States. He based this estimate on the fact that 160 million ha of cropland received a very heavy dose of pesticides per year (3 kg a.i./ha on average—including a number of very toxic pesticides), a breeding density of 4.2 birds/ha (from census plot data), and a conservative kill estimate of 10% of exposed birds. This estimate ignores kills of wintering birds, which could be substantial. Also, some of the largest kills recorded in North America have been of migrants (e.g., Lapland long- spurs (*Calcarius lapponicus*)), which would not be captured in estimates based on breeding densities in farmland.

The carbamate insecticide carbofuran (Furadan™) was very broadly used in North America and has been studied more than any other insecticide.[12] The manufacturer's own studies on a granular formulation of carbofuran as well as search efficiency and scavenging studies were used to provide an estimate of bird mortality per treated surface.[13] Two major field studies, both from the United States, were retained for purposes of extrapolation. Estimated kill rates were 3.05 birds/ha for an Iowa site (once raw carcass counts were corrected for scavenging and for unsearched areas of the field) and 15.9 birds/ha for an Illinois site with better off-field habitat nearby. A third study gave estimates that were simply too high to lead to a kill rate that could safely be extrapolated; fully 799 carcasses of a single species (horned lark—*Eremophila alpestris*) were recovered from slightly more than 100 ha of crop. Based on the two lesser kill rates, it was estimated that, at the height of its popularity, in the late 70s to mid-80s, this single pesticide was killing approximately 17 to 91 million songbirds annually in the 32 million ha of U.S. corn (maize) fields alone.[13]

Fortunately, several of the more toxic organophosphorous and carbamate insecticides are being phased out and replaced in North America and Europe—although their use may still be increasing in the developing world. Their cancellation was not out of a concern for birds[14] but rather an attempt to reduce risks to consumers and applicators under new legislation that demanded the assessment of cumulative impacts from pesticides with the same mode of action. The result of these product cancellations, however, has been a definite reduction in the proportion of our crop area where birds are at risk of lethal poisoning.[15] The authors estimated that the cumulative number of cropped hectares over which avian mortality was probable decreased from about 17 million ha in 1997 to about 6 million ha 5 years later.

The measurement of brain cholinesterase levels was an extremely useful (although certainly not foolproof) diagnostic tool for bird mortality from these classes of pesticides. The test has the advantage of being economical and relatively easy to carry out. Wildlife kills resulting from newly developed insecticides will be harder to elucidate in the absence of such a convenient biomarker. Diagnosis will hinge on sophisticated and costly residue analyses without the benefit of the "smoking gun," which cholinesterase titers represented.

Secondary Poisoning

Secondary poisoning occurs when predators, such as hawks or owls, consume prey contaminated by pesticides. Such predators are few because of their position at the top of the food chain. Therefore, the death of one predator may constitute a significant reduction in the local population of that species. Historically, researchers have associated secondary poisoning with persistent organochlorine insecticides and other substances that are not readily metabolized and that accumulate in tissues. However, other currently registered pesticides can cause secondary poisoning when the predator encounters the pesticide in a high concentration on the surface or in the gastrointestinal tract of its prey. Also, predators capture birds debilitated by insecticides much more easily.

Sublethal Effects and Delayed Mortality

Many pesticides can affect the normal functioning of exposed individuals at doses insufficient to kill them directly. At high doses, the organophosphorus and carbamate pesticides previously described cause respiratory failure and death. However, wild birds exposed to these agents in lesser amounts have experienced impaired coordination, weight loss, an inability to maintain body temperature, and loss of appetite. Also, exposed birds may spend less time at the nest, provide less food for their young, be less able to escape predation, and be more aggressive with their mates. Finally, exposure to some pesticides may reduce resistance to disease.

Effects on Reproduction

A high proportion of pesticides currently registered have the potential to affect reproduction by reducing egg production, hatching, or fledging success, although the extent to which this actually happens in the wild is not known.[16,17] A few products cause embryonic mortality when sprayed directly onto eggs.

Routes of Pesticide Exposure in Birds

Birds ingest pesticides through their food or through preening or grooming. Despite being feathered, they absorb pesticides through their skin, encountering droplets directly or by rubbing against foliage and other contaminated surfaces. Birds are also exposed through their feet. Finally, they have a very high ventilation rate and inhale vapor and fine droplets. The degree to which each of these routes of exposure contributes to the total dose depends on the crop being sprayed, the chemical, the species exposed, and environmental factors. Evidence to date suggests that the dietary route is not necessarily the dominant route of exposure in birds under most situations. Yet, this is the only route currently assessed by regulators worldwide—a mind-set that clearly needs to change if we want to be proactive in protecting birds from pesticides.

Although the relative importance of different exposure routes is difficult to ascertain on a case-by-case basis, it is possible to recognize different situations that arise where birds are massively exposed to pesticides and are often poisoned as a result.

Abuse and Misuse

Pesticide abuse is the deliberate use of a pesticide in a non-authorized fashion, usually to poison wildlife species considered to be pests. In the United Kingdom, as well as in several European countries, officials estimate that deliberate bird kills due to pesticide abuse outnumber cases where label instructions were strictly followed. Between 1978 and 1986, officials in the United Kingdom estimate that, on average, 71% of incidents were the result of abuse. For birds of prey alone, more than 90% of cases recorded between 1985 and 1994 in the United Kingdom were abuse cases. On the other hand, for raptors in the United States during the same period, kills involving labeled uses of pesticides were almost as frequent as abuse cases. This difference appears to be wholly attributable to the high toxicity of insecticides used in the United States. Abuse generally involves baits of some kind, the only limit being the imagination of the perpetrator. Typically, liquid insecticides are poured or injected and applied to seed, bread, meat, etc., and granules are sprinkled or mixed into a paste. Because of the high concentration of pesticide involved in abuse cases, carcasses are usually found in close proximity to the site of baiting, thus biasing the kill record through a higher recovery of carcasses. The choice of chemicals used in abuse cases reflects availability and toxicity. Pesticides typically used in deliberate poisoning attempts include carbofuran, aldicarb, monocrotophos, parathion, mevinphos, diazinon, and fenthion, chemicals that are all recognized as being inherently very toxic to vertebrates in general and birds in particular. The main problem of course is that the baits are often indiscriminate in the species that they kill. Secondary poisoning is also frequent when predators or scavengers take dead or debilitated prey with highly concentrated bait in their gut.

Pesticide misuse refers to a pesticide application that is not exactly as specified by the label. This may be an application at a rate that is higher than specified or an application to a crop or pest other than those listed. Alternatively, the user may not have the legal permission to use a certain product even if he followed label directions to the letter. Pesticide misuse is difficult to establish, especially after the fact. In many cases, it becomes very difficult to distinguish a misuse from a normal agronomic use when the label contains instructions that are vague, difficult, or impossible to follow. What constitutes a misuse in one jurisdiction may indeed be an approved use elsewhere.

Granular Formulations and Treated Seed

Granular insecticides and treated seeds are frequent routes of exposure and intoxication in birds. Granular insecticides were designed for convenience, safety to applicators, and time release of the chemical, yet for birds, granular formulations of the more toxic insecticides such as aldicarb, parathion, carbofuran, fensulfothion, phorate, terbufos, fonofos, disulfoton, diazinon, and bendiocarb are repeatedly associated with bird mortality. Several bird and small mammal species have a fatal attraction to granular formulations, mistaking them either as dietary grit or as a food source. The most attractive granules are those made of sand (silica) or an organic base such as dried corn (maize) cobs. Somewhat less attractive are clay, gypsum, and coal granules. Exposure can also occur via invertebrates, especially earthworms to which granules easily adhere. Secondary toxicity is likely in predators and scavengers that eat their prey whole or ingest their gastrointestinal tract contents. In Canada and the United States, there have been cases of poisoning of waterfowl foraging in puddles in fields more than 6 months after applications of granular insecticides because of specific soil conditions that may retard break- down.[18] Granules that are friable and disintegrate quickly when exposed to moisture are best for birds, but they are the products least convenient to farmers. Regardless of the type of carrier, a pesticide granule is likely to be a problem if a lethal dose can be obtained in a few granules only.

To date, *no* agricultural machinery or application technique can achieve complete incorporation of granular insecticides below the soil horizon. Birds have also been known to probe the soil for granules or to pull up germinating seeds with granules attached. The worst applications are those made above the soil surface and “banded” or “side dressed” over or to the side of the seed furrow. In carefully controlled engineering trials, between 6% and 40% of applied granules were left on the soil surface. The same equipment can achieve radically different soil incorporation when used by different individuals under different conditions.

Treated seeds present a similar engineering problem. As with granules, more seeds are left on the surface wherever the seeders have to turn or negotiate obstacles. Small spills are part of normal farming practice and can occur anywhere depending on topography and soil conditions but more often at field edge. Historically, seed dressings were one of the main sources of bird exposure to organochlorine and mercurial compounds. Poisoning incidents with seed dressings are still relatively frequent because several bird species make heavy use of waste (or even planted) grain in fields. The size and type of seed dictate which bird species are at risk. Since use of organochlorines and organomercurials has declined, kills have been recorded with cholinesterase-inhibiting insecticides such as carbophenothion, chlorfenvinphos, isofenphos, bendiocarb, disulfoton, furathiocarb, and fonofos. Some kills have been recorded with newer insecticides as well, e.g., the neonicotinoid imidacloprid, although it is not yet known how serious or frequent a problem this will become.

Liquid Formulations on Vegetation: The Grazing Problem

Grazing birds are particularly vulnerable to foliar applications of pesticides. Kills have been recorded with several cholinesterase-inhibiting pesticides such as parathion, diazinon, carbofuran, isofenphos, dimethoate, and triazophos. Grazers typically include geese, ducks, and coots (families Anatidae and Rallidae). These birds eat large quantities of foliage because they do not digest cellulose. Fertilized areas are particularly attractive to grazing species that can detect the high nitrogen levels. Golf courses attract grazers because the turf is cut frequently, watered, and fertilized, and courses often have other attractions such as ponds and drainage streams. More than 100 cases of waterfowl mortality were recorded due to the use of diazinon on turf [19] before the pesticide was withdrawn from golf courses and sod farms in the United States. Other well-documented problems are kills of ducks and geese in alfalfa fields treated with carbofuran and of sage grouse (*Centrocercus urophasianus*) feeding on alfalfa crops treated with dimethoate or on potato foliage and weeds in potato fields sprayed with methamidophos.

Liquid Formulations on Insect Prey: The Gorging Problem

Bird species that feed on agricultural pests such as grasshoppers, leatherjackets (larvae of the crane fly), grubs, and cutworms are at high risk of poisoning. Kills of these species are all the more tragic because they are beneficial to agriculture. Some species are particularly vulnerable because they specialize in insect outbreaks. These birds take advantage of pest control operations that result in insects becoming either debilitated or more visible following treatment. In a well-studied case in Argentina, approximately 20,000 Swainson's hawks were poisoned within the span of a few weeks after feeding on grasshoppers sprayed with monocrotophos. As with carbofuran, the extreme toxicity of this product means that it is difficult to find use patterns that do not result in bird kills.

Vertebrate Control Agents: Unintended Victims

Rodenticides as a rule are not specific to their intended targets and cause direct impacts to non-target species. Only a detailed knowledge of the habits of the target species and use of specific baiting locations or specialized bait holders can reduce kills of non-target species. More problematic is secondary poisoning. Historically, the use of thallium and endrin to control rodents has had disastrous consequences on raptors. Recently, the trend has been to use more efficacious "single feed" anticoagulants; these present a greater hazard to predators than the older products (e.g., warfarin, diphacinone, chlorophacinone). The new "super coumarin"-type products include compounds such as difenacoum, brodifacoum, bromadiolone, difethialone, and flocoumafen—all extremely toxic and very long lived in liver tissue, thereby increasing the likelihood of secondary poisoning. A recent analysis of Canadian data suggests that approximately 11% of the great horned owl (*Bubo virginianus*) population of southern inhabited Canada is at risk of fatal poisoning from anticoagulants.[20]

Fenthion, an organophosphorus "insecticide" used to control pest birds in Africa (e.g., *Quelea quelea*) and in North America (e.g., by means of the Rid-a-Bird™ perch system), has given rise to frequent secondary poisoning.[21] Secondary poisoning is also very likely following the use of toxic organophosphorous or carbamate products for the control of parakeets, doves, and other seed eaters. The use of organophosphorus pesticides such as famphur and fenthion for the treatment of parasites in livestock frequently leads to wildlife kills. Famphur, which was one of the leading causes of eagle poisonings in the American Southwest, persists on the hair of cattle up to 100 days after treatment. Magpies are poisoned when they eat the hair, and eagles are poisoned when they scavenge the magpies. Medicated feed at livestock feed yards is another high-exposure situation. Sparrows, starlings, and other birds pick up the feed and subsequently are scavenged by hawks and eagles.

Forestry Insecticides

Forestry uses of toxic insecticides deserve special consideration because the terrain and method of application result in kills being difficult or impossible to detect. In a forestry situation, critical wildlife habitat is sprayed directly, and a large number of individuals of many species are exposed to the chemical. In Canada, the forestry insecticides phos-phamidon and fenitrothion were canceled after impacts on birds were judged unacceptable. Although fenitrothion is not as acutely toxic as a number of other anti-cholinesterase insecticides used in agriculture, its use in forestry led to severe and widespread inhibition of brain acetylcholinesterase in a number of songbird species as well as some reports of kills.

Measuring Bird Impacts

Incident Monitoring

Incident monitoring refers to the capacity of competent authorities to investigate reported kills or conduct spot checks of use conditions. Even if a pesticide has been studied extensively under controlled conditions, unforeseen problems and situations often arise following commercialization of the product. An absence of incident reports does not necessarily mean there are no problems but, conversely, well-investigated incidents and kills can reveal unforeseen aspects of a pesticide or reinforce a suspicion that arose in the course of laboratory or field testing. An incident monitoring scheme will require a network of individuals trained in carrying out pesticide investigations and in proper handling of carcasses and tissue samples, as well as access to a laboratory equipped to perform the required chemical and biochemical analyses.[22]

Even where relatively efficient incident monitoring systems are in place, only a very small proportion of kills are ever uncovered. There are several reasons for this: affected wildlife are often dispersed and at relatively low density in farm fields, they often leave the treated area to die, they are likely to seek cover and hide when overcome by the pesticide, they are often cryptic and hard to see, and their carcasses are scavenged rapidly after death. Typical rates of carcass removal by scavengers are 40%-90% in the first 24 hr. Farm fields are large; the mechanization and sheer size of modern agricultural machinery often remove the farmer from any "close contact" with the land. The increasing size of farms also means that, when kills occur, often in the few days that follow a pesticide application, the farmer is busy elsewhere, treating another part of the farm. Pesticide intoxication may be a causal factor in a kill visibly caused by something else—e.g., intoxicated birds hitting fences, utility wires, cars, or buildings—and not be recognized as a pesticide kill. Also, there is a large difference between casual searching of fields and a well-organized intensive search effort. An intensive search effort consisting of several trained individuals, transects, and repeated, well-timed searches have produced between 10- and 500-fold improvements in carcass detection rates over field inspections carried out once or a few times only by single individuals. Equally important is the motivation and training of the search teams. Finally, a proper investigation of kills can be expensive and out of the reach of many jurisdictions despite the availability of inexpensive biomarkers—currently at least.[23]

Even when incidents are uncovered, they are often not reported. If the kills involve only one or a few individuals, not much importance is attached to the incident even though, for reasons just outlined, these few carcasses likely represent the "tip of the iceberg." Even if the kills are reported, the information is often not centralized and made available to national pesticide regulatory bodies. It is important to understand and recognize biases inherent in any incident reporting system. Some of those biases will depend on how the incident monitoring system is set up and which persons/organizations are responsible. Some biases can be reduced over time, but others are unavoidable. Common biases relate to body size and color of the casualties, numbers and density of the species in any given area, "status" of the species, and individual and institutional interests and sensitivities. We expect most kills to be of small-bodied birds widely dispersed in field margins. Yet, such kills are seldom reported.

Despite these limitations, it is important for countries to investigate wildlife kills and make the information available.[24] Registration decisions are made on the basis of very limited information. There are large differences in toxicological and ecological vulnerability among species. The ways in which wildlife species are exposed to pesticides are varied and sometimes difficult to predict or study. The behavior of pesticides depends on local conditions although pesticides are often tested under standardized conditions only. The outcome of exposure is also much more variable in the wild. Pesticide exposure can interact with weather, the condition or health of the animal, etc. Therefore, whether or not pesticides are routinely field tested to look for environmental impacts, it is essential to have a good incident monitoring system in place. An incident monitoring system can also be useful to warn manufacturers if their products are abused or used incorrectly.

Field Testing: Active Monitoring

Carrying out a field study to measure the impact (or lack of impact) of a specific pesticide usually consists of the surveillance of a group of birds prior to, during, and after the application of the pesticide according to label instructions. Researchers observe or count individuals of one or more species within and outside the treated area and record their behavior. Frequently, they search for carcasses in order to determine the extent of pesticide-induced mortality.[25] for an example of how these data can yield useful predictive models.) They may capture birds to ascertain the health of individuals or to collect samples, for example, blood or brain tissue for biochemical assays or feathers and foot rinses for residue determinations. Agricultural engineering studies (e.g., measurements of granular insecticides or treated seed remaining on the surface) or monitoring of pesticide residues remaining on avian food items over time provides valuable information on expected exposure levels. The most sophisticated field studies will involve monitoring nests, as well as banding, marking, or radio-tagging individuals in order to assess turnover rates and help locate sick and dead birds. Rare, vulnerable, or ecological keystone species can be used as indicator species where relevant and feasible.

It is not always feasible to investigate the effects of a single pesticide. In a number of cropping situations, several pesticides are used as a mix or in quick succession, making the identification of compound-specific impacts difficult. In agricultural systems, the mosaic of treated fields can be so complex as to make it difficult to assess exposure to any one pesticide. Two approaches then suggest themselves: 1) treated sites or landscapes are compared to non-treated areas, provided those can be found, and 2) the "severity" of treatment (the *a priori* expectation of toxicity) for any given site is used as a variable against which a number of different parameters (such as reproductive success) are regressed. Great care must be taken in comparing treated to non-treated areas because they are likely to differ in other ways as well.

Surveys

Data from regional or national surveys of bird population levels are rarely adequate to demonstrate specific pesticide impacts, although surveys can point to a general situation of bird declines in farmland. In order to carry out wildlife monitoring in treated areas, it is necessary to have a good knowledge of the normal complement of species for the area of concern and to be able to assess the vulnerability of each of these species during and after pesticide treatments. The diversity or abundance of species may already have been affected by previous pesticide use so that only a complement of the more insensitive species remains available for testing.

Regardless of the strategy employed, more attention needs to be paid to the impact of pest control practices on bird species if we are to reverse the current trends of population declines.

Modeling

The probability of finding a bird kill (of any size) following a pesticide application was derived from models based on a large sample of empirical field studies where known insecticides were applied, and searching was carried out to detect casualties.[25] Models were developed for field and orchard crops separately. Because few of the studies were quantitative in nature, logistic modeling was used and the output of the models is the likelihood that a kill of undefined size would occur and be found assuming an adequate search effort.

Species most frequently implicated in kills are those that are cosmopolitan, closely associated with agriculture, and reasonably common, e.g., mourning doves (*Zenaida macroura*); several sparrows, horned larks, and meadowlarks (*Sturnella* spp.), American robins (*Turdus migratorius*); house sparrows (*Passer domesticus*); and several blackbird species. However, the sheer diversity of birds potentially killed by pesticides is impressive and suggests that toxicological or ecological susceptibility is less important than being simply in the wrong place at the wrong time.

Conclusion: The Way Forward

The pesticide industry has clearly shown that it is incapable of policing itself when it comes to reducing or eliminating impacts on birds. Elsewhere,[14,26] I have reviewed some of the more egregious cases, showing how it often takes decades of legal wrangling to remove clear problem pesticides from the market in developed countries while the use of those same compounds continues to increase in the developing world, sometimes affecting the same bird populations (e.g., neotropical migrants). The extent to which birds are protected from pesticide impacts in various countries depends very much on public opinion and on the effectiveness of bird conservation groups. The pesticides responsible for most of the impacts on birds around the world (at least the direct impacts) tend to be the same group of depressingly familiar products. Fortunately, they usually can be replaced by better alternatives without risk to the livelihood of farmers or food security. Adopting better laws and regulations to protect the environment against pesticide use is part of the answer; enforcing those regulations in the face of a very strong pesticide lobby is undoubtedly the biggest hurdle.

References

1. Askins, R.A. Population trends in grassland, shrubland, and forest birds in eastern North America. Curr. Ornithol. **1993**, 11, 1–34.
2. Sirawardena, G.; Baillie, S.R.; Buckland, S.T.; Fewster, R.M.; Marchant, J.H.; Wilson, J.D. Trends in the abundance of farmland birds: A quantitative comparison of smoothed Common Birds Census indices. J. Appl. Ecol. **1998**, *35,* 24–43.
3. Mineau, P.; Downes, C.M.; Kirk, D.A.; Bayne, E.; Csizy, M. Patterns of bird species abundance in relation to granular insecticide use in the Canadian prairies. Ecoscience **2005**, *12* (2), 267–278.
4. Tennekes, H. The Systemic Insecticides: A Disaster in the Making; Weevers Walburg Communicatie, Zutphen, Netherlands, 2010, 72 pp.
5. Kirk, D.A.; Evenden, M.D.; Mineau, P. Past and current attempts to evaluate the role of birds as predators of insect pests in temperate agriculture. Curr. Ornithol. **1997**, *13,* 175–269.
6. Campbell, L.H.; Avery, M.I.; Donald, P.; Evans, A.D.; Green, R.E.; Wilson, J.D. A review of the indirect effect of pesticides on birds. Joint Nature Conservation Committee Report No. 227. JNCC: Peterborough, England, U.K., 1997.
7. Mineau, P.; Baril, A.; Collins, B.T.; Duffe, J.; Joerman, G.; Luttik, R. Reference values for comparing the acute toxicity of pesticides to birds. Rev. Environ. Contam. Toxicol. **2001**, *170,* 13–74.
8. Hooper, M.J.; Mineau, P.; Zaccagnini, M.E.; Winegrad, G.W.; Woodbridge, B. Monocrotophos and the Swainson's hawk. Pestic. Outlook **1999**, *10* (3), 97–102.
9. Stafford, T.R.; Best, L.B. Bird response to grit and pesticide granule characteristics: Implications for risk assessment and risk reduction. Environ. Toxicol. Chem. **1999**, *18* (4), 722–733.
10. Mineau, P.; Fletcher, M.R.; Glazer, L.C.; Thomas, N.J.; Brassard, C.; Wilson, L.K.; Elliott, J.E.; Lyon, L.A.; Henny, C.J.; Bollinger, T.; Porter, S.L. Poisoning of raptors with organophosphorous pesticides with emphasis on Canada, U.S. and U.K. J. Raptor Res. **1999**, 33, (1), 1–37.
11. Pimentel, D.; Acquay, H.; Biltonen, M.; Rice, P.; Silva, M.; Nelson, J.; Lipner, V.; Giordano, S.; Horowitz, A.; D'Amore, M. Environmental and economic costs of pesticide use. BioScience **1992**, *42,* 750–760.
12. Richards, N., Ed. *Carbofuran and Wildlife Poisoning: Global Perspectives and Forensic Approaches*; Wiley-Blackwell, Chichester, U.K. 2012, 277 pp.
13. Mineau, P. Direct losses of birds to pesticides—Beginnings of a quantification. In *Bird Conservation Implementation and Integration in the Americas*, Proceedings of the Third International Partners in Flight Conference 2002; Ralph, C.J., Rich, T.D., Eds.; U.S.D.A. Forest Service, GTR-PSW-191, Albany, CA, 2005; Vol. 2, 1065–1070.
14. Mineau, P. Birds and pesticides: Are pesticide regulatory decisions consistent with the protection afforded migratory bird species under the Migratory Bird Treaty Act? William Mary Environ. Law Policy Rev. **2004**, *28* (2), 313–338.
15. Mineau, P.; Whiteside, M. The lethal risk to birds from insecticide use in the U.S.—A spatial and temporal analysis. Environ. Toxicol. Chem. **2006**, *25* (5), 1214–1222.
16. Mineau, P.; Boersma, D.C.; Collins, B. An analysis of avian reproduction studies submitted for pesticide registration. Ecotoxicol. Environ. Saf. **1994**, *29,* 304–329.
17. Mineau, P. A review and analysis of study endpoints relevant to the assessment of "long term" pesticide toxicity in avian and mammalian wildlife. Ecotoxicology **2005**, *14* (8), 775–799.

18. Elliott, J.E.; Wilson, L.K.; Langelier, K.M.; Mineau, P.; Sinclair, P. Secondary poisoning of birds of prey by the organophosphorus insecticide, phorate. Ecotoxicology **1996**, *5*, 1–13.
19. Frank, R.; Mineau, P.; Braun, H.E.; Barker, I.K.; Kennedy, S.W.; Trudeau, S. Deaths of Canada geese following spraying of turf with diazinon. Bull. Environ. Contam. Toxicol. **1991**, *46,* 852–858.
20. Thomas, P.J.; Mineau, P.; Shore, R.F.; Champoux, L.; Martin, P.; Wilson, L.; Fitzgerald, G.; Elliott, J.E. Second generation anticoagulant rodenticides in predatory birds: Probabilistic characterisation of toxic liver concentrations and implications for predatory bird populations in Canada. Environ. Int. **2011**, *37* (5), 914–920.
21. Hunt, K.A.; Bird, D.M.; Mineau, P.; Shutt, L. Secondary poisoning hazard of fenthion to American Kestrels. Arch. Environ. Contam. Toxicol. **1991**, *21,* 84–90.
22. ASTM. *Standard Guide for Fish and Wildlife Incident Monitoring and Reporting*; American Society for Testing and Materials: Philadelphia, PA, Standard E 1997; 1849–1896.
23. Mineau, P.; Tucker, K.R. Improving detection of pesticide poisoning in birds. J. Wildl. Rehab. **2002**, Part 1: *25* (2), 4–13; Part 2: *25* (3), 4–12.
24. Greig-Smith, P.W. Understanding the impact of pesticides on wild birds by monitoring incidents of poisoning. In *Wildlife Toxicology and Population Modeling*, SETAC Special Publication Series; Kendall, R.J., Lacher, T.E., Eds.; CRC Press, Inc.: Boca Raton, FL, 1994; 301–319.
25. Mineau, P. Estimating the probability of bird mortality from pesticide sprays on the basis of the field study record. Environ. Toxicol. Chem. **2002**, *24* (7), 1497–1506.
26. Mineau, P. Birds and pesticides: Is the threat of a silent spring really behind us? 2009 Rachel Carson Memorial Lecture, Pesticide News **2009**, *86,* 12–18, available at http://www.pan-uk.org/pestnews/Free%20Articles/PN86/Birds%20and%20pesticides.pdf.

6

Insect Growth Regulators

Meir Paul Pener

Introduction

Insect growth regulators, abbreviated as IGRs, are chemical compounds that interfere with insect-specific physiological systems that do not exist in vertebrates. The body of insects and related arthropods is covered by a more or less hard, often very hard, exoskeleton, named “cuticle”. Once produced, the cuticle is fixed and it cannot grow larger. To allow growth of the insect, the old cuticle is shed and a new, larger, often different cuticle is formed. This change of the old cuticle to a new one is named “molt” or “molting” and a stage between two consecutive molts is named an “instar”. To reach the adult stage, insects molt several times, i.e. they have several instars and they undergo “metamorphosis”, from egg to several larval or nymphal stages, then to adult (“incomplete metamorphosis”), or from egg to several larval stages then to pupa and finally to adult (“complete metamorphosis”). Both molting and metamorphosis are controlled by hormones. Through molecular processes, the effect of these hormones is upregulation or downregulation of the expression of many enzymes, some of which are specific to insects or arthropods. These physiological systems constitute the targets of the IGRs as they do not exist in vertebrate animals, including humans. Most IGRs used today in insect control are either interfering with the formation of the new cuticle, disturbing the molting process, or interfering with metamorphosis. The name “insect growth regulators” is somewhat unfortunate; IGRs rather “deregulate” insect development. An attempt made by Pener and Dhadialla[1] to use the name “insect growth disruptors” instead of “insect growth regulators” turned out to be unsuccessful. It seems that the name “insect growth regulators” is deeply embedded in the literature and is difficult to uproot it.

An Overview of IGRs

There are thousands of scientific publications on IGRs. ISI Web of Knowledge lists over 2,300 articles on the subject that appeared in peer-reviewed scientific journals. This list does not include books, book chapters, conference proceedings volumes, encyclopedia entries, and journals not covered by ISI. For comparison, in my former chapter of 2013,[2] this number was 1400.

Reviews on IGRs are numerous but usually deal with one or two kinds (categories) of IGRs and not with all categories. Also, in most instances, reviews are either applied oriented with effect(s) on whole organism and/or tissues, or related to basic science, focused on the mode of action of IGRs at cellular, biochemical, and more recently at molecular levels. Additionally, IGR research often constitutes an approach or method to study the mode of action of insect hormones and/or enzymes, again mostly in relation to basic research.

A review of 2012[1] presents a detailed account of the applied aspects of all categories of IGRs. A slightly earlier review[3] is comprehensive, and it is a remarkable example of discussion of all categories of IGRs, from both the applied and basic research aspects. It lists the most important literature on IGRs. Another review devoted to all categories of IGRs was published in 2016.[4] Some other reviews also deal with all categories of IGRs but aim to a specific insect group or biotope; for example, IGRs against stored product insects[5] or storage mites.[6] Also, tens of different studies were published on the mode of action of IGRs on a single harmful species, exemplified by *Rhodnius prolixus*.[7]

This chapter is an updated and enlarged revision of my former chapter on IGRs[2] and is application oriented, as this aspect is more important from the standpoint of environmental management. The presented list of references is based mostly on reviews, literature from the last two decades, and on publications with historical importance. Basic research on IGRs is not detailed; however, reviews relevant to basic research are mentioned. These reviews, as well as the comprehensive review mentioned,[3] cite a detailed literature on relevant basic research.

As targeted against insect-specific physiological processes, IGRs exhibit no or mild toxicity to vertebrates. In some cases, lack of susceptibility of vertebrates is so marked that IGRs are administered by the "feed-through" method, i.e. applied orally to a vertebrate animal against insects that are harmful as adults which in the larval stages feed on the feces of that animal. For example, cyromazin is fed to poultry for controlling housefly (*Musca domestica*) maggots in the manure.[8,9] Methoprene is fed to cattle to control horn fly (*Haematobia irritans*) larvae that develop in cattle dung.[10,11] In a more recent example,[12,13] novaluron and pyriproxyfen were fed to hamsters and gerbils in the laboratory, to control sand fly (*Phlebotomus papatasi* and *Phlebotomus doboscqi*) larvae; in nature, these larvae live in burrows of rodents. Adult sand flies are vectors of leishmaniasis, an infection caused by protozoan parasites.

Although the feed-through method is very convenient, in most instances, IGRs are administered for pest control by contact or oral application. For social insects (for example, termites), IGRs mixed with food serve as a bait, which is carried by the workers to the colony. These are the practical methods of application, though it was reported that injection of an IGR may be up to thousand times more efficacious than its topical application.[14]

The effect of IGRs on insects is often delayed, becoming overt at the next molt or at one of the next stages of metamorphosis. Up to the effect, the pest inflicts further damage. Therefore, the ideal cases for usage of IGRs are insects that are harmless in the larval stages and only the adults are harmful (mosquitoes, horn flies, sand flies, etc.).

IGRs are often very selective. A compound may be lethal to an insect species but may have no effect on another. For example, buprofezin well affects sweet potato whitefly (*Bemisia tabaci*), but it has no effect on cotton bollworm (*Helicoverpa armigera*), whereas lufenuron affects both.[15] The selectivity is extremely important in practical pest control. Insect pests usually have specific or nonspecific insect predators and/or parasitoids ("parasitoid" means that only the larvae are parasites, the adults are not).

If some predators or parasitoids are more susceptible to an IGR than some pests, instead of eliminating pests, the natural enemies of some pests may be eliminated by using that IGR.[16] Larvae of Chrysopidae are beneficial insects, predators that feed on small insects, mostly on harmful aphids. *Chrysoperla externa* is highly susceptible to chitin synthesis inhibitors (CSIs) though not affected by a juvenile hormone analog (JHA).[17] There are also other beneficial insects; especially pollinators and their possible susceptibility must be taken into account. Pests may develop resistance and even cross-resistance to IGRs, and further application may be ineffective. Integrated pest management (IPM) tries to find the best solutions to these problems.[18,19]

It is important to note that usage of any IGR (like any insecticide or acaricide) needs an official permit. Such permits may be specific to a compound used against a specific pest or may be more general. Different countries and often different states within the same country may issue different permits. Therefore, the legal situation should be made clear before usage of an IGR. Thoms et al.[20] report a good example of bureaucratic troubles concerning the usage of an IGR as termite bait in Florida.

IGR Categories

IGRs are usually classified according to their mode of action.[1–4] Sometimes, however, terminology related to their chemical structure is also practiced. For example, most CSIs are benzoylphenyl urea derivatives, and this term is often used in the literature.[1–4,21] However, some other CSIs are not related to benzoylphenyl ureas. Juvenile hormone analogs (JHAs), also termed juvenoids or juvenile hormone (JH) mimics, constitute another category of IGRs. JHAs disturb metamorphosis,[1–4,22] resulting in nymph-adult, or larva-pupa, or pupa-adult, intermediate creatures that are not viable or at least unable to reproduce. Anti-JH agents also exist,[23,24] but these have not been developed to commercial insecticides.

Ecdysone agonists (EAs) constitute the relatively most recent category of IGRs[1,25]; they induce an untimely molt, which cannot be completed by the insect, eventually resulting in death.

There are also insecticides of botanical origin that interfere with insect-specific physiological processes. Azadirachtin[26,27] extracted from the seed kernels of the neem tree, *Azadirachta indica*, is an antifeedant and/or molt inhibitor, and it is often toxic to insects. Azadirachtin is not discussed here because it is a botanical insecticide, like natural pyrethrins or nicotine.

Naming IGRs

IGR names may confuse nonprofessionals. In early stages of development, a commercial firm uses a code, usually constructed from the abbreviated company name and a product number. Later, a name (sometimes called "common name" and after commercial production termed as "active ingredient") is given to the compound and its "chemical name", reflecting its chemical structure is revealed. The commercial product is a formulation that has a "trade name". Different formulations, often directed against different pests, but based on the same active ingredient, may have different trade names. For example, the JHA, pyriproxyfen, was developed by Sumitomo Chemical Company, Japan, under the code "S-31183". It received the common name "pyriproxyfen", and its chemical name is "2-[1-methyl-2-(4-phenoxyphenoxy)ethoxy] pyridine", or "4-phenoxyphenyl (*RS*)-2-(2-pyridyloxy) propyl ether", according to two different naming systems of chemicals. Its formulations were registered under the trade names "Sumilarv", "Admiral", and "Knack". Sometimes, production of an IGR is transferred to another commercial body that may come up with other trade names. For example, the firm Valent gave the trade names "Distance", "Esteem", and "Seize" to pyriproxyfen formulations. The label on a formulation should include the name and amount of the active ingredient. Details are found in *The Pesticide Manual*[28] under the name of this active ingredient (always use the latest editions).

For each IGR listed in this chapter, some (but not necessarily all) trade names are given in brackets.

Chitin Synthesis Inhibitors (CSIs)

As the name of these IGRs indicates, they inhibit chitin synthesis. Chitin is a major component of the insect body wall (cuticle), and it is present in the peritrophic membrane.[29] Chitin is a linear homopolymer of the aminosugar, N-acetyl-D-glucosamine. The molecules of the aminosugar are linked together, like a chain, in microfibrils, and these are arranged in layers (named "lamella"), constituting the chitin component of the cuticle. The biosynthesis of chitin from the aminosugar is catalyzed by the enzyme "chitin synthase".[30–32] CSIs inhibit the action of chitin synthase or interfere with insect chitin production by other ways, resulting in abnormal, malformed, cuticle. Consequently, the insect dies in or after the molt.[21,30,31] CSIs also have ovicidal effects; see the article by Boiteau and Noronha[33] for example. Insect embryos undergo molt within the eggs and/or at hatching, and chitin is an important component of the embryonic cuticle.

Vertebrates do not have chitin. However, chitin is present in many other living organisms; in some microbes, fungi, and in several animal phyla. There are many CSIs, but some of them are not effective as insecticides. CSIs and the spectrum of their activity are summarized by Merzendorfer.[34]

From the chemical aspect, two groups of CSIs may be considered affecting arthropods, benzoylphenyl ureas and non-benzoylphenyl ureas.

Benzoylphenyl Urea CSIs

There are several reviews on benzoylphenyl urea CSIs,[21,31] and some of them are more recent.[1,3,35,36] The latest review by Sun et al.[37] is very comprehensive. It discusses the history of discovering commercial benzoylphenyl ureas, starting with the diflubenzuron (under the trade name, Dimilin in 1975), structure–activity relationship, environmental fate, ecotoxicology, and action mechanism of CSIs. The review concludes [37] that "…the exact action site of benzoylurea insecticides has not been identified". By the way these CSIs do not inhibit chitin production in fungi.

A short description of the CSIs is presented below under the common name of each compound and in brackets some of the trade names and commercial formulations. This description was compiled from data in the above reviews, research articles on the subject in the last decades, and reliable Internet sources. References are usually not presented, except in some specific instances. Figure 1 shows the chemical structure of most benzoylphenyl urea CSIs.

Diflubenzuron [Dimilin, Micromite, Adept, and other trade names]. It is the oldest CSI with insecticidal activity, discovered in the early 1970s. It was also the first registered commercial CSI under the trade name Dimilin. It affects insects by topical application and/or ingestion. Diflubenzuron is used against a wide range of insects, especially against leaf-chewing species. A review is devoted to its use against locusts and grasshoppers.[38]

Bistrifluron [DBI-3204]. It is a relatively new CSI, which is used against pests from Lepidoptera (moths) and Hemiptera (whiteflies). Recently, it was tried successfully as a bait component against termites.[39]

Chlorbenzuron [Chlorbenzuron]. It is a Chinese product. Its common name is approved in China as a trade name and sold by several Chinese firms. It is not listed in *The Pesticide Manual*. Chlorbenzuron is used in China against insect pests, especially against pine caterpillars with little effect on parasitoids. Chlorbenzuron is essentially harmless to humans, but its degradation product, 2-chlorobenzamide, is suspected of being carcinogenic.[40]

Chlorfluazuron [Atabron, Ishipron, Aim, Helix, Jupiter]. It is a CSI developed in Japan. It is active through ingestion and used against chewing pests on cotton and vegetables, as well as against pests on fruits, potatoes, and tea. Interestingly, bait formulation with chlorfluazuron increased the soldiers/workers ratio in the subterranean termite, *Coptotermes curvignathus*, in Indonesia; colony elimination took about 6–8 weeks.[41]

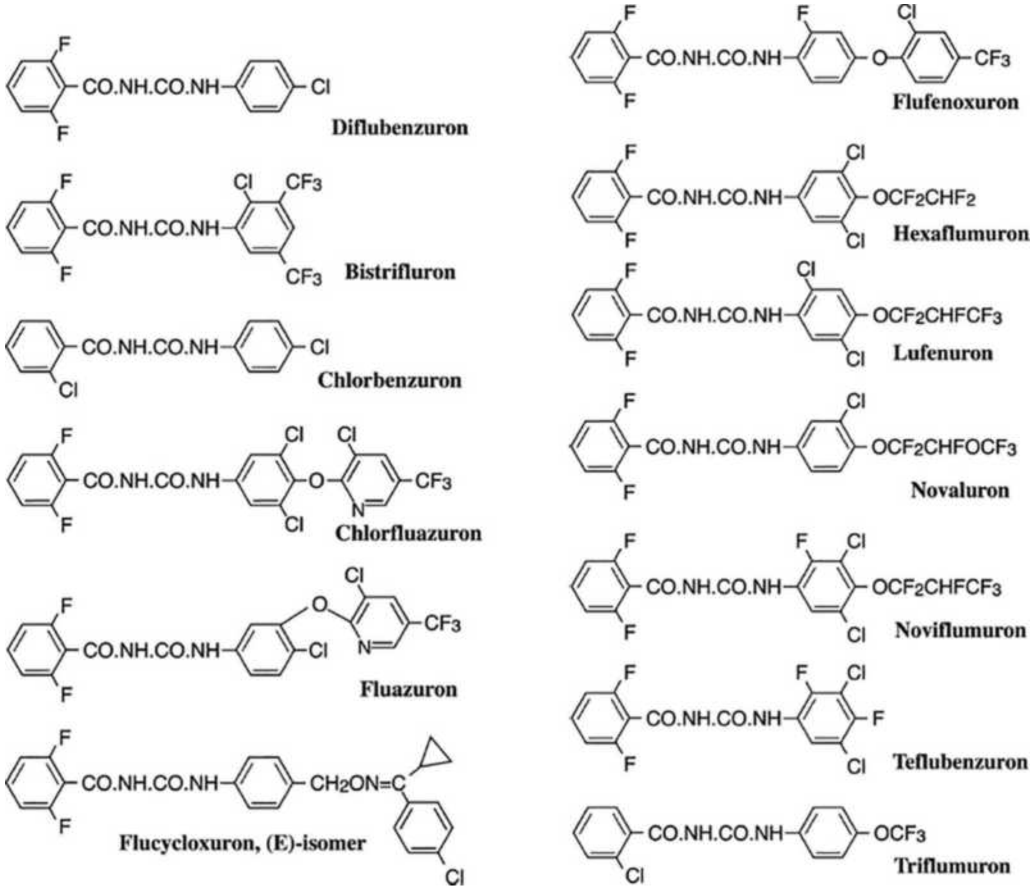


FIGURE 1 The chemical structure of benzoylphenyl urea CSIs.

Dichlorbenzuron [Dichlorbenzuron]. Although developed by Philips–Duphar, it is not listed in *The Pesticide Manual*. However, like chlorbenzuron (see above), it is approved and used as an insecticide in China.

Fluazuron [Acatak]. It is used mainly against ticks of the genus *Boophilus* and mites; therefore, it may be considered as an acaricide. Application of a pour-on formulation of fluazuron to cattle exerted no negative effects on the dung beetle, *Onthophagus gazella* in South Africa.[42] It is also used against ectoparasites, like fleas, on mammals.[43]

Flucycloxuron [Andalin]. This compound is both an acaricide and an insecticide. As an acaricide, it is effective against phytophagous mites. As an insecticide, it affects leaf-munching and leaf-rolling insects, mainly moth larvae on fruits, vegetables, cotton, and ornamentals.

Flufenoxuron [Cascade] (sometimes abbreviated as "flurox"). It is both an acaricide and an insecticide. The compound was found to be the most effective IGR against stored product mites; it also affects phytophagous mites. It controls many insect pests on fruit, tea, cotton, maize, and vegetables crops. In an artificial diet, it was lethal to the plant-sucking pea aphid, *Acyrthosiphon pisum*.[44]

Hexaflumuron [Consult, Recruit II, Trueno, Shatter, SentriTech as termite bait]. This compound is used mostly as an efficacious bait toxicant against subterranean termites. It is effective for control of insect larvae of Lepidoptera (including the Asiatic rice borer, also named striped stem borer, *Chilo suppressalis*), Coleoptera (including the Colorado potato beetle *Leptinotarsa decemlineata*), Homoptera, and Diptera.

Lufenuron [Match, Axor, Luroner, Luster, Manyi]. It is used mostly as a bait toxicant against subterranean termites, and in some instances, it was found more efficacious than hexaflumuron. Lufenuron is effective against larvae of moths (cotton bollworm, *H. armigera*, among others) and beetles on maize, vegetables, fruit, and cotton. It is effective against fleas and also acts as an acaricide.

Novaluron [Rimon, Diamond]. This compound acts by contact or ingestion. It is effective against whiteflies (*B. tabaci*), moth larvae (including species of the highly harmful genus, *Spodoptera*), and a wide range of insect pests on cotton, maize, certain fruits, citrus, potato, and vegetables. Its administration to adult female insects often reduces egg viability, up to complete ovicidal action. By feed-through method, it affects sand fly larvae in rodents' burrows.[12] Novaluron may also act as a mosquito larvicide.

Noviflumuron [Recruit III, Recruit IV, T-max as termite baits]. It is a relatively new (2004) CSI. It is active mostly as bait toxicant against several species of subterranean termites. Noviflumuron also affects cockroaches. In the German cockroach (*Blattella germanica*), it acts as a larvicide, and its ingestion as a bait by adults induces up to 100% ovicidal effect.

Teflubenzuron [Nomolt, Nemolt, Nobelroc, Dart, Diamond, Teflurate]. It is effective against whiteflies (*B. tabaci*), moth larvae (genera *Heliothis, Spodoptera*), and other pests on cotton. It also affects pests on fruits and potatoes (good effect on the Colorado potato beetle, *L. decemlineata*). An undesirable effect was observed on some sediment processing Polychaete worms in water.[45]

Triflumuron [Alsystin, Baycidal, Certero, Poseidon]. It affects pests, from various orders of insects, on fruits, soybean, cotton, and vegetables. It is used against ectoparasites on mammals, sheep lice, and others. Pour-on application was found highly effective against the biting louse, *Werneckiella equi*, of horses. It affected the blood-sucking bug, *R. prolixus* by oral, topical, or continuous contact (in coated Petri dishes) application. It affected larvae of mosquitoes.[46]

Search for novel benzoylphenyl urea derivatives. Research to discover novel benzoylphenyl urea-related CSIs has been maintained,[36,37] especially in China.[47–50] The usual procedure is to test the effect of the newly synthesized compound on certain insect(s) and compare the effect of one or more benzoylphenyl urea CSIs on the same insect(s).

Non-Benzoylphenyl Urea CSIs

These compounds chemically differ from benzoylphenyl urea CSIs. They inhibit chitin biosynthesis, not necessarily by inhibiting chitin synthase. Their exact mode of action is rather blurred, but the eventual effect is interference with chitin production and molting. This chapter shows the chemical structure of three compounds (Figure 2).

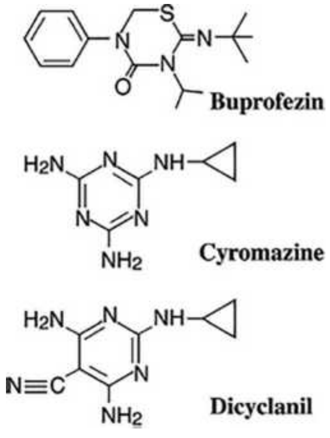


FIGURE 2 The chemical structure of non-benzoylphenyl urea CSIs.

Buprofezin [Applaud, Courier, Maestro, Mercao]. This compound is an insecticide that inhibits molt by contact or ingestion. It is used against homopteran insects, such as leafhoppers, plant hoppers, including the notorious rice plant hopper, *Nilaparvata lugens*, mealy bugs, jumping plant lice, and whiteflies, especially *B. tabaci*, which easily develop resistance.[51] Considerable resistance was found also in *N. lugens* in China.[52] It did not affect the ability of a predatory mite to suppress a pest mite in Californian vineyards.[53]

Cyromazine [Larvadex, Trigard, Neporex, Cliper, Ciromate, Vetrazin, Garland]. The most frequent use of cyromazine is against the housefly, *M. domestica*, in poultry farms. It is used by feed-through application or spray-on the manure.[8,9] It is also used to control leaf miners in vegetables, potatoes (including the Colorado potato beetle, *L. decemlineata*, and mushroom sciarid fly, *Lycoriella ingenua*).

Dicyclanil [Clik]. It is used mostly as a veterinary pharmaceutical agent against maggots (fly larvae) that infect traumatic myiasis in livestock. Dicyclanil is employed for prophylactic treatment against flesh flies, screwworm flies, Australian blowfly, and other flies, the maggot of which infect and/or live in wounds. Effective by pour-on-formulation induces hepatocarcinogenesis (liver cancer) in mice.

Etoxazole [Baroque, Sorado]. It was shown that etoxazole inhibits chitin biosynthesis.[54] Etoxazole is used as an acaricide; it affects eggs, larvae, and nymphs but not adults (which do not molt) of the spider mite species belonging to the genera *Tetranychus* and *Panonychus*. Unfortunately, resistance is reported in Japan, South Korea, and Australia.[55]

Based on genetic studies, Demaeght et al.[56] concluded that **hexythiazox** and **clofentezine**, well-known acaricides, like etoxazole, affect chitin synthase 1. If so, these may be considered as CSIs.

Juvenile Hormone and Its Analogs (JHAs)

The Roles of Juvenile Hormone in Insects

Insect metamorphosis is under endocrine control. The presence of JH in subadult stages (nymphs or larvae) prevents metamorphosis to adult. JH is secreted by a pair of endocrine glands, the corpora allata. In the last preadult instar, cessation of the activity of the corpora allata results in a temporary absence of JH, allowing completion of the metamorphosis. In the adult, the corpora allata resumes activity and the JH is involved in reproduction. There are several natural JHs in insects, the most common is JH III (Figure 3), but all have a similar sesquiterpenoid basic structure. No JH exists in vertebrates.

JH and its mode of action have been extensively reviewed[57–61] from the level of organism, including effects on cast differentiation in some social insects, to the level of molecular biology.

Juvenile Hormone Analogs

Juvenile hormone analogs (JHAs), also termed “JH mimics”, or “juvenoids”, are compounds that exert similar effects to that of the natural JH. Treatment of late nymphal or larval instars with JHAs prevents metamorphosis to adults. The chemical structure of JHAs may or may not resemble natural JH. There are thousands of chemically different compounds that exert JHA activity, many of them known for long time.[62] A JHA is often specific to a group of insects. Only some JHAs were developed to commercial products.

In the course of normal development, JH disappears only in the last larval instar and this instar constitutes the target of JHAs. However, treatment of earlier instars may also be effective as environmental degradation of JHAs takes some time and meanwhile the insect may reach the last larval instar. Also, metabolic inactivation of JHAs is usually slower than that of natural JH, and a JHA absorbed in an earlier instar is not necessarily inactivated before the insect reaches the last larval instar. Often sustained (slow) release formulations are used. JHAs are usually effective by contact; they penetrate the cuticle. JHAs prevent development of normal adults; larva- or pupa-adult intermediate creatures are obtained

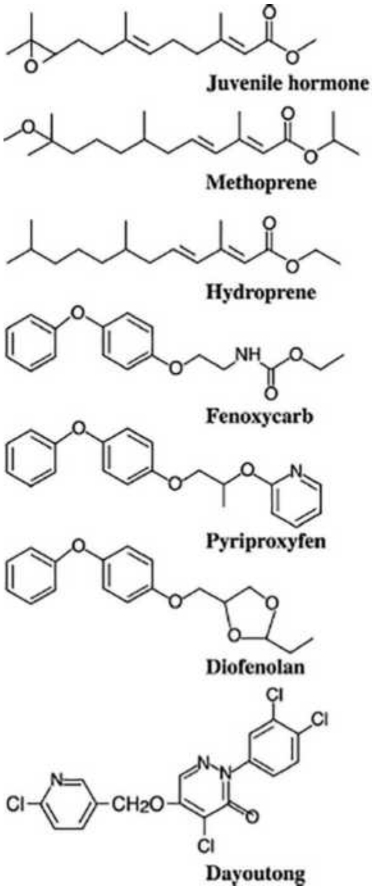


FIGURE 3 The chemical structure of the most common JH in insects and JHAs.

that die in the molt, or in the next (supernumerary) molt, or survive but are unable to reproduce. JHAs do not prevent damage caused by larvae and in some instances induce viable supernumerary larvae with prolonged life span and increased damage.[63]

JHAs, their effects, and their role in insect pest management were repeatedly reviewed.[1,3,22,25,62,64–70] These reviews also outline the roles of JH to explain the effects of JHAs. JHAs are also used for molecular basic research related to metamorphosis.[63,67–70]

Methoprene [Altosid, Precor, Apex, Dianex, Aquaprene, Kabat, Pharrorid, Fleatrol, Ovitrol, Extinguish, Diacon]. There are many formulations of methoprene, such as wettable powder, tossits, granules, and briquettes. Together with hydroprene, it was developed by Zoecon Company, constituting the first JHA approved for commercial usage. The first commercial products were racemic mixtures of the isomers of methoprene, but later only the 7S-isomer has been used. It is effective against mosquitoes, horn flies (feed-through method), sciarid flies (pests in mushroom cultures), fleas, some stored product insects, and fire ants, *Solenopsis invicta*. The compound has a low persistence in most field conditions; therefore, usually slow release technology is employed.[11] Methoprene was reviewed by Henrick.[11] Recently, control of stored product insects by methoprene was reviewed.[71]

Hydroprene [Altozar, Gencor, Gentrol, Raid]. It was approved at the same time together with methoprene for commercial usage. The 7S-isomer is used. It affects cockroaches, stored product insects, fruit flies, bed bugs, and some beetles. For stored product insects, pyriproxyfen (see below) seems to be more efficacious than hydroprene.[72]

Kinoprene [Enstar II]. The 7S-isomer is used. It is effective against homopterous pests. It was reported that kinoprene affects aphid polymorphism.[73]

Fenoxycarb [Insegar, Logic, Torus, Pictyl, Varikill, Comply]. It affects a very wide range of insects, disturbing metamorphosis and/or acting as an ovicide, directly applied to the eggs or parent females. Susceptible species were reviewed already in 1993.[74] It may be toxic to non-target species, possibly including honey bee brood.[75] Its chemistry and environmental fate were recently reviewed.[76]

Pyriproxyfen [Sumilarv, Admiral, Knack, Distance, Esteem, Seize, Tiger, Nylar, Intracure]. It is used against scale insects, whiteflies, aphids, pear psylla, thrips, cockroaches, codling moth, stored product insects, mosquitoes, and fire ants. The literature on its effect on the whitefly, *B. tabaci*, is especially rich, including resistance-related publications. The compound disturbs metamorphosis and/or interferes with reproduction. It may be toxic to some aquatic organisms. Its properties and environmental fate were reviewed.[77]

Diofenolan [Aware]. This JHA is less widely used than the above compounds. It was classified by some authors as a “molt inhibitor”, but it is a JHA. Insects that are susceptible to diofenolan were listed by Streibert et al.[78]

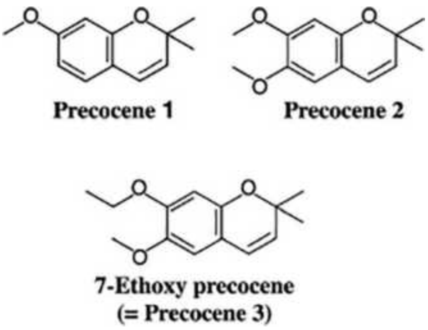
Dayoutong. This compound has no trade name. It is a recently developed JHA approved under the name dayoutong in China.

Others. Triprene and Epofenonane are superseded JHAs.

Anti-Juvenile Hormone Agents

These are compounds that interfere with JH production,[23,24] but none of them have been developed to a commercial insecticide. Only the case of precocenes is mentioned here, because of their specific mode of action. The anti-JH effects of precocenes 1 and 2 (Figure 4) were discovered by Bowers and coworkers[79]; they extracted it from the plant *Ageratum houstonianum*. Many chemical derivatives, including the more active 7-ethoxy precocene (Figure 4), also called precocene 3, were synthesized and tested in several laboratories. Precocenes do not affect JH but selectively causes atrophy of the corpora allata,[80] the endocrine glands that secrete the JH. Susceptible species are limited, mostly hemipteran and orthopteran insects. Applied topically to early preadult nymphal instars, miniature nymph-adult intermediate creatures are obtained. Application of precocene to young adults also leads to the absence of JH that interferes with reproduction, especially with oocyte development. Precocenes are not mutagenic, but they are hepatotoxic and nephrotoxic to vertebrates and suspected to be carcinogenic.

Interference with JH biosynthesis, or enzymatic JH regulation, is considered as possible anti-JH procedures.[66]



Precocene 1

Precocene 2

7-Ethoxy precocene
(= Precocene 3)

FIGURE 4 The chemical structure of precocenes that act as anti-juvenile hormone agents.

Ecdysteroids and Their Agonists

The Role of Ecdysteroids in Insects

Ecdysone, a steroid, serves as a prohormone of the molting hormone. It is secreted by a pair of endocrine glands, termed “prothoracic glands” (though in some insects they are not located in the prothorax but in the ventro-posterior part of the head). In certain tissues, ecdysone is converted to 20-hydroxyecdysone (Figure 5) which is the actual molting hormone. To initiate molting, a peak titer of 20-hydroxyecdysone is necessary, but for the full process of molting, a complete decline of this peak is needed. This decline leads to activation of certain genes that control production of necessary enzymes and secretion of additional hormones, which induce molting-related behavior. Ecdysone, 20-hydroxyecdysone, and chemically related substances, collectively termed as “ecdysteroids” are present in many plants and non-vertebrate animals.

The prothoracic glands usually degenerate in adult insects; if persist, they are inactive. In adult insects, the ovaries become the source of ecdysteroids and they transfer ecdysteroids, usually as conjugates, to the developing oocytes. An outline of the chemistry and physiological role of ecdysteroids, aimed primarily to under- or postgraduate students, is presented in Klowden’s book.[81] A review dealing mostly with the mode of action of ecdysteroids at the molecular level was recently published.[82]

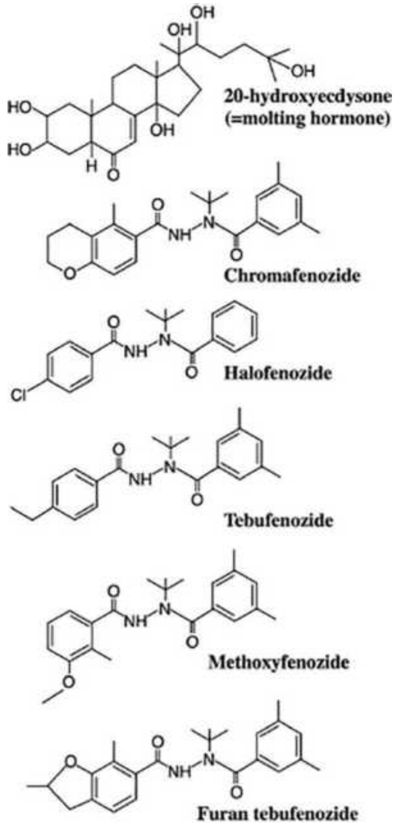


FIGURE 5 The chemical structures of 20-hydroxyecdysone (molting hormone) and nonsteroidal ecdysone (ecdysteroid) agonists.

Ecdysone (Ecdysteroid) Agonists (EAs)

Reviews devoted to EAs[1,3,25,83] usually describe and discuss the role of ecdysone in normal molting. As already outlined, a peak level and then a decline of the peak of ecdysone are needed for molting. The presently existing commercial EAs (Figure 5) bind strongly to ecdysteroid receptors, conveying a false message that no decline of 20-hydroxyecdysone occurred. Consequently, these EAs induce upregulation of all genes that occur in normal molting due to the 20-hydroxyecdysone peak; however, they do not activate those genes, which are upregulated in normal molting by the decline of the peak. The result is an incomplete, unfinished molt which is lethal. EAs are used against larvae, but they often also exert ovicidal effect.

From the chemical standpoint, commercial EAs are bisacylhydrazines, also termed diacylhydrazines. In contrast to ecdysteroids, bisacylhydrazines are not steroids. *In vivo* they are more stable than ecdysteroids and their effect may be regarded as "hyperecdysonism".[3,84] EAs are usually applied by ingestion. Their application is effective any time, except at natural molting. An advantage of the EAs is feeding inhibition of susceptible larvae, preventing further damage. Stopping of feeding is one of the first events in normal molting. The current commercial EAs are very selective and affect only some groups of insects. The reason for this selectivity seems to be differences in the ligand-binding domain of ecdysone receptors in different insects.

Chromafenozide [Matric, Killot, Pestanal]. This compound was developed in Japan, and it specifically affects highly harmful lepidopteran pests on rice, vegetables, fruit, vine, and other crops.

Halofenozide [Mach 2]. It was developed by Rohm and Haas Company. Halofenozide has a broader spectrum of susceptible insects than other commercial EAs. It affects both lepidopteran and coleopteran larvae. However, specific selectivity exists even within these two orders of insects. Despite its activity on some coleopteran larvae, halofenozide has a relatively low binding affinity to coleopteran ecdysteroid receptors.

Methoxyfenozide [Intrepid, Runner, Prodigy, Falcon]. Again it was developed by Rohm and Haas Company. Similar to chromafenozide and tebufenozide, it controls only lepidopteran larvae. Like in other EAs, efficacy depends on the species. Methoxyfenozide is effective against bollworms, leaf miners, and diamond moth. Recently, it was claimed that methoxyfenozide has some potential for mosquito control.[85]

Tebufenozide [Confirm, Mimic, Fimic, Romdan]. This EA was again developed by Rohm and Haas. It affects lepidopteran larvae, similar to methoxyfenozide, but species-dependent differences may exist. Tebufenozide has poorer ovicidal activity than methoxyfenozide.

Fufenozide [Fuxian]. This compound was developed in China under the code name JS-118; it is also named as furan tebufenozide. It is effective against the Asiatic rice borer, *C. suppressalis* and other lepidopteran larvae. Its degradation by hydrolysis and photolysis was recently studied.[86]

Research on bisacylhydrazine-type compounds were published,[87] and steroid-like compounds were investigated[88] for obtaining new insecticides. Non-bisacylhydrazine nonsteroidal EAs are available,[89,90] but none of them developed to commercial insecticide.

Advantages of IGRs

IGRs are nontoxic or slightly toxic (often as low as at g/kg level) to most vertebrates, especially to mammals and birds. Fish may be more susceptible to toxic effects of IGRs. Details of toxicity are presented in *The Pesticide Manual*[28] under the name of each compound separately.

Parasitoid wasps are usually less susceptible to IGRs than to conventional insecticides.[91–95] Moreover, in most (but not all) instances, parasitoid and predatory insects are less affected by IGRs than their respective harmful insect hosts and preys. This seems to be the case also for mites; the pest mite is more susceptible to IGR toxicity than its predatory mite.[53,96] Treatment of livestock with IGRs by feed-through or pour-on methods of application has no notable effect on dung beetles.[42]

The most important advantage of IGRs is that they affect only a relatively narrow range of susceptible species. IGRs are certainly friendlier to the environment and make less damage to non-target species than conventional insecticides that are usually toxic to a wide range of animals, including humans.

Disadvantages of IGRs

The environmental advantages of IGRs (see above) may be disadvantageous economically. The more specific the IGR, the more limited its marketing. Development of an IGR to a commercial product, with permits in many countries, is expensive, and if the expected marketing is limited, a commercial firm would not invest the cost of development.

Although in most instances, parasitoid and predatory insects exhibit low susceptibility to IGRs, this is not always so. Some parasitoid insects may be affected by some IGRs, even when non-IGR insecticides are more toxic than IGRs.[93] IGRs may be harmful to predatory insects.[17,97,98] Also, IGRs may affect other non-target organisms, especially crustaceans.

Besides the direct role of IGRs in parasitoid–pest or predator–pest relations, lethal and sublethal effects of IGRs on beneficial insects, such as pollinators, should also be considered. The sublethal effects of insecticides, including IGRs, on beneficial arthropods were reviewed.[99]

The best-known pollinators are the honey bees, bumblebees, and other bee species. Claims concerning effects of IGRs on bees are contradictory. CSIs, including novaluron, did not affect adult bumblebees (*Bombus terrestris*) but dramatically reduced brood production.[100] Other authors, however, found that novaluron did not affect the brood of this species.[101] An earlier publication reported that some CSIs are toxic to immature stages of bees.[102] Recent studies confirmed that diflubenzuron and novaluron (CSIs) are toxic to the progeny of honey bees and alfalfa leafcutting bees (*Megachile rotundata*).[103–106] Mommaerts and coauthors concluded that fenoxycarb (JHA) and methoxyfenozide (EAs) are safe to bumblebees, whereas pyriproxyfen and kinoprene (JHAs) affect the brood at the maximum field-recommended concentration.[107] Pyriproxyfen was found to be toxic to honey bees in recent studies too.[108,109] In contrast, fenoxycarb was claimed to negatively affect adult emergence in honey bees.[75] An older review on the effect of IGRs on some bees[110] distinguishes between different species and between different doses.

An additional disadvantage is that IGRs are effective only in special stages or instars of the pests: CSIs and EAs before natural molt, and JHAs in the last larval instar. Therefore, IGRs do not prevent some or all damage by the larvae if these larvae are harmful. In extreme instances, JHAs induce supernumerary larval instars[63] with longer-than-normal larval life span and additional damage.

Resistance to IGRs

Insects are capable of developing resistance to any kind of insecticides, and IGRs are not exceptions. Hundreds of scientific publications are devoted to acquired resistance of insects to IGRs.

Several interrelated approaches to IGR resistance may be considered (these do not differ from those to conventional insecticides). One of them is studying field-collected insects; the susceptibility to an IGR of a sample of insects collected in the field is assessed and compared to that of a non-resistant strain maintained in the laboratory or supplied by international or national organizations (WHO, etc.). A much higher dose to affect the field sample than that induces the same effect in the non-resistant strain means that the insects of the field sample are resistant. The effective dose for the field sample divided by that obtained for the non-resistant strain yields the “resistance ratio” (RR), which can be low but can be as high as several thousands. RR values may differ for different populations of the same species within the same country, as exemplified by Kristensen and Jespersen.[111]

Another approach is to expose a non-resistant laboratory strain to an IGR and select the survivors over subsequent generations with increasing doses. The effective dose obtained for the last generation divided by that of the initial dose results in the relevant RR. This reveals the resistance potential of the pest.[111] Obviously, this approach can be combined with the former one by taking an already resistant strain from the field, then expose subsequent generations to increasing doses. The RR for pyriproxyfen resistance of the whitefly, *B. tabaci*, was found to be over 2000[112] and can be as high as 7000.[113]

The third approach is to investigate cross-resistance, meaning that a strain resistant to an IGR shows resistance to other IGRs without formerly being exposed to these IGRs.[111,112] Sometimes, resistance to a conventional, non-IGR insecticide may result in cross-resistance to IGRs.[18]

Other approaches are studies on the inheritance of resistance,[113,114] residual activity,[115] mechanisms of resistance at the molecular level, and gene expressions.[116,117]

Reversal of resistance is studied by maintaining a resistant strain in the laboratory for many subsequent generations, without exposure to the IGR to which the strain is resistant, and assessing the residual resistance of the last generation.[18,19] Reversal of resistance does occur, but it is seldom absolute. Also, when a strain that experienced reversal of resistance is exposed again to the same IGR, it develops resistance more quickly than a strain that is originally non-resistant.

Integrated Pest Management

IPM means employment of multiple control tactics to reduce pest population and usage of chemicals, as well as decrease development of resistance, all with minimal economic cost and environmental damage.[118] The IPM for each pest, in each environment, should be worked out. Efficacy and resistance potential of different insecticides, effects on non-target species, including parasitoids and/or predators of the pest, screening the pest (below a certain level of population density no treatment is needed), and effects on the environment should be considered. Usually IGRs are more efficacious and less harmful to the environment than conventional insecticides. However, this generalization should be verified by carefully gathered actual data. IPM partially solves the problem of resistance, especially resistance to a novel IGR, by alternating IGRs, for example, pyriproxyfen and buprofezin.[51] Such alternation tactics may also include non-IGR insecticides, for example, alternation of novaluron, pyriproxyfen, and neonicotinoids as suggested by Ishaaya and coauthors.[119] The effects of each alternating compound on non-target species and the environment should be studied; the alternating compounds should not induce cross-resistance. In many instances, there is more than one pest harmful to the same object. If so, it is preferable to work out such alternation of IGRs in which one or more compound exert toxic activity on more than just one pest.[15,119]

Obviously, IPM includes not only chemical factors but also biological ones, as exemplified by Mouden et al.[120] and Wilson et al.[121] In the European Union, IPM is compulsory, though its implementation is problematic.[122]

Conclusions

IGRs are insecticides that interfere with insect-specific physiological systems that do not exist in vertebrates, including humans. Consequently, IGRs are either nontoxic or slightly toxic to vertebrates. Within the insect world, IGRs may affect different species differently, even in the same family.[123] The narrow range of susceptibility to IGRs is a great environmental advantage over conventional insecticides, which are usually toxic to a much wider spectrum of species and much more harmful to the environment. However, IGRs may also affect non-target species, parasitoids, and/or predators of the pest, or of other pests, or other beneficial insects, such as pollinators, and even some non-insect species. Therefore, adequate research is needed to estimate the effect of an IGR on one or more pests, beneficial and non-target organisms, and the surrounding environment.

Insects may develop resistance or even cross-resistance to one or more IGRs, which can be prevented, or at least delayed, by IPM. In this respect, IGRs and conventional insecticides are similar.

Despite the advantages of IGRs, there are some disadvantages. Interference with a physiological system means that the insect is susceptible when that physiological system is active. IGRs interfere with production of new cuticle at molt, or disturb metamorphosis, or prevent reproduction. So, the toxic effects of IGRs become overt at the next molt, or in the next or subsequent instars, and meanwhile the pest inflicts additional damage. The ideal targets for IGRs are insects harmless in preadult stages and harmful only as adults (for example, mosquitoes). The belated action is an additional disadvantage; the farmer, or the sanitarian, or anybody who uses IGRs, do not see immediate knockdown of the pest, and doubt, therefore, the efficacy of the IGR.

New IGRs aimed to attack currently susceptible physiological systems are synthesized, tested, and possibly developed to commercial products. However, it is feasible to assume that in the future, additional IGRs and their effects, as well as various approaches attacking other insect-specific physiological or molecular systems, will be discovered and implemented.[124,125]

References

1. Pener, M. P., and T. S. Dhadialla. 2012. An overview of insect growth disruptors; applied aspects. *Adv. Insect Physiol.* 43: 1–162.
2. Pener, M. P. 2013. Insect growth regulators. In *Encyclopedia of Environmental Management*, ed. S. E. Jorgensen, vol. II, 1459–1470. New York: Taylor & Francis.
3. Dhadialla, T. S., A. Retnakaran, and G. Smagghe. 2010. Insect growth- and development-disrupting insecticides. In *Insect Control: Biological and Synthetic Agents*, ed. L. I. Gilbert, and S. S. Gill, 121–184. New York: Elsevier.
4. Subramanian, S., and K. Shankarganesh. 2016. Insect hormones (as pesticides). In *Ecofriendly Pest Management for Food Security*, ed. B. K. Omkar, 613–650. London: Academic Press.
5. Oberlander, H., D. L. Silhacek, E. Shaaya, and I. Ishaaya. 1997. Current status and future perspectives of the use of insect growth regulators for the control of stored product insects. *J. Stored Prod. Res.* 33: 1–6.
6. Collins, D. A. 2006. A review of alternatives to organophosphorus compounds for the control of storage mites. *J. Stored Prod. Res.* 42: 395–426.
7. Alzogaray, R. A., and E. N. Zerba. 2017. *Rhodnius prolixus* intoxicated. *J. Insect Physiol.* 97: 93–113.
8. Axtell, R. C., and T. D. Edwards. 1983. Efficacy and nontarget effects of Larvadex® as a feed additive for controlling house flies in caged-layer poultry manure. *Poult. Sci.* 62: 2371–2377.
9. Anderson, J. F., A. Spandorf, L. A., Magnarelli, and W. Glowa. 1986. Control of house flies in commercial poultry houses in Connecticut. *Poult. Sci.* 65: 837–844.
10. Harris, R. L., E. D. Frazer, and R. L. Younger. 1973. Horn flies, stable flies and house flies: development in feces of bovines treated orally with juvenile hormone analogues. *J. Econ. Entomol.* 66: 1099–1102.
11. Henrick, C. A. 2007. Methoprene. *J. Am. Mosq. Control Assoc.* 23 (supplement 2): 225–239.
12. Mascari, T. M., and L. D. Foil. 2010. Laboratory evaluation of novaluron as a rodent feed-through insecticide against sand fly larvae. *J. Med. Entomol.* 47: 205–209.
13. Mascari, T. M., R. W. Stout, and L. D. Foil. 2012. Evaluation of three feed-through insecticides using two rodent and two sand fly species as models. *J. Am. Mosq. Control Assoc.* 28: 260–262.
14. Pener, M .P., A. Ayali, G. Kelmer, B. Bennettová, V. Němec, M. Rejzek, and Z. Wimmer. 1997. Comparative testing of several juvenile hormone analogues in two species of locusts, *Locusta migratoria migratorioides* and *Schistocerca gregaria*. *Pestic. Sci.* 51: 443–449.
15. Gogi, M. D., R. M. Sarfraz, L. M. Dosdall, M. J. Arif, A. B. Keddie, and M. Ashfaq. 2006. Effectiveness of two insect growth regulators against *Bemisia tabaci* (Gennadius) (Homoptera: Aleyrodidae) and *Helicoverpa armigera* (Hübner) (Lepidoptera: Noctuidae) and their impact on population densities of arthropod predators on cotton in Pakistan. *Pest Manag. Sci.* 62: 982–990.
16. Hattingh, V. 1996. The use of insect growth regulators – Implications for IPM with citrus in southern Africa as an example. *Entomophaga* 41: 513–518.

17. Zotti, M. J., A. D. Grutzmacher, I. H. Lopes, and G. Smagghe. 2013. Comparative effects of insecticides with different mechanisms of action on *Chrysoperla externa* (Neuroptera: Chrysopidae): lethal, sublethal and dose-response effects. *Insect Sci*. 20: 743–752.
18. Reuveny, H., and E. Cohen. 2004. Resistance of the codling moth *Cydia pomonella* (L.) (Lep., Tortricidae) to pesticides in Israel. *J. Appl. Entomol.* 128: 645–651.
19. Wilson, M., P. Moshitzky, E. Laor, M. Ghanim, A. R. Horowitz, and S. Morin. 2007. Reversal of resistance to pyriproxyfen in the Q biotope of *Bemisia tabaci* (Hemiptera: Aleyrodidae). *Pest Manag. Sci*. 63: 761–768.
20. Thoms, E. M., J. E. Eger, M. T. Messenger, E. Vargo, B. Cabrera, C. Riegel, S. Murphree, J. Mauldin, and P. Scherer. 2009. Bugs, baits and bureaucracy: completing the first termite bait efficacy trials (quarterly replenishment of noviflumuron) initiated after adoption of Florida rule, chapter 5E-2.0311. *Am. Entomol*. 55: 29–39.
21. Wright, J. E., and A. Retnakaran. ed. 1987. *Chitin and Benzoylphenyl Ureas*. Series.Entomologica 38. Dordrecht: Dr W. Junk Publisher.
22. Minakuchi, C., and L. M. Riddiford. 2006. Insect juvenile hormone action as a potential target of pest management. *J. Pestic. Sci*. 31: 77–84.
23. Staal, G. B. 1986. Anti juvenile hormone agents. *Annu. Rev. Entomol*. 31: 391–429.
24. Ghoneim, K., and R. F. Bakr. 2018. Physiological activities of anti-juvenile hormone agents against insects and their role devising fourth generation insecticides: a comprehensive review. *Egypt. Acad. J. Biol. Sci. A. Entomol*. 11: 45–138.
25. Dhadialla, T. S., G. R. Carlson, and D. P. Le. 1998. New insecticides with ecdysteroidal and juvenile hormone activity. *Annu. Rev. Entomol*. 43: 545–569.
26. Mordue (Luntz), A. J., E. D. Morgan, and A. J. Nisbet. 2010. Azadirachtin, a natural product in insect control. In *Insect Control: Biological and Synthetic Agents*, ed. L. I. Gilbert and S. S. Gill, 185–203. London: Elsevier, Academic Press.
27. Mordue (Luntz), A. J., E. D. Morgan and A. J. Nisbet. 2010. Addendum: azadirachtin, a natural product in insect control: an update. In *Insect Control: Biological and Synthetic Agents*, ed. L. I. Gilbert, and S. S. Gill, 204–206. London: Elsevier, Academic Press.
28. Turner, J. A. ed. 2018. *The Pesticide Manual*, 18th ed. Aldershot Hampshire: British Crop Production Council.
29. Kelkenberg, M., J. Odman-Naresh, and S. Muthukrishnan. 2015. Chitin is a necessary component to maintain the barrier function of the peritrophic matrix in the insect midgut. *Insect Biochem. Mol. Biol*. 56: 21–28.
30. Cohen, E. 1987. Chitin biochemistry: synthesis and inhibition. *Annu. Rev. Entomol*. 32: 71–93.
31. Cohen, E. 2001. Chitin synthesis and inhibition: a revisit. *Pest Manag. Sci*. 57: 946–950.
32. Muthukrishnan, S., H. Merzendorfer, Y. Arakane, and K. J. Kramer. 2012. Chitin metabolism in insects. In *Insect Molecular Biology and Biochemistry*, ed. L. I. Gilbert, London: Elsevier, Academic Press, 193–235.
33. Boiteau, G., and C. Noronha. 2007. Topical, resudial and ovicidal contact toxicity of three reduced-risk insecticides against the European corn borer, *Ostrinia nubilalis* (Lepidoptera: Crambidae), on potato. *Pest Manag. Sci*. 63: 1230–1238.
34. Merzendorfer, H. 2013. Chitin synthesis inhibitors: old molecules and new developments. *Insect Sci*. 20: 121–138.
35. Matsumura, F. 2010. Studies on the action mechanism of benzoylurea insecticides to inhibit the process of chitin synthesis in insects: a review on the status of research activities in the past, the present and the future prospects. *Pestic. Biochem. Physiol*. 97: 133–139.
36. Doucet, D., and A. Retnakaran. 2012. Insect chitin: metabolism, genomics and pest management. *Adv. Insect Physiol*. 43: 437–511.
37. Sun, R., C. Liu, H. Zhang, and Q. Wang. 2015. Benzoylurea chitin synthesis inhibitors. *J. Agric. Food Chem*. 63: 6847–6865.

38. Weiland, R.T., F. D. Judge, T. Pels, and A. C. Grosscurt. 2002. A literature review and new observations on the use of diflubenzuron for control of locusts and grasshoppers throughout the world. *J. Orthoptera Res.* 11: 43–54.
39. Webb, G. A. 2017. Efficacy of bistrifluron termite bait on *Coptotermes lacteus* (Isoptera: Rhinotermitidae) in southern Australia. *J. Econ. Entomol.* 110: 1705–1712.
40. Liu, G., Q. Zhou, and W. Lu. 2001. The formation of 2-chlorobenzamide upon hydrolysis of the benzoylphenylurea insecticide 1-(2-chlorobenzoyl)-3-(4-chlorophenyl) urea in different water systems. *J. Agric. Food Chem.* 49: 1304–1308.
41. Sukartana, P., G. Sumarni, and S. Broadbent. 2009. Evaluation of chlorfluazuron in controlling the subterranean termite *Coptotermes curvignathus* (Isoptera: Rhinotermitidae) in Idonesia. *J. Trop. For. Sci.* 21: 13–18.
42. Kryger, U., C. Deschodt, A. L. V. Davis, and C. H. Scholtz. 2007. Effects of cattle treatment with fluazuron pour-on on survival and reproduction of the dung beetle species *Onthophagus gazella* (Fabricius). *Vet. Parasitol.* 143: 380–384.
43. Davis, R. M., E. Cleugh, R. T. Smith, and C. L. Fritz. 2008, Use of a chitin synthesis inhibitor to control fleas on wild rodents important in the maintenance of plague, *Yersinia pestis* in California. *J. Vector Ecol.* 33: 278–284.
44. Sadeghi, A., E. J. M. Van Damme, and G. Smagghe. 2009. Evaluation of the susceptibility of the pea aphid, *Acyrthosiphon pisum*, to a selection of novel biorational insecticides using an artificial diet. *J. Insect Sci.* 9: article 65.
45. Méndez N. 2005. Effects of teflubenzuron on larvae and juveniles of the polychaete *Capitella* sp. B from Barcelona, Spain. *Water Air Soil Poll.* 160: 259–269.
46. Belinato, T. A., A. J. Martins, J. B. P. Lima, and D. Valie. 2013. Effect of triflumuron, a chitin synthesis inhibitor, on *Aedes aegypti, Aedes albopictus* and *Culex quinquefasciatus* under laboratory conditions, *Parasit. Vectors* 6: article 83.
47. Chen, L., Q. Wang, R. Huang, C. Mao, J. Shang, and F. Bi. 2005. Synthesis and insecticidal evaluation of propesticides of benzoylphenylureas. *J. Agric. Food Chem.* 53: 38–41.
48. Sun, R., M. Lü, L. Chen, Q. Li, H. Song, F. Bi, R. Huang, and Q. Wang. 2008. Design, synthesis, bioactivity and structure–activity relationship (SAR) studies of novel benzoylphenylureas containing oxime ether group. *J. Agric. Food Chem.* 56: 11376–11391.
49. Sun, R., Y. Zhang, F. Bi, and Q. Wang. 2009. Design, synthesis, and bioactivity study of novel benzoylpyridazyl ureas. *J. Agric. Food Chem.* 57: 6356–6361.
50. Zhang, J., X. Tang, I. Ishaaya, S. Cao, J. Wu, J. Yu, H. Li, and X. Qian. 2010. Synthesis and insecticidal activity of heptafluoroisopropyl-containing benzoylphenylurea structures. *J. Agric. Food Chem.* 58: 2736–2740.
51. Horowitz, A. R., G. Forer, and I. Ishaaya. 1994. Managing resistance in *Bemisia tabaci* in Israel with emphasis on cotton. *Pestic. Sci.* 42: 113–122.
52. Zhang, X., X. Liu, F. Zhu, J. Li, H. You, and P. Lu. 2014. Field evaluation of insecticide resistance in the brown plathopper (*Nilaparvata lugens* Stål) in China. *Crop Prot.* 58: 61–66.
53. Stavrinides, M. C., and N. J. Mills. 2009. Demographic effects of pesticides on biological control of Pacific spider mite (*Tetranychus pacificus*) by the western predatory mite (*Galendromus occidentalis*). *Biol. Control* 48: 267–273.
54. Nauen, R., and G. Smagghe. 2006. Mode of action of etoxazole. *Pest Manag. Sci.* 62: 379–382.
55. Herron, G. A., L. K. Wooley, K. I. Langfield, and Y. Chen. 2018. First detection of etoxazole resistance in Australian two-spotted mite *Tetranychus urticae* Koch (Acarina: Tetranychidae) via bioassay and DNA methods. *Austral. Entomol.* 57: 365–368.
56. Demaeght, P., E. J. Osborne, J. Odman-Naresh, M. Grbić, R. Nauen, H. Merzendorfer, R. M. Clark, and T. Van Leeuwen. 2014. High resolution genetic mapping uncovers chitin synthase-1 as the target site of the structurally diverse mite growth inhibitors clofentezine, hexythiazox and etoxazole in *Tetranychus urticae*. *Insect Biochem. Mol. Biol.* 51: 52–63.

57. Tobe, S. S., and B. Stay. 1985. Structure and regulation of the corpus allatum. *Adv. Insect Physiol.* 18: 305–432.
58. Riddiford, L. M. 1994. Cellular and molecular actions of juvenile hormone I. General considerations and premetamorphic actions. *Adv. Insect Physiol.* 24: 213–274.
59. Wyatt, G. R., and K. G. Davey. 1996. Cellular and molecular actions of juvenile hormone. II. Roles of juvenile hormone in adult insects. *Adv. Insect Physiol.* 26: 1–155.
60. Riddiford, L. M. 2008. Juvenile hormone action: a 2007 perspective. *J. Insect Physiol.* 54: 895–901.
61. Jindra, M., S. R. Palli, and L. M. Riddiford. 2013. The juvenile hormone signaling pathway in insect development. *Annu. Rev. Entomol.* 58: 181–204.
62. Sláma, K., M. Romaňuk, and F. Šorm. 1974. *Insect Hormones and Bioanalogues*. Wien: Springer-Verlag.
63. Parthasarathy, R., and S. R. Palli. 2009. Molecular analysis of juvenile hormone analog action in controlling the metamorphosis of the red flour beetle, *Tribolium castaneum*. *Arch. Insect Biochem. Physiol.* 70: 57–70.
64. Staal, G. B. 1975. Insect growth regulators with juvenile hormone activity. *Annu. Rev Entomol.* 20: 417–460.
65. Palli, S. R. 2009. Recent advances in the mode of action of juvenile hormones and their analogs. In *Biorational Control of Arthropod Pests*, ed. I. Ishaaya, and A. R. Horowitz, 111–129. Dordrecht: Springer.
66. Guerrero, A. and G. Rosell. 2005. Biorational approaches for insect control by enzymatic inhibition. *Curr. Medicinal Chem.* 12: 461–469.
67. Kostyukovsky, M., B. Chen., S. Atsmi, and E. Shaaya. 2000. Biological activity of two juvenoids and two ecdysteroids against three stored product insects. *Insect Biochem. Mol. Biol.* 30: 891–897.
68. Krämer, B., U. Körner, and P. Wolbert. 2002. Differentially expressed genes in metamorphosis and after juvenile hormone application in the pupa of *Galleria*. *Insect Biochem. Mol. Biol.* 32: 133–140.
69. Suzuki, Y., J. W. Truman, and L. M. Riddiford. 2008. The role of Broad in the development of *Tribolium castaneum*: implications for the evolution of the holometabolous insect pupa. *Development* 135: 569–577.
70. Ramaseshadri, P., R. Farkas, and S. R. Palli. 2012. Recent progress in juvenile hormone analogs (JHA) research. *Adv. Insect Physiol.* 43: 353–436.
71. Wijayaratne, L. K. W., F. H. Arthur, and S. Whyard. 2018. Methoprene and control of stored-product insects. *J. Stored Prod. Res.* 76: 161–169.
72. Arthur, F. H., S. Liu, B. Zhao, and T. W. Phillips. 2009. Residual efficacy of pyriproxyfen and hydroprene applied to wood, metal and concrete for control of stored-product insects. *Pest Manag. Sci.* 65: 791–797.
73. Mittler, T. E., S. G. Nassar, and G. B. Staal. 1976. Wing development and parthenogenesis induced in progenies of kinoprene-treated gynoparae of *Aphis fabae* and *Myzus persicae. J. Insect Physiol.* 22: 1717–1725.
74. Grenier, S., and A.-M. Grenier. 1993. Fenoxycarb, a fairly new insect growth regulator: a review of its effects on insects. *Ann. Appl. Biol.* 122: 369–403.
75. Aupinel, P., D. Fortini, B. Michaud, F. Marolleau, J.-N. Tasei, and J.-F. Odoux. 2007. Toxicity of dimethoate and fenoxycarb to honey bee brood (*Apis mellifera*), using a new *in vitro* standardized feeding method. *Pest Manag. Sci.* 63: 1090–1094.
76. Sullivan, J. J. 2010. Chemistry and environmental fate of fenoxycarb. In *Reviews of Environmental Contamination and Toxicology*, Vol. 202, ed. D. M. Whitacre, 155–184. New York: Springer-Verlag.
77. Sullivan, J. J., and K. S. Goh. 2008. Environmental fate and properties of pyriproxyfen. *J. Pestic. Sci.* 33: 339–350.
78. Streibert H. P., M. L. Frischknecht, and F. Karrer. 1994. Diofenolan – a new insect growth regulator for the control of scale insects and important lepidopterous pests in deciduous fruit and citrus. *Br. Crop Prot. Conf. Pests Dis.* 1: 23–30.

79. Bowers, W. S., T. Ohta., J. S. Cleere, and P. A. Marsella. 1976. Discovery of insect anti-juvenile hormones in plants. *Science* 193: 542–547.
80. Pener, M. P., L. Orshan, and J. De Wilde. 1978. Precocene II causes atrophy of corpora allata in *Locusta migratoria*. *Nature* 272: 350–353.
81. Klowden, M. J. 2013. *Physiological Systems in Insects*, 3rd ed. San Diego: Academic Press.
82. Nakagawa, Y., and H. Sonobe. 2016. Ecdysteroids. In *Handbook of Hormones*, ed. Y. Takei, H. Ando, and K. Tsutsui, 557–559 and e98-1–e98-14. San Diego: Academic Press.
83. Retnakaran, A., P. Krell, Q. Feng, and B. Arif. 2003. Ecdysone agonists: mechanism and importance in controlling insect pests of agriculture and forestry. *Arch. Insect Biochem. Physiol.* 54: 187–199.
84. Smagghe, G., L. E. Gomez, and T. S. Dhadialla. 2012. Bisacylhydrazine insecticides for selective pest control. *Adv. Insect Physiol.* 43: 163–249.
85. Hamaidia, K., F. Tine-Djebbar, and N. Soltani. 2018. Activity of a selected insecticide (methoxyfenozide) against two mosquito species (*Culex pipiens* and *Culiseta longiareolata*): toxicological, biometrical and biochemical study. *Physiol. Entomol.* 43: 315–323.
86. Hu, J-Y., C. Liu, Y.-C. Zhang, and Z.-X. Zheng. 2009. Hydrolysis and photolysis of diacylhydrazines-type insect growth regulator JS-118 in aqueous solutions under abiotic conditions. *Bull. Environ. Contam. Toxicol.* 82: 610–615.
87. Huang, Z., Q. Cui, L. Xiong, Z. Wang, K. Wang, Q. Zhao, F. Bi, and Q. Wang, . 2009. Synthesis and insecticidal activities and SAR studies on novel benzoheterocyclic diacylhydrazine derivatives. *J. Agric. Food Chem.* 57: 2447–2456.
88. Zou, C., G. Liu, S. Liu, S. Liu, Q. Song, J. Wang, Q. Feng, Y. Su, and S. Li. 2018. Cucurbitacin B acts as a potential insect growth regulator by antagonizing 20-hydroxyecdysone activity. *Pest Manag. Sci.* 74: 1394–1403.
89. Dinan, L., Y. Nakagawa, and R. E. Hormann. 2012. Structure–activity relationships of ecdysteroids and non-steroidal ecdysone agonists. *Adv. Insect Physiol.* 43: 251–298.
90. Kitamura, S., T. Harada, H. Hiramatsu, R. Shimizu, H. Miyagawa, and Y. Nakawaga. 2014. Structural requirement and stereospecificity of tetrahydroquinolines as potent ecdysone agonists. *Bioorg. Medic. Chem. Lett.* 24: 1715–1718.
91. Naranjo, S. E., P. C. Ellsworth, and J. R. Hagler. 2004. Conservation of natural enemies in cotton: role of insect growth regulators in management of *Bemisia tabaci*. *Biol. Control* 30: 52–72.
92. Vianna, U. R., D. Prattisoli, J. C. Zanuncio, E. R. Lima, J. Brunner, F. F. Pereira, and J. E. Serrão. 2009. Insecticide toxicity to *Trichogramma pretiosum* (Hymenoptera: Trichogrammatidae) females and effect on descendant generations. *Ecotoxicology* 18: 180–186.
93. Suma, P., L. Zappalà, G. Mazzeo, and G. Siscaro. 2009. Lethal and sub-lethal effects of insecticides on natural enemies of citrus scale pests. *BioControl* 54: 651–661.
94. Wang, Y., C. Wu, T. Cang, L. Yang, W. Yu, X. Zhao, Q. Wang, and L. Cai. 2014. Toxicity risk of insecticides to the insect egg parasitoid *Trichogramma evanescens* Westwood (Hymenoptera: Trichogrammatidae). *Pest Manag. Sci.* 70: 398–404.
95. Khan, M. A., and J. R. Ruberson. 2017. Lethal effects of selected novel pesticides on immature stages of *Trichogramma pretiosum* (Hymenoptera: Trichogrammatidae). *Pest Manag. Sci.* 73: 2465–2472.
96. Döker, I., M. L. Pappas, K. Samaras, A. Triantafllou, C. Kazak, and G. D. Broufas. 2015. Compatibility of reduced-risk insecticides with the non-target predatory mite *Iphiseius degerans* (Acari: Phytoseiidae). *Pest Manag. Sci.* 71: 1267–1273.
97. Cabral, S., P. Garcia, and A. O. Soares. 2008. Effects of pirimicarb, buprofezin and pymetrozin on survival, development and reproduction of *Coccinella undecimpunctata* (Coleoptera: Coccinellidae). *Biocontrol Sci. Technol.* 13: 307–318.
98. He, F., S. Sun, X. Sun, S. Ji, X. Li, J. Zhang, and X. Jiang. 2018. Effects of insect growth-regulator insecticides on the immature stages of *Harmonia axyridis* (Coleoptera: Coccinellidae). *Ecotoxicol. Environ. Safety* 164: 665–674.

99. Desneux, N., A. Decourtye, and J.-M. Delpuech. 2007. The sublethal effects of pesticides on beneficial arthropods. *Annu. Rev. Entomol.* 52: 81–106.
100. Mommaerts, V., G. Sterk, and G. Smagghe. 2006. Hazards and uptake of chitin synthesis inhibitors in bumblebees *Bombus terrestris*. *Pest Manag. Sci.* 62: 752–758.
101. Malone, L. A., C. D. Scott-Dupree, J. H. Todd, and P. Ramankutty. 2007. No sub-lethal toxicity to bumblebees, *Bombus terrestris*, exposed to Bt-corn pollen, captan and novaluron. *New Zealand J. Crop Horticult. Sci.* 35: 435–439.
102. Chandel, R. S., and P. R. Gupta. 1992. Toxicity of diflubenzuron and penfluron to immature stages of *Apis cerena indica* F. and *Apis mellifera* L. *Apidologie* 23: 465–473.
103. Fine, J. D., C. A. Mullin, M. T. Frazier, and R. D. Reynolds. 2017. Field residues and effects of the insect growth regulator novaluron and its major co-formulant *N*-methyl-2-pyrrolidone on honey bee reproduction and development. *J. Econ. Entomol.* 110: 1993–2001.
104. Wade, A., C.-H. Lin, C. Kurkul, E. R. Regan, and R. M. Johnson. 2019. Combined toxicity of insecticides and fungicides applied to California almond orchards to honey bee larvae and adults. *Insects* 10: article 20.
105. Hodgson, E. W., T. L. Pitts-Singer, and J. D. Barbour. 2011. Effects of the insect growth regulator, novaluron on immature alfalfa leafcutting bees, *Megachile rotundata*. *J. Insect Sci.* 11: article 43.
106. Pitts-Singer, T. L., and J. D. Barbour. 2017. Effects of residual novaluron on reproduction in alfalfa leafcutting bees, *Megachile rotundata*. F. (Megachilidae). *Pest Manag. Sci.* 73: 153–159.
107. Mommaerts, V., G. Sterk, and G. Smagghe. 2006. Bumblebees can be used in combination with juvenile hormone analogues and ecdysone agonists. *Ecotoxicology* 15: 513–521.
108. Chen, Y.-W., P.-S. Wu, E.-C. Yang, Y.-S. Nai, and Z. Y. Huang. 2016. The impact of pyriproxyfen on the development of honey bee (*Apis mellifera* L.) colony in field. *J. Asia-Pacific Entomol.* 19: 589–594.
109. Fisher II, A., C. Colman, C. Hoffmann, B. Fritz, and J. Rangel. 2018. The effects of the insect growth regulators methoxyfenozide and pyriproxyfen and the acaricide bifenazate on honey bee (Hymenoptera: Apidae) forager survival. *J. Econ. Entomol.* 111: 510–516.
110. Tasei, J.-N. 2001. Effects of insect growth regulators on honey bees and non-*Apis* bees. A review. *Apidologie* 32: 527–545.
111. Kristensen, M., and J. B. Jespersen. 2003. Larvicide resistance in *Musca domestica* (Diptera: Muscidae) populations in Denmark and establishment of resistant laboratory strains. *J. Econ. Entomol.* 96: 1300–1306.
112. Ishaaya, I., S. Kontsedalov, and A. R. Horowitz. 2005. Biorational insecticides: mechanism and cross-resistance. *Arch. Insect Biochem. Physiol.* 58: 192–198.
113. Horowitz, A. R., K. Gorman, G. Ross, and I. Denholm. 2003. Inheritance of pyriproxyfen resistance in the whitefly *Bemisia tabaci* (Q biotype). *Arch. Insect Biochem. Physiol.* 54: 177–186.
114. Cao, G.-C., and Z.-J. Han. 2015. Tebufenozide resistance is associated with sex-linked inheritance in *Plutella xylostella*. *Insect Sci.* 22: 235–242.
115. Pavan, F., E. Cargnus, G. Bigot, and P. Zandigiacomo. 2014. Residual activity of insecticides applied against *Lobesia botrana* and its influence on resistance management strategies. *Bull. Insectol.* 67: 273–280.
116. Ghanim, M., and S. Kontsedalov. 2007. Gene expression in pyriproxyfen-resistant *Bemisia tabaci* Q biotype. *Pest Manag. Sci.* 63: 776–783.
117. Jia, B., Y. Liu, Y. C. Zhu, X. Liu, C. Gao, and J. Shen. 2009. Inheritance, fitness cost and mechanism of resistance to tebufenozide in *Spodoptera exigua* (Hübner) (Lepidoptera: Noctuidae). *Pest Manag. Sci.* 65: 996–1002.
118. Castle S., and S. E. Naranjo. 2009. Sampling plans, selective insecticides and sustainability: the case for IPM as ‘informed pest management’. *Pest Manag. Sci.* 65: 1321–1328.
119. Ishaaya, I., S. Kontsedalov, and A. R. Horowitz. 2003. Novaluron (Rimon), a novel IGR: potency and cross-resistance. *Arch. Insect Biochem. Physiol.* 54: 157–164.

120. Mouden, S., K. F. Sarmiento, P. G. L. Klinkhamer, and K. A. Leiss. 2017. Integrated pest management in western flower thrips: past, present and future. *Pest Manag. Sci*. 73: 813–822.
121. Wilson, L. J., M. E. A. Whitehouse, and G. E. Herron. 2018. The management of insect pests in Australian cotton: an evolving story. *Annu. Rev. Entomol*. 63: 215–237.
122. Matyjaszczyk, E. 2019. Problems implementing compulsory integrated pest management. *Pest Manag. Sci*. doi:10.1002/ps.5357. In Press.
123. Cabrera, P., D. Cormier, and É. Lucas. 2017. Differential sensitivity of an invasive and an indigenous ladybeetle to two reduced-risk insecticides. *J. Appl. Entomol*. 141: 690–701.
124. Smagghe, G., M. Zotti, and A. Retnakaran. 2019. Targeting female reproduction in insects with birational insecticides for pest management: a critical review with suggestions for future research. *Curr. Opin. Insect Sci*. 31: 65–69.
125. Ishaya, I., S. R. Palli, and A. R. Horowitz, eds. 2012. *Advanced Technologies for Managing Insect Pests*. Dordrecht: Springer.

II

COV: Comparative Overviews of Important Topics for Environmental Management

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

7

Biodiversity and Sustainability

Odo Primavesi

Introduction

One of the greatest concerns related to sustainability today is the degradation and switching off of essential ecosystem services caused by anthropogenic activities like land degradation and pollution. These human activities, by producing solid, liquid, gaseous, and radiating (heat, light, radioactive emissions, and others) wastes and contaminants, lead to physical, chemical, and biological pollution, affecting the whole planet. For example, global warming and climate change result from a combination of air pollutants, such as the excess of greenhouse gases (carbon dioxide, methane, nitrous oxide, and others) and the excess of long-wave infrared or heat radiation (>300 W/m^2) from degraded and dry landscapes, and other factors.

Another example is the pollution of water due to poor management of soils (erosion, siltation of water bodies, and contamination with nitrates, phosphates, and other pollutants) or direct discharge of household, laboratory, or industrial waste. Water pollution and air pollution to a great extent contribute to levels of different human sicknesses and death rates in the very contaminated areas. Lack of good management practices regarding the conservation of natural resources (soil, water, air, and biodiversity) and ecosystem services and the lack of adequate management of waste are also very important issues.

In this entry, the goals are to improve the awareness and understanding of the importance of concepts or objectives like biodiversity, environmental health, and sustainability, and to give some suggestions on how the reader may contribute to individual and communal well-being.

Biodiversity

What Is Biodiversity?

Biodiversity or biological diversity is defined by the United Nations Convention on Biological Diversity as "the variability among living organisms from all sources, including terrestrial, marine, and other aquatic ecosystems and the ecological complexes of which they are part; this includes diversity within species, among species, and of natural and altered ecosystems,"[1,2] including species, ecosystem, morphology, gene, and molecular diversity.

Ecosystem means a dynamic complex of plant, animal, and microorganism communities and the inorganic environment, interacting as a functional unity. Different levels of diversity are considered, among individuals, subspecies, species (most useful level), biological communities, and ecosystems. Species richness increases from colder to warmer latitudes. This is also true for the deep-sea species diversity.[1]

The number of species on Earth is estimated to be 5 to 30 million, from which 1.4 to 2 million were identified in a formal system. The majority are invertebrates (1 million; mainly insects and myriapods), followed by microorganisms (5760; e.g., fungi), chelicerates, protists, nematodes, plants (250,000; around 50,000 as trees), molluscs, crustaceans, and vertebrates (19,100 fishes, 9000 birds, 6300 reptiles, 4200 amphibians, and 4000 mammals).

Occurrence of species diversity follows in this decreasing order: tropical and subtropical moist broadleaf forests >>> tropical and subtropical grasslands, savannas, and shrub lands > deserts and xeric shrub lands = tropical and subtropical dry broadleaf forests > mountain grassland and shrub lands > temperate broadleaf and mixed forests > flooded grasslands and savannas = tropical and subtropical coniferous forests > temperate grasslands, savannas, and shrub lands = mangroves = temperate coniferous forests > Mediterranean forests, woodland, and scrub >> boreal forests or taiga > tundra.[1]

The tropics are the home of most of the species. An example is the Amazonian rainforest, where 60% of all life-forms (e.g., 60,000 plant species) reside.[3] In one hectare of Atlantic rainforest in the southern Bahia state in Brazil, there are up to 454 tree species recorded.[4]

The reason for the species richness of the tropics is not well known. Some ideas proposed are the longer time available to develop new species and potentially greater supply of solar energy, allowing more biomass production or more organisms per unit of area.[1]

Furthermore, soils in temperate climates show greater chemical fertility, water-holding capacity, and clay activity, and the cold switch off of biological activity controls these processes. Under tropical conditions, deep soils with mostly low chemical fertility, low water-holding capacity, and low clay activity, and the higher temperatures throughout the year allow for a greater number of interactions of water/drought × water table depth × nutrient availability × salinity × temperature/altitude/shade × strong rains × wind × fire × plant residues × organic matter content × photoperiod × oxygen (because of faster respiration rates and heat).

Therefore, habitat variability occurs, with specificities settled in by the different plant species, the first component of the food web and net.

Biological diversity is organized in a food web, with plants as base, harvesting sun energy freely available to them, and humans as top of the pyramidal net, where the individuals act as producer or consumer, or as recycler or decomposer.[5] The diversity of litter, defense substances, and root exudates produced by these different plant species, and correlated fauna, need to be chopped up by, e.g., invertebrates and decomposed by a greater number of microorganisms in soil, because of their specificity in producing degradation enzymes. The great recycling activities in soil need to be considered, because of the big importance of organic material as nutrient source for higher plants, as energy source for the microbial activity, and as a factor in improving soil structure and its water-holding capacity. In the tropics, with the great variability of habitats, the diversity of species is the keystone for high biomass yield per unit of area, making the food net of an ecosystem very complex.

Biodiversity reaches the maximum level when the environment offers enough water and energy and low to medium level of nutrients, such as nitrogen and phosphorus. With high levels of nitrogen and phosphorus occurs a very stiff inter- and intraspecific competition, mainly for light, from some more responsive or demanding species, similar to that occurring in high-fertility and very high-fertility soils or in eutrophic water bodies. There are also growth-restrictive conditions like in very low-fertility soils, with very low phosphorus and/or high aluminum content, or saline soils.

What Is the Importance of Biodiversity?

The greatest importance of biodiversity is still the optimized ecosystem services it provides.[1,6–9] Ecosystem services are flows of material, energy, and information on environmental structures (natural capital), which, combined with services, products, and human capital, generate human well-being. Usually, most of these processes could not be substituted in needed scale by any human technology, and if the ecosystem is extinguished, reversion is in general very difficult to realize in an economic way. The best way is conservation, including the costs in the price of products and services. Usually, ecosystem services occur in an imperceptible way, similarly to the involuntary and vital processes in our organism such as pumping of oxygenated blood or breathing.

Biodiversity provides three functional ecosystem services (production, regulation, and cultural) and a support service. Production and supply include mainly food, freshwater, timber, fiber, fuel, energy, genetic resources, medicines, wildlife, and others. Regulation includes maintenance of climate, carbon sequestration, soil conservation, wind and sea wave power, biodegradation and recycling of wastes, biological remediation of soils, and decontamination and cleaning up of water, cleaning up of air, maintaining soil permeability, and others. Cultural services could include cultural and heritage diversity; aesthetic, ethical, medicinal, and health knowledge; inspiration; educational, spiritual, and religious values; leisure; ecotourism; and so on. Support of life is carried out by maintaining a stratospheric oxygen-ozone layer to filter ultraviolet radiation and a greenhouse gas layer to filter infrared sun radiation and to retain partial infrared or heat radiation from surface, as well as by clouds, which reduce sunshine incidence on Earth surface, avoiding the burning of life. It is also carried out by maintaining a long water cycle (rain–interception–infiltration–storage–internal flow–evaporation and transpiration–air humidity–clouds) with distributed soft rains, by stabilizing air temperature and air humidity, and others.[1,2,10]

Biological diversity also provides resilience or a stabilizing effect on the food web, which sustains the human species, when an environmental disturbance occurs.[11]

Nature uses biodiversity to produce the maximum of life and biomass per square meter and year by optimal use of the available sun energy. Biodiversity is also the result of this settlement process of nature, with a great variety of abiotic conditions, mainly in the tropics.

At the same time, nature, by developing a food chain into a complex food net, allows for greater food availability and diversity for the individuals on the top of the food web or food pyramid, such as the human species, and also ensures their sustainability. However, when biodiversity of the food web is disrupted by the establishment of a monoculture (industrial cropping system) and/or when the environment is under subjection of a degradation process, a population outbreak of the more resistant or adapted members of the food web may occur such as the so-called parasites and pathogens.[12]

The soil is one of the most diverse habitats on Earth and contains one of the most diverse assemblages of living organisms, mainly in the humid tropics,[13] due to its plant diversity. In a broader view, it is advisable to consider soil as the undisassociable soil–plant interaction, mainly in the tropics and subtropics. Soil without a permanent living plant cover will lose its main function of harvesting rain and storing resident available freshwater, essential for life and biological production (food, fiber, wood, biofuel, etc.). This interaction will improve the degree of soil biodiversity, including also the rooting system architecture of the so-called weeds. Their rooting structure, as well as that of crops, may be used as a visible indicator of the degree of soil health.[14,15]

In both natural and agricultural ecosystems, the different groups of soil biota interacting with plants and their debris are responsible for, or strongly influence, the soil properties and optimize processes or ecosystem services such as soil genesis, soil structure, carbon, nutrient and water cycles, agrochemical movement or breakdown, plant protection, growth, and production.[5,16]

Soil organisms act in processes of synthesis or production, transformation and decomposition, or consumption of organic material, affecting abiotic and biotic components, transportation, and soil engineering. Therefore, soil biodiversity is a keystone for sustainable agriculture and it could be used as a good indicator of agro-ecosystem or soil health. Soil biodiversity does not necessarily refer to the number of individuals or species, but to the ratio of functional groups,[17,18] and the result or the tool of their activities, such as the presence and intensity of enzymatic activity.[12] It is necessary to remember the importance of soils.

The settlement of terrestrial environments by life was only possible by storing rainwater in rocks that nature developed to permeable soils, and these soils were maintained permeable by a triple-protection layer: plant canopy, litter, and surface rooting system. To succeed in our activities and also to meet our quality of life, it is necessary to maintain, restore, or mimic this structure and the processes involved using artificial technology.

It can be said that the disappearance of several ancient human civilizations with populations concentrated in cities was partially caused by food insecurity due to soil degradation, combined with freshwater shortage due to destruction of forests and soil permeability, therefore reducing the long water cycle, and by lack of sanitation and waste disposal affecting public health.[19] These problems of slowing down or disrupting ecosystem services are at present global in scale, with the emergence of a new problem—global warming and climate change. Besides, there occurs programmed or accidental inclusion of poisons, toxic substances, heavy metals, nitrates, phosphates, hormones, etc.

To have an idea of the current importance of the whole biodiversity, the monetary value of 17 ecosystem services required to sustain life and the biological production capacity of landscapes was estimated to be about $33 trillion/yr, against the $18 trillion of the global gross domestic product (GDP).[9]

What Destroys Biodiversity and Ecosystem Services?

This can be answered by the simple elimination of plants covering the soil and the prevention of their complete regrowth and occupation of soil surface, as well as by turning the soil impervious, by crusting, compaction, pavement, etc.

How Do We Take Care of Biodiversity and Essential Ecosystem Services?

An interesting point is that, instead of considering the natural climax ecosystem as reference for good environmental practices, we use the primary natural ecosystem (rocky landscape) as reference for characteristics we do not want in our agricultural or urban ecosystems. The following are environmental characteristics we have to avoid in our management program: the primary environment has no capacity to store water; it has no biological carrying capacity for higher species; it sustains no food chain nor web; it presents a very short water cycle (rain, evaporation and runoff); and it shows high temperature and air humidity amplitude during the day.

Hence, primary environments are unsuitable for life and production. For example, land and soil degradation will turn life- and production-friendly environmental characteristics into life- and production-unsuitable conditions (with impervious, compacted, dry, hard soil, like a rock), similar to that occurring in primary natural ecosystems.

Considering the growing soil and landscape degradation, for example, due to further erosion and salinity, we need to be aware that the process of desertification[23–25] is a great challenge in dry lands. These dry lands support 44% of all cultivated systems and are the origin of 30% of world's cultivated plants.[26,27]

Thus, first of all, we need to stop the landscape and ecosystem services degradation, by conservation practices. The second step is to recover soils and landscapes by simple harvesting and storing rainwater, and reducing water losses, e.g., by runoff or evaporation in excess, by protecting the soil surface against erosion, and by allowing the growth of a diverse plant cover, to turn the soil permeable and to stabilize air temperature and humidity.

It is necessary to prevent the destruction of the whole plant cover and their debris. The biologically diverse green areas in rural and urban environments are mandatory. Under tropical and subtropical conditions, this means that we have to avoid large-scale areas of pasture or cropland or buildings (cities) without trees. A permanent tree cover is important to maintain the evapotranspiration and windbreak service and to stabilize air temperature and humidity.[8]

Considering that the greatest land-use change, with massive deforestation, is mainly for agricultural purposes (as well as for wood harvest, coal production, and mining activity), it will significantly alter the micro- and mesoclimate in a region, increasing infrared radiation and heating up of lower atmosphere, and because it will promote the significant degradation of other essential ecosystem services due to its scale, agricultural practices based on ecological principles are advisable.

For large areas, the first step should be the practice of conservation agriculture[20] complemented by windbreaks and vaporizing tree cluster. For small areas, agroforestry[21,22] production systems are useful, mainly under tropical and subtropical conditions, considering that the tropics are the engine of global climate dynamics. Without trees, the heat production over terrestrial areas will be greater, the cloud production will be smaller, and the climate dynamics will be faster or stronger, more dangerous.

Examples of successful ecosystem services restoration are those in conservation agriculture[20,28–30] agro-ecological production systems,[31–33] and agroforestry,[34–39] aside from forest management[40] and those biodiversity conservation practices[41,42] that substitute paid environmental services for free ecosystem functions or services, by using ecological principles.

Also, the need to reduce the use of toxic substances, as well as the production of wastes and their random release to the environment, is urgent. Gaseous wastes affect human and biodiversity health, especially when considering the troposphere ozone and acid rain production, the increasing carbon footprint, and the large amount of smog production, brought about by landscape fire, fire from furnaces (used for coal production), or fire from household wood-burning stoves. Liquid waste degrades freshwater, turns it dangerous for human health, and increases freshwater shortages. Solid waste released in large amounts on the landscape will reduce land and water quality and will increase the ecological footprint.

The best indicator of biodiversity that provides adequate ecosystem services is when we have a greater number of plants per unit of area. Plants are associated with fauna and microorganism.[6]

Which Tools Should Be Used to Improve Biodiversity?

Considering the process nature uses to develop a certain site, or to recover a degraded soil or land under fallow, it could be seen that plant diversity is the key tool used. This is because it allows the complementary activity of individuals with different structures, functions, wastes (debris, root exudations, and others), and needs to flourish in one of the different habitats created by the diverse interaction of the abiotic and emergent biotic factors occurring.

Therefore, vegetative techniques to recover the permanent diverse plant cover together with the improvement of environmental legislation and education are in place.

Environmental Health

What Is Environmental Health?

According to the World Health Organization,[43,44] environmental health addresses all the physical, chemical, and biological factors external to a person, and all the related factors impacting behavior. It encompasses the assessment and control of those environmental factors that can potentially affect health. It is targeted towards preventing disease and creating health-supportive environments.

Which Environmental Conditions Will Affect Health?

Ecosystem degradation or disruption can impact on health in a great variety of ways.[43–46] The most important ones are freshwater pollution (lack of sanitation), contamination (heavy metals, hormones, poisons, and excess of medicines), and degradation (siltation and increase of nitrates and phosphates).

Perhaps we need to have a broader view of what is environmental health, considering that persons are also participants of our environment. In a health environment, for example, aside from waterborne diseases due to pollution and acute and chronic respiratory diseases due to air pollution, we need to also consider malnutrition due to food lacking micronutrients, extreme heat and very low air humidity, lack of education and training, low (or no) income, and so forth.

What Is the Importance of Environmental Health?

A healthy environment, with its natural resources and main structures conserved or improved, will allow the ecosystem services to run in order to benefit our well-being. A healthy environment will provide enough clean, fresh water (150 to 200 L/person/day); it will secure food (1500 to 2000 kcal/person/day, without toxic or dangerous substances and contaminants); it will provide clean air with enough humidity (around 10 g water as vapor/m^3 air, or 40%–60% relative humidity in a range of 20–24°C, the comfort range, without solid microparticles, smog, dangerous gases and substances, and inconvenient odor); and it will lead to stabilized temperature.

An excess or low level of air humidity will bring about and increase the occurrence of several diseases.[47] Unprotected surface/soil will produce higher temperature (Figure 1) and therefore greater amount of long-wave infrared radiation, heating up the lower atmosphere. In addition, increasing air temperature will mainly result in decreasing air humidity, because of the increase of atmospheric demand for water, when no water is available to be evaporated or transpired by plants (Figure 2).

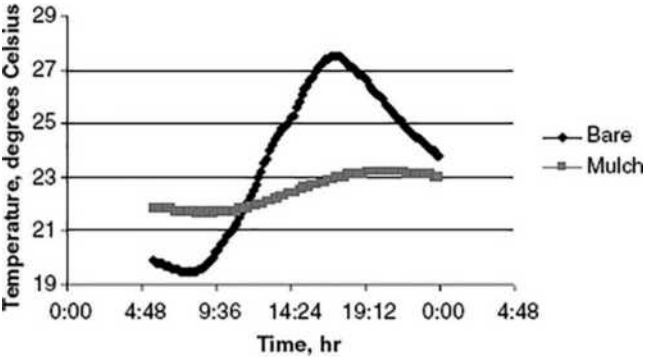


FIGURE 1 Temperature variation (amplitude), at 15 cm depth, in a bare and mulch-covered soil, under tropical conditions (November). The graph shows that it is possible to manage temperature extremes using mulching technique.
Source: Adapted from Torres (1997) in Primavesi.[8]

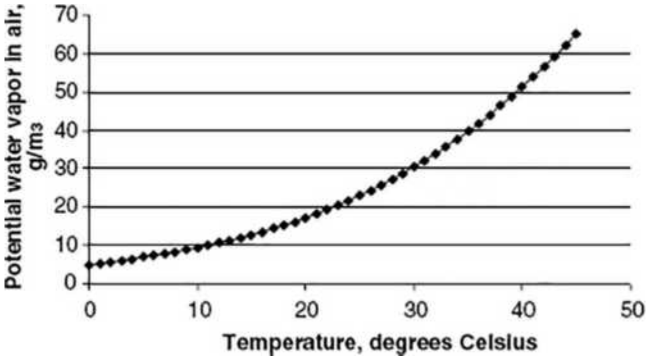


FIGURE 2 Relation between air temperature and potential water demand for air saturation.
Source: Adapted from Addams et al.[52]

Heat in excess, above 30°C, will sharply increase the productivity loss of labor and increase the mistakes or errors in the production line,[48] while polluted and dry air will increase respiratory diseases.[47] Polluted water will cause several gastroenteritis processes.[49–51] Human health degradation will cause problems for the production system, the health care service, and to our well being, with expectation of an earlier death.

What to Do?

The reactivation of the ecosystem services is needed, with the restoration of biodiversity, on a permeable soil, increasing rainwater yield and storage. It is also necessary to reduce waste production, by reduction, reuse, and recycling, as well as by neutralizing and adequate disposal of all harmful materials. We need to use water more efficiently.[52]

Another very important point is to stabilize air temperature (avoiding extremes) and air humidity amplitude, by restoring green evapotranspiring areas, by establishing shade (Table 1) or soil cover (Figure 1), or by managing the surface albedo.[54–62] Nature manages albedo or sunshine reflectivity in function of available liquid water. Water bodies, humid soils, or live plants are darker (with lower albedo) and may absorb more sun energy than dry surfaces, dry leaves (straw), or solid water (snow and ice) fields (with greater albedo).

A notable complementary problem is when fire use is routine in landscape management,[63] especially in dry periods, turning the albedo very low (black surfaces), increasing the heat, and therefore knocking down air humidity and quality, affecting health. Since we have global areas experiencing water shortage[64] [being dry for moderate (3 to 5 months) to long (>6 months) periods], this may result in

TABLE 1 Temperature of Shaded and Unshaded Surfaces in the Tropics

	Temperature		
	Shaded	Unshaded	Increase
Surface	°C		%
Green lawn	32	35	9
Dry lawn	35	52	48
Concrete	37	52	40
Asphalt	37	57	54

Source: Addams et al.[5]

heating up of soils and surfaces, with temperatures above 52°C,[65] and this will produce atmospheric heat in excess (>300 W/ m^2),[66] surprisingly where there were no forests.[67]

Trees, due to their darker color, will absorb more sun energy and produce more heat, warming up a cold environment, but when air temperature rises above a certain level, it starts the evapotranspiration process, keeping away the heat in excess. Trees may warm up and cool down an environment, depending on the need. Therefore, it is advisable to make a global effort to plant trees (forests) and manage agroforestry systems to restore ecosystem services such as temperature and air humidity stabilization, as well as to maintain a longer water cycle with more and better- distributed rain.[68–76]

It is necessary to take into account that there are biophysical limits for economic growth (adjusted to nature's biological carrying capacity), which are mainly responsible for pollution as well as for biodiversity and land degradation. From the nine main limitations (biodiversity loss, nitrogen and phosphorus cycle, climate change, acidification of seawater, ozone reduction in stratosphere, freshwater availability, change of soil use, chemical pollution, and aerosol pollution in atmosphere), we did trespass the first three, endangering our health and livelihood.[77]

Indicators of Improvement

The presence of biodiversity developing on permeable soils and the absence of pollution, waste disposal in landscapes, and contaminants result in clean freshwater, soil, air, and food.

Tools to Be Used

Processes and tools based on ecological principles and processes to recover soil permeability (Figures 3 and 4), biodiversity, and ecosystem services need to be used. The keystone of the food web we depend on is plants, especially those with healthy roots. Roots need to be in aerated and humid soil, protected (against temperatures above 33°C, which harms root health), and rich in organic material.

The concept of integrated natural resource management considering the watershed scale, aside from integrated and efficient water, fertilizer, and pest management, is advisable. In relation to waste disposal, pollution, and contamination, environmental technology, eco-technology, and cleaner technology are all in place. In the case of fossil fuel use, the reduction of wasteful use and the substitution by alternative renewable energy sources are advisable. In all cases, improvement in environmental legislation and education is necessary.

FIGURE 3 Crusting of an unprotected prepared seedbed after tropical rainshower, turning soil impervious to rainwater and aeration.
Source: Author's personal archive.

FIGURE 4 Left: permeable healthy soil from a natural ecosystem, plenty of visible roots. Right: the same soil type, compacted, impervious, and dry from the inter row of a sugarcane field after 5 years of continuous intensive cultivation.
Source: Author's personal archive.

Sustainability

What Is Sustainability?

In biology, sustainability means the processes running to maintain biological systems diverse and productive over time.[5] In 1987, the World Commission on Environment and Development established a definition of sustainability, known as the Brundtland Report. It stated that sustainable development should reach the needs of the present without compromising the ability of future generations to meet their own needs. Although this definition has become widely publicized, the term *sustainability* is not limited to one precise definition.[78]

Several authors have discussed the real meaning of sustainability and sustainable development.[16,28,79] Munoz[80] has presented a theoretical model of true sustainability or sustainable development ideas with global to local implications based on the need to balance social, economic, and environmental goals; stakeholders' interests; and issues to induce or determine fairer or more appropriate development solutions, options, and actions.

However, the social component can be seen as part of the environmental component, and the interaction of a health and developed environment (essential ecosystem services running) with educated, trained, organized, and health persons will generate a long-lasting economic component. Hence, the environmental component is the keystone, to sustain the virtual world as well, and only its improvement and quality will allow us to reach a stable social welfare and a sustainable economic profit.

What Is the Importance of Sustainability?

Historically, the first known examples of worry on sustainability, and the establishment of a definition, were from those growing up in the forestry sector, and this is because of wood shortage for a salt mining activity in Germany or for house building and coal production in Japan. Both cases resulted from poor forest management and overexploitation. The question was, "What should we do to maintain a constant wood yield and cash flow for current and future forest owners', managers', and local workers' livelihood?" This question was discussed since 1442 in various regions of Germany, where wood consuming salt-mining activities did occur, and the concept was formulated firstly in 1650 in Saxonia (Germany) and later, independently, in 1666 in some deforested regions of Japan.[81]

The idea was that regenerative living resources, like wood/trees, could only be used/harvested in the same amount of the natural regrowth/recover/refill (with no use of external inputs, like water, energy, or fertilizer), and this by maintaining productivity, vitality, rejuvenation capacity, and biological diversity, in a time span of around 120 years (the time needed by trees to grow for cutting), to avoid natural resource shortage, labor and cash shortage, and livelihood and health problems (mainly related to erosion and flooding).[81]

In 1732, the idea on sustainable use of forest was published for the first time by von Carlowitz, and in 1795, Georg Ludwig Hartig described how to manage a sustainable forest.[81] History shows us also that a sustainable activity may last for 5000 years and more.[82]

Related to sustainable forest management, the current definition is as follows: "The stewardship and use of forests and forest lands in a way, and at a rate, that maintains their biodiversity, productivity, regeneration capacity, vitality and their potential to fulfill, now and in the future, relevant ecological, economic and social functions, at local, national, and global levels, and that does not cause damage to other ecosystems."[40]

Nature teaches us that the real development process of a natural primary environment occurs with the development or restoration of a permeable soil, protected by a diversified vegetation cover with an active rooting system, and the return to soil of diversified organic material, the energy source for the diversified and active soil life.

The soil–vegetation and associated biodiversity interaction could improve the available resident water of a site and also a longer water cycle. The more resident water, the more vegetation and the more permeable the soil, in a growing feedback loop. With greater amounts of available resident water-permeable soil-diversified vegetation and soil life, there is also an improvement in micro- and mesoclimate, with an increase in relative air humidity and a decrease in the maximum temperature and thermal amplitude, characteristic for desert environments.

This friendlier mesoclimate helps more sensible plant and animal species to establish and improves biodiversity, with their additive and emergent characteristics, mainly observed in the humid tropics. Ecology considers desert ecosystems sometimes as sustainable as drier tropical forest ecosystems, due to their richness in biodiversity. Thus, what level of environmental sustainability is desirable? It depends on the biological carrying capacity we want.

The biological carrying capacity represents the concept of primary productivity of an ecosystem, or the rate and amount in which energy is stored by photosynthetic or chemosynthetic activity of the producer organisms as organic substances, food for the food web.[5] The biological carrying capacity also considers the feeding capacity of grazing cattle, or grain equivalent available for humans (4 or 16 persons/ha/yr, with a minimum need of 1000 kcal/day), calculated as available digestive energy or calories per surface unit and year. The biological carrying capacity depends on the recovering capacity of a site (resilience) to produce biomass, after yield, extraction, degradation, or pollution activities.

Hence, considering the exuberant flora and biomass production by the Amazonian forest, the question arises: What is the biological carrying capacity of the around 40% sandy soils (<15% clay content) in the Amazonian basin without that great vegetation and the aggregated or dependent mesoclimate? Something similar to the Sahara? This example shows that the biological carrying capacity may be managed in a certain range by improving the natural resource structural tripod of resident water (harvested and stored rainwater) in permeable soil (organic matter rich) under biodiverse, permanent, alive plant cover and the ecosystem services.

Some agricultural production systems—although their processes result in improvement in environmental characteristics, income, and social inclusion—are not sustainable, because of their great dependence on external inputs [fertilizer, energy, water (including fossil underground water), and technical support]. We could observe, however, that, with time, it is possible to reduce part of this dependence, by switching to more organic and biological processes, after building up a minimum of fertility and organic matter level in soil, and introducing nitrogen-fixing leguminous trees. The goal of such process is to be more efficient and productive while maintaining or improving environment (natural resources) quality.[29]

What Needs to Be Focused on Sustainability?

Considering that the main objective of all human activity, from a global to a local scale, is to promote life and its quality, environmental sustainability will be reached when the biological carrying capacity is adequate to supply the minimal health life requirements of a given human population.

An increase in the biological carrying capacity level will allow a rise in human population density. Instead of this, what occurs now is the destruction of the main natural resource structures and functions (or ecosystem services), a decrease in the biological carrying capacity with an increase in human population density, and an increase in the production of solid, liquid, gaseous, and radiative wastes, thrown randomly in landscapes and marine ecosystems. We are currently watching a global ecological regression process of terrestrial and marine ecosystems, back to conditions unsuitable for human health, livelihood, and life.[8]

We need to reduce losses and wasting of materials, as well as to avoid pollution and contamination of natural resources and products.

Which Indicators Are Usable?

First, we need to reduce our ecological footprint (optimal land use),[83–88] our water footprint (optimal and efficient water use),[89–93] and our carbon footprint (carbon equivalent production in processes and carbon cycle)[94–96] by turning our production systems or life system more efficient and adequate to the global ecosystem carrying capacity. The measure of total sun energy use[97–100] is perhaps the best way to measure and turn the processes more sustainable.

Which Tools Should Be Used?

The following are the main tools we need to use: primarily, knowledge on ecological principles and processes;[5,97] a better understanding of the processes using energy, land, water, and primary products; a retooling of the processes to turn them more efficient, using materials more harmless to the environment and health; and a reduction of waste disposal in ecosystems. All of these are geared towards improving environmental legislation and education.

Conclusion

It is necessary to stop the degradation or regression and pollution processes of terrestrial and marine ecosystems and restore their health. The processes nature uses to develop complex resilient ecosystems with great biological (including human) carrying capacity are known. To conserve, restore, or improve a sustainable and healthy planet with the immense carrying capacity of the human species, biodiversity and ecosystem services should be considered, based on the ecological knowledge of essential natural structures and processes.

Moreover, as in nature, all kinds of wastes need to be reduced and recycled. Our technologies and processes need to maintain, restore, or mimic natural structures and processes to succeed and improve sustainability. A global awareness on the ecological principles and processes for a regional planning of the integrated local participatory activities to restore, maintain, or even improve the environmental health for sustainability of human societies is necessary.

References

1. Millennium Ecosystem Assessment. *Ecosystems and Human Well-Being. Biodiversity Synthesis;* World Resources Institute: Washington, 2005; 1–100. Available at http://www.millenniumassessment.org/documents/document.354.aspx.pdf (accessed March 2010).
2. Millennium Ecosystem Assessment. *Ecosystem and Human Well-Being: Synthesis;* Island Press: Washington, DC, 2005; 1–155. Available at http://www.millenniumassessment.org/en/Synthesis.aspx and http://www.millenniumassessment.org/documents/document.356.aspx.pdf (accessed March 2010).
3. Peneireiro, F.M.; Rodrigues, F.Q.; Brilhante, M.O.; Ludewigs, T. *Apostila do Educador Agroflorestal: Introdução aos Sistemas Agroflorestais—Um Guia Técnico*; Universidade Federal do Acre, Parque Zoobotânico: Rio Branco-AC, Brazil, 1999; 1–76. Available at http://www.agrofloresta.net/artigos/apostila_do_educador_agroflorestal-arboreto.pdf and http://www.slideshare.net/FlaviaCremonesi/apostila-do-educador-agroflorestal-arboreto-1353733 (accessed March 2010) (in Portuguese).
4. Girardi, E.P. *Atlas da Questão Agrária Brasileira*; Unesp: Presidente Prudente-SP, Brazil, 2008; 1–259. Available at http://www4.fct.unesp.br/nera/atlas/configuracao_territorial.htm (accessed March 2010) (in Portuguese).
5. Odum, E.P.; Barrett, G.W. *Fundamentals of Ecology,* 5th. Ed.; Brooks/Cole: New York, 2005; 1–612.
6. Convention on Biological Diversity. COP 6 Decision VI/7. Identification, monitoring, indicators and assessments; UNEP: Montreal, Canada, 2002. Available at http://www.cbd.int/decision/cop/?id=7181 (accessed March 2010).
7. Chivian, E., Ed. Biodiversity: Its importance to human health. *Interim Executive Summary*; Center for Health and Global Environment/Harvard Medical School/WHO/UNDP/UNEP: Harvard, USA, 2003; 1–59. Available at http://chge.med.harvard.edu/publications/documents/Biodiversity_v2_screen.pdf (accessed March 2010).
8. Primavesi, O.; Arzabe, C.; Pedreira, M.S. *Mudanças Climáticas: Visão Tropical Integrada das Causas, Dos Impactos e de Possíveis Soluções para Ambientes Rurais ou Urbanos*; Embrapa Pecuaria Sudeste: Sao Carlos-SP, Brazil, 2007; 1–200 (Embrapa Pecuaria Sudeste. Documentos, 70). Available at http://www.cppse.embrapa.br/080servicos/070publicacaogratuita/documentos/Documentos70.pdf (accessed March 2010) (in Portuguese, with a 3-page Executive Summary in English).
9. Costanza, R.; D'Arge; Groot, R.; Farber, S.; Grasso, M.; Hannon, B.; Limburg, K.; Naeem, S.; O'Neill, R.V.; Paruelo, J.; Raskin, R.G.; Sutton, P.; Belt, M. The value of the world's ecosystem services and natural capital. Nature **1997**, *15* (387, May), 253–260. Partially available at http://myweb.facstaff.wwu.edu/~medlerm/classes/08_09/502/nature-paper.pdf and http://earthmind.net/marine/docs/session2c-on-costanza-global-valuation.ppt (accessed March 2010).
10. Daily, G.C.; Alexander, S.; Ehrlich, P.R.; Goulder, L.; Lub-chenco, J.; Matson, P.; Mooney, H.A.; Postel, S.; Schneider, S.H.; Tilman, D.; Woodwell, G.M. Ecosystem services: Benefits supplied to human societies by natural ecosystems. Issues Ecol. **1997**, *1* (2), 1–18. Available at http://www.esa.org/science_resources/issues/FileEnglish/issue2.pdf (accessed March 2010).
11. Fischer, J.; Lindenmayer, D.B.; Manning, A.D. Biodiversity, ecosystem function, and resilience: Ten guiding principles for commodity production landscapes. Front. Ecol. Environ. **2006**, *4* (2), 80–86. Available at http://people.anu.edu.au/adrian.manning/ten_guiding_principles.pdf (accessed March 2010).

12. Primavesi, A.M. *Manejo Ecologico do Solo: A Agricultura em Regioes Tropicais*; Nobel: São Paulo-SP, Brazil, 1980; 1–541 (in Portuguese).
13. Bunning, S.; Jiménez, J.J. *Soil Biodiversity Portal: Conservation and Management of Soil Biodiversity and Its Role in Sustainable Agriculture*; FAO: Rome, Italy, 2004. Available at http://www.fao.org/ag/agl/agll/soilbiod/default.stm (accessed March 2010).
14. Food and Agriculture Organization. *Biological Management of Soil Ecosystem for Sustainable Agriculture*; Report of the International Technical Workshop: Londrina-PR, Brazil, 2002. FAO: Rome, Italy, 2003; 1–37 (World Soil Resources Report, 101). Available at ftp://ftp.fao.org/docrep/fao/006/Y4810E/Y4810E00.pdf (accessed March 2010).
15. Food and Agriculture Organization. *Integrated Crop Management: An International Technical Workshop Investing in Sustainable Crop Intensification—The Case for Improving Soil Health*; FAO: Rome, Italy, 2008; Vol. 6, 1–139. Available at http://www.fao.org/docrep/012/i0951e/i0951e.pdf (accessed March 2010).
16. Dumanski, J.; Gameda, S.; Pieri, C. *Indicators of Land Quality and Sustainable Land Management: An Annotated Bibliography*; The World Bank: Washington, DC, 1998; 1–126. Available at http://books.google.com.br/books?id=vY3HjKEJitkC&printsec=frontcover&dq=Indicators+of+and+Quality+and+Sustainable+Land+Management:+An+Annotated+Bibliography+dumanski&source=bl&ots=y67bbN2_L-&sig=hveLwdTFCNXfoPEpXFRTYdNRIMo&hl=pt-BR&ei=iIvxS9HQLsX_lgfW-cGzCA&sa=X&oi=book_result&ct=result&resnum=1&ved=0CBoQ6AEwAA (accessed March.2010).
17. Tilman, D.; Cassman, K.G.; Matson, P.A.; Naylor, R.; Polansky, S. Agricultural sustainability and intensive production practices. Nature **2002**, *418*, 671–677. Available at http://pangea.stanford.edu/research/matsonlab/members/PDF/TilmanNaylorMatson2002.pdf (accessed March 2010).
18. Scherer-Lorenzen, M.; Palmborg, C.; Prinz, A.; Schulze, E.-D. The role of plant diversity and composition for nitrate leaching in grasslands. Ecology **2003**, *84* (6), 1539–1552. Abstract available at http://www.jstor.org/pss/3107974 (accessed March 2010).
19. Liebmann, H. *Ein Planet wird unbewohnbar: ein Suenderegister der Menschheit von der Antike bis zur Gegenwart*; R. Piper & Co: Munchen, Germany, 1973; 1–181 (in German).
20. United Nations Environment Programme. *The United Nations Convention to Combat Desertification: A New Response to an Age-Old Problem*; UNDPI: New York, 1997. Available at http://www.un.org/ecosocdev/geninfo/sustdev/desert.htm (accessed March 2010).
21. United Nation Convention to Combat Desertification. GBO-3's Significant Findings Also in What's Unsaid about Landbased Biological Diversity. Press Release, 12/10, 2010. Available at http://www.unccd.int/publicinfo/pressrel/showpressrel.php?pr=press11_05_10 (accessed March 2010).
22. Menne, B.; Bertollini, R. The health impacts of desertification and drought. Down to Earth (Newsletter of the Convention to Combat Desertification) 2000 (14), 4–7. Available at http://www.google.com.br/url?sa=t&source=web&ct=res&cd=1&ved=0CB4QFjAA&url=http%A%2F%2Fwww.unccd.int%2Fpublicinfo%2Fnewsletter%2Fno14%2Fnews14eng.pdf&rct=j&q=The+health+impacts+of+desertification+and+drought&ei=EpnxS5y2KoL6lwf-9Ki0CA&usg=AFQjCNHNGLIR1oJFAQZQzDiEUeqHzC9k-Q (accessed March 2010).
23. Food and Agriculture Organization. *Carbon Sequestration in Dryland Soils. Chapter 2. The World's Dry Land* (World Resources Report, 102). FAO/Natural Resources Management and Environment Department: Rome, Italy, 2004; 1-129. Available at http://www.fao.org/docrep/007/y5738e/y5738e06.htm and http://www.fao.org/ag/agl/agll/wrb/soilres.stm (accessed March 2010).
24. United Nation Convention to Combat Desertification. *Only One Earth—Drylands Are Vital*; UNCCD: Bonn, Germany, 2010. Available at http://www.unccd.int/publicinfo/announce/earth_day.php (accessed March 2010).
25. Food and Agriculture Organization. *Conservation Agriculture*. Available at http://www.fao.org/ag/ca/ (accessed March 2010).

26. Food and Agriculture Organization. Realizing the economic benefits of agroforestry: Experience, lessons and challenges. In *State of the World's Forests*; FAO: Rome, Italy, 2005; 88–97. Available at ftp://ftp.fao.org/docrep/fao/007/y5574e/y5574e09.pdf (accessed March 2010).
27. Hailu, M.; Landford, K.; Selvarajah-Jaffery, R.; Vanhoutte, K. (Coord.) *Agroforestry—A Global Land Use*. World Agroforestry Centre: Nairobi, Quenia, 2009; 1–53 (Annual Report 2008–2009). Available at http://www.worldagroforestry.org/downloads/publications/PDFs/B16416.PDF (accessed March 2010).
28. Dumanski, J.; Peiretti, R.; Benites, J.R.; McGarry, D.; Pieri, C. The paradigm of conservation agriculture. Proceedings of the World Association on Soil and Water Conservation, 2006, 58–64. Available at http://www.unapcaem.org/publication/ConservationAgri/ParaOfCA.pdf (accessed March 2010).
29. García-Torres, L.; Martínez-Vilela, A.; Holgado-Cabrera, A.; Gónzalez-Sánchez, E. Conservation agriculture, environment and economic benefits. International Congress of the European Society for Soil Conservation, 2000, 4, Valencia-Spain. Proceedings, 2000; 1–10. Available at http://www.unapcaem.org/publication/ConservationAgri/CA1.pdf (accessed March 2010).
30. Ngandwe, T. Conservation agriculture boosts yields and incomes. SciDev.Net 2006 (January 26). Available at http://www.scidev.net/en/news/conservation-agriculture-boosts-yields-and-incom.html (accessed March 2010).
31. Altieri, M.A. The ecological role of biodiversity in agroecosystems. Agric., Ecosyst. Environ. **1999**, *74*, 19–31. Available at http://comunidades.mda.gov.br/o/1540391 (accessed March 2010).
32. Uphoff, N. *Agroecologically Sound Agricultural Systems: Can They Provide for the World's Growing Populations?* International Research of Food Security, Natural Resource Management and Rural Development: Tropentag, 2005. University of Hohenheim: Stuttgart, Germany, 2005. Available at http://www.tropentag.de/2005/proceedings/node181.html and http://www.tropentag.de/2005/proceedings/node3.html (accessed March 2010).
33. Uphoff, N. Agricultural futures: What lies beyond modern agriculture? Trop. Agric. Assoc., Newsl., London, U.K., **2007**, *27* (3; September), 13–19. Available at http://www.slideshare.net/SRI.CORNELL/0906-agricultural-development-what-comes-after-modern-agriculture (accessed March 2010).
34. Wilkinson, K.M.; Elevitch, C.R. *Integrating Understory Crops with Tree Crops: An Introductory Guide for Pacific Islands*; Permanent Agriculture Resources: Holualoa, Hawaii, 2000; 1–50. Available at http://www.agroforestry.net/pubs/Understory.pdf (accessed March 2010).
35. Elevitch, C.R.; Wilkinson, K.M. *Agroforestry Guides for Pacific Islands*; Permanent Agriculture Resources: Holualoa, Hawaii, 2000; 1–239. Available at http://www.agroforestry.net/afg/ (accessed March 2010).
36. Beetz, A. *Agroforestry Overview*; ATTRA: Fayetteville, Arkansas, 2002; 1–16. Available at http://attra.ncat.org/attra-pub/PDF/agrofor.pdf and http://attra.ncat.org/attra-pub/agroforestry.html (accessed March 2010).
37. Davies, K. *Indian Agroforestry: Some Ecological Aspects of Northeastern India Agroforestry Practices*, 1984. Available at http://www.daviesand.com/Papers/Tree_Crops/Indian_Agroforestry/ (accessed March 2010).
38. Dixon, R.K. Agroforestry systems: Sources or sinks of greenhouse gases? Agroforestry Syst. **1995**, *31* (2), 99–116. Abstract available at http://www.springerlink.com/content/g106873871666629/ (accessed March 2010).
39. Leakey, R.R.B.; Tchoundjeu, Z. Diversification of tree crops: Domestication of companion crops for poverty reduction and environmental services. Exp. Agric. **2001**, *37* (3), 279–296. Available at http://www.wanatca.org.au/acotanc/Papers/Leakey-2/index.htm (accessed March 2010).
40. Convention on Biological Diversity. *A Good Practice Guide: Sustainable Forest Management, Biodiversity and Livelihoods*; CBD: Montreal, Canada, 2009; 1–47. Available at http://www.cbd.int/development/doc/cbd-good-practice-guide-forestry-booklet-web-en.pdf (accessed March 2010).

41. Hopper, K.; Summers, D., Eds. *Protecting the Source: Land Conservation and the Future of America's Drinking Water;* The Trust for Public Land and American Water Works Association: San Francisco, 2004; 1–56 (Water Protection Series). Available at http://earthtrends.wri.org/pdf_library/feature/eco_fea_value.pdf (accessed March 2010).
42. Rand Corporation. *New York City Depends on Natural Water Filtration;* Rand Co.: Santa Monica, California, 2007. Available at http://www.rand.org/scitech/stpi/ourfuture/NaturesServices/sec1_watershed.html (accessed March 2010).
43. World Health Organization. *Environmental Health;* WHO: Geneva, Switzerland, 2010. Available at http://www.who.int/topics/environmental_health/en/ (accessed March 2010).
44. World Health Organization. *Ecosystems and Health;* WHO: Geneva, Switzerland, 2010. Available at http://www.who.int/globalchange/ecosystems/en/ (accessed March.2010).
45. World Health Organization. *Water, Sanitation and Hygiene;* WHO: Geneva, Switzerland, 2010. Available at http://www.who.int/water_sanitation_health/en/ (accessed March 2010).
46. World Health Organization. *Water for Health: Taking Charge;* WHO: Geneva, Switzerland, 2001; 1–40. Available at http://www.who.int/water_sanitation_health/takingcharge/en/ (accessed March 2010).
47. Skuttle. *Impact of Relative Humidity on Air Quality;* Skuttle: Marietta, OH, USA, 2010. Available at http://www.skuttle.com/pdfs/skuttlehumidity.pdf (accessed March 2010).
48. Ciocci, M.V. *Reflexos do excesso de calor na saúde, e na redução da produtividade;* Cabano Engenharia: Belém-PA, Brazil, 2003. Available at http://www.cabano.com.br/excesso_de_calor.htm (accessed March 2010) (in Portuguese).
49. Corocoran, E.; Nellermann, C.; Baker, E.; Bos, R.; Osborn, D.; Savelli, H., Eds. *Sick Water? The Central Role of Wastewater Management in Sustainable Development. A Rapid Response Assessment*; United Nations Environment Programme: Nairobi, Quenia, 2010; 1–88. Available at http://www.cbsnews.com/htdocs/pdf/SickWater_screen.pdf (accessed March 2010).
50. Palaniappan, M.; Gleick, P.H.; Aallen, L.; Cohen, M.J.; Christian-Smith, J.; Smith, C.; Ross, N. *Clearing the Waters: A Focus on Water Quality Solutions*; United Nations Environment Programme: Nairobi, Quenia, 2010; 1–91. Available at http://www.unep.org/PDF/Clearing_the_Waters.pdf (accessed March 2010).
51. United Nations Environment Programme. *World Water Day 2010 Highlights Solutions and Calls for Action to Improve Water Quality Worldwide;* UNEP: Nairobi, Quenia, 2010 (March). Available at http://www.unep.org/Documents.Multilingual/Default.asp?DocumentID=617&ArticleID=6505&l=en&t=long (accessed March 2010).
52. Addams, L.; Boccaletti, G.; Kerlin, M.; Stuchtey, M. *Charting our Water Future: Economic Frameworks to Inform, Decision-Making;* WRG—2030. Water Resources Group/ McKinsey and Company: Chicago, USA, 2009; 1–198. Available at http://www.mckinsey.com/App_Media/Reports/Water/Charting_Our_Water_Future_Full_Report_001.pdf and http://www.nestle.com/InvestorRelations/Events/AllEvents/2030+Water+Resources.htm (accessed March 2010).
53. Gonçalves, C.E.C. *Ruas confortáveis, ruas com vida.: proposição de diretrizes de desenho urbano bioclimático para vias públicas. Av. Juscelino Kubitscheck, Palmas, TO*; UNB: Brasí0lia, Brazil, 2009; 1–149. (MSc Dissertation). Available at http://repositorio.bce.unb.br/bitstream/10482/3901/2/2009_CarlosEduardoCavalheiroGoncalves_pag_82_ate_final.pdf and http://repositorio.bce.unb.br/bitstream/10482/3901/1/2009_CarlosEduardoCavalheiroGoncalves_ate_pag_81.pdf (accessed March 2010).
54. Gash, J.H.C.; Shuttleworth, W.J. Tropical deforestation: Albedo and the surface-energy balance. Clim. Change **1991**, *19* (1–2), 123–133. Partially available at http://books.google.com.br/books?id=XB16EtxyZPUC&pg=PA123&lpg=PA123&dq=Tropical+deforestation:+albedo+and+the+surface-energy+balance&source=bl&ots=2PFRYbSs0l&sig=LOW2DnQ-d4-ZNgthrnFY4BxzYhu0&hl=pt-BR&ei=ccX1S9S3JoL-8Aavm9njCg&sa=X&oi=book_result&ct=result&resnum=5&ved=0CDwQ6AEwBA#v=onepage&q=Tropical%20deforestation%3A%20albedo%20and%20the%20surface-energy%20balance&f=false (accessed March 2010).

55. Gordeau, J. *Albedo*; Environmental Science Published for Everybody Round the Earth—ESPERE; Educational Network on Climate: Cracow, Poland, 2004. Available at http://www.atmosphere.mpg.de/enid/25w.html (accessed March 2010).
56. Russell, R. *Global Warming, Clouds, and Albedo: Feedback Loops*; National Earth Science Teachers Association: Boulder, CO, USA, 2007. Available at http://www.windows.ucar.edu/tour/link=/earth/climate/warming_clouds_albedo_feedback.html (accessed March 2010).
57. Hartmann, D.L. *Global Physical Climatology*; Academic Press: California, USA, 1994; 1–386 (International Geophysics, v. 56). Partially available at http://books.google.com.br/books?id=Zi1coMyhlHoC&pg=PA90&lpg=PA90&dq=albedo+straw&source=bl&ots=_SiR7JL67l&sig=8XHCaS4O43ICJR0XkagXhXN_5rs&hl=pt-BR&ei=urRgS9vqMsWtgeipuzYDQ&sa=X&oi=book_result&ct=result&resnum=10&ved=0CDgQ6AEwCQ#v=onepage&q=albedo%20 straw&f=false (accessed March 2010).
58. Yang, Z.L. *Physical Climatology: The Global Energy Balance*, 2006; 1–22. Available at http://www.geo.utexas.edu/courses/387H/Lectures/chap2.pdf and http://www.geo.utexas.edu/courses/387H/default.htm (accessed March 2010).
59. Sharratt, B.S.; Campbell, G.S. Radiation balance of a soil- straw surface modified by straw color. Agron. J. **1994**, *86;* 200–203. Abstract available at http://agron.scijournals.org/cgi/content/abstract/86/1/200 (accessed March 2010).
60. Triparthi, R.P.; Katiyara, T.P.S. Effect of mulches on the thermal regime of soil. Soil Tillage Res. **1984**, *4* (4), 381–390. Abstract available at http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6TC6-48XDCKB-1B&_user=10&_coverDate=07%2F31%2F1984&_rdoc=1&_fmt=high&_orig=search&_sort=d&_docanchor=&view=c&_searchStrId=1183152286&_rerunOrigin=google&_acct=C000050221&_version=1&_urlVersion=0&_userid=10&md5=19d963847840c192c30e24092fba1ad2 (accessed March 2010).
61. Heat Island Group. *Cool Roofs*; University of California: Berkeley, USA, 2000. Available at http://eetd.lbl.gov/HeatIsland/ (at cool roofs) (accessed March 2010).
62. Levinson, R.; Berdahl, P.; Akbari, H. Solar spectral optical properties of pigments—Part II: Survey of common colorants. Sol. Energy Mater. Sol. Cells **2005**, *89* (4), 351–389. Available at http://www.odulo.com/eric/Enviroglobal/email%20documentsfPigments2.pdf (accessed March 2010).
63. European Space Agency. World fire maps now available online in near-real time. ESA Observing the Earth, News, 2006 (May 26). Available at http://www.esa.int/esaEO/SEMRBH9ATME_index_0.html (accessed March 2010).
64. United States Department of Agriculture. *Soil Moisture Regimes Map*. USDA/Natural Resources Conservation Service/Soil Survey Division/World Soil Resources: Washington, DC, 1999. Available at http://soils.usda.gov/use/worldsoils/mapindex/smr.html (accessed March 2010).
65. Prata, F. *Global Distribution of Maximum Land Surface Temperature Inferred from Satellites*; CSIRO Atmospheric Research, Aspendale: Victoria, 2000; 1–13 (see page 8). Available at http://www.eoc.csiro.au/associates/aatsr/lst_atlas.pdf (accessed March 2010).
66. The Centre for Australian Weather and Climate Research. *Current Climate Charts: Outgoing Longwave Radiation (OLR) Products—Latest Day and Night-Pass Data*; Bureau of Meteorology Research Centre: Australia, 2009. Available at http://cawcr.gov.au/bmrc/clfor/cfstaff/matw/maproom/index.htm (accessed March 2010).
67. Greenpeace. The World's Last Intact Forest Landscapes (poster A1), 2006. Available at http://www.intactforests.org/pdf.publications/World.IFL.2006.poster_low.pdf and http://www.greenpeace.org/international/campaigns/forests/our-disappearing-forests (accessed March 2010).
68. Natural Environment Research Council. *Satellites Reveal that Green Means Rain in Africa*; ScienceDaily, Sept. 27, 2006. Available at http://www.sciencedaily.com/releases/2006/09/060925064922.htm (accessed March 2010).

69. Coghla, A. More crops for Africa as trees reclaim the desert. New Sci. 2006 (October 14) (Article preview). Available at http://www.newscientist.com/article/dn10293-more-crops-for-africa-as-trees-reclaim-the-desert.html (accessed March 2010).
70. Makarieva, A.M.; Gorshkov, V.G.; Li, B.-L. Conservation of water cycle on land via restoration of natural closed-canopy forests: Implications for regional landscape planning. Ecol. Res. **2006**, *21,* 897–906. Available at http://www.biotic-regulation.pl.ru/offprint/wat_pr1.pdf (accessed March 2010).
71. Makarieva, A.M.; Gorshkov, V.G. Biotic pump of atmospheric moisture as driver of the hydrological cycle on land. Hydrol. Earth Syst. Sci. **2007**, 11, 1013–1033. Available at http://www.biotic-regulation.pl.ru/offprint/hess07.pdf (accessed March 2010).
72. Rowntree, P.R. Review of general circulation models as a basis for predicting the effects of vegetation change on climate (Ch. 8). In *Forests, Climate, and Hydrology*: *Regional Impacts;* Reynolds, E.R.C., Thompson, F.B., Eds.; The United Nations University: Tokyo, Japan, 1988. Available at http://www.unu.edu/unupress/unupbooks/80635e/80635E00.htm#Contents (accessed March 2010).
73. Sheil, D.; Murdiyarso, D. How forests attract rain: An examination of a new hypothesis. BioScience **2009**, *59* (4), 341–347. Abstract available at http://www.bioone.org/doi/abs/10.1525/bio.2009.59.4.12?journalCode=bisi and http://www.ncriverwatch.org/wordpress/2009/12/14/forests-attract-rain-an-examination-of-a-new-hypothesis/ (accessed March 2010).
74. Bonan, G.B. Forests and climate change: Forcing feedbacks and the climate benefits of forests. Science **2008**, *320* (5882), 1444–1449. Abstract available at http://www.sciencemag.org/cgi/content/abstract/sci;320/5882/1444 (accessed March 2010).
75. Takata, K.; Saito, K.; Yasunari, T. Changes in the Asian monsoon climate during 1700–1850 induced by pre-industrial cultivation. PNAS Online **2009**, *106* (24; June). Available at http://www.nagoya-u.ac.jp/en/pdf/research/activities/090529_hyarc_yasunari.pdf (accessed March 2010).
76. Pearce, F. Rainforests may pump wind worldwide. New Sci. 2009 (2702; April 1), 1–3. Available at http://www.ideastransformlandscapes.org/media/uploads/File/Rainforests%20may%20pump%20winds%20worldwide.pdf (accessed March 2010).
77. Rockström, J.; Steffen, W.; Noone, K.; Persson, A.; Chapin, F.S.; Lambin, E.F.; Lenton, T.M.; Scheffer, M.; Folke, C.; Schellnhuber, H.J.; Nykvist, B.; de Wit, C.A.; Hughes, T.; van der Leeuw, S.; Rodhe, H.; Sörlin, S.; Snyder, P.K.; Costanza, R.; Svedin, U.; Falkenmark, M.; Karlberg, L.; Corell, R.W.; Fabry, V.J.; Hansen, J.; Walker, B.; Liverman, D.; Richardson, K.; Crutzen, P.; Foley, J.A. A safe operating space for humanity. Nature **2009**, *461* (September), 472–475. Abstract available at http://www.nature.com/news/specials/planetaryboundaries/index.html (accessed March 2010).
78. Towers, P.T.; Lumper, K. *Definitions of Sustainability*; The University of Reading: Reading, U.K., 2010. Available at http://www.ecifm.rdg.ac.uk/definitions.htm (accessed March 2010).
79. Food and Agriculture Organization. *Land Quality Indicators and Their Use in Sustainable Agriculture and Rural Development*. Land and Water Bulletin 5, FAO, UNDP, UNEP, World Bank: Rome, 1996; 1–217. Available at http://www.mpl.ird.fr/crea/taller-colombia/FAO/AGLL/pdfdocs/landqual.pdf (accessed March 2010).
80. Munoz, L. Understanding sustainability versus sustained development by means of a WIN Development Model. In *Sustainability Review;* Flint, W., Ed.; Five E's Unlimited: Pungoteague, VA, Canada, 1999 (1; September 6); 1–18. Available at http://theomai.unq.edu.ar/artmunoz001.htm (accessed March 2010).
81. Wikipedia. *Nachhaltigkeit* (Sustainability). Available in German at: http://de.wikipedia.org/wiki/Nachhaltigkeit and http://de.wikipedia.org/wiki/Nachhaltigkeit_%28Forstwirtschaft%29 (accessed March 2010).
82. *Kanshie, T.K. Five Thousand Years of Sustainability? A Case Study on Gedeo Land Use (Southern Ethiopia); Kanshie:* Wageningen, Netherlands, 2002; 1–295 (Treebook, 5). Available at http://www.treemail.nl/books (accessed March 2010).

83. Chambers, N.; Simmons, C.; Wackernagel, M. *Sharing Nature's Interest: Ecological Footprints as an Indicator of Sustainability;* James and James, Earthscan: London, 2001; 1–200. Partially available at http://en.book2down.com/Sharing-Natures-Interest-Using-Ecological-Footprints-As-An-Indicator-Of-Sustainability/186751 (accessed March.2010).
84. Wikipedia. *List of Countries by Ecological Footprint.* 2008. Available at http://en.wikipedia.org/wiki/List_of_countries_by_ecological_footprint (accessed March 2010).
85. Schaefer, F.; Luksch, U.; Steinbach, N.; Cabeça, J.; Hanauer, J. *Ecological Footprint and Biocapacity: The World's Ability to Generate Resources and Absorb Wastes in a Limited Time Period* (Working paper and studies). Office for Official Publications of the European Communities; Luxembourg, 2006; 1–11. Available at http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-AU-06-001/EN/KS-AU-06-001-EN.PDF (accessed March 2010).
86. Global Footprint Network. *Footprint for Nations. 2009;* Global Footprint Network: Oakland, CA, USA, 2010. Available at http://www.footprintnetwork.org/en/index.php/GFN/page/footprint_for_nations/ (accessed March 2010).
87. World Wide Fund for Nature. *Living Planet Report 2008;* WWF International: Gland, Suisse, 2008; 1–48. Available at http://assets.panda.org/downloads/living_planet_report_2008.pdf (accessed March 2010).
88. Best, A.; Giljun, S.; Simmons, C.; Blobel, D.; Lewis, K.; Hammer, M.; Cavalieri, S.; Lutter, S.; Magguirre, C. *Potential of the Ecological Footprint for Monitoring Environmental Impacts from Natural Resources Use: Analysis of the Potential of the Ecological Footprint and Related Assessment Tools for use in the EU's Thematic Strategy on the Sustainable Use of Natural Resources;* Report to the European Commission, DG Environment, 2008; 1–312. Available at http://ec.europa.eu/environment/natres/pdf/footprint.pdf (accessed March 2010).
89. Waterfootprint Network. *Water Footprint versus Water Scarcity, Self-Sufficiency and Water Import Dependency per Country*; Water Footprint Network/University of Twente: Enschede, Netherlands, 2001. Available at http://www.waterfootprint.org/?page=files/NationalStatistics (accessed March 2010).
90. Waterfootprint Network. *Water Footprints of Crop and Livestock Products (m3/t) for Some Selected Countries (1997–2001)*; Water Footprint Network/University of Twente: Enschede, Netherlands, 2001. Available at http://www.waterfootprint.org/?page=files/Productwaterfootprint-statistics (accessed March 2010).
91. Hoekstra, A.Y.; Chapagain, A.K. Water footprint of nations: Water use by people as a function of their consumption pattern. Water Resour. Manage. **2007**, *21,* 35–48. Available at http://www.waterfootprint.org/Reports/Report18.pdf and http://www.waterfootprint.org/Reports/Hoekstra_and_Chapagain_2007.pdf (accessed March 2010).
92. Hoekstra, A.Y. *Water Neutral: Reducing and Offsetting the Impacts of Water Footprints.* Research report series nr. 28—Value of Water, UNESCO-IHE, 2008; 1–42. Available at http://www.waterfootprint.org/Reports/Report28-WaterNeutral.pdf (accessed March 2010).
93. Hoekstra, A.Y.; Chapagain, A.K.; Aldaya, M.M.; Mekonnen, M.M. *Waterfootprint Manual: State of the Art 2009;* Netherlands: Water Footprint Network: Enschede, Netherlands, 2009; 1–131. Available at http://www.waterfootprint.org/downloads/WaterFootprintManual2009.pdf (accessed March 2010).
94. Wikipedia. *Carbon Footprint*. Available at http://en.wikipedia.org/wiki/Carbon_footprint (accessed March 2010).
95. Wikipedia. *List of Countries by Carbon Dioxide Emission per Capita.* Available at http://en.wikipedia.org/wiki/List_of_countries_by_carbon_dioxide_emissions_per_capita (accessed March 2010).
96. Global Footprint Network. *Carbon Footprint*; Global Footprint Network: Oakland, CA, USA, 2009. Available at http://www.footprintnetwork.org/en/index.php/GFN/page/carbon_footprint/ (accessed March 2010).

97. Odum, H.T. *Environmental Accounting, Emergy and Decision Making;* John Wiley: New York, USA, 1996; 1–370. Partially available at http://dieoff.org/page170.htm (accessed March 2010).
98. Hau, J.L.; Bakshi, B.R. *Promise and Problems of Emergy Analysis;* Ohio State University: Ohio, USA, 2003; 1–13. Available at http://www.che.eng.ohio-state.edu/~bakshi/EcolModel3.pdf (accessed March 2010).
99. Ferreyra, C. Emergy analysis of one century of agricultural production in the rolling pampas of Argentina. Int. J. Agric. Resour. Governance Ecol. **2006**, *5* (2–3), 185–205. Available at http://snre.ufl.edu/graduate/files/publicationsbyalumni/Ferreyra%202006.pdf (accessed March 2010).
100. Martin, J.F.; Tilley, D.R. Accounting for environmental sustainability with emergy analysis. In: *Encyclopedia of Soil Science*; Lal, R., Ed.; Taylor and Francis Group: New York, 2007. Abstract available at http://www.mformaworld.com/smpp/content~db=all~content=a788663399 (accessed March 2010).

8

Biofertilizers

J. R. de Freitas

Introduction

Biofertilizers include microorganisms and their metabolites that are capable of enhancing soil fertility, crop growth, and/or yield. These include both indigenous microbes and microbial inoculants, that is, microorganisms that replace fertilizers or increase a crop's fertilizer use efficiency. Soil microorganisms such as bacteria, ectomycorrhiza, arbuscular mycorrhizal fungi, and soil algae, especially the N_2-fixing cyanobacteria have potential as biofertilizers. Nitrogen-fixing inoculants based on *Rhizobium* species were among the first biofertilizers introduced into agroecosystems back in the 19th century. In the 21th century, biofertilizers will become an increasingly important area of research and development.[1] The use of fertilizers and pesticides has increased steadily since the 1970s; consequently, concerns about the impacts of these chemicals on land, air, and water have become significant environmental issues. Biofertilizers provide an alternative to agricultural chemicals as more sustainable and ecologically sound practices to increase crop productivity. Biofertilizer sales forecasts in the United States for the years 2001 and 2006 represent $690 million and $1.6 billion, respectively. Examples of some biofertilizers currently in use worldwide are shown in Table 1.[2]

TABLE 1 Organisms, Mode of Action, Crops, and Producers of Biofertilizers Currently in Use for Agriculture

Type	Mode of Action	Crop	Used in
Rhizobium spp.	N_2 fixation	Legumes	Russia; several countries
Cyanobacteria	N_2 fixation	Rice	Japan; several countries
Azospirillum spp.	N_2 fixation	Cereals	Several countries
Mycorrhizae	Nutrient acquisition	Conifers	Several countries
Penicillium bilaii	P solubilization	Cereals, legumes	Western Canada
Directed compost	Soil fertility	All plants	Several countries
Earthworm	Humus formation	Vegetables, flowers	Cottage industry

Source: Adapted from Tengerdy and Szakacs.[2] Copyright 1998 Elsevier Science.

TABLE 2 Market Price and Potential Price Increments with Yield Increases of 5% and 25% in Selected Vegetable Crops

		Price Increments ($ per acre)	
Vegetable Crop	Market Price ($ per acre)	5% Yield Increase	25% Yield Increase
Carrot	4,520	226	1130
Cauliflower	4,179	209	1045
Celery	10,132	507	2533
Cucumber	3,296	165	824
Lettuce	5,882	294	1471
Tomato	9,966	498	2492

Source: Adapted from USDA.[3]

Market for Biofertilizers

The market potential for biofertilizers includes the high value vegetable industry. A comparison of the base value of various crops and the increased value that can be obtained as the crop yield rises is illustrated in Table 2. Due to high nutrient requirements and high susceptibility to diseases, vegetable growers spend substantial amounts to protect this valuable produce. For example, average broccoli and tomato crops grown in California require ca. $62 and $170 worth of fertilizer and/or fungicide per acre, respectively. When the U.S. government prohibits the use of methyl bromide as a soil fumigant, as anticipated in 2005, development of biological products will be stimulated as an alternative to the use of chemicals.[3]

Mechanisms of Growth Promotion

The mechanisms covered in this entry are those that show commercial market potential; thus, it does not include all modes of action by which biofertilizers promote crop growth. Biofertilizers promote crop growth using several mechanisms with the primary one varying as a function of environmental conditions. Although the mechanisms of commercially available biofertilizers are not always entirely understood, growth promotion has been classified as the result of indirect or direct mechanisms. Indirect plant growth promotion may be associated with the repression of negative effects caused by phytopathogenic organisms, that is, biological control. Conversely, direct growth promotion mechanism may either provide some compound essential to crop development and/or stimulate nutrient uptake. Biofertilizers based on biological control agents *Mycorrhizae* and *Rhizobium* will be discussed in more detail elsewhere in this encyclopedia (e.g., Biological Pest Controls; Mycorrhiza; Rhizobia).

Phytohormones

Production of phytohormones is a commonly noted direct mechanism of plant growth promotion.[4] The nature of growth response may be the result of phytohormone production in the rhizosphere. Phytohormones are produced by many biofertilizers and include a list of plant growth regulators that are important in the plant's metabolism. For example, auxins such as indole-3-acetic acid are known for their ability to stimulate root cell division, differentiation, and promote cell elongation. Other phytohormones such as cytokinin, gibberellin, and ethylene also play key roles in plant development and have been reported to increase the growth of various commercial crops. The horticulture market for biofertilizer products based on gibberellin and other auxins is currently estimated at $600 million per year.

Plant Nutrient Acquisition

Several direct mechanisms are responsible for increased nutrient acquisition.

Biological Nitrogen Fixation (BNF)

Nitrogen (N) is an essential macronutrient, that is, it is the key building block of proteins, thus an indispensable component of the protoplasm of microorganisms, animals, and plants. The supply of biologically available N to agriculture through BNF represents ca. 140×10^6 ton/year, globally.[1] Therefore, BNF represents an economy of millions of dollars. N_2-fixation by free-livingbacteria such as *Azospirillum, Azotobacter, Bacillus,* and *Derxia* species have beenexploited in agricultural systems for many decades and constitute an important source of Ninput into agro-ecosystems.[5] Other BNF associations include the water fern *Azolla* that forms asymbiosis with the heterocystous cyanobacterium *Anabaena azollae.* The *Azolla–Anabaena* system has been used as a biofertilizer in Vietnam and China for rice production and has the potential to supply the entire N requirement (30–50 kg N/ha) for a rice crop during the growing season.[6] Another diazotroph, the N2-fixing actinomycete *Frankia,* forms nodules (actinor- rhizae) in ca. 17 genera of nonlegume wood species with *Alnus* (alder) and the genus *Casuarina* being the most important for forestry and agriculture. Estimates of total N2 fixed range between 50–250 kg/ha/year, depending on the plant species and region. However, in some cases, inoculation with *Frankia* is necessary for nodulation to occur. Ac- tinorrhizal plant species have been successfully inoculated with *Frankia* on a large scale. For example, millions of actinorrhizal trees, especially *Alnus* spp., inoculated with *Frankia* were used in land reclamation programs established in Canada.[7]

Phosphorus Solubilization

Certain microorganisms are very effective in solubilizing phosphorus (P) from insoluble phosphate compounds such as hydroxyapatite through the action of organic acids. Numerous claims have been made about biofertilizers that can enhance plant growth by solubilizing P. A classical example is the bacterium *Bacillus megaterium,* which was formulated into an inoculant under the name of Phosphobacterin in the former Soviet Union. A similar biofertilizer based on P-solubilizing fungi is currently marketed in Canada as JumpStart™ for use on wheat, canola, mustard, and N2-fixing legumes.

Microbial Siderophore Uptake

Iron (Fe) is an important plant micronutrient. Plants assimilate iron by acidifying the rhizosphere and/or secreting phyto-siderophores with subsequent reassimilation of the iron–siderophore complex. However, plants also may benefit from the direct uptake of microbial siderophor–iron complexes. For example, some biofertilizers synthesize siderophores that can solubilize and sequester Fe from soil and provide it to plant cells, thus contributing to the nutrition and development of crops. In fact, studies demonstrate that ferric pseudobactin 358 may stimulate chlorophyll synthesis in carnation and barley.[8]

Other Nutrients

Studies with *Azospirillum* spp. and plant growth-promoting rhizobacteria (PGPR) have demonstrated the ability of these biofertilizers to promote enhancement of nutrient and water uptake into the plant. For example, inoculation of winter wheat seeds with pseudomonad PGPR stimulated the uptake of soil-Fe and fertilizer-^{15}N by winter wheat cultivated in two Canadian soils.[9] Similarly, inoculation of canola seeds with a *Pseudomonas putida* increased phosphate uptake from nutrient solution.[10] In these cases, the authors speculated that plant growth regulators produced by the biofertilizers in the plant's rhizosphere stimulated root development which, in turn, enhanced nutrient acquisition.

Future Research Directions

It is clear that commercial crops can benefit directly from biofertilizers. Certainly, with the development of molecular biology and genetic manipulation of biofertilizers to improve N_2-fixation, rhizosphere competence and ability to be used together with specific chemicals, will contribute to an integrated strategy to reduce the total amount of chemicals used in agriculture. Although biofertilizer products are currently available on the market, consistency is still the major factor that limits their use. Elucidation of mechanisms, development of stable formulations, effective delivery systems, and field demonstration of effective biofertilizers will definitely improve reliability and enhance their use as commercial biofertilizers.

References

1. Killham, K. *Soil Ecology*; Cambridge University Press: Cambridge, U.K., 1994; 242.
2. Tengerdy, R.P.; Szakács, G. Perspectives in agrobiotechnology. J. Biotechnol. **1998**, *66,* 91–99.
3. USDA. USDA Economics and Statistics System. National Agricultural Statistics Service; Cornell University. http://mann77.mannlib.cornell.edu/reports/nassr/fruit/pvg-bban/vegetables_annual_summary-01.16.98 (accessed June 1999).
4. Glick, B.R. The enhancement of plant growth by freeliving bacteria. Can. J. Microbiol. **1995**, *41,* 109–117.
5. Pankhurst, C.E.; Lynch, J.M. The role of soil microbiology in sustainable intensive agriculture. Adv. Plant Pathol. **1995**, *11*, 230–247.
6. Zuberer, D.A. Biological Dinitrogen Fixation: Introduction and Nonsymbiotic. In *Principles and Applications of Soil Microbiology;* Sylvia, D.M., Fuhrmann, J.J., Hartel, P.G., Zuberer, D.A., Eds.; Prentice Hall: Upper Saddle River, NJ, 1998; 295–321.
7. Périnet, P.; Brouillette, J.G.; Fortin, J.A.; Lalonde, M. Large scale inoculation of actinorrhizal plants with frankia. Plant Soil **1985**, *87,* 175–183.
8. Duiff, B.J.; de Kogel, W.J.; Bakker, P.A.H.M.; Schipper, B. Significance of Pseudobactin 358 for the Iron Nutrition of Plants. In *Improving Plant Productivity with Rhizosphere Bacteria;* Ryder, M.H., Stephens, P.M., Bowen, G.D., Eds.; Third International Workshop on Plant-Growth Promoting Rhizobacteria, Adelaide, Australia, March 7–11, CSIRO Division of Soils: South Australia, 1994; 142–144.
9. De Freitas, J.R.; Germida, J.J. Growth promotion of winter wheat by fluorescent pseudomonads under growth chamber conditions. Soil Biol. Bioch. **1992**, *24,* 1127–1135.
10. Lifshitz, R.; Kloepper, J.W.; Kozlowiski, M.; Simonson, C.; Carlson, J.; Tipping, E.M.; Zaleska, I. Growth promotion of canola (rapeseed) seedlings by a strain of *pseudomonas putida* under gnotobiotic conditions. Can. J. Microbiol. **1987**, *33,* 390–395.

9

Ecosystems: Large-Scale Restoration Governance

Shannon Estenoz,
Denise Vienne,
and Alka Sapat

Introduction

Large-scale ecosystem restoration involves layers of complexity including, at the most fundamental level, ecological cause/effect relationships that may not be well understood, immediately observable, or measurable. Ecological complexity alone ensures that ecosystem restoration will be characterized by difficulty and uncertainty. However, in the context of large-scale ecosystem restoration, ecological complexity is compounded by other complexities including, for example, governance systems and regulatory frameworks not designed to approach restoration at the ecosystem scale and complex political interactions associated with resource consumption, private property rights, and economic development. Restoration leaders have created new governance structures, alternatively termed "collaborative resource governance" (CRG) or "collaborative environmental management," designed to work through and around such layered complexity.[1] While these structures vary in form and scope across restoration programs, there are common threads among them that include the scale and complexity of problems they are designed to address, emphasis on locally tailored solutions, the allowance for experimentation and dynamic adaption, and the importance of information sharing and collaborative problem solving at multiple overlapping levels.[2] The main focus of this entry is on the nature of CRG as a means of managing complexity and cross-jurisdictional issues in large-scale ecosystem restoration. In doing so, we review the evolution of these approaches in environmental and ecosystem management and the characteristics, challenges, and opportunities that such approaches present for governance and ecosystem restoration. More specifically, we discuss collaborative governance structures that have been adopted to manage two of the largest ecosystem restoration programs in the country, Gulf Coast ecosystem restoration and Everglades restoration.[3,4] Similar institutional mechanisms, i.e., ecosystem restoration taskforces, were put in place for both these ecosystems. These two programs present an opportunity for ascertaining whether synergies can be created in an ecosystem restoration program by cross-pollinating engagement at many levels with another restoration program.

We begin our discussion of this topic by examining how collaborative governance has emerged as one of the predominant mechanisms for addressing complex policy problems. Next, we note the evolution of environmental management in the United States toward large-scale ecosystem management through employment of collaborative governance. Although ecosystem restoration projects increasingly rely on this approach for governance and implementation, they are extremely diverse as a reflection of their political, administrative, and environmental contexts. Thus, the following section attempts to derive important lessons gleaned from these specific projects. Then, CRG will be defined as a distinct type of governance model that exhibits several general characteristics. A specific example of CRG, the ecosystem restoration task force model is the focus of the final sections. The Florida Everglades and the Gulf of Mexico, two of the world's largest ecosystem restoration projects that employ the task force model, will be discussed and compared. The conclusion elaborates how experience with these models can serve as the basis for future comparative research and as a guide for governance structures and environmental management.

Governance Models: Collaborative Environmental Management

Since the 1980s, collaborative governance has emerged as a predominant paradigm to resolve various policy issues, and this shift has been part of a gradual and broader movement in both the theory and praxis of public administration. Public managers have increasingly been addressing complex problems through networks, strategic partnerships, alliances, coalitions, contractual relationships, committees, consortia, and councils; these mechanisms have been used to develop collaborative mechanisms across jurisdictions, governments, and sectors. Collaborative governance approaches are thus increasingly being used in response to public demands and to deal with boundary-spanning policy problems that cannot be addressed by traditional regulatory approaches relying on single agencies and jurisdictions.[5,6]

Governance models for natural resource management have mirrored and, in some instances, even led to the gradual shift in collaborative governance, due to the emergence of various forms of partnerships and civic environmentalism that emerged to deal with environmental problems.[7,8] Reasons for the emergence of collaborative environmental management are many, overlapping, and varied. Some types of collaborative environmental governance emerged as a result of federal cutbacks, leading to civic environmentalism and citizens pushing for environmental changes.[8] For the most part, though, forms of collaborative environmental management came about as a result of the problems stemming from command-and-control measures, which often resulted in protracted conflict or failure to achieve key objectives. This led to calls for change from bureaucratic, adversarial, technology-based regulatory approaches, which were the basis for many environmental policies in the 1970s and early 1980s, to "results-based" and voluntary approaches to regulation.[9] Similar to other policy problems, calls for more collaborative measures also stemmed from the complexity of environmental problems that were not effectively addressed by conventional rule-based, top-down, and hierarchical approaches. The latter were seen as problematic in that they did not allow for more democratic forms of participation, stymied potential innovation, were ineffective in addressing multimedia environmental hazards and those stemming from nonpoint sources of pollution, and relied on unrealistic models of administrative and individual rationality.[7,9–11] To deal with these cross-jurisdictional problems, there were calls for more holistic and integrated approaches to deal with ecosystems such as large watersheds and forests. The adoption of more holistic approaches through collaborative environmental management grew during the 1990s particularly the Clinton administration; during this period, the ecosystem approach was greatly expanded both to better administer large restoration projects (such as the Everglades and the Northwest Forest plans) and to address smaller ecosystem and habitat conservation planning.[12]

Collaborative environmental management takes many forms: some entail collaboration between multiple agencies only, while others are characterized by the presence of multiple stakeholders such as actors in the private and nonprofit sectors. In some cases of collaborative environmental management, government has led the efforts; in others, government has encouraged it through grants and other incentives; and, at times, the efforts at collaborative management have been pioneered by other stakeholders such as citizens or nonprofit groups.[13] What distinguishes these collaborative approaches from earlier forms of environmental management is the movement away from command-and-control policies toward those that are both more inclusive of and seek to incentivize public participation. The search for mutually beneficial policy solutions by encouraging broad participation from local stakeholders, underscoring the importance of consensus and voluntary approaches, and building trust-based policy networks are defining characteristics of collaborative environmental approaches. In addition, collaborative environmental governance also involves improvements in scientific understanding of how ecological processes affect resource outcomes across artificial jurisdictional and political boundaries[14] and careful scientific monitoring to allow for managerial adaptation as necessary.[12]

Since collaborative environmental governance is an institutional mechanism for natural resource management, its development and success may depend on the set of institutional rules applied in ecosystem management. As noted by Elinor Ostrom, successful institutions tend to have certain design principles, such as rules adapted to local circumstances, clearly defined resource boundaries, information about resource variability, monitoring and sanction mechanisms, and local conflict resolutions forums.[15] While these principles may not ensure success, they could be contributory factors, as noted in our discussion about CRG structures adopted for large-scale ecosystem restoration.

In the next section, we discuss how collaborative governance approaches have been used in ecosystem management, which has seen a shift from place-based to species-based conservation.

Collaborative Governance Approaches and Ecosystem Management

Over the past 30 years, place-based and species-based approaches to environmental conservation have been giving way to more holistic and collaborative approaches to ecosystem management.[16] The earlier strategies involved prescriptive, or command-and-control, regulatory approaches that targeted discrete, identifiable sources of environmental damage.[17] Traditional regulatory approaches target the "low-hanging fruit" with respect to environmental problems, but they are insufficient for resource management and the restoration of ecosystems that are subject to multiple interdependent and interacting conditions, lacking clear or easily discernable sources of damage.[17] Collaborative ecosystem governance is a response to seemingly intractable environmental problems that do not fall neatly within the domain of any one governmental jurisdiction or agency.[1] Comparatively speaking, there is a great deal of variation in how such collaborative governing bodies are convened, their constellation of participants, the type and pattern of interactions, and the ecosystem problems they manage.

On the surface, these diverse institutional arrangements appear to be distinct problem-specific environmental responses because they conform to the particular resource systems they are designed to manage. However, collaborative ecosystem governance models share common characteristics that set them apart as a distinctive alternative to the traditional legal and institutional approaches relied upon for environmental management.[2] Furthermore, since they are characteristically more flexible, participatory, deliberative, and heterarchical forms of organization, they are considered to be more responsive, legitimate, and effective. Yet, little is known about the preconditions for their effectiveness or the ultimate consequences for democratic processes.[16]

The concern for democratic process is not inconsequential, and it is twofold. In the most general sense, the nature of these new governance structures precludes direct lines of political accountability. The decision-making process encompasses a diverse range of public and private actors along with nonelected administrators, and it supersedes jurisdictional authority. This not only threatens political accountability but also can be an impediment to collaboration to the extent that it undermines the political autonomy of participants. For example, governors are likely to be faced with choosing which to subordinate—the political will of their state or the will of the collaborative governance structure. In the case of interstate water compacts, this dilemma lies at the heart of legal conflicts. Despite commitment to interstate water agreements, legal recourse is often the only viable mechanism for conflict resolution when public officials are trapped between the terms of their compact agreements and evolving demands of their constituencies.[18] Given the long-term nature of these governance models, the challenge is to address both the issue of democratic process and strategies for maintaining broad-based public support and the political will necessary to sustain and inform them.

Opportunities and Challenges of Large-Scale Ecosystem Restoration

There are numerous challenges presented by collaborative ecosystem governance that stem from the complexity inherent both in the nature of large-scale ecosystem restoration and in the governance models themselves. Conceptualizing resource management as ecosystem management expands the range of considerations to include multiple competing uses and users, nonpoint sources of damage, nonlinear environmental reactions, and uncertain interactive effects and threshold points. The corresponding governance models are no less complex because they involve designing governing arrangements to accommodate specific environmental issues that are often framed by competing economic, political, and social contexts. As such, they necessarily overlap political jurisdictions and the functional boundaries of agencies, and they encompass multiple levels of governmental decisionmakers, regulatory bodies, and nongovernmental participants. Sustaining active cooperation, maintaining public and political support, and securing sufficient commitment resources necessary for uncertain and changing environmental problems over an indeterminate time frame are just some of the practical challenges. At the same time, embracing postsovereign governing arrangements, "rolling-rule" regulatory models, and adaptive management techniques represent paradigmatic shifts that entail philosophical challenges, which depend on more fundamental changes.[16,17,19,20] Reliance on intersovereign agreements and sovereign-dominated approaches has not been an adequate match for the scale of ecosystem problems particularly when they merely result in lowest-common-denominator agreements on uses and measures.[19] Nevertheless, expanding the range of issues addressed and the number of parties involved in governance raises questions about the role of government in facilitating effective decisions operating outside direct political control and in the "shadow of hierarchy."[16] In addition to overcoming philosophical and practical challenges, there is the additional challenge of designing the appropriate governance structure given the context and the particular ecosystem problem to be addressed.[20]

The propensity to focus on the challenges overlooks important opportunities presented by large-scale ecosystem governance. Collaborative environmental management offers a highly dynamic and participative approach to policy implementation that integrates localized knowledge and scientific learning into future decisionmaking.[2] Adaptive management approaches are more conducive to experimentation required in conditions of uncertainty, allowing for creativity, flexibility, and learning.[17] Broad-based participation improves the public and political buy-in necessary for credibility and legitimacy, the commitment and pooling of resources, and long-term collaboration.[20] Moreover, collaborative governance models offer an opportunity to address long-recognized challenges presented by the decentralization and devolution of government and related concerns about democratic control associated with the increasing discretion afforded networks of administrators and nongovernmental actors in performing functions of the state. The state has an important role to play in preventing dysfunction and facilitating these governance models by ensuring democratic processes and integrating technical and normative considerations toward effective implementation.[16] Distinct from networks of policy actors, public–private partnerships, or devolved federal responsibilities, CRG is a much narrower, issue-specific, model of democratic governance.

Collaborative Resource Governance

Collaborative resource governance can be defined as a diverse group of public and private stakeholders working together to address shared problems that extend beyond their individual capacities.[1,16] Unlike other policy configurations common in the literature, such as public–private partnerships, policy networks that seek to influence government, or networks of government actors that informally coalesce around "wicked problems," they are distinctive problem-solving, polyarchic governance models focused on complex ecological problems that overwhelm the capacity of the sovereign state.[17] Although examples of CRG vary in scale, focus, and structure, they are an increasingly important focus of research for the purpose of ascertaining factors that determine their effectiveness, the specific role of the state, and their level of success in managing large-scale ecological problems.[20,1]

The form a collaborative governance model assumes depends on environmental, political, and economic contexts. Comparative research is beginning to identify specific factors that contribute to how they function, as well as emergent concerns. For the most part, the diversity of these initiatives belies similarities among them. Some defining characteristics of CRG are its hybrid public-private structure, the scale and complexity problems they are designed to address, emphasis on locally tailored solutions, the allowance for experimentation and dynamic adaption, and the importance of information sharing and collaborative problem solving at multiple overlapping levels.[2] Success, on the other hand, is highly dependent on how they are organized, funded, and governed.[21] Integrating science into the decision-making process has also been seen as critically important, necessitating broad-based participatory processes that preclude science from being manipulated, or trumped, by the political process.[20] Process is just as important as structure for ensuring flexibility and adaptability in order to incorporate new scientific information and to make necessary course corrections. Despite the dependence on nongovernmental actors and decentralized and fragmented approaches, there is a critical role for government in providing definitional guidance, participatory incentives, and enforcement capabilities.[16] One common problem is that most initiatives severely underestimate the expense (financial, time, and personnel investment) of collaborative approaches to ecosystem management relative to traditional regulatory approaches.[20] Examples of CRG programs are indicative of specific lessons learned.

The Chesapeake Bay Program is generally acknowledged as one of the oldest and most organizationally successful of these programs despite questionable improvements in water quality.[17] It exemplifies the importance of having reliable mechanisms in place in order to facilitate participation, public outreach, and the integration of scientific information into decisionmaking.[20] Diffuse governance that is not contained by traditional political authority requires special effort not only to capitalize on the respective strengths of participants but also to cultivate legitimate and credible processes.

The California Bay-Delta Program (CALFED) has demonstrated the importance of generating broad-based scientific input, instituting internal and external peer review of scientific proposals, and devising conceptual models for communicating scientific information to decision makers and stakeholders. The CALFED has also shown the benefit of dedicating facilitators and planners to ensure vertical integration and secure long-term funding. However, problems did emerge with CALFED; for instance, several independent reports criticized its governance structure, stakeholders moved water management decisions outside the CALFED process using environmental litigation, the executive director and lead scientist resigned in 2004, and faced with mounting criticism, the CALFED Bay-Delta Authority voted to disband itself in 2005.[14]

The Comprehensive Everglades Restoration Plan (CERP) is an example of how clearly defined problems and agreement over the urgency of an environmental issue can coalesce political and financial support. However, CERP is also an example of how political tensions between state and federal levels can impede cooperation. These tensions stem in part from prematurely determining CERP's organizational structure and possible alternative solutions for Everglades restoration prior to scientific involvement, which constrained options by favoring methods preferred by the United States Army Corps of Engineers (USACE).[20] The CERP has demonstrated the benefit of adaptive monitoring in the early stages of new strategies and also, for the sake of long-term goals, the need for clarifying performance measures and indicators.

Since the 2010 oil spill, the Louisiana Coastal Area Ecosystem Restoration Program (LCA) has transitioned to the task-force model followed by CERP. However, problems with its previous model are instructive. The LCA, like CERP, experienced state-federal tensions for differing reasons but which were also exacerbated by restrictive control under the USACE. Options available to the LCA were severely constrained by political pressures by powerful stakeholders. Still, the LCA's structure and processes were not adequately designed to facilitate broad-based participation, generate public buy-in, or integrate science in ongoing decision-making.[20] The result was an excessive focus on local symptoms to the exclusion of root causes tied to resource practices in the Mississippi Ohio–Missouri river basin.

Lastly, the Glen Canyon Adaptive Management Program exemplifies the potential for adaptive management in these new governance models.[20] Adaptive management emphasizes learning, adjustment, and the acceptance of rolling rules rather than grounding in static regulatory or managerial approaches. However, there is still much misunderstanding about these approaches that requires education in order to facilitate a transition from traditional, static approaches to adaptive methods of regulation and localized policy implementation.[17] The experiences of these CRG programs are indicative of jurisdictional issues relating to ecosystem management that do not easily correspond with traditional forms governance or regulatory approaches.

Cross-Jurisdictional Issues

Jurisdictional issues are of particular importance to CRG because the geographic dimensions of ecosystem restoration defy traditional political and functional boundaries, presenting new challenges for environmental regulation and management. The geographic scale of an ecosystem means that it often will encompass some combination of municipalities, counties, states, regions, and nations.[17] Similarly, resource users and sources of environmental degradation are not contained by any particular jurisdiction or subsumed under any one political authority. Thus, problems neither fall neatly within the control of any particular authority nor the functional realm of any particular agency. Nevertheless, the magnitude of environmental problems exceeds the capacity, resources, and expertise of any one governmental entity.

Traditional regulatory methods are not equipped to deal with uncertainty, complexity, and continuous change. They are appropriate for targeting point-source problems but cannot account for the diversity presented by numerous local circumstances.[16] Legal scholars are examining the implications of CRG for the future of environmental regulation and the use of adaptive rolling rules more suitable to uncertainty and change.[16,17,19] Similarly, the decentralized and fragmented governance structure necessitated by the nature and magnitude of resource problems presents an enormous challenge for management. There are numerous public and private parties involved, and addressing problems of this scope depends on long-term commitment and public support. This requires managing collaborative activities of numerous parties from multiple jurisdictions and diverse functional backgrounds that often have competing interests. The challenge is not only to coordinate participants but also to sustain cooperation. Although CALFED demonstrated the importance of broad participation and "bottom-up" policy approaches, the problems that emerged highlighted the importance of having a clear direction and identifiable goals. Authority may be diffuse, but leadership is imperative in order to maintain momentum, negotiate compromises among competing interests, and translate and communicate across functions (e.g., political decisionmakers and scientists, scientists and the public, and across agency missions and cultures). Adaptive management is appropriate for CRG because it allows for experimentation, learning, adaptation, and course corrections.[17,20] Adaptive management represents a paradigm shift, and possibly a hurdle, for politicians, administrators, and regulators used to operating within defined jurisdictions. As researchers consider the list of factors that determine the success or failure of CRG as a general approach to large-scale ecosystem restoration, it is helpful to highlight one specific CRG structure that has been both in place for a considerable period of time and recently replicated. As such, may provide important and unique opportunities to increase understanding of CRG.

Focus: Ecosystem Restoration Task Force Model

The restoration of the Gulf of Mexico and the restoration of the Florida Everglades are two of the largest and most complex ecosystem restoration programs in the world.[22–24] The highest level of intergovernmental coordination for both programs occurs under the auspices of intergovernmental task forces created by the federal government. This model warrants closer consideration because although the two task forces differ somewhat in constitution and scope, there are enough similarities between them to make the case that they are two examples of a single intergovernmental coordination and governance structure. This "duplication" of a specific CRG structure in two large-scale ecosystem restoration programs provides an opportunity to observe the successes and challenges for this institutional design and structure over time in a way that cannot occur when focused solely on a structure that is unique to a specific ecosystem. In addition, the model warrants closer consideration because of its longevity in the case of the Everglades and, in the case of the Gulf Coast, its relevance to the most current and high profile issues in large-scale ecosystem restoration.

South Florida Ecosystem Restoration Task Force

The origins of the South Florida Ecosystem Restoration Task Force (SFERTF) can be traced to the fractured state of intergovernmental relationships that existed in the late 1980s due to contentious litigation between the federal government and the state of Florida over degraded water quality in the Everglades.[23] The litigation, which began in 1988 and continues today, produced a settlement agreement in 1992 that, for the time being, opened a window of opportunity to improve intergovernmental relations in order to begin addressing a broader range of Everglades restoration issues.[23,25] At the beginning of the Clinton administration, then secretary of the U.S. Department of the Interior, Bruce Babbitt, exploited the litigation lull and ushered in a new era of collaboration on Everglades restoration issues, which included the first incarnation of the SFERTF, created by Bruce Babbitt.[23,25]

Membership of this original SFERTF was limited to six federal agencies, but 3 years later, Congress created a new task force (also called the SFERTF) and expanded its membership to include nonfederal representatives.[23] The new Task Force was to be chaired by the Secretary of the Department of the Interior and was to include the secretaries of the Commerce, Army, Agriculture, and Transportation departments, the Administrator of the Environmental Protection Agency (EPA), and the attorney general. Congress allowed these Presidential Cabinet members to designate appointees to represent them; however, the statute requires that designees be at the assistant secretary or equivalent level of authority.[4] The SFERTF also includes representation for the state of Florida, the Miccosukee Tribe of Florida, the Seminole Tribe of Florida, the South Florida Water Management District, and local governments. In the same legislation, Congress directed the Secretary of the Army to develop a "... proposed comprehensive plan for the purpose of restoring, preserving and protecting the South Florida ecosystem" [(4) §§ 528(b)(1)(A)(i)-528(f)(3)].

Initially, the SFERTF was charged with a number of coordination and oversight responsibilities during the restoration plan development phase; however the statute did not include a sunset provision to dissolve or transform the SFERTF at the completion of that phase. On the contrary, over the decade that followed, the Congress wove SFERTF oversight into the implementation phase of restoration, as did federal regulations governing the restoration program that were developed in 2002.[26,27] The responsibilities of the SFERTF include consultation with the Secretary of the Army; coordination of restoration policy, strategy, priorities, and programs; the exchange of information among task force members; facilitation of intergovernmental dispute resolution; coordination of restoration science;

the support of implementing agencies; the coordination of financial reporting and budget requests; and reporting biennially to Congress on its own activities and on the progress of restoration efforts and results.[4]

Gulf Coast Ecosystem Restoration Task Force

In the spring of 2010, the Deepwater Horizon oil spill became one of the worst man-made environmental disasters in American history.[22] In total, an estimated 4.9 million gallons of oil was released into the Gulf of Mexico, resulting in short- and long-term environmental and economic impacts that may not be fully quantified or understood for years to come.[22] The emergency response to the Deep-water Horizon crisis required significant intergovernmental coordination to address rescue and recovery, well closure, and cleanup.[3,22] During the response period, U.S. President Barack Obama ordered the Secretary of the Navy, Ray Mabus, to develop a vision for moving from response to recovery and restoration in the Gulf of Mexico.

The so-called "Mabus Report" was released in September 2010 and took an expansive view of Gulf restoration that went beyond ecological damage caused by the oil spill and included consideration of the broad range of ecosystem challenges in the Gulf and Gulf Coast that preceded the Deepwater Horizon crisis.[22] Secretary Mabus issued a set of recommendations, which in part focused on short- and long-term intergovernmental coordination in the recovery and restoration effort. The secretary recommended that Congress establish the Gulf Coast Recovery Council to coordinate federal, state, and tribal restoration and recovery actions and to coordinate with and support activities conducted under the Natural Resources Damage Assessment (NRDA) process.[22] However, the Secretary also recommended the immediate designation of a lead federal restoration agency and the immediate creation of the Gulf Coast Restoration Task Force. The executive branch of government could carry out these recommendations without Congressional action. Mabus suggested that the task force initiate the development of a restoration and recovery strategy for the Gulf and pointed out that if Congress acted to establish the recommended council, it could subsume the task force.[22]

A month after the release of the Mabus Report, President Barack Obama created the Gulf Coast Ecosystem Restoration Task Force (GCERTF) by Executive Order.[3] The President's order states that "[t]o effectively address the damage caused by the BP Deepwater Horizon Oil Spill, address the [sic] long-standing ecological decline, and begin moving toward a more resilient Gulf Coast ecosystem, ecosystem restoration is needed" (Section 1). The Executive Order specifies federal membership of the GCERTF as including "senior officials" [Section 2 (a)(1)] of the Departments of Defense (Army Civil Works), Justice, Interior, Agriculture, Commerce, and Transportation; the EPA; the White House Offices of Management and Budget and Science and Technology Policy; and the White House Councils on Environmental Quality and Domestic Policy. The task force includes representatives of the five Gulf coast states and has the authority to add representatives of affected tribal governments.[3] The GCERTF was created as an advisory body to coordinate intergovernmental restoration efforts, support the NRDA process, present to the President a strategy for Gulf restoration, coordinate scientific research, engage stakeholders and the public, and report to the President biennially on the progress of the restoration strategy.[3]

Common Threads between the SFERTF and GCERTF

The similarities of structure and scope are evident between these two governance bodies. Every federal agency represented on the SFERTF is also represented on the GCERTF, a reflection of the significant jurisdictional and geographic overlap between the Everglades and the Gulf Coast; both include state representation, and both are responsible for strategic planning, the coordination of intergovernmental activities and science, stakeholder engagement, and biennial reporting on restoration progress.[3,4] Both task forces are administered by senior executives of the federal government who supervise full-time staff dedicated exclusively to task force administration.

More than 700 miles of Gulf coastline and 100% of the Everglades ecosystem are located in the state of Florida, and in fact, the Greater Everglades Ecosystem includes significant Gulf Coast resources such as the Caloosa-hatchee Estuary.[24,28] This shared geography not only ensures a great deal of intergovernmental cross-pollination at both the political and staff levels, but it also ensures high levels of cross-pollination among stakeholders, nongovernmental organizations, scientists, journalists, and elected officials involved in both programs. Operational or "day-to-day" collaboration, above and beyond the collaboration that flows from formal governance structures like the GCERTF and SFERTF, is an important factor in many large-scale ecosystem restoration programs.[1] Over time, the cross-pollination occurring between the Gulf Coast and Everglades restoration programs may have synergistic effects on operational collaboration for both. These two programs may help operationalize the premise that collaboration among ecosystem restoration efforts creates synergies, efficiencies, and benefits across programs, a premise central to the mission of efforts such as America's Great Waters Coalition.[29]

Conclusion

As discussed in this entry, large-scale ecosystem restoration, as an approach to solving complex, multidimensional environmental problems, presents challenges of scale, causality, and jurisdiction that can overwhelm the capacity of a single state, jurisdiction, or authority. Configurations of governmental, private sector, and nongovernmental actors have produced a variety of collaborative arrangements in response to addressing high levels of complexity. While models vary across restoration programs and tend to pursue strategies tailored to specific environmental problems, there are identifiable common features among models, such as the integration of adaptive management principles, stakeholder engagement processes, and the incorporation of science in decisionmaking. The government's comprehensive approach to ecosystem restoration in the aftermath of the Deepwater Horizon oil spill may be an indication that CRG is an approach to environmental management that is in its ascendancy. The Gulf ecosystem restoration program borrows its model of collaborative governance from a model that has been in place in the Everglades restoration program for 16 years. This is an opportunity for comparative research on collaborative resource management that can inform future applications of CRG about specific challenges and opportunities relating to large-scale ecosystem restoration. Importantly, future research must also account for several issues that are beyond the scope of this entry. For example, a nuanced understanding of the political, administrative, and logistical challenges requires an in-depth understanding of the histories and the legal challenges confronted in each of these cases, as well as assessments of the level of success in meeting respective environmental goals.

References

1. Gerlack, A.; Heikkla, T. Comparing collaborative mechanisms in large-scale ecosystem governance. Nat. Resour. J. **2006**, *46* (Summer), 657–707.
2. Karkkainen, B. Collaborative Ecosystem Governance: Scale, Complexity, and Dynamism, available at http://www.law.virginia.edu/lawweb/lawweb2.nsf/0/2ba27078dc464a84852569700060de96/$FILE/HDOCSscalecomplex.pdf.2002 (accessed September 2011).
3. Obama, B. Executive Order 13554 of October 5, 2010, Establishing the Gulf Coast Ecosystem Restoration Task Force. Federal Register. Vol. 75 (No. 195), Presidential Documents; 2010.
4. *Water Resources Development Act,* Pub. L. No. 104–303, 110 Stat. Section 528 3767–3773; 1996.
5. Agranoff, R.; M. McGuire. *Collaborative Public Management: New Strategies for Local Governments*; Georgetown University Press: Washington, DC, 2003.
6. O'Leary, R.; Bingham, L., Eds. *The Collaborative Public Manager: New Ideas for the 21st Century*; Georgetown University Press: Washington, DC, 2009.
7. Koontz, T.M.; Thomas, C.W. What do we know and need to know about the environmental outcomes of collaborative management? Pub. Admin. Rev. **2006**, *66*, 111–121.

8. John, DeWitt. *Civic Environmentalism: Alternatives to Regulation in States and Communities*; CQ Press: Washington, DC, 1994.
9. Durant, R.F.; Fiorino, D.; O'Leary, R., Eds. *Environmental Governance Reconsidered: Challenges, Choices, and Opportunities*; The MIT Press: Cambridge, MA, 2004.
10. National Academy of Public Administration. *Resolving the Paradox: EPA and the States Focus on Results*; NAPA: Washington, DC, 1997.
11. O'Leary, R.; Durant, R.F.; Fiorino, D.; Weiland, P.S. *Managing for the Environment: Understanding the Legal, Organizational, and Policy Challenges*; Jossey-Bass: San Francisco, 1999.
12. Vig, N.J.; Kraft, M., Eds. *Environmental Policy: New Directions for the 21st Century*; CQ Press: Washington, DC, 2010.
13. Koontz, T.M.; Steelman, T.A.; Carmin, J.; Korfmacher, K.S.; Mosely, C.; Thomas, C.W. *Collaborative Environmental Management: What Roles for Government?* Resources for the Future: Washington, DC, 2004.
14. Lubell, M.; Segee, B. Conflict and cooperation in natural resource management. In *Environmental Policy: New Directions for the 21st Century*; Vig, N., Kraft, M., Eds.; CQ Press: Washington, DC, 2010; 171–196.
15. Ostrom, E. *Governing the Commons: The Evolution of Institutions for Collective Action*; Cambridge University Press: New York, 1990.
16. Gunningham, N. The new collaborative environmental governance: The localization of regulation. J. LawSoc. **2009**, *36* (1), 145–166.
17. Ruhl, J. Regulation by adaptive management—Is it possible? Minn. J. Law, Sci., Tech. **2006**, *7* (1), 21–57.
18. Heikkila, T.; Schlager, E.; Davis, M. The role of cross-scale institutional linkages. In common pool resource management: Assessing interstate river mangement. Policy Stud. J. **2011**, *39* (1), 121–145.
19. Karkkainen, B. Post-sovereign environmental governance. Global Environ. Polit. **2004**, *4* (1), 72–96.
20. Van Cleve, A.; Simenstad, C.; Geotz, F.; Mumford, T. Application of the "best available science" in ecosystem restoration: Lessons learned from large-scale restoration project efforts in the USA. Tech. Rep. **2004**, *1* (May), 1–29.
21. Wiley, H.; Canty, D. *Regional Environmental Initiatives in the United States: A Report to the Puget Sound Shared Strategy*; Evergreen Funding Consultants: Seattle, Washington, 2003, available at http://www.sharedsalmonstrategy.org/files/Final_regional%20initiatives.pdf (accessed October 2011).
22. Mabus, R. America's Gulf Coast: A Long Term Recovery Plan After the Deepwater Horizon Oil Spill 2010, available at http://www.restorethegulf.gov (accessed November 2011).
23. Salt, T.; Langton, S.; Doyle, M. The Challenges of Restoring the Everglades Ecosystem. In *Large-scale Ecosystem Restoration: Five case Studies from the United States*; Doyle, M., Drew, C., Eds.; Island Press: Washington, DC, 2008.
24. Gulf Coast Ecosystem Restoration Task Force. *Gulf of Mexico Regional Ecosystem Restoration Strategy* (Preliminary); 2011.
25. Grunwald, M. *The Swamp: The Everglades, Florida and the Politics of Paradise*; Simon and Schuster:New York, 2006.
26. *Water Resources Development Act*, Public Law 106–541, 114 Stat. Section 601; 2000; 2680–2693.
27. *Programmatic Regulations for the Comprehensive Everglades Restoration Plan*, Final Rule 33 CFR Part 385; 2002.
28. South Florida Ecosystem Restoration Task Force. *Strategy and Biennial Report*; July 2008–June 2010.
29. National Wildlife Federation. America's Great Waters Coalition, available at http://www.nwf.org/Wildlife/What-We-Do/Waters/Great-Waters-Restoration/Great-Waters-Coalition.aspx (accessed November 2011).

10

Ecosystems: Planning and Trade-offs

Ioan M. Ciumasu,
Keith Culver,
Mihai Costica,
and Jean-Paul
Vanderlinden

Introduction

This chapter presents an updated and extended version of prior work, while retaining an earlier definition and overview of the notion of ecosystem planning—and the related prospect of reconciliation of disciplinary tensions and realities slowing progress in understanding of the merits of ecosystem planning.[1] This chapter extends the original discussion of natural landscapes and urban systems to the potentially positive roles of ecosystem planning in communities of actors and stakeholders dealing with the concurrent effects of urbanization and climate changes.

Humans have a long history of modifying their environment to suit their own perceived needs. Science and technology provide a certain, limited capacity to anticipate the behavior of ecosystems and plan a chosen set of desired ecosystem responses. However, because ecosystems are complex systems, i.e., governed by nonlinear behavior via self-organization and thresholds, true planning activities are bound to have a very limited relevance and effect. In various ecosystem contexts, some planning may or may not be possible—depending on the type and scale of the intended ecosystem responses. Therefore, it is important to define the scope of ecosystem planning. The term ecosystem planning is very rarely used in the literature, and this formulation might well be regarded as arriving with negative connotations likely to hamper the future use of the concept in disciplinary contexts. However, its simplicity and straightforwardness have a certain appeal for environmental managers in their search for a synthetic vocabulary to help them in the necessary integration of knowledge from many fields and disciplines. The term "ecosystem planning" may, for this reason, be usefully adopted in practice, albeit in epistemologically controversial ways. An eventual adoption of the term by the growing community of problem-solving-oriented environmental managers will require some substantial definition clarification efforts and responses to epistemological difficulties.

Context

Any observer of the use of the expression "ecosystem planning" in the extant literature will see it frequently used as shorthand for "ecosystem management planning," most notably in situations of management of environmental resources. For example, silvicultural practices often employ the term "planning" in the context of forest management, sometimes referring to "forest planning" or even "ecological forest planning."[2] Such contexts are typically described in terms of managerial objectives such as "multi-objective managerial planning," to which descriptors such as "ecological" or "ecosystem" are often applied to account for the natural processes involved.[3] In this context, planning seems to refer to procedures applied to attain some desired, pre-established objectives in terms of exploitation of resources. One can observe that, despite occasional claims of the contrary, the literature tends to recognize that planning refers to some empirical knowledge-related procedure to obtain a benefit *from* ecosystems, not a procedure to *make* ecosystems. Yet, this distinction is not always sufficiently clear, with certain confusions persisting in the authors' understanding of the fundamental differences between the aspiration to plan and the nature of ecosystems. Ecosystem planning is often mentioned in contexts that imply manipulating ecosystems to the purpose of obtaining desired outcomes. This fact, together with the lack of agreement on what ecosystem planning is supposed to designate, is a warning that more work is needed to clarify the concept and the practice. This entry takes on this objective and retains the original definition as baseline for further developments and reviews.

To various extents, properties of ecosystems may be changed. At one end of the spectrum (gradient) of human intervention in ecosystems is the domain of conservation biology, which mainly aims at protecting–preserving a given ecosystem (or components of it) as is.[4,5] At the other end of the gradient is "ecological engineering," which essentially pursues objectives related to "ecological reconstruction" or "ecosystem restoration" of heavily degraded areas or even aims at developing new ecosystems.[6] The term ecosystem planning incorporates an inbuilt contradiction between the self-organizing, nonlinear character of ecosystems[7,8] and the linear, human-organizing character of planning. Despite this, the incontestable global reality is that most terrestrial and water ecosystems are under some form of modification by humans and thus subject to management involving both natural ecosystem dynamics and planning. Even though nature conservation's original aim was the prioritized protection of pristine and near-pristine environments, and ecology seeks to understand natural processes, i.e., those undisturbed by human interventions, we have to deal every day, and in most situations, with ecosystems that are heavily influenced by human choice and action. Indeed, given the fact of global changes, we cannot just continue to talk about "untouched" or "wild" nature. The term ecosystem planning might, therefore, gain widespread use, in diverse circumstances, notwithstanding the epistemological difficulties it carries along.

Definition

In order to reconcile the nature of ecosystems with human planning reflexes and with the need for a practically useful, integrative understanding of the term "ecosystem planning," we propose a working definition: *ecosystem planning refers to the human activity of anticipating and inducing the generation of a set of ecosystem goods and services (EGS), within the limits allowed by the intrinsic dynamics of ecosystems.*

For our purposes, EGS are "the capacity of natural processes and components to provide goods and services that satisfy human needs, directly or indirectly." EGS involve the translation of ecological complexity (structures and processes) into a more restricted set of ecosystem functions that ground the provision of goods and services that are valued by humans. Admittedly, the term "ecosystem function" has been subject to conflicting interpretations, sometimes referring to internal functioning of an ecosystem (e.g., energy and matter fluxes, food chains, and food webs) and sometimes relating to the benefits derived by humans from the properties and processes of ecosystems (e.g., food production).[9,10]

Such semantic hesitations illustrate or perhaps reproduce the epistemological difficulties that tend to be encountered at the human/nature interface. Indeed ecosystem planning is implicit in the literature concerned with Coupled Human and Natural Systems (CHANS).[11] In the definition above, ecosystem planning is explicitly ascribed to, and constrained by, the limited human capacity to influence ecosystems. This attribute should be read as the crux of the concept and a key idea of this work piece.

Dynamics intrinsic to ecosystems, thresholds and nonlinearity, are explained by the laws of physics, and ecosystems can be understood as complex systems—with the support of complexity science. Ecosystem properties are subject to some limited influence, but they are not responsive to human planning imperatives. To use a somewhat less precise but more intuitive illustration, an ecosystem (as a multi-individual system, in terms of system biology) is analogous to a human body (an individual system) in that it can be trained and improved, and a certain number of its parts may be changed, but it cannot be changed as a system—any attempt to do so will cross organizational and functional thresholds that are steps towards its destruction. The system has its own character. It can be enlarged, squeezed, accelerated, or made more productive or more beautiful—but its essential characteristics as a system cannot be changed. Ecosystem planning is, therefore, like education of a human individual: it can aim at reaching the utilization of its maximum potential, but it cannot change its genetic endowment.

Term Uses

Viewed in isolation and out of context, the term ecosystem planning makes little sense. Its meaning and value become apparent only once it is considered in association with other types of planning and with attention to the concept's carrying of a specific understanding of intrinsic ecosystem dynamics.

Our working definition has several consequences and lessons. One is that the term ecosystem planning captures only partially the relation between humans and ecosystems. Employment of unduly broad or general meanings is harmful because they would allow (or inadvertently convey) the mistaken assumption that human action can compensate for ecosystem destruction by sheer planning. In reality, once destroyed (e.g., by overexploitation), the capacity of ecosystems to provide EGS (or "ecosystem carrying capacity" in the ecological economics vocabulary) cannot be simply restored by further planning. Carrying capacity is grounded in the ecosystem's intrinsic properties, and these develop during the long-term ecological history of the place. In economically equivalent terminology, the carrying capacity represents the "natural capital" (K_N), or, differently put, the ensemble of EGS extractable from a given ecosystem. Unlike the physical (K_P) and the human capital (K_H), natural capital is irreplaceable since it is dependent on ecological thresholds.

At this point, we should recall that one fundamental characteristic of biological–ecological systems dynamics is their irreversibility, meaning that once an internal threshold is crossed, an essential set of characteristics of that ecosystem are irremediably lost because passage across biological–ecological thresholds cannot be reversed. This is a consequence of the laws of thermodynamics and can be taken either as a source of ultimate ecological pessimism or simply a stark warning that the sustainability of human civilization allows no major mistake. In any case, the recognition of this reality is a foundation stone for many research threads that are bundled together in the heterodox tradition of ecological economics and the related concept of bioeconomy.[12–17] In the second half of the 19th century, Ernst Haeckel, while trying to understand the consequences of the Darwinian multifaceted struggle for survival, coined the term "ecology" to denote "the body of knowledge concerning the economy of nature." But this term only became popular in the second half of the 20th century after the introduction of the notion of "ecosystem" in 1935 by the British plant ecologist A.G. Tansley to describe the system of interactions between a biocoenosis (i.e., a community of living organisms) and their biotope (i.e., their environment), followed by the introduction of the notion of "ecosystem ecology" in 1953 by the American brother ecologists Eugene Odum and Howard T. Odum in the aftermath of the development of a slurry of mathematical models of interactions between species in the first half of the 20th century.[18–23]

Indeed, there is an even older notion which deserves renewed attention in the context of ecosystem planning and sustainability practice: that of geonomy (or geonomie), a term introduced by the Romanian biologist–ecologist and oceanographer Grigore Antipa, a student of Ernst Haeckel and the father of modern dioramas currently used in the museums of natural sciences. Antipa first used "geonomy" in 1909 to describe a system of rational management of natural resources within the watershed of Danube River and the Black Sea.[24] French juridical expert, economist, and urban planner Maurice-François Rouge in turn used the concept of geonomy to explain the organization of physical space and the structures and equilibria that determine the land uses from the triple perspective of geography, sociology, and economy.[25] This concept is known in English as "geonomics," a term which American economic geographer Ray Hugues Whitbeck used in 1926 to redefine human economy in its biophysical context of planet Earth.[26–28]

The concept of geonomics emerged initially from the holistic perspective of evolutionary ecology and then later that of economic geography. Both departure points lead to a shared insight which is now central to sustainability studies, namely that we must cease conceiving of humanity as distinct from or in opposition to nature, and instead recognize humanity as part of nature. This insight may be stated in more formal terms as the recognition that any economic system is a subsystem depending for its existence on a larger social system, in turn a subsystem of a natural (biophysical) system on which its existence depends.[29,30] The concept is roughly equivalent to—but is larger than, and includes—what is now known in the literature as the CHANS evoked in the previous subchapter. It is *prima facie* plausible to regard all of these terms as elements of the domain of sustainability studies—also known as sustainability science.[31,32] Yet when considered in light of the aim of ecosystem planning to achieve long-term sustainability, the term geonomy holds the special potential to structure a more coherent body of knowledge—i.e., more like a conception of a traditional discipline—than the idea of "sustainability studies" which stands as a broad reference to a collection of contributions to problem definition which nonetheless lack *inter alia* the internal consistency characteristic of recognized disciplines. This is a unique potential to foster systematic and comprehensive approaches to a problem domain. Use of such a unifying term would rally those ready to add to the early days of problem analysis a complementary conceptual picture of the landscape to be inhabited by an array of solutions. This is much more important than usually acknowledged in complex projects where agreements on values and knowledge are a prerequisite for project success.[33]

The three forms of capital mentioned above form together the total capital stock,[34] recognizing as a crucial constraint that natural capital cannot be replaced by the other forms of capital, and that local increases of the total capital stock available may occur while the utility of that stock decreases. Thus, when the natural capital on which human activity depends is depleted irreversibly, the value of the other two major forms of capital also decreases—they are rendered unsustainable by the destruction of the underpinning natural capital.

As argued by Costanza et al.[35] in a widely cited article, it is possible to imagine generating human welfare without natural capital/EGS in artificial "space colonies," but this possibility is still too remote and impractical for all but conceptual experimentation with the bounds of possibility. One additional, more practically relevant way to conceive the value of EGS is to determine what it would cost to replicate them in a technologically produced, artificial biosphere. Past experience with manned space missions and with Biosphere 2 (a vivarium greenhouse, i.e., a materially closed ecological system, built on 1.27 ha between 1987 and 1991) in Arizona indicates that this would be an exceedingly complex and expensive project. By contrast, Biosphere 1 (the Earth) is incomparably more efficient as least-cost provider of life support services for humans. Nevertheless, given the extent of constant human impact on biosphere, Earth's biosphere is maintained in equilibrium between its natural dynamics and the human factor (exerted through human technology), a situation which is likely to also characterize any sustainable future of humanity on Earth. To take this into account, it has been proposed that we should give great attention to the conditions required for a sustainable (or unsustainable) technobiosphere.[36] Nonetheless, greenhouse-based approaches enhanced with computer-controlled sensor systems (model laboratory terrestrial ecosystems called "ecotrons") can provide insights about impact of forcing of biogeochemical cycles (e.g., CO_2) and climate changes (e.g., temperature) on ecosystems[37–40] and thus support science-based planning of human interventions on ecosystems at planetary and local scales—aimed at mitigation of, and adaptation to, climate and other environmental changes.

Discussions on costs open, of course, the issue of potential valuation of EGS in a market. The concept of "payment for ecosystem (goods and) services" (PES/PEGS) remains controversial.[41,42] Consideration of an EGS-inclusive economy is nonetheless useful to the extent that it brings attention to the importance of incorporating into a comprehensive analysis such previously marginal concepts like "intangible values" (social and intellectual aspects—they are also connected with ecosystems) and "externalities" (ecosystem resources usually taken as "given" and not factored in by the current, neoclassical economics). In a review of this increasingly urgent matter, as current human consumption of EGS outstrips ecosystem production capacities, Baggethun et al.[42] have shown that the recent advances towards monetization of EGS indicate the slow progress of the concept. While monetization has helped draw policy and economic attention to ecosystems and EGS, it has also absorbed it into the logic of the aging yet still dominant neoclassical theory of economics in ways that deprive the concept of many of its initial virtues related to ecological and social values. Reviewers of the literature may observe that the incomplete incorporation of EGS into mainstream policy and decision-making is attributable to the prevalence of an approach that remains essentially disciplinary. EGS was repackaged or translated from ecology to current economics, along the way losing many of the virtues of EGC. Successful incorporation of EGS into policy decision-making will require a problem-driven (rather than a tool- or discipline-driven) transdisciplinary approach capable of synthesizing tools, skills, and methodologies. As their review suggests—reflecting wider opinions in the literature—valuation needs not always amount to an exclusively monetary valuation. This recommendation is naturally applicable to ecosystem planning in the sense discussed here.

Planning efficiency is nonetheless dependent on financial costs; EGS and PES represent essential advancements towards a solid grounding and meaningful use of the concept of ecosystem planning. However, as the initial concept has become diluted by disciplinary habits of subdivision and specialization, current developments and achievements are limited in their scope and practical utility. They are usefully and positively regarded as important initial forays into the conditions of sustainability.[43] On the socioeconomically transformative journey that is the transition to sustainability, ecosystem planning needs to factor in how much exploitation of EGS ecosystems can bear. In other words, how much of the carrying capacity of ecosystems can be consumed by humans without risking the collapse of the ecosystem carrying capacity. In this sense, ecosystem planning can draw support from the concept of "ecological footprint," an accounting tool for assessing the natural capital and its degree of depletion at various scales—individuals, cities, nations, or the entire planet. This approach uses estimates of consumption of natural resources for food, shelter, transportation, personal care, pollution absorption capacity, and other uses of the natural capital and compares it with the carrying capacity of ecosystems. This is one highly relevant tool for overall assessment of ecosystem planning: a value of less than 1.0 indicates that the ecological footprint remains below the carrying capacity of the ecosystems (at the scale taken into consideration), which means that ecosystem exploitation is within the limits of sustainability. On the contrary, a value greater than 1.0 indicates that a person, city, nation, or the entire humanity lives beyond the natural support capacity limits of the ecosystems/biosphere.[44–47] Values that surpass 1.0 are of utmost importance because any excess represents "eating up" the regenerative capacity of the ecosystems/biosphere, with the consequence that the overall carrying capacity decreases. At the planetary scale, it appears that humanity is already close to this tipping point or already beyond it, and the possibility of collapse is very real and a reason for accelerating and improving our efforts to understand and respond to this possibility.

The solutions for an effective transition to sustainability are still to be developed. At epistemic and moral levels, one option is to aim for a steady-state economy. Such a steady state would amount to a kind of plateau of development—as permitted by the Earth's carrying capacity—and would maintain this constant value in the future.[48] Another approach is to commit to "degrowth," meaning a collective decision of humanity to consume less, especially in those affluent countries that consume beyond the ecosystem carrying capacity available to them.[49] This idea is likely to encounter major social acceptability obstacles, certainly in developed countries, and especially in developing countries. The degrowth proposal appears to express a chosen rejection of the excessive monetization of intangible assets (social and cultural values) and ecosystem services. However, the grounds to be considered for sustainable development are first and foremost biophysical: natural limits cannot be overcome as they are thermodynamic realities already known for decades—since the origin of ecological economics.[50] The relevance of ecosystem planning relates to the immutable character of the physical laws: planning for a sustainable equilibrium of humans with ecosystems may be a matter of social decision-making and acceptance, but the consequences of those decisions are not. In other words, ecosystem planning will need to help humans accomplish sustainable development (and thus, survive). If this fails, collapse will follow. This is not science fiction: rather, it is a prediction grounded in history as civilizations have in fact overused their natural resources and faced collapse. At the planetary scale, this event is the natural outcome in the absence of a downward adjustment in the ecological footprint/carrying capacity balance.[51–53]

Taking this emergency seriously, Daily et al.[54] propose a conceptual frame for operationalizing the relation between humans and ecosystems, conceived as a never-ending circuit of five links (rendered in all caps here) and types of relations between them (rendered in italics here): ECOSYSTEMS → *biophysical models* → SERVICES → *economic and cultural models* → VALUES → *information* → INSTITUTIONS → *incentives* → DECISIONS → *actions and scenarios* → ECOSYSTEMS. Scenarios are developed for an applied case in Hawaii, called The Natural Capital Project, which tries to develop a scientific basis and connect it with policy and finance mechanisms, aiming to incorporate natural capital into resource- and land-use decisions on local and larger scales. This can be regarded as a useful example of current attempts to link ecosystem conservation with development and to render EGS concepts useful in policy and business. Yet, at the same time, we are far from the desirable situation when this would become mainstream practice. In the light of these considerations, the relevance of the term ecosystem planning remains ambiguous in the literature dedicated to the theory and applications of environmental/ecosystem management.

Mac Nally et al.[55] delimit ecosystem planning in a more cautious manner. By "ecosystem-based planning," they mean planning activities for biodiversity conservation purposes. Often, the expression "conservation planning" is used to the same effect.[5] Margules and Pressey[56] identify different stages of biodiversity conservation planning in relation to ecosystem services. Conservation biology (a term designating the aim of biodiversity and ecosystem conservation) is gradually converging with the concept of EGS.[57–59] Sometimes, an equally cautious alternative concept and expression are used, that of "ecosystem-based management," where planning activities are discussed in terms of the constraints imposed by ecosystem properties and dynamics in general and in specific contexts and case studies. In an extensive book on the matter,[60] Randolph uses terminology including "environmental protection," "land conservation," "environmental management," "environmental land use management," and "ecosystem management," as well as "land-use planning," "environmental planning," and "habitat conservation planning," to systematically describe the intricacies of human–nature interactions.

Among the rationales behind a potentially useful concept of "ecosystem planning," one can count earlier efforts to integrate environment planning and development via ecosystem approaches. Thus, the necessity to extend planning activities from human-created environment and modified environments to the natural environment arises from the expansion of human activities themselves. Therefore, the ideal situation of environmental planning would be one where it was not needed. This applies *mutatis mutandis* to ecosystem planning as well. The point we want to make here is that ecosystem planning may reveal itself to be not a goal per se, but a necessary practice towards an aspiration that is not yet concretized into a specific, readily measurable, or quantifiable goal. Implemented well, it will embody a managerial compromise between human activity expansion and the need to control, limit, and minimize the impact of human activities on the carrying capacity of biophysical systems.

One common feature is evident throughout the literature surveyed above, namely, the avoidance of a direct association between "ecosystem" and "planning." Instead, authors seem to be searching for new concepts to describe those actual situations where "ecosystem" and "planning" coincide. The authors surveyed are generally preoccupied with selection of the most apt definitions and ascription of the most appropriate meanings, hoping that the reality of human–nature interactions is being neither hastily misconceived nor ignored. For example, expressions such as "environmental planning" or "land-use planning" bear a lesser epistemological burden than "ecosystem planning." Such terms are general enough to avoid asserting an unnecessarily specific, contentious relation between planning and ecosystems, yet they clearly do include reference to sheer planning of use/occupancy of natural resources, especially in the case of water or land resources. Instead of being merely a matter of semantics, the expression ecosystem planning cannot be a useful concept unless it is associated with a clear distinction between employing knowledge about ecosystems and the effects of potential human action on ecosystems. The concept of ecosystem planning must help illuminate the border area between what humans know and can do about environment and ecosystems. Various disciplines and approaches can contribute to such a new understanding of "ecosystem planning." However, the lack of a compact body of literature makes the epistemological and technical reviewing process a highly demanding endeavor—a potential subject for future examinations.

In practice, however, understanding of the relation between humans and ecosystems may not always precede norms and decisions, as social rules and behaviors may not wait for detailed scientific clarifications. Kagan[61] examines ecosystem planning from a legal perspective, using ecosystem planning when referring to human activities related to environmental management and decision process over land/resources use, seemingly unaware of the ecological implications of putting together the words "ecosystem" and "planning." The main point of his perspective, however, is that it reflects a situation where ecosystems are being viewed, consciously or not, as something "out there," exterior to social matters. This is not very different from the currently dominant neoclassical view in economics, according to which natural resources and social matters are external to the economic processes and associated accounts, in the sense that they are "externalities"—their values or influences need not be accounted for in setting and operating a core process towards a desired outcome. All those narrow perspectives, however, are now being challenged by the widening community of practitioners for sustainable development. Nicholas Stern, the author of the homonymous report on the cost of climate changes due to greenhouse gas emissions, has arguably provided the best synthesis of the man–nature relations in a conference at the Royal Economic Society in 2007: "Climate change is a result of the greatest market failure the world has seen. The evidence on the seriousness of the risks from inaction or delayed action is now overwhelming. We risk damage on a scale larger than the two world wars of the last century. The problem is global and the response must be collaboration on a global scale." The more urgent it becomes, the great the need to reassemble disciplinary perspectives into a common understanding of the relations between natural, social, political, and economic realms.

At the borders between legislation, natural sciences, economy, and social sciences, one often-used terminology is "forest planning," meaning, in silvicultural practices, planning for forest ecosystem exploitation. The literature on forest management is among the earliest to acknowledge the tension between the deterministic approach of forest exploitation planning and the uncertainties intrinsic to ecosystem dynamics. Given the importance of forests as resource generators for human populations, uncertainty was "accommodated" in forest exploitation planning[62] in terms of managerial approaches that accounted for disturbances and other unpredictable, less deterministic phenomena. In this sense, forests represent a good case study for the relation between the management of EGS, ecosystems, and socioeconomic systems, where natural resource management planning is often seen as akin if not identical to "ecosystem planning." An extensive overview on the matter is beyond the scope of this entry. Some examples, however, may be illustrative. Exploitation of ecosystems is obviously related to the property bearing lands and waters. In Central and Eastern Europe, for example, Romania has served as a natural experiment in this regard. The recovering of the individual property rights to forests (lost at the time of military imposition of communism by the Soviet Union, when 23% of the country's forests were in private hands) in a context of a still-recovering economy and society has led to massive forest overexploitation. Large areas of forests were clear-cut in the 1990s for the immediate purpose of selling the wood (120,000 ha, almost half of the first wave of forest restitutions). Obviously, firm property rights are necessary for effective management of forests, because they provide marketability, i.e., options for future planning conservation of EGS and various uses—timber and other biomass, recreation, carbon sequestration, water retention and decontamination, and so forth. These property rights are therefore not in question, but their consequences are, in this perilous situation where the postcommunist socioeconomic transition has left a management and regulatory vacuum. Fortunately, the last decades have seen a gradual socioeconomic recovery of the country and improvements in the definition of the property rights, giving rise to remarkable improvements in forest management. Entrepreneurial activities are now related to uses of forests, based on the (1) right to access; (2) (resource) withdrawal right; (3) management right; (4) exclusion right; and (5) alienation right.[63] Along the way, a process of reforestation is taking place in the entire eastern half of the European Union. The region's forests are now recovering from some of the losses suffered during the previous years of harsh transition, with various countries at various levels of reforestation—now matching the levels of socioeconomic convergence with the more developed western European countries.[64] In such situations, ecosystem planning bears the full weight of socioeconomic contexts.

A related example that is relevant to ecosystem planning is linked to the ecological reconstruction that may follow a cycle of ecosystem overexploitation and partial ecosystem recovery. Reforestation may be difficult on highly degraded and low-accessible lands, which may become prone to soil erosion. However, knowledge of natural processes may provide the means to plan for revegetation of degraded slopes. To remain within our example from Central Europe, past experiences show how the natural vegetation succession can be used as a tool. Thus, field experiments with planting a certain shrub like the common sea-buckthorn (*Hippophae rhamnoides*) on degraded slopes in Romania have led to fast revegetation with the shrub as quasi-exclusive species. In later stages, this monospecific vegetation allows an evolution of the structure of the soil and the advent of other species (grasses and tree seedlings) to grow under its shade. In a next stage, oak (*Quercus*) and beech (*Fagus*) tree juveniles, as well as other shrubs, have grown taller and shaded the sea-buckthorn and gradually replaced it as new forest, which, even if it is characteristic of the area, would not have grown there because the land was too degraded and the conditions were suboptimal prior to the establishment of the sea-buckthorn in the first place.[65] Application of such lessons from the field, however, is only possible under favorable socioeconomic conditions. Recent advances in the country and the region may reopen the possibility of using older field ecological experiments. In such situations, ecosystem planning means planning for a desired favoring of certain natural processes, with the purpose of halting the loss of natural capital and the recovery of certain ecosystem functions and services.

A fresh question now arises: to what extent can we take advantage of our knowledge to manipulate natural processes towards recovery, albeit partial, of some initial ecosystem structures and functions? The discipline of ecological engineering is one way to explore answers to this question. Its original principles have been updated to reflect the improving understanding of the limited planning powers of humans upon ecosystems.[6,66] (1) It is based on ecosystems' self-organization capacity; (2) it can be a field test of ecological theory; (3) it relies on integrated system approaches; (4) it conserves nonrenewable energy; and (5) it supports biological conservation. Such developments seem to indicate broad acceptance that planning, and for that matter "ecosystem planning," must incorporate the idea of adaptive management, both at the scales of landscapes and those of local settlements and despite epistemological uncertainties.[67,68]

Landscape Planning

The first and most obvious application of a future-oriented, broadly accepted concept of ecosystem planning is in the area of landscape planning. However, many authors in this area may understand ecosystem planning as "the process of land-use decision-making that considers organisms and processes that characterize the ecosystem as a whole." In other words, ecosystem planning is used as shorthand for ecosystem-based planning of land uses across landscapes. Much of the literature deploys a concept that may have a certain utility in its own right, albeit narrow, yet it is often used in epistemically irregular and unstudied constructions. Musacchio et al.[69] use the expression "landscape ecological planning (LEP)." There are major challenges in integrating the concept of EGS in landscape planning and, for that matter, also in the concept of ecosystem planning as defined in this entry, simply because there is as yet no coherent and integrated approach for practical applications of EGS in planning, management, and decision-making in general.[53,70] Even in the hypothetical case where the literature will start displaying decreasing pluralism (which is not necessarily desirable) and increasing use of a shared concept (which is in principle desirable)—even a cluster concept admitting of many conceptions—it remains practically the case that there are no developments in the literature associated with substantive changes in practice. This stands in interesting contrast to other experiments regarding the place of individual human flourishing in the context of larger systems, such as the "living wage" movement's experiment in Finland with a minimum guaranteed income.[71,72] Humans are unquestionably engaged in large-scale experiments regarding new ways of living together in conditions distributing available resources with attention to the conditions of human dignity; yet those experiments are not substantially engaging the scope and scale of the sustainability challenge—which warrants special attention from researchers and managers.

Recent research programs have tried to tackle this obstacle. For example, De Groot et al.[73] have proposed a framework that involves a chain of five major links: (1) understanding and quantifying how ecosystems provide services; (2) valuing ecosystem services; (3) use of ecosystem services in trade-off analysis and decision-making; (4) use of ecosystem services in planning and management; and (5) financing sustainable use of ecosystem services. Inevitably, such a succession of actions bears the intellectual virtue of a coherent synthesis of what we can imagine to be eventually done for successfully approaching the desired equilibrium between humans and nature. But it also carries a certain naivety with respect to what may be possible—for example, it sees ecosystems as "still poorly understood" and proposes more detailed quantification as the best way to achieve a better grasp of how ecosystems function and provide EGS, as if it is only a matter of time until we get it and equally a matter which must be resolved more fully than at present if meaningful downstream action is to be taken. This way of approaching the issue, however, seems to ignore (1) the serious epistemic hurdles involved in seeking total, comprehensive understanding complex systems, especially with a reductionist mindset and (2) the practical feasibility of such an endeavor in terms of available time and resources. To give only one example, tens of thousands of chemicals exist on the market,[74] and thousands of new chemicals are introduced to the market (and released into the broader environment) every year, and there is no chance that ecotoxicological test batteries could ever be developed for each of them to "better grasp" how ecosystems function, not to mention the inherently limited relevance of each bioassay. In general, it is thought that different patterns of urbanization have a direct impact on ecosystem dynamics through feedback mechanisms and complex interactions between urban activities, land/water uses, and environmental changes.[75] Some integrative, multitier studies can be done to account for changes in ecosystem parameters in space and time, for some specific indicators, for example, water quality along a river basin, upstream and downstream from a city that is a source of water pollution.[76–78] Notably, urbanization facilitates the transition of rivers from transporters to transformers of carbon.[79]

But even those are very far from being sufficient for understanding the state of ecosystems and the quality status of EGS. What seems to be missing in the five-link type of framework proposed in the analysis cited above is the acknowledgment of the facts that we only have a limited amount of time and material resources to be allotted to advancing towards the equilibrium between humanity and the ecosystems in which it is embedded, and that those resources are themselves dependent upon our mankind–ecosystems relation that we want to address. We simply cannot plan everything we would like to. As such, the mentioned paper takes a tool-based approach rather than a problem-based approach, being concerned more with understanding than understanding sufficiently for remedial action. Apart from this difficulty, and to be fair to the mentioned authors, the framework provided by De Groot et al.[73] does appear to us to have an essentially "color-test" value for the concept of ecosystem planning proposed hereby. Under the definition of ecosystem planning employed in their paper, the chain of actions proposed by the abovementioned team at Wageningen University can actually be understood as applied to "ecosystem planning," as a useful organizer of what needs to be done by humans within the limits of the possible. Without the definition above, the same chain of actions proposed by De Groot et al. allows "everything and nothing" in a progressively increasing demand for finite. Thus, mapping landscape functions for planning the management of EGS is potentially a smart and justifiable approach, on the condition that it is integrated in a management process that leads to less destruction of EGS and, thus, interrupts the series of blind actions upon ecosystems.

A necessary feature of "ecosystem planning," as a synthetic concept, must be simplicity in the sense of operational readiness. Proposed methods should maintain sensitivity to the complexity of the phenomenon but avoid creating supplementary complications, since those would exacerbate the problem by postponing practical action. Ecosystem planning should recognize limits of human knowledge and actions and seek genuinely integrative solutions, i.e., simple yet responding to a complexity of pressures. Humanity has not managed to integrate the fragmented knowledge we have produced so far into a coherent whole, and the tension between humanity and ecosystem continues to increase. We cannot realistically hope to solve this problem by producing more piles of disparate (and mutually untranslatable disciplinary) knowledge. Ecosystem planning, if effective, will need to consist in solutions that are relevant (adaptable or instructive) to situations as various as floodplains, polders, warm and cold deserts, tropical and boreal forests, marine and fluvial ecosystems, and so forth.

Although the literature explicitly using the concept of ecosystem planning in a substantial way is scarce, there is a large and eclectic body of knowledge on implicit ecosystem planning attempts. In a study on the potential overlap between EGS and conservation policies in the coast eco-region of central California, Chan et al.[58] have extracted a set of key insights into the convergence of conservation and EGS planning: (1) both suitability of sites and demand from citizens are main drivers in what ecosystem planning can do for protecting biodiversity and EGS—near cities demand is higher but suitability is lower; (2) spatial scales may vary for optimum of biodiversity conservation or/and different types of EGS—pollination services result from a variety of small locations, while water quality from entire river basins; (3) population centers yield tensions in planning between estimated value and demand—for example, a low-value/high-demand situation can occur near cities; (4) data are usually lacking, and thorough research and analysis are necessary; (5) it must involve multidisciplinary and transdisciplinary teams, for the integration of theoretical and empirical expertise from diverse fields; (6) efficiency is conditioned by considering both trade-offs and side effects of biodiversity conservation and EGS conservation—often trade-offs are the most efficient way, and often side effects reveal points of common relevance for various goals of planning.

While the concept of EGS appears to prioritize or privilege focus on human–ecosystem relations, challenges for future applications involving the concept of ecosystem planning can be related (without being equivalent) to situations as various as land-use planning; degraded landscape management, saltmarshes and river regulation, flood prevention and management of river catchments, management of freshwater quality, silviculture (with related concepts like "ecosystem stewardship planning," "forest stewardship plans," and "forest stewardship"), and planning for carbon sequestration in terrestrial and water ecosystems. In the current context of increasing need for mitigation and adaptation to climate change, forms of ecosystem planning have been imagined at planetary scale. Thus, geo-engineering approaches propose that the planetary ocean should undergo fertilization with limiting substances (e.g., iron ions) to allow faster carbon sequestration in phytoplankton biomass. This, however, involves large-scale, unpredictable risks that require thorough investigation in a race against time.[80]

Urban Planning

In spite of some controversy, the concept of ecosystem is appropriate in the study of cities.[81] Rees and Wackernagel[44] have proposed estimations of the ecological footprint of cities and show why cities cannot be sustainable per se, even while they are actually central to sustainability. Cities can only be considered sustainable together with their hinterland, from where they draw their resources and on which they depend. In the current era of globalization, cities have become part of a planetary network of cities, with a "common hinterland," which is the Earth's biosphere. Cities play preeminent roles in the global exchanges of information and matter energy and have an immediate impact on, and are immediately affected by, the entire biosphere. Ecological footprint estimates allow cities to track the necessary resources and the plan for development according to this. Consequently, urban development planning for sustainability will require evaluations of the ecological demand of a given city (disaggregated by type of activities) and comparison with available ecosystem resources. In the same logic, ecological deficits may determine the way the city may develop, if it intends to remain sustainable.[46] Urban ecosystem services can be identified and described in urban areas just as as in nonurban areas.[82] When proposing an integrated planning tool for sustainable cities, Rotmans et al. implicitly refer to "ecosystem planning." In their paper, the concept of environmental capital seems to be equated with natural capital, as part of a wider planning for sustainable city.[83]

Niemelä describes urban ecology in the context of the relation between ecology and urban planning, where urban planning is a type of land-use planning.[84] He makes the essential observation that urban nature has been regarded in prior studies as a true field experiment about human impact on ecosystems. This resonates with the wider preoccupations with the place of urban areas within the landscapes, for example, with the degrees of vegetation cover[85] or habitat patch corridors to avoid excessive habitat fragmentation in highly urbanized areas and to allow landscape connectivity between local plant/animal populations as part of the wider metapopulations (network of local populations) within landscapes. In a previous communication concerned with the place of cities as a human social construct within local ecosystems, we have described the city in terms of an ecosystem disturbance.[86] Applying current ecological theory—the intermediate disturbance hypothesis (IDH) and island biogeography theory (IB)—we have identified an ecological taxonomy of cities and city areas as a function of disturbance intensity in a local biogeographical context: small city/city periphery (low disturbance intensity), medium-sized city/city near-center (intermediate disturbance intensity), and large city/city center (high disturbance intensity). Within this framework, any urban unit can be described on a gradient of ecological disturbance and in the local biogeographical context, with biodiversity theories and indices serving as proxies for the state and dynamics of ecosystems and ecosystems–humans dynamics.[87,88]

An abundant urban biodiversity literature has emerged providing the means for use of urban ecology-biodiversity studies in urban planning and for planning for biodiversity conservation in urban areas.[89–92] Further literature points to the transformative impact of urbanization, particularly the reduction of native habitat and the creation of new habitats, and ecosystem homogenization.[93,94] These complex aspects must be accounted for by urban/ecosystem planning. In effect, city planning emerges as a form of planning the impact (types and intensity of disturbance) of humans upon ecosystems. The management of urban areas per se and the management of city hinterlands become a matter of planning for extraction of EGS for use in cities and human settlements in general. The place of the city within the surrounding ecosystems is, therefore, a matter of "ecosystem planning."

At the level of immediate impacts, urban planning may require intra-urban or peri-urban planning, which may involve a diversity of terrestrial/water ecosystems, coastal ecosystems, and inclusion of the effects of climate changes, grazing, and agricultural land. City planners must face the concomitant emergence of new knowledge domains and a plethora of new challenges. The central feature of urban ecology is that “cities are emergent phenomena of local-scale, dynamic interactions among socioeconomic and biophysical forces” that give rise “to a distinctive ecology and to distinctive forcing functions.”[95]

The role of a city as disturbance of the local ecosystems is the result of a complex dynamic involving major concerns as disparate as urban energy systems[96] and global change,[97] solid waste and wastewater,[98,99] and infrastructure reliability engineering,[100] to cite only a few. All this requires effective community coordination, based on social multi-criteria conception and evaluation of urban sustainability policies.[101–103] In this context, the use of some multiple criteria decision aid (MCDA) methodologies seems to be unavoidable for any urban and associated ecosystem planning, and effective coordination is yet another step further. The best candidate to providing an organic integration of all these issues and perspective is the systemic view of sustainability, according to which any economic system is a subsystem of a social system, itself a subsystem of a biophysical system.[29,30] However, translation from such ecological–economic insight to management strategies, actually tried and revised, seems to be waiting.

Elsewhere, starting from the systemic conception of sustainability, we have proposed an integrative approach for problem structuring and identifying sustainable/unsustainable management scenarios for solid wastes.[104] The method consists in translating the nested inclusion relationship between economy, society, and ecosystems into an ordered set of sustainability filters to be respected in managerial practices and proposed solutions *in illo ordine:* ecological sustainability filter (EcSF), social sustainability filter (SoSF), and economic sustainability filter (EnSF). Successfully passing all filters (i.e., meeting threshold values and indicators negotiated/listed in the technical definition of each filter) indicates that a given policy or solution proposed for the stated purpose (e.g., solving the problem of an old landfill) is consistent with sustainable development. This would be considered a sustainable scenario of action, under the conditions and knowledge available at the given place and time. Under this framework, the role of ecosystem planning can be recognized at the priority level of ecosystem carrying capacity (as reflected in EcSF) prior to action being taken towards further exploitation of EGS. Alternative methods can be identified to assess sustainability of solutions proposed for urban landfills, like the so-called sustainability potential analysis (SPA), and extension of the bioecological potential analysis (BEPA) that was originally developed for local and regional landscape analyses,[105] a reminder that any urban planning is part of planning for an equilibrium between humans and the natural support systems.

Trade-offs

One idea unites all discussions about ecosystem planning, namely the need for trade-offs. For example, in the case of both landscapes and waterscapes, river drainage basins (catchment areas) are considered necessary because they represent the minimal unit by which natural forces (starting with gravity) unite to shape the surface of the planet as understood in geological and ecological terms. The European Union’s Directive 2000/60/EC, the so-called Water Framework Directive (WFD) legally requires each EU member state to have river basin management plans as a common EU-wide tool for implementing a drainage basin-based approach to integrated water resources management. Although this approach has both advantages and disadvantages, it is considered a best trade-off between the need for locally optimized performance of technical approaches to natural resource management and the need for EU-wide comparisons, which requires a common understanding and coordination, which in turn requires a commonly agreed set of procedures between over 300 methods for monitoring water quality across Europe.[106,107] Creating a common methodological body required an intercalibration exercise (which is an example of agreement on knowledge) at the scale of the European Union. So planning across large areas requires coordination which requires agreements on values and knowledge. While the WFD is an example of agreement on values achieved at political level, the intercalibration exercise is an example of agreement on knowledge achieved at the level of experts in water physics, chemistry, and biology but directed towards managerial objectives aiming at obtaining a set of EGS bundled in terms of “ecologically clean water status.”

In the case of human communities, cities can be regarded as the best candidate scale and unit for modeling and testing states and processes in the transition to sustainability, because cities present the best trade-off between system-level relevance and managerial feasibility. The city is large enough to incorporate all fundamental functions that exist in a human system, yet small enough to still allow an intelligent agent to have a holistic overview so that action can be coordinated. Lower scales, for example, city districts, are too small because they are too specialized to be representative for an entire local community (e.g., as a residential area or industrial area or commercial area). Higher scales, for example, regional or national scales, are typically too large and administered through administrative divisions like specialized departments (ministries, agencies, and offices). Systemic changes necessary in the transition to sustainability may be subject to experimentation at smaller scales, and it will be cheaper and easier than at whole-city scales, but the results will not be relevant for how whole human systems work. At scales larger then cities, experimentation is not feasible without major disruptions from which the communities may be unable to recover.

In terms of ecosystem planning, the various instances of urban green and blue infrastructure (green areas and surface water bodies) represent planning trade-offs between the natural dynamics of ecosystems and the local needs of cities, notably mitigation of urban heat island effects and flood risks.[108,109] Thus, planting trees (to create a canopy cover and dissipate heat through biological processes) can help limit the heat island effect during the day, but other factors become important for temperature relief during the night (which is important for health, as human body needs to recover during the night): reducing soil sealing (impervious surfaces).[110] At the same time, tree canopy per se and that of particular species may be part of the local culture, which adds a social component that must be taken into account in urban ecosystem planning; failing to do so can result in strong rejection, civil disobedience, and reversal of the measures. This is what happened in 2013 in the city of Iasi, Romania, where a number of lime trees (*Tilia* sp., a species common in temperate climates of Europe and with strong cultural resonance in the city and in the country) from the city's historical center have been fell for sanitary-security (related to the old age of the trees) and esthetic reasons (related to the desire to clear up facades of historic buildings and respecting the initial plans of architects), but this managerial measure was executed in a manner that was perceived as an unacceptable bureaucratic imposition—namely, without public consultation. In addition, the lime trees were replaced with Japanese pagoda tree (*Sophora japonica*)—a world-popular tree for urban forests, one that is resistant to urban pollution and heat island effect and fits with the good practice of high diversity of genera and species of urban trees[111] but which was not part of the city's *spiritus loci* (spirit of the place). Less than 2 years later, following public revolt against what was perceived as a savage massacre against a local living emblem, and after an online local referendum which showed an overwhelming support of 94% for lime trees and in spite of the large sunken cost involved, *Sophora* was dug out and replanted elsewhere in the city while the city center was replanted with lime trees as predominant species.

Such situations show that a complex balance between factors must be reached through adaptive planning of land uses and vegetation structure in each sector of the city, within the limits of what is known at a certain moment in time and, most importantly, within the local cultural frame. Often this requires combining historic and democratic traditions, certain species bearing strong symbolism codifying common cultural references and social values that reflect the local history and details of the relations between people and nature in certain biographic provinces and cultural regions of the world.

Perspectives

It is not clear yet whether the term ecosystem planning will acquire a widely recognized meaning in its own right. Nevertheless, the concept often appears to function as a convenient stylistic contraction employed in the description of complex, multilevel, multidiscipline-fed issues and topics related to environmental management. Should the term enter mainstream terminology, it will need to overcome some major epistemic difficulties presently overlooked while the term is used as an aspirational gesture unconstrained by specific operational meaning and associated specific performance measures. A variety of methodologies are expected to be developed in the future, and further redefinitions of the term are likely. No matter how methodological reflections and terminological development occur, however, it remains the case that all reflection and working terms must face the fundamental tension between human-centered planning and human-independent characteristics of ecosystems. Ecosystems are an autonomous part of the environment, with their own, conspicuously nonlinear, dynamics. Planning regarding these dynamics cannot be made fully operational, if at all, by managerial task subdelegation let alone semantic reduction. The fact that planning may work effectively in certain sectors of environmental management does not translate into effective planning at the level of ecosystems. Rather, it typically means that (1) the concept of environmental planning has been limited to engineering- and management-tractable issues; (2) it has benefitted a certain capacity of environment as a whole (i.e., ecosystem included) to absorb disturbances by human activity; and (3) environmental planning needs further refining to clarify the limitations of planning and account for the reality of nonlinear behavior of ecosystems.

In the current context of vast and fast environmental changes, which are calling for coherent responses from science and management, certain knowledge domains may assimilate the concept of ecosystem planning. Notably, conservation biology aims at more effective nature protection, going well beyond the traditional focus on biodiversity conservation, with planning emerging as a necessary component of any comprehensive approach (for a review and discussion on the matter, see Reyers et al.[112]). The imperative of a global transition to sustainable development commands a profound transformation of the economy and society, which entails effective knowledge integration and management. A recent global survey that included experts in both social and natural sciences and that was carried out by the International Council for Science and the International Social Science Council has revealed a set of five grand challenges that need to be tackled within the coming years[113] and which can be summarized in four points as follows: (1) higher management-relevant capacity to anticipate environmental changes; (2) coordinated observation of environmental changes; (3) comprehension and anticipation of disruptive environmental changes; and (4) social–institutional transformation for the transition to sustainable development. As discussed here, ecosystem planning is a focal concept for current knowledge integration and management and, despite some challenging epistemic questions, has the potential to amalgamate into a core vocabulary of sustainability. Among the multiple approaches and contributions towards sustainability, landscape and urban planning emerge as particularly effective platforms for intellectual developments involving ecosystem planning.

References

1. Ciumasu, I.M.; Buzdugan, L.; Stefan, N.; Culver, K. Ecosystems: Planning. In *Encyclopedia of Environmental Management*, Jorgensen, S.E. (ed.), 4 Volume Set, pp. 632–642. Boca Raton: CRC Press, 2012.
2. Zagas, T.D.; Raptis, D.I.; Zagas, D.T. Identifying and mapping the protective forests of southeast Mt. Olympus as a tool for sustainable ecological and silvicultural planning, in a multi-purpose forest management framework. *Ecol. Eng.* **2011**, *37*, 286–293.
3. Palahi, M.; Trasobares, A.; Pukkala, T.; Pascual, L. Examining alternative landscape metrics in ecological forest planning: A case for capercaillie in Catalonia. *For. Syst.* **2004**, *13* (3), 527–538.
4. Soulé, M.E. What is conservation biology? *BioScience* **1985**, *35* (11), 727–734.
5. Naidoo, R.; Balmford, A.; Ferraro, P.J.; Polasky, S.; Ricketts, T.H.; Rouget, M. Integrating economic costs into conservation planning. *Trends Ecol. Evol.* **2006**, *21* (12), 681–687.
6. Mitsch, W.J.; Jorgensen, S.E. *Ecological Engineering: An Introduction to Eco Technology*. John Wiley & Sons: New York, 1989.
7. Koch, E.W.; Barbier, E.B.; Silliman, B.R.; Reed, D.J.; Perillo, G.M.E.; Hacker, S.D.; Granek, E.F.; Primavera, J.H.; Muthiga, N.; Polasky, S.; Halpern, B.S.; Kennedy, C.J.; Kappel, C.V.; Wolanski, E. Non-linearity in ecosystem services: Temporal and spatial variability in coastal protection. *Front. Ecol. Environ.* **2009**, *7* (1), 29–37.

8. Walther, G.R. Community and ecosystem responses to recent climate change. *Philos. Trans. R. Soc. B* **2010**, *365* (1549), 2019–2024.
9. De Groot, R.S. *Functions of Nature: Evaluation of Nature in Environmental Planning, Management and Decision Making*. Wolters-Noordhoff: Groningen, 1992.
10. De Groot, R.S.; Wilson, M.A.; Boumans, R.M.J. A typology for the classification, description and valuation of ecosystem functions, goods and services. *Ecol. Econ.* **2002**, *41*, 393–408.
11. Kandziora, M.; Burkhard, B.; Müller, F. Interactions of ecosystem properties, ecosystem integrity and ecosystem service indicators—A theoretical matrix exercise. *Ecol. Indic.* **2013**, *28*, 54–78.
12. Georgescu-Roegen, N. Inequality, limits and growth from a bioeconomic viewpoint. *Rev. Soc. Econ.* **1977**, *35* (3), 361–375.
13. Daly, H.E. On Nicholas Georgescu-Roegen's contributions to economics: An obituary essay. *Ecol. Econ.* **1995**, *13* (3), 149–154.
14. Gowdy, J.; Mesner, S. The evolution of Georgescu-Roegen's bioeconomics. *Rev. Soc. Econ.* **1998**, *56* (2), 136–156.
15. Mayumi, K., Giampietro, M.; Gowdy, J.M. Georgescu-Roegen/daly versus solow/stiglitz revisited. *Ecol. Econ.* **1998**, *27* (2), 115–117.
16. Mayumi, K. *The Origins of Ecological Economics: The Bioeconomics of Georgescu-Roegen*. Routledge: London, 2002.
17. Kleidon, A. Life, hierarchy, and the thermodynamic machinery of planet Earth. *Phys. Life Rev.* **2010**, *7* (4), 424–460.
18. Haeckel, E. *Generelle Morphologie der Organismen. Allgemeine Grundzüge der organischen Formen-Wissenschaft, mechanisch begründet durch die von C. Darwin reformirte Descendenz-Theorie*, Berlin: G. Reimer, Vol. 2. 1966.
19. Tansley, A.G. The use and abuse of vegetational concepts and terms. *Ecology* **1935**, *16* (3), 284–307.
20. Tansley, A.G. British ecology during the past quarter-century: The plant community and the ecosystem. *J. Ecol.* **1939**, *27* (2), 513–530.
21. Odum, E.; Odum, H.T. *Fundamentals of Ecology*. XV. B. Saunders Co.: Philadelplhia, 1953.
22. Tauber, A.I. The immune system and its ecology. *Philos. Sci.* **2008**, *75* (2), 224–245.
23. Egerton, F.N. History of ecological sciences, part 47: Ernst Haeckel's ecology. *The Bull. Ecol. Soc. Am.* **2013**, *94* (3), 222–244.
24. *Encyclopédie Roumaine*, édition de 1900, t. *II*, p. 528.
25. Gohier, J. Un cours à l'EPHE, la géonomie de Maurice François Rouge. *Ann. Rech. Urb.* **1988**, *37* (1), 94–97.
26. Whitbeck, R.H. A science of geonomics. *Ann. Assoc. Am. Geogr.* **1926**, *16* (3), 117–123.
27. Finch, V.C. Ray Hughes Whitbeck 1871–1939. *J. Geogr.* **1939**, *38* (6), 252–253.
28. Floyd, B. On reversing the image of a simplistic geography: The task of socio-economic geography (geonomics). *J. Geogr.* **1971** *70* (2), 84–90.
29. Giddings, B.; Hopwood, B.; O'Brien, G. Environment, economy and society: Fitting them together into sustainable development. *Sustain. Dev.* **2002**, *10* (4), 187–196.
30. Gowdy, J.; Erickson, J.D. The approach of ecological economics. *Camb. J. Econ.* **2005**, *29*, 207–222.
31. Brandt, P.; Ernst, A.; Gralla, F.; Luederitz, C.; Lang, D.J.; Newig, J.; Reinert, F., Abson, D.J.; Von Wehrden, H. A review of transdisciplinary research in sustainability science. *Ecol. Econ.* **2013**, *92*, 1–15.
32. Lam, J.C., Walker, R.M.; Hills, P. Interdisciplinarity in sustainability studies: A review. *Sustain. Dev.* **2014**, *22* (3), 158–176.
33. Castle, D.; Culver, K. Getting to 'no': The method of contested exchange. *Sci. Public Policy* **2013**, *40* (1), 34–42.
34. Barbier, E.B. *Natural Resources and Economic Development*. Cambridge University Press: Cambridge, 2005.

35. Costanza, R.; D'Arge, R.; De Groot, R.; Farber, S.; Grasso, M.; Hannon, B.; Limburg, K.; Naeem, S.; O'Neill, R.V.; Paruelo, J.; Raskin, R.G.; Sutton, P.; Van Den Belt, M. The value of the world's ecosystem services and natural capital. *Nature* **1997**, 387 (6630), 253–260.
36. Turner, D.P. Global vegetation monitoring: Toward a sustainable technobiosphere. *Front. Ecol. Environ.* **2011**, *9* (2), 111–116.
37. Lawton, J.H.; Naeem, S.; Woodfin, R.M.; Brown, V.K.; Gange, A.; Godfray, H.J.C.; Heads, P.A.; Lawler, S.; Magda, D.; Thomas, C.D.; Thompson, L.J. The Ecotron: A controlled environmental facility for the investigation of population and ecosystem processes. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* **1993**, *341* (1296), 181–194.
38. Lawton, J.H. The Ecotron facility at Silwood Park: The value of "big bottle" experiments. *Ecology* **1996**, *77* (3), 665–669.
39. Jones, T.H.; Thompson, L.J.; Lawton, J.H.; Bezemer, T.M.; Bardgett, R.D.; Blackburn, T.M.; Bruce, K.D.; Cannon, P.F.; Hall, G.S.; Hartley, S.E.; Howson, G. Impacts of rising atmospheric carbon dioxide on model terrestrial ecosystems. *Science* **1998**, *280* (5362), 441–443.
40. Lange, M.; Eisenhauer, N.; Sierra, C.A.; Bessler, H.; Engels, C.; Griffiths, R.I.; Mellado-Vázquez, P.G.; Malik, A.A.; Roy, J.; Scheu, S.; Steinbeiss, S. Plant diversity increases soil microbial activity and soil carbon storage. *Nat. Commun.* **2015**, *6*, 6707.
41. Redford, K.H.; Adams, W.M. Payment for ecosystem services and the challenge of saving nature. *Conserv. Biol.* **2009**, *23* (4), 785–787.
42. Gómez-Baggethun, E.; De Groot, R.; Lomas, P.L.; Montes, C. The history of ecosystem services in economic theory and practice: From early notions to markets and payment schemes. *Ecol. Econ.* **2010**, *69*, 1209–1218.
43. Daily, G.C.; Matson, P.A. Ecosystem services: From theory to implementation. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105* (28), 9455–9456.
44. Rees, W.; Wackernagel, M. Urban ecological footprints: Why cities cannot be sustainable—and why they are a key to sustainability. *Environ. Impact Assess. Rev.* **1996**, *16* (4–6), 223–248.
45. Wackernagel, M.; Onisto, L.; Bello, P.; Callejas Linares, A.; Lopez Falfan, I.S.; Mendez Garcia, J.; Suarez Guerrero, A.I. National natural capital accounting with the ecological footprint concept. *Ecol. Econ.* **1999**, *29*, 375–390.
46. Wackernagel, M.; Kitzes, J.; Moran, D.; Goldfinger, S.; Thomas, M. The ecological footprint of cities and regions: Comparing resource availability with resource demand. *Environ. Urbanization* **2006**, *18* (1), 103–112.
47. Goldfinger, S.; Wackernagel, M.; Galli, A.; Lazarus, E.; Lin, D. Footprint facts and fallacies: A response to Giampietro and Saltelli (2014) "Footprints to Nowhere". *Ecol. Indic.* **2014**, *46*, 622–632.
48. Czech, B.; Daily, H.E. The steady state economy: What it is, entails, and connotes. *Wildl. Soc. Bull.* **2004**, *32* (2), 598–605.
49. Martínez-Alier, J.; Pascual, U.; Vivien, F.D.; Zaccai, E. Sustainable de-growth: Mapping the context, criticisms and future prospects of an emergent paradigm. *Ecol. Econ.* **2010**, *69*, 1741–1747.
50. Georgescu-Roegen, N. The steady state and ecological salvation: A thermodynamic analysis. *BioScience* **1977**, *27* (4), 266–270.
51. Liu, J.; Diamond, J. China's environment in a globalizing world. *Nature* **2005**, *435*, 1179–1186.
52. Costanza, R.; Graumlich, L.; Steffen, W.; Crumley, C.; Dearing, J.; Hibbard, K.; Leemans, R.; Redman, C.; Schimel, D. Sustainability or collapse: What can we learn from integrating the history of humans and the rest of nature? *Ambio* **2007**, *36* (7), 522–527.
53. Costanza, R.; de Groot, R.; Braat, L.; Kubiszewski, I.; Fioramonti, L.; Sutton, P.; Farber, S.; Grasso, M. Twenty years of ecosystem services: How far have we come and how far do we still need to go? *Ecosys. Serv.* **2017**, *28*, 1.
54. Daily, G.C.; Polasky S.; Goldstein, J.; Kareiva, P.M.; Mooney, H.A.; Pejchar, L.; Ricketts, T.H.; Salzman, J.; Shallenberger, R. Ecosystem services in decision making: Time to deliver. *Front. Ecol. Environ.* **2009**, *7*, 21–28.

55. Mac Nally, R.; Bennett, A.F.; Brown, G.W.; Lumsden, L.F.; Yen, A.; Hinkley, S.; Lillywhite, P.; Ward, D. How well do ecosystem-based planning units represent different components of biodiversity? *Ecol. Appl.* **2002**, *12* (3), 900–912.
56. Margules, C.R.; Pressey, R.L. Systematic conservation planning. *Nature* **2000**, *405*, 243–253.
57. Kremen, C. Managing ecosystem services: What do we need to know about their ecology? *Ecol. Lett.* **2005**, *8*, 468–479.
58. Chan, K.M.A.; Shaw, M.R.; Cameron, D.R.; Underwood, E.C.; Daily, G.C. Conservation planning for ecosystem services. *PLoS Biol.* **2006**, *4* (6), 2138–2152.
59. Egoh, B.; Rouget, M.; Reyers, B.; Knight, A.T.; Cowling, R.M.; Van Jaarsveld, A.S.; Weltz, A. Integrating ecosystem services into conservation assessments: A review. *Ecol. Econ.* **2007**, *63*, 714–721.
60. Randolph, J. *Environmental Land Use Planning and Management.* Island Press: Washington, DC, 2004.
61. Kagan, R.A. Political and legal obstacles to collaborative ecosystem planning. *Ecol. Law Q.* **1997**, *24*, 871–876.
62. McCarthy, M.A.; Burgman, M.A. Coping with uncertainty in forest wildlife planning. *For. Ecol. Manage.* **1995**, *74*, 23–36.
63. Nichiforel, L.; Schenz, H. Property rights distribution and entrepreneurial rent-seeking in Romanian forestry: A perspective of private forest owners. *Eur. J. For. Res.* **2011**, *130* (3), 369–381.
64. Taff, G.N.; Müller, D.; Kuemmerle, T.; Ozdeneral, E.; Walsh, S.J. Reforestation in Central and Eastern Europe after the breakdown of socialism. *Reforesting Landscapes* **2010**, *10,* 121–147.
65. Stefan, N. Contributii la studiul sindinamicii asociatiei Hyppophaetum *rhamnoides* in Subcarpatii de Curbura. *Academia Româna—Memoriile Sectiilor Stiintifice* **1991**, seria IV, tom *XIV* (1), 223–233.
66. Mitsch, W.J.; Jorgensen, S.E. Ecological engineering: A field whose time has come. *Ecol. Eng.* **2003**, *20,* 363–377.
67. Lessard, G. An adaptive approach to planning and decisionmaking. *Landscape Urban Plann.* **1998**, *40,* 81–87.
68. Williams, B.K. Adaptive management of natural resources—Framework and issues. *J. Environ. Manage.* **2011**, *92* (5), 1346–1353.
69. Musacchio, L.R.; Coulson, R.N.; Robert, N. Landscape ecological planning process for wetland, waterfowl, and farmland conservation. *Landscape Urban Plann.* **2001**, *56*, 125–147.
70. ICSU, UNESCO, UNU. Ecosystem change and human wellbeing. Research and Monitoring Priorities Based on the Findings of the Millennium Ecosystem Assessment, 2008, Paris, International Council for Science, ISBN 978-0-930357-67-2, available at http://www.icsu.org/icsu-asia/news-centre/news/archive-2006-2010/ICSUUNESCO-UNU_Ecosystem_Report.pdf (accessed December 12, 2011).
71. Koistinen, P.; Perkiö, J. Good and bad times of social innovations: The case of universal basic income in Finland. *Basic Income Stud.* **2014**, *9* (1–2), 25–57.
72. Marchal, S.;Van Mechelen, N. A new kid in town? Active inclusion elements in European minimum income schemes. *Soc. Policy Adm.* **2017**, *51* (1), 171–194.
73. De Groot, R.S.; Alkemade, R.; Braat, L.; Hein, L.; Willemen, L. Challenges in integrating the concept of ecosystem services and values in landscape planning, management and decision making. *Ecol. Complexity* **2010**, *7,* 260–272.
74. Gustavsson, M.B.; Hellohf, A.; Backhaus, T. Evaluating the environmental hazard of industrial chemicals from data collected during the REACH registration process. *Sci. Total Environ.* **2017**, *586*, 658–665.
75. Alberti, M. Maintaining ecological integrity and sustaining ecosystem function in urban areas. *Curr. Opin. Environ. Sustain.* **2010**, *2* (3), 178–184.
76. De Pauw, N.; Heylen, S. Biotic index for sediment quality assessment of watercourses in Flanders, Belgium. *Aquat. Ecol.* **2001**, *35*, 121–133.

77. Neamtu, M.; Ciumasu, I.M.; Costica, N.; Costica, M.; Bobu, M.; Nicoara, M.N.; Catrinescu, C.; Becker van Slooten, K.; De Alencastro, L.F. Chemical, biological, and ecotoxicological assessment of pesticides and persistent organic pollutants in the Bahlui River, Romania. *Environ. Sci. Pollut. Res.* **2009**, *16*, S76–S85.
78. Von der Ohe, PC.; De Deckere, E.; Prüß, A.; Muñoz, I.; Wolfram, G.; Villagrasa, M.; Ginebreda, A.; Hein, M.; Brack, W. Toward an integrated assessment of the ecological and chemical status of European river basins. *Integr. Environ. Assess. Manage.* **2009**, *5*, 50–61.
79. Smith, R.M.; Kaushal, S.S. Carbon cycle of an urban watershed: Exports, sources, and metabolism. *Biogeochemistry* **2015**, *126* (1–2), 173–195.
80. Ciumasu, I.M.; Costica, M.; Secu, C.V.; Gurjar, B.R.; Ojha, C.S.P. Adapting to climate change. In *Greenhouse Gas Emissions and Climate Change*; Surampalli, R.Y., Zhang, T., Ojha, C.S.P., Gurjar, B.R., Eds. ASCE Press – American Society of Civil Engineers: Reston, Virginia, 2013.
81. Pickett, S.T.; Grove, J.M. Urban ecosystems: What would Tansley do? *Urban Ecosyst.* **2009**, *12* (1), 1–8.
82. Bolund, P.; Hunhammar, S. Ecosystem services in urban areas. *Ecol. Econ.* **1999**, *29*, 293–301.
83. Rotmans, J.; Van Asselt, M.; Vellinga, P. An integrated planning tool for sustainable cities. *Environ. Impact Assess. Rev.* **2000**, *20*, 265–276.
84. Niemelä, J. Ecology and urban planning. *Biodiversity Conserv.* **1999**, *8*, 119–131.
85. Pauleit, S.; Duhme, F. Assessing the environmental performance of land cover types for urban planning. *Landscape Urban Plann.* **2000**, *52*, 1–20.
86. Ciumasu, I.M.; Culver, K. The city as disturbance of the local ecosystems. ISEE2011—Advancing Ecological Economics—Theory and Practice. *The 9th International Conference of the European Society of Ecological Economics*, Istanbul, June 14–17, 2011.
87. Marzluff, J.M. Island biogeography for an urbanizing world: How extinction and colonization may determine biological diversity in human-dominated landscapes. *Urban Ecosyst.* **2005**, *8*, 157–177.
88. Kattel, G.R.; Elkadi, H.; Meikle, H. Developing a complementary framework for urban ecology. *Urban For. Urban Gree.* **2013**, *12* (4), 498–508.
89. Wittig, R. The origin and development of the urban flora of Central Europe. *Urban Ecosyst.* **2004**, *7*, 323–339.
90. Dearborn, D.C.; Kark, S. Motivations for conserving urban biodiversity. *Conserv. Biol.* **2010**, *24* (2), 432–440.
91. Tzoulas, K.; James, P. Making biodiversity measures accessible to non-specialists: An innovative method for rapid assessment of urban biodiversity. *Urban Ecosyst.* **2010**, *13,* 113–127.
92. MacGregor-Fors, I.; Morales-Perez, L.; Schondube, J.E. Does size really matter? Species–area relationships in human settlements. *Diversity Distrib.* **2011**, *17,* 112–121.
93. McKinney, M.L. Urbanization as a major cause of biotic homogenization. *Biol. Conserv.* **2006**, *126,* 247–260.
94. Kühn, I.; Klotz, S. Urbanization and homogenization—Comparing the floras of urban and rural areas in Germany. *Biol. Conserv.* **2006**, *127* (3), 292–300.
95. Alberti, M.; Marzluff, J.M.; Shulenberger, E.; Bradley, G.; Ryan, C.; Zumbrunnen, C. Integrating humans into ecology: Opportunities and challenges for studying urban ecosystems. *BioScience* **2003**, *53* (12), 1169–1179.
96. Manfred, M.; Caputo, P.; Costa, G. Paradigm shift in urban energy systems through distributed generation: Method and models. *Appl. Energy* **2011**, *88*, 1032–1048.
97. Grimm, N.B.; Faeth, S.H.; Golubiewski, N.E.; Redman, C.L.; Wu, J.; Bai, X.; Briggs, J.M. Global change and the ecology of cities. *Science* **2008**, *319* (5864), 756–760.
98. Dyson, B.; Chang, N.B. Forecasting municipal solid waste generation in a fast-growing urban region with system dynamics modeling. *Waste Manage.* **2005**, *25*, 669–679.
99. Muga, H.E.; Mihelcic, J.R. Sustainability of wastewater treatment technologies. *J. Environ. Manage.* **2008**, *88*, 437–447.

100. Zio, E. Reliability engineering: Old problems and new challenges. *Reliability Eng. Syst. Saf.* **2009**, *94,* 125–141.
101. Lahdelma, R.; Saminen, P.; Hokkanen, J. Using multicriteria methods in environmental planning and management. *Environ. Manage.* **2000**, *26* (6), 595–605.
102. Munda, G. Social multi-criteria evaluation for urban sustainability policies. *Land Use Policy* **2006**, *23*, 86–94.
103. Ling, C.; Hanna, K.; Dale, A. A template for integrated community sustainability planning. *Environ. Manage.* **2009**, *44*, 228–242.
104. Ciumasu, I.M.; Costica, M.; Costica, N.; Neamtu, M.; Dirtu, A.C.; De Alencastro, L.P.; Buzdugan, L.; Andriesa, R.; Iconomu, L.; Stratu, A.; Popovici, O.A.; Secu, C.V.; Paveliuc- Olariu, C.; Dunca, S.; Stefan, M.; Lupu, A.; Stingaciu-Basu, A. ; Netedu, A.; Dimitriu, R.I.; Gavrilovici, O.; Talmaciu, M.; Borza, M. Complex risks from old urban waste landfills—A sustainability perspective from Iasi, Romania. *J. Hazard., Toxic Radioact. Waste* **2012**, *16* (2), 158–168.
105. Lang, D.J.; Scholz, R.W.; Binder, C.R.; Wiek, A.; Stäubli, B. Sustainability Potential Analysis (SPA) of landfills—A systemic approach: Theoretical considerations. *J. Cleaner Prod.* **2007**, *17,* 1628–1638.
106. Birk, S.; Bonne, W.; Borja, A.; Brucet, S.; Courrat, A.; Poikane, S.; Solimini, A.; Van De Bund, W.; Zampoukas, N.; Hering, D. Three hundred ways to assess Europe's surface waters: An almost complete overview of biological methods to implement the Water Framework Directive. *Ecol. Indic.* **2012**, *18*, 31–41.
107. Poikane, S.; Zampoukas, N.; Borja, A.; Davies, S.P.; van de Bund, W.; Birk, S. Intercalibration of aquatic ecological assessment methods in the European Union: Lessons learned and way forward. *Environ. Sci. Policy* **2014**, *44*, 237–246.
108. Demuzere, M.; Orru, K.; Heidrich, O.; Olazabal, E.; Geneletti, D.; Orru, H.; Bhave, A.G.; Mittal, N.; Feliu, E.; Faehnle, M. Mitigating and adapting to climate change: Multi-functional and multi-scale assessment of green urban infrastructure. *J. Environ. Manage.* **2014**, *146*, 107–115.
109. Žuvela-Aloise, M.; Koch, R.; Buchholz, S.; Früh, B. Modelling the potential of green and blue infrastructure to reduce urban heat load in the city of Vienna. *Clim. Change* **2016**, *135* (3–4), 425–438.
110. Ziter, C.D.; Pedersen, E.J.; Kucharik, C.J.; Turner, M.G. Scale-dependent interactions between tree canopy cover and impervious surfaces reduce daytime urban heat during summer. *Proc. Natl. Acad. Sci.* **2019**, *116* (15), 7575–7580.
111. Santamour Jr, F.S. Trees for urban planting: Diversity uniformity, and common sense. In *The Overstory Book: Cultivating Connections with Trees*; C. Elevitch, Ed., pp. 396–399. Permanent Agriculture Resources, 2004.
112. Reyers, B.; Roux, D.J.; Cowling, R.M.; Ginsburg, A.E.; Nel, J.L.; O'Farrell, P. Conservation planning as a transdisciplinary process. *Conserv. Biol.* **2010**, *24* (4), 957–965.
113. Reid, W.V.; Chen, D.; Goldfarb, L.; Hackmann, H.; Lee, Y.T.; Mokhele, K.; Ostrom, E.; Raivio, K.; Rockström, J.; Schellnhuber, H.J.; Whyte, A. Environment and development. Earth system science for global sustainability: Grand challenges. *Science* **2010**, *330* (6006), 916–917.

11

Natural Enemies: Conservation

Cetin Sengonca

Introduction

After the successful utilization of two methods of biological control, classical biological control (importation and establishment of exotic natural enemies against either exotic or native pests) and augmentation of natural enemies (either inundative or inoculative releases of mass reared natural enemies), the third method—conservation and enhancement—has become more and more important during recent years.[1,2] Conservation of natural enemies is probably the most important concept in the practice of biological control and, fortunately, is one of the easiest to understand and readily available to growers. Most authors consider conservation as an environmental modification to protect and enhance natural enemies.[3] This definition will be the main subject of this entry.

Natural enemies of arthropod pests, also known as biological control agents, include predators and parasitoids that occur in all production systems from commercial fields to backyards where they have adapted to the local environment and target pests. Their conservation is generally noncomplicated and cost-effective.[4] With relatively little effort the activities of these natural enemies can be observed. Natural control agents are a major factor in controlling agricultural pests and need to be considered when making pest management decisions. Today, therefore, the conservation of natural enemies is considered inseparable from enhancement and together they represent a successful biological control method.

Avoid Harmful Practices of Pesticides

Pesticides used in agriculture are not only killing target pests but they also can have direct effects on natural enemies by killing them, or indirectly by eliminating their hosts or preys and causing them to starve. In contrast, the conservation concept of natural enemies attempts to avoid the application of particularly broad-spectrum, highly disruptive pesticides. Applying selective or specific and beneficially safe pesticides may contribute much toward preserving natural enemies. Pesticide selectivity to beneficial arthropods has been broadly classified into two forms. The first of these is physiological selectivity, that is, pesticides are less toxic to natural enemies than to their target pest when applied at the recommended rate. The second form is ecological selectivity that pertains to the means and domains in which pesticides are used. Systemic pesticides killing leaf-feeding herbivores, for example, may have little or no effect on the many natural enemies that have contact only with the leaf surface. In some cases, pesticides can be successfully integrated into pest management systems with little or no detrimental effect on natural enemies, and this trend is likely to increase substantially in the future. Nowadays, the pesticide industry places increasing emphasis on the development of beneficially safe and environmentally friendly pesticides that exhibit greater selectivity for natural enemies and have minimal environmental impacts. In the same way, governmental regulatory agencies increasingly consider the adverse effects of pesticides on natural enemies in their registration process for pesticides, reflecting the growing concern over negative effects on beneficial insects. Despite these important steps and the great progress that has been made, the latent effects of pesticides and the impact of "cocktail applications" on natural enemy populations are still not fully understood. Further research and implementation of research results are urgently needed.

Alternatively, when selective pesticides are unavailable, recommended conservation tactics usually involve exact timing of pesticide applications. Careful forecasting and observation of the occurrence and growth of pest populations can substantially reduce the number of pesticide applications. Forecasting systems should be based on a defined economic threshold for each pest, considering also the presence of natural enemies.[5] Another approach is the selective placement of pesticides in agricultural fields. Limiting pesticide application only to infested parts of the field will reduce costs and also conserve natural enemies. An ideal alternative approach is the use of microbial insecticides, such as commercially available *Bacillus thuringi- ensis* and fungal and viral products that have little adverse effects on natural enemies and the environment.

Habitat and Environmental Manipulation

Another form of natural enemy conservation is habitat and environmental manipulation. The agricultural landscape is currently so intensively managed that the species diversity of many natural habitats has disappeared or become endangered. A similar reduction can be observed among natural enemies too. It has been strongly suggested by many experts that natural enemies also can be conserved by simply encouraging vegetational diversity of the agroecosystem. In this context, hedgerows, cover crops, strips inside and bordering fields, and even in-field balks provide important refuges for parasitoids and predators of many pest species. There they find and benefit from safe shelters, sources of pollen and nectar, and also alternative prey or hosts in case of food scarcity in cultivated fields. And thus, at such sites, a long-lasting and self-regulating biocoenosis will develop. A higher acceptance of weeds in agricultural crops may also increase the efficacy of natural enemies. Similarly, mixed plantings, for example, Umbelliferae, mustard, and *Phacelia tanacetifolia*,[6] growing weed strips even within fields, and providing flowering field borders significantly increase habitat diversity. *P. tanacetifolia* has been cultivated widely between crops in the production system in Germany for more than a decade. At the same time this will provide shelters and alternative food sources for natural enemies. Another important concept in conservation by habitat manipulation is that of connectivity. Natural or less disturbed habitats are often scattered and isolated within the agricultural landscape. Connecting these habitats, for example, by hedgerows and woods, will establish a continuous network of corridors allowing movement of natural enemies between fields.

Experiments have shown that a constant population of natural enemies can be established and conserved by releasing their prey or hosts during periods of scarcity, for example, releasing the red mite, *Tetranychus urticae* Koch, to support the establishment of its predatory mite, *Phytoseiulus persimilis* Athias-Henriot, in cucumber and bean cultures widely in greenhouses in middle Europe.[7] Similarly, distributing sterilized *Eupoecelia ambiguella* Hb. eggs between the two generations of this lepidopteran pest preserved its egg parasitoid *Trichogramma semblidis* (Auriv.) in the Ahr valley in Germany. A classical example is the black scale, *Saissetia oleae* (Oliv.), which interrupts its development for a short period during summer in the hot arid areas of central California. Planting irrigated oleander plants adjacent to citrus orchards allowed the black scale a continuous development. As a result, this enabled its specific parasitoid *Metaphycus helvolus* (Comp.) to maintain its population, particularly during the hot summer months.[7] In another example, it has been found that preservation of nettles, *Urtica* spp., an important host plant of *Aglais urticae* L., can enhance the efficacy of *T. semblidis* on the second generation of *E. ambiguella.* The reason is that the parasitoid maintains its population on this alternative host during the two nonoverlapping generations of *E. ambiguella.*[8]

The manipulation of some simple cultural measures can also conserve natural enemies. A famous example is strip harvesting hay alfalfa, allowing mobile natural enemies to disperse from cut strips to half-grown strips. Similarly, *Trissolcus vasilievi* (Mayr) and *T. semistriatus* Nees, two important egg-parasitoids of the sunn pest, *Eurygaster integriceps* Put., were successfully conserved in Turkey by growing shade trees in hot arid areas, providing shade and thus suitable climatic conditions for these parasitoids.

Overwintering and Shelter Sites

Natural enemies build up considerably high population densities during summer periods, but then suffer from lack of overwintering or shelter sites and unfavorable climatic conditions during winter. As a result of these detrimental environmental factors, extreme low entomophagous arthropod densities are often present the following year. This permits pest populations to explode and in consequence requires more pesticide applications. In contrast, by preserving existing or providing artificial hibernation sites or shelters, natural enemies can be conserved during overwintering periods. For example, planting of trees and perennial bunch grasses near agricultural sites, the use of burlap or cloth trees, and wrap and stones provided as hiding places at overwintering time allow coccinellid lady beetles higher survival rates during hibernation. In the same way “trunk traps” and “trap bands” can be used as artificial overwintering sites for predatory bugs and lace-wings,[9] and felt belts can be wrapped around the trunks of fruit trees and vines for the predatory mite *Typhlodromus pyri Scheuten*. The green lacewing, *Chrysoperla carnea* (Stephens), overwinters as an adult in barns, roof trusses, houses, and under the bark of trees, where mortality rates during hibernation in middle European climatic conditions may still reach 60%–90%. By using specially designed, simple wooden shelters (hibernation boxes) the overwintering mortality was reduced to only 4%–8%.[10] On this ground, these hibernation boxes are now being accepted and commonly used by farmers, gardeners, and also environmental protectionists in Germany and Switzerland.

Future Concerns

The majority of pest problems in agriculture are due to the elimination of natural enemies by the indiscriminate and intensive use of pesticides. Improper habitat manipulation and mismanagement of ecosystems have further intensified this problem by reducing the available flora and fauna. Conservation and enhancement of natural enemies is the easiest and least costly method of biological control offering solutions to most pest problems without harming and disturbing the natural ecosystem. Unfortunately, however, this field has received little attention and very little investment has been made in research. There is a serious need for research into the areas of direct conservation and enhancement of natural enemies during the vegetation period and hibernation. In a self-regulating mechanism focusing on conservation and enhancement, natural enemies can keep agricultural pests below their economic threshold and help to reduce the number and frequency of pesticide applications. Furthermore, integration of conservation and enhancement of natural enemies into existing IPM programs[11] will lead to a more sustainable and cost-effective agriculture system.

See also Cosmetic Standards, pages 152–154; Conservation of Biological Controls, pages 138–140; Augmentative Controls, pages 36–38; Biological Controls, pages 57–60; pages 61–63; pages 64–67; pages 68–70; pages 71–73; pages 74–76; pages 77–80; pages 81–84.

References

1. Ehler, L.E. Conservation Biological Control: Past, Present and Future. In *Conservation Biological Contro*l; Barbosa, P., Ed.; Academic Press: New York, 1998; 1–8.
2. Bugg, R.L.; Pickett, C.H. Introduction: Enhancing Biological Control–Habitat Management to Promote Natural Enemies of Agricultural Pests. In *Enhancing Biological Contro*l; Pickett, C.H., Bugg, R.L., Eds.; University of California Press: Berkeley, 1998; 1–23.
3. DeBach, P. *Biological Control of Insect Pests and Weeds*; Chapman and Hall: New York, 1964; 844.
4. Barbosa, P. Agroecosystems and Conservation Biological Control. In *Conservation Biological Control*; Barbosa, P., Ed.; Academic Press: New York, 1998; 39–59.
5. Sengonca, C. Conservation and enhancement of natural enemies in biological control. Phytoparasitica **1998**, *26* (3), 187–190.
6. Sengonca, C.; Frings, B. Einfluss von phacelia tanacetifolia auf schaedlings- und nuetzlingspopulation in Zuckerruebe. Pedobiologia **1988**, *32* (5/6), 311–316.
7. Krieg, A.; Franz, J.M. *Lehrbuch der biologischen Schaedlingsbekaempfung*; Verlag Paul Parey: Berlin, 1989; 302.
8. Schade, M.; Sengonca, C. Foerderung des traubenwickler- eiparasitoiden Trichogramma semblidis (Auriv.) (Hym., Trichogrammatidae) durch bereitstellung von ersatzwirten an brennesseln im weingebiet ahrtal. Vitic. Enol. Sci. **1998**, *53* (4), 157–161.
9. Beane, K.A.; Bugg, R.L. Natural and Artificial Shelter to Enhance Arthropod Biological Control Agents. In *Enhancing Biological Control*; Pickett, C.H., Bugg, R.L., Eds.; University of California Press: Berkeley, 1998; 240–253.
10. Sengonca, C.; Frings, B. Enhancement of green lacewing Chrysoperla carnea (Stephens) by providing artificial facilities for hibernation. Turk. Entomol. Derg. **1989**, *13* (4), 245–250.
11. *CRC Handbook of Pest Management in Agricultur*e; Pimentel, D., Ed.; CRC Press: Boca Raton, FL, 1991; *2*, 757.

12

Pests: Landscape Patterns

F. Craig Stevenson

Pest and Landscape Patterns

Small Patches

Certain areas within a field may be more conducive to pest survival, establishment, and development relative to other areas. The size and proximity of localized infestations with a given field differs among the three major pest groups. Past research has shown that weed and disease patches generally have a radius of about 25 m (if you assume that they have a circular shape), but may be as large as 100 m.[1–3] In the case of wild oat *(Avena fatua* L.), the patch may be associated areas that have higher soil water and nutrient availability (see Figure 1). Weeds such as perennial sowthistle (*Sonchus arvensis* L.), with seeds adapted to wind dissemination, may occur in areas of higher elevation where seeds are trapped as they are blown across a field. Foliar disease patches generally are associated with low wet areas of fields or sheltered field margins that have higher canopy humidity and lower wind speeds. Landscape patterns for insects, however, are less persistent and harder to predict because of rapid reproductive rates and pest movement.[4] To a lesser extent, this same level of complexity also occurs for polycyclic diseases that easily become airborne. However, the overwintering phases in northern agricultural regions and

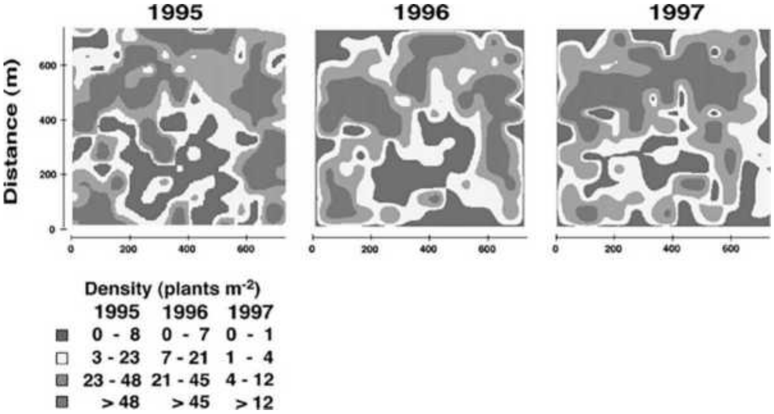


FIGURE 1 Landscape pattern of wild oat in a 65 ha field.
Source: Thomas.[6]

nonflying phases of insects may be associated with areas of a field with a more friable soil structure (e.g., grasshoppers) or other soil and microclimatic factors. In addition, for insects and to a lesser extent diseases, landscape patterns are complicated pest–predator interactions that can affect infestation of that pest in the given space in the next growing season. For example, the landscape pattern of nematodes has been shown to be partly a function of nematophagous fungi.[5] Therefore, the proximity of areas more or less conducive for the pest is ultimately dependent on the topography of the field and spatial variation of microclimatic conditions, especially for those pests with restricted mobility.

Larger Scale Landscape Patterns

These patterns are seen where larger areas of a field, or numerous fields, are being infested by a pest. These types of landscape patterns generally are dependent on external factors that move the pest from a smaller patch to adjacent areas. Climatic processes, such as wind, that blow weed seeds, disease spores, and insects can rapidly increase the area of pest infestation. Farming practices such as tillage and harvest operations also cause larger scale landscape patterns. For example, a study showed that *Polygonum* spp. of weeds tended to vary with the direction of tramlines up to distances of 635 m, a distance about 600 m more than that observed in the direction perpendicular to tramlines.[1] The spatial distribution of pests can be influenced by tillage practices and harvest operations that move crop residues and soil across a field. For example, nematodes are spread by cultivation from initial infestation foci.[5] Also, grain combine harvesters and shank-type tillage implements can move weed seeds and other reproductive structures across a field quickly.

Temporal Stability of Landscape Patterns

Time tends to complicate and/or obscure distinct landscape patterns for pests within a field. Varying climatic patterns (rainfall, temperature, etc.), within and among growing seasons can have a profound effect on the ability of pests to survive, establish, and develop. This complexity, interacting with polycyclic and multigeneration reproductive strategies within a growing season and factors moving pests across or between fields, poses a major hurdle to a holistic understanding of pests and landscape patterns. Extensive field research combined with predictive modeling may be a fruitful avenue for such a complex phenomenon.

Future Developments

The advent of global positioning systems (GPS), satellite imagery, and geographic information systems (GIS) software have heightened awareness and provided insight into landscape patterns and pests. Ultimately, these technologies may allow for the site-specific management of pests in accordance with their landscape patterns.[7] For example, spatially referenced maps could be linked to GPS on pesticide applicators to reduce the total amount of pesticide applied, thus improving economic returns and resulting in farming systems less dependent on pesticides. Our current understanding clearly shows that pest infestations often occur in patterns across a field, however, a great deal more effort will be necessary to provide information to accurately predict where pests will occur and affect crop production most extensively. These challenges will be especially difficult considering the highly variable and dynamic nature of current climatic conditions and farm management systems.

References

1. Nordbo, E.; Christensen, S.;Kristensen, K.;Walter, M. Patch spraying of weed in cereal crops. Asp. Appl. Biol. **1994**, *40,* 325–334.
2. Zadoks, J.C.; van den Bosch, F. On the spread of plant disease: a theory on foci. Ann. Rev. Phytopathol. **1994**, *32,* 503–521.

3. Zanin, G.; Berti, A.; Riello, L. Incorporation of weed spatial variability into the weed control decision-making process. Weed Res. **1998**, *38,* 107–118.
4. Hassell, M.P.; Comins, H.N.; May, R.M. Spatial structure and chaos in insect population dynamics. Nature **1991**, *353,* 255–258.
5. Webster, R.; Boag, B. Geostatistical analysis of cyst nematodes in soil. J. Soil Sci. **1992**, *43,* 583–595.
6. Thomas, A.G. *Agriculture and Agri-Food Canada*; Saskatoon, SK, Canada, 1999, unpublished data.
7. *The State of Site Specific Management for Agriculture*; Sadler, E.J.,Pierce,F.J.,Eds.; American Society of Agronomy, Inc., Crop Science Society of America, Inc., Soil Science Society of America, Inc.: Madison, WI, 1997; 423.

III

CSS: Case Studies of Environmental Management

13

Biological Control of Vertebrates: Myxoma Virus and Rabbit Hemorrhagic Disease Virus as Biological Controls for Rabbits

Peter Kerr and
Tanja Strive

Myxomatosis

Myxomatosis is a generalized, lethal disease of European rabbits classically characterized by gross skin swellings, sometimes termed tumors or myxomas, which appear mucinous on cut sections, mucopurulent blepharoconjunctivitis, swollen face and head, thickened, edematous, often drooping ears and grossly swollen perineum. In acute infections, death occurs 8–12 days after infection or 3–5 days after development of obvious clinical signs. The case fatality rate (CFR) can approach 100%. Myxomatosis was not a disease of European rabbits in their natural range; it is caused by myxoma virus (MYXV), which originally circulated in the South American jungle rabbit or tapeti (*Sylvilagus menensis* also called

Sylvilagus brasiliensis). In this species, MYXV causes a relatively inconsequential cutaneous fibroma at the site of inoculation. Myxomatosis is an example of an emerging disease that was only seen when a susceptible host species, the European rabbit, was imported into South America where MYXV was circulating.

Myxoma Virus

MYXV is a poxvirus (family: *Poxviridae*; subfamily: *Chordopoxvirinae*; genus: *Leporipoxvirus*; species: *Myxoma virus*). Like other chordopoxviruses, MYXV is a large brick-shaped particle (286 × 230 × 75 nm). The type virus (Lausanne strain) has a double-stranded DNA genome of 161.8 kb encoding 158 unique genes, 12 of which are duplicated in terminal inverted repeats of 11.5 kb at each end of the genome. The virus replicates in the cytoplasm of infected cells and so must encode the necessary enzymes for DNA replication and RNA transcription. Like other large DNA viruses, many genes encode virulence factors which suppress host immune responses or antiviral mechanisms (Cameron et al. 1999, Kerr 2012, Kerr et al. 2015). In the natural host, these virulence factors are probably essential for local suppression of immune responses, thus maintaining virus in the fibroma for sufficient time to allow transmission. However, in European rabbits, the combination of these immune suppressive factors and the ability of the virus to replicate in lymphocytes and monocytes of this species allow the virus to disseminate from the site of initial infection and cause generalized systemic disease with lymphoid tissue destruction and profound immunosuppression (Best et al. 2000, Best and Kerr 2000, Jeklova et al. 2008). MYXV is passively transmitted on the mouthparts of biting arthropods such as mosquitoes or fleas that probe through the virus-rich skin lesions in search of a blood meal. The virus does not replicate in the arthropod vector. In European rabbits, transmission may also occur via virus in conjunctival or nasal discharges.

The Use of Myxoma Virus as a Biological Control for Rabbits in Australia

The use of MYXV to control rabbits was suggested as early as 1919 (Fenner and Ratcliffe 1965) and some early experiments were done in Australia. However, it was not until the 1930s that systematic studies were undertaken, initially in Britain and subsequently in Australia, to examine the lethality for rabbits, the host range, and the likely impact of MYXV on other Australian species (Martin 1936, Bull and Dickinson 1937). Field studies in Australia were undertaken in 1937–1939 in enclosures on Wardang Island off the coast of South Australia and subsequently, in the summer of 1940–1941, in an enclosure in low-rainfall country on the mainland where native stickfast fleas (*Echidnophaga* spp.) were present as a vector (Bull and Mules 1944). These studies were followed by field releases in the same area. The virus had a substantial impact on local rabbit populations when fleas were present and predators such as foxes were removed to prevent them from killing sick rabbits. However, it did not spread beyond the study sites and was regarded as unlikely to control rabbits on any scale (Bull and Mules 1944).

Recognizing the need for mosquitoes as vectors, experimental studies were undertaken in 1950 in higher rainfall districts in the Murray River valley near Albury on the border between New South Wales (NSW) and Victoria and near Gunbower in north-central Victoria (Myers 1954). The virus appeared to spread initially but then died out at the release sites. However, myxomatosis reappeared in December of that year at one of the sites and simultaneously at other locations along the Murray River and at locations up to 650 km away (Ratcliffe et al. 1952, Myers 1954). Transmission was driven by mosquitoes, particularly *Culex annulirostris*, which were in high numbers due to flooding in the northern Murray–Darling river system and the vast numbers of rabbits which were completely susceptible to this novel pathogen. The scale of the epizootic was unprecedented with estimates of spread covering an area of 1600 km south to north and 1750 km east to west (Ratcliffe et al. 1952). The virus trickled on in local outbreaks through the winter and then reappeared in epizootics the following summer once mosquitoes emerged. Over the next 3 years, MYXV spread, either naturally or by deliberate introduction, wherever rabbits had established and there were enough mosquitoes or other vectors to transmit the virus.

Myxomatosis reduced rabbit numbers by perhaps 90% to 99% with CFRs due to the released virus (termed the standard laboratory strain; SLS) estimated at 99.8% (Fenner et al. 1953, Myers et al. 1954). However, in some local epizootics, the estimated CFR was as low as 40% with surviving rabbits identified by scarring and the presence of serum antibodies to MYXV (Fenner et al. 1953).

Rabbits that recover from myxomatosis are probably immune for life, with immunity potentially boosted by reexposure to the virus. In addition, recovered females will transfer maternal antibodies across the placenta to kittens, which may provide limited passive protection in the early weeks of life (Fenner and Marshall 1954, Kerr 1997).

It soon became apparent that slightly attenuated strains of MYXV were occurring in the field (Fenner et al. 1953, Myers et al. 1954, Marshall et al. 1955). While still highly lethal, with CFRs of 90%–99%, these strains had a selective advantage because they allowed infected rabbits to survive for a slightly longer time, thus providing more opportunity for mosquito transmission of the virus (Marshall et al. 1955, Fenner et al. 1956).

Field strains of virus were subsequently classified into five virulence grades (1–5) based on CFR, survival times, and clinical signs of small groups of infected laboratory rabbits (Fenner and Marshall 1957). Grade 3 viruses, with CFRs of 70%–95% and survival times of 17–28 days, became the predominant strains in the field within 4 years (Figure 1) and remained so for the next 30 years after which there are no systematic data available (Fenner and Ross 1994). However, both highly virulent and highly attenuated strains can still be recovered from the field (Kerr et al. 2017, unpublished data).

With such high CFRs, MYXV was exerting very high selection pressure on the rabbit population and the emergence of somewhat attenuated viruses probably increased the frequency with which rabbits

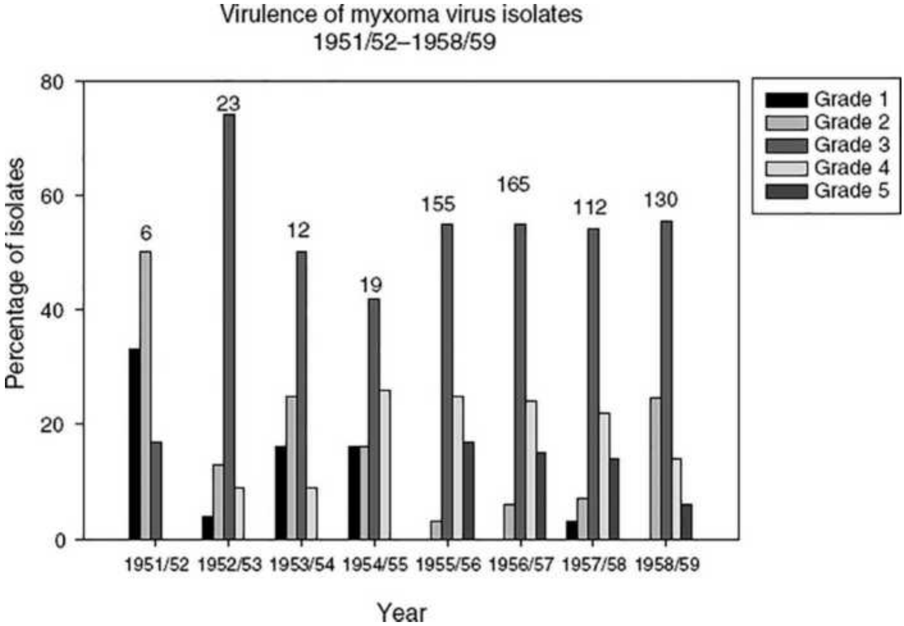


FIGURE 1 Virulence changes in Australian isolates of MYXV from 1950–1959. Virulence of MYXV field isolates is shown as the percentage of isolates made each season classified into grades 1–5. The number of isolates tested each year is shown above the bars. Data for 1951/1952–1957/1958 are from Marshall and Fenner (1960) and data for 1958/1959 are from Fenner and Woodroofe (1965). Virulence was classified as: Grade 1 average survival time (AST): ≤ 13 days; CFR: 99.5%; Grade 2 AST: > 13 ≤ 16 days; CFR: 99%; Grade 3 AST > 16 ≤ 28 days; CFR: 70%–95%; Grade 4: AST > 28 ≤ 50 days; CFR: 50%–70%; Grade 5 AST not applicable; CFR < 50%.

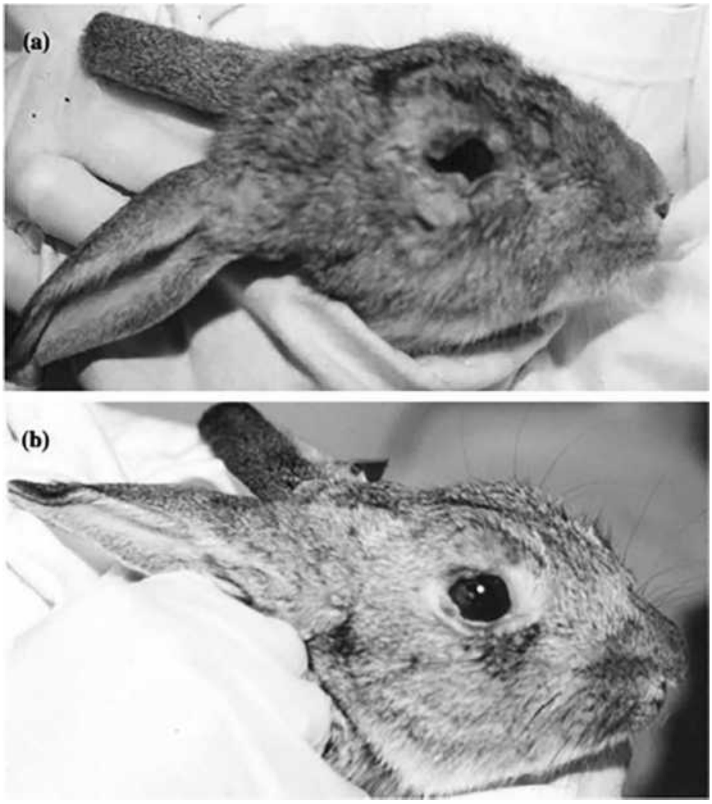


FIGURE 2 Genetic resistance to MYXV in Australian wild rabbits. (a) A wild rabbit infected 10 days previously with the grade 1 virulent SLS MYXV introduced into Australia in 1950. The rabbit shows relatively mild clinical signs of myxomatosis. (b) A wild rabbit 20 days after infection with the same virus. This rabbit has virtually fully recovered. SLS originally killed essentially 100% of infected rabbits with an average survival time of ≤13 days. (Figures are from Best SM and Kerr PJ (2000) Virology 267, 36–48, Elsevier publishing, with permission.)

with alleles conferring some resistance to myxomatosis survived infection (Figure 2) (Alves et al. 2019). In addition, high environmental temperatures enhanced survival of rabbits infected with moderately attenuated virus strains (Marshall 1959, Marshall and Douglas 1961).

A dramatic example of the selection for resistance was provided by a study on rabbits from Lake Urana in NSW. Within 7 years of natural selection by annual epizootics of myxomatosis, the CFR in these rabbits, infected under laboratory conditions with the KM13 strain of MYXV, dropped from 90% to 26% (Marshall and Fenner 1958, Marshall and Douglas 1961).

The emergence of attenuated virus strains and genetically resistant rabbits in the decade from 1950–1960 reduced the impact of myxomatosis. However, there is no doubt that the disease continued to suppress rabbit populations especially in the agricultural areas of Australia where myxomatosis, in occasional widespread epizootics and more regular local outbreaks, complemented conventional control measures. In an experiment where myxomatosis was experimentally controlled, rabbit numbers increased by 8–12 fold in a region where myxomatosis had previously been regarded as not contributing to rabbit control (Parer et al. 1985).

In the early 1990s, 40 years after the initial epizootic, it was estimated that rabbit numbers in the agricultural areas of Australia were on average fluctuating around 5% of their pre-myxomatosis numbers (Fenner and Ross 1994, Williams et al. 1995). However, in the low rainfall rangeland and semiarid areas, where conventional management was more difficult or nonexistent, populations were about 25%

of pre-1950 numbers and population irruptions were occurring (Fenner and Ross 1994, Williams et al. 1995). The rabbit breeding cycle and climatic variation means that there will be large fluctuations in population sizes particularly in the arid zones where the natural decline in rabbit populations due to drought may be as much as 99%, but it is also in these areas that rabbits have their greatest impact on preventing regeneration of vegetation (Williams et al. 1995, Cooke 2012).

Since 1995/1996, the impact of myxomatosis on rabbit numbers has been more difficult to discern because of the presence of rabbit hemorrhagic disease (RHD), which may also alter the timing of epizootics of myxomatosis by removing the seronegative young rabbits needed to sustain an outbreak (Mutze et al. 2002, 2010a). Although difficult to quantify, it is likely that myxomatosis continues to suppress rabbit numbers across Australia at essentially no economic cost.

The Introduction of Novel Insect Vectors to Australia to Enhance Biological Control by Myxomatosis

MYXV in Australia was largely vectored by mosquitoes with some local spread by direct contact and native stickfast fleas. In dry seasons or arid zones, there might be no transmission because there was little or no mosquito breeding. There were areas such as Tasmania, the northern tablelands of NSW, the western district of Victoria, or south-western West Australia where mosquito activity was low (Williams et al. 1995) with a consequent lower impact of myxomatosis.

In Britain and parts of Europe, the main vector of MYXV, following its release in 1952 (see below), was the European rabbit flea (*Spilopsyllus cuniculi*), however, this parasite had not been introduced to Australia with the rabbit. Rabbit fleas are present on rabbits all year round and move freely between rabbits, thus potentially allowing transmission of MYXV when mosquitoes are absent or in low numbers.

S. cuniculi was imported to Australia in 1966 and released from quarantine in 1968 after rather minimal testing for environmental impact (Sobey and Menzies 1969). The fleas were widely released throughout Australia, and in many regions, this produced a drop in rabbit numbers that was comparable to the earlier impact of myxomatosis (Cooke 1983, King et al.1985, Mutze et al. 2002, Cooke et al. 2013). The overall result appears to have been an increase in the efficacy of myxomatosis despite attenuated field strains of virus and genetic resistance in rabbits. In particular, fleas could drive epizootics when mosquito activity was low. *S. cuniculi* was also introduced into rabbit populations on the subantarctic Macquarie Island and the Kerguelen archipelago.

Despite its initial establishment, *S. cuniculi* tended to die out in rabbit populations in the hot dry areas of Australia (annual rainfall <200 mm) (Cooke 1984). To overcome this limitation, the Spanish rabbit flea *Xenopsylla cunicularis*, which is adapted to hot, arid conditions, was introduced (Cooke 1990). This flea was released from quarantine in 1993 after studies on species specificity showed that native animals were not a preferred host. Large-scale breeding was undertaken and over one million pupae or newly emerged fleas were subsequently released at some 500 sites (Fenner and Fantini 1999, Saunders et al. 2010). Unfortunately, due to the emergence of RHD in 1995, virtually no follow-up studies were done on the impact of this flea on the epidemiology of myxomatosis although there were reports of unseasonable outbreaks of myxomatosis (Fenner and Fantini 1999). More recent anecdotal evidence indicates that this flea remains established in rabbit populations.

Introduction of More Virulent Strains of Myxoma Virus to Australia

From the time of the first epizootic of myxomatosis in Australia, MYXV was deliberately introduced into rabbit populations. Preparations of SLS were made available for field release, and in NSW and later in other states, a field isolate derived from SLS known as the Glenfield strain was used. The distinctive

Lausanne strain of MYXV (used for release in Europe, see below) was widely used in release campaigns in Australia from the 1970s particularly in association with *S. cuniculi* release (Fenner and Ross 1994). Lausanne and Glenfield are more virulent than SLS when tested in genetically resistant wild rabbits although the virulence of the three viruses is indistinguishable in laboratory rabbits. Releases of Lausanne continued until the late 1990s even though it was recognized that released viruses were outcompeted by field strains of virus and at best might create local epizootics (Myers et al. 1954, Marshall et al. 1955, Fenner et al. 1957, Berman et al. 2006). Virulent virus is no longer produced for release in Australia, and releases of MYXV are not occurring (rabbit hemorrhagic disease virus (RHDV) is produced for release). Proposals for new introductions of virulent strains of MYXV have been made (Henzell et al. 2008, Mutze et al. 2010a), although it is difficult to see how these would compete in the field based on previous releases of virulent virus. For example, despite its widespread release in Australia, no circulating strains of MYXV derived from the Lausanne strain have been identified (Saint et al. 2001, Kerr et al. 2019).

Myxomatosis in Europe and Other Regions

The Lausanne strain of MYXV was released in France in June 1952 by a landholder seeking to control rabbits. The virus established and spread, killing an estimated 90% of wild rabbits in France (Fenner and Ross 1994). From France, MYXV spread to the rest of Europe wherever rabbits were present in the wild and into domestic rabbitries. It was illegally introduced into Britain in 1953 and spread through the British wild rabbit population with estimates of population reduction as high as 99% (Fenner and Ross 1994). As in Australia, there was emergence of both attenuated virus strains and genetically resistant rabbits as the virus and its new host coevolved, although this seems to have happened more slowly in Britain (Fenner and Ross 1994).

In Europe, the rabbit is a keystone species in some ecosystems and a valued game animal for hunting. Collapse of rabbit populations, initially due to myxomatosis and later due to RHD, has added to pressure on endangered predators such as the Spanish imperial eagle (*Aquila adalberti*) and Iberian lynx (*Lynx pardinus*) (Rogers et al. 1994). In 2018, a widespread epizootic of myxomatosis occurred in hares (*Lepus granatensis*) in the Iberian Peninsula potentially signaling a species shift (Dalton et al. 2019).

MYXV was also introduced into Chile in 1954 (and later into Argentina) to control feral rabbit populations. It has also been released onto the subantarctic Macquarie Island and Kerguelen Archipelago.

Rabbit Hemorrhagic Disease as a Biological Control for European Rabbits

RHD was first described in 1984 as a novel disease outbreak in domestic Angora rabbits imported into China from Germany. It subsequently spread to, or emerged in, Europe, the Americas, Korea, and other parts of the world causing substantial losses in domestic and wild rabbit populations (Cooke and Fenner 2002).

In its original form, RHD typically presented as a sudden onset of high mortality in subadult and adult rabbits. Kittens less than 4–6 weeks of age were generally unaffected. In the peracute form of RHD, rabbits often died with no premonitory signs and some still had food in their mouths. More normally, the disease manifested as a short period of malaise with fever and elevated respiratory rate for up to 24 hours prior to death, which might be accompanied by ataxia, posterior paresis, convulsions and coma, hematuria, and a bloody discharge from the nostrils. Death is due to fulminant hepatitis complicated by disseminated intravascular coagulation with hemorrhages in lungs, kidneys, and other tissues. Subacutely affected rabbits that survive for a few days may show signs of jaundice.

Rabbit Hemorrhagic Disease Virus

RHD is caused by a calicivirus (Family: *Caliciviridae*; genus: *Lagovirus*; species: *Rabbit hemorrhagic disease virus*; RHDV). RHDV is a small non-enveloped virus approximately 40 nm in diameter with icosahedral symmetry. It has a single-stranded positive-sense RNA genome of 7.4 kb and replicates in the cytoplasm of infected cells; it cannot be routinely isolated in cell culture.

Two distinct genotypes of the virus now circulate: viruses derived from the original RHDV (here referred to as RHDV1) and a more recently emerged genotype, which will be referred to as RHDV2. A new nomenclature for the lagoviruses has been proposed with a single species *Lagovirus europaeus* that subsumes RHDV and the related *European brown hare syndrome virus* (EBHSV) each of which would become a genogroup. Genogroups are then further subdivided into genotypes and variants based on phylogenetic relationships. In this proposal, RHDV1 is genotype G1.1 and RHDV2 becomes genotype G1.2 (Le Pendu et al. 2017).

RHDV probably evolved from an apathogenic or mildly pathogenic rabbit calicivirus (RCV). These viruses replicate in the small intestine (Hoehn et al. 2013) and are widely distributed in domestic and wild rabbits in Europe and Australia and probably other parts of the world where farmed, feral, or wild European rabbits occur (Capucci et al. 1996, Cooke et al. 2004, Marchandeau et al. 2005, Forrester et al. 2007, Strive et al. 2009, Bergin et al. 2009, Jahnke et al. 2010, Strive et al. 2010, Marchandeau et al. 2010, Le Gall Reculé et al. 2011b, Nicholson et al. 2017). RCVs are closely related to RHDV, induce cross-reactive antibodies and, in some cases, appear to provide cross-protection to RHDV.

Following the initial spread of RHDV1, an antigenic variant termed RHDVa emerged in Europe (Capucci et al. 1998) and appeared to outcompete RHDV. The more divergent RHDV2 emerged in 2010 and appears to have evolved independently (Le Gall Reculé et al. 2011a, Silvério et al. 2018). RHDV2 overcomes the age resistance to disease in young kittens (Dalton et al. 2014, Neimanis et al. 2018) and has a broader host range spilling over from rabbits to cause disease in multiple species with some evidence of circulation in hares (Puggioni et al. 2013, Camarda et al. 2014, Hall et al. 2017, Neimanis et al. 2018). RHDV1 and a recombinant of RHDV1 have also been retrospectively demonstrated in Iberian hares with clinical signs of RHD (Lopes et al. 2014, Lopes et al. 2017).

Recombination of RHDV1, RHDV2, and low virulence RCVs is a significant feature of the ongoing evolution of these viruses (Lopes et al. 2015, Silvério et al. 2018, Hall et al. 2018).

Infection with RHDV can occur by most routes of inoculation, but infection via the oropharynx is probably most common in the field. RHDV replicates to very high titres in the liver causing massive hepatic necrosis. Virus RNA is present in liver, spleen, mesenteric lymph nodes, bile, thymus, urine, blood, and feces (Shien et al. 2000). In experimental studies, where relatively high doses of virus are used, death generally occurs within 2–3 days of infection. The CFR in susceptible rabbits following infection with RHDV1 is over 90%.

Kittens less than 4 weeks of age inoculated with RHDV1 support virus replication and can transmit virus (Lenghaus et al. 1994, Mikami et al. 1999, Matthaei et al. 2014) but generally do not develop clinical disease and are subsequently immune to reinfection. This age-related resistance is gradually lost between 4 and 9 weeks of age (Robinson et al. 2002b). However, kittens infected with RHDV2 succumb to disease with high CFRs.

The titres of RHDV in the blood are sufficiently high for fleas and mosquitoes to passively transmit the virus; it does not replicate in these insects (Lenghaus et al. 1994). However, the duration of viremia prior to death is very short so it is unlikely that biting insects play a significant role in transmission. In contrast, carrion-feeding flies play a major role in transmission by feeding on carcasses and passively shedding virus in feces and regurgita (Asgari et al. 1998, McColl et al. 2002a, Schwensow et al. 2014).

Rabbits that have recovered from RHD are probably immune for life to the infecting antigenic type with potential boosting by natural reexposure to the virus (Cooke et al. 2002). Immune females will transfer protective antibodies across the placenta to their kittens. These passive maternal antibodies can protect kittens for up to 11 weeks after birth depending on the titre of antibody in the doe. Importantly, kittens that are challenged with RHDV while passive antibody titres are waning may become infected but survive and be subsequently immune (Robinson et al. 2002b). It is uncertain how much immunological cross-protection occurs between RHDV1 and RHDV2. Anecdotally, vaccination against RHDV1 seems poorly protective against RHDV2 (Dalton et al. 2012) and limited field studies showed little cross-protection (Peacock et al. 2017). However, in a study with experimentally infected rabbits, some cross-protection appeared to occur (Calvete et al. 2018).

RHDV is relatively resistant to environmental degradation, and infective virus can persist in carcasses for at least 20 days at 22°C and perhaps as long as 3 months, although this would be affected by the climatic conditions (McColl et al. 2002b, Henning et al. 2005). Thus, RHDV has two modes of transmission: direct contact between susceptible and infected animals shedding virus and environmental contamination from carcasses and flies that feed on infected carcasses. Predators such as foxes that feed on infected rabbits or scavenge carcasses do not support virus replication but may also pass infectious RHDV in their feces and potentially spread the virus.

Introduction of RHDV to Australia as a Biological Control Agent

The potential of RHDV as a biological control for European rabbits was quickly recognized following epizootics in wild rabbits in Spain (Cooke and Fenner 2002), and in 1991, the Czech 351 strain of RHDV was imported into the high-security Australian Animal Health Laboratory at Geelong for evaluation. Studies of virulence, pathogenesis, and transmission in laboratory and wild rabbits of different ages were undertaken (Lenghaus et al. 1994). Species-specificity testing indicated that rabbits were the only susceptible species of those tested. Approval was then given for the virus to be evaluated in pens on Wardang Island 5 km off the coast of South Australia; the same location that had been used nearly 60 years earlier for trials with MYXV.

Testing on Wardang Island commenced in March 1995. Despite careful security, including insecticide treatments, RHDV escaped from the island and appeared in rabbit populations on the adjacent mainland in October 1995 (Cooke and Fenner 2002). It was probably transmitted by blowflies (*Calliphora* spp.) or bushflies (*Musca vetustissima*), which had not been previously suspected as vectors. Contingency plans to contain the virus by eradicating rabbits in the area were put into action, but it quickly became obvious that the virus was not just spreading between warrens but was making large jumps to new populations hundreds of kilometres away (Cooke and Fenner 2002). This long-distance spread, together with the escape from Wardang Island, clearly implicated flying insects in dissemination of the virus (Cooke 2002).

Following this accidental release of RHDV, regulatory inquiries, including public submissions, and further species-specificity testing suggested that the virus was unlikely to infect other species and it was registered as a pest control agent (Cooke and Fenner 2002). This allowed government agencies to produce virus for deliberate release as an inoculation for rabbits or for distribution on carrot baits.

Impact of RHDV as a Biological Control

Within 12 months of its escape, RHDV was reported in all the mainland states of Australia. As with myxomatosis 45 years earlier, the impact on rabbit populations was dramatic with some populations being reduced by 95% (Mutze et al. 1998, Saunders et al. 1999, Story et al. 2004, Bruce and Twigg 2004). This was particularly so in the low rainfall regions of Australia. In the more temperate regions, the impact was mixed. Generally, populations were considerably reduced, but at some sites, there was little or no impact following RHD (Saunders et al. 1999, Henzell et al. 2002, Bruce and Twigg 2005, Richardson et al. 2007).

In the subsequent years, there was a gradual reduction in the impact of RHDV on rabbit populations but numbers were still at lower levels than pre-RHD (Mutze et al. 2002, 2014). In southern West Australia, RHDV disappeared and was successfully reintroduced (Bruce and Twigg 2005). Reduction in rabbit populations in some arid regions was strongly associated with recovery of endangered mammals (Pedler et al. 2016). RHDV was also released on Macquarie Island in 2011 as part of a successful rabbit eradication campaign.

Emergence and Release of Novel RHDV Strains

In 2014, an exotic RHDVa strain was detected in south-eastern Australia (Mahar et al. 2018b); it is not known how this virus entered Australia. The virus is a recombinant between an RCV-like and an RHDVa virus most closely related to a virus from China. In 2015, routine surveillance detected an exotic RHDV2 strain in south-eastern Australia (Hall et al. 2015). Phylogenetic studies indicated that it had been circulating for some months prior to detection; this virus spread across the rabbit-infested areas of Australia largely replacing RHDV1 (Mahar et al. 2018a). A novel recombinant between this virus and RCV has also been reported (Hall et al. 2018). These RHDV2 strains also spill over into introduced European hares causing some lethal infections (Hall et al. 2017). Studies with the originally isolated Australian RHDV2 virus showed that it caused similar disease processes in kittens and adults, as had already been shown in Europe and, in this way, was quite distinct from RHDV1 (Neave et al. 2018, Neimanis et al. 2018). It appears to have substantially reduced rabbit populations (Mutze et al. 2018, Ramsey et al. 2020) and, as of 2019, was the dominant virus in the environment. It is possible that the ability of RHDV2 to kill very young rabbits has interfered with the spread of RHDV1 by removing a susceptible cohort from the population. Combined with the ability to overcome immunity to RHDV1, this has likely given RHDV2 a selective advantage, which may wane as the breeding female population becomes dominated by survivors of RHDV2, which will pass passive antibody protection to their offspring.

From 2010 onwards, a large research program in Australia examined other globally occurring variants of RHDV for their potential to augment the impact of circulating RHDV and slow the growth in rabbit populations. An RHDVa strain from Korea was selected because it overcame partial cross-protection conferred by recent infection with benign RCV (see below) and had a better ability to infect genetically resistant wild rabbits than other strains. This virus (termed K5) was released in a nationwide campaign in March 2017 (Strive and Cox 2019).

Although K5 achieved a local reduction in rabbit populations at some release sites, there is so far no evidence that it has established in the rabbit population. This is likely due to the competing RHDV2. K5 remains available as a commercial product for rabbit control but probably only has impact as a local biocide rather than disseminating widely. Whether it provides additional genetic potential for recombination with other strains remains to be seen.

Adult wild rabbits are likely to have survived RHD and, therefore, be immune to further challenge with the same antigenic type (McPhee et al. 2009). This means that the virus will have its main impact on the susceptible kittens and subadults each breeding season. However, depending on the timing of epizootics, the impact of RHD may be reduced by age resistance, passively transferred maternal antibodies, and potentially by cross-protection from apathogenic caliciviruses (see below) (McPhee et al. 2009, Mutze et al. 2010a, 2010b, Cox et al. 2017, Cooke et al. 2018). The co-circulation of two antigenic types of RHDV that can at least partially overcome existing immunity may be altering the epidemiology in complex ways (Dalton et al. 2012, Peacock et al. 2017).

Interference with RHDV by Rabbit Caliciviruses

The impact of RHD was substantially less in the higher rainfall, more temperate areas of Australia (Henzell et al. 2002, Cooke 2002, McPhee et al. 2002). This may be due, at least in part, to an apathogenic RCV (Rabbit Calicivirus Australia-1; RCV-A1) (Strive et al. 2009, Strive et al. 2010). Infection with RCV-A1 induces antibodies that cross-react with RHDV (Nagesha et al. 2000, Cooke et al. 2002, Robinson et al. 2002a). Recent infection with RCV-A1 provides some protection from experimental challenge with

RHDV (Strive et al. 2013). However, this protection is relatively short-lived and appears to be independent of the titres of cross-reacting antibody. In addition to attenuating RHDV infection, previous exposure to RCV has been suggested to reduce the rate of RHDV transmission at a population level (Cooke et al. 2018).

As already noted, similar caliciviruses are widely distributed in farmed and wild European rabbits around the world. Recombination between RCV and RHDV, in particular exchange of the nonstructural and structural genes, generating new variants appears to be relatively common (Lopes et al. 2015, Hall et al. 2018).

Impact of RHDV in New Zealand

RHDV was illegally imported into New Zealand from Australia and released into rabbit populations in October 1997. Following its release, RHDV spread over the rabbit infested areas of New Zealand. Decrease in rabbit populations ranged from 90% in some areas to less than 20%, essentially no effect, in others (Sanson et al. 2000, Parkes et al. 2002, Norbury et al. 2002, Henning et al. 2006). Cross-reacting antibodies to RCV and possible protection from RHDV were also observed in New Zealand wild rabbits (O'Keefe et al. 1999, Nicholson et al. 2017). As in Australia, once the initial impact on the naïve population had passed, there was a gradual recovery in some rabbit populations. However, the rate of increase has been much slower in the presence of RHD than occurred following conventional controls (Parkes et al. 2008). Predation, particularly by cats, may be locally important in maintaining suppression in low density populations in the presence of RHD (Reddiex et al. 2002, Henning et al. 2008). In 2018, RHDV2 emerged in New Zealand via an unknown entry pathway. In the same year, New Zealand also released K5 as an additional tool for rabbit control. However, the effects of RHDV2 and K5 on New Zealand rabbit populations have yet to be quantified.

What Is the Future for Biological Control with RHDV?

Adult rabbits experimentally infected with RHDV1 generally only survive for 2–3 days. Based on the myxomatosis model, it might be expected that selection would occur for virus that allowed the rabbit to survive for longer and shed virus for longer. There has been no systematic study of RHDV virulence in the way MYXV was studied. However, there is no compelling epidemiological evidence to suggest that the virus is attenuating either in Australia or Europe and some data showing the emergence of strains of higher virulence compared to the released virus (Elsworth et al. 2014). The avirulent RCVs identified so far are clearly not derived from RHDV (Capucci et al. 1996, Strive et al. 2009, Jahnke et al. 2010, Mahar et al. 2016).

It is probable that RHDV is not attenuating because high virulence, leaving carcasses with extremely high titres of virus as a source of contamination for insect vectors or for direct transmission, is an effective transmission strategy that offsets the advantages of shedding for longer (Elsworth et al. 2014, Schwensow et al. 2014). By causing such rapid death, the virus almost completely avoids the adaptive immune response (Cooke and Berman 2000), which may also reduce selective pressure. In addition, very young rabbits can be infected but are resistant to disease caused by RHDV1 and, thus, could maintain virulent virus without selection pressure. However, this may be of limited importance since in most populations kittens will have passive maternal antibodies which would prevent infection or limit shedding during this period of age resistance (Robinson et al. 2002b). The emergence of RHDV2 has also considerably altered the role of kittens in RHDV epidemiology.

It has been reported that some populations of wild rabbits in Australia exhibit resistance to infection with RHDV (Saunders et al. 2010, Elsworth et al. 2012, Schwensow et al. 2017). Similar observations have been made in Europe (Fouchet et al. 2009). Studies on the mechanism of resistance have focused on the expression of complex carbohydrate molecules called histo-blood group antigens (HBGAs) on the surface of epithelial cells (Guillon et al. 2009). Binding of RHDV to these molecules appears important in the infection process although other cell-surface receptors must also be utilized since HBGAs are not expressed on hepatocytes.

Rabbits that do not express particular HGBAs on intestinal epithelium are significantly overrepresented in populations following epizootics of RHDV suggesting a strong selection advantage (Guillon et al. 2009, Nyström et al. 2011). Experimentally, such rabbits were significantly more likely to survive oral challenge with low doses of virus although they still became infected, and challenge with higher doses of virus was lethal (Nyström et al. 2011). Regulation of HBGA expression in rabbits is complex and not fully understood (Nyström et al. 2015).

Different strains of RHDV utilize different HGBAs for attachment suggesting that there is likely to be a complex and ongoing selection in rabbit populations for expression of particular HBGAs (and potentially other receptors) and in virus populations for recognition of different HBGAs. It should not be assumed that this is the only mechanism of resistance that might emerge. Indeed, other factors such as major histocompatibility molecules type I (MHC-I) have been implicated in genetic studies (Schwensow et al. 2017). This ongoing host–pathogen coevolution may have unpredictable consequences for viral virulence and biological control (Fouchet et al. 2009).

The emergence of RHDV2 may also be altering the ongoing host–pathogen dynamic in unpredictable ways. One could envisage a scenario of alternating dominance of RHDV1 and RHDV2 variants depending on the degree of immunological cross-protection. As with myxomatosis, RHD is likely to continue to suppress rabbit populations at essentially no economic cost but with gradually diminishing impact. Economic modeling has suggested that the combined impact of myxomatosis and RHD has accrued benefits to agriculture of 70 billion Australian dollars (at 2011 values) (Cooke et al. 2013). Population suppression also means that other natural processes such as predation, parasitism, and other disease agents have more impact.

Bibliography

Alves, J. M., M. Carneiro, J. Y. Cheng, A. Lemos de Matos, M. M. Rahman et al. 2019. Parallel adaptation of rabbit populations to myxoma virus. *Science* 363:1319–1326.

Asgari, S., J. R. E. Hardy, R. G. Sinclair, and B. D. Cooke. 1998. Field evidence for mechanical transmission of rabbit haemorrhagic disease virus (RHDV) by flies (Diptera: Calliphoridae) among wild rabbits in Australia. *Virus Res.* 54:123–132.

Bergin, I. L., A. G. Wise, S. R. Bolin, T. P. Mullaney, M. Kiupel et al. 2009. Novel calicivirus identified in rabbits, Michigan, USA. *Emerg. Infect. Diseases* 15:1955–1962.

Berman, D., P. J. Kerr, R. Stagg, B. H. van Leeuwen, and T. Gonzalez. 2006. Should the 40-year-old practice of releasing virulent myxoma virus to control rabbits (*Oryctolagus cuniculus*) be continued? *Wildl. Res.* 33:549–556.

Best, S. M., S. V. Collins, and P. J. Kerr. 2000. Coevolution of host and virus: cellular localization of myxoma virus infection of resistant and susceptible European rabbits. *Virology* 277:76–91.

Best, S. M., and P. J. Kerr. 2000. Coevolution of host and virus: the pathogenesis of virulent and attenuated strains of myxoma virus in resistant and susceptible European rabbits. *Virology* 267:36–48.

Bruce, J. S., and L. E. Twigg. 2004. Rabbit haemorrhagic disease virus: serological evidence of a non-virulent RHDV-like virus in south-western Australia. *Wildl. Res.* 31:602–612.

Bruce, J. S., and L. E. Twigg. 2005. The reintroduction, and subsequent impact of rabbit haemorrhagic disease virus in a population of wild rabbits in south-western Australia. *Wildl. Res.* 32:139–150.

Bull, L. B., and C. G. Dickinson. 1937. The specificity of the virus of rabbit myxomatosis. *J CSIR* 10:291–294.

Bull, L. B., and M. W. Mules. 1944. An investigation of *myxomatosis cuniculi* with special reference to the possible use of the disease to control rabbit populations in Australia. *J. CSIR* 17:79–93.

Calvete, C., M. Mendoza, A. Alcaraz, M. P. Sarto, M. P. Jiménez-de-Bagüéss et al. 2018. Rabbit haemorrhagic disease: cross-protection and comparative pathogenicity of G1.2/RHDV2/b and G1.1/RHDV lagoviruses in a challenge trial. *Vet Microbiol.* 219:87–95.

Camarda, A., N. Pugliese, P. Cavadini, E. Circella, L. Capucci et al. 2014. Detection of the new emerging rabbit haemorrhagic disease type 2 virus (RHDV2) in Sicily from rabbit (*Oryctolagus cuniculus*) and Italian hare (*Lepus corsicanus*). *Res. Vet. Sci.* 97:642–645.

Cameron, C., S. Hota-Mitchell, L. Chen, J. Barrett, J-X. Cao et al. 1999. The complete DNA sequence of myxoma virus. *Virology* 264:298–318.

Capucci, L., F. Fallacara, S. Grazioli, A. Lavazza, M. L. Pacciarini et al. 1998. A further step in the evolution of rabbit hemorrhagic disease virus: the appearance of the first consistent antigenic variant. *Virus Res.* 58:115–126.

Capucci, L., P. Fusi, A. Lavazza, M. L. Pacciarini, and C. Rossi. 1996. Detection and preliminary characterization of a new calicivirus related to rabbit haemorrhagic disease virus but non pathogenic. *J. Virol.* 70:8614–8623.

Cooke, B. D. 1983. Changes in the age-structure and size of populations of wild rabbits in South Australia, following the introduction of European rabbit fleas, *Spilopsyllus cuniculi* (Dale), as vectors of myxomatosis. *Aust. Wildl. Res.* 10:105–120.

Cooke, B. D. 1984. Factors limiting the distribution of the European rabbit flea *Spilopsyllus cuniculi* (Dale) (Siphonoptera) in inland South Australia. *Aust. J. Zool.* 32 (4):493–506.

Cooke, B. D. 1990. Notes on the comparative reproductive biology and the laboratory breeding of the rabbit flea *Xenopsylla cunicularis* Smit (Siphonaptera: Pulicidae). *Aust. J. Zool.* 38:527–534.

Cooke, B. D. 2002. Rabbit haemorrhagic disease: field epidemiology and the management of wild rabbit populations. *Rev. Sci. Tech. Off. Int. Epiz.* 21:347–358.

Cooke, B. D. 2012. Rabbits: manageable environmental pests or participants in new Australian ecosystems? *Wildl. Res.* 39:279–289.

Cooke, B. D., and D. Berman. 2000. Effect of inoculation route and ambient temperature on the survival time of rabbits, *Oryctolagus cuniculus* (L.) infected with Rabbit haemorrhagic disease virus. *Wildl. Res.* 27:137–142.

Cooke, B. D., J.-L. Chapuis, V. Magnet, A. Lucas, and J. Kovaliski. 2004. Potential use of myxoma virus and rabbit haemorrhagic disease virus to control feral rabbits in the Kerguelen Archipelago. *Wildl. Res.* 31:415–420.

Cooke, B. D., P. Chudleigh, S. Simpson, and G. Saunders. 2013. The economic benefits of the biological control of rabbits in Australia, 1950–2011. *Aust. Econ. Hist. Rev.* 53:1–17.

Cooke, B. D., R. P. Duncan, I. McDonald, J. Liu, L. Capucci et al. 2018. Prior exposure to non-pathogenic calicivirus RCV-A1 reduces both infection rate and mortality from rabbit haemorrhagic disease in a population of wild rabbits in Australia. *Transbound. Emerg. Dis* 65:e470–e477.

Cooke, B. D., and F. Fenner. 2002. Rabbit haemorrhagic disease and the biological control of wild rabbits, *Oryctolagus cuniculus*, in Australia and New Zealand. *Wildl. Res.* 29:689–706.

Cooke, B. D., S. McPhee, A. J. Robinson, and L. Capucci. 2002. Rabbit haemorrhagic disease: does a pre-existing RHDV-like virus reduce the effectiveness of RHD as a biological control in Australia. *Wildl. Res.* 29:673–682.

Cox, T., J. Liu, R. van de Ven, and T. Strive. 2017. Different serological profiles to co-occurring pathogenic and nonpathogenic caliciviruses in wild European rabbits (*Oryctolagus cuniculus*) across Australia. *J. Wildl. Dis.* 53:472–481.

Dalton, K. P., J. M. Martin, I. Nicieza, A. Podadera, D. de Llano et al. 2019. Myxoma virus jumps species to the Iberian hare. *Transbound. Emerg. Dis.* 66(6):2218–2226.

Dalton, K. P., I. Nicieza, J. Abrantes, P. J. Esteves, and F. Parra. 2014. Spread of new variant RHDV in domestic rabbits on the Iberian Peninsula. *Vet Microbiol.* 169:67–73.

Dalton, K. P., I. Nicieza, A. Balseiro, M. A. Muguerza, J. M. Rossell et al. 2012. Variant Rabbit Haemorrhagic Disease Virus in young rabbits in Spain. *Emerg. Inf. Dis.* 18:2009–2012.

Elsworth, P., B. D. Cooke, J. Kovaliski, R. Sinclair, E. C. Holmes et al. 2014. Increased virulence of rabbit haemorrhagic disease virus associated with genetic resistance in wild Australian rabbits (*Oryctolagus cuniculus*). *Virology* 464–465:415–423.

Elsworth, P. G., J. Kovaliski, and B. D. Cooke. 2012. Rabbit haemorrhagic disease: are Australian rabbits (*Oryctolagus cuniculus*) evolving resistance to infection with Czech CAPM 351 RHDV? *Epidemiol. Infect*. 140:1972–1981.

Fenner, F., M. F. Day, and G. M. Woodroofe. 1956. Epidemiological consequences of the mechanical transmission of myxomatosis by mosquitoes. *J. Hyg. (Camb.)* 54:284–303.

Fenner, F., and B. Fantini. 1999. *Biological Control of Vertebrate Pests. The History of Myxomatosis – An Experiment in Evolution*. New York: CAB International.

Fenner, F., and I. D. Marshall. 1954. Passive immunity in myxomatosis of the European rabbit (*Oryctolagus cuniculus*): the protection conferred on kittens by immune does. *J. Hyg. (Camb.)* 52:321–336.

Fenner, F., and I. D. Marshall. 1957. A comparison of the virulence for European rabbits (*Oryctolagus cuniculus*) of strains of myxoma virus recovered in the field in Australia, Europe and America. *J. Hyg. (Camb.)* 55:149–191.

Fenner, F., I. D. Marshall, and G. M. Woodroofe. 1953. Studies in the epidemiology of infectious myxomatosis of rabbits. I. Recovery of Australian wild rabbits (*Oryctolagus cuniculus*) from myxomatosis under field conditions. *J. Hyg. (Camb.)* 51:225–244.

Fenner, F., W. E. Poole, I. D. Marshall, and A. L. Dyce. 1957. Studies in the epidemiology of infectious myxomatosis of rabbits. VI. The experimental introduction of the European strain of myxoma virus into Australian wild rabbit populations. *J. Hyg. (Camb.)* 55:192–206.

Fenner, F., and F. N. Ratcliffe. 1965. *Myxomatosis*. Cambridge: Cambridge University Press.

Fenner, F., and J. Ross. 1994. Myxomatosis. In *The European Rabbit. The History and Biology of a Successful Colonizer*, edited by H. V. Thompson and C. M. King, 205–240. Oxford: Oxford University Press.

Fenner, F., and G. M. Woodroofe. 1965. Changes in the virulence and antigenic structure of strains of myxoma virus recovered from Australian wild rabbits between 1950 and 1964. *Aust. J. Exptl. Biol. Med. Sci*. 43:359–370.

Forrester, N. L., R. C. Trout, and E. A. Gould. 2007. Benign circulation of rabbit haemorrhagic disease virus on Lambay Island, Eire. *Virology* 358:18–22.

Fouchet, D., J. Le Pendu, J.-S. Guitton, M. Guiserix, S. Marchandeau et al. 2009. Evolution of microparasites in spatially and genetically structured host populations: the example of RHDV infecting rabbits. *J. Theor. Biol*. 257:212–227.

Guillon, P., N. Rovoen-Clouet, B. le Moullac-Valdye, S. Marchandeau, and J. Le Pendu. 2009. Association between expression of the H histo-blood group antigen, α1,2 fucosyltransferases polymorphism of wild rabbits, and sensitivity to rabbit haemorrhagic disease virus. *Glycobiology* 19:21–28.

Hall, R. N., J. E. Mahar, S. Haboury, V. Stevens, E. C. Holmes et al. 2015. Emerging rabbit hemorrhagic disease virus 2 (RHDVb) in Australia. *Emerg. Inf. Dis*. 21:2276–2278.

Hall, R. N., J. E. Mahar, A. J. Read, R. Mourant, M. Piper et al. 2018. A strain-specific multiplex RT-PCR for Australian rabbit haemorrhagic disease viruses uncovers a new recombinant virus variant in rabbits and hares. *Transbound. Emerg. Dis* 65:e444–e456.

Hall, R. N., D. E. Peacock, J. Kovaliski, J. E. Mahar, R. Mourant et al. 2017. Detection of RHDV2 in European brown hares (*Lepus europaeus*) in Australia. *Vet. Record* 180:121.

Henning, J., J. Meers, P. R. Davies, and R. S. Morris. 2005. Survival of rabbit haemorrhagic disease virus (RHDV) in the environment. *Epidemiol. Infect*. 133:719–730.

Henning, J., D. U. Pfeiffer, P. R. Davies, J. Meers, and R. S. Morris. 2006. Temporal dynamics of rabbit haemorrhagic disease virus infection in a low-density population of wild rabbits (*Oryctolagus cuniculus*) in New Zealand. *Wildl. Res*. 33:293–303.

Henning, J., D. U. Pfeiffer, P. R. Davies, M. A. Stevenson, and J. Meers. 2008. Mortality patterns over 3 years in a sparse population of wild rabbits (*Oryctolagus cuniculus*) in New Zealand, with an emphasis on rabbit haemorrhagic disease (RHD). *Eur. J. Wild. Res*. 54:619–626.

Henzell, R. P., B. D. Cooke, and G. D. Mutze. 2008. "The future biological control of pest populations of European rabbits, *Oryctolagus cuniculus*." *Wildl. Res*. 35:633–650.

Henzell, R. P., R. B. Cunningham, and H. M. Neave. 2002. Factors affecting the survival of Australian wild rabbits exposed to rabbit haemorrhagic disease. *Wildl. Res*. 29:523–542.

Hoehn, M., P. J. Kerr, and T. Strive. 2013. In situ hybridisation assay for localisation of rabbit calicivirus Australia-1 (RCV-A1) in European rabbit (*Oryctolagus cuniculus*) tissues. *J. Virol. Meth*. 188:148–152.

Jahnke, M., E. C. Holmes, P. J. Kerr, J. D. Wright, and T. Strive. 2010. Evolution and phylogeography of the non-pathogenic calicivirus RCV-A1 in wild rabbits in Australia. *J. Virol*. 84:12397–12404.

Jeklova, E., L. Leva, J. Matiosovic, K. Kovarcik, H. Kudlackova et al. 2008. Characterisation of immunosuppression in rabbits after infection with myxoma virus. *Vet. Microbiol*. 129:117–130.

Kerr, P.J. 1997. An ELISA for epidemiological studies of myxomatosis: persistence of antibodies to myxoma virus in European rabbits (*Oryctolagus cuniculus*). *Wildl. Res*. 24:53–65.

Kerr, P. J. 2012. Myxomatosis in Australia and Europe: a model for emerging infectious diseases. *Antiviral Res*. 93:387–415.

Kerr, P. J., I. M. Cattadori, J. Liu, D. G. Sim, J. W. Dodds et al. 2017. Next step in the ongoing arms race between myxoma virus and wild rabbits in Australia is a novel disease phenotype. *Proc. Natl. Acad. Sci. USA*. 114:9397–9402.

Kerr, P. J., J-S. Eden, F. Di Giallonardo, D. Peacock, J. Liu, et al. 2019. Punctuated evolution of myxoma virus: rapid and disjunct evolution of a recent viral lineage in Australia. *J. Virol*. 93:e01994–18.

Kerr, P. J., J. Liu, I. M. Cattadori, E. Ghedin, A. F. Read et al. 2015. Myxoma virus and the leporipoxviruses: an evolutionary paradigm. *Viruses* 7:1020–1061.

King, D. R., A. J. Oliver, and S. H. Wheeler. 1985. The European rabbit flea, *Spilopsyllus cuniculi*, in southwestern Australia. *Aust. Wildl. Res*. 12:227–236.

Le Gall Reculé, G., F. Zwinglestein, S. Boucher, B. Le Normand, G. Plassiart et al. 2011a. Detection of a new variant of Rabbit Haemorrhagic Disease Virus in France. *Vet. Record* 168:137–138.

Le Gall Reculé, G., F. Zwinglestein, M. P. Fages, S. Bertagnoli, J. Gelfi et al. 2011b. Characterisation of a non-pathogenic and non-protective infectious rabbit lagovirus related to RHDV. *Virology* 410:395–402.

Le Pendu, J., J. Abrantes, S. Bertagnoli, J.-S. Guitton, G. Le Gall Reculé et al. 2017. Proposal for a unified classification system and nomenclature of lagoviruses. *J. Gen. Virol*. 98:1658–1666.

Lenghaus, C., H. Westbury, B. Collins, N. Ratnamohan, and C. Morrissy. 1994. Overview of the RHD project in the Australian Animal Health Laboratory. In *Rabbit Haemorrhagic Disease: Issues in Assessment for Biological Control*, edited by R. K. Munro and R. T. Williams, 104–129. Canberra: Bureau of Resource Sciences.

Lopes, A. M., K. P. Dalton, M. J. Magalhães, F. Parra, P. J. Esteves et al. 2015. Full genome analysis of new variant rabbit haemorrhagic disease virus revealed multiple recombination events. *J. Gen. Virol*. 96:1309–1319.

Lopes, A. M., S. Marques, E. Silva, M. J. Magalhães, A. Pinheiro et al. 2014. Detection of RHDV strains in the Iberian hare (*Lepus granatensis*): earliest evidence of rabbit lagovirus cross-species infection. *Vet. Res*. 45:94.

Lopes, A. M., D. Silvério, M. J. Magalhães, H. Areal, P. C. Alves et al. 2017. Characterization of old RHDV strains by complete genome sequencing identifies a novel genetic group. *Scientific Reports* 7:13599.

Mahar, J. E., R. N. Hall, D. Peacock, J. Kovaliski, M. Piper et al. 2018a. Rabbit hemorrhagic disease virus 2 (RHDV2; G1.1) is replacing endemic strains of RHDV in the Australian landscape within 18 months of its arrival. *J. Virol*. 92:e01374-17.

Mahar, J. E., L. Nicholson, J-S. Eden, S. Duchêne, P. J. Kerr et al. 2016. Benign rabbit caliciviruses exhibit evolutionary dynamics similar to those of their virulent relatives. *J. Virol*. 90:9317–9329.

Mahar, J. E., A. J. Read, X. Gu, N. Urakova, R. Mourant et al. 2018b. Detection and circulation of a novel rabbit hemorrhagic disease virus in Australia. *Emerg. Inf. Dis*. 24:22–31.

Marchandeau, S., S. Bertagnoli, Y. Leonard, H. Santin-Janin, B. Peralta et al. 2010. Serological evidence for the presence of non-pathogenic rabbit haemorrhagic disease virus-like strains in rabbits (*Oryctolagus cuniculus*) of the Kerguelen archipelago. *Polar Biol*. 33:985–989.

Marchandeau, S., G. Le Gall-Recule, S. Bertagnoli, J. Aubineau, G. Botti et al. 2005. Serological evidence for a non-protective RHDV-like virus. *Vet. Res*. 36:53–62.

Marshall, I. D. 1959. The influence of ambient temperature on the course of myxomatosis in rabbits. *J. Hyg. (Camb.)* 57:484–497.

Marshall, I. D., and G. W. Douglas. 1961. Studies in the epidemiology of infectious myxomatosis of rabbits. VIII. Further observations on changes in the innate resistance of Australian wild rabbits exposed to myxomatosis. *J. Hyg. (Camb.)* 59:117–122.

Marshall, I. D., A. L. Dyce, W. E. Poole, and F. Fenner. 1955. Studies in the epidemiogy of infectious myxomatosis of rabbits. IV. Observations of disease behaviour in two localities near the northern limit of rabbit infestation in Australia, May 1952 to April 1953. *J. Hyg. (Camb.)* 53:12–25.

Marshall, I. D., and F. Fenner. 1958. Studies in the epidemiology of infectious myxomatosis of rabbits. V. Changes in the innate resistance of wild rabbits exposed to myxomatosis. *J. Hyg. (Camb.)* 56:288–302.

Marshall, I. D., and F. Fenner. 1960. Studies in the epidemiology of infectious myxomatosis of rabbits. VII. The virulence of strains of myxoma virus recovered from Australian wild rabbits between 1951 and 1959. *J. Hyg. (Camb.)* 58:485–488.

Martin, C. J. 1936. Observations on *Myxomatosis cuniculi* (Sanarelli) made with a view to the use of the virus in the control of rabbit plagues. *CSIR Australia Research Bulletin, No. 96*:1–28.

Matthaei, M., P. J. Kerr, A. J. Read, P. Hick, S. Haboury et al. 2014. Comparative quantitative monitoring of rabbit haemorrhagic disease viruses in rabbit kittens. *Virol. J*. 11:109.

McColl, K. A., J. C. Merchant, J. Hardy, B. D. Cooke, A. Robinson et al. 2002a. Evidence for insect transmission of rabbit haemorrhagic disease virus. *Epidemiol. Infect*. 129:655–633.

McColl, K. A., C. J. Morrissy, B. J. Collins, and H. A. Westbury. 2002b. Persistence of rabbit haemorrhagic disease virus in decomposing rabbit carcases. *Aust. Vet. J*. 80:298–299.

McPhee, S. R., D. Berman, A. Gonzales, K. L. Butler, J. Humphrey et al. 2002. Efficacy of a competitive enzyme-linked immunosorbent assay (cELISA) for estimating prevalence of immunity to rabbit haemorrhagic disease virus (RHDV) in populations of Australian wild rabbits (*Oryctolagus cuniculus*). *Wildl. Res*. 29:635–647.

McPhee, S. R., K. L. Butler, J. Kovaliski, G. Mutze, L. Capucci et al. 2009. Antibody status and survival of Australian wild rabbits challenged with rabbit haemorrhagic disease virus. *Wildl. Res*. 36:447–456.

Mikami, O., J. H. Park, T. Kimura, K. Ochial, and C. Itakura. 1999. Hepatic lesions in young rabbits experimentally infected with rabbit haemorrhagic disease virus. *Res. Vet. Sci*. 66:237–242.

Mutze, G., P. I. Bird, J. Kovaliski, D. Peacock, S. Jennings et al. 2002. Emerging epidemiological patterns in rabbit haemorrhagic disease, its interactions with mxyomatosis, and their effects on rabbit populations in South Australia. *Wildl. Res*. 29:577–590.

Mutze, G., P. Bird, D. Peacock, N. De Preu, J. Kovaliski, B. D. Cooke et al. 2014. Recovery of South Australian rabbit populations from the impact of rabbit haemorrhagic disease. *Wildl. Res*. 41:552–559.

Mutze, G., B. Cooke, and P. Alexander. 1998. The initial impact of rabbit hemorrhagic disease on European rabbit populations in South Australia. *J. Wildl. Dis*. 34:221–227.

Mutze, G., N. De Preu, T. Moody, D. Koerner, D. McKenzie et al. 2018. Substantial numerical decline in South Australian rabbit populations following the detection of rabbit haemorrhagic disease virus 2. *Vet. Rec*. 182(20):574.

Mutze, G., J. Kovaliski, K. Butler, L. Capucci, and S. McPhee. 2010a. The effect of rabbit population control programmes on the impact of rabbit haemorrhagic disease in south-eastern Australia. *J. App. Ecol*. 47:1137–1146.

Mutze, G., R. Sinclair, D. Peacock, J. Kovaliski, and L. Capucci. 2010b. Does a benign calicivirus reduce the effectiveness of rabbit haemorrhagic disease virus (RHDV) in Australia? Experimental evidence from field releases of RHDV on bait. *Wildl. Res*. 37:311–319.

Myers, K. 1954. Studies in the epidemiology of infectious myxomatosis of rabbits. II. Field experiments, August-November 1950, and the first epizootic of myxomatosis in the Riverine plain of south-eastern Australia. *J. Hyg. (Camb.)* 52:47–59.

Myers, K., I. D. Marshall, and F. Fenner. 1954. Studies in the epidemiology of infectious myxomatosis of rabbits. III. Observations on two succeeding epizootics in Australian wild rabbits on the Riverine plain of south-eastern Australia. *J. Hyg. (Camb.)* 52:337–360.

Nagesha, H. S., K. A. McColl, B. J. Collins, C. J. Morrissy, L. F. Wang et al. 2000. The presence of cross-reactive antibodies to rabbit haemorrhagic disease virus in Australian wild rabbits prior to the escape of virus from quarantine. *Arch Virol* 145:749–757.

Neave, M. J., R. N. Hall, N. Huang, K. A. McColl, P. Kerr et al. 2018. Robust innate immunity of young rabbits mediates resistance to rabbit hemorrhagic disease caused by *Lagovirus europaeus* G1.1 but not G1.2. *Viruses* 10:512.

Neimanis, A., U. Larsson Pettersson, N. Huang, D. Gavier-Widén, and T. Strive. 2018. Elucidation of the pathology and tissue distribution of *Lagovirus europaeus* G1.2/RHDV2 (Rabbit haemorrhagic disease virus 2) in young and adult rabbits (*Oryctolagus cuniculus*). *Vet. Res.* 49:46.

Nicholson, L. J., J. E. Mahar, T. Strive, T. Zheng, E. C. Holmes et al. 2017. Benign rabbit calicivirus in New Zealand. *Appl. Environ. Microbiol.* 83:e00090–17.

Norbury, G., R. Heyward, and J. Parkes. 2002. Short-term ecological effects of rabbit haemorrhagic disease in the short-tussock grasslands of the South Island, New Zealand. *Wildl. Res.* 29:599–604.

Nyström, K., J. Abrantes, A. M. Lopes, B. Le Moullac-Vaidye, S. Marchandeau et al. 2015. Neofunctionalization of the Sec1 α1,2fucosyltransferase paralogue in leporids contributes to glycan polymorphism and resistance to rabbit haemorrhagic disease. *PLoS Pathog* 11:e1004759.

Nyström, K., G. Le Gall Reculé, P. Grassi, J. Abrantes, N. Ruvoën-Clouet et al. 2011. Histo-blood group antigens act as attachment factors of rabbit hemorrhagic disease virus infection in a virus strain-dependent manner. *PLoS Pathog* 7:e1002188.

O'Keefe, J. S., J. E. Tempero, M. X. J. Motha, M. F. Hansen, and P. H. Atkinsona. 1999. Serology of rabbit haemorrhagic disease virus in wild rabbits before and after release of the virus in New Zealand. *Vet. Microbiol.* 66:29–40.

Parer, I., D. Conolly, and W. R. Sobey. 1985. Myxomatosis: the effects of annual introductions of an immunizing strain and a highly virulent strain of myxoma virus into rabbit populations at Urana, New South Wales, Australia. *Aust. Wildl. Res.* 12:407–424.

Parkes, J. P., B. Glentworth, and G. Sullivan. 2008. Changes in immunity to Rabbit haemorrhagic disease virus, and in abundance and rates of increase of wild rabbits in Mackenzie Basin, New Zealand. *Wildl. Res.* 35:775–779.

Parkes, J. P., G. L. Norbury, R. P. Heyward, and G. Sullivan. 2002. Epidemiology of rabbit haemorrhagic disease (RHD) in the South Island, New Zealand, 1997–2001. *Wildl. Res.* 29:543–555.

Peacock, D., J. Kovaliski, R. Sinclair, G. Mutze, A. Iannella et al. 2017. RHDV2 overcoming RHDV immunity in wild rabbits (*Oryctolagus cuniculus*) in Australia. *Vet. Record.* doi:10.1136/vr.104135.

Pedler, R. D., R. Brandle, J. L. Read, R. Southgate, P. Bird et al. 2016. Rabbit biocontrol and landscape-scale recovery of threatened desert mammals. *Conserv. Biol.* 30:774–782.

Puggioni, G., P. Cavadini, C. Maestrale, R. Scivoli, G. Botti et al. 2013. The new French 2010 *Rabbit haemorrhgic disease virus* cases an RHD-like disease in the Sardinian Cape hare (*Lepus capensis mediterraneus*). *Vet. Res.* 44:96.

Ramsey, D. S. L., T. Cox, T. Strive, D. M. Forsyth, I. Stuart et al. 2020. Diverse impacts of two novel strains of rabbit haemorrhagic disease virus on wild rabbit populations in Australia. *J. App. Ecol.* 57:630–641.

Ratcliffe, F. N., K. Myers, B. V. Fennessy, and J. H. Calaby. 1952. Myxomatosis in Australia. A step towards the biological control of the rabbit. *Nature* 170:7–19.

Reddiex, B., G. J. Hickling, G. L. Norbury, and C. M. Frampton. 2002. Effects of predation and rabbit haemorrhagic disease on population dynamics of rabbits (*Oryctolagus cuniculus*) in North Canterbury, New Zealand. *Wildl. Res.* 29:627–633.

Richardson, B. J., S. Philips, R. A. Hayes, S. Sindhe, and B. D. Cooke. 2007. Aspects of the biology of the European rabbit (Oryctolagus cuniculus) and rabbit haemorrhagic disease virus (RHDV) in coastal eastern Australia. *Wildl. Res.* 34: 398–407.

Robinson, A. J., P. D. Kirkland, R. I Forrester, L. Capucci, B. D. Cooke et al. 2002a. Serological evidence for the presence of a calicivirus in Australian wild rabbits, *Oryctolagus cuniculus*, before the introduction of rabbit haemorrhagic disease virus (RHDV): its potential influence on the specificity of a competitive ELISA for RHDV. *Wildl. Res*. 29:655–652.

Robinson, A. J., P. T. M. So, W. J. Muller, B. D. Cooke, and L. Capucci. 2002b. Statistical models for the effect of age and maternal antibodies on the development of rabbit haemorrhagic disease in Australian wild rabbits. *Wildl. Res*. 29:663–671.

Rogers, P. M., C. P. Arthur, and R. C. Soriguer. 1994. The rabbit in continental Europe. In *The European Rabbit. The History and Biology of a Successful Colonizer*, edited by H. V. Thompson and C. M. King, 22–63. Oxford: Oxford University Press.

Saint, K. M., N. French, and P. Kerr. 2001. Genetic variation in Australian isolates of myxoma virus: an evolutionary and epidemiological study. *Arch. Virol*. 146:1105–1123.

Sanson, R. L., H. V. Brooks, and G. W. Horner. 2000. An epidemiological study of the spread on rabbit haemorrhagic disease virus amongst previously non-exposed rabbit populations in the North Island of New Zealand. *N. Z. Vet. J*. 48:105–110.

Saunders, G., D. Choquenot, J. McIlroy, and R. Packwood. 1999. Initial effects of rabbit haemorrhagic disease on free-living rabbit (*Oryctolagus cuniculus*) populations in central-western New South Wales. *Wildl Res* 26:69–74.

Saunders, G., B. Cooke, K. McColl, R. Shine, and T. Peacock. 2010. Modern approaches for the biological control of vertebrate pests: an Australian perspective. *Biol. Control* 52:288–295.

Schwensow, N., B. Cooke, J. Kovaliski, R. G. Sinclair, D. Peacock et al. 2014. Rabbit haemorrhagic disease virus: persistence and adaptation in Australia. *Evolutionary Adaptations* 7:1056–1067.

Schwensow, N. I., H. Detering, S. Pederson, C. Mazzoni, R. Sinclair et al. 2017. Resistance to RHD virus in wild Australian rabbits: comparison of susceptible and resistant individuals using a genomewide approach. *Mol. Ecol*. 26:4551–4561.

Shien, J. H., H. K. Shieh, and L. H. Lee. 2000. Experimental infections of rabbits with rabbit haemorrhagic disease virus monitored by polymerase chain reaction. *Res. Vet. Sci*. 68:255–259.

Silvério, D., A. M. Lopes, J. Ferriera-Melo, M. J. Magalhães, P. Monterroso et al. 2018. Insights into the evolution of the new variant rabbit haemorrhagic disease virus (G1.2) and the identification of novel recombinant strains. *Transbound. Emerg. Dis*. 65:983–982.

Sobey, W. R., and W. Menzies. 1969. Myxomatosis: the introduction of the European rabbit flea *Spilopsyllus cuniculi* (Dale) into Australia. *Aust. J. Sci*. 31:404–405.

Story, G., D. Berman, R. Palmer, and J. Scanlan. 2004. The impact of rabbit haemorrhagic disease on wild rabbit (*Oryctolagus cuniculus*) populations in Queensland. *Wildl. Res*. 31:183–193.

Strive, T., and T. Cox. 2019. Lethal biological tools for landscape-scale mitigation of rabbit impacts in Australia. *Aust. Zool.* doi/10.7882/AZ.2019.016.

Strive, T., P. Elsworth, J. Liu, J. D. Wright, J. Kovaliski et al. 2013. The non-pathogenic Australian rabbit calicivirus RCV-A1 provides temporal and partial cross-protection to lethal rabbit haemorrhagic disease virus infection which is not dependent on antibody titres. *Vet. Res*. 44:51.

Strive, T., J. Wright, J. Kovaliski, G. Botti, and L. Capucci. 2010. The non-pathogenic Australian lagovirus RCV-A1 causes a prolonged infection and elicits partial cross-protection to rabbit haemorrhagic disease virus. *Virology* 398:125–134.

Strive, T., J. D. Wright, and A. J. Robinson. 2009. Identification and partial characterisation of a new lagovirus in Australian wild rabbits. *Virology* 384:97–105.

Williams, K., I. Parer, B. Coman, J. Burley, and M. Braysher. 1995. *Managing Vertebrate Pests: Rabbits. Canberra: Bureau of Resource Sciences*. Australian Government Publishing Services.

14

Cabbage Disease Ecology and Management

Anthony P. Keinath,
Marc A. Cubeta,
and David B.
Langston, Jr.

Introduction

Cabbage (*Brassica oleracea* 'Capitata Group') has long been cultivated as an important vegetable crop and a source of fiber, minerals, and vitamins, particularly during cold seasons in temperate climates. More recently, cabbage and other cruciferous vegetables (members of the Brassicaceae) have been recognized as important sources of dietary chemoprotective phytochemicals. Cabbage and other brassica vegetables were grown on 2.4 million hectares worldwide in 2018 (Food and Agriculture Organization, 2020). Cabbage is a productive vegetable based on biomass per area of cultivation, which averaged 28.7 ton/ha in 2018. However, many diseases, particularly those caused by bacteria and fungi, affect the production of this crop. This chapter focuses on six diseases of worldwide importance in cabbage production. These diseases also affect other cole crops, i.e., vegetables derived from *B. oleracea*, including broccoli, Brussels sprouts, cauliflower, collard, kale, and kohlrabi, and other genetically related cruciferous vegetables, such as Chinese cabbage, mustard, rutabaga, and turnip. Emphasis will be placed on pathogen life cycle stages susceptible to management techniques. Control measures will be presented in an integrated pest management (IPM) framework.

Major Diseases and Pathogen Ecology

Black Rot

Black rot is caused by the bacterium *Xanthomonas campestris* pathovar *campestris*. *X. campestris* pv. *raphani* also has been isolated from necrotic spots on cabbage seedlings (Lange et al. 2016). Because this bacterium can be seedborne, black rot occurs in most areas of the world where cabbage and other crucifers are grown. The pathogen produces V-shaped chlorotic and necrotic lesions starting at the margins of leaves, but it also causes wilting of plants if it reaches the vascular system in the stem (systemic infection). Blackening of the leaf veins is a helpful diagnostic symptom. The pathogen survives in infested crop debris and usually survives only for a few months in soil in the absence of host debris.

Clubroot

Clubroot is caused by the protozoan-like organism *Plasmodiophora brassicae*. This soilborne organism is an obligate biotroph, completing its unique life cycle within the root cells of crucifers. Infected root cells enlarge and divide to produce the diagnostic swollen, club-like roots. The pathogen produces resting spores in the clubs that persist in soil for at least 10 years after they decay. Isolates of *P. brassicae* differ in host range, and races have been found that are pathogenic on the few resistant cultivars of cabbage. Resting spores are extremely resistant to desiccation and easily spread with wind-blown soil and irrigation water.

Black Spot, Dark Leaf Spot

Five species of *Alternaria*, but primarily *A. brassicae* and *A. brassicicola*, infect cabbage and other crucifers. *A. brassicicola* has a higher optimal temperature for growth, sporulation, and spore germination (20°C–30°C) than *A. brassicae* (18°C–24°C). Both fungi can be seedborne and airborne but do not survive apart from infested host debris in soil. Infested debris left on the soil surface can be a significant source of pathogen spores for up to 12 weeks after harvest (Humpherson-Jones 1989). Seedborne inoculum can lower seed germination and vigor but usually is not damaging to seedlings.

Downy Mildew

Crucifer downy mildew is caused by the oomycete *Hyaloperonospora parasitica*. This biotroph produces airborne sporangia on the underside of leaves and oospores inside infected leaves, roots, and stems. The pathogen is believed to survive as dormant oospores in roots and stems and on living hosts in frost-free regions. Cabbage is affected by downy mildew particularly during the seedling and heading growth stages. High relative humidity, dew, and fog are favorable for infection.

Watery Soft Rot, Sclerotinia Stem Rot, White Mold

Sclerotinia sclerotiorum has a wide host range but is especially damaging to cabbage, because it not only infects the head in the field but also can cause decay in transit and storage. Infection occurs primarily at head maturity when wrapper leaves shade the soil, providing a cool, moist environment that favors the pathogen. This fungus produces airborne sexual spores (ascospores) that infect plants, and soilborne survival structures (sclerotia) cause infection when they germinate near a plant. The closely related species *Sclerotinia minor* also has been reported on brassica vegetables.

Wirestem

Wirestem, a postemergence disease, is caused by the soilborne fungus *Rhizoctonia solani* anastomosis groups (AGs) 2-1 and 4 (Budge and Shaw 2009). In soils cropped repeatedly to crucifers, AG 2-1 predominates. At low pathogen levels, wirestem is more prevalent or severe than preemergence damping-off. Seedlings may be killed by wirestem when lesions girdle stems. Older plants may be killed later or be stunted and fail to produce a marketable-sized head. Root rot also occurs when infection is severe but is absent when discrete stem lesions are the only symptoms.

Control

General Control Principles

Exclusion. It is extremely important to prevent contamination of clubroot-free land by excluding the pathogen. Movement of field-grown transplants, soil, and equipment from clubroot-infested fields or farms should be avoided (Hwang et al. 2014). Growers in clubroot-free areas should avoid purchasing field-grown transplants or equipment from infested areas. Before purchase, greenhouse-grown transplants should be inspected for symptoms of downy mildew, black rot, and black spot.

Eradication. Testing seed for the fungal pathogen *Plenodomus* (=*Phoma*) *lingam* has greatly reduced outbreaks of black leg disease of crucifers. Cruciferous weeds should be eradicated to eliminate sources of the pathogens causing black rot, downy mildew, and clubroot (Table 1). Cruciferous ornamentals also can be infected by the same species of *Alternaria*, *Hyaloperonospora*, *Plasmodiophora,* and *Xanthomonas* that infect cabbage. The parasitic fungus *Paraconiothyrium minitans*, the active ingredient in the biofungicide Contans®, infects and destroys sclerotia of *S. sclerotiorum* and *S. minor.*

Avoidance. Cabbage crops should be rotated to a different field or plot each season to avoid contact with infested debris remaining from previous crops. Using a shallow planting depth for transplants reduces wirestem incidence, because the susceptible hypocotyl is not directly in contact with *Rhizoctonia*-infested soil. Avoid wounding plants to prevent black rot, bacterial soft rot, and watery soft rot.

Resistance. Host plant resistance is widely available in green (white) and red cabbage for yellows (caused by *Fusarium oxysporum* f. sp. *conglutinans*). Some hybrid cultivars have partial resistance to black rot that restricts lesions to the wrapper leaves. A few cabbage cultivars (mostly red cabbage) have moderate resistance to *Alternaria*. Cabbage cultivars available in the United States are susceptible to clubroot, downy mildew, white mold, and wirestem.

Protection. Seed treatment with fungicides is very effective in preventing damping-off caused by species of *Pythium* and *R. solani*. Protectant fungicides are effective against black spot and downy mildew and used for managing black rot, clubroot, and wirestem with varying degrees of success. Potassium phosphite is an alternative fungicide that is effective against downy mildew. The fungicides boscalid and penthiopyrad are registered in the United States to control *Sclerotinia* on cole crops.

TABLE 1 Management Practices for Common Diseases of Cabbage

Disease	Plant-Resistant Cultivars	Use Healthy Seed or Transplants	Eliminate Weeds	Avoid Wounding	Bury Crop Residue	Rotate with Nonhosts	Apply Protectant Fungicide or Bactericide
Black spot	+	+	+	–	+	+	+
Bacterial soft rot	–	–	–	+	–	–	–
Black leg	–	+	–	–	+	+	+
Black rot	+	+	+	+	+	+	–
Clubroot	–	+*	+	–	–	–	+/–
Downy mildew	–	+	+	–	+	+	+
Yellows	+	–	–	–	–	–	–
Sclerotinia stem rot	–	–	+	+	+	–	+/–
Damping-off	–	–	–	–	–	–	+
Wirestem	–	+*	–	–	+	–	+

+ = Practice can be used to manage the disease; – = Practice is ineffective or inappropriate, based on pathogen life cycle; +/– = Practice may be useful under certain conditions.

* The pathogen is not seedborne but can be spread on infected, field-grown transplants.

Therapy. The only measure to control cabbage diseases postinfection is the application of systemic fungicides for downy mildew.

Examples of Integrated Disease Management

Cabbage scouting guides have been developed and are useful for surveying production fields for diseases, insects, and weeds. Controlling weeds, especially ragweed (*Ambrosia artemisiifolia*), can reduce the incidence of watery soft rot. Ascospores of *Sclerotinia* can also infect ragweed flowers that fall onto cabbage leaves and heads because flower parts provide nutrients for the pathogen (Dillard and Hunter 1986). Lastly, it is important to control flea beetles (*Phyllotreta cruciferae*), which carry spores of *Alternaria brassisicola* on their bodies and in their frass and transmit conidia while feeding.

Managing Seedborne Pathogens

Seed should originate from seedlots that have tested negative for the presence of the black leg and black rot pathogens. Hot water seed treatment is useful to control seedborne black rot bacteria, provided the water temperature is monitored carefully so it remains at 50°C for 25 minutes to avoid reducing seed germination and quality. Minimize leaf wetness periods when producing transplants in glasshouses, because of the ease of spreading pathogens. Apply protectant fungicides to seed crops to prevent infection of seed by *Alternaria*.

Managing Soilborne Pathogens

Soil fumigants generally are not used against soilborne pathogens in cabbage production because of the high cost, although they may be used to disinfest seedbeds and suppress clubroot. Field-grown transplants may be sources of the clubroot and wirestem pathogens and spread them to noninfested fields. Because of this risk, transplants should be produced in soilless mixes in greenhouses when possible. Do not plant any cruciferous cover crops or vegetables in fields before or after cropping to cabbage. Use monocots as rotation crops, because *R. solani* AG-4 and *S. sclerotiorum* have wide host ranges among dicotyledonous crops. The resting spores of the clubroot organism cannot be eradicated by rotation. Instead, liming soil to raise the pH above 7.2 with calcium oxide or hydrated lime prevents infection of roots.

Managing Foliar Pathogens

Diseases caused by foliar pathogens, such as *Alternaria* and *Xanthomonas*, can be managed with crop rotation during the period when infested host debris is decaying in affected fields, because these foliar pathogens of cabbage do not survive longer than 1 or 2 years in soil, respectively. Disk and bury or compost unmarketable cabbage heads. Apply protectant fungicides as needed based on environmental conditions and host susceptibility. Black spot can be reduced by increasing row width and plant spacing to promote air circulation that dries leaves, because *Alternaria* spp. require relatively long periods of leaf wetness for infection (a minimum of 5–9 hours).

Conclusion

The diseases dark spot, downy mildew, watery soft rot, and wirestem often can be managed successfully using a combination of biological, chemical, and cultural control measures. The biological and cultural methods listed in Table 1 are amenable to organic production systems. Management of black rot and clubroot remains challenging. In the future, resistance to downy mildew and improved resistance to black rot may be available in cabbage cultivars. It may be possible to transfer downy mildew resistance from broccoli to cabbage using molecular genetics methods. Additional molecular genetics research is needed to clarify the identity of races of the downy mildew and clubroot organisms.

Acknowledgments

Technical Contribution No. 5100 of the Clemson University Experiment Station. We thank Richard Morrison, Sakata Seed Company, and J. Powell Smith, Clemson University, for reviewing this chapter.

References

Budge, G.E., and Shaw, M.W. 2009. Characterization and origin of infection of *Rhizoctonia solani* associated with *Brassica oleracea* crops in the UK. *Plant Pathol.* 58:1059–1070.

Dillard, H.R., and Hunter, J.E. 1986. Association of common ragweed with Sclerotinia rot of cabbage. *Plant Dis.* 70 :26–28. Available at http://www.fao.org/statistics/en/.

Humpherson-Jones, F.M. 1989. Survival of *Alternaria brassicae* and *Alternaria brassicicola* on crop debris of oilseed rape and cabbage. *Ann. Appl. Biol.* 115:45–50.

Hwang, S.F., Howard, R.J., Strelkov, S.E., Gossen, B.D., and Peng, G. 2014. Special Issue: Management of clubroot (*Plasmodiophora brassicae*) on canola (*Brassica napus*) in western Canada. *Can. J. Plant Pathol.* 36:49–65.

Lange, H. W., Tancos, M. A., Carlson, M. O., and Smart, C. D. 2016. Diversity of *Xanthomonas campestris* isolates from symptomatic crucifers in New York State. *Phytopathology* 106:113–122.

15

Natural Enemies and Biocontrol: Artificial Diets

Simon Grenier and
Patrick De Clerq

Introduction

Arthropod parasitoids and predators used in biological control strategies are at present mainly produced on natural or alternative hosts or prey. However, their large-scale production may be more convenient and cost-effective when using artificial diets/media. Studies aiming at the successful development of arthropod parasitoids and predators under artificial conditions have started a long time ago, but the practical use of insects and mites grown on artificial diets is still in its infancy. Besides their use for the production of natural enemies, artificial media may be valuable tools for physiological and behavioral studies of entomophagous arthropods due to a simplification of their environment. Different types of artificial diets with or without insect additives can support the development and/ or reproduction of natural enemies. Successes have been achieved for several species of parasitoids and predators but these have mainly been restricted to an experimental level. Comparisons of the performances of artificially vs. naturally reared natural enemies (as quality control) have primarily been conducted in the laboratory, and only very rarely in the field. The promising results achieved in recent years open up new prospects for natural enemy producers.

Artificial Diets for Predators and Parasitoids

The culture of entomophagous insects and mites involves rearing not only of the host/prey, but often also of the host's/prey's plant food, and thus requires a tritrophic level system. Different steps were taken to try to reduce the production line for entomophagous arthropods. The complete line comprises plant growing, host/prey rearing, and parasitoid/predator rearing. The simplified line includes the use of artificial diets instead of plants for the phytophagous host/prey, or of factitious hosts/prey that are easier to rear in the laboratory than the natural food (e.g., eggs of *Ephestia kuehniella* or *Sitotroga cerealella*, larvae of *Galleria mellonella* or *Tenebrio molitor*). The ultimate reduction of the production line consists only of an artificial diet for direct parasitoid/predator rearing. Mass rearing entomophagous insects on artificial media, first suggested 60 years ago, holds the promise to increase the ease and flexibility of insect production, including automation of procedures, and to reduce cost. The early and subsequent efforts at developing artificial diets have extensively been reviewed.[1–3] The basic qualitative nutritional requirements of parasitoids and predators are similar to those of free-living insects. But the very fast growth of some parasitoids such as tachinid larvae requires a perfectly well-balanced diet[4] to minimize intermediate metabolism and toxic waste product accumulation.

Essentially, two types of artificial diets can be distinguished: Those including and those excluding insect components. The availability of media without insect components offers a greater independence from insect hosts/prey, even if in some countries insect components are cheap and easily available by-products, e.g., from silk production in Asia or South America.[5] In diets containing insect additives, such varied components as hemolymph, body tissue extract, bee brood extract or powder, egg juice, or homogenate of the natural host have been used. Products of insect cell culture have also been incorporated into diets as host factors. The composition of most media for in vitro rearing of *Trichogramma* egg parasitoids is based on lepidopterous hemolymph.[6] Media for the tachinid fly *Exorista larvarum*, the chalcid wasp *Brachymeria intermedia*, and the ichneumonid wasp *Diapetimorpha introita* contain various insect components. Bee extracts or bee brood have been commonly added in diets for predatory coccinellids.[1,5] Only few diets devoid of insect additives are composed of ingredients that are fully chemically defined in their composition and structure. Besides proteins or protein hydrolysates, most of such diets contain crude or complex components, e.g., hen's egg yolk, chicken embryo extract, calf serum, cow's milk, yeast extract or hydrolysate, meat or liver extract, or plant oils. Beef or pork meat and liver have extensively been used as basic components of diets for feeding coccinellids and several predatory heteropterans.[1,5]

Successes and Failures with Artificial Diets

Both biochemical and physical aspects determine the success of an artificial diet. Artificial diets should be nutritionally adequate to support development and reproduction of an insect and should be formulated in such a manner that the medium is easily recognized and accepted for feeding or oviposition; the food should be readily ingested, digested, and absorbed.[7] For parasitoids, the diet must also allow the growing larvae to satisfy other physiological needs like respiration and excretion without diet spoiling. The best results on artificial media were obtained with idiobiontic parasitoids such as egg or pupal parasitoids and with polyphagous predators. Different tachinid species were also successfully grown in vitro, but the koinobiontic Hymenoptera appear the most difficult group to be reared in vitro, probably because of a close relationship with their living host that supplies them with crucial growth factors. Ectoparasitoids are generally easier to culture in vitro than endoparasitoids for which the diet is also the living environment of the immature stages.[2] Several predatory insects have been reared for successive generations on artificial diets, including heteropterans (e.g., *Geocoris punctipes, Orius laevigatus, Podisus maculiventris*), coccinellids (e.g., *Coleomegilla maculata, Harmonia axyridis),* and chrysopids (e.g., *Chrysoperla carnea, Chrysoperla rufilabris*).[3]

Artificial rearing of natural enemies has mostly remained at an experimental level, and the practical experience with natural enemies produced in artificial conditions has remained quite limited. Wasps of the genus *Trichogramma* reared on factitious host eggs are the most common agents used worldwide in biological control in many field crops and forests. In China, *Trichogramma* spp. and *Anastatus* spp. produced on a large scale in artificial host eggs have been released on thousands of hectares of different crops with a parasitization rate above 80%, leading to an effective pest control level equal to that of naturally reared parasitoids.[5] In the U.S.A., field tests with encouraging first results were conducted using the pteromalid parasitoid *Catolaccus grandis* reared for successive generations on artificial diet for the control of the cotton boll weevil *Anthonomus* grandis.[5] Since the late 1990s, biocontrol companies in the U.S.A. and Europe have started producing a number of natural enemies (partially) on artificial diets.

Quality Control of Natural Enemies Produced on Artificial Diets

Long-term rearing on artificial diets could lead to genetic bottleneck effects inducing high selection pressure on the entomophages and possible reduction of their effectiveness. Periodic population renewals from nature may circumvent this drawback. The use of natural enemies in augmentative biological control requires a reliable mass production of good quality insects. Therefore, quality control is a key element for the efficiency and the long-term viability of biological control. The quality control procedures developed for in vivo production of entomophages could be recommended as a first approach for in vitro production.[8] Many parameters can be used as quality criteria. Size, weight, life cycle duration, survival rate, and especially fecundity, longevity, and predation/parasitization efficiency are the most relevant characters.[5] Besides its value as a quality criterion, the biochemical composition (based upon carcass analyses) of the insects produced on artificial diets may be a powerful tool for improving the composition and performance of the diets through the detection of excess or deficiency of some nutrients. Often, different criteria are closely linked; hence, the quality control process may be simplified if one easily measured parameter can be used to predict another one that is more complex or time consuming to determine (e.g., fecundity). Arguably, excellent field performance of the artificially produced natural enemy against the target pest remains the ultimate quality criterion. However, quality assessments of artificially reared natural enemies have mostly been performed at a laboratory scale or in semifield conditions, and only rarely so in practical field conditions.

Conclusions

At present, rearing systems using natural or factitious foods remain the only effective way for industrial production of most entomophagous insects and mites. However, success achieved for a restricted number of species of parasitoids (e.g., *Trichogramma* spp., *Exorista larvarum, Catolaccus grandis)* and predators (e.g., *Orius* spp., *Geocoris punctipes, Chrysoperla* spp., *Harmonia axyridis)* has prompted producers to increasingly incorporate artificial diets into their mass rearing systems. Further behavioral and physiological investigations may lead to significant improvements in artificial rearing through a better knowledge of the host-parasitoid and predator-prey relationships. Besides an easier mechanization of the production line, the use of artificial diets opens new possibilities for preimaginal conditioning of parasitoids/predators to targeted hosts/prey by adding specific chemicals in their food. Artificial diets also seem the only way of mass rearing for some middle-sized egg parasitoids (Encyrtidae, Eulophidae, Eupelmidae, Scelionidae to name a few) that are promising pest control agents but are unable to develop normally in the small lepidopteran substitution host eggs commonly used nowadays *(Ephestia kuehniella, Sitotroga cerealella).*

References

1. Thompson, S.N. Nutrition and culture of entomophagous insects. Annu. Rev. Entomol. **1999**, *44*, 561–592.
2. Grenier, S.; Greany, P.D.; Cohen, A.C. Potential for mass release of insect parasitoids and predators through development of artificial culture techniques. In *Pest Management in the Subtropics: Biological Control—A Florida Perspective;* Rosen, D., Bennett, F.D., Capinera, J.L., Eds.; Intercept: Andover, U.K., 1994; 181–205.
3. Thompson, S.N.; Hagen, K.S. Nutrition of entomophagous insects and other arthropods. In *Handbook of Biological Control: Principles and Applications;* Bellows, T.S., Fisher, T.W., Eds.; Academic Press: San Diego, CA, 1999; 594–652.
4. Grenier, S.; Delobel, B.; Bonnot, G. Physiological interactions between endoparasitic insects and their hosts— Physiological considerations of importance to the success of in vitro culture: an overview. J. Insect Physiol. **1986**, *32* (4), 403–408.

5. Grenier, S.; De Clercq, P. Comparison of artificially vs. naturally reared natural enemies and their potential for use in biological control. In *Quality Control and Production of Biological Control Agents: Theory and Testing Procedures*; van Lenteren, J.C., Ed.; CABI Publishing: Wallingford, U.K., 2003; 115–131.
6. Grenier, S. Rearing of *Trichogramma* and other egg parasit- oids on artificial diets. In *Biological Control with Egg Para- sitoids;* Wajnberg, E., Hassan, S.A., Eds.; CAB International: Wallingford, U.K., 1994; 73–92.
7. Cohen, A.C. *Insect Diets—Science and Technology;* CRC Press: Boca Raton, U.S.A., 2003.
8. van Lenteren, J.C.; Hale, A.; Klapwijk, J.N.; van Schelt, J.; Steinberg, S. Guidelines for quality control of commercially produced natural enemies. In *Quality Control and Production of Biological Control Agents: Theory and Testing Procedures;* van Lenteren, J.C., Ed.; CABI Publishing: Wallingford, U.K., 2003; 265–303.

IV

DIA: Diagnostic Tools: Monitoring, Ecological Modeling, Ecological Indicators, and Ecological Services

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

16

Animals: Toxicological Evaluation

Vera Lucia S.S.
de Castro

Introduction

Developmental toxicants may be a chemical, microorganism, physical agent, or deficiency state that alters the morphology or a physiological process of a developing organism, both before and after birth. Exposure of the developing embryo or fetus to some environmental agents like gamma irradiation and thalidomide is known to produce anatomical anomalies leading to in utero death or structural birth defects. Developmental toxicology studies the causes, mechanisms, manifestation, and prevention of developmental deviations produced by developmental toxicants. Several environmental agents are established as causing developmental toxicity in humans, while many others are suspected of causing developmental toxicity in humans on the basis of data from experimental animal studies.

Nowadays, all the possible manifestations of developmental toxicity (death, structural abnormalities, growth alterations, and behavioral and functional deficits) are of concern, while in the past there has been a tendency to consider only malformations and death as end points of concern. Developmental toxicity usually results from prenatal exposures to toxicants experienced by the mother, but it can also result from paternal exposures, e.g., in rural workers exposed to pesticides. Effects can include birth defects, reduced body weight at birth, growth and developmental retardation, organ toxicity, death, abortion, and functional dysfunctions. It can also lead to behavioral deficits that become manifest as the organism develops since the chemicals can impair postnatal development up to pubertal development. Further considerations related to the efforts to assemble an internationally harmonized source of common nomenclature for use in describing observations of fetal and neonatal external, visceral, and skeletal abnormalities can be found in Makris.[1]

In May 2001, more than 90 nations adopted the Stockholm Convention on Persistent Organic Pollutants (POPs), with significant contributions from non-governmental organizations, trade unions, and private companies. The Stockholm Convention is a global treaty created to protect human health and the environment from chemicals that remain intact in the environment for long periods, become widely distributed geographically, accumulate in the fatty tissue of humans and wildlife, and have adverse effects to human health or to the environment. Exposure to POPs can lead to serious health effects, including certain cancers, birth defects, dysfunctional immune and reproductive systems, greater susceptibility to disease, and even diminished intelligence. Given their long-range transport, no one government acting alone can protect is citizens or its environment from POPs. In response to this global problem, the Stockholm Convention entered into force in 2004, requiring parties to take measures to eliminate or reduce the release of POPs into the environment. The convention is administered by the United Nations Environment Programme and is based in Geneva, Switzerland. The substances covered initially are eight pesticides (aldrin, chlordane, DDT, dieldrin, endrin, heptachlor, mirex, and toxaphene), two industrial chemicals (hexachlorobenzene and polychlorinated biphenyls), and two POP by-products (dioxins and furans). In 2009, the Conference of the Parties decided to undertake a work program to provide guidance to parties on how best to restrict and eliminate nine newly listed POPs and invited parties to support work on the evaluation of alternatives and other work related to the restriction and elimination of these new POPs [α-hexachlorocyclohexane, ß-hexachlorocyclohexane, lindane, pentachlorobenzene, perfluorooctane sulfonic acid (PFOS), PFOS salts, perfluorooctane sulfonyl fluoride, tetrabromodiphenyl ether, and pentabromodiphenyl ether (commercial pentabromodiphenyl ether)] (http://chm.pops.int/default.aspx).[2]

Exposure of the developing embryo or fetus to some environmental agents like gamma irradiation and thalidomide is known to produce anatomical anomalies leading to in utero death or structural birth defects, commonly termed teratogenesis. Perhaps less well appreciated is that such environmental exposures also can cause functional disorders that persist postnatally and into adult life. The spectrum of such postnatal consequences is growing, and more recently is thought to include disorders of the immune system, brain function, obesity, and diseases such as diabetes and cancer, to name a few.[3]

Importance of Oxidative Stress

Xenobiotics such as phenytoin and benzo[*a*]pyrene can be bioactivated by enzymes like the cytochromes P450 (CYPs); however, the developing embryo and fetus have relatively low levels of most CYP isozymes. The xenobiotic bioactivation to a free radical intermediate by enzymes associated with peroxidase activities within the embryo or fetus can be a critical determinant of teratogenesis. In the developing embryo and fetus, enhanced formation of reactive oxygen species by xenobiotics may adversely alter development by oxidatively damaging cellular lipids, proteins, and DNA, and/or by altering signal transduction. The postnatal consequences may include an array of birth defects, postnatal functional deficits, and diseases.[3]

Also, estrogen-like substances, such as several organochlorine pesticides, have been demonstrated to induce defeminization, miscarriages, malformations, and transplacental carcinogenesis. This variety of effects results from the interference of these substances with the metabolism of steroid and protein hormones, therefore altering a whole spectrum of complex developmental functions.[4]

Importance of Toxic Effects on Nervous System Development

Only about 200 chemicals out of more than 80,000 registered with the U.S. Environmental Protection Agency (EPA) have undergone extensive neurotoxicity testing, and many chemicals found in consumer goods are not required to undergo any neurodevelopmental testing. The cumulative effects of co-contaminants and the difficulties in analyzing biomarkers of exposure in human tissues have complicated comprehensive risk assessment. Furthermore, population-based studies that measure subtle effects on neurobehavioral outcomes are challenging to interpret and costly to conduct.[5]

The effects observed following developmental exposure may be different both quantitatively and qualitatively from adult exposure because of the potential to affect processes in the developing child that have no parallel process in adults. Central nervous system development consists of a series of processes that occur in sequence and are dependent on each other, such that interference with one stage may also affect later stages of development. This makes the timing of a potential environmental neurotoxicant a critical parameter in the risk for subsequent neurologic effects. The sequence includes proliferation, migration, differentiation, synaptogenesis, apoptosis, and myelination.

There are multiple windows of vulnerability during which environmental exposures can interfere with normal development. For many developmental toxicants, there is a spectrum of adverse outcome depending on dose timing of exposure, maternal and fetal susceptibility, and interactions with other environmental factors.[6]

Therefore, environmental exposure to a toxic agent that affects synaptogenesis, such as lead, may affect brain areas differentially depending on timing of exposure. Since different brain areas develop on different time lines during prenatal and postnatal life, an environmental neurotoxic agent may produce impairment in different functional domains depending on the time of exposure. For example, the same exposure at different points in development could result in an adverse effect on motor systems versus memory or executive functions. Similarly, exposures at different concentrations or for different lengths of time could potentially produce differential effects. Therefore, the constellation of observed effects should not be expected to be the same in different children exposed to the same neurotoxic agent.[7]

Concern on Developmental Impairment in Children

The findings of the some studies imply that children's exposure to pesticides may bring about impairments in their neurobehavioral development. The long-term neurotoxicity risks caused by prenatal exposures to pesticides are sometimes unclear. Effective control of exposure is complicated by variable exposure sources and variable contaminant levels in food and environment. This awareness has also been extended to effect(s) of toxic contaminants on breastfeeding women and their children.[8]

The information deriving from epidemiological studies indicate a need to increase awareness among people and children exposed to pesticides about the association between the use of pesticides and neurodevelopmental impairments. There are modest epidemiological evidence on occupational exposures of female workers to industrial chemicals and the consequences in regard to the child's neurodevelopment. The majority of the occupational studies identified aimed to assess organic solvents and organophosphate pesticide effects in the offspring, and neurobehavioral impairments were reported. In some reports, however, the evidence suffers from a variety of shortcomings and sources of imprecision. These problems would tend to cause an underestimation of the true extent of the risks. Due to the vulnerability of the brain during early development, a precautionary approach to neurodevelopmental toxicity needs to be applied in occupational health.[9,10]

With increasing evidence of the high prevalence of pesticide use and the considerable risk it poses to children, it is of concern that there has been little research into the health implications of household pesticide use. Children are exposed to pesticides in various ways, not only environmentally, but also through food and through use of pesticides at home and in public areas. The exposure depends on a large variety of factors related to chemical characteristics and use, and children's activities. In spite of its potential health and environmental risks and contribution to agribusiness, the use of agricultural chemicals for yard care has not been well studied. The probability that a household chooses a mix of do-it-yourself and hired applications of synthetic chemicals increases with income, age, and the presence of preschoolers.[11]

Also, children who live in farming communities are furthermore exposed to both agricultural and household pesticides. Farmworkers bring home pesticide residues on their clothing, boots, and skin, placing other household members at risk, particularly children.[12] Recent studies on in utero exposure to the organochlorine pesticide dichlorodiphenyltrichloroethane and its breakdown product, dichlorodiphenyltrichloroethane, indicate that exposure is associated with poorer infant (6 mo and older) and child neurodevelopment depending on the end point evaluated. Research on organophosphate pesticide exposure and neurodevelopment also suggests some negative association of exposure and neurodevelopment at certain ages. About abnormal reflexes in neonates and in young children (2–3 years), two separate studies observed an increase in maternally reported pervasive developmental disorder with increased levels of organophosphate exposure.[13]

For example, children whose mothers worked in the flower industry during pregnancy scored lower on communication and fine motor skills, and had higher odds of having poor visual acuity, compared with children whose mothers did not work in the flower industry during pregnancy, after adjusting for potential confounders. These facts showed that maternal occupation in the cut-flower industry during pregnancy may be associated with delayed neurobehavioral development of children aged 3–23 mo. However, possible hazards associated with working in the flower industry during pregnancy include pesticide exposure, exhaustion, and job stress.[14]

The increased emphasis on children's exposures to pesticides and other organic pollutants has led to a surge in recent years in the number of research studies aimed at this specific susceptible population. Continuous strong investment in research plus strong preventive action by the government is required for further progress in environmental pediatrics and for better control of the diseases caused in children by environmental toxic exposures. This research will have high costs and demanding long-term multiyear studies.[15]

Developmental Animal Experimental Studies

Before 1960, governmental recommendations for the assessment of chemical effects on the reproductive cycle involved limited animal testing. During the early 1960s, the thalidomide disaster evidenced, on the one hand, the greater vulnerability of the embryo and fetus, and on the other, that the complexity of the mother-child unity warranted special consideration. Thalidomide was used as a sedative drug in the 1950s, also in pregnant women. Later, it was found that it could cause malformations in newborns of mothers who ingested the drug during the sensitive period.[16] This disaster fostered the establishment of formal laboratory animal-testing procedures for assessing fetal development.

In contrast to most other toxicological tests, developmental studies are usually required in rodent and nonrodent subjects. One of the reasons for this requirement is the thalidomide disaster. When the developmental toxicity of thalidomide was studied in experimental animals, large interspecies differences were found in effective doses and in the types of effects. However, the discrimination between direct and indirect (i.e., as a consequence of maternal toxicity) developmental effects was often doubtful, and is one of the factors that could explain the apparent differences between species.[16]

Despite the limitations, animals can be useful predictors of chemical hazards to humans. A specific animal model might be chosen for any conjunction of widely varying reasons. Accessibility of embryos, cost of acquiring or maintaining animals, availability of genomic analyses or probes, and/or close similarity to human physiology might factor in the design of a laboratory experiment. Growth and development are compressed into a shorter period in animals, which makes interpretation of animal testing inherently more difficult. Each experimental species has its own advantages and the use of laboratory animal models is based on diverse practical grounds arising from an assumption of generalizability across species. The conjunction of evolutionary and developmental biology shows that the timing and sequence of early events in brain development are remarkably conserved across mammals.[17] During mammalian development, the fetal organism is exposed to its own gonadal hormones, placental steroids, and maternal hormones that may cross the placental barrier.[18]

Manifestations of developmental toxicity observed in humans are not always reproduced in experimental animals, and in general there is at least one experimental species that mimics the types of effects seen in humans. The fact that every species may not react in the same way could be due to species-specific differences in critical periods, differences in timing of exposure, metabolism, developmental patterns, placentation, or mechanisms of action.[6]

The vast majority of the laboratory studies on the developmental effects of environmental contaminants use the pregnant or lactating animal as the conduit to deliver the contaminant to the developing offspring. The traditional approach in developmental toxicology adopts a linear perspective for the interpretation of the effects of contaminants on the offspring as a particular one is given to the mother in order to expose the fetus in utero or postnatally via lactation. Variations of this approach may manipulate time of exposure or use cross-fostering strategies to separate the effects of in utero from those of lactational exposure.

Maternal behavior is not always monitored during the treatment period, even though there is evidence of the effects of these compounds on adult behaviors, particularly on behaviors that are sensitive to hormonal manipulations. As is the case with all mammals, maternal care in female rodents comprises very specific behaviors that help ensure the survival of the offspring by providing nourishment, warmth, sensory stimulation, cleaning, and protection. Maternal behavior begins even before parturition as the dam builds a nest in order to provide warmth and protection for the coming offspring. The fact that many developmental outcomes are determined or modulated by the amount and quality of maternal care raises the question of the importance of possible changes in maternal behavior in determining the consequences of their exposure during early development.[19] There are some guidelines proposed in the literature in order to evaluate the maternal behavior and its effects on litter development.[20,21]

Reproductive and Developmental Protocols

Developmental toxicology bioassays are designed to identify agents with the potential to induce adverse effects and include dose levels that induce maternal toxicity. In reproduction toxicity studies, the determination of the high dose is important since changes in body weight are often used as an index of toxicity. The highest dose level should be chosen with the aim to induce some parental toxicity (e.g., clinical signs, decreased body weight gain, not more than 10%) and/or evidence of toxicity in a target organ. Consequently, the knowledge on maternal (and to some degree, paternal) toxicity is important as a natural limiter to prevent underdosing. A comparison between doses causing effects in adults and offspring can also be used, although a direct comparison is difficult since the level of observation applied in offspring is often much higher than in adults. It is also useful in order to obtain information on the influence of pregnancy and/or lactation on the susceptibility to a test compound. However, if dosing was high enough to cause maternal toxicity, these doses often also cause some effects in the offspring.[22]

Prenatal developmental toxicity studies are designed to provide general information concerning the effects of exposure to the pregnant test animal on the developing organism. Although exposures during a typical guideline prenatal developmental toxicity study are designed to include either the entire period of gestation or limited species-specific gestation periods, developmental end points are considered to be an integral concern in the assessment of potential health effects from continuous lifetime exposures to a toxicant. Pregnancy and fetal development are thus considered to represent a potentially susceptible life stage that should be considered in lifetime or chronic assessments. If studies and information on reproductive or developmental toxicity are absent in a health-effect assessment, specific uncertainty factors (e.g., database) may be applied to the final point of departure used in risk calculations. It is also well established, however, that developmental toxicity may occur in response to single exposure, such as during specific developmental windows of susceptibility. Whereas this circumstance does not influence the relevance of typical guideline developmental studies to the evaluation of chronic or lifetime assessments of health effects, it does signify that developmental end points observed in these repeated dose studies are relevant in health-effect assessments of shorter-term exposures, including acute exposures.[23] Results from developmental toxicology bioassays have significant predictive value in identifying potential health risks to the human embryo/fetus.[24]

Data Quality Control of Experimental Studies

To accomplish good experimental planning, some points before the beginning of the study should be observed for adequate data interpretation in view of the experimental delineation.[25,26] Animals to be used in laboratory research should experience an acceptable welfare (for ethical reasons) and show normal behavioral and physiological reaction patterns to guarantee the quality of research.[27] In this sense, the performing laboratory should maintain a historical control database to track any changes in the data over time in the animals and/or in the equipment. The value of historical data depends on its quality and its reliability at the side of contemporary controls.

International Protocols

Different international organizations have developed protocols for testing the reproductive and developmental toxicity of chemicals. The information produced by them, related to toxic effects produced by environmental pollutants, is very useful worldwide. Besides, new protocols are being evaluated viewing to reduce the number of animals used and to improve the predictability for human health hazards identification.

Developmental study methodology has been extensively reviewed and evaluated over the last 25 years. This has included the conduct of a number of meetings and collaborative studies involving experts from academic, industry, regulatory, and public interest groups. For example, in recent years, the International Life Sciences Institute (ILSI), under a cooperative agreement with the EPA, established a working group of scientists from government, industry, and academia, to discuss developmental neurotoxicity test protocol ending with a public workshop in which occasion the conclusions of the working group were presented.[28]

Food Quality Protection Act

Developmental and reproductive toxicity testing protocols such as those recommended by the EPA, Food and Drug Administration, and Organization for Economic Cooperation and Development are useful for characterizing toxicity in developing animals and for assessing risks to children that might arise from in utero and postnatal exposures.[7] However, there is a global interest in reducing, refining, and replacing (3Rs) the use of animals in research.[29]

The Food Quality Protection Act (FQPA) of 1996 enactment by U.S. Congress amended the Federal Insecticide, Fungicide, and Rodenticide Act and the Federal Food, Drug, and Cosmetic Act by fundamentally changing the way the EPA regulates pesticides. The major requirements of the FQPA include stricter safety standards, especially for infants and children, and a complete reassessment of all existing pesticide tolerances. They include an additional safety factor to account for developmental risks and incomplete data when considering a pesticide's effect on infants and children, and any special sensitivity and exposure to pesticide chemicals that infants and children may have (http://www.epa.gov/opp00001/regulating/laws/fqpa/).

The FQPA mandated that all pesticides in the United States undergo re-registration with a focus on reducing cumulative risk of exposure to pesticides sharing a common mode of action. Enforcement of FQPA has resulted in the modification of use patterns and removal (or pending removal) of many organophosphate insecticides that had previously seen wide use.[30]

The FQPA also requires the EPA to consider the cumulative effects of exposure to pesticides having a common mechanism of toxicity, considering for a cumulative risk assessment the exposure to all chemicals that act by a common mechanism of toxicity, as well as the exposure to each chemical via various routes and sources in an aggregate risk assessment. To support the grouping of different chemicals together for purposes of cumulative risk assessment, there must be sufficient evidence to support a common adverse effect that is associated with a common mechanism of action in specific target tissues. However, the criteria that are required to establish a common mechanism of toxicity with a specific toxic effect have not always been achieved for various pesticides as the common mechanism of toxicity of organophosphorus and carbamate insecticides (inhibition of acetylcholinesterase activity) that can be associated with adverse effects (cholinergic signs of intoxication). For example, a determination of common mechanism of toxicity in mammals is complicated by the number of potential biological target sites and effects expressed by various pyrethroid insecticides on these targets. Probably, the differences of action on neuronal ion channels among the pyrethroid insecticides contribute to the diversity of neurologic and behavioral manifestations of acute toxicity that are evident in the whole animal.[31]

European Union's Reach Legislation

REACH is the European Community Regulation on chemicals and their safe use. It deals with the Registration, Evaluation, Authorization, and Restriction of Chemicals. The law entered into force on June 1, 2007. The REACH regulation places greater responsibility on industry to manage the risks from chemicals and to provide safety information on the substances. Manufacturers and importers are required to gather information on the properties of their chemical substances, which will allow their safe handling, and to register the information in a central database in Helsinki (http://ec.europa.eu/environment/chemicals/reach/reach_intro.htm). It transfers responsibility for risk assessment from government to the manufacturers and importers, and includes downstream uses in the registration and management process. It introduces authorization and restriction procedures for the most hazardous chemicals and creates a new European Chemicals Agency. The legal permission to market products is conditional on the firms testing them for toxicity. If firms do not provide data required by the program, their products will not be permitted to enter (or remain in) the market. This program holds some promise for detecting developmental toxicants before they enter commerce and cause adverse effects. Whether this will work for subclinical neurotoxic and other developmental effects depends on tests the European Union requires. The REACH testing strategy is to require fewer tests for products produced in lesser amounts and to require more tests and more detailed tests as the production volume increases.[32]

In particular, large numbers of industrial chemicals are unlikely to be tested under this paradigm, and there is no specific requirement for developmental neurotoxicological testing under the new European Union law governing chemical regulation (REACH), which was passed in 2006 and went into effect in 2007. In response to the need for broader screening for developmental neurotoxicity, efforts are under way to develop additional developmental neurotoxicity screening paradigms. Additional efforts will be needed to focus on identifying possible chemical class- specific targets and biomarkers of effect, and on ways to differentiate normal variability in response from changes that are adverse.[33]

Perspectives

There is a need to expand risk assessment paradigms to evaluate exposures relevant to children from preconception to adolescence, taking into account the specific susceptibilities at each developmental stage. Risk assessment approaches for exposures in children must be linked to life stages. Establishing causal links between specific environmental exposures and complex, multifactorial health outcomes is difficult and challenging, particularly in children. For children, the stage in their development when the exposure occurs may be just as important as the magnitude of exposure. Very few studies have characterized exposures during different developmental stages. Some examples of health effects resulting from developmental exposures include those observed prenatally and at birth (e.g., miscarriage, stillbirth, low birth weight, birth defects), in young children (e.g., infant mortality, asthma, neurobehavioral and immune impairment), and in adolescents (e.g., precocious or delayed puberty). Emerging evidence suggests that an increased risk of certain diseases in adults (e.g., cancer, heart disease) can result in part from exposures to certain environmental chemicals during childhood. Advancing technology and new methodologies now offer promise for capturing exposures during these critical windows. This will enable investigators to detect conceptions early and estimate the potential competing risk of early embryonic mortality when considering children's health outcomes that are conditional on survival during the embryonic and fetal periods.[34]

Furthermore, emerging technologies such as gene expression, electrical activity measurements, and meta-bonomics have been identified as promising tools for evaluating neurotoxicity. In a combination with other assays, the in vitro approach could be included into a developmental neurotoxicological intelligent testing strategy to speed up the process of developmental neurotoxicity evaluation mainly by initial prioritization of chemicals with developmental neurotoxicity potential for further testing. Also, emerging nano/microtechnologies (piezoelectric spotting and microcontact printing of different biomolecules to create protein microarrays) can be used to promote cell differentiation and make the model suitable for developmental neurotoxicity screening.[35]

Developmental Toxic Effects on Nontarget Organisms in Environment

In addition to their potential role as human reproductive toxicants, pesticides are also implicated in reproductive failure of wildlife species exposed to pesticide sprays and residues. One main example of the ecological consequences of teratogenic pesticides is related to the organochlorines, which, besides inducing malformations in embryos, cause calcification problems in eggshells and impair reproduction in several wild bird species. This problem still remains in numerous areas, due to organochlorine residue accumulation through food chains and to wild populations exposed to organochlorine-contaminated sites.[36] Also, organochlorine pesticides can be maternally transferred to the developing eggs of alligators. This maternal exposure is associated with reduced clutch success and increased embryonic mortality.[37]

Toxicity assays are available for the evaluation of pesticide impact on wildlife or their surrogates, and many compounds have been shown to cause teratogenesis in fish, amphibian, avian, and mammalian species. Some chemicals that have been detected in the environment may be disrupting of both target and non-target systems in exposed populations of wildlife and fish.[38] Environmental compounds can also interfere with the endocrine systems of wildlife. Surface waters are the main sinks of endocrine disrupters, which are mainly of anthropogenic origin. Thus, aquatic organisms, especially lower vertebrates such as fish and amphibians, are the main potential targets for endocrine disrupters at direct or indirect risk via ingestion and accumulation of endocrine disrupters, direct exposure, or via the food chain. The impact of these compounds on reproductive biology can be mediated through four principal mechanisms of action: estrogenic, anti-estrogenic, androgenic, and anti-androgenic.[39]

Furthermore, population studies have revealed disruptions in crustacean growth, molting, sexual development, and recruitment that are indicative of environmental endocrine disruption. However, environmental factors other than pollution (i.e., temperature, parasitism) also can elicit these effects, and definitive causal relationships between endocrine disruption in field populations of crustaceans and chemical pollution is generally lacking.[40]

Amphibians are considered reliable indicators of environmental quality, in particular due to their biphasic life (aquatic and terrestrial) and semipermeable skin. These vertebrates are sensitive to a great number of pollutants dispersed in the environment, such as pesticides, heavy metals, and polychlorinated biphenyls. Field studies on frogs from polluted and reference sites have provided information on the effects of chronic exposure to contaminants.[41]

Despite the fact that almost all environmental chemical exposure is to mixtures, the current understanding of environmental health risks is based almost entirely on the evaluation of chemicals studied in isolation. Consequently, it is essential to develop and validate methods to accurately predict effects of endocrine-disrupting mixtures beyond the individual exposure to a single chemical, in order to protect humans and wildlife from the risk associated with potentially cumulative effects of these mixtures.[36,42] Its focus should be on the biological system or the target tissue rather than on the mechanism of toxicity or even a single signaling pathway.[43]

Although the importance of multiple stressors is widely recognized in aquatic ecotoxicology, pesticide mixture studies pose some major challenges such as experimental design difficulties (e.g., near-insurmountable factorial complexity for large numbers of chemicals), poorly understood pathways for chemical interaction, potential differences in response among species, and the need for more sophisticated statistical tools for analyzing complex data.[44]

Conclusion

The ability of a species to reproduce successfully requires the careful orchestration of developmental processes during critical time points, particularly the late embryonic and early postnatal periods. Standard developmental toxicology bioassays are designed to identify agents with the potential to induce adverse effects. Government agencies that regulate the use of pesticides and various industrial chemicals evaluate the toxicity of these agents to the developing embryo/fetus as an integral part of the testing protocol used to assess potential dangers to the public. However, an adequate experimental delineation is important to the data interpretation.

Humans and other non-target organisms are exposed to a mixture of chemicals. Various chemicals may target the organism's development during the same critical developmental period. Although developmental studies have been historically conducted on a chemical-by-chemical basis, the interest on considering cumulative risks of chemicals is growing. Chemicals as pesticides represent a risk for the reproduction and development of children and non-target organisms of terrestrial and aquatic ecosystems.

References

1. Makris, S.L.; Solomon, H.M.; Clark, R.; Shiota, K.; Barbellion, S.; Buschmannf, J.; Emag, M.; Fujiwarah, M.; Grotei, K.; Hazeldenj, K.P.; Hewk, K.; Horimotol, M.; Ooshimam, Y.; Parkinsonn, M.; Wiseo, L.D. Terminology of developmental abnormalities in common laboratory mammals (version 2). Reprod. Toxicol. **2009,** *28,* 371–434.
2. Rodan, B.D.; Pennington, D.W.; Eckley, N.; Boethlin, R.S. Screening for persistent organic pollutants: Techniques to provide a scientific basis for POPs criteria in international negotiations. Environ. Sci. Technol. **1999**, *33,* 3482–3488.
3. Wells, P.G.; McCallum, G.P.; Chen, C.S.; Henderson, J.T.; Lee, C.J.J.; Perstin, J.; Preston, T.J.; Wiley, M.J.; Wong, A.W. Oxidative stress in developmental origins of disease: Teratogenesis, neurodevelopmental deficits, and cancer. Toxicol. Sci. **2009**, *108* (1), 4–18.
4. Chernoff, N. The reproductive toxicology of pesticides. In *Toxicology of Pesticides: Experimental, Clinical and Regulatory Perspectives;* Costa, L.G., Galli, C.L., Murphy, S.D., Eds.; NATO ASI-Cell Biology; Springer-Verlag: Berlin, Germany, 1987; H13, 109–123.
5. Miodovnik, A. Environmental neurotoxicants and developing brain. Mt. Sinai J. Med. **2011**, *78,* 58–77.
6. U.S. EPA. *Guidelines for Developmental Toxicity Risk* Assessment; U.S. Environmental Protection Agency, Risk Assessment Forum: Washington, D.C.; EPA/600/FR-91/001, 1991.
7. Mendola, P.; Selevan, S.G.; Gutter, S.; Rice, D. Environmental factors associated with a spectrum of neurodevelopmental deficits. Mental retardation and developmental disabilities. Res. Rev. **2002**, *8,* 188–197.

8. Lakind, J.S.; Berlin, C.M.; Naiman, D.Q. Infant exposure to chemicals in breast milk in the United States: What we need to learn from a breast milk monitoring program. Environ. Health Perspect. **2001**, *109,* 75–88.
9. Jurewicz, J.; Hanke, W. Prenatal and childhood exposure to pesticides and neurobehavioral development: Review of epidemiological studies. Int. J. Occup. Med. Environ. Health **2008**, *21* (2), 121–132.
10. Julvez, J.; Grandjean, P. Neurodevelopmental toxicity risks due to occupational exposure to industrial chemicals during pregnancy. Ind. Health **2009**, *47,* 459–468.
11. Templeton, S.R.; Zilberman, D.; Yoo, S.J.; Dabalen, A.L. Household use of agricultural chemicals for soil-pest management and own labor for yard work. Environ. Resour. Econ. **2008**, *40,* 91–108.
12. Strong, L.L.; Starks, H.E.; Meischke, H.; Thompson, B. Perspectives of mothers in farmworker households on reducing the take-home pathway of pesticide exposure. Health Educ. Behav. **2009**, *36* (5), 915–929.
13. Rosas, L.G.; Eskenazi, B. Pesticides and child neurodevelopment. Curr. Opin. Pediatr. **2008**, *20* (2), 191–197.
14. Handal, A.J.; Harlow, S.D.; Breilh, J.; Lozoff, B. Occupational exposure to pesticides during pregnancy and neurobehavioral development of infants and toddlers. Epidemiology **2008**, *19* (6), 851–859.
15. Landrigan, P.J.; Miodovnik, A. Children's health and the environment: An overview. Mt. Sinai J. Med. **2011**, *78,* 1–10.
16. Janer, G.; Slob, W.; Hakkert, B.C.; Vermeire, T.; Piersma, A.H. A retrospective analysis of developmental toxicity studies in rat and rabbit: What is the added value of the rabbit as an additional test species? Regul. Toxicol. Pharmacol. **2008**, *50,* 206–217.
17. Clancy, B.; Finlay, B.L.; Darlington, R.B.; Anand, K.J.S. Extrapolating brain development from experimental species to humans. NeuroToxicology **2007**, *28,* 931–937.
18. Gore, A.C. Developmental programming and endocrine disruptor effects on reproductive neuroendocrine systems. Front. Neuroendocrinol. **2008**, *29*, 358–374.
19. Cummings, J.A.; Clemens, L.G.; Nunez, A.A. Mother counts: How effects of environmental contaminants on maternal care could affect the offspring and future generations. Front. Neuroendocrinol. **2010**, *31,* 440–451.
20. Champagne, F.; Francis, D.; Mar, A.; Meaney, M. Variations in maternal care in the rat as a mediating influence for the effects of environment on development. Physiol. Behav. **2003**, *79,* 359–371.
21. Slamberová, R.; Charousová, P.; Pometlová, M. Maternal behavior is impaired by methamphetamine administered during pre-mating, gestation and lactation. Reprod. Toxicol. **2005**, *20,* 103–110.
22. Buschmann, J. Critical aspects in reproductive and developmental toxicity testing of environmental chemicals. Reprod. Toxicol. **2006**, 22, 157–163.
23. Davis, A.; Gift, J.S.; Woodall, G.M.; Narotsky, M.G.; Foureman, G.L. The role of developmental toxicity studies in acute exposure assessments: Analysis of single-day vs. multiple-day exposure regimens. Regul. Toxicol. Pharmacol. **2009**, *54,* 134–142.
24. Chernoff, N.; Rogers, E.H.; Gage, M.I.; Francis, B.M. The relationship of maternal and fetal toxicity in developmental toxicology bioassays with notes on the biological significance of the "no observed adverse effect level". Reprod. Toxicol. **2008**, *25* (2), 192–202.
25. Slikker, Jr., W.; Acuff, K.; Boyes, W.; Chelonis, J.; Crofton, K.; Dearlove, G.; Li, A.; Moser, V.; Newland, C.; Rossi, J.; Schantz, S.; Sette, W.; Sheets, L.; Stanton, M.; Tyl, S.; Sobotka, T. Behavioral test methods workshop. Neurotoxicol. Teratol. **2005**, 27, 417–427.
26. Festing, M.; Altman, D. Guidelines for the design and statistical analysis of experiments using laboratory animals. ILAR J. **2002**, *43* (4), 244–258.
27. Olsson, I.A.S.; Westlund, K. More than numbers matter: The effect of social factors on behaviour and welfare of laboratory rodents and non-human primates. Appl. Anim. Behav. Sci. **2007**, *103,* 229–254.

28. Tyl, R.W.; Crofton, K.; Moretto, A.; Moser, V.; Sheets L. P.; Sobotka, T.J. Identification and interpretation of developmental neurotoxicity effects. A report from the ILSI Research Foundation/Risk Science Institute expert working group on neurodevelopmental endpoints. Neurobehav. Toxicol. **2008**, *30,* 349–381.
29. Matthews, E.J.; Kruhlak, N.L.; Benz, R.D.; Contrera, J.F. A comprehensive model for reproductive and developmental toxicity hazard identification: I. Development of a weight of evidence QSAR database. Regul. Toxicol. Pharmacol. **2007**, *47,* 115–135.
30. Jones, V.P.; Steffan, S.A.; Hull, L.A.; Brunner, J.F.; Biddinger, D.J. Effects of the loss of organophosphate pesticides in the US: Opportunities and needs to improve IPM Programs. Outlooks Pest Manage. **2010**, *21* (4), 161–166.
31. Weiner, M.L.; Nemec, M.; Sheets, L.; Sargent, D.; Breckenridge, C. Comparative functional observational battery study of twelve commercial pyrethroid insecticides in male rats following acute oral exposure. NeuroToxicology **2009**, *30S,* S1–S16.
32. Cranor, C. The legal failure to prevent subclinical developmental toxicity. Basic Clin. Pharmacol. Toxicol. **2008**, *102,* 267–273.
33. Raffaele, K.C.; Rowland, J.; May, B.; Makris, S.L.; Schumacher, K.; Scarano, L.J. The use of developmental neurotoxicity data in pesticide risk assessments. Neurotoxicol. Teratol. **2010**, *32*, 563–572.
34. Louis, G.B.; Damstra, T.; Díaz-Barriga, F.; Faustman, E.; Hass, U.; Kavlock, R.; Kimmel, C.; Kimmel, G.; Krishnan, K.; Luderer, U.; Sheldon, L. Principles for evaluating health risks in children associated with exposure to chemicals. Environ. Health Criteria **2006**, *237,* 1–327.
35. Bal-Price, A.K.; Hogberg, H.T.; Buzanska, L.; Lenas, P.; van Vliet, E.; Hartung T. In vitro developmental neurotoxicity (DNT) testing: Relevant models and endpoints NeuroToxicology **2010**, *31,* 545–554.
36. Hotchkiss, A.K.; Rider, C.V.; Blystone, C.R.; Wilson, V.S.; Hartig, P.C.; Ankley, G.T.; Foster, P.M.; Gray, C.L.; Gray, L.E. Fifteen years after "Wingspread"—Environmental endocrine disrupters and human and wildlife health: Where we are today and where we need to go. Toxicol. Sci. **2008**, *105* (2), 235–259.
37. Rauschenberger, R.H.; Wiebe, J.J.; Buckland, J.E.; Smith, J.T.; Sepulveda, M.S.; Gross, T.S. Achieving environmentally relevant organochlorine pesticide concentrations in eggs through maternal exposure in *Alligator mississippiensis.* Mar. Environ. Res. **2004**, *58,* 851–856.
38. Ankley, G.T.; Brooks, B.W.; Huggett, D.B.; Sumpter, J P. Repeating history: Pharmaceuticals in the environment. Environ. Sci. Technol. **2007**, *41*, 8211–8217.
39. Kloas, W.; Lutz, I. Amphibians as model to study endocrine disrupters. J. Chromatogr. A **2006**, *1130,* 16–27.
40. LeBlanc, G.A. Crustacean endocrine toxicology: A review. Ecotoxicology **2007**, *16,* 61–81.
41. Falfushinska, H.I.; Romanchuk, L.D.; Stolyar O.B. Different responses of biochemical markers in frogs (*Rana ridibunda*) from urban and rural wetlands to the effect of carbamate fungicide. Comp. Biochem. Physiol. C **2008**, *148*, 223–229.
42. Jia, Z.; Misra H.P. Developmental exposure to pesticides zineb and/or endosulfan renders the nigrostriatal dopamine system more susceptible to these environmental chemicals later in life. NeuroToxicology **2007**, *28*, 727–735.
43. Rider, C.V.; Furr, J.R.; Wilson, V.S.; Gray, Jr., L.E. Cumulative effects of in utero administration of mixtures of reproductive toxicants that disrupt common target tissues via diverse mechanisms of toxicity. Int. J. Androl. **2010**, *33*, 443–462.
44. Laetz, C.A.; Baldwin, D.H.; Collier, T.K.; Hebert, V.; Stark, J.D.; Scholz, N.L. The synergistic toxicity of pesticide mixtures: Implications for risk assessment and the conservation of endangered pacific salmon. Environ. Health Perspect. **2009**, *117* (3), 348–353.

17

Bioindicators for Sustainable Agroecosystems

Joji Muramoto
and Stephen
R. Gliessman

Introduction

The growth of larger-scale monocultures that heavily depend on use of synthetic pesticides, chemical fertilizers, and fossil fuel-based agricultural machinery has significantly increased crop yield. However, this has also brought about both direct and indirect negative consequences on ecological (e.g., soil erosion, nitrate and pesticide contamination of groundwater, loss of agrobiodiversity), economic (e.g., increase of production and marketing costs and decrease of the net income of farmers), and social (e.g., loss of agricultural communities due to farm consolidations) sustainability of agriculture.[1,2] Especially, agricultural intensification is one of the major drivers of global biodiversity loss as a result of associated habitat fragmentation, land conversion, and agrochemical applications.[3,4] Since the 1970s, and more so after the 1980s, therefore, various alternative farming practices have been developed in pursuit of sustainable agriculture.[5–12] Farmers around the world have adopted these practices to varying degrees, but evaluation of their success needs to be conducted locally. To implement sustainable agricultural practices and policies, managers and policy makers need tools for assessing changes in agroecosystems in various time and spatial scales. Numerous indicators for agricultural sustainability including physicochemical, socioeconomic, and biological indicators, or bioindicators, have been developed.[13–17] Bioindicators are important because they are direct measures of the desired outcome, i.e., sustained or increased biodiversity. They also are living, dynamic, and active indicators, often responding quickly to the way farming is carried out. As evidence regarding the role of biodiversity in maintaining agroecosystem structure and function accumulates, its role as a bioindicator increases.

This chapter will first briefly review the background, definitions, concepts, and history of bioindicators. Then, we discuss the current status of developing bioindicators with two case studies: Europe and Latin America (shaded coffee systems). The former case represents temperate agroecosystems with 4000 years of farming history. The second case characterizes upland tropical agroecosystems known to occur in regions of the world with some of the highest, yet most threatened, biodiversity. Furthermore, examples of recent studies on use of aquatic organisms as bioindicators for agricultural sustainability and an example of the participatory approach in implementing agrobiodiversity policy in Japan are presented.

Agricultural Sustainability and Agrobiodiversity

Although many definitions of sustainable agriculture exist, most of them address ecological, economic, and social goals.[18,19] From an ecological perspective, sustainable farming practices (1) maintain their natural resource base; (2) rely on minimal artificial inputs from outside the farm system; (3) manage pests and diseases through internal regulating mechanisms; and (4) allow the recovery from the disturbances caused by cultivation and harvest.[20] Ecological characteristics of an agroecosystem, such as diversity, trophic structure, energy flow, nutrient cycles, population-regulating mechanisms, stability, and resilience, are metrics to help determine if a particular farming practice, input, or management decision is sustainable.[2]

Functions of biodiversity in agroecosystems, or agrobiodiversity, are foundations of sustainable farming practices. Ever since agriculture began some 10,000–12,000 years ago, biodiversity has allowed farming systems to evolve by providing genetic resources including edible plants, crop species, and livestock species.[21] However, diversities of crops and livestock have been decreasing rapidly since the introduction of monocultures of high-yielding cultivars and industrialization.[22] Insect and plant biodiversity in an agroecosystem provide multiple ecological services for agriculture. Examples include pollination,[23,24] insect pest management by intercropping,[25] and disease control by mixed planting of multiple cultivars.[26] Soil biodiversity, even though perhaps fewer than 10% of the species have been identified, is vital to soil fertility, decomposition of organic matter, soil structure, and soil health.[27,28] With higher diversity, there is greater microhabitat differentiation; there are more opportunities for coexistence and beneficial interference between species; there are more possible kinds of beneficial interactions between herbivores and their predators; there is more efficient use of resources of soil, water, and light; and there is reduced risk for the farmer. Examples of farming practices that can enhance agrobiodiversity and agroecosystem sustainability are listed in Table 1. Studies show that species and taxonomic groups respond to varying degrees to environmentally friendly management.[29,30]

Bioindicators

Heink and Kowarik[31] proposed an all-encompassing definition of indicators: "An indicator in ecology and environmental planning is a component or a measure of environmentally relevant phenomena used to depict or evaluate environmental conditions or changes or to set environmental goals. Environmentally relevant phenomena are pressures, states, and responses as defined by OECD (2003)"[32] (see Figure 1).

Bioindicators are based on biota that serve as indicators of the quality of the environment, the biotic component, or human impacts within an ecosystem.[34] The concept of bioindicators has been around for a long time—for example, the use of canaries to detect deadly carbon monoxide and methane gas buildup in mines has a long history. In the United Kingdom, it had been used since 1911 until it was completely replaced with gas detectors in 1986.[35] Other early uses of bioindicators include aquatic organisms to evaluate water quality in the United States[36] and the melanic form of moths to detect air pollution in the United Kingdom.[37,38] The number of publications about environmental indicators including bioindicators has been increasing since the 1980s.[39] The inaugural issues of the journals *Ecological Indicators* and *Environmental Bioindicators* (changed to *Environmental Indicators* in 2010) were published in 2001 and 2006, respectively. At the 2002 Johannesburg World Summit on Sustainable Development, representatives from 190 countries committed to achieving a significant reduction of the

TABLE 1 Farming Practices That Can Enhance Agrobiodiversity and Agroecosystem Sustainability

Habitat diversification	
Spatial	Temporal
Intercropping	Rotations
Trap crops	Fallow
Hedgerows	Cover crops
Shelterbelts	
Windbreaks	
Agroforestry	
Mosaic landscape	
Organic amendment applications	
Compost, organic mulch	
Green manure	
Conservation or minimum tillage	
Biological pest management	
No or reduced use of pesticides, fungicides, herbicides, and fumigants	
Use of beneficial insects	
Anaerobic soil disinfestation[a]	
Physical pest management	
Solarization	
Flooding	
Plant resistance	

[a] A biological method to control a range of soilborne pathogens using a principle of acid fermentation in anaerobic soil. See Rosskopf et al.[33] for more details.

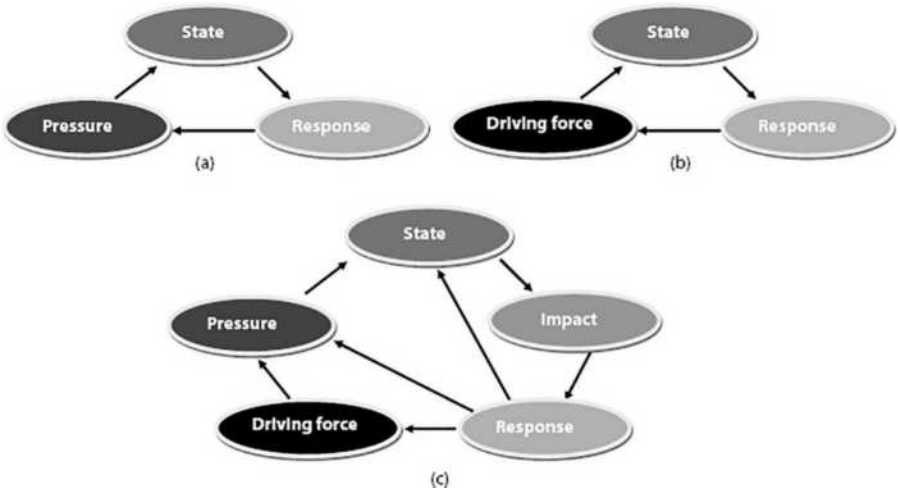


FIGURE 1 Varying causal chain frameworks. (a) PSR (pressure–state–response), (b) DSR (driver–state–response), and (c) DPSIR (driving force–pressure–state–impact–response). (Niemeijer and de Groot.[54])

current rate of biodiversity loss at the global, regional, and national levels by 2010.[40] Since then, the use of bioindicators for evaluating the status and functions of ecosystems and sustainability of agroecosystems has risen considerably,[41,42] yet the majority of recent studies of bioindicators has focused on the detection of pollution.[39]

Successful bioindicators should not only have biological and methodological relevance but also societal relevance because indicators require long-term monitoring to separate real change from natural fluctuations. Features of bioindicators are listed in Table 2. Inventory and classification are the foundations for developing indicators.[43] Developing an indicator then involves several steps: define objectives; determine end uses; construct indicators; determine norms and thresholds; and test sensitivity, probability, and usefulness.[44] Compared to abiotic indicators, however, developing a bioindicator has greater challenges for several reasons. First, biotic parameters are highly variable both temporally and spatially. There is a paucity of information on temporal and spatial variability of most natural species populations. Second, due to high variability, it costs more to collect additional replications to achieve the same statistical power in data. Standardized sampling methods for bioindicators are often lacking. Databases for bioindicators are critically limited especially in developing countries[45] although number of such studies are increasing.[46,47] Finally, it is more difficult with bioindicators to define background levels, norms, and thresholds.[42]

Nevertheless, demand for bioindicators as evaluative tools at diverse levels is increasing. Some of these are policy driven and others are market driven. For example, international organizations, such as the Organisation for Economic Co-operation and Development (OECD)[48–50] and the

TABLE 2 Features of Bioindicators for Environmental and Ecological Health Assessment

Biological relevance	• Provides early warning • Exhibits changes in response to stress • Changes are measured on appropriate time scale • Intensity of changes related to intensity of stressors • Change occurs when effect is real • Changes are biologically important and occur early enough to prevent catastrophic effects • Changes can be attributed to a cause • Changes indicate effects on ecosystem services • Can be used as sentinels for humans
Methodological relevance	• Easy to use in the field • Can be used by nonspecialists • Easy to analyze and interpret data • Measures what it is supposed to measure • Useful to test management questions • Can be used for hypothesis testing • Can be conducted in a reasonable time • Does not require expensive or complicated equipment • Easily repeatable with little training
Societal relevance	• Of interest to the public • Of interest to regulators and public policy makers • Easily understood by the public • Methods transparent to the public • Measures related to environment, ecological integrity, and human health • Cost-effective • Adds measurably to other indicators • Complements other indicators

Source: Modified from Burger.[34]

European Union (EU),[51] promote development of standardized bioindicators as a policy to compare biodiversity worldwide. Among commercial sectors, bioindicators have been used to certify "environmentally sound" products, such as migratory birds for shade-grown coffee.[52] Ecologists are using bioindicators and various other indicators at different levels of organization to evaluate the health of ecosystems (Table 3) and make recommendations for future management.[53,54] Bioindicators are often integrated into the frameworks of sustainability indicators of agroecosystems.[55,56]

TABLE 3 Usefulness of Indicators at Different Biological Levels of Organization to Ecological Health

Ecological Level	Type of Indicator	Ecological Health
Individual	Contaminant levels	Used to evaluate health of individuals For evaluation of risk to higher-level consumers As an indicator of health of its foods, including prey
	Lesions	
	Disease	
	Tumors	
	Infertility	
	Growth	
	Longevity	
	Reproduction	
	Age of reproduction	
	Hormonal balance	
	Proper development and maturation	
Population	Reproductive rates	Used to evaluate health of populations of species, particularly endangered or threatened species
	Growth rates	
	Survival rates	For comparison among populations
	Movements	For temporal comparisons
	Population genetics related to the breadth of the gene bank	Sources of resistance and pressure of natural selection
Community	Foraging guilds	Measures health of species using the same niche, such as colonial birds nesting in a colony or foraging animals such as dolphins and tuna
	Breeding guilds (groups of related species)	
	Predator–prey interactions	Indicates relationship among different species within guilds or assemblages
	Competitive interactions	
	Pathogen–host, pest–host relationships	For spatial and temporal comparisons
	Species richness	For evaluating efficacy of management options
Ecosystem	Decomposition rates	Measure changes in relative presence of species, how fast nutrients and energy will become available, how fast nutrients in soil will no longer be available, how much photosynthesis is occurring
	Erosion rates	
	Primary productivity	
	Energy transfer	Examines overall structure of the ecosystem in terms of the relationships among trophic levels
	Nutrient flow	
	Relationship among different trophic levels	For evaluating efficacy of management options
	Biomass	
	Energy flow	
Landscape	Relative amounts of different habitats	Measures dispersion of different habitat types, indicates relative species diversity values
	Patch size	Measures the difference among habitats
	Corridors between habitat types or different ecosystems	Measures distribution of corridors and refugia within the landscape
	The extent of uniform genetics	Also can measure the relationship between developed and natural areas for evaluating the importance of specific ecosystems within the landscape

Source: Modified from Burger.[34]

Bioindicator Development in European Union Agroecosystems

The importance of agro-environmental indicators has been highlighted by the EU[51] where current agricultural policies aim to increase multifunctionality of agricultural production. Intensive studies on bioindicators have been conducted in European countries. Bioindicators demonstrated to be sensitive to farm management intensities in European and some other agroecosystems are listed in Table 4. Generally, it is observed among invertebrate species that with less intensive management, there are more specialists and less generalists (as a result of ecological succession), greater biodiversity, and greater resilience.[42] An example of a bioindicator based on these correlations is European spiders; habitat preferences of spiders, particularly the ratio of "pioneer species (mostly Linyphiidae)" versus "wolf spiders (Araneae: Lycosidae)," can be a sensitive indicator for the assessment of farming intensity.[56] Many bioindicators listed in Table 4, however, have critical use limitations due to technically complex sampling methods and greater temporal and spatial variability. Although special instruments are required, recent advancements in molecular techniques may make some highly sensitive bioindicators such as nematodes much more accessible.[57,58]

TABLE 4 Examples of Potential Bioindicators for Sustainability of Farming Practices in European and Other Agroecosystems

Bioindicator	Parameter	Comments	References
		Arthropods	
Ground beetles (Carabidae)	Abundance	• Sensitive to management intensity but needs intensive data collection	[91–96]
Spiders (Araneae)	Habitat preferences Percent pioneer species	• Highly sensitive to management intensity and database is available on ecological characteristics of central European spiders	[97–105]
Hoverflies (Syrphidae)	Percent stenotopic species	• Diversity of landscape structure adjacent to the field enhances species numbers	[106–111]
Pollinators	Individuals Populations Ecological guilds	• Environmental stress brought about by pesticides and habitat modification reduce pollinators	[112–114]
Arthropod community	Abundance of key species	• Diverse arthropod communities are affected by landscape diversity and environmental stress	[98,115–116]
		Soil fauna	
Ants	Diversity Community composition	• Good indicator for rangeland monitoring and land-use changes	[117–122]
Earthworm	Biomass Species number Ecological guilds	• Suitable indicator for soil structure or compaction, tillage practice, pesticides, and other toxic chemicals	[123–131]
Collembola Protozoa Nematode Micro-arthropods Mites Soil animal	Physiotype Biodiversity Trophic index Maturity index Food web	• Highly sensitive to management intensity but time consuming and special skills are required for identification. Recent advancement in molecular approach makes these more accessible	[58,132–157]
		Soil microbiota	
Soil enzymes (e.g., glucosinases for cellulose decomposition, phenol oxidases for lignin decomposition)	Activities	• Moderately sensitive to management intensity and relatively easy to measure but moderately variable both temporally and spatially	[58,158–160]

(*Continued*)

TABLE 4 (*Continued*) Examples of Potential Bioindicators for Sustainability of Farming Practices in European and Other Agroecosystems

Bioindicator	Parameter	Comments	References
Soil protists (e.g., algae, amoeba, diatom)	Species diversity Photosynthetic activity	• Less studied but sensitive to tillage, pesticides, and fertilizers	[161–163]
Microbial communities	Composition Functional diversity PLFA[a] profiles qPCR[b]	• Moderately sensitive to management intensity, but special skills and equipment are required and difficult to interpret	[58,164–166]
Functional groups	Mycorrhizae Nitrification Root pathogens	• Highly sensitive to management intensity, but special skills and instruments are required. DNA-based approaches (e.g., TRFLP[C], qPCR) became popular	[58,167–171]
Microbial activity	Soil respiration Mineralization Multiple substrate-induced respiration (MSIR)	• Relatively easy to measure but highly variable both temporally and spatially	[58,159,160,171]
Microbial biomass	C, N, and P biomass	• Relatively easy to measure but highly variable both temporally and spatially	[159,160]
		Birds	
Terrestrial birds Farmland birds	Species abundance of focal species and taxonomic composition at landscape scales	• Bird communities can be used to evaluate the effect of agriculture on surrounding ecosystems and hydrology and to explore sustainable land-use scenarios at regional scales	[62,172–175]
		Plants	
Higher plants	Numbers of "characteristic" species, weeds, functional groups, and endangered species Cover of litter in vegetation Diversity Evenness indices Habitat age	• Capable of being integrated into sophisticated floristic diversity at the habitat scale but requires intensive data collection	[176–180]

[a] PLFA, phospholipid fatty acid.

[b] qPCR, quantitative polymerase chain reaction.

[c] TRFLP, terminal restriction fragment length polymorphism.

To practically implement agro-environmental policy in the EU, efforts have been made to develop relatively easy-to-measure surrogate indicators (e.g., length of borders, farm size, area managed with organic farming).[42] Another practical bioindicator is a list of indicator plant species to evaluate species richness of a farm. A total of 28 indicator flower species for meadows and pastures that can be easily identified by local farmers were selected in Baden-Württemberg, Germany. Agro-environmental payments are granted to farms that have at least 4 of these 28 indicator species in all of the meadows and pastures on the farm.[59] In designing more efficient agro-environmental schemes, advantages of result-oriented remuneration (e.g., payment towards species-rich meadows) over action-oriented remuneration (e.g., payment towards manuring and mowing once per year) have been discussed.[60] It has been further recognized that the preservation of biodiversity is only possible through the (re)establishment of a mosaic of habitat patches at the landscape level. To meet this need, Geographical Information System

(GIS)-based landscape-oriented indicators have been examined.[61] GIS approaches are powerful tools for special analysis of bioindicators such as studies on correlations of bioindicator distributions (e.g., birds) and environmental and social factors at a landscape level.[62] Landscape context can have significant effect on biodiversity.[63,64]

To select objective, broad-scale, and unbiased indicators, conceptual frameworks for agro-environmental indicators have been proposed. Varying causal chain frameworks such as pressure–state–response (PSR), driving force–state–response (DSR),[65] and driving force–pressure–state–impact–response (DPSIR)[66] have been developed (Figure 1). Enhanced DPSIR (eDPSIR) was proposed as a way to provide improved conceptual guidance in indicator selection based upon the DPSIR approach, systems analysis, and causal networks.[54] Sustainability Assessment of Farming and the Environment (SAFE), a hierarchical framework for assessing the sustainability of agricultural systems in Belgium, was created as a consistent and comprehensive framework of principles, criteria, and indicators.[55] These frameworks integrate bioindicators as a component of sustainability indicators of agroecosystems.

Biodiversity has multiple dimensions, and it is difficult to measure by a single indicator. Feest et al. proposed the "biodiversity quality" approach using multiple indicators (e.g., species richness, evenness/dominance, density/population, relative biomass, and species conservation value index) in combination to assess the balance between a range of indices and their relative magnitude.[67,68] Four different types of indicators of biodiversity change were suggested as a tool to facilitate indicator development (Figure 2).[69] With a goal of stable or increasing populations in all species associated with agricultural landscapes, Butler et al. developed a cross-taxonomic index for quantifying the health of farmland biodiversity by which the detrimental impacts of agricultural change to a broad range of taxonomic groupings can be assessed.[70]

For nationwide soil monitoring in the United Kingdom, a framework in selecting soil bioindicators for balancing scientific and technical opinion to assist policy development was established.[58]

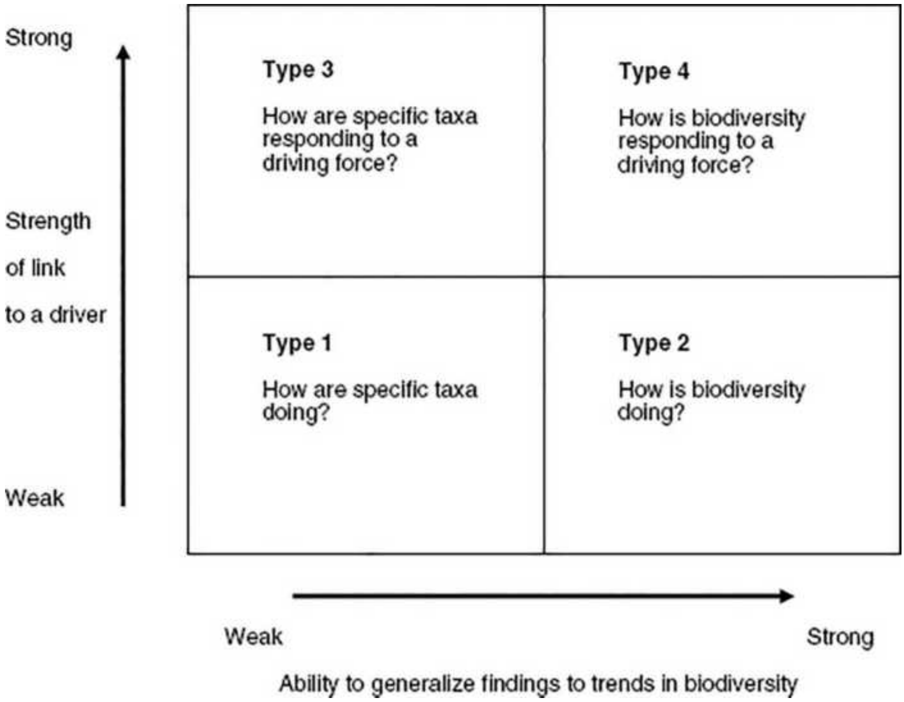


FIGURE 2 Typology of biodiversity change indicators. (van Strien et al.[69])

This semi-objective approach using "logical sieve" yielded 17 bioindicators that cover a range of genotypic-, phenotypic-, and functional-based indicators for different trophic groups out of 183 candidate bioindicators. This framework allows transparency in the decision-making process in selecting soil bioindicators as well as flexibility in including other indicators depending on priorities of the monitoring. The need for unambiguous and broad definitions of terms such as *indicators* and *biodiversity* was addressed for improving communication among interdisciplinary researchers, policy makers, and stakeholders.[71] To make indicators relevant to potential users and have long-term public support, the necessity of stakeholder participation in developing ecological indicators was also emphasized.[72]

Bioindicator Development in Shaded Coffee in Latin America

Coffee is the second most traded commodity in the world after petroleum and forms the principal economic activity of more than 20 million people in farming communities throughout much of the developing world.[73] Traditionally grown in the understory of forest cover or planted shade trees, throughout the 1970s and 1980s, many coffee farmers adopted more modern production practices, planting higher-yielding varieties in full sun, eliminating shade trees, and increasing pesticide and fertilizer applications.[74] The loss of shade cover, and associated biodiversity, has led to many environmental problems such as soil erosion, loss of water capture and recharge ability, and contamination from the excessive fertilizer and pesticide use associated with sun-grown coffee.

Recent research has documented the high levels of agrobiodiversity and ecosystem services (such as pollination, soil conservation, and natural pest control) associated with diverse shade coffee production.[75,76] From this research, several very important bioindicators are being developed. The number of tree species in the shade canopy is an important indicator of both conservation potential of coffee plantations and increased options for farmer livelihood alternatives.[77–79] The species richness of ants and insect feeding birds are greater in shade coffee than in sun coffee, especially ground-foraging ants that act as important predators and birds that are bark gleaners and leaf surface foragers. It also appears that the higher diversity of predaceous ants restricts the development of pest ants such as the fire ant *Solenopsis* sp. (Hymenoptera), a very common pest in open landscapes and sun-grown coffee. Associated species such as orchids in the shade tree canopy can also be important bioindicators. Local farmers can be trained to recognize orchids, demonstrating how the development of bioindicators of sustainability can be an accessible and local methodology that adds value to more sustainable farming practices. Orchids have an intrinsic value from a conservation perspective and, at the same time in this region of northern Nicaragua, have added an attractive value for an emerging agro-ecotourism industry associated with shade-grown coffee landscapes.[80]

Aquatic Organisms as Bioindicators

Recent studies showed that a range of aquatic organisms that inhabit water bodies in agroecosystems or in watersheds including agroecosystems can be used as bioindicators for evaluating sustainability of agroecosystems or their impacts on the environment. A study in organic and conventional paddy fields in Costa Rica found that some aquatic macroinvertebrates (*Baetis* sp., *Fallceon* sp., *Leptohyphes* sp., *Tricorythodes* sp., *Farrodes* sp., *Phyllogomphoides* sp., *Hydroptila* sp., *Mayatrichia* sp., *Neotrichia* sp., *Oxyethira* sp., *Nectopsyche* sp.1, *Nectopsyche* sp.2, and *Oecetis* sp.) were found only in organic paddy fields and can be used as sensitive bioindicators of water quality in these agroecosystems.[81]

Rhinella arenarum, a toad species, inhabiting ponds associated with agroecosystems, showed a smaller egg diameter, a higher frequency of abnormal eggs, and a higher mortality of oviposition compared to those inhabiting a reference pond without agricultural influences.[82] The δ15N signature of large, long-lived, freshwater mussels (Mollusca: Bivalvia: Unionidae) was successfully used as an integrated bioassessment tool for tracking agricultural nitrogen inputs into watersheds.[83]

Improving Awareness of Farmers and Consumers

Often, farmers themselves are not fully aware of the biodiversity in their farms. Moreover, to economically sustain environmentally friendly farming practices, recognition and support from consumers are necessary. To implement sustainable agricultural policy, therefore, improved awareness of both farmers and consumers on agrobiodiversity is required.

An example of such activity is the participatory biodiversity inventory in paddy rice ecosystems in Japan, where paddy fields occupy ~50% of the cultivated lands of the country. Japanese paddy agroecosystems contain a diverse group of organisms: Recent inventory found 5668 species including birds (189 spp.), fish (143 spp.), reptiles and amphibian (61 spp.), arthropods (1726 spp.), and plants (2075 spp.).[84] Since the 1960s, however, many native species inhabiting paddy fields have decreased in abundance mainly due to nontarget effects of pesticide applications and the construction of concrete ditches. Consequently, 3 mammal species, 15 insect species, 5 bird species, 27 species of fish, 3 amphibian species, 3 mollusks, and 26 plant species are listed as endangered species inhabiting paddy fields.[84] Nongovernmental organizations (NGOs), scientists, farmers, the general public, and administrators collaborated to create nationwide inventories of biodiversity in paddy agroecosystems and proposed a broad range of bioindicators (237 animal species[85] and 223 plant species[86]) to enhance the awareness of both farmers and consumers on paddy field biodiversity (Figure 3).

FIGURE 3 A portion of Japanese poster for improving farmer and consumer awareness of agrobiodiversity in paddy fields. Central figure indicates a bowl of rice being equivalent to 3000–4000 grains of rice, 3 stubs of rice plants, and 35 tadpoles in a paddy. This can be interpreted as eating a bowl of rice supports 35 tadpoles in a paddy. Numbers of species are averages from the national survey of agrobiodiversity in paddy fields in Japan. Clockwise from the upper left, the species listed are *Daphnia pulex* (water flea), *Pantala flavescens* (Globe Skimmer), *Sympetrum frequens* (Autumn Darter), *Triops longicaudatus* (longtail tadpole shrimp), *Branchinella kugenumaensis* (fairy shrimp), *Hyla japonica* (Japanese tree frog), *Rhabdophis tigrinus* (tiger keelback), *Misgurnus anguillicaudatus* (Dojo Loach), *Oryzias latipes* (Medaka), *Rana nigromaculata* (Dark-spotted frog), *Rana rugosa* (Japanese wrinkled frog), and *Ceriagrion melanurum* (damselfly). (Une.[181])

Conclusions

Indicators represent a compromise between scientific knowledge of the moment and simplicity of use.[44] Compared to assessment systems for natural ecosystems,[43,87] the current status of developing bioindicators for sustainability of farming practices appears to still be in its early stages but is making good progress as seen in increased number of studies from places outside of Europe. For using bioindicators to implement sustainable agricultural policy, however, we need not only more research on the science of bioindicators but also better awareness on agrobiodiversity among farmers and consumers.

Future studies on bioindicators for sustainable farming practices should address the following: (1) the importance of stability and reproductive potential of not only pest and beneficial species but also other species typical of agroecosystems;[42,70,88] (2) standardization of sampling methods;[42] (3) expansion of databases and improvement of the statistical techniques to minimize potential bias of indicators;[68,89] (4) improved understanding of the mechanisms linking the status of biodiversity, Earth system processes, human decisions, and ecosystem services impacting human welfare through interdisciplinary studies;[52,89] (5) use of participatory approaches in the processes of developing bioindicators;[72,90] (6) development of multiple sets of bioindicators tailored to different end users (general public, farmers, policy makers, and scientists);[42] and (7) constructing a hierarchical system that integrates different types of bioindicators and ecological, social, cultural, and economic indicators to evaluate sustainability of farming practices.[55].

References

1. Pretty, J. Agricultural sustainability: Concepts, principles and evidence. *Philos. Trans. R. Soc., B* **2008**, *363* (1491), 447–465.
2. Gliessman, S.R. *Agroecology: The Ecology of Sustainable Food Systems*, 3rd ed.; CRC Press/Taylor and Francis: Boca Raton, FL, 2015.
3. Green, R.E.; Cornell, S.J.; Scharlemann, J.P.W.; Balmford, A. Farming and the fate of wild nature. *Science* **2005**, *307* (5709), 550–555.
4. Tscharntke, T.; Klein, A.M.; Kruess, A.; Steffan-Dewenter, I.; Thies, C. Landscape perspectives on agricultural intensification and biodiversity—Ecosystem service management. *Ecol. Lett.* **2005**, *8* (8), 857–874.
5. International Federation of Organic Agriculture Movements (IFOAM). *Towards a Sustainable Agriculture: Papers*; Wirz: Aarau, Switzerland, 1978.
6. Douglass, G.K. *Agricultural Sustainability in a Changing World Order*; Westview Press: Boulder, CO, 1984.
7. Jackson, W.; Berry, W.; Colman, B. *Meeting the Expectations of the Land: Essays in Sustainable Agriculture and Stewardship*; North Point Press: San Francisco, CA, 1984.
8. Dover, M.J.; Talbot, L.M. *To Feed the Earth: Agro-ecology for Sustainable Development*; World Resources Institute: Washington, DC, 1987.
9. Altieri, M.A. Beyond agroecology: Making sustainable agriculture part of a political agenda. *Am. J. Altern. Agric.* **1988**, *3* (4), 142–143.
10. Andow, D.A.; Hidaka, K. Experimental natural history of sustainable agriculture: Syndromes of production. *Agric. Ecosyst. Environ.* **1989**, *27*, 447–462.
11. MacRae, R.J.; Hill, S.B.; Henning, J.; Mehuys, G.R. Agricultural science and sustainable agriculture—A review of the existing scientific barriers to sustainable food-production and potential solutions. *Biol. Agric. Hort.* **1989**, *6* (3), 173–219.
12. Soule, J.D.; Piper, J.K. *Farming in Nature's Image: An Ecological Approach to Agriculture*; Island Press: Washington, DC, 1992.
13. Liverman, D.M.; Hanson, M.E.; Brown, B.J.; Merideth, R.W.J. Global sustainability toward measurement. *Environ. Manage.* **1988**, *12* (2), 133–144.

14. Jansen, D.M.; Stoorvogel, J.J.; Schipper, R.A. Using sustainability indicators in agricultural land use analysis: An example from Costa Rica. *Neth. J. Agric. Sci.* **1995**, *43* (1), 61–82.
15. Food Agriculture Organization of the United Nations (FAO). Land Quality Indicators and Their Use in Sustainable Agriculture and Rural Development: Proceedings of the Workshop; FAO: Rome, Italy, 1997.
16. MAFF U.K. *Toward Sustainable Agriculture: A Pilot Set of Indicators*; MAFF Publications: London, UK, 2000.
17. Bell, S.; Morse, S. Sustainability Indicator: Measuring the Immeasurable, 2nd Ed.; Earthscan: London, UK, 2008.
18. Spedding, C.R.W. Agriculture and the Citizen, 1st Ed.; Chapman and Hall: London, UK and New York, 1996.
19. Ikerd, J.E. Rethinking the first principles of agroecology: Ecological, social, and economic. In Sustainable Agroecosystem Management: Integrating Ecology, Economics, and Society; Bohlen, P.J., House G.J., Eds.; CRC Press/Taylor & Francis: Boca Raton, FL, 2009; 41–52.
20. Edwards, C.A. Sustainable agricultural practices. In Encyclopedia of Pest Management; Pimentel, D., Ed.; Marcel Dekker/Taylor & Francis: Boca Raton, FL, 2002; 812–814.
21. Hillel, D.; Rosenzweig, C. The role of biodiversity in agronomy and agroecosystem management in the coming century. In Sustainable Agroecosystem Management: Integrating Ecology, Economics, and Society; Bohlen, P.J., House, G.J., Eds.; CRC Press/Taylor & Francis: Boca Raton, FL, 2009; 167–193.
22. Thrupp, L.A. The importance of biodiversity in agroecosystems. *J. Crop Improv.* 2004, *11* (1/2), 315–338.
23. Kearns, C.A.; Inouye, D.W.; Waser, N.M. Endangered mutualisms: The conservation of plant–pollinator interactions. *Annu. Rev. Ecol. Syst.* **1998**, *29*, 83–112.
24. Buchmann, S.L.; Nabhan, G.P. *The Forgotten Pollinators*; Island Press/Shearwater Books: Washington, DC, 1996.
25. Ogol, C.; Spence, J.R.; Keddie, A. Maize stem borer colonization, establishment and crop damage levels in a maize-leucaena agroforestry system in Kenya. *Agric. Ecosyst. Environ.* **1999**, *76* (1), 1–15.
26. Zhu, Y.Y.; Chen, H.R.; Fan, J.H.; Wang, Y.Y.; Li, Y.; Chen, J.B.; Fan, J.X.; Yang, S.S.; Hu, L.P.; Leung, H.; Mew, T.W.; Teng, P.S.; Wang, Z.H.; Mundt, C.C. Genetic diversity and disease control in rice. *Nature* **2000**, *406* (6797), 718–722.
27. Pankhurst, C.E.; Doube, B.M.; Gupta, V.V.S.R., Eds. *Biological Indicators of Soil Health;* CAB International: Wallingford, UK, 1997.
28. Orgiazzi, A., Bardgett, R. D., Barrios, E., Behan-Pelletier, V., Briones, M. J. I., Chotte, J.-L., De Deyn, G. B., Eggleton, P., Fierer, N., Fraser, T., Hedlund, K., Jeffery, S., Johnson, N. C., Jones, A., Kandeler, E., Kaneko, N., Lavelle, P., Lemanceau, P., Miko, L., Montanarella, L., Moreira, F. M. S., Ramirez, K. S., Scheu, S., Singh, B. K., Six, J., van der Putten, W. H., and Wall, D. H. Eds. Global *Soil Biodiversity Atlas*; Publications Office of the European Union: Luxembourg, 2016.
29. Critchley, C.N.R.; Fowbert, J.A.; Sherwood, A.J.; Pywell, R.F. Vegetation development of sown grass margins in arable fields under a countrywide agri-environment scheme. *Biol. Conserv.* **2006**, *132* (1), 1–11.
30. Marshall, E.J.P.; West, T.M.; Kleijn, D. Impacts of an agrienvironment field margin prescription on the flora and fauna of arable farmland in different landscapes. *Agric. Ecosyst. Environ.* **2006**, *113* (1–4), 36–44.
31. Heink, U.; Kowarik, I. What are indicators? On the definition of indicators in ecology and environmental planning. *Ecol. Indic.* **2010**, *10* (3), 584–593.
32. Organisation for Economic Co-operation and Development (OECD). *Core Environmental Indicators: Development, Measurement, and Use*; OECD: Paris, France, 2003.

33. Rosskopf, E., Serrano-Pe´rez, P., Hong, J., Shrestha, U., del Carmen Rodrı´guez-Molina, M., Martin, K., Kokalis-Burelle, N., Shennan, C., Muramoto, J., and Butler, D. Anaerobic soil disinfestation and soilborne pest management. In *Organic Amendments and Soil Suppressiveness in Plant Disease Management*; Meghavansi, M. K., Varma, A., Eds.; Springer International Publishing: Switzerland, 2015; 277–305.
34. Burger, J. Bioindicators: Types, development, and use in ecological assessment and research. *Environ. Bioindic.* **2006**, *1* (1), 22–39.
35. BBC. 1986: Coal Mine Canaries Made Redundant. On This Day: 30th of December, available at http://news.bbc.co.uk/onthisday/hi/dates/stories/december/30/newsid_2547000/2547587.stm (accessed June 2019).
36. Richardson, R.E. The bottom fauna of the middle Illinois river, 1913–1925. *Ill. Nat. Hist. Surv. Bull.* **1928**, *17*, 387–475.
37. Kettlewell, H.B.D. How industrialisation can alter species. *Discovery* **1955**, *16* (12), 507–511.
38. Heck, W.W. The use of plants as indicators of air pollution. *Air Water Pollut.* **1966**, *10* (2), 99–111.
39. Burger, J. Bioindicators: A review of their use in the environmental literature 1970–2005. *Environ. Bioindic.* **2006**, *1* (2), 136–144.
40. United Nations Environment Programme (UNEP). Decisions Adopted by the Conference of the Parties to the Convention on Biological Diversity at Its Sixth Meeting, The Hague, 7–19 April 2002.
41. Paoletti, M.G. Using bioindicators based on biodiversity to assess landscape sustainability. *Agric. Ecosyst. Environ.* 1999, 74 (1–3), 1–18.
42. Buchs, W. Biodiversity and agri-environmental indicators—General scopes and skills with special reference to the habitat level. *Agric. Ecosyst. Environ.* **2003**, *98* (1–3), 35–78.
43. Innis, S.A.; Naiman, R.J.; Elliott, S.R. Indicators and assessment methods for measuring the ecological integrity of semi-aquatic terrestrial environments. *Hydrobiologia* 2000, 422–*423*, 111–131.
44. Girardin, P.; Bockstaller, C.; Van der Werf, H. Indicators: Tools to evaluate the environmental impacts of farming systems. *J. Sustainable Agric.* **1999**, *13* (4), 5–21.
45. Dudley, N.; Baldock, D.; Nasi, R.; Stolton, S. Measuring biodiversity and sustainable management in forests and agricultural landscapes. *Philos. Trans. R. Soc., B* **2005**, *360* (1454), 457–470.
46. Diame, L., Blatrix, R., Grechi, I., Rey, J. Y., Sane, C. A. B., Vayssieres, J. F., de Bon, H., and Diarra, K. Relations between the design and management of Senegalese orchards and ant diversity and community composition. *Agric. Ecosys. Environ.* **2015**, *212*, 94–105.
47. Rousseau, L., Fonte, S. J., Tellez, O., van der Hoek, R., and Lavelle, P. Soil macrofauna as indicators of soil quality and land use impacts in smallholder agroecosystems of western Nicaragua. *Ecol. Indic.* **2013**, *27*, 71–82.
48. Neher, D. Ecological sustainability in agricultural systems: Definition and measurement. *J. Sustainable Agric.* **1992**, 2, 51–61. https://doi.org/10.1300/J064v02n03_05
49. Organisation for Economic Co-operation and Development (OECD). *Environmental Indicators for Agriculture: Methods and Results*. OECD Publishing: Paris, France, 2001; Vol. 3.
50. Organisation for Economic Co-operation Development (OECD). Agriculture and Biodiversity: Developing Indicators for Policy Analysis: Proceedings from an OECD Expert Meeting, Zurich, Switzerland. OECD Publishing: Paris, France, 2001.
51. The Commission of the European Communities. Evaluation of Agri-environment Programmes. State of Application of Regulation. (EEC) no. 2078/92. DGVI Commission Working Document VI/7655/98, 1998.
52. Mas, A.H.; Dietsch, T.V. Linking shade coffee certification to biodiversity conservation: Butterflies and birds in Chiapas, Mexico. *Ecol. Appl.* **2004**, *14* (3), 642–654.
53. The Millennium Assessment Board. The Millennium Ecosystem Assessment, 2005.
54. Niemeijer, D.; de Groot, R.S. A conceptual framework for selecting environmental indicator sets. *Ecol. Indic.* **2008**, *8* (1), 14–25.

55. Van Cauwenbergh, N.; Biala, K.; Bielders, C.; Brouckaert, V.; Franchois, L.; Cidad, V.G.; Hermy, M.; Mathijs, E.; Muys, B.; Reijnders, J.; Sauvenier, X.; Valckx, J.; Vanclooster, M.; Van der Veken, B.; Wauters, E.; Peeters, A. SAFE—A hierarchical framework for assessing the sustainability of agricultural systems. *Agric. Ecosyst. Environ*. 2007, 120 (2–4), 229–242.
56. Buchs, W.; Harenberg, A.; Zimmermann, J.; Wei, B. Biodiversity, the ultimate agri-environmental indicator? Potential and limits for the application of faunistic elements as gradual indicators in agroecosystems. *Agric. Ecosyst. Environ.* **2003**, *98* (1–3), 99–123.
57. Donn, S.; Griffiths, B.S.; Neilson, R.; Daniell, T.J. DNA extraction from soil nematodes for multi-sample community studies. *Appl. Soil Ecol.* **2008**, *38* (1), 20–26.
58. Ritz, K.; Black, H.I.J.; Campbell, C.D.; Harris, J.A.; Wood, C. Selecting biological indicators for monitoring soils: A framework for balancing scientific and technical opinion to assist policy development. *Ecol. Indic*. **2009**, *9* (6), 1212–1221.
59. Oppermann, R. Nature balance scheme for farms-evaluation of the ecological situation. *Agric. Ecosyst. Environ*. **2003**, *98* (1–3), 463.
60. Matzdorf, B.; Kaiser, T.; Rohner, M.-S. Developing biodiversity indicator to design efficient agri-environmental schemes for extensively used grassland. *Ecol. Indic*. **2008**, *8* (3), 256–269.
61. Osinski, E. Operationalisation of a landscape-oriented indicator. *Agric. Ecosyst. Environ*. **2003**, *98* (1–3), 371.
62. Gottschalk, T.K.; Dittrich, R.; Diekötter, T.; Sheridan, P.; Wolters, V.; Ekschmitt, K. Modelling land-use sustainability using farmland birds as indicators. *Ecol. Indic*. **2010**, *10* (1), 15–23.
63. Merckx, T.; Feber, R.E.; Riordan, P.; Townsend, M.C.; Bourn, N.A.D.; Parsons, M.S.; Macdonald, D.W. Optimizing the biodiversity gain from agri-environment schemes. *Agric. Ecosyst. Environ*. 2009, 130 (3–4), 177–182.
64. Rundlof, M.; Bengtsson, J.; Smith, H.G. Local and landscape effects of organic farming on butterfly species richness and abundance. *J. Appl. Ecol*. **2008**, *45* (3), 813–820.
65. Organisation for Economic Co-operation and Development (OECD). *Environmental Indicators for Agriculture: Volume 1, Concepts and Frameworks*; OECD Publisher: Paris, France, 1999.
66. Smeets, E.; Weterings, R. *Environmental Indicators: Typology and Overview*; European Environmental Agency: Copenhagen, Denmark, 1999.
67. Feest, A.; Aldred, T.D.; Jedamzik, K. Biodiversity quality: A paradigm for biodiversity. *Ecol. Indic*. **2010**, *10* (6), 1077–1082.
68. Feest, A.; van Swaay, C.; van Hinsberg, A. Nitrogen deposition and the reduction of butterfly biodiversity quality in the Netherlands. *Ecol. Indic.* **2014**, *39*, 115–119.
69. van Strien, A.J.; van Duuren, L.; Foppen, R.P.B.; Soldaat, L.L. A typology of indicators of biodiversity change as a tool to make better indicators. *Ecol. Indic*. **2009**, *9* (6), 1041–1048.
70. Butler, S.J.; Brooks, D.; Feber, R.E.; Storkey, J.; Vickery, J.A.; Norris, K. A cross-taxonomic index for quantifying the health of farmland biodiversity. *J. Appl. Ecol*. **2009**, *46* (6), 1154–1162.
71. Sieg, C.H.; Flather, C.H. Applicability of Montreal Process Criterion 1—Conservation of biological diversity—to rangeland sustainability. *Int. J. Sustainable Dev. World Ecol*. **2000**, *7* (2), 81.
72. Turnhout, E.; Hisschemöller, M.; Eijsackers, H. Ecological indicators: Between the two fires of science and policy. *Ecol. Indic*. **2007**, *7* (2), 215–228.
73. Bacon, C.; Méndez, V.E.; Gliessman, S.R.; Goodman D.; Fox, J. *Confronting the Coffee Crisis: Fair Trade, Sustainable Livelihoods and Ecosystems in Mexico and Central America*; MIT University Press: Cambridge, MA, 2008; 390pp.
74. Moguel, P.; Toledo, V.M. Biodiversity conservation in traditional coffee systems of Mexico. *Conserv. Biol.* **1999**, *13,* 11–21.
75. Perfecto, I.; Vandermeer, J.H.; Wright, A. *Nature's Matrix: Linking Agriculture, Conservation, and Food Sovereignty*; Earthscan: New York, 2010; 237pp.

76. Coral-Acosta, N.; Perez-Torres, J. Diversity of diurnal butterflies (Lepidoptera: Papilionoidea) associated with a shade coffee agroecosystem (Curiti, Santander). *Revista Colombiana De Entomologia* **2017**, *43*, 91–99.
77. Gliessman, S.R.; Putnam, H.; Cohen, R. Chapter 8: Agroecology and participatory knowledge production and exchange as a basis for food system change: The case of the Community Agroecology Network. In *Agroecological Practices for Sustainable Agriculture: Principles, Applications, and Making the Transition*; Wezel, A., Ed.; World Scientific Publishing: London, UK, 2017; 201–228.
78. IPES-Food. *Breaking Away from Industrial Food and Farming Systems: Seven Case Studies of Agroecological Transition*; International Panel of Experts in Sustainable Food Systems: Brussels, Belgium, 2018; 109pp.
79. Mendez, V.E.; Gliessman, S.R.; Gilbert, G.S. Tree biodiversity in farmer cooperatives of a shade coffee landscape in western El Salvador. *Agric. Ecosyst. Environ.* 2007, 119 (1–2), 145–159.
80. Bacon, C.M. Confronting the coffee crisis: Nicaraguan farmers' use of cooperative, fair trade, and agroecological networks to negotiate livelihoods and sustainability. PhD Dissertation, Department of Environmental Studies. University of California Santa Cruz, California, 2005.
81. Rizo-Patron, F.; Kumar, A.; Colton, M. B. M.; Springer, M.; Trama, F. A. Macroinvertebrate communities as bioindicators of water quality in conventional and organic irrigated rice fields in Guanacaste, Costa Rica. *Ecol. Indic.* **2013**, *29*, 68–78.
82. Babini, M.S.; Bionda, C.D.; Salinas, Z.A.; Salas, N.E.; Martino, A. L. Reproductive endpoints of Rhinella arenarum (Anura, Bufonidae): Populations that persist in agroecosystems and their use for the environmental health assessment. *Ecotoxic. Environ. Safety*, **2018**, *154*, 294–301.
83. Atkinson, C.L.; Christian, A.D.; Spooner, D.E.; Vaughn, C.C. Long-lived organisms provide an integrative footprint of agricultural land use. *Ecol. Appl.* **2014**, *24*, 375–384.
84. Kiritani, K.; Ed. *A Comprehensive List of Organisms Associated with Paddy Ecosystems in Japan*, Revised Ed.; Nou-to shizen-no kenkyujo, Seibutsu-tayousei-nougyou shien senta: Fukuoka, Tokyo, Japan, 2010.
85. Kiritani, K., Ed. *Bioindicators for Paddy Ecosystems*; Nou-to shizen-no kenkyujo, Seibutsu-tayousei-nougyou shien senta: Fukuoka, Tokyo, Japan, 2009 (in Japanese).
86. Ito, K.; Mineta, T., Eds. *Plant Bioindicators for Paddy Ecosystems*; Nou-to shizen-no kenkyujo, Seibutsu-tayousei-nougyou shien senta: Fukuoka, Tokyo, Japan, 2009 (in Japanese).
87. Karr, J.R.; Fausch, K.D.; Angermeier, P.L.; Yant, P.R.; Schlosser, I.J. *Assessing Biological Integrity in Running Waters: A Method and Its Rationale. Special Publication 5*; Illinois Natural History Survey: Urbana, IL, 1986.
88. Kiritani, K. Integrated biodiversity management in paddy fields: Shift of paradigm from IPM toward IBM. *Integr. Pest Manage. Rev.* **2000**, *5* (3), 175–183.
89. Balmford, A.; Crane, P.; Dobson, A.; Green, R.E.; Mace, G.M. The 2010 challenge: Data availability, information needs and extraterrestrial insights. *Philos. Trans. R. Soc., B* **2005**, *360* (1454), 221–228.
90. Burger, J. Stakeholder involvement in indicator selection: Case studies and levels of participation. *Environ. Bioindic.* **2009**, *4* (2), 170–190.
91. Irmler, U. The spatial and temporal pattern of carabid beetles on arable fields in northern Germany (SchleswigHolstein) and their value as ecological indicators. *Agric. Ecosyst. Environ.* **2003**, *98* (1–3), 141–151.
92. Heyer, W.; Huelsbergen, K.J.; Wittmann, C.; Papaja, S.; Christen, O. Field related organisms as possible indicators for evaluation of land use intensity. *Agric. Ecosyst. Environ.* **2003**, *98* (1–3), 453–461.
93. Doring, T.F.; Hiller, A.; Wehke, S.; Schulte, G.; Broll, G. Biotic indicators of carabid species richness on organically and conventionally managed arable fields. *Agric. Ecosyst. Environ.* **2003**, *98* (1–3), 133.
94. Cavaliere, F.; Brandmayr, P.; Giglio, A. DNA damage in haemocytes of Harpalus (Pseudophonus) rufipes (De Geer, 1774) (Coleoptera, Carabidae) as an indicator of sublethal effects of exposure to herbicides. *Ecol. Indic.* **2019**, *98*, 88–91.

95. Pizzolotto, R.; Mazzei, A.; Bonacci, T.; Scalercio, S.; Iannotta, N.; Brandmayr, P. Ground beetles in Mediterranean olive agroecosystems: Their significance and functional role as bioindicators (Coleoptera, Carabidae). *PLoS ONE*, **2018**, *13*, e0194551.
96. Magagula, C.N.; Nzima, B.A. Interaction between habitat characteristics and insect diversity using ground beetles (coleoptera: carabidae) and ants (hymenoptera: formicidae) within a variety of agricultural habitats. *Appl. Ecol. Environ. Res.* **2015**, *13*, 863–876.
97. Oberg, S. Influence of landscape structure and farming practice on body condition and fecundity of wolf spiders. *Basic Appl. Ecol.* **2009**, *10* (7), 614–621.
98. Fernandez, D.E.; Cichon, L.I.; Sanchez, E.E.; Garrido, S.A.; Gittins, C. Effect of different cover crops on the presence of arthropods in an organic apple (*Malus domestica* Borkh) orchard. *J. Sustainable Agric.* **2008**, *32* (2), 197–211.
99. Birkhofer, K.; Fliessbach, A.; Wise, D.H.; Scheu, S. Generalist predators in organically and conventionally managed grass-clover fields: Implications for conservation biological control. *Ann. Appl. Biol.* **2008**, *153* (2), 271–280.
100. Clough, Y.; Holzschuh, A.; Gabriel, D.; Purtauf, T.; Kleijn, D.; Kruess, A.; Steffan-Dewenter, I.; Tscharntke, T. Alpha and beta diversity of arthropods and plants in organically and conventionally managed wheat fields. *J. Appl. Ecol.* **2007**, *44* (4), 804–812.
101. Isaia, M.; Bona, F.; Badino, G. Influence of landscape diversity and agricultural practices on spider assemblage in Italian vineyards of Langa Astigiana (northwest Italy). *Environ. Entomol.* **2006**, *35* (2), 297–307.
102. Schmidt, M.H.; Roschewitz, I.; Thies, C.; Tscharntke, T. Differential effects of landscape and management on diversity and density of ground-dwelling farmland spiders. *J. Appl. Ecol.* **2005**, *42* (2), 281–287.
103. Hough-Goldstein, J.A.; Vangessel, M.J.; Wilson, A.P. Manipulation of weed communities to enhance ground-dwelling arthropod populations in herbicide-resistant field corn. *Environ. Entomol.* **2004**, *33* (3), 577–586.
104. Marc, P.; Canard, A.; Ysnel, F. Spiders (Araneae) useful for pest limitation and bioindication. *Agric. Ecosyst. Environ.* **1999**, *74* (1–3), 229–273.
105. Mazzia, C.; Pasquet, A.; Caro, G.; Thenard, J.; Cornic, J.F.; Hedde, M.; Capowiez, Y. The impact of management strategies in apple orchards on the structural and functional diversity of epigeal spiders. *Ecotoxicology* **2015**, *24*, 616–625.
106. Kleijn, D.; van Langevelde, F. Interacting effects of landscape context and habitat quality on flower visiting insects in agricultural landscapes. *Basic Appl. Ecol.* **2006**, *7* (3), 201–214.
107. Schweiger, O.; Maelfait, J.P.; Van Wingerden, W.; Hendrickx, F.; Billeter, R.; Speelmans, M.; Augenstein, I.; Aukema, B.; Aviron, S.; Bailey, D.; Bukacek, R.; Burel, F.; Diekotter, T.; Dirksen, J.; Frenzel, M.; Herzog, F.; Liira, J.; Roubalova, M.; Bugter, R. Quantifying the impact of environmental factors on arthropod communities in agricultural landscapes across organizational levels and spatial scales. *J. Appl. Ecol.* **2005**, *42* (6), 1129–1139.
108. Thomas, J.A. Monitoring change in the abundance and distribution of insects using butterflies and other indicator groups. *Philos. Trans. R. Soc., B* **2005**, *360* (1454), 339–357.
109. Henle, K.; Dziock, F.; Foeckler, F.; Follner, K.; Husing, V.; Hettrich, A.; Rink, M.; Stab, S.; Scholz, M. Study design for assessing species environment relationships and developing indicator systems for ecological changes in floodplains—The approach of the RIVA project. *Int. Rev. Hydrobiol.* **2006**, *91* (4), 292–313.
110. Dormann, C.F.; Schweiger, O.; Augenstein, I.; Bailey, D.; Billeter, R.; de Blust, G.; DeFilippi, R.; Frenzel, M.; Hendrickx, F.; Herzog, F.; Klotz, S.; Liira, J.; Maelfait, J.P.; Schmidt, T.; Speelmans, M.; van Wingerden, W.; Zobel, M. Effects of landscape structure and land-use intensity on similarity of plant and animal communities. *Global Ecol. Biogeogr.* **2007**, *16* (6), 774–787.
111. Villa, M.; Santos, S.A.P.; Marrao, R.; Pinheiro, L.A.; Lopez-Saez, J.A.; Mexia, A.; Bento, A.; Pereira, J.A. Syrphids feed on multiple patches in heterogeneous agricultural landscapes during the autumn season, a period of food scarcity. *Agric. Ecosyst. Environ.* **2016**, *233*, 262–269.

112. Kevan, P.G. Pollinators as bioindicators of the state of the environment: Species, activity and diversity. *Agric. Ecosyst. Environ*. 1999, 74 (1–3), 373–393.
113. Pe'er, G.; Settele, J. The rare butterfly *Tomares nesimachus* (Lycaenidae) as a bioindicator for pollination services and ecosystem functioning in northern Israel. *Isr. J. Ecol. Evol*. **2008**, *54* (1), 111–136.
114. Matos, M.C.B.; Sousa-Souto, L.; Almeida, R.S.; Teodoro, A.V. Contrasting patterns of species richness and composition of solitary wasps and bees (Insecta: Hymenoptera) according to land-use. *Biotropica* **2013**, *45*, 73–79.
115. Paoletti, M.G.; Hu, D.X.; Marc, P.; Huang, N.X.; Wu, W.L.; Han, C.R.; He, J.H.; Cai, L.W. Arthropods as bioindicators in agroecosystems of Jiang Han Plain, Qianjiang City, Hubei China. *Crit. Rev. Plant Sci*. **1999**, *18* (3), 457–465.
116. Vitanovic, E.; Ivezic, M.; Kacic, S.; Katalinic, M.; Durbesic, P.; Barcic, J.I. Arthropod communities within the olive canopy as bioindicators of different management systems. *Spanish J. Agric. Res*. **2018**, *16*, e0301.
117. de Bruyn, L.A.L. Ants as bioindicators of soil function in rural environments. *Agric. Ecosyst. Environ*. 1999, 74 (1–3), 425–441.
118. Hoffmann, B.D. Using ants for rangeland monitoring: Global patterns in the responses of ant communities to grazing. *Ecol. Indic*. **2010**, *10* (2), 105–111.
119. Groc, S.; Delabie, J.H.C.; Fernandez, F.; Petitclerc, F.; Corbara, B.; Leponce, M.; Cereghino, R.; Dejean, A. Litter-dwelling ants as bioindicators to gauge the sustainability of small arboreal monocultures embedded in the Amazonian rainforest. *Ecol. Indic*. **2017**, *82*, 43–49.
120. Magagula, C.N.; Nzima, B.A. Interaction between habitat characteristics and insect diversity using ground beetles (Coleoptera: Carabidae) and ants (Hymenoptera: Formicidae) within a variety of agricultural habitats. *Appl. Ecol. Environ. Res*. **2015**, *13*, 863–876.
121. Sanabria, C.; Lavelle, P.; Fonte, S.J. Ants as indicators of soil-based ecosystem services in agroecosystems of the Colombian Llanos. *Appl. Soil Ecol*. **2014**, *84*, 24–30.
122. Crepaldi, R.A.; Portilho, I.I.R.; Silvestre, R.; Mercante, F.M. Ants as bioindicators of soil quality in integrated crop-livestock system. *Ciencia Rural* **2014**, *44*, 781–787.
123. Springett, J.A.; Gray, R.A.J.; Bakker, L. Influence of agriculture on Enchytraeidae fauna of soils in the south-west of the north island of New Zealand. *Pedobiologia* **1996**, *40* (5), 461–466.
124. Buckerfield, J.C.; Lee, K.E.; Davoren, C.W.; Hannay, J.N. Earthworms as indicators of sustainable production in dry-land cropping in southern Australia. *Soil Biol. Biochem*. 1997, 29 (3–4), 547–554.
125. Paoletti, M.G. The role of earthworms for assessment of sustainability and as bioindicators. *Agric. Ecosyst. Environ*. **1999**, 74 (1–3), 137–155.
126. Suthar, S. Earthworm communities a bioindicator of arable land management practices: A case study in semiarid region of India. *Ecol. Indic*. **2009**, *9* (3), 588–594.
127. Bartlett, M.D.; Briones, M.J.I.; Neilson, R.; Schmidt, O.; Spurgeon, D.; Creamer, R.E. A critical review of current methods in earthworm ecology: From individuals to populations. *Eur. J. Soil Biol*. **2010**, *46* (2), 67–73.
128. Fusaro, S.; Gavinelli, F.; Lazzarini, F.; Paoletti, M. G. Soil biological quality index based on earthworms (QBS-e). A new way to use earthworms as bioindicators in agroecosystems. *Ecol. Indic*. **2018**, *93*, 1276–1292.
129. Campani, T.; Caliani, I.; Pozzuoli, C.; Romi, M.; Fossi, M.C.; Casini, S. Assessment of toxicological effects of raw and bioremediated olive mill waste in the earthworm Eisenia fetida: A biomarker approach for sustainable agriculture. *Appl. Soil Ecol*. **2017**, *119*, 18–25.
130. Bartz, M.L.C.; Pasini, A.; Brown, G.G. Earthworms as soil quality indicators in Brazilian no-tillage systems. *Appl. Soil Ecol*. **2013**, *69*, 39–48.
131. Pelosi, C.; Rombke, J. Enchytraeids as bioindicators of land use and management. *Appl. Soil Ecol*. **2018**, *123*, 775–779.
132. Filser, J.; Fromm, H.; Nagel, R.F.; Winter, K. Effects of previous intensive agricultural management on microorganisms and the biodiversity of soil fauna. *Plant Soil* **1995**, *170* (1), 123–129.

133. Heisler, C. Collembola and Gamasina: Bioindicators for soil compaction. *Acta Zool. Fenn.* **1995**, *196*, 229–231.
134. Wardle, D.A.; Yeates, G.W.; Watson, R.N.; Nicholson, K.S. The detritus food-web and the diversity of soil fauna as indicators of disturbance regimes in agro-ecosystems. *Plant Soil* **1995**, *170* (1), 35–43.
135. Porazinska, D.L.; McSorley, R.; Duncan, L.W.; Graham, J.H.; Wheaton, T.A.; Parsons, L.R. Nematode community composition under various irrigation schemes in a citrus soil ecosystem. *J. Nematol.* **1998**, *30* (2), 170–178.
136. Porazinska, L.; McSorley, R.; Duncan, L.W.; Gallaher, R.N.; Wheaton, T.A.; Parsons, L.R. Relationships between soil chemical status, soil nematode community, and sustainability indices. *Nematropica* **1998**, *28* (2), 249–262.
137. van Straalen, N.M. Evaluation of bioindicator systems derived from soil arthropod communities. *Appl. Soil Ecol.* 1998, 9 (1–3), 429–437.
138. Enami, Y.; Shiraishi, H.; Nakamura, Y. Use of soil animals as bioindicators of various kinds of soil management in northern Japan. *JARQ* **1999**, *33* (2), 85–89.
139. Foissner, W. Soil protozoa as bioindicators: Pros and cons, methods, diversity, representative examples. *Agric. Eco- syst. Environ.* **1999**, *74* (1–3), 95–112.
140. Koehler, H.H. Predatory mites (Gamasina, Mesostigmata). *Agric. Ecosyst. Environ.* 1999, 74 (1–3), 395–410.
141. Paoletti, M.G.; Hassall, M. Woodlice (Isopoda: Oniscidea): Their potential for assessing sustainability and use as bioindicators. *Agric. Ecosyst. Environ.* **1999**, *74* (1–3), 157–165.
142. Yeates, G.W.; Bongers, T. Nematode diversity in agroecosystems. *Agric. Ecosyst. Environ.* **1999**, *74* (1–3), 113–135.
143. Lenz, R.; Eisenbeis, G. Short-term effects of different tillage in a sustainable farming system on nematode community structure. *Biol. Fertil. Soils* **2000**, *31* (3–4), 237–244.
144. Neher, D.A. Role of nematodes in soil health and their use as indicators. *J. Nematol.* **2001**, *33* (4), 161–168.
145. Li, Y.; Wu, J.; Chen, H.; Chen, J. Nematodes as bioindicator of soil health: Methods and applications. *Yingyong Shengtai Xuebao* **2005**, *16* (8), 1541–1546.
146. Briar, S.S.; Jagdale, G.B.; Cheng, Z.; Hoy, C.W.; Miller, S.A.; Grewal, P.S. Indicative value of soil nematode food web indices and trophic group abundance in differentiating habitats with a gradient of anthropogenic impact. *Environ. Bioindic.* **2007**, *2* (3), 146–160.
147. Gulvik, M.E. Mites (Acari) as indicators of soil biodiversity and land use monitoring: A review. *Pol. J. Ecol.* **2007**, *55*, 415–440.
148. Menta, C.; Leoni, A.; Bardini, M.; Gardi, C.; Gatti, F. Nematode and microarthropod communities: Comparative use of soil quality bioindicators in covered dump and natural soils. *Environ. Bioindic.* **2008**, *3* (1), 35–46.
149. Gergocs, V.; Hufnagel, L. Application of oribatid mites as indicators (review). *Appl. Ecol. Environ. Res.* **2009**, *7* (1), 79–98.
150. Aspetti, G.P.; Boccelli, R.; Ampollini, D.; Del Re, A.A.M.; Capri, E. Assessment of soil-quality index based on microarthropods in corn cultivation in northern Italy. *Ecol. Indic.* **2010**, *10* (2), 129–135.
151. Sánchez-Moreno, S.; Jiménez, L.; Alonso-Prados, J.L.; García-Baudín, J.M. Nematodes as indicators of fumigant effects on soil food webs in strawberry crops in southern Spain. *Ecol. Indic.* **2010**, *10* (2), 148–156.
152. Pothula, S.K.; Grewal, P.S.; Auge, R.M.; Saxton, A.M.; Bernard, E.C. Agricultural intensification and urbanization negatively impact soil nematode richness and abundance: A meta-analysis. *J. Nematol.* 2019, *51*, UNSP e2019–11.
153. du Preez, G.C.; Daneel, M.S.; Wepener, V.; Fourie, H. Beneficial nematodes as bioindicators of ecosystem health in irrigated soils. *Appl. Soil Ecol.* **2018**, *132*, 155–168.

154. Ozden, O.; Hodgson, D.J. The impact of tillage and chemical management on beneficial arthropods in Mediterranean olive groves. *Isr. J. Ecol. Evol.* **2017**, *63*, 14–18.
155. Rudisser, J.; Tasser, E.; Peham, T.; Meyer, E.; Tappeiner, U. The dark side of biodiversity: Spatial application of the biological soil quality indicator (BSQ). *Ecol. Indic.* **2015**, *53*, 240–246.
156. Wang, K.H.; Radovich, T.; Pant, A.; Cheng, Z.Q. Integration of cover crops and vermicompost tea for soil and plant health management in a short-term vegetable cropping system. *Appl. Soil Ecol.* **2014**, *82*, 26–37.
157. Nesbitt, J.E.; Adl, S.M. Differences in soil quality indicators between organic and sustainably managed potato fields in Eastern Canada. *Ecol. Indic.* **2014**, *37*, 119–130.
158. Antonious, G.F. Impact of soil management and two botanical insecticides on urease and invertase activity. *J. Environ. Sci. Health, Part B* 2003, B38 (4), 479–488.
159. Benintende, S.M.; Benintende, M.C.; Sterren, M.A.; De Battista, J.J. Soil microbiological indicators of soil quality in four rice rotations systems. *Ecol. Indic.* **2008**, *8* (5), 704–708.
160. Lagomarsino, A.; Moscatelli, M.C.; Di Tizio, A.; Manci-nelli, R.; Grego, S.; Marinari, S. Soil biochemical indicators as a tool to assess the short-term impact of agricultural management on changes in organic C in a Mediterranean environment. *Ecol. Indic.* 2009, *9* (3), 518–527.
161. Geisen, S.; Mitchell, E.A.D.; Adl, S.; Bonkowski, M.; Dunthorn, M.; Ekelund, F.; Fernández, L.D.; Jousset, A.; Krashevska, V.; Singer, D.; Spiegel, F.W.; Walochnik, J.; Lara, E. Soil protists: A fertile frontier in soil biology research. *FEMS Microbiol. Rev.* **2018**, *42*, 293–323.
162. Berard, A.; Rimet, F.; Capowiez, Y.; Leboulanger, C. Procedures for determining the pesticide sensitivity of indigenous soil algae: A possible bioindicator of soil contamination? Arch. *Environ. Contam. Toxicol.* **2004**, *46* (1), 24–31.
163. Heger, T.J.; Straub, F.; Mitchell, E.A.D. Impact of farming practices on soil diatoms and testate amoebae: A pilot study in the DOK-trial at Therwil, Switzerland. *Eur. J. Soil Biol.* **2012**, *49*, 31–36.
164. Stenberg, B. Monitoring soil quality of arable land: Microbiological indicators. *Acta Agric. Scand., Sect. B* **1999**, *49* (1), 1–24.
165. Kaur, A.; Chaudhary, A.; Choudhary, R.; Kaushik, R. Phospholipid fatty acid—A bioindicator of environment monitoring and assessment in soil ecosystem. *Curr. Sci.* **2005**, *89* (7), 1103–1112.
166. Jordan, D.; Miles, R.J.; Hubbard, V.C.; Lorenz, T. Effect of management practices and cropping systems on earthworm abundance and microbial activity in Sanborn field: A 115-year-old agricultural field. *Pedobiologia* **2004**, *48* (2), 99.
167. Subbarao, K.V.; Kabir, Z.; Martin, F.N.; Koike, S.T. Management of soilborne diseases in strawberry using vegetable rotations. *Plant Dis.* **2007**, *91* (8), 964–972.
168. Smukler, S.M.; Jackson, L.E.; Murphree, L.; Yokota, R.; Koike, S.T.; Smith, R.F. Transition to large-scale organic vegetable production in the Salinas Valley, California. *Agric. Ecosyst. Environ.* **2008**, *126* (3–4), 168–188.
169. Oehl, F.; Laczko, E.; Bogenrieder, A.; Stahr, K.; Bosch, R.; van der Heijden, M.; Sieverding, E. Soil type and land use intensity determine the composition of arbuscular mycorrhizal fungal communities. *Soil Biol. Biochem.* **2010**, *42*, 724–738.
170. Bell, N.L.; Adam, K.H.; Jones, R.J.; Johnson, R.D.; Mtandavari, Y.F.; Burch, G.; Cave, V.; Cameron, C.; Maclean, P.; Popay, A.J.; Fleetwood, D. Detection of invertebrate suppressive soils, and identification of a possible biological control agent for meloidogyne nematodes using high resolution rhizosphere microbial community analysis. *Front. Plant Sci.* 2016, *7*, 1946.
171. Elmholt, S. Microbial activity, fungal abundance, and distribution of *Pénicillium* and *Fusarium* as bioindicators of a temporal development of organically cultivated soils. *Biol. Agric. Hort.* **1996**, *13* (2), 123–140.
172. Padoa-Schioppa, E.; Baietto, M.; Massa, R.; Bottoni, L. Bird communities as bioindicators: The focal species concept in agricultural landscapes. *Ecol. Indic.* **2006**, *6* (1), 83–93.
173. Bar, A.; Loffler, J. Ecological process indicators used for nature protection scenarios in agricultural landscapes of SW Norway. *Ecol. Indic.* **2007**, *7* (2), 396–411.

174. Robledano, F.; Esteve, M.A.; Farinós, P.; Carreño, M.F.; Martínez-Fernández, J. Terrestrial birds as indicators of agricultural-induced changes and associated loss in conservation value of Mediterranean wetlands. *Ecol. Indic.* **2010**, *10* (2), 274–286.
175. Morelli, F.; Girardello, M. Buntings (Emberizidae) as indicators of HNV of farmlands: A case of study in Central Italy. *Ethol. Ecol. Evol.* **2014**, *26*, 405–412.
176. Albrecht, H. Suitability of arable weeds as indicator organisms to evaluate species conservation effects of management in agricultural ecosystems. *Agric. Ecosyst. Environ.* **2003**, *98* (1–3), 201–211.
177. Sturz, A.; Matheson, B.; Arsenault, W.; Kimpinski, J.; Christie, B. Weeds as a source of plant growth promoting rhizobacteria in agricultural soils. *Can. J. Microbiol.* **2001**, *47,* 1013–1024.
178. Aavik, T.; Liira, J. Agrotolerant and high nature-value species—Plant biodiversity indicator groups in agroecosystems. *Ecol. Indic.* **2009**, *9* (5), 892–901.
179. Höft, A.; Müller, J.; Gerowitt, B. Vegetation indicators for grazing activities on grassland to be implemented in outcome-oriented agri-environmental payment schemes in north-east Germany. *Ecol. Indic.* **2010**, *10* (3), 719–726.
180. Fanfarillo, E.; Kasperski, A.; Giuliani, A.; Cicinelli, E.; Latini, M.; Abbate, G. Assessing naturalness of arable weed communities: A new index applied to a case study in central Italy. *Biol. Agric. Hort.* **2018**, *34*, 232–244.
181. Une, Y. Illustrated Poster of Rice and Living Organisms; Nou-to-shizen-no-kenkyujo: Fukuoka, Japan, 2004 (in Japanese).

18

Ecological Indicators: Eco-Exergy to Emergy Flow

Simone Bastianoni,
Luca Coscieme, and
Federico M. Pulselli

Introduction

Living systems can be viewed as dissipative structures, self-organizing systems that dissipate energy for the maintenance of organization.[1] They exist by the transforming available energy potentials, building new structures as a consequence of the process.[2] To study the ecosystem in this view, we need proper tools able to take into account the biological time[3] that has been necessary for the "creative" planning and construction of dissipative structures in the system, through the concentration of energies and matter.

The study of the behavior of a single species in a system provides knowledge of the system's parts. Similar to a jigsaw puzzle, we need to know the single drawing on each piece, as well as the boundaries/shapes of the pieces, in order to properly choose where each piece must be placed, so that we can reconstruct the entire drawing. However, in a very complex context, we are not capable of reconstructing the whole drawing if we do not have a reference, a picture of the whole system that can guide us, e.g., the drawing on the box of the puzzle. Holistic indicators, orientors, or goal functions give us this kind of information, guiding the disciplinary research in a conjunction of efforts, in order to have the best description of the system. In fact, extending the study from a "simple" description of the system in a given time to a description of the system evolving through various states, the holistic indicators allow us to understand if the system under study is globally following a path that will take the system to a "better" or to a "worse" state.[4]

As maintained by Müller and Leupelt,[5] goal functions and orientors have been developed as holistic measures of the global performance of ecosystems, of what we could call, in a broad sense, complexity. Fath, Patten, and Choi[6] showed that 10 extremal principles involving orientors (power, storage, empower, emergy, ascendency, dissipation, cycling, residence time, specific dissipation, and empower/exergy ratio) can be unified by ecological network notation.

In its evolution, a system changes properties and behaviors, like human beings or other organisms change their "reactions" to the surrounding "signals" during their evolution. Every different holistic indicator is appropriate to describe the system in a particular state or phase of its evolution or a particular characteristic of the same system. When studying a dynamic system in evolution, holistic indicators (and the joint use of them) seem to be more powerful than nonsystemic indicators. Only this approach seems to be truly useful to describe the system and its dynamics, i.e., from a "young system" state, through intermediate states (i.e., of reorganization after possible perturbations), toward a climax mature stage.

This holistic approach is relatively new, but it is very efficacious and promising for ecosystem investigation and management.

In this entry, we describe the combined use of two holistic indicators, eco-exergy and emergy, composing a ratio that is able to describe the system during its evolution. Moreover, it adds information to that obtained from the use of the two orientors separately. This kind of approach is in line with the book *A New Ecology: Systems Perspective* proposed by Sven Jørgensen and coauthors[4] and with what Fath, Patten, and Choi[6] maintained in their fundamental paper on the use of a plurality of "goal functions," highlighting their complementarity and interdependency.

A Description of the System Using Emergy

A class of indicators, based on the concept of emergy, are able to evaluate the convergence of matter and energy (several inputs) to a system. On the basis of a thermodynamic hierarchy of energy, and starting from solar energy input to the earth, emergy provides a measure of the environmental work necessary to generate an item or a flow (which could be, for example, an input to any system).

The concept of emergy derives from a reflection about the concept of energy quality.[7] The second principle of thermodynamics states that energy transformations imply an irreversible degradation of energy. In living systems and ecosystems, inputs of energy are transformed, and the energy quality changes. A portion of diluted sunlight is lost as heat, and a portion is concentrated into forms that are more able to do work, and/or more flexible to be used. According to Odum, many joules of low quality are needed for a few joules of high quality. In the case of a typical web of connections, like the food chain, "at each stage, energy is degraded as a necessary part of transforming a lower quality energy to a higher quality one in lesser quantity. The energy flows decrease as one goes up the food chain."[7] "A joule of sunlight, a joule of coal, a joule of human effort are of different quality and represent vastly different convergences of energy in their making."[8] Therefore, in many cases, it is not correct to use energy to describe the dynamics and behavior of systems. Solar energy can be considered the basis upon which energy transformations in the biosphere occur. Emergy, expressed in solar emergy joule (seJ or, according to a recent proposal, semj) is defined as the quantity of solar energy directly or indirectly necessary to produce a flow or a product. In the case of ecosystems, the emergy flow (empower) is considered, which is the quantity of solar energy directly or indirectly necessary to support the system and its level of organization.[9–11]

To compare all kinds of energy on the common basis of solar energy, solar transformity has been utilized, defined as the solar energy directly and indirectly required to generate 1 J of a product. It is a conversion factor that takes into account the position of one energy form in a sort of thermodynamic hierarchy in the biosphere. The solar transformity of the sunlight absorbed by the earth is 1 seJ/J by definition.[12]

To quantify the emergy of a product or system, all the inputs to the system or production process must be quantified and expressed in seJ by means of suitable transformities, which are used to convert different flows of energy into equivalent solar energy.[9] In case of matter inputs, the specific emergy (expressed in seJ/g or another unit) is used to convert mass into equivalent solar energy. Recently, the concept of unit emergy value has been used, independently of the unit in which the flows are expressed (energy or mass). The total emergy of a system (EmS) is given, approximately, by the sum of the energy content (*Ei*) of the *i*th input to the system multiplied by the corresponding transformity (Tr*i*), while avoiding double counting any inputs (see Bastianoni[13] for a thorough analysis of this calculation method):

$$\mathrm{Em_S} = \Sigma E_i \cdot \mathrm{Tr}_i \tag{1}$$

Every flow can be expressed by means of its solar equivalent, and a system of environmental accounting based on emergy can be implemented.[14]

Emergy represents the convergence of different kinds of energy to a system (E_i) times the quantity of solar energy that has been necessary to make available one unit (Tr_i,). It is not a state function because it depends on the kinds of energy and the process that is used to obtain a certain item (that can be a product or a given state of the system). Therefore, emergy enables us to identify, quantify, and weigh the inputs that feed an ecosystem. If natural selection has been given time to operate, the higher the emergy flux necessary to sustain a system, component, or a process, the higher the hierarchical level and the usefulness that can be expected for that entity (maximum empower principle[7]).

In a pristine natural system, self-organizing from young states, this maximum empower principle (derived from the maximum power principle[15]) is realized, and the continuous increase of emergy is an indication of a proper evolution toward a mature system (climax stage).[16] Therefore, empower has been proposed as an ecosystem health indicator.[17]

In "real-world" cases, where natural systems interact with human systems and dynamics, the increase of emergy (as a consequence of the increase in energetic inputs that reach the ecosystem) is not always "good" in the sense that it will support the evolution of the system toward a climax stage. In fact, a portion of the inputs that the ecosystem receives is not used to build structures in order to maintain the nonequilibrium state (e.g., nutrient overflow, pollutants). Emergy flow alone cannot be used as a "reference direction" indicator[18] if the system is not a pure pristine system.

A Description of the System Using Eco-Exergy

The usable energy input to a system is converted into genetic information, biomass storages, and a relation network.[7,19,20] Structural complexity and biodiversity influence the possible evolutions of the system toward another (more or less) stable state.[21]

Eco-exergy expresses the development of ecosystems by the increase in the work capacity,[4] considering the biomass stored in a system and its genetic information.[22–25]

According to Jørgensen,[26,27] we can distinguish between technological exergy and eco-exergy: the former uses the environment as a reference state and is able to measure the useful energy provided by a production process; the latter uses as a reference state the same ecosystem at thermodynamic equilibrium. Eco-exergy thus estimates the distance of an ecosystem from thermodynamic equilibrium and is given by the formula

$$\mathrm{Eco-Ex} = \Sigma \beta_i \cdot c_i \tag{2}$$

where c_i is the concentration of the ith component of the ecosystem and β_i is the weighting factor that accounts for the genetic information that the component carries (for a list of β values.[28]

When the available inputs to the system are used to build up new biomass and/or complexity, the system is tending to its climax stage (maximum exergy principle[4]). Eco-exergy is a state-based descriptor of a system's structure (and functions, networks, interactions) based on usable energy and information. It has also been used as an ecosystem health indicator.[24,25,29]

Ratio of Eco-Exergy to Emergy Flow

Emergy and eco-exergy can be considered as complementary entities, with the former accounting for the amount of basic energy (solar) required to support a process or an ecosystem and the latter being the level of organization reached by a system.

The need to compare the emergy flow that supports an ecosystem with the consequent ecosystem reaction was already clear to Odum, who tried to assess the ecosystem response using the emergy-to-information ratio[7] as a measure of information hierarchy. Odum used the mass of DNA as information carrier and looked at the ratio of emergy to information measured in bits, considering that the emergy required per bit may give an indication on the usefulness of the information in an adapted system.[7,30] The emergy-to-information ratio, also used by Keitt,[31] has a problem that consists in the arbitrariness of the choice of the system's basic element: an atom or an individual of a species, a letter of an alphabet, or a gene as the basic "symbol." Bastianoni and Marchettini[30] have proposed joining emergy with eco-exergy in order to measure the ecosystem structure. The ratio between eco-exergy and emergy flow indicates the organization or structure of an ecosystem per unit of solar emergy flow required to produce or maintain it.[30] Bastianoni and Marchettini[30] first introduced this- relation as the ratio of emergy (flow) to eco-exergy. This choice was made in order to maintain coherence with the definition of transformity (i.e., the emergy that contributes to a production system divided by the energy content of a product). In fact, the ratio of emergy flow to eco-exergy represents an empower converging to a certain system divided by the eco-exergy of the whole system. Actually, the inverse seems more comprehensible, where the effect (eco-exergy) is at the numerator and the requirement is at the denominator, as in any efficiency indicator.

The role of information and structure is fundamental when we approach the study of complex systems, such as an ecosystem. The use of eco-exergy adds something to the classical exergy approach (for an overview, see Jørgensen and Mejer,[22] Jørgensen,[27] and Sciubba and Wall[32]), which does not take into account information content. For instance, the difference between a living organism and a dead one is not related to the classical exergetic content that is in fact the same, but is related to the capability of the living system to use the information content in its DNA.[33]

The ratio of eco-exergy to emergy flow has also been considered as an indicator of the efficiency[30] in transforming available inputs, evaluated in emergy terms, into the structure, information, and ecosystem organization, evaluated in eco-exergy terms. In fact, it represents the state of the system (as eco-exergy) per unit input (as empower). Strictly speaking, its unit is JyrseJ^{-1} (we maintain this representation even if β values may influence the pure thermodynamic nature of eco-exergy and its unit). Since the dimensions are those of time, it cannot be regarded as a real efficiency (which is dimensionless) but more as a proxy of efficiency. The higher its value, the higher the efficiency of the system; if the eco-exergy/empower ratio tends to increase (apart from oscillations due to normal biological cycles), it means that natural selection is making the system follow a thermodynamic path that will bring the system to a higher organizational level.[33,34] As an efficiency indicator, the ratio of eco-exergy to empower enlarges the viewpoint of a pure exergetic approach, where the exergy degraded and the eco-exergy stored for various ecosystems are compared: using emergy, there is a recognition of the fact that solar radiation is the driving force of all the energy (and exergy) flows on the biosphere, which is essential when important "indirect" inputs (of solar energy) are also present in a process, and must be identified, weighted, and finally, taken into account.

Fath, Patten, and Choi[6] have identified the ratio of eco-exergy to emergy flow as one of the possible orientors of an ecosystem: they link the emergy flow to the total system throughput and eco-exergy to the total system storage, therefore connecting the maximization of the ratio of eco-exergy to emergy flow with the maximization of residence time.

Applications in Environmental Investigation and Management

A Definition of Ecosystem Development

Ecological orientors and goal functions can indicate some aspects of the degree of naturalness of ecosystems; they provide a good basis for finding usable indicators for ecosystem health, ecological integrity, and sustainability;[5] they can also be used to evaluate the strength of human impact and an ecosystem's structural carrying capacity.

The ratio of eco-exergy to empower has often been applied in order to assess ecosystem health: in fact, ecosystems with different empower and eco-exergy can be compared with each other, also regarding their behavior and performance. In general, we can say that in natural systems, where selection has acted undisturbed for a long time, the ratio of eco-exergy to empower is higher and decreases with the progressive introduction of artificial inputs and stress factors that make the emergy flow higher and lower the eco-exergy content of the ecosystem. In the evolutionary process, close to the steady state (climax), the ratio of eco-exergy to empower tends to increase, which means that the system uses all the materials and energy available to reach a higher eco-exergy content. The same systems, once the climax has been reached, will remain in such a state for some time and can then grow/develop again only if further energy and/or materials are available. In the latter case, new sources of energy (or better emergy) can be used to build up new biomass and/or complexity of the ecosystem (stored eco-exergy). In terms of the ratio of eco-exergy to empower, when a system is relatively young and acquires new inputs, the ratio tends to be lower; when the system is developing toward the climax stage, the ratio tends to rise.[2,35]

This approach is helpful when different kinds of ecosystems are compared, natural and artificial. The former might have different quantities and qualities of available inputs, while the latter cannot be compared only on the basis of its "state" but also considering its requirements to develop and sustain the state itself. The results of a study on eight different aquatic ecosystems[30,35–37] demonstrated that the highest level of efficiency (in the exergy/empower sense) is obtained by a seminatural system within the lagoon of Venice, a farming basin developed over several centuries. Its efficiency is of the same order of magnitude as natural systems, but it is higher than systems with limited human input. Furthermore, the efficiency of the seminatural system is two orders of magnitude greater than artificial ecosystems by virtue of a higher level of organization and less need for external input.[36,38] Application to agroecosystems can be found in Bastianoni et al.[39] The ratios of eco- exergy to emergy and eco-exergy to empower have been also proposed to assess the self-organization efficiency of forest ecosystems.[40]

A Definition of Ecosystem Health and Pollution

A qualitative or quantitative change in the set of inputs can contribute to a change in a system's self-organization pattern and a system's different responses. Moreover, controlled human intervention can make a positive contribution to the system in terms of organization, information complexity, etc., (the eco-exergy of the system) that more than offset the environmental cost of the same intervention (the emergy flow corresponding to the human-induced inputs to the system).

If we consider the emergy flow to a system to vary between two equal and contiguous intervals, we will indicate the variation of emergy flow with ΔEM. Consequent changes in system organization can be measured by the variation of the exergy content of the system ΔEX. The quantity $\sigma = \Delta EX/\Delta EM$, with the dimensions of $J{\cdot}s{\cdot}seJ^{-1}$, represents the change of level of organization (exergy) of the system under study, when it is related to a change of the emergy flow. It is a quantity that is specific to the inputs that are subtracted or added. If σ is positive, the addition of emergy input gives rise to further organization (increasing eco-exergy), whereas a lowering of emergy has a negative effect on the system (decreasing eco-exergy). On the other hand, when σ is negative, a higher emergy flow implies a decrease in organization, whereas a lower quantity of one or more inputs causes increasing organization. In these two last cases, the inputs can generally be viewed as pollutants: in an evolutionary perspective, if they are removed, the system self-organizes; if they are added, the system is damaged. This provides a definition of pollution based on two holistic orientors representing system dynamics. A first-level observation of the behavior of σ (and of the system as a whole) gives information on the existence of pollutants; a deeper analysis can identify the intensity of pollution that is given by the sensibility of eco-exergy relative to a change in emergy flow.[41]

Holistic Interpretation of Ecosystem Services

Emergy and eco-exergy can be related to the concept of "ecosystem services." Ecosystem services are the benefits human populations derive, directly or indirectly, from ecosystem functions (e.g., food provision or waste assimilation).[42] This concept derives from a reconceptualization of ecosystem functioning from an anthropocentric viewpoint.[43]

In general, ecosystems utilize flows of energy and matter from the environment to maintain themselves as far as possible from thermodynamic equilibrium and to survive, grow, and/or develop. The degree of development and the efficacy with which these flows are used up and processed depend on the state/structure/organization of the ecosystem, which is a particular configuration of the abiotic–biotic system components, characterized by specific relationships between living organisms and nonliving surroundings. The outputs of an ecosystem are all flows of energy and matter moving from the system to the environment, as well as all goods and services useful for humans.

A relationship between the inflows of energy and matter supporting the ecosystem and the services it provides has been investigated by Pulselli, Coscieme, and Bastianoni,[21] who noted that in this input–output representation of ecosystems, the emergy flow (input) supporting an ecosystem and the value of the services (output) it provides are rather independent from each other, because the former depends on natural dynamics and the latter on the utility humans (decide to) draw from nature, which may vary from case to case. Despite this, at the global level, it has been calculated that nature contributes to humans not only more (as Costanza et al.[42] demonstrated) but also in a more efficient way than do all the world economic infrastructures. In fact, if we divide the world ecosystem service value by the emergy flow to the biosphere, we obtain the amount of money that is, on average, produced by 1 seJ of solar emergy feeding the global ecosystem. This ratio combines an amount of money that is not really circulating in the global economy and the flow of all renewable resources that feed the planet (sunlight, geothermal heat, rain, wind, etc.). It can be considered as a potential efficiency of the entire biosphere in providing a kind of economic wealth for humans (since at least a portion of it can be converted into real economic utility/benefit) based only on its natural functioning.[21]

The reciprocal of the above relation, i.e., the ratio of the global emergy flow to the total value of ecosystem services, is between 5.09×10^{11} and 1.51×10^{n} seJ/€ (depending on the minimum and maximum values calculated by Costanza et al.[42]). Both the maximum and minimum values are lower than values traditionally calculated for national economies, which in emergy theory are known as the emergy-to-money ratio (EMR). The EMR is given by the ratio of the emergy flow of a country (including both natural and commercial man-induced flows) to its GDP, and its order of magnitude is, in general, 10^{12} seJ/€ or more (for an overview of national values.[44] This means that the global ecosystem uses, on average, less emergy than a national economy per unit money provided to humans.

Jørgensen[26] proposed an approach to connect an ecosystem's structure and organization descriptor (eco-exergy) to a user-side measure (the value of ecosystem services), highlighting a relation between a biophysical and an economic evaluation of the environment. The calculation of ecosystem services through eco-exergy resulted in values higher than those proposed by Costanza et al.,[42] because eco-exergy represents the annual work capacity increase of an ecosystem that can be translated into the set of the possible services it can offer (not only the services that anthropic systems actually utilize) and can be compared with the actual flow of services utilized.

Ecosystem dynamics can be represented through a generic and complete input–state–output scheme. In this sense, we can imagine a kind of 3-D diagram, with the inflows of resources, measured in terms of solar emergy, on the x axis; the work capacity embodied in the system biota, expressed in terms of eco-exergy, on the y axis; and the useful services for humans, valuable in economic terms, on the z axis.[45] This multidimensional holistic approach makes it clear that inputs are used up, directly or indirectly, to produce services in output and/or to develop the system, and enables us to have an indication of changes in ecosystem dynamics, structure, and services. Within this framework, a thermodynamic/

socio-ecological evolutionary time path can be acknowledged: from young systems, to climax-stage systems, to socio-ecological integrated systems.

Conclusion

The ratio of eco-exergy to emergy flow is the combination of two thermodynamics-based orientors: emergy flow, which quantifies the amount of resources necessary for the system to survive, and eco-exergy, which represents the actual state of the ecosystem in terms of work capacity and distance from thermodynamic equilibrium. The joint use of these two entities adds information that is useful for the investigation of the behavior of the system during its evolution. It is a measure of the ability of a system to reach and maintain a given structure (as eco-exergy) per unit input (as emergy); it is therefore a measure of the efficiency of the system in transforming available resources into organization. The use of the ratio of eco-exergy to emergy flow is important when investigating the evolution of an ecosystem or the effects on ecosystem dynamics of human intervention and infrastructures. Two further applications of this entity have been described too, which can be useful in the field of environmental management: according to the change in exergy due to a change in emergy flow, we can identify potential pollutants and define the intensity of pollution; the combined use of eco-exergy and emergy can also help in assessing the role of ecosystem services in human well-being.

References

1. Prigogine, I. *From Being to Becoming: Time and Complexity in the Physical Sciences*; Freeman: New York, 1980.
2. Bastianoni, S.; Pulselli, F.M.; Rustici, M. Exergy versus emergy flow in ecosystems: Is there an order in maximization? Ecol. Indic. **2006**, *6*, 58–62.
3. Tiezzi, E. *The End of Time*; Wit Press: Southampton, U.K., 2003.
4. Jørgensen, S.E.; Fath, B.D.; Bastianoni, S.; Marques, J.C.; Müller, F.; Nielsen, S.N.; Patten, B.C.; Tiezzi, E.; Ulano- wicz, R.E. *A New Ecology: Systems Perspective*; Elsevier: Amsterdam, 2007.
5. Müller, F.; Leupelt, M. *Eco Targets, Goal Functions, and Orientors*; Springer-Verlag: New York, 1998.
6. Fath, B.D.; Patten, B.C.; Choi, J.S. Complementarity of ecological goal functions. J. Theor. Biol. **2001**, *208*, 493–506.
7. Odum, H.T. Self organization, transformity and information. Science **1988**, *242*, 1132–1139.
8. Odum, H.T. Emergy and biogeochemical cycles. In *Ecological Physical Chemistry*; Rossi, C., Tiezzi, E., Eds.; Elsevier: Amsterdam, 1991; 25–26.
9. Odum, H.T. *Environmental Accounting. Emergy and Environmental Decision Making*; John Wiley and Sons: New York, 1996.
10. Odum, H.T. *Emergy of Global Processes, Folio #2. Handbook of Emergy Evaluation*; Center for Environmental Policy, University of Florida: Gainesville, FL, 2000.
11. Odum, H.T.; Brown, M.T.; Brandt-Williams, S. *Introduction and Global Budget, Folio #1. Handbook of Emergy Evaluation*; Center for Environmental Policy, University of Florida: Gainesville, FL, 2000.
12. Brown, M.T.; Ulgiati, S. Emergy analysis and environmental accounting. In *Encyclopedia of Energy*; Cleveland, C., Ed.; Academic Press, Elsevier: Oxford, 2004; 329–359.
13. Bastianoni, S.; Morandi, F.; Flaminio, T.; Pulselli, R.M.; Tiezzi, E.B.P. Emergy and emergy algebra explained by means of ingenuous set theory. Ecol. Modell. **2011**, 222, 2903–2907.
14. Hau, J.L.; Bakshi, B.R. Promise and problems of emergy analysis. Ecol. Modell. **2004**, *178*, 215–225.
15. Lotka, A.J. Contribution to the energetics of evolution. Proc. National Acad. Sci. U. S. A. **1922**, 8, 147–151.

16. Campbell, D.E. Proposal for including what is valuable to ecosystems in environmental assessments. Environ. Sci. Technol. **2001**, *35*, 2867–2873.
17. Campbell, D.E. Using energy systems theory to define, measure, and interpret ecological integrity and ecosystem health. Ecosyst. Health **2000**, *6* (3), 181–204.
18. Samhouri, J.F.; Levin, P.S.; James, C.A.; Kershner, J.; Williams, G. Using existing scientific capacity to set targets for ecosystem-based management: A Puget Sound case study. Mar. Policy **2011**, *35*, 508–518.
19. Jørgensen, S.E.; Patten, B.C.; Straškraba, M. Ecosystems emerging: 4. Growth. Ecol. Modell. **2000**, *126*, 249–284.
20. Ulanowicz, R.E. *Ecology, the Ascendent Perspective*; Columbia University Press: New York, 1997.
21. Pulselli, F.M.; Coscieme, L.; Bastianoni, S. Ecosystem services as a counterpart of emergy flows to ecosystems. Ecol. Modell. **2011**, *222*, 2924–2928.
22. Jørgensen, S.E.; Mejer, H.F. A holistic approach to ecological modelling. Ecol. Modell. **1979**, 7, 169–189.
23. Jørgensen, S.E.; Mejer, H.F. Application of exergy in ecological models. In *Progress in Ecological Modelling*; Dubois, D., Ed.; Editions CEBEDOC: Liege, Belgium, 1981; 311–347.
24. Jørgensen, S.E. Application of holistic thermodynamic indicators. Ecol. Indic. **2006**, *6*, 24–29.
25. Jørgensen, S.E. Eco-exergy as an ecosystem health indicator. In *Encyclopedia of Ecology*; J0rgensen, S.E., Fath, B., Eds.; Elsevier: Amsterdam, 2008; 977–979.
26. Jørgensen, S.E. Ecosystem services, sustainability and thermodynamic indicators. Ecol. Complexity **2010**, 7, 311–313.
27. Jørgensen, S.E. Exergy. In *Encyclopedia of Ecology*;, Jørgensen, S.E., Fath B.D., Eds.; Elsevier: Oxford, 2008; 1498–1509.
28. Jørgensen, S.E.; Ladegaard, N.; Debeljak, M.; Marques, J.C. Calculations of exergy for organisms. Ecol. Modell. **2005**, *185* (2–4), 165–175.
29. Zaldivar, J.M.; Austoni, M.; Plus, M.; De Leo, G.A.; Giordani, G.; Viaroli, P. Ecosystem health assessment and bioeconomic analysis in coastal lagoon. In *Handbook of Ecological Indicator for Assessment of Ecosystem Heath*; Jørgensen, S.E., Costanza, R., Xu, F.L., Eds.; CRC Press: Boca Raton, FL, 2005; 163–184.
30. Bastianoni, S., Marchettini, N. Emergy/exergy ratio as a measure of the level of organization of systems. Ecol. Modell. **1997**, *99*, 33–40.
31. Keitt, T.H. *Hierarchical Organization of Energy and Information in a Tropical Rain Forest Ecosystem*; M.S. Thesis, University of Florida: Gainesville, FL, 1991.
32. Sciubba, E.; Wall, G. A brief commented history of exergy from the beginnings to 2004. Int. J. Thermodyn. **2007**, *10* (1), 1–26.
33. Tiezzi, E. *Steps Towards an Evolutionary Physics*; Wit Press: Southampton, U.K., 2006.
34. Pulselli, F.M.; Gaggi, C.; Bastianoni, S. Eco-Exergy to emergy flow ratio for the assessment of ecosystem health. In *Handbook of Ecological Indicators for Assessment of Ecosystem Health*, 2nd Ed.; Jørgensen, S.E., Xu, F.L., Costanza, R., Eds.; CRC Press: Boca Raton, FL, 2010; 113–124.
35. Bastianoni, S. Eco-Exergy to emergy flow ratio. In *Encyclopedia of Ecology*; Jørgensen, S.E., Fath, B., Eds.; Elsevier: Oxford, 2008; 979–983.
36. Bastianoni, S. Use of thermodynamic orientors to assess the efficiency of ecosystems: A case study in the lagoon of Venice. Sci. World J. **2002**, *2*, 255–260.
37. Bastianoni, S. Emergy, empower and the eco-exergy to empower ratio: A reconciliation of H.T. Odum with Prigogine? Int. J. Ecodyn. **2006**, *1* (3), 226–235.
38. Bastianoni, S.; Marchettini, N.; Pulselli, F.M.; Rosini, M. The joint use of exergy and emergy as indicators of ecosystems performances. In *Handbook of Ecological Indicators for Assessment of Ecosystem Health*; Jørgensen, S.E., Costanza, R., Xu, F.L., Eds.; CRC Press: Boca Raton, FL, 2005; 239–248.
39. Bastianoni, S.; Nielsen, S.L.; Marchettini N.; J0rgensen, S.E. Use of thermodynamic functions for expressing some relevant aspects of sustainability. Int. J. Energy Res. **2005**, *29*, 53–64.

40. Lu, H.F.; Wang, Z.H.; Campbell, D.E.; Ren, H.; Wang, J. Emergy and eco-exergy evaluation of four forest restoration modes in southeast China. Ecol. Eng. **2011**, *37*, 277–285.
41. Bastianoni, S. A definition of 'pollution' based on thermodynamic goal functions. Ecol. Modell. **1998**, *113*, *2* (1–3), 163–166.
42. Costanza, R.; d'Arge, R.; De Groot, R.; Farber, S.; Grasso, M.; Limburg, K.; Naeem, S.; O'Neill, R.V.; Paruelo, J.; Raskin, R.G.; Sutton, P.; van den Belt, M. The value of the world's ecosystem services and natural capital. Nature **1997**, *387*, 253–260.
43. De Groot, R.S.; Wilson, M.A.; Boumans, R.M.J. A typology for the classification, description and valuation of ecosystem functions, goods and services. Ecol. Econ. **2002**, *41*, 393–408.
44. Sweeney, S.; Cohen, M.J.; King, D.; Brown, M.T. Creation of a global emergy database for standardized national emergy synthesis. In *Emergy Synthesis 4: Theory and Application of Emergy Methodology*; Brown, M., Ed.; University of Florida: Gainesville, FL, 2007; 23.1–23.15.
45. Pulselli, F.M.; Coscieme, L.; Jørgensen, S.E.; Bastianoni, S. Thermodynamics-based orientors for holistic interpretation of ecosystem services. In *Acts of the XXIV Congress of the Italian Chemical Society*; University of Salento: Lecce, Italy, 11–16 September 2011; 101.

19

Ecological Indicators: Ecosystem Health

Felix Müller,
Benjamin Burkhard,
Marion Kandziora,
Claus Schimming,
and Wilhelm
Windhorst

Introduction

Environmental management has to operate in an extremely complex framework, which can be characterized by a multitude of different components and an even higher number of interrelations between these parts.[1] Ecological as well as human and societal influences have to be taken into account, and the dominant role of indirect effects has to be realized.[2] Consequently, there is a very high demand for holistic management concepts, which approach the management object from an ecosystem-based starting point. The biggest challenge of such concepts arises from the enormous complexity of human–environmental systems.[3,4] To cope with this challenge, indicators can be suitable tools. They enable quantitative statements in spite of the complex environment, but this potential is attained due to simplification, aggregation, modeling, and abstraction. Therefore, indicators often are correlated with a high uncertainty, i.e., whether the indicated object is really well represented by the indicator.[5]

Ecosystem health is an environmental management concept that directly meets these problems.[6] It follows the important demand to manage the environment from a system-based viewpoint, acknowledging inherent complexities. Therefore, the selection of suitable indicators plays a major role for quantitative applications of the health concept.[7] In the following text, some focal approaches on how these challenging demands are met are documented. The leading questions are the following:

- What are the basic features and requirements for ecosystem-based indicators, and how are they related to the concept of ecosystem health?
- What are the focal ideas of ecosystem health, and how can they be translated into ecological indication concepts?

- What are the problems and experiences of health indication in different ecosystem types?
- How do recent indicator approaches cope with the challenging demands for ecosystem-based indication?

The entry starts with a short statement on the general requirements for ecological indicators. Thereafter, the actual indicandum—ecosystem health—is sketched, and a literature survey of established ecosystem health indicators is presented. In the final sections, indicator applications and aggregations are discussed, and demands for future development are formulated based on a comparison of identified requirements and the state of the art.

Basic Features of Ecological Indicators

Environmental management should be based upon qualitative or quantitative key variables that can be used to demonstrate the demand for management actions and the outcomes of the respective activities. Such indicators provide aggregated information on certain complex phenomena,[8–10] which often are not directly accessible.[11–13] Indicators are developed on the basis of specific management purposes; they often include an integrating, synoptical value, and they should be capable of showing the differences between an existing situation and an aspired-to target state.[14,15] Thus, indicators are signals for attracting attention on changes in complex humanenvironmental systems. Heink and Kowarik[16] propose the following indicator definition: "An indicator in ecology and environmental planning is a component or a measure of environmentally relevant phenomena used to depict or evaluate environmental conditions or changes or to set environmental goals."

Being applied for the management of human–environmental systems, indicators have to satisfy very different and challenging demands: On the one hand, scientific correctness is a major requirement, and on the other, transparency and public utility in the decision making processes are significant demands (see Tables 1 and 2). Therefore, indicator applications should be based on satisfying scientific hypotheses and relevant cause–effect relations, while they also have to translate the high complexity of ecosystems in a scientifically sound way to meet the needs of politicians and decision makers for common acceptance. Furthermore, the indicator should comprise an optimal sensitivity for the related disturbance, and it should be characterized by a clear representation of the indicandum by the indicator—in this case, it should represent the challenging properties of ecosystem health.

Basic Components of Ecosystem Health

The pioneering ecologist Aldo Leopold's writings about land sickness[17] created the basic ideas for the ecosystem health concept. In the following decades, definitions of ecosystem health have been constantly evolving toward an increasing integration of human and societal contexts in order to understand what is

TABLE 1 Scientific Demands on Good Indicators

Good indicator sets should provide scientific correctness basing upon the following:

- A clear representation of the indicandum by the indicator.
- Clear proof of relevant cause–effect relations.
- An optimal sensitivity of the representation.
- Information for adequate spatio-temporal scales.
- A very high transparency of the derivation strategy.
- A high degree of validity and representativeness of the available data sources.
- A high degree of comparability in and with indicator sets.
- An optimal degree of aggregation.
- Good fulfillment of statistical requirements concerning verification, reproduction, representativity, and validity.

Source: Wiggering an Müller.[9]

TABLE 2 Applied Demands on Good Indicators

Good indicator sets should provide practical applicability basing upon the following:

- Information and estimations of the normative loadings.
- High political relevance concerning the decision process.
- High comprehensibility and public transparency.
- Direct relations to management actions.
- An orientation toward environmental targets.
- A high utility for early warning purposes.
- A satisfying measurability.
- A high degree of data availability.
- Information on long-term trends of development.

Source: Wiggering an Müller.[9]

considered to be a healthy ecosystem.[18,19] The concept has gained special popularity in the United States and in Canada, where ecosystem health has been integrated in legislation. Today, ecosystem health is part of various international political programs, like the Rio convention on sustainable development.[20] Here, it has been demanded in principle 7 that "states shall cooperate in a spirit of global partnership to conserve, protect and restore the health and integrity of the Earth's ecosystem." Ecosystem health does not only take into account ecological components but also requires a linkage with social, economic, and cultural dimensions (see Table 3).

De Kruijf and Van Vuuren[25] analyzed that "the present definitions of ecosystem health contain several of the following elements:

- Healthy ecosystems are free from ecosystem distress syndrome …;
- Healthy ecosystems are resilient … they recover from natural perturbations and disturbances;
- Healthy ecosystems are self-sustaining and can be perpetuated without subsidies or drawing down natural capital;
- Healthy ecosystems do not impair adjacent ecosystems …;
- Healthy ecosystems are free from risk factors;
- Healthy ecosystems are economically viable;
- Healthy ecosystems sustain healthy human communities."

Summarizing the different approaches, a working definition of ecosystem health can be given as follows: Ecosystem health refers to the ability of an ecosystem to maintain its structure and function over time under external stress, safeguarding a sustainable provision of ecosystem goods and services contributing to human well-being.

The facts mentioned above have to be reflected while defining appropriate sets of indicators that can be applied for environmental management.[16] Suitable ecosystem health indicator sets have to consider ecological structures as well as ecological functions on different spatial and temporal scales. As shown by the Millennium Ecosystem Assessment,[26] most ecosystems on our planet have already been

TABLE 3 Different Approaches Defining Ecosystem Health

Haskell et al.[21]: An ecological system is healthy and free from "distress syndrome" if it is stable and sustainable—that is, if it is active and maintains its organization and autonomy over time and is resilient to stress.'

Karr[22]: Ecosystem health is related to "the condition in which a system realizes its inherent potential, maintains a stable condition, preserves its capacity for self-repair when perturbed, and needs minimal external support for management."

Rapport et al.[23]: Ecosystem health refers to a "condition where the parts and functions of an ecosystem are sustained over time and where the system's capacity for self-repair is maintained, such that goals for uses, values, and services of the ecosystem are met."

Xu and Mage[24]: Ecosystem health refers to "the system's ability to realize its functions desired by society and to maintain its structure needed both by its functions and by society over a long time."

degraded under the pressure of increasing human demands. If ecosystem health shall be achieved on a long-term perspective, preventative and restorative environmental management strategies are needed. When looking at different typical ecosystem health indicators, like species diversity or water quality, many of the ecosystems on our planet can be considered unhealthy.[18] As a consequence, many ecosystem functions needed for the provision of ecosystem services have been altered. Ecosystem health refers to systems that are manipulated to satisfy human needs.[17] Therefore, ecosystem health provides a suitable conceptual framework describing the linkages between ecosystem functions, services, and human well-being.

The explicit integration of societal components makes it different from other ecosystem management concepts, for example, ecological integrity. Ecological integrity refers to the functioning of ecosystems based on self-organized processes, while ecosystem health also includes resilience and sustainability with regard to the provision of ecosystem services. Therefore, different ecological concepts like homeostasis, diversity, complexity, emergent properties, or hierarchy principles are closely related to the health concept.[19]

Basic Requirements for the Indication of Ecosystem Health

A focal problem of these concepts is the complexity of ecosystems that arises from the high number of components, relations, and interactions. Hence, for environmental practice and decision making, this complexity has to be reduced. Ecosystem theories provide an applicable basis for such a reduction. Some of the respective theoretical fundamentals of the ecosystem health concept are listed in Table 4.

These requirements are summarized in the "V-O-R model,"[21] describing the ecosystem vigor, organization, and resilience. Vigor is indicated by activity, metabolism, or primary productivity, while organization represents the diversity and number of interactions between the system components. Resilience is understood as a system's capacity to maintain structure and function in the presence of (external) stress. When resilience is exceeded, the system can shift to an alternate state. By including this approach, ecosystem health is closely related to the concepts of stress ecology, where vigor, system organization, resilience, and the absence of ecosystem distress are the main factors for a system's condition.

This model is correlated with a very high demand for comprehensive data sets and long-time series to determine resilience. Additionally, linkages of environmental and social–economic attributes and attributes representing structures as well as functions and organization have to be included. All these demands can hardly be fulfilled. Therefore, quantification deficits have to be expected as one of the main problems of the ecosystem health approach.

Utilized Indicators of Ecosystem Health

In the following paragraphs, a short literature survey of health indicators is presented, whereby different approaches have been distinguished: In the beginning, community-based indicators are listed, followed by aggregated theoretical indicators and indicators based on ecosystem analysis. Finally, a link will be developed toward the indication of ecosystem services.

TABLE 4 Axioms of Ecosystem Health

Dynamism: Nature is a set of processes, more than a composition of structures.
Relatedness: Nature is a network of interactions.
Hierarchy: Nature is built up by complex hierarchies of spatio-temporal scales.
Creativity: Nature consists of self-organizing systems.
Different fragilities: Nature includes various sets of different resiliences.

Source: Wiggering an Müller.[9]

Species- and Community-Based Indicators and Indices

Biodiversity loss is one of the characteristic signs for ecosystems under stress[18] and thus is a major issue in environmental management.[26] Consequently, there are many initiatives and concepts to describe and assess biodiversity.[27] Biodiversity indicators and indices are based on the abundance, absence, or composition of selected species or communities. They vary from single-species indicators to complex composite indicators. Suitable indicator species have to be selected in order to be representative for certain phenomena or sensitive to particular environmental changes.[28] Therefore, the appearance and dominance of certain communities can be associated with states of ecosystem health.[7]

The parameters used to quantify respective indicators can be derived from direct measurements and observations of selected species' abundance. Species- and community- based indicators can be linked to numerous international and national policy instruments, for which biodiversity indicators need to be derived (e.g., Bern Convention 1979; Bonn Convention 1979; Convention on Biological Diversity CBD 1992; the Millennium Development Goals to be achieved by 2015). Most policy and decision makers rely on indices that aggregate biodiversity data across large numbers of species,[29] but also, a limited number of key taxa are frequently used to indicate ecosystem health.[30] Some of these biodiversity indicator concepts—mainly related to the indicator collections of Marquez et al.[31] and Joergensen et al.[6]—are presented in the following.

Species Richness

The most established way to indicate biodiversity is based on species counts and composite indices. One advantage is that the number of species in a certain area is a measure that is easily established and understood by a broad range of people. The Shannon–Wiener index and the Simpson index are the most commonly applied indicators.[29] The Shannon-Wiener index H' originates in information theory and integrates species' number and evenness:

$$H' = -\sum p_i \log 2 p_i \tag{1}$$

p_i is the proportion of individuals found in species i. The values of this index can vary between 0 and 5. H' has a maximum value if the individuals of all species occur with the same density. The Simpson index refers to the number of species present and the relative abundance of each species.[32] Species richness provides important information on ecosystem conditions. However, its application in ecosystem health assessments can be misleading, for example, concerning nonnative (exotic or "invasive") species. Their abundance will increase the value of standard biodiversity indicators, but their increased dominance in biotic communities can be a typical sign of ecosystem stress.[18]

Indicator Taxa

Indicator taxa (or bioindicators, indicator species) are species or higher taxonomic groups whose properties can be used as proxies for assessments of ecosystem health.[33,34] Respective species have to be selected in order to react on ecosystem alterations by changing their abundance, density, conditions, or activities. Therefore, indicator species have to be selected objectively and must represent clear indicator-indicandum relationships. Pollinators have been suggested as useful bioindicators for ecosystem health by Kevan[35] as they are crucial for the functioning of almost all terrestrial ecosystems. Moreover, pollinators are needed for the provision of manifold ecosystem services. Further indicators based on the abundance of selected species have been suggested by Jørgensen et al.[6,36] and Burkhard et al.[7]

- Saprobic Classification: The saprobe index gives information about the degree of water pollution.[37] The different saprogenic stages are related to certain indicator organisms like bacteria, fungi, algae, amoeba, mussels, worms, insect larvae, or fishes. The stages range from polysaprobic (very highly polluted), α-mesosaprobic (highly polluted), β-mesosaprobic (medium polluted), to oligosaprobic (rather clean and clear water).

- Bellan's Pollution Index: Aquatic species like *Platynereis dumerilii*, *Theosthema oerstedi*, *Cirratulus cirratus,* and *Dodecaria concharum* are used as indicators for water pollution, whereas species like *Syllis gracillis* and *Typosyllis prolifera* indicate clear water conditions.[38] Bellan's pollution index equation is

$$\mathrm{IP} = \sum \text{dominance of pollution indicator species/clear water indicator} \tag{2}$$

  Index values higher than 1 indicate a pollution-based disturbance in the community.
- AZTI Marine Biotic Index: AZTI Marine Biotic Index (AMBI) distinguishes the soft bottom macrofauna into five groups in accordance with their sensitivity to increasing stres[39]:
  I. Species very sensitive to organic enrichment and eutrophication, present only under unpolluted conditions.
  II. Species indifferent to organic enrichment, occurring in low densities only and with no significant variations over time.
  III. Species tolerant to excess organic matter enrichment, usually supported by organic enrichment, that can also be found under normal conditions.
  IV. Second-order opportunist species.
  V. First-order opportunist species (deposit feeders).
  The coefficient is calculated as follows:

$$\mathrm{AMBI} = \left\{(0\times\%\mathrm{I})+(1.5\times\%\mathrm{II})+(3\times\%\mathrm{III})+(4.5\times\%\mathrm{IV})+(6\times\%\mathrm{V})\right\}/100 \tag{3}$$

  The AMBI values vary among the following: normal (0.0–1.2), slightly polluted (1.2–3.2), moderately polluted (3.2–5.0), highly polluted (5.0–6.0), or very highly polluted (6.0–7.0).
- BENTIX
  BENTIX is based on the AMBI but uses three groups only[40]:
  I. Species generally sensitive to disturbances.
  II. Species tolerant to stress or disturbance. Populations may respond to organic enrichment or other pollution sources.
  III. First-order opportunistic species (pioneer, colonizers, or species that are tolerant to hypoxia).
  The indicator values are calculated as follows:

$$\mathrm{Benix} = \left\{(6\times\%\ \mathrm{I})+2(\%\ \mathrm{II}+\%\ \mathrm{III})\right\}/100 \tag{4}$$

  The results represent different states of aquatic ecosystems: normal (4.5–6.0), slightly polluted (3.5–4.5), moderately polluted (2.5–3.5), highly polluted (2.0–2.5), or very highly polluted (Bentix = 0).

- Macrofauna Monitoring Index: Twelve indicator species are included in the macrofauna monitoring index. Each indicator species is assigned a score, based on the ratio of its abundance. The actual index value is the average score of species that are present in the sample.[41]
- Umbrella, Flagship, and Keystone Species: Umbrella species have high demands for their habitat conditions with regard to habitat size and quality. When protecting these species, many other species will be supported automatically. Flagship (or charismatic) species are organisms whose necessity for protection can be easily communicated. Keystone species provide an extraordinary importance for the maintenance of ecosystem structures and functions as well as for other species in the same ecosystem. Therefore, the identification and protection of keystone species can be crucial for the management of ecosystem health.[42]

Examples for the utilization of these indicator types are the Species Trend Index,[43] Red Lists,[44] or the Living Planet Index (LPI), which has been developed for land, freshwater, and marine vertebrate species. The average population trends over time are documented in the LPI. The actual calculations are based on a data set of more than 2500 species and 8000 population time series over the past 30 years. Three indices are calculated: 1) terrestrial species population index; 2) freshwater species population index, and 3) marine species population index. Each of these indices is set to a baseline of 100 in 1970, and all are given an equal weighting.[45,46]

Ratios between Different Classes of Organisms or Elements

The increase or decrease of one species in relation to others provides information about changes in ecosystems, for example, Nygard's algal index[36] or the diatoms/nondiatoms ratio.[47]

Indicators Based on Ecological Strategies

Different ecological strategies are altered by human activities or during different stages of natural development. Hence, indicators for the distinct behavior of different taxonomic groups under environmental stress situations were developed, e.g., the nematodes/copepods index, the polychaetes/ amphipods ratio, and the index of *r*/*k* strategists, which considers different taxa: Most communities in ecosystems in rather late developmental stages show dominance of *k*- selected or conservative species with large body sizes and long life spans. *R*-selected or opportunistic species have shorter life spans and are often numerically dominant. After a significant disturbance and during the following reorganization, the opportunistic species can become dominant in biomass as well as in number, whereas the conservative species are usually less favored.[36,48] Another strategy-related indicator is the trophic infaunal index, which refers to organisms' different feeding strategies (distinction of macrobenthos species into suspension feeders, interface feeders, surface deposit feeders, and subsurface deposit feeders).[49]

Additionally, there are several attempts aiming at harmonizing existing biodiversity indicator initiatives. Two examples are sketched here.

- Streamlining European 2010 Biodiversity Indicators: The Streamlining European Biodiversity Indicators (SEBI) were established in 2005 under the umbrella of the Convention on Biological Diversity (CBD). It is a process to select a set of biodiversity indicators to monitor progress toward the 2010 target of halting biodiversity loss and help achieve progress toward the target.[50,51] The SEBI is a regionally coordinated program that has been initiated in Europe as collaboration between the European Environment Agency and other European and United Nations institutions. The SEBI proposes a list of 26 indicators within the 7 CBD focal areas: status and trends of the components of biological diversity, threats to biodiversity, ecosystem integrity and ecosystem goods and services, sustainable use, status of access and benefit sharing, status of resource transfers, and public opinion.[50,51]
- Group on Earth Observations Biodiversity Observation Network: The Group on Earth Observations Biodiversity Observation Network (GEO BON) is a global partnership helping to collect, manage, analyze, and report biodiversity data.[52] It is a voluntary partnership of 73 national governments and 46 participating organizations and was launched in 2002. The GEO BON aims at providing a framework to coordinate data and observations within the Global Earth Observation System of Systems (GEOSS). Biodiversity has been named as one of nine GEOSS priority societal benefit areas. The GEO BON will integrate key ecosystem functional parameters into a Terrestrial Ecosystem Function Index (TEFI). The TEFI will integrate data of measurements of the energy, carbon, and nutrient balance.[52]

Aggregated Theory-Based Indicators and Indices

In contrast to the biotic approaches, which mainly can be used as structural indicators, the health component "vigor" is included in ecosystem theory based-indicators and indicator sets. The following four aggregations stem from thermodynamics, network, and information theories. Their basic target is a holistic aggregation of ecosystem properties into one guiding variable.

- Exergy and Exergy Indices: Exergy is that energy fraction that can be transformed into useful mechanical work. In ecological terms, it can, for example, be measured by the total biomass of the system. Eco-exergy is a refinement of the exergy concept in which biomass is weighted by the genetic complexity of the species observed.[4,36] Further holistic indicators are the exergy index and the specific exergy.[4] In ecosystems, the captured exergy is used to build up biomass and structures during successions.[48] Therefore, more complex systems also have more built-in exergy than simpler ones. Both exergy and specific exergy have been used as indicators for ecosystem health.[6] Relations between the exergy values and other ecosystem health characteristics like diversity, structure, or resilience can be found. For example, a very eutrophic ecosystem has a very high exergy due to the high biomass, but the specific exergy is low as the biomass is dominated by algae with low β values. The combination of exergy index and specific exergy provides a satisfactory structural and holistic description of ecosystem health.
- Entropy: Entropy production is one result of any metabolic activity. It can be measured by the system's respiration or the total system's export. As life is a very effective producer of entropy, this indicator has been proposed as an ecological orientor to represent maturity as well as ecosystem stress.[53,54]
- Emergy: Emergy (embodied energy) accounts for the differences between distinct biomass fractions in ecosystems basing upon the energy that has been used to build up the respective structure.[55] Conversion values, called transformities, have been derived to allow the calculation of emergy values for many ecological entities as well as socio-ecological products.[56]
- Ascendency: Ascendency is a holistic indicator that is based upon the energy flows in ecological systems and the information associated with the particular network configuration.[57,58] It represents the total system throughput and the flow diversity as a result of the food web structure. The respective network configuration is indicated by the average mutual information. Ascendency is measured by the total system throughput times the average mutual information, providing helpful information on an ecosystem's energy flow schemes and efficiencies.

Ecosystem Analytical Indicators and Indices

While most of the approaches mentioned before are aiming at one focal dimension and one value to characterize the state of an ecosystem, the following indicators have been constructed as multidimensional approaches. They try to represent ecosystem structures (biotic and abiotic structures), functions (water, matter, energy flows), and (in some cases) their relevance for human systems.

- Integrity Indicators and Orientors: Several ecosystem assessments are based on the concept of ecosystem integrity, which is closely related to ecosystem health.[18,59] The focal difference can be found in the origins of the concepts: While integrity was related to wilderness as a target function, health has been referring to ecosystems under human pressures from the beginning.[60] Meanwhile, the core conceptions have become rather similar. Therefore, one approach of integrity indication will be included in the following paragraphs.

  Taking into account the focal ideas of the sustainable development concept, "meet the needs of future generations "means "keep available ecosystem services on a long-term, intergenerational and broad scale, intragenerational level." From a synoptic viewpoint, all ecosystem services are strongly dependent on the performance of the system's regulation capacity (see the section on "Ecosystem Service Indicators"). Taking into account that the integral of regulating ecosystem services represents self-organized processes,[61] it becomes clear that the respective benefits are dependent on the degrees and potentials of self-organization. To maintain these services, the ability for future self-organizing processes has to be preserved.[53,62] Under this viewpoint, Barkmann et al.[63] have defined ecological integrity as a "management target for the preservation against

> nonspecific ecological risks, that are general disturbances of the self-organising capacity of ecological systems. Thus, the goal should be a support and preservation of those processes and structures which are essential prerequisites of the ecological ability for self-organisation".

In ecosystem theory, many different approaches (see Joergensen[4]) are highly compatible with the theory of self-organization. The consequences have been condensed within the orientor approach,[64–66] a system-based theory about ecosystem development, which is founded on the ideas of nonequilibrium thermodynamics,[54,62] and network development on the one hand and succession theory on the other.[67,68] The basic idea is that throughout the undisturbed complexifying development of ecosystems, certain characteristics are increasing steadily, developing toward an attractor state, which is restricted by the specific site conditions. For instance, the food web will become more and more complex; heterogeneity, species richness, and connectedness will be rising; and many other attributes will follow a similar long-term trajectory.

Many of these orientors cannot be easily measured or modeled under usual circumstances. Therefore, the selected indicators have to be represented by variables that are accessible by "traditional" methods of ecosystem quantification. Furthermore, it has to be reflected that the number of indicators should be reduced as far as possible, providing a small set consisting of the most important items that can be calculated or measured in many local instances. The focal subsystems that should be taken into account to represent ecosystem organization are ecosystem structures with the biotic and abiotic diversity and functions, represented by the energy, water, and matter balances (for a detailed justification, see Müller[61]).

On the basis of these features, a general indicator set to describe the ecosystem or landscape state in terrestrial environments has been derived. It is shown in Table 5. The basic hypothesis concerning this set is that a holistic representation of the degree of and the capacity for complexifying ecological processes on the basis of an accessible number of indicators can be fulfilled by these variables. They also represent the basic trends of ecosystem development; thus, they show the developmental stage of an ecosystem or a landscape. As a whole, this variable set represents the degree of self-organization in the investigated system. For quantifications, see Müller[69] or Müller and Burkhard.[70]

- The holistic ecosystem health indicator: An expansion toward an integration of human items is provided by the holistic ecosystem heath indicator (HEHI) system. It was developed in 1999 in Costa Rica as an integrative indicator that might be an appropriate tool for assessing and

TABLE 5 Proposed Indicators to Represent the Organizational State of Ecosystems and Landscapes

Orientor Group	Indicator	Potential Key Variable
Biotic structure	Biodiversity	Number of species
Abiotic structures	Biotope heterogeneity	Index of heterogeneity
Energy balance	Exergy capture	Gross or net primary production
	Entropy production	Entropy production[71] Entropy production[72] Output by evapotranspiration and respiration
	Metabolic efficiency	Respiration per biomass
Water balance	Biotic water flows	Transpiration per evapotranspiration
Matter balance	Nutrient loss	Nitrate leaching
	Storage capacity	Intrabiotic nitrogen
	Soil organic carbon	Soil organic matter

Source: Müller.[69]

TABLE 6 Elements of the Holistic Ecosystem Health Indicator Set

Ecological Elements	Social Elements	Interactive Elements
Soil quality	Income	Land use and distribution
Riparian zone	Access to Services	Watershed protection
Water quality	Job stability	Land degradation
Biomass	Gender roles	Citizen involvement
Land use	Demographics	Implementation of legislation
Primary production	Community	Environmental awareness
Regeneration	Strength	
Biodiversity		
Erosion		

Source: Aguilar.[73]

evaluating health of managed ecosystems.[73] The HEHI follows a hierarchical structure starting with three main branches: ecological, social, and interactive. The interactive branch includes measures relating to land use and management decisions that characterize the interactions between the human communities and the ecosystem. Furthermore, each branch is subdivided into categories or criteria (see Table 6). Each category is given a target or a benchmark, which is based on references available in scientific literature, policies, etc.

Ecosystem Service Indicators

As the explicit target figure of ecosystem health indicators has more and more been moved toward human wellbeing, the respective indicanda—ecosystem goods and services—are mentioned here. Their implementation seems to be very significant as a criterion of success in ecosystem health management. Ecosystem services are the benefits people obtain from natural structures and functions. Since ecosystems are dynamic and complex units, the assessment of their services is strongly facilitated by the categorization into functional groups, which are exemplarily listed in Table 7, referring to the following:

- Provisioning services: products obtained from ecosystems, e.g., food, water.
- Regulating services: benefits from regulating ecosystem processes, e.g., flood regulation, disease regulation.
- Cultural services: nonmaterial benefits, e.g., recreation, spiritual benefits, information.

Ecosystem structures and functions can be indicated by ecological integrity, as described above. Regulating ecosystem services is strongly related to ecosystem functions, and some regulating services are even overlapping with ecological integrity processes (e.g., processes related to nutrient or water regulation.[74] Thus, clear definitions of ecological integrity variables and regulating ecosystem service indicators are mandatory. Most ecosystem functions are difficult to quantify under natural conditions, but the application of ecological models can help. Perhaps the best data are available for provisioning ecosystem services. Normally, production and trade quantities and their market prices are used. Cultural ecosystem services again are rather difficult to quantify due to each individual's subjective and situation-dependent appreciation of related values.[26]

Ecosystem services are not a linear chain from means to ends because ecosystems as well as societal systems are complex, dynamic, and adaptive.[3] There exist multiscale relationships between services and benefits. When ecosystems are stressed and degraded, their service provision is affected too, which in turn has impacts on human activities and health.[75] The fact that there is a high correlation between decline in ecosystem health and service provision leads to the suggestion that ecosystem services are good integrative and aggregate measures, showing the consequences of the respective ecosystem health conditions.[76]

TABLE 7 List of Ecosystem Services

	Regulating Ecosystem Services
Local climate regulation	Effects on temperature, wind, radiation, precipitation
Global climate regulation	For example, carbon sequestration, greenhouse gas emission
Flood protection	Extreme flood event dampening
Groundwater recharge	Runoff, flooding, aquifer recharge
Air quality regulation	Removal of toxic and other elements from the atmosphere
Erosion regulation	Soil retention and prevention of landslides
Nutrient regulation	(Re)cycling of, e.g., N, P, or other elements
Water purification	Removal of impurities from fresh water
Pollination	For example, by wind and bees
	Provisioning Ecosystem Services
Crops	Edible plants
Livestock	Edible animals
Fodder	Animal fodder
Capture fisheries	Fish accessible for fishermen
Aquaculture	Terrestrial or marine aquaculture
Wild foods	For example, berries, mushrooms, hunting, fishing
Timber	Trees or plants for construction
Wood fuel	Trees or plants for heating, cooking
Energy (biomass)	Trees or plants for energy generation
Biochemicals and medicine	Production of biochemicals, medicines
Freshwater	For example. for drinking, irrigation
	Cultural Ecosystem Services (selection)
Recreation and aesthetic values	Landscape and visual qualities
Intrinsic value of biodiversity	Value of nature and species themselves

Source: Belcher and Boehm,[87] Schönthaler et al.[111] and Reuter et al.[112]

Health Indicators in Different Ecosystem Types

To illuminate the wide field of health indicators, in the following sections, some utilizations of the concept in different ecosystem types are presented.

Agroecosystem Health

An agroecosystem can be defined as "a socio-ecological system, managed primarily for the purpose of producing food, fiber and other agricultural products, comprising domesticated plants and animals, biotic and abiotic elements of the underlying soils, drainage networks, and natural vegetation and wildlife."[77–79] The health status of agro ecosystems has been described by Rapport,[80] and Rapport et al.[81] In this entry, a valuable overview about various approaches to indicate agroecosystem health on various spatial scales is provided. For example, Zhang et al.[82] merge geographical information systems (GIS)-based land use analysis data with the use of pesticides and their pathways through the environment. This approach allows the assessment of ecosystem health as a ratio between the amounts of pesticides applied in one grid to the maximum dose applied in the study area. Kaffka et al.[83] suggest that the capacity to retain nutrients like N and P might be useful to evaluate the ecosystem health of a catchment area, while Mitchell et al.[84] consider the content of soil organic matter (SOM) of agricultural areas as a focal indicator to assess ecosystem health in the foreground.

Hopkins[85] indicates ecosystem health via the number and size of wildlife patches in an area dominated by intensive agriculture.

In comparison to the aforementioned authors, Altieri and Nicholls[86] base their suggestions to achieve healthy agroecosystems on the avoidance capacity of diseases—indicated by optimal recycling of nutrients, closed energy flows, water and soil conservation, and biological pest regulation. Belcher and Boehm[87] use yield, soil N and P, soil water, SOM, soil erosion, and CO_2 emissions as major attributes in their sustainable agroecosystem model. However, the assessments of agroecosystem health regularly focus on resources that have to be classified as internal to the system.[88] Hence, the options to assess the health of an agroecosystem in relation to its ability to adapt to variations in its changing socio-economic and ecological context are rarely realized according to Waltner-Toews[79] or Ikerd.[89]

Forest Ecosystem Health

In the United States, 20 years ago, a sound definition of forest health had already been derived to sustain healthy conditions of ecosystem development and productivity in a long-term perspective.[90] The term "forest health" increasingly found evidence in mandates concerning environmental management and protection, mostly supported by the idea that the conventional measures for describing forest states (e.g., crown conditions, tree growth, loss of nutrients, soil potentials, biodiversity) can also be used to indicate forest ecosystem health. Regarding forests, only recently in Europe, Ecosystem Health has been advertised in the context of deposition of air pollutants[91] and the consequent change of chemical states, particularly the degree of eutrophication and acidification which relate to the holistic aspects of productivity and biogeochemical cycling. Information on nutrient recycling, imbalance with the inputs and outputs and on energy use are the essentials of the Ecosystem Assessment Health concept and the adequate indicators proposed by Jørgensen.[6] This kind of concept, basically productivity related and holistic, does greatly comply with ecosystem theories and opens on for indicating Forest Ecosystem Health combing both, utilitarian and ecosystem perspectives[92] respecting the conditions under which selforganization of forest ecosystems can take place.[50]

Whereas atmospheric deposition of acidifying air pollutants and eutrophic nitrogen is identified as one of the major environmental problems, an ecosystem process-orientated indicator has been demanded for quantifying marginal loads for damaging structures and for interference with ecosystem functions.[93–95] Since the respective concept of critical loads is a stoichiometric approach and a function of the load quantity on chemical effects on ecosystems, intensity criteria are needed to provide adequate threshold values.[96–98] The conduction of the concept depends on combined balances of mass and charge provided by mineral elements, nitrogen, and free acidity completing the ion composition of internal transfers and the matter–flux relationship with the abiotic environment. Imbalance between input and recycling respecting mass and ionic charge indicates the efficiency of nutrient use while the quantities can be related with the intensities of effective concentrations and free acidity. In this regard soil chemistry is a function of the extensity of production on the intensities of nutrient availability, free acidity and effectiveness of toxic concentrations under influence of ecosystem self-organization.

The performance of cycling and imbalance between input rates and the degree of recycling provide useful information on the Ecosystem Health aspect. Moreover, as imbalance of mineral element cycling and related losses from the ecosystem are irreversible processes for terrestrial ecosystems, decreasing alkalinity in combination with increasing free acidity are directing to maturity of ecosystems. Regarding abiotic structures, maturity emerges by development of soil structures, which is related to loss of potentials. Based on these principles, Ulrich[99] suggested the indication of stages of forest ecosystems, for instance, by the structural properties of soil constituents providing acid neutralization and buffer capacity.

Aquatic Ecosystem Health

Aquatic ecosystems (wetlands, marine and estuarine zones, lakes, groundwater, lagoons, and rivers) consist of complex structures and fulfill important functions for the provision of numerous ecosystem services.[26] Some of these aquatic ecosystems are heavily endangered, e.g., wetlands were turned into agricultural land with dramatic consequences due to the loss of the buffer capacity for pesticides, nutrients, and floods as well as the loss of habitat functions. Marine and estuarine ecosystems are heavily influenced by humans due to population growth and the associated consequences such as pollution, growing demands for resources, eutrophication, overfishing, and habitat modification (e.g., mangrove clearing). Estuarine and marine ecosystems are interdependent as estuarine areas are the nursery grounds for many species, which are then provided as successful functioning marine commercial stocks. Probably the most important function of lakes is the freshwater storage, which provides ecosystem services for society and economy, but these systems also are heavily endangered.

Aquatic ecosystems are focal areas of ecosystem health assessments. Therefore, many of the previously mentioned indicator concepts mainly provide information on aquatic ecosystem health (see Utilized indicators of ecosystem health or Joergensen et al.[6]). Besides these long indicator listings, Boesch and Paul[100] highlight the following traditional indicators: contaminant levels, material input (e.g., nutrients, sediments), water quality (e.g., dissolved oxygen), fish catch, extent of certain habitats (e.g., wetlands), community structure, toxicity biomarkers, and indicators of human pathogens).

Urban Ecosystem Health

A very special aspect is provided by the health concept in urban systems, i.e., because the human factor plays a dominant role in the relevant literature.[101–105] For instance, Su et al.[101] state that “an urban ecosystem consists of residents and their environment in certain time and space scales, in which, ecologically-speaking, consumers are the dominant component lacking producers and decomposers.”

Therefore, urban ecosystem health must be assessed by very comprehensive, integrative indicator sets, which also include variables of human health. Consequently, Hancock[104] has determined six basic elements for healthy cities:

1. Population health and distribution.
2. Societal well-being.
3. Government, management, and social equity.
4. Human habitat quality and convenience.
5. Natural environment quality.
6. Impact of the urban ecosystem on the larger-scale natural ecosystem.

A similar approach can be found in the indicator sets of Su et al.,[101] who generally distinguish human and environmental subsystems (see Table 8).

TABLE 8 Some Indicanda for Urban Health

Human Subsystem	Environmental Subsystem
Public health	Provisioning services
Health expenditure	Environmental quality
Nutrition	Atmospheric quality
Budget and finance situation	Water quality
Urban infrastructure	Forest coverage
Human housing conditions	Farmland area
Education	Emergy density
Employment	Carrying capacity

Source: Su et al.[101]

Indicator Application and Aggregation in Management

Management toward ecosystem health is directed to improve human well-being. However, in order to environmentally manage, alternatives to achieve higher levels of human well-being and ecosystem health or to stabilize the present level in a changing world have to be identified. Hence, indicators have to facilitate the comparability of states in space and time, in order to allow informed decisions based on assessments of the present state and the state achievable by management options in respect to the power and competences in the hands of the manager. Concerning ecosystem health, at least, decision makers acting on the following levels have to be equipped with indicators:

1. Site management, e.g., field, forest, lake
2. Unit management, e.g., farm, forest district, catchment, nature sanctuary
3. Public management, e.g., community, county, nation
4. Public, e.g., citizens, interest groups (nongovernmental organizations)

All levels are interrelated. For example, 1) different kinds of pest management on the site level change not only the local ecosystem but also the food quality and availability. 2) The fodder quality available on the farm level depends on the productivity of the fields and impacts the economic efficiency of the farm. 3) The socio-economic state on the community level depends, on the one hand, on the economic viability of the hosted unit (item 2), but the provision of public services likewise constrains the range of activities of the units. 4) Adaptation to changing global constraints of public (item 3), unit (item 2), and site (item 1) management largely depends on the public awareness and level of satisfaction, e.g., human well-being realized at present and achievable in the future.

To account for all levels requires a nested approach in indication of ecosystem health plus mirroring the mutual synergistic and antagonistic interactions between the different levels. The need for such an approach has been articulated by Walter-Toewe and Wall[88] and Rapport and Singh.[106] However, a broadly accepted scheme meeting these requirements is not in place yet.[107] The present state of the art is largely influenced by the concept of ecosystem services and human well-being presented by the Millennium Assessment[26] and the TEEB Study.[108] With respect to "management," the further development of indicators should be constrained by the range of management options in space and time available to decision makers on the indicated levels (1–4) and be limited to parameters relatable to indicators of human well-being.

The need for indicator aggregation evolves out of the definition of ecosystem health. According to Waltner-Toews and Wall,[88] a nested approach is required in this context also. Another constraining aspect to be considered is societal interests, which are embedded in the definition as well. In addition, in case that the state of ecosystem health is considered to be poor, new or changed management activities have to be initiated and monitored by indicators. Thus, the goal of indicator aggregation is to transfer the information about the ecosystem state to those spatiotemporal scales on which management is possible and performed. The respective levels range from site management via communities, counties, states, nations up to the international institutions. Hence, a satisfying overlap between the spatial extent of the ecosystems at stake and the respective management unit is required. Systems theory and hierarchy theory provide a theoretical background to facilitate indicator aggregation for such nested systems.

To practically deal with this means to accept generalizations, including losses of information, as never can all interactions causing emergent properties relevant for human well-being be known. Hence, bottom-up aggregation is normally severely restricted. Furthermore, the focus on ecosystems enforces the integration of components, which are measured with parameters that cannot be aggregated on the base of concise units, e.g., the mortality rate of a specific species in forests of a watershed is already challenging to determine but analytically impossible to fuse with the number of pathogens endangering the fish population in a lake of the same catchment. A feasible approach to deal with this challenge is to work with relative indicator values, i.e., to indicate relative changes of the selected variables. A suitable option to facilitate spatiotemporal comparisons is to define a particular ecosystem state as reference state and to study the relative alterations from this reference state. Examples harnessing this approach have been presented by Windhorst et al.[109,110] A suitable method to facilitate aggregation and to deal with spatiotemporal interactions is to use simulation models (Belcher and Boehm,[87] Schönthaler et al.[111] and Reuter et al.[112]).

In any case, all ecosystem health indicators and respective aggregation approaches should be connectable to the different components of human well-being. The Millennium Assessment's[26] categories were "security, basic material for good life, health, good social relations and freedom of choice and action." They can be indicated for the management units at stake. Hence, multiple interactions to be considered take place and further are conceivable for future situations, creating a fuzzy environment, obstructing the development of generally applicable aggregation procedures. However, progress in identifying suitable aggregation procedures can be achieved by answering the following questions for each management unit at stake:

- Can the aggregated indicator of ecosystem health be addressed and modified by at least one management action?
- Are changed values of the aggregated indicator indicative for different ecosystem states of the management unit at stake?
- Do changed values of the aggregated indicator indicate betterment of at least one category of human well-being?

Conclusions

The previous paragraphs have shown that there is an enormous variety of indicators proposed to represent ecosystem health. Many of these indicators have been used to assess ecosystem health in different ecosystem types, and in many of those cases, the indicators have been useful tools for environmental management. On the other hand, these applications illuminate some general problems of health indicators that should be solved in the future:

- The health approach is very challenging, i.e., due to its comprehensive character. Therefore, health status can hardly be represented by one variable alone. Instead, comprehensive indicator sets are necessary to include the basic elements of vigor, organization, and resilience.
- Indicators or sets selected have often failed in fulfilling the comprehensive criteria of ecosystem health assessments because they are mostly specific for one particular environmental problem to be solved[6] and do not adequately represent ecosystem complexity. Thus, there are difficulties in satisfying the original ecosystem health idea advertised.
- Consequently, the meaningfulness of solely structural indicators is limited, as they do not reflect processes of ecosystem functions. Therefore, many of the listed indicators should be understood as elements of indicator sets, not as single indicators of ecosystem health.
- As the health concept has been outlined in strict contact with human systems, the linkage between man and environment should be included, at least in indicator selection. That linkage up to now can hardly be found in the literature.
- Due to the metaphoric character of the "health" approach, it has been very successful in some areas, while in other nations, it has not been applicable due to critical viewpoints on the concept's theoretical or even philosophical character.
- Furthermore, the indication of ecosystem health also covers the health status of the human population, as well as socio-economic and cultural dimensions in relation to the vigor, organization, and resilience of ecosystems.[18] Hence, parameters to indicate security, basic material for good life, health, and good social relations are indispensable to analyze overall ecosystem health, which is in line with the assessment of Rapport et al.[75] An integrative approach to bundle a suite of indicators and to attach meaningful values is the concept of ecosystem distress syndromes,[113]

which can be seen as a forerunner of the environmental degradation syndromes elaborated by the German Advisory Council on Global Change (WBGU). The three major syndromes have been named by WBGU[105]: 1) utilization, which includes the overexploitation of marginal land; 2) development which includes the destruction of ecosystems as a result of large-scale projects; and 3) sink, comprising environmental degradation resulting from large-scale diffusion of long-lived substances.

Coming back to our initial questions, we can summarize the following:

- The basic features and requirements of the ecosystem health approach demand for comprehensive and integrative indicator sets, which is a big scientific and practical challenge.
- Therefore, several proposals exist, and the health concept is used as a reference in several cases, but very often, the interdisciplinary demands of the approach are not fulfilled.
- Applications can be found mainly in aquatic ecosystems, mostly being quite distant from the involvement of human factors.
- Good chances for future development can be seen by enhancing the integration with integrity and ecosystem services.

References

1. Wilson, G.A.; Bryant, R.L. *Environmental Management: New Directions for the 21st Century;* Taylor and Francis Group: London, 2002.
2. Fath, B.D.; Patten, B.C. Review of the foundations of network environ analysis. Ecosystems **1999**, *2*, 167–179.
3. Allen, C.R.; Holling, C.S. *Discontinuities in Ecosystems and Other Complex Systems*; Columbia University Press, New York: 2008.
4. Joergensen, S.E.; Fath, B.; Bastianoni, S.; Marquez, J.; Müller, F.; Nielsen, S.N.;Patten, B.; Tiezzi, E.; Ulanowicz, R. *A New Ecology—The Systems Perspective;* Elsevier Publishers: Amsterdam, 2007.
5. Bossel, H. Indicators for sustainable development: Theory, method, applications. A report to the Balaton Group. International Institute for Sustainable Development: Winnipeg, 1999.
6. Jørgensen, S.E.; Xu, F.L.; Costanza, R., Eds. *Handbook of Ecological Indicators for Assessment of Ecosystem Health*, 2nd Ed. (Applied Ecology and Environmental Management); CRC Press: London, 2010.
7. Burkhard, B.; Müller, F.; Lill, A. Ecosystem health indicators: Overview. In *Ecological Indicators. Vol. [2] of Encyclopedia of Ecology,* 5 vols; Jørgensen, S.E., Fath, B.D., Eds.; Elsevier: Oxford, 2008; 1132–1138.
8. Dale, V.H.; Beyeler, S.C. Challenges in the development and use of ecological indicators. Ecol. Indic. **2001**, *1*, 3–10.
9. Wiggering, H.; Müller, F., Eds. *Umweltziele und Indikatoren;* Springer-Verlag: Berlin, Heidelberg, New York, 2003.
10. Müller, F.; Lenz, R. Ecological indicators: Theoretical fundamentals of consistent applications in environmental management. Ecol. Indic. **2006**, *6,* 1–5.
11. Walz, R. Development of environmental indicator systems: Experiences from Germany. Environ. Manage. **2000**, *26*, 613–623.
12. Turnhout, E.; Hisschemoller, M.; Eijsackers, H. Ecological indicators: Between the two fires of science and policy. Ecol. Indic. **2007**, *7,* 215–228.
13. Niemeijer, D.; De Groot, R.S. A conceptual framework for selecting environmental indicators sets. Ecol. Indic. **2008**, *8,* 14–25.
14. Girardin, P.; Bockstaller, C.; van der Werf, H.M.G. Indicators: Tools to evaluate the environmental impacts of farming systems. J. Sustainable Agric. **1999**, *13,* 5–21.
15. EEA. Environmental Indicators: Typology and Use in Reporting; EEA: Copenhagen, 2003; 20 pp.

16. Heink, U.; Kowarik, I. What are indicators? On the definition of indicators in ecology and environmental planning. Ecol. Indic. **2010**, *10,* 584–593.
17. Leopold, A. Wilderness as a land laboratory. Living Wilderness **1941**, *6*, 3.
18. Rapport D.J. Regaining healthy ecosystems: The supreme challenge of our age. In *Managing for Healthy Ecosystems*; Rapport, D.J., Ed.; Lewis Publisher: Boca Raton, 2003; 5–10.
19. Costanza, R.; Norton, B.G.; Haskell, B.D., Eds. *Ecosystem Health: New Goals for Environmental Management*; Island Press: Washington, DC, 1992; 279 pp.
20. UNCED (United Nations Conference for Environment and Development). *Agenda 21*; United Nations: New York, 1992.
21. Haskell B.D.; Norton, B.G.; Costanza, R. Introduction: What is ecosystem health and why should we worry about it? In *Ecosystem Health: New Goals for Environmental Management*; Costanza, R., Norton, B.G., Haskell B.D., Eds.; Island Press: Washington, DC, 1992; 3–20.
22. Karr, J.D. Measuring biological integrity: Lessons from streams. In *Ecological Integrity and the Management of Ecosystems*; Woodley, S., Kay, J., Francis, G., Eds.; CRC Press: Boca Raton, FL, 1993; 83–104.
23. Rapport, J.D.; Rolston, E.D.; Qualset, O.C.; Damania, B.A.; Lasley, L.W. *Managing for Healthy Ecosystems*; CRC Press: Boca Raton, London, New York, Washington, DC, 2002.
24. Xu, W.; Mage, J.A. A review of concepts and criteria for assessing agroecosystem health including a preliminary case study from southern Ontario. Agric., Ecosyst. Environ. **2001**, *83*, 215–233.
25. De Kruijf, H.A.M.; Van Vuuren, D.P. Following sustainable development in relation to the North–South dialogue: Ecosystem health and sustainability indicators. Ecotoxicol. Environ. Safety **1989**, *40*, 414.
26. MA (Millennium Ecosystem Assessment). *Ecosystems and Human Wellbeing*: Synthesis. Island Press, World Resources Institute: Washington, DC, 2005.
27. Scholes, R.J.; Mace, G.M.; Turner, W.; Geller, G.N.; Jürgens, N.; Larigauderie, A.; Muchoney, D.; Walther, B.A.; Mooney, H.A. Toward a global biodiversity observing system. Science **2008**, *321*, 1044–1045.
28. van Strien, A.J.; van Duuren, L.; Foppen, R.P.B.; Soldaat, L.L. A typology of indicators of biodiversity change as a tool to make better indicators. Ecol. Indic. **2009**, *9*, 1041–1048.
29. Lamb, E.G.; Bayne, E.; Holloway, G.; Schieck, J.; Boutin, S.; Herbers, J.; Haughland, D.L. Indices for monitoring biodiversity change: Are some more effective than others? Ecol. Indic. **2009**, *9*, 432–444.
30. Rossi, J.P. Extrapolation and biodiversity indicators: Handle with caution! Ecol. Indic. **2010**, doi:10.1016/ j.ecolind.2010.09.002.
31. Marquez, J.C.; Salas, F.; Patricio, J.; Teixera, H.; Neto, J.M. *Ecological Indicators for Coastal and Estuarine Environmental Assessment: A User Guide*; WIT Press: Southampton, 2009.
32. Simpson, E.H. Measurement of diversity. Nature **1949**, *163*, 688.
33. Hilty, J.; Merenlender, A. Faunal indicator taxa selection for monitoring ecosystem health. Biol. Conserv. **2000**, *92*, 185–197.
34. Pinto, R.; Patricio, J.; Baeta, A.; Fath, B.; Neto, J.M.; Marques, J.C. Review and evaluation of estuarine biotic indices to assess benthic condition. Ecol. Indic. **2009**, *9*, 1–25
35. Kevan, P.C. Pollinators as bioindicators of the state of the environment: Species, activity and diversity. Agric., Eco-syst. Environ. **1999**, *74*, 373–393.
36. Jørgensen, S.E. Introduction. In *Handbook of Ecological Indicators for Assessment of Ecosystem Health*, 2nd Ed. (Applied Ecology and Environmental Management); Jørgensen, S.E., Xu, F.L., Costanza, R., Eds.: CRC Press: London, 2010.Jørgensen, S.E.; Xu, F.L.; Salas, F.; Marques, J.C. *Application of Indicators for the Assessment of Ecosystem Health. Handbook of Ecological Indicators for the Assessment of Ecosystem Health*; CRC Press, Boca, Raton: 2005.
37. Kolkwitz, R.; Marsson, M. Grundsätze für die biologische Beurteilung des Wassers nach seiner Flora und Fauna. Mitt. aus d. kgl. Prüfungsanstalt für Wasserversorgung und Abwässerbeseitigung **1902**, *1*, 33–72.

38. Bellan, G. Pollution Indices. In *Encyclopedia of Ecology*, 5 vols; Jørgensen, S.E., Fath, B.D., Eds.; Elsevier: Oxford, 2008; 2861–2868.
39. Borja, A.; Franco, J.; Perez, V. A marine biotic index to establish the ecological quality of soft-bottom benthos within European estuarine and coastal environments. Mar. Pollut. Bull. **2000**, *40* (12), 1100–1114.
40. Simboura, N.; Zenetos, A. Benthic indicators to use ecological quality classification of Mediterranean soft bottom marine ecosystems, including a new biotic index. Mediterr. Mar. Sci. **2002**, *3* (2), 77–111.
41. Roberts, R.D.; Gregory, M.G.; Foster, B.A. Developing an efficient macrofauna monitoring index from an impact study—A dredge spoil example. Mar. Pollut. Bull. **1998**, *36* (3), 231–235.
42. Simberloff, D. Flagships, umbrellas, and keystones: Is single-species management passé in the landscape era? –Biol. Conserv. **1998**, *83* (3), 247–257.
43. Cocciufa, C.; Petriccione, B.; Framstad, E.; Bredemeier, M. Biodiversity Assessment in LTER sites. An EC Report (Deliverable 3.R2.D1) from ALTER Net (A Long Term Biodiversity, Ecosystem and Awareness Research Network); 2007.
44. Brito, D.; Ambal, R.G.; Brooks, T.; De Silva, N.; Foster, M.; Hao, W.; Hilton-Taylor, C.; Paglia, A.; Rodríguez, J.P.; Rodríguez, J.V. How similar are national red lists and the IUCN Red List? Biol. Conserv. **2010**, *143*, 1154–1158.
45. UNEP-WWF. *The Living Planet Report*; 2004.
46. Loh, J. *Living Planet Report 2000*. Gland, World Wide Fund for Nature; 2000.
47. Lenhart, H.J. Effects of river nutrient load reduction on the eutrophication of the North Sea, simulated with the ecosystem model ERSEM. Senckenbergiana maritima **2001**, *31* (2), 299–311.
48. Burkhard, B.; Fath, B.D.; Müller, F. Adapting the adaptive cycle: Hypotheses on the development of ecosystem properties and services. Ecol. Modell. **2011**, doi:10.1016/ j.ecolmodel.2011.05.016.
49. Word, J.Q. Classification of benthic invertebrates into infaunal trophic index feeding groups. Coastal Water Research Project Report 1979–1980; Los Angeles, 1979; 103–121.
50. EEA. Halting the Loss of Biodiversity by 2010: Proposal for a First Set of Indicators to Monitor Progress in Europe, EEA Technical Report No 11/2007; 2007.
51. EEA. *Progress towards the European 2010 Biodiversity Target*, EEA Technical Report No 5/2009; 2009.
52. GEO BON (Group on Earth Observations Biodiversity Observation Network). Detailed Implementation Plan. GEO BON. Version 1.0, May 22, 2010.
53. Kay, J.J. On the nature of ecological integrity: Some closing comment. In *Ecological Integrity and the Management of Ecosystems*; Woodley, S., Kay, J., Francis, G., Eds.; St. Lucie Press: Delray, FL, 1993; 210–212.
54. Schneider, E.D.; Kay, J.J. Life as a manifestation of the second law of thermodynamics. Math. Comput. Modell. **1994**, *19*, 25–48.
55. Odum, H.T. Environmental Accounting: Emergy and Environmental Policy Making; John Wiley and Sons: New York, 1996; 370 pp.
56. Brown, M.T.; Ulgiati, S. Updated evaluation of exergy and emergy driving the geobiosphere: A review and refinement of the emergy baseline. Ecol. Modell. **2010**, *221*, 2501–2508.
57. Ulanowicz, R.E. *Growth and Development: Ecosystems Phenomenology*; Springer-Verlag: NY, 1986; 203 pp.
58. Ulanowicz, R.E. *Ecology the Ascendent Perspective*; Columbia University Press: NY, 1997; 201 pp.
59. Callicott, J.B. A review of some problems with the concept of ecosystem health. Ecosyst. Health **1995**, *1* (2), 101–112.
60. Westra, L. The ethics of ecological integrity and ecosystem health: The interface. In *Managing for Healthy Ecosystems*; Rapport, D.J., Ed.; Boca Raton, London, New York, Washington, DC, 2003; 31–40.

61. Müller, F. Ecosystem indicators for the integrated management of landscape health and integrity. In *Ecological Indicators for Assessment of Ecosystem Health*; Joergensen, S.E., Costanza, R., Xu, F.L., Eds.; Taylor and Francis: Boca Raton, FL, 2004; 277–304.
62. Kay, J.J. Ecosystems as self-organised holarchic open systems: Narratives and the second law of thermodynamics. In *Handbook of Ecosystem Theories and Management*; Joergensen, S.E., Müller, F., Eds.; Lewis Publishers: Boca Raton, 2000; 135–160.
63. Barkmann, J.; Baumann, R.; Meyer, U.; Müller, F.; Windhorst, W. Ökologische Integrität: Risikovorsorge im Nachhaltigen Landschaftsmanagement. Gaia **2001**, *10* (2), 97–108.
64. Bossel, H. Sustainability: Application of systems theoretical aspects to societal development. In *Handbook of Ecosystem Theories and Management*; Jørgensen, S.E., Müller, F., Eds.; CRC Press, Boca Raton: 2000; 519–536.
65. Müller, F.; Leupelt, M. *Eco Targets, Goal Functions, and Orientators*; Springer: Berlin, 1998; 619 pp.
66. Müller, F.; Jørgensen, S.E. Ecological orientors: A path to environmental applications of ecosystem theories. In *Handbook of Ecosystem Theories and Management*; Jørgensen, S.E., Müller, F., Eds.; CRC Publishers: New York, 2000; 561–576.
67. Odum, E.P. The strategy of ecosystem development. Science **1969**, *164*, 262–270.
68. Odum, E.P. *Fundamentals of Ecology*; WB Saunders WB: Philadelphia, 1971.
69. Müller, F. Indicating ecosystem and landscape organization. Ecol. Indic. **2005**, J (4), 280–294
70. Müller, F.; Burkhard. B. Ecosystem indicators for the integrated management of landscape health and integrity. In *Handbook of Ecological Indicators for Assessment of Ecosystem Health*, 2nd Ed.; Jorgensen, S.E., Xu, L., Costanza, R., Eds.; Taylor and Francis, 2010; 391–423.
71. Aoki, I. Entropy and exergy in the development of living systems: A case study of lake ecosystems. J. Phys. Soc. Jpn. **1998**, *67*, 2132–2139
72. Svirezhev, Y.M.; Steinborn, W. Exergy of solar radiation: Thermodynamic approach. Ecol. Modell. **2001**, *145*, 101–110.
73. Aguilar, B.J. Applications of ecosystem health for the sustainability of managed systems in Cost Rica. Ecosyst. Health **1999**, *5*, 1–13.
74. Burkhard, B.; Kroll, F.; Nedkov, S.; Müller, F. Mapping supply, demand and budgets of ecosystem services. Ecol. Indic. **2011**, doi:10.1016/j.ecolind.2011.06.019.
75. Rapport, D.J.; Costanza, R.; McMichael, A.J. Assessing ecosystem health. TREE **1998**, *13* (10), 397–402.
76. Rapport, D.J. Ecosystem services and management options as blanket indicators of ecosystem health. J. Aquat. Ecosyst. Health **1995**, *4*, 97–105.
77. Gallopin, G. The potential of agroecosystem health as a guiding concept for agricultural research. Ecosyst. Health *1*, 129–140.
78. Waltner-Toews, D. Ecosystem health: A Framework for implementing sustainability in agriculture. BioScience *46* (9), 686–689.
79. Waltner-Toews, D. Agro-ecosystem health: Concept and principles. In *Agro-Ecosystem Health*, Proceedings of a seminar held in Wageningen, Sep 26, 1996; van Bruchem, J., Ed.; NRLO-rapport nr. 97/31; 1997 9–22.
80. Rapport, D.J. Evaluating ecosystem health. J. Aquat. Eco- syst. Health **1992**, *1*, 15–24.
81. Rapport, D.J.; Lalsley, W.L.; Rolston, E.R.; Nielsen, N.O.; Qualset, C.Q.; Damania, A.B. *Managing for Healthy Ecosystems*, Proceedings of the 1st Ecosystem Management Congress, 1999; CRC Press LLC, 2003; 1510 pp.
82. Zhang, M.; Smallwood, K.S.; Anderson, E. Relating indicators of ecosystem health and ecological integrity to assess risks to sustainable agriculture and native biota—A case study of Yolo County, California. In *Managing for Healthy Ecosystems*, Proceedings of the 1st Ecosystem Management Congress, 1999; Rapport, D.J., et al., Eds.; CRC Press LLC, 2003; 757–768.

83. Kaffka, S.R.; Dhawan, A.; Kirby, D.W. Irrigation, agricultural drainage, and nutrient loading in the Upper Klamath Basin. In *Managing for Healthy Ecosystems*, Proceedings of the 1st Ecosystem Management Congress, 1999; Rapport, D.J., et al., Eds.: CRC Press LLC, 2003; 1011–1026.
84. Mitchell, J.P.; Lanini, W.T.; Temple, S.R.; Brostrom, P.N.; Herrrero, E.V.; Miyao, E.M.; Prather, S.; Hembree, K.J. Reduced-disturbance agroecosystems in California. In *Managing for Healthy Ecosystems*, Proceedings of the 1st Ecosystem Management Congress, 1999; Rapport, D.J., et al., Eds.; CRC Press LLC, 2003; 993–997.
85. Hopkins, D.H. Fallow LAND Patches and ecosystem health in California's Central Valley agroecosystem. In *Managing for Healthy Ecosystems*, Proceedings of the 1st Ecosystem Management Congress, 1999; CRC Press LLC, 2003; 981–991.
86. Altieri, M.A.; Nicholls, C.I. Ecologically based pest management: A key pathway to achieving agroecosystem health. In *Managing for Healthy Ecosystems*, Proceedings of the 1st Ecosystem Management Congress, 1999; Rapport, D.J., et al., Eds.; CRC Press LLC, 2003; 999–1010.
87. Belcher, K.W.; Boehm, M. Evaluating agroecosystem sustainability using an integrated model. In *Managing for Healthy Ecosystems*, Proceedings of the 1st Ecosystem Management Congress, 1999; Rapport, D.J., et al., Eds.; CRC Press LLC, 1209–1226.
88. Waltner-Toews, D.; Wall, E. Emergent perplexity of postnormal questions for community and agroecosystem health. Soc. Sci. Med. **1997**, *45* (11), 1741–1749.
89. Ikerd, J. Assessing the health of agroecosystems: A socioeconomic perspective. In Proceedings of the 1st International Ecosystem Health and Medicine Symposium, University of Missouri; 1994; 9 pp.
90. Kolb, T.E.; Wagner, M.R.; Covington, W.W. Utilitarian and ecosystem perspectives. Concepts of forest health. J. For. **1994**, *92* (2), 10–15.
91. Percy, K.E.; Ferretti, M. Air pollution and forest health: Towards new monitoring concepts. Environ. Pollut. **2004**, 130, 113–126.
92. USDA Forest Service. Healthy forests for America's future: A strategic plan, USDA Forest Service MP-1513; 1993; 58 pp.
93. Spranger, T.; Lorenz, U.; Gregor, H.D., Eds. *Manual on Methodologies and Criteria for Modelling and Mapping Critical Loads and Levels and Air Pollution effects, Risks and Trends*; Federal Environmental Agency (Umweltbundesamt): Berlin, 2004.
94. Gauger, T.; Haenel, H.D.; Rösemann, C.; Dämmgen, U.; Bleeker, A.; Erisman, J.H.; Vermeulen, A.T.; Schaap, M.; Timmermanns, R.M.A.; Builtjes, P.J.H.; Duyzer, J.H. *National Implementation of the UNECE Convention on Long-range Transboundary Air Pollution (Effects)—Part 1: Deposition Loads: Methods, Modelling and Mapping Results, Trends*, Environmental Research of the Federal Ministry of the Environment, Natural Conservation and Nuclear Safety Research Report 204 63 252, No. 38/2008, UBA-FB 001189E; Federal Environmental Agency (Umweltbundesamt): Dessau, 2010a.
95. Gauger, T.; Haenel, H.D.; Rösemann, C.; Nagel, H.D.; Becker, R.; Kraft, P.; Schlutow, A.; Schütze, G.; Weigelt- Kirchner, R.; Anshelm, F. Nationale Umsetzung UNECE- Luftreinhaltekonvention (Wirkungen) Teil 2: Wirkungen und Risiokoabschätzungen Critical Loads, Biodiversität, Dynamische Modellierung, Critical Levels Überschreitungen, Materialkorrosion, Environmental Research of the Federal Ministry of the Environment, Natural Conservation and Nuclear Safety Research Report 204 63 252l No. 39/2008, UBA-FB 001189; Federal Environmental Agency (Umweltbundesamt): Dessau, 2010b.
96. UNECE Convention on Long-range Transboundary Air Pollution. Manual on methodologies for mapping critical loads/levels and geographical areas where they are exceeded, http://www.rivm.nl/thema/images/mapman-2004_tcm61-48383.pdf.
97. Warfinge, P.; Sverdrup, H. Calculating critical loads of acid deposition with profile—A steady soil chemistry model. Water, Air Soil Pollut. **1992**, *63*, 119–143.

98. Warfinge, P.; Sverdrup, H. Critical loads of acidity to Swedish forest soils, methods, data and results. Dep. Chem. Eng., Rep. Ecol. Environ. Eng. 5; Lund University: Lund, Sweden, 1995; 104 pp.
99. Ulrich, B. Stability, elasticity, and resilience of terrestrial ecosystems with respect to matter balance. In *Potentials and Limitations of Ecosystems Analysis*; Schulze, E.D., Zwölfer, H., Eds.; Springer: Berlin, 1987; 435 pp.
100. Boesch, D.F.; Paul, J.F. An overview of coastal environmental health indicators. Hum. Ecol. Risk Assess. **2001**, *7* (5), 1–9.
101. Su, M.; Fath, B.D.; Yang, Z. Urban ecosystem health assessment: A review. Sci. Total Environ. **2010**, *408*, 2425–2434.
102. Bell, M.L.; Cifuntes, L.A.; Davis, D.L.; Cushing, E.; Gusman Telles, A.; Gouveia, N. Environmental health indicators and a case study of air pollution in Latin American Cities. Environ. Res. **2011**, 111, 57–66.
103. Alberti, M. Maintaining ecological integrity and sustaining ecosystem function in urban areas. Curr. Opin. Environ. Sustainability **2010**, 2, 178–184.
104. Hancock, T. Health and sustainability in the urban environment. Environ. Impact Assess. Rev. **1996**, *16*, 259–277.
105. Tzoulas, K.; Korpela, K.; Venn, S.; Ylipelkonen, V.; Kazmierczak, A.; Niemela, J.; James, P. Promoting ecosystem and human health in urban areas using green infrastructure: A literature review. Landscape Urban Plann. **2007**, *81*, 167–178.
106. Rapport, D.J.; Singh, A. An EcoHealth-based framework for state of environment reporting. – Ecol. Indic. **2005**, *6*, 409–428.
107. Li, B.; Xie, H.L.; Guo, H.H.; Hou, Y. Study on the assessment method of agroecosystem health based on the pressure-state-response mode. In Proceedings of the IEEE International Conference on Industrial Engineering and Engineering Management, Hong Kong, Dec 8–11, 2009; 2458–2462.
108. Kumar, P., Ed. *The Economics of Ecosystems and Biodiversity: Ecological and Economic Foundations*; Earthscan, 2010; 456 pp.
109. Windhorst, W; Müller, F.; Wiggering, H. Umweltziele und Indikatoren für den Ökosystemschutz. In *Umweltziele und Indikatoren*; Müller, F., Wiggering, H., Eds.; Springer Verlag, 2003; 345–373.
110. Windhorst, W.; Colijn, F.; Kabuta, S.; Laane, R.P.; Lenhart, H. Defining a good ecological status of coastal waters—A case study for the Elbe plume. In *Managing European Coasts;* Vermaat, J.E., Bouwer, L., Turner, K., Salomons, W., Eds.; Springer Verlag, 2005; 59–74.
111. Schönthaler, K.; Meyer, U.; Pokorny, D.; Reichenbach, M.; Schuller, D.; Windhorst, W. *Ökosystemare Umweltbeobachtung—Vom Konzept zur Umsetzung*; Erich-Schmidt-Verlag: Berlin, 2004; 370 pp.
112. Reuter, H.; Middelhoff, U.; Schmidt, G.; Windhorst, W.; Schröder, W.; Breckling, B. Up-scaling the environmental effects of genetically modified plants—Assessing potential impact on nature conservation areas in Northern Germany. In *Multiple Scales in Ecology*; Schröder, B., Reuter, H., Reineking, B., Eds.; Peter Lang Internationaler Verlag der Wissenschaften: Frankfurt am Main. Band 13 edition. Theorie in der Ökologie, 2007; 95–109.
113. Rapport, D.J.; Regier, H.A.; Hutchinson, T.C. Ecosystem behavior under stress. Am. Nat. **1985**, *125* (5), 617–640.

20

Sustainable Fisheries: Models and Management

Fabian Zimmermann and Katja Enberg

Introduction: Population Dynamics and Fisheries

Population dynamics models are a key component of fisheries sciences to describe the changes in populations over time and their responses to fishing. The need for modeling approaches originates from the difficulties to observe fish directly. Together with the large socioeconomic relevance of fisheries, this has put fisheries models at the forefront of modeling biological systems. The focus has traditionally been on describing changes in population biomass through the growth and decay of a population's biomass. Historically, models have been divided into biomass models that lump entire populations into one biomass and models that are structured by age or size, allowing for more specific dynamics such as growth of body size, maturation, reproduction, recruitment and mortality. In this section, we will contrast biomass models with their structured counterparts, present models of growth, stock-recruitment and mortality, and outline current applications in a fisheries context.

Biomass, Age and Size Structure

In the 1950s, progress in industrial fishing triggered an increasing need to analytically describe the dynamics of fish populations in response to fishing, serving as basis to estimate the productivity of populations and maximize the yield of fisheries. Some of the first models to address these questions were the surplus production models, notably the Schaefer model by Milner B. Schaefer. These types of models describe a population as undifferentiated biomass that grows in response to the population size. Specifically, the Schaefer model (Schaefer 1954) assumes that changes in biomass are governed by population growth rate r, the carrying capacity K and the fisheries catch C:

$$\frac{dB}{dt} = rB\left(1 - \frac{B}{K}\right) - C \tag{1}$$

C can be defined in different ways, most typically it is introduced as a product of B and a fishing rate (which in turn is usually fishing effort times a catchability coefficient). The core of the equation is, however, the logistic growth of the population, which levels out to zero at $B = K$ (and at $B = 0$) and reaches a maximum of $\frac{rK}{4}$ at $B=\frac{K}{2}$. Consequently, in this model, B remains in perpetuity at $\frac{K}{2}$ if $C=\frac{rK}{4}$, which therefore maximizes the sustainable yield of a fishery. This idea builds the foundation of the concept of a maximum sustainable yield (MSY) that to date dominates global fisheries policies (Hey 2012).

There are various modifications of the Schaefer model (Haddon 2010), such as, for instance, the Fox model:

$$\frac{dB}{dt}=rB\left(1-\frac{\log(\mathrm{B})}{\log(K)}\right) \tag{2}$$

which accounts for the fact that productivity in many fish species is assumed to be not a symmetric parabola but exhibit a maximum at $B<\frac{K}{2}$. An attempt to integrate different functional forms into one generalized surplus production model is the Pella and Tomlinson model:

$$\frac{dB}{dt}=\frac{r}{s}B\left(1-\left(\frac{B}{K}\right)^{s}\right) \tag{3}$$

Here the parameter s defines the shape of the relationship between biomass and productivity, corresponding to the Schaefer model $s=1$ and displaying a left- or right-skewed parabola when $s<1$ or $s>1$, respectively.

Surplus production models can be considered as the first analytical models that have been applied systematically in fisheries science, both to assess the state of fish stocks as well as to determine reference points for management or explore conceptual questions. The biggest advantage of this class of models is their simplicity that requires comparatively little information or input data to generate (somewhat) meaningful predictions. It is for this reason why biomass-based models are still widely in use today, mostly in data-poor fisheries where the lack of information prevents the use of more sophisticated models. However, the simplicity is also the biggest caveat since omitting largely the biology and ecology of a population tends to oversimplify true dynamics. The demographic composition of a population and traits linked to demographics such as growth, maturity and fecundity are important factors of population dynamics. Ignoring these will, therefore, often result in biased or even completely wrong predictions. In the following sections, we will therefore focus on structured models of fisheries dynamics and their different components.

Growth

Besides the recruitment of new individuals to the population, increases in population biomass are caused by increasing body size of individual fish, making individual growth one of two major factors of a population's production. Fish typically grow from millimeter-sized eggs to adults that can reach up to several meters in some species while undergoing dramatic changes in their ecology and morphology. By constraining the available food and potential predation, the size of a fish is a major determinant of its ecological niche and habitat choice, while food acquisition and environmental conditions drive changes in size. Growth is a trait that is shaped by the entire life history and ecology of a species (Enberg et al. 2012), such that, typically, short-lived species tend to grow faster than long-lived ones. Furthermore, the size of fish commonly determines their fecundity and reproductive potential, and thus, the size composition of a population may be important for its productivity (Hixon Johnson, and Sogard 2014). Understanding growth is, therefore, essential to model population dynamics.

Based on the model design and their application, growth models in fish can be separated into two groups: process-based growth models and statistical growth models (Enberg, Dunlop, and Jørgensen 2008).

The former type of models defines growth through the underlying biology of an organism and the environment it experiences, shaping the processes that govern growth. Typically, process-based growth models use insights from bioenergetics to predict growth as a product of energy acquisition and transformation, accounting for metabolic processes, behavior, food intake, temperature or life-history trade-offs, as for instance between growth and reproductive investment. Process-based growth models aim at a fundamental understanding of growth, providing an approximation to the mechanisms that determine growth and enabling them to predict more accurately growth and how it may change in response to environmental and anthropogenic disturbances. In contrast, the second group of growth models focuses on a statistical description of observed sizes with little to none of the underlying biology of growth included. These models contain often fewer parameters and assume a direct relationship between age or size of an organism with growth. Examples are the logistic growth model

$$G_t = K\left(1 - \frac{W_t}{W_\infty}\right) \tag{4}$$

or the Gompertz growth model

$$G_t = K\left(\log\left(W_\infty\right) - \log\left(W_t\right)\right) \tag{5}$$

that assume that the growth rate G_t is a linear function of absolute or log-transformed weight W at time t subject to growth parameter K and an asymptotic maximum weight W_∞. While ignoring the biology of growth processes, statistical growth models can be applied to size-at-age data that is readily available for many commercially harvested fish populations. Process-based growth models, on the other hand, require more knowledge of the growth processes and data to fully parameterize the model. Particularly the necessary data on bioenergetic processes is often missing, turning the larger flexibility and predictive power of such models into a disadvantage in a practical context. Consequently, statistical growth models tend to be more common in many areas of fisheries science such as stock assessment, as they are sufficient to describe the observed size compositions of a population.

There is, however, an overlap between the two groups of growth models. Statistical growth models can be extended with additional physical or biological parameters, approximating better the actual drivers behind growth. On the other hand, process-based growth models are often used in purely statistical approaches. A good example for such applications is the von Bertalanffy growth model (VBGM), which stems from a model that includes anabolic and catabolic processes but is commonly fitted to size data like a statistical model.

Ludwig von Bertalanffy originally proposed a general growth model describing the change in length as a differential equation of length at time t, the maximum length L_∞, and a growth parameter r:

$$\frac{dL}{dt} = r\left(L_\infty - L_t\right). \tag{6}$$

Mechanistically, this is founded on the differences in how anabolic and catabolic processes scale with body weight W:

$$\frac{dW}{dt} = aW_t^{m_1} - cW_t^{m_2} \tag{7}$$

where a and c are the proportionality coefficients for anabolism and catabolism, respectively, and m_1 and m_2 the corresponding scaling exponents. Assuming that the acquisition and thus anabolic processes are proportional to body mass by $m_1 = \frac{2}{3}$ and catabolic processes (metabolism and maintenance) by $m_1 = 1$, this results with increasing body size in a higher proportion of available resources spent on catabolic processes, leaving less for growth.

The VBGM has later been introduced by Beverton and Holt into fisheries, where it has become and remained the dominating model for fish growth, mainly because it provides typically good fits to length or weight data of most fish species for both individuals and population averages. The standard form of the model is used to calculate the length or weight at time t and results in an asymptotic shape, i.e., size approaches a maximum while growth increments decrease over time:

$$L_t = L_\infty \left(1 - e^{-k(t-t_0)}\right) \tag{8}$$

$$W_t = aW_\infty \left(1 - e^{-k(t-t_0)}\right)^b \tag{9}$$

Length or weight are, therefore, determined by the growth coefficient k and an asymptotic (maximum) length L_∞ or weight W_∞. The exponent b is derived from the age–length relationship $L_t = aW_t^b$ (and in most applications simplified to a cubic relationship), and t_0 is an (hypothetical, negative) age when size is zero. The latter is included to avoid that length or weight equal to zero at hatching ($t = 0$).

The VBGM assumes discrete time steps that correspond in most applications to years. However, this may be inadequate from a biological perspective since temperate and boreal species typically do not grow equally throughout the year but show distinct intra-annual growth patterns that align with the seasonal food availability. To account for such dynamics, the standard VGBM can be modified by introducing cyclical growth patterns:

$$L_t = L_\infty \left(1 - e^{-\left(s_1 sin\left(\frac{2\pi(t-s_2)}{s_3}\right)+k(t-t_0)\right)}\right), \tag{10}$$

with s_1 and s_2 determining the shape of the oscillation and s_3 the frequency by subdividing the time step, e.g., in the most common case of annual time steps into months ($s_3 = 12$) or weeks ($s_3 = 52$).

Such modifications also reveal the main limitation of the standard VBGM: lacking functional flexibility and very simplified or erroneous biology. Specifically, the underlying scaling of anabolic and catabolic processes with size has been empirically shown as very similar, with values ranging between 0.7 and 0.8 for both exponents in most fish species. Furthermore, the VBGM neglects crucial insights from life-history theory and, therefore, captures juvenile growth less accurately than adult growth. The reason for this is that the model does not account for maturation and reproduction. Reproductive investment is very energy-demanding and, thus, requires organisms to allocate a major share of the acquired energy to it. Since resources for basic maintenance processes can only be reduced to a very limited extent, reproductive investment mainly comes at the expense of growth. Consequently, there is a trade-off between growth and reproduction, resulting in different growth trajectories before and after maturation. The VBGM does not incorporate these dynamics and, therefore, tends to underestimate the growth rates of juvenile fish, which typically approximates linear growth. The following two models attempt to address these limitations by allowing for more functional flexibility or specifically incorporating life-history considerations.

Generalized models serve the purpose of aggregating different models into one equation that can take different functional forms depending on the parameter values. The model developed by Schnute and Richards generalizes several of the growth models used in fisheries, including the VBGM and logistic model as well as models previously proposed by Gompertz, Chapman, Richards, or Schnute:

$$L_t = L_\infty \left(1 - ae^{-bt^c}\right)^{-\frac{1}{d}} \tag{11}$$

Setting parameters c and d equal to one, for instance, reduces the model to the VBGM. A generalized model such as the one from Schnute and Richards allows therefore for a better representation of observed data and biology of a population. Nevertheless, this presents a statistical approach rather than a processed-based one, possibly explaining data well but without biological understanding. Additionally, the large number of parameters may make the fitting process challenging, particularly if no prior (mechanistic) knowledge is included.

A different approach has been taken by biphasic growth models that build explicitly on life-history theory and incorporate the trade-off between growth and reproduction. These build on bioenergetics and the assumption that the change in somatic weight W depends on the acquired energy E_t and the energy invested in gonadal weight G_{t+1}:

$$W_{t+1} = W_t + E_t - G_{t+1} \tag{12}$$

Provided that weight is a cubic function of length, $W = L^3$, this can be used to model the growth in length. The model by Roff (1983) is one application of this idea, assuming a linear growth for juvenile fish, i.e., when age a is lower than the maturation age a_{mat}, whereas growth will depend on the gonado-somatic index (GSI) from the onset of maturation:

$$L_{t+1} = \begin{cases} L_t + l_0 \\ L_t + l_0 \Big/ (1 + R_{t+1})^{\frac{1}{3}} \end{cases} \text{for} \begin{cases} a < a_{mat} \\ a \geq a_{mat} \end{cases} \tag{13}$$

GSI is gonad mass divided by somatic mass, which means the larger the investment in R_{t+1}, the more will growth in length be reduced.

A very similar approach has been taken by the modified model based on Quince et al. (2008) in which the growth is also shaped by the reproductive investment R_{t+1} from of the GSI (Boukal et al 2014):

$$L_{t+1} = \sqrt[(1-\alpha)\beta]{\frac{L_t^{(1-\beta)\alpha} + (1-\beta)cb^{-(1-\beta)}}{1 + q^{-1}(1-\beta)R_{t+1}}} \tag{14}$$

For fish below a_{mat}, R_{t+1} equals zero, whereas from a_{mat} onward $R_{t+1} > 0$. The modified Quince et al. model resolves some limitations of Roff's model and incorporates a larger functional flexibility. This includes a less constrained length–weight relationship than Roff's model by assuming $W_t = aL_t^{\alpha}$, a conversion factor between somatic and gonadic investment q, and by not enforcing strictly linear growth for juvenile individuals. Both models, however, incorporate the same key feature that results in biphasic growth trajectories, separating quasi-linear growth prior to maturation from a decreased growth after the onset of maturation that depends on the reproductive investment. These models illustrate how our understanding of life-history processes can be included in growth models to represent better how growth can vary between different life stages and achieve better fits to empirical data.

Besides the trade-off between growth and reproduction, there are other trade-offs that may directly or indirectly affect growth. A major driver is survival and, therefore, everything that affects mortality, most notably predation and fishing. For instance, very size-selective mortality can increase survival for fish that invest more into growth (instead into basic maintenance processes such as the immune system) to grow faster through the size range of increased mortality. Similarly, behavioral adaptations may result in decreased or increased growth. For instance, passive behavior such as hiding can be used to reduce predation risks at the expense of reduced foraging, reducing the resources available for growth and reproduction. In contrast, growth may be increased through more active foraging and a bolder behavior; however, this may also expose fish to higher predation and reduce the probability of survival (Claireaux, Jørgensen, and Enberg 2018).

Essential life-history traits such as growth, reproductive investment, survival, and behavior are to a large part determined by an individual's inherited genotype. This means that traits are shaped by evolution and subject to evolutionary change that depends on the reproductive success of a specific life history within a given environment. Consequently, how resources are acquired and allocated into growth or reproduction is less of an individual decision than the result of an inherited life-history strategy. The trade-offs between growth, reproductive investment, and survival are key to this process and determine the success of a specific strategy under the current environmental conditions in reproducing and thus inheriting the same strategy to the next generation. Because mortality is a major driver of natural selection, changes in the degree or selectivity of mortality may affect the selection landscape and result in evolutionary adaptations in growth or traits that influence growth. This means growth trajectories within a population are not stationary over time but may change in response to environmental change and anthropogenic perturbation such as fishing or increasing sea water temperatures.

Although the growth of an organism is fundamentally shaped by life-history evolution, most observed changes in growth and thus size-at-age occur in the short term as a result of phenotypic plasticity. The most important driver is environmental variability, specifically physical and ecological conditions that influence metabolism and food availability. The latter is particularly relevant, since it determines directly the available energy that can be acquired by an organism and invested into processes such as growth. Food availability per capita is the combined result of food production through the food web and the competition for the available food sources. On a seasonal or annual time scale, growth of an organism can therefore be determined by bottom-up effects through variation in the ecosystem productivity, e.g., through varying nutrient inflow or temperatures, as well as through the abundance of its own population and other competing species. Feedbacks occur between the environmental variability and density dependence, for instance, when high food availability leads to increasing population abundance(s) and thus to competition in the future. Although the causes of environmental variability are often difficult to determine and parametrize in models, density-dependent growth has been empirically established (Zimmermann, Ricard, and Heino 2018) and may affect the sustainability of fisheries (van Gemert and Andersen 2018). A simple approach to implement density dependence in growth models is the use of an asymptotic length $L_{\infty,t}$ that decreases as a function of a density dependence coefficient d and the population biomass B_t in each year: $L_{\infty,t} = L_\infty - dB_t$ (Lorenzen and Enberg 2002). This example illustrates that not only insights from bioenergetics, physiology, and life-history theory are important to modeling growth but also insights from population ecology.

Recruitment

Recruitment is a key component of population dynamics and contributes together with body growth to the increase in biomass within a population. Because of the enormous reproductive potential of most fish species, recruitment tends to be the most important factor for the overall productivity of a population and the major driver of fluctuation in population size. Recruitment as such is the combined result of the total number of eggs produced by the mature part of a population and the survival throughout the early life stages from egg to juveniles, which explains the large variation in number of recruits observed in most fish. Typically, fish produce very large numbers of eggs per individuals, reaching millions per spawning event. At the same time, the early life stages are very vulnerable to unsuitable physical conditions, predation or insufficient food, causing large inter-annual variation in survival. Consequently, recruitment can result in favorable years in very large cohorts that sustain a population for many years during which recruitment may be average or fail completely. This variability in recruitment, however, poses also a challenge for any attempt to model and predict recruitment. Nevertheless, because recruitment is fundamental for population dynamics and thus fisheries, various recruitment models have been established. Most of these models rely on the basic assumption that recruitment must be related to the mature part of the population and is subject to some form of density-dependent reduction. The two models that remain most widely used until today are the stock-recruitment models developed by Ricker (1954) and Beverton and Holt (1957).

The Beverton–Holt model assumes a stock–recruitment relationship that increases with increasing biomass of mature fish B at time t, however, with decreasing number of recruits per spawning individual and thus approaching an asymptotic maximum of recruitment:

$$R_t = \frac{\alpha B_t}{\beta + B_t} \tag{15}$$

with α representing the asymptotic maximum for a given B_t and β the population biomass where $\alpha/2$ is reached, defining the steepness of the curve. Biologically, α stands for the maximum spawning output that linearly increases with population biomass, whereas β defines the density-dependent regulation in recruitment and therefore the productivity of a population at specific biomass levels. The underlying mechanism is the density-dependent survival of early life stages, which is assumed to decrease with increasing amount of eggs spawned by a larger population biomass due to intra-cohort competition for resources, particularly food (Van Poorten, Korman, and Walters 2018), and stronger predation pressure.

The Ricker model takes a similar approach as the Beverton–Holt model, except that it assumes an overcompensatory effect of increasing population biomass on recruitment. This implies that the Ricker stock-recruitment curve reaches a peak recruitment after which the realized number of recruits decreases again, instead of simply approaching an asymptotic maximum. The typical equation to describe this relationship is denoted as:

$$R_t = \alpha B_t e^{-\beta B_t} \tag{16}$$

Here, α defines the recruitment at a low biomass of the spawning population and scales the total number of recruits, whereas β determines the density-dependent decrease in recruits per spawning biomass. As in the Beverton–Holt model, α represents the reproductive output of the mature population and β the density-dependent mortality experienced by early life stages after spawning. The key difference between the two models is that the recruits per spawning biomass in the Ricker model do not remain at a value larger than zero but approach zero, suggesting that a population biomass above a certain level has such detrimental effects on recruitment that it overcompensates the marginal increase in spawning output. Such effects can occur through substantial negative inter-cohort interactions through cannibalism or competition (Ricard, Zimmermann, and Heino 2016), i.e., older cohorts that deplete the resources of following cohorts or prey directly on them, or other negative feedbacks, for instance when growth at early life stages is density-dependent while predation is size-dependent. This may add up to a substantially increased mortality when cohort density decreases growth rates.

It is noteworthy that the Ricker curve can take an almost identical shape as the Beverton–Holt curve for an observed range of population biomass and number of recruits, making the Ricker model more flexible in representing populations with different recruitment patterns. A step further in this direction is taken by generalized recruitment models that allow for a large functional flexibility with other models as special cases. One example for such an approach is a model suggested by Deriso and later modified by Schnute:

$$R_t = \alpha B_t (1 - \beta \gamma B_t)^{\frac{1}{\gamma}} \tag{17}$$

Here α and β take equivalent roles as in the Beverton–Holt or Ricker models, while parameter γ determines the form of the recruitment curve. For instance, when γ goes to 0, the Deriso model corresponds to the Ricker model, and $\gamma = -1$ transforms it into a Beverton–Holt-type model. A generalized model of this kind avoids the need for prior assumptions on the type of relationship, enabling better fits to data or to test effects of gradual changes in the functional form. The downside of such an approach is, however, that it takes mainly a statistical and not a process-orientated approach to incorporate biological knowledge. Furthermore, recruitment data typically turns out to be very noisy, which may make it very difficult to find reasonable parameter estimates for a model with a high degree of functional freedom.

Empirically, the biomass of the spawning population is in most cases an insufficient predictor of recruitment. As a consequence, models such as Beverton–Holt or Ricker typically fit poorly to data.

Recruitment data is in general very noisy and shows a lot of variation. Because survival at early life stages tends to be much more important than the total reproductive output of the mature individuals, other factors such as environmental conditions, food availability and predation that directly or indirectly affect mortality of eggs, larvae and juveniles are key drivers of recruitment (Zimmermann, Claireaux, and Enberg 2019). Especially large-scale atmospheric and oceanographic processes with cyclical patterns have been identified as important forcing factors. A simple approach to model such cyclical patterns in recruitment is to include an autoregressive term in a stock-recruitment model:

$$R = \alpha B_t e^{-\beta B_t} + u_t \tag{18}$$

representing a Ricker model that includes a $AR(1)$ process defined as $u_t = \varphi u_{t-1}$ with φ as autocorrelation coefficient. This allows for capturing temporal autocorrelation in recruitment time series and simulating inter-annual cyclical patterns.

A different approach to extend a Ricker model is to include explicitly an additional variable that is related with recruitment:

$$R = \alpha B_t e^{-\beta B_t} e^{-\gamma X_t} \tag{19}$$

Here X is the second variable besides mature biomass B and γ is the corresponding coefficient. Examples for X could include any environmental or ecological factor that is expected to affect the recruitment of a specific population, such as annual sea surface temperature, zooplankton indices or the biomass of a predator. This approach can be further extended with additional, equally defined terms.

Classic stock-recruitment models assume that the relationship between mature population biomass and recruitment remains stationary over time. However, this assumption may often not hold because the reproductive potential of a population as well as the mechanisms of density regulation can change over time due to external factors, both anthropogenic and natural. Regime shifts can, for instance, occur when changes in the environment or anthropogenic impacts alter the ecosystem productivity with implications on the reproductive success of a population and thus the relationship between the mature population and recruitment (Vert-pre et al. 2013). To incorporate two different regimes, a Ricker model can be modified:

$$R_t = \alpha_i B_t e^{-\beta_i B_t} \tag{20}$$

α and β depend now on $i = 1,2$ which can be defined as two different periods in a time series.

The number of regimes can be extended further if required. This comes, however, with an equally increasing number of parameters as caveat. Considering that time series of recruitment data usually only cover a few decades, the number of parameters can become easily disproportional compared to the number of data points. A solution to this problem is to introduce a time-invariant parameter that captures gradual changes in the relationship between population and recruitment over time (Perälä and Kuparinen 2015). For instance, if we assume that the productivity of the mature biomass varies with time, time-variant α_t can be introduced:

$$R_t = \alpha_t B_t e^{-\beta B_t} \tag{21}$$

where parameter α follows a random walk process $\alpha_t = \alpha_{t-1} + \sigma_t$, with σ assumed to be normally distributed. This enables the productivity term in the stock-recruitment relationship to vary gradually over time, representing changes in the population's reproductive output for instance through changes in the population's demographic structure due to fishing or because of increasing temperatures, as well as subsequent evolutionary adaptation (Enberg et al 2010).

Mortality and Population Dynamics

Whereas individual growth and recruitment represent the increase in biomass of a population, mortality constitutes the loss term in the equation. Survival is therefore a crucial component in fisheries models, making them very sensitive to the underlying assumptions and specifications of mortality. Mortality is typically given in rates that define the survival of individual fish over time. Two main sources of mortality are distinguished: mortality from natural causes, most notably predation and diseases, commonly denominated as natural mortality M, and mortality from anthropogenic harvesting, commonly termed fishing mortality F. These instantaneous rates of mortality can be summed to total mortality $Z = F + M$, which translates into proportional survival after a given time step as e^{-Z} and, reciprocally, relative mortality as $1 - e^{-Z}$. In an age-structured population, the abundance reduces in one time step t by:

$$N_t = N_{t-1}e^{-Z} \tag{22}$$

or alternatively an abundance at a given age a of:

$$N_a = N_0 e^{-Za}, \tag{23}$$

with N_0 as initial abundance of a cohort, typically corresponding to recruitment R_t.

Mortality is, however, in most fish species dependent on size and, thus, age of an individual. This applies both for F and M. While other sources of mortality such as diseases, parasite or starvation can be relevant as well, M is to a large degree shaped by predation. Vulnerability to predation is influenced by various traits, including the individual behavior, but most importantly body size because most predators have a size window for prey based on limitations in perception and handling. For a given stock, smaller and younger fishes tend to be much more vulnerable to predation than bigger and older ones, which means that mortality decreases substantially with increasing size and age.

F is subject to similar size selectivity, indirectly because parts of populations such as pre-recruits often do not share the same habitat as the ones targeted by fisheries and directly through size-selective fishing gear. In contrast to M, F increases in most cases with age and size: while small, young fishes are usually excluded from fisheries and experience very low fishing, the F for the targeted age and size classes can be substantial and outweigh M. These dynamics can be captured by using age-specific mortality rates F_a and M_a, leading to

$$N_{a,t} = N_{a,t-1}e^{-(F_a + M_a)} \tag{24}$$

This model can be further extended by allowing F and M also to vary in time:

$$N_{a,t} = N_{a,t-1}e^{-(F_{a,t} + M_{a,t})} \tag{25}$$

The two-dimensional mortality matrices here are age- and time-specific, which is common for stock assessment models where annual F_a's are estimated.

Stock Assessment

Managing fish stocks depends on knowing the state of the stock and its trajectory. Because fish are typically numerous, mobile and difficult to observe, this can be much more difficult than for classic terrestrial resources such as forests. Finding methods to assess a fish stock and use this information as the basis for regulative decisions is therefore at the core of sustainable fisheries management. This so-called stock assessment process relies commonly on observations, notably from commercial catches or standardized scientific sampling, and mathematical models to transform these often nonrepresentative, biased or snapshot-like observations into information on changes in the stock. The type and complexity of the model used is determined by the quantity, quality and structure of the available data.

The main distinction follows our previous differentiation of model complexity: biomass or age-/length-structured models. The latter require consistent information on the age or size of individual fish in the catch or scientific samples, which implies comparatively large and costly efforts in data collection over many years. Consequently, only a minority of global fish stocks are assessed in such a framework, mostly in commercially valuable fisheries in industrialized countries. All other fish stocks are classified as data-limited and are assessed with a wide range of methods (Chrysafi and Kuparinen 2015). Typically, these produce time series of absolute or relative biomass based on combining data on fisheries catch and effort with life-history information. However, to inform on the key parameters of population growth r and K (see Eq. 1), data of stock sizes at very high and low levels are necessary to give good indications on maximum growth rate and carrying capacity, respectively, which typically requires relatively long time series that ought to include unfished and overfished periods, what is rarely available. Alternatively, life-history parameters can act as predictors of stock productivity and resilience and therefore indicate adequate catch levels. This comes with the benefit that information gaps even in cases with very little or no data can be bridged by tapping into existing knowledge on ecology and life-history theory to incorporate it as priors into Bayesian assessment models (Kindsvater et al. 2018).

Age- or length-structured models can be divided into two major categories: statistical methods and cohort dynamics models. Both are based on the population dynamics detailed before in Eqs. 10 and 24 and, most importantly, the catch equation that relates total catch to age- or size-specific abundance and mortality:

$$C_{a,t} = \frac{F_a}{F_a + M_a} N_{a,t} \left(1 - e^{-(F_a + M_a)}\right) \tag{26}$$

In stock assessment, this equation provides an approach to determine the unknown variables F_a and N_a from catch data. Because time series of catch exist in essentially all commercial fisheries where assessments are conducted, the catch equation represents the core of almost all assessment models.

The main difference between statistical methods and cohort dynamics is the way these variables and other parameters are derived. Cohort dynamics, most commonly known as virtual population analysis (VPA) or more advanced derivatives such as extended survivor analysis (XSA), do not estimate their parameters but are instead based on an iterative process where the size of each cohort is calculated backwards in time and the result reseeded until the values of $N_{a,t}$ and $F_{a,t}$ approximate the observed $C_{a,t}$ sufficiently well. For this approach to work, other parameters, especially M_a, need to be given as fixed values. Later versions such as XSA were mostly developed to allow for the inclusion of catch indices from commercial or scientific surveys. VPA-type assessment models were particularly widespread in the assessments of European fisheries but have been increasingly replaced by statistical models in recent years.

Statistical catch-at-age models estimate their parameters directly by fitting Eq. 26 and other underlying equations to data by optimizing a given objective function. This offers greater flexibility and actual statistical evaluation compared to cohort dynamics models, yet comes with risks of over-parametrization and over-fitting when using ill-defined models. This can easily be illustrated with a model that assumes all $N_{a,t}$ and $F_{a,t}$ as independent parameters, which would result with, e.g., 10 age classes and a time series of only 20 years in 200 parameters alone, twice as many as data point. The problem is typically circumvented by linking parameters such as $F_{a,t}$ among each other through, e.g., a selectivity-at-age function or restricting changes over time. Together with increasing computational power and efficient optimization algorithms, this has elevated the utility of these models considerably and therefore also the number of stocks assessed with such, even though limitations remain (Subbey 2018).

Two of the most common frameworks for statistical assessment models are Stock Synthesis (Methot and Wetzel 2013) and, in recent years, State-space Assessment Model (SAM, Nielsen and Berg 2014). The strength of Stock Synthesis and reason for its popularity is mainly its flexible adjustment to a range of data situations, from relatively data-limited stocks to such with many data series and process knowledge available. SAM on the other hand is a state-space model framework which allows, thus, for a separation of observation and process error and the estimation of the development in F_a over time. Benefits include that time-varying selectivity can be directly estimated and uncertainty incorporated, representing possible solutions for two major problems in stock assessment. Both Stock Synthesis and SAM utilize comparable automatic differentiation algorithms that can estimate a large number of parameters efficiently based on maximum likelihood (or, in case of stock synthesis, also Bayesian inference).

Despite technical advances in assessment methods and increasing data availability, major challenges persist in stock assessment. Generally, there's a widespread lack of process knowledge and true mechanistic understanding that leads to biased or erroneous estimates (Maunder and Piner 2015). This includes insufficient knowledge on key biological processes, notably natural mortality, which are often described and parametrized in highly simplified ways that build a stark contrast to the increasing wealth of data and complexity when it comes to commercial and scientific catch indices. Natural mortality is in the large majority of assessed stocks kept constant over all age classes and over time. Similarly, maturity at age and growth are often described with time-invariant equations, despite existing knowledge that all these traits and the underlying processes can show large variation over time with significant impacts on stock dynamics. Ultimately, this is also reflected in the fact that most stock assessments maintain an isolated perspective on one specific stock and ignore its ecological context (Skern-Mauritzen et al. 2016), even though this excludes important information on drivers of dynamics and stock productivity. Incorporating ecosystem dynamics, for instance, by integrating them into natural mortality and growth in assessment models and their forecasts may therefore provide an important step forward in improving the prediction power and subsequently also the management of fisheries. Currently, the downside is the required data and effort that adds to an already work-intensive assessment process. One potential solution to this problem could be reduced assessment frequencies to free up capacities (Zimmermann and Enberg 2017) and focus on collecting the necessary ecosystem data to improve the quality of assessments.

Advice and Management

A key goal of modeling fish stocks and assessing them is to inform management strategies and shape regulative fisheries policies. The main outputs of the assessment process are therefore indicators of a stock's current state to evaluate them against reference points, forming the basis of the management advice. Reference points are important targets or thresholds in the state of a stock or the catch levels, which are typically linked to desirable objectives (notably, the stock size or fishing mortality that produces MSY) or unwanted situation (e.g., the stock size below which the risk of collapse exceeds a predefined limit). These thresholds are estimated from the available information, i.e., the specific approach depends again on data availability. If a complete analytical assessment exists, the estimated parameters (such as stock–recruitment relationship, length-at-age, natural mortality, fisheries selectivity-at-age, uncertainty) can be applied in stochastic simulation models of the population dynamics to calculate the desired reference points. In data-limited situations, there is often the need for different approaches due to the lack of a full assessment model. These utilize either catch and effort data or qualitative measurements of stock status (Froese et al. 2017) or are derived from biological information (Brooks, Powers, and Cortés 2009). It is noteworthy that in data-limited stocks, methods that rely on time series of catches tend to not perform better than simpler methods with lower data requirements (Carruthers et al. 2014).

Management advice to the regulatory agents is provided as management strategies and their outcomes, often with clear recommendations accompanied by a decision table that lists different options and how they perform in respect to reference points or policy objectives. Risk assessment is an important part of management advice, typically focusing on estimated probabilities of falling below limit thresholds of stock size or yield. Ideally, this translates into risk management that implements a precautionary approach to mitigate harmful impacts on fish stocks or fisheries (Hilborn et al. 2001). To develop and test management strategies, the most common framework today are management strategy evaluations (MSEs) that apply a suite of management strategies in simulation models to evaluate their long-term performance. These can include a broad range of (generic) management strategies or specifically developed ones. Main challenges hereby are the representation of uncertainty and the comprehensive transfer of results to managers and decision makers (Punt et al. 2016). The dimensions of a MSE increase with each management strategy or reference point considered, easily resulting in very large and complex outputs. It is therefore crucial that the results of MSE are summarized and presented in the best possible way, otherwise they may fail to convey the necessary information to managers and lead to suboptimal decision-making.

Ultimately, the key to successful fisheries management is the clear definition of objectives. A common reason for management failure is ill-specified or conflicting objectives (Hilborn 2007). Many objectives in fisheries management do not align or are even mutually exclusive. Classic trade-offs among objectives include food production and conservation or economic benefits and employment in the fisheries sector. It is, therefore, crucial that stakeholders such as policy makers, managers, fishermen, scientists and the public agree on the objectives that fisheries ought to achieve, to facilitate the design of suitable management strategies and their evaluation. Nevertheless, the success also hinges on the entire process of assessing stocks and, therefore, the adequate use of models to represent and predict stock dynamics sufficiently well. The major challenge of fisheries science remains therefore to collect the necessary information, use data efficiently and maximize the knowledge gain through suitable model choices.

The ecosystem approach to fisheries management has increasingly been recognized as a next step towards a more holistic and sustainable use of marine resources. Fish stocks are part of an ecosystem that is shaped by its environment, species composition and anthropogenic pressures. Climatic and oceanographic processes and trophic interactions among fish stocks and their prey and predators are important sources of variation in population dynamics that need to be accounted for if these dynamics ought to be accurately explained or predicted. Interactions between stocks are, for instance, important for management strategies because they not only affect recruitment and natural mortality but also result in trade-offs in reference points such as MSY (Voss et al. 2014) and rebuilding strategies of overfished stocks (Zimmermann and Yamazaki 2017). Furthermore, adaptations to selective harvesting (Heino et al. 2013), density effects (van Gemert and Andersen 2018) or climate change (Britten, Dowd, and Worm 2016) can affect the productivity of fish stocks and, thus, shift reference points. Ignoring such dynamics may therefore result in flawed reference points that increase the risk stock collapse and overfishing or underfishing, that is managing a fish stock suboptimally in respect to the biological or economic objectives. Finding robust and beneficial ways of increasingly integrating ecological information into assessment and management is therefore a major goal in our quest towards sustainable fisheries.

Currently, ecological interactions and the variation caused by them are still rarely included in stock assessment and management (Skern-Mauritzen et al. 2016). A major obstacle in the process towards ecosystem-based management has been that higher model complexity results in the curse of dimensionality and increasing uncertainty (Collie et al. 2016). Models that were meant to build the basis for ecosystem-based management by replicating entire food webs or ecosystems have turned out too complex, lack uncertainty quantification and are difficult to parametrize on a year-to-year basis, particularly for the many fisheries where data or knowledge are limited. For tactical fisheries management, current ecosystem models are therefore too data hungry and, most importantly, show poor estimation and prediction capabilities (Planque 2016). This is problematic because for management, the short-term prediction skills are crucial (Dietze et al. 2018) and there is no gain from incorporating ecosystem dynamics into management without sufficient prediction power (Basson 1999). A way forward is, therefore, an approach that focuses on integrating specific ecological or environmental interactions that can be effectively monitored, have high predictive power and significantly impact the dynamics of a population. Future research will subsequently focus on how such ecological interactions can in a systematic way be identified and incorporated into fisheries management processes.

References

Basson, M. 1999. “The importance of environmental factors in the design of management procedures.” *ICES Journal of Marine Science* 56 (6):933–942. doi:10.1006/jmsc.1999.0541.

Beverton, Raymond J.H., and Sidney J. Holt. 1957. “On the dynamics of exploited fish populations.” In Fisheries and Food, Fisheries Investigations *Series 2*, Vol. 19. London: UK Ministry of Agriculture.

Boukal, D., Dieckmann, U., Enberg, K., Heino, M., Jørgensen, C. (2014) Life-history implications of the allometric scaling of growth. *Journal of Theoretical Biology* **359**: 199–207.

Britten, Gregory L., Michael Dowd, and Boris Worm. 2016. "Changing recruitment capacity in global fish stocks." *Proceedings of the National Academy of Sciences* 113 (1):134–139.

Brooks, Elizabeth N., Joseph E. Powers, and Enric Cortés. 2009. "Analytical reference points for age-structured models: Application to data-poor fisheries." *ICES Journal of Marine Science* 67 (1):165–175. doi:10.1093/icesjms/fsp225.

Carruthers, Thomas R., André E. Punt, Carl J. Walters, Alec MacCall, Murdoch K. McAllister, Edward J. Dick, and Jason Cope. 2014. "Evaluating methods for setting catch limits in data-limited fisheries." *Fisheries Research* 153:48–68. doi:10.1016/j.fishres.2013.12.014.

Chrysafi, Anna, and Anna Kuparinen. 2015. "Assessing abundance of populations with limited data: Lessons learned from data-poor fisheries stock assessment." *Environmental Reviews* 24 (1):25–38. doi:10.1139/er-2015-0044.

Claireaux, M., Jørgensen, C., Enberg, K. (2018) Evolutionary effects of fishing gear on foraging behavior and life-history traits. *Ecology and Evolution* **8**, 10711–10721.

Collie, Jeremy S., Louis W. Botsford, Alan Hastings, Isaac C. Kaplan, John L. Largier, Patricia A. Livingston, Éva Plagányi, Kenneth A. Rose, Brian K. Wells, and Francisco E. Werner. 2016. "Ecosystem models for fisheries management: Finding the sweet spot." *Fish and Fisheries* 17:101–125.

Dietze, M.C., Fox, A., Beck-Johnson, L.M., et al. (2018) Iterative near-term ecological forecasting: Needs, opportunities, and challenges. *Proceedings of the National Academy of Sciences* **115**, 1424–1432.

Enberg, K., E.S. Dunlop, and C. Jørgensen. 2008. Fish growth. In Encyclopedia of Ecology, Edited by S.E. Jørgensen and B.D. Fath. Elsevier, Oxford. Pp 1564–1572.

Enberg, K., Jørgensen, C., Mangel, M. (2010) Fishing-induced evolution and changing reproductive ecology of fish: the evolution of steepness. *Canadian Journal of Fisheries and Aquatic Sciences* **67**: 1708–1719.

Enberg, K., C. Jørgensen, E.S. Dunlop, Ø. Varpe, D.S. Boukal, L. Baulier, S. Eliassen, and M. Heino. 2012. "Fishing induced-evolution of growth: Concepts, mechanisms and the empirical evidence." *Marine Ecology* 33 (1):1–25. doi:10.1111/j.1439–0485.2011.00460.x.

Froese, Rainer, Nazli Demirel, Gianpaolo Coro, Kristin M. Kleisner, and Henning Winker. 2017. "Estimating fisheries reference points from catch and resilience." *Fish and Fisheries* 18 (3):506–526. doi:10.1111/faf.12190.

Haddon, Malcolm. 2010. *Modelling and Quantitative Methods in Fisheries*. Second ed. Boca Raton, FL: CRC Press.

Heino, Mikko, Loïc Baulier, David S. Boukal, Bruno Ernande, Fiona D Johnston, Fabian M Mollet, Heidi Pardoe, Nina O Therkildsen, Silva Uusi-Heikkilä, and Anssi Vainikka. 2013. "Can fisheries-induced evolution shift reference points for fisheries management?" *ICES Journal of Marine Science: Journal du Conseil* 70 (4):707–721.

Hey, Ellen. 2012. "The persistence of a concept: Maximum sustainable yield." *The International Journal of Marine and Coastal Law* 27 (4):763–771.

Hilborn, R. 2007. "Defining success in fisheries and conflicts in objectives." *Marine Policy* 31 (2):153–158.

Hilborn, R., J.J. Maguire, A.M. Parma, and A.A. Rosenberg. 2001. "The precautionary approach and risk management: Can they increase the probability of successes in fishery management?" *Canadian Journal of Fisheries and Aquatic Sciences* 58 (1):99–107.

Hixon, Mark A., Darren W. Johnson, and Susan M. Sogard. 2014. "BOFFFFs: On the importance of conserving old-growth age structure in fishery populations." *ICES Journal of Marine Science* 71 (8):2171–2185. doi:10.1093/icesjms/fst200.

Kindsvater, Holly K., Nicholas K. Dulvy, Cat Horswill, Maria-José Juan-Jordá, Marc Mangel, and Jason Matthiopoulos. 2018. "Overcoming the data crisis in biodiversity conservation." *Trends in Ecology & Evolution*. doi:10.1016/j.tree.2018.06.004.

Lorenzen, K., and K. Enberg. 2002. "Density-dependent growth as a key mechanism in the regulation of fish populations: Evidence from among-population comparisons." *Proceedings of the Royal Society B: Biological Sciences* 269 (1486):49–54.

Maunder, Mark N., and Kevin R. Piner. 2015. "Contemporary fisheries stock assessment: Many issues still remain." *ICES Journal of Marine Science* 72 (1):7–18. doi:10.1093/icesjms/fsu015.

Methot, Richard D., and Chantell R. Wetzel. 2013. "Stock synthesis: A biological and statistical framework for fish stock assessment and fishery management." *Fisheries Research* 142:86–99. doi:10.1016/j.fishres.2012.10.012.

Nielsen, Anders, and Casper W. Berg. 2014. "Estimation of time-varying selectivity in stock assessments using state-space models." *Fisheries Research* 158: 96–101.

Perälä, Tommi, and Anna Kuparinen. 2015. "Detecting regime shifts in fish stock dynamics." *Canadian Journal of Fisheries and Aquatic Sciences* 72 (11):1619–1628.

Planque, Benjamin. 2016. "Projecting the future state of marine ecosystems, 'la grande illusion'?" *ICES Journal of Marine Science: Journal du Conseil* 73 (2):204–208.

Punt, André E., Doug S. Butterworth, Carryn L. de Moor, José A. A. De Oliveira, and Malcolm Haddon. 2016. "Management strategy evaluation: Best practices." *Fish and Fisheries* 17 (2):303–334. doi:10.1111/faf.12104.

Quince, C., P.A. Abrams, B.J. Shuter, and N.P. Lester. 2008. "Biphasic growth in fish I: Theoretical foundations." *Journal of Theoretical Biology* 254 (2):197–206.

Ricard, D., F. Zimmermann, and M. Heino. 2016. "Are negative intra-specific interactions important for recruitment dynamics? A case study of Atlantic fish stocks." *Marine Ecology Progress Series* 547:211–217. doi:10.3354/meps11625.

Ricker, William E. 1954. "Stock and recruitment." *Journal of the Fisheries Board of Canada* 11 (5):559–623.

Roff, Derek A. 1983. "An allocation model of growth and reproduction in fish." *Canadian Journal of Fisheries and Aquatic Sciences* 40 (9):1395–1404.

Schaefer, M.B. 1954. "Some aspects of the dynamics of populations important to the management of the commercial marine fisheries." *Inter-American Tropical Tuna Commission Bulletin* 1 (2):27–56.

Skern-Mauritzen, Mette, Geir Ottersen, Nils Olav Handegard, Geir Huse, Gjert E. Dingsør, Nils C. Stenseth, and Olav S. Kjesbu. 2016. "Ecosystem processes are rarely included in tactical fisheries management." *Fish and Fisheries* 17 (1):165–175.

Subbey, Sam. 2018. "Parameter estimation in stock assessment modelling: Caveats with gradient-based algorithms." *ICES Journal of Marine Science*. doi: 10.1093/icesjms/fsy044.

van Gemert, Rob, and Ken H. Andersen. 2018. "Challenges to fisheries advice and management due to stock recovery." *ICES Journal of Marine Science*. doi:10.1093/icesjms/fsy084.

Van Poorten, B., Korman, J., Walters, C. (2018) Revisiting Beverton–Holt recruitment in the presence of variation in food availability. *Reviews in Fish Biology and Fisheries* **28**: 607–624.

Vert-pre, Katyana A., Ricardo O. Amoroso, Olaf P. Jensen, and Ray Hilborn. 2013. "Frequency and intensity of productivity regime shifts in marine fish stocks." *Proceedings of the National Academy of Sciences of the United States of America* 110 (5):1779–1784. doi:10.1073/pnas.1214879110.

Voss, Rudi, Martin Quaas, Joern O. Schmidt, and Julia Hoffmann. 2014. "Regional trade-offs from multi-species maximum sustainable yield (MMSY) management options." *Marine Ecology Progress Series* 498: 1–12.

Zimmermann, Fabian, Marion Claireaux, and Katja Enberg. 2019. "Common trends in recruitment dynamics of north-east Atlantic fish stocks and their links to environment, ecology and management." *Fish and Fisheries*. doi:10.1111/faf.12360.

Zimmermann, Fabian, and Katja Enberg. 2017. "Can less be more? Effects of reduced frequency of surveys and stock assessments." *ICES Journal of Marine Science* 74 (1):56–68. doi:10.1093/icesjms/fsw134.

Zimmermann, Fabian, Daniel Ricard, and Mikko Heino. 2018. "Density regulation in Northeast Atlantic fish populations: Density dependence is stronger in recruitment than in somatic growth." *Journal of Animal Ecology* 87 (3):672–681. doi:10.1111/1365-2656.12800.

Zimmermann, Fabian, and Satoshi Yamazaki. 2017. "Exploring conflicting management objectives in rebuilding of multi-stock fisheries." Ocean & Coastal Management 138:124–137. doi: 10.1016/j.ocecoaman.2017.01.014.

V

ENT: Environmental Management Using Environmental Technologies

21

Bioremediation: Contaminated Soil Restoration

Sven Erik Jørgensen

Introduction

Contaminated soil is an increasing problem in industrialized countries, and the high cost of remediation has driven the interest in the direction of ecological engineering applications of bioremediation technologies. They apply biological processes, mainly microorganisms or plants. It is often possible to solve the pollution problem satisfactorily by this ecotechnologically based methodology without the hazard and expense involved in removing polluted materials for treatment elsewhere for the use of traditional environmental technological methods.

Ecotechnological bioremediation may be applied for both organic waste and heavy metals, although the methods applied in practice may differ and they are therefore treated below in two different sections, on organic compounds and heavy metals. The success of any bioremediation technology depends on a number of factors including site characteristics, environmental factors such as temperature, pH, redox potential, concentrations of nutrients, the contaminant, the presence of microorganisms, and bioavailability. It is therefore necessary to look into these factors to comprehend the applicability of these methods.

Bioavailability of Toxic Organic Compounds

Bioavailability is a crucial factor for the application of bioremediation. It is defined as the amount of contaminant present that can be readily taken up by organisms. The bioavailability controls the biodegradation rates for organic contaminants because microbial cells must expend energy to induce the catabolic processes used in biodegradation. If the contaminant concentration is too low, induction will not occur. Soil microbial populations are typically slow-growing organisms and often exposed to nutrient-poor environments.[1] Bioavailability also determines the toxicity of both organic and inorganic contaminants to organisms other than those applied for bioremediation. There is therefore an increased need for bioremediation when bioavailability is high, which fortunately makes bioremediation more attractive. Three cases can be envisioned that would result from different bioavailabilities of contaminants:[2]

1. Biodegradation will not occur because the concentration of the bioavailable contaminant is insufficient and/or the biodegradation rate for the contaminant is too low to justify the energy expenditure to induce biodegradation.
2. Microbial cells may degrade the contaminant at low bioavailable concentrations and/or low biodegradation rates, but in a resting or maintenance stage rather than in a growing stage.
3. At a sufficient bioavailability and biodegradation rate, there is enough bioavailable contaminant to induce biodegradation in a growing stage. That will allow for optimal rates of remediation.

The biodegradability of organic contaminants is highly dependent on the physical and chemical structure[3] of the contaminant and the soil. The section on "Biodegradation" will discuss this topic. A coarse but still applicable rule for a very first estimation of the biodegradability of organic compounds is given in the entry entitled "Biodegradation." Moreover, the software EEP (Estimation of Ecotoxicological Parameters) is able to give some first estimation of biodegradability.

The bioavailability of heavy metals is also a significant factor for the applicability of bioremediation. Heavy metals are of course not degraded but removed, mainly by plant uptake. The uptake by organisms of heavy metals, which determines the overall removal efficiency, is entirely controlled by the bioavailable amount of heavy metals. An ecological model presented in the section on *Uptake of Heavy Metals by Plants* will illustrate the strong dependence of the bioavailability of heavy metals.[4]

Bioavailability is influenced by a number of factors:

1. Low water solubility
2. Sorption on the solid phase of soil
3. Physical makeup of the soil (pore size distribution)
4. Microbial adaptations

Low water solubility can limit availability of the substrate to bacterial cells and hence constrain biodegradation.[5,6] Microbial cells are 70%–90% water, and the food they utilize comes from the water surrounding the cells. Plants take up water to cover the evapotranspiration needed for the maintenance of their life functions. Therefore, uptake and transport are only feasible for water-soluble material. If first-order biodegradation kinetics is presumed, the biodegradation rate becomes proportional to the concentration in the water phase. It means for components with low water solubility that they are biodegraded very slowly. There are clear relationships between the water solubility of an organic compound and the chemical structure that can be utilized to estimate the water solubility.[3] EEP (a software containing many equations to estimate ecotoxicological parameters) and other estimation equations utilize these relationships to make estimation of the water solubility and of K_{ow}. Side reactions may change the water solubility. This is of particular interest for heavy metal ions, which can increase the solubility by the formation of complexes either with organic or with inorganic compounds. The formation of complexes with humic acid and fluvic acid plays a major role for the solubility of metal ions in soil water. Hydrocarbons that are frequently found as soil contaminant have a low water solubility: 2–6 μg/L for pentacyclic aromatic hydrocarbons and *n*-alkanes of chain length 18–30. The solubility decreases with increasing molecular weight.[3]

The state of the contaminant in combination with the water solubility is also of importance. There is evidence that liquid-phase hydrocarbons are more bioavailable than solid-phase hydrocarbons.[7] In practical terms, this means that the maximum growth rate occurs in different solubility ranges for liquid-phase (0.01–1 mg/L) and solid-phase (1–10 mg/L) components. Degradation can be described by a Michaelis–Menten expression. Water solubility increases with increasing temperature and usually an Arrhenius expression can be applied with the temperature coefficient 1.06 or 1.07.

Many authors have found that surfactants increase mineralization rate due to increased dissolution. Also, surfactants may provide an additional carbon source, which is preferentially utilized by the bacteria. There may, how-ever, also be a negative effect by surfactants due to their toxicity to the bacterial population.

Sorption on the solid phase of soil may be a limiting factor for biodegradation of microorganisms and uptake by plants. There are several reports that suggest that organic chemicals are not mineralized while associated with solid phases.[8,9] Experiments by Robinson et al.[10] show that sorbed-phase substrate was not degraded and that longterm biodegradation was limited by the slowly desorbing fraction of substrate. These results suggest that rate-limited mass transfer processes (primary desorption) may significantly affect the rate at which a compound is degraded in the presence of a solid phase.

The model presented in the section on "Uptake of Heavy Metals by Plants" uses the fraction soluble in the soil water of heavy metal ions to determine the uptake. The sorption is dependent of pH, redox potential, and humus, clay, and sand fractions in the soil. The relationship between these factors and the sorption is included in the model. If the sorption of organic compounds to soil is not known, the soil–water partition coefficient, K_{oc}, can be estimated from the octanol–water partition coefficient by the following equations:

$$\log K_{oc} = -0.006 + 0.937 \log K_{ow} \tag{1}$$

$$\log K_{oc} = -0.35 + 0.99 \log K_{ow} \tag{2}$$

In the case that the carbon fraction of organic carbon in soil is f, the distribution coefficient, K_D, for the ratio of the concentration in soil and in water can be found as $K_D = K_{oc} f$.

It has been suggested that there are different stages of sorption processes and that newly sorbed material is more labile and therefore more bioavailable than aged sorbed material. Numerous experiments have demonstrated that aging affects bioavailability in soil due to changes in the soil structure, resulting in slower desorption processes. The sorption can frequently be described by either Freundlich or Langmuir adsorption isotherm, expressed respectively by the following equations:

$$a = kc^b \tag{3}$$

$$a = k'c/(c + b') \tag{4}$$

where a is the concentration in soil, c is the concentration in water, and k, k', b, and b' are constants. Equation 3, corresponding to Freundlich adsorption isotherm, is a straight line with slope b in a log–log diagram, since $\log a = \log k + b \log c$.[11] The Langmuir adsorption isotherm is an expression similar to the Michaelis–Menten equation. If $1/a$ is plotted versus $1/c$,[11] we obtain a straight line, Lineweaver-Burk's plot, as $1/a = 1/k' + b'/k'c$. When $1/a = 0$, $1/c = -1/b'$ and when $1/c = 0$, $1/a = 1/k'$. This plot can be applied to assess the expression of the type used in Michaelis–Menten's equation and in Langmuir's adsorption isotherm; it is observed that b is often close to 1 and c is for most environmental problems small. This implies that the two adsorption isotherms get close to $a/c = k$, and k becomes a distribution coefficient. k for 100% organic carbon is usually denoted K_{oc} (see above).

The sorption determines the uptake of organic contaminants by plants as it is expressed in the following equation:[11]

$$\mathrm{BCF} = f_{\mathrm{lipid}} K_{ow}{}^{b} / hfK_{ow}{}^{a}, \tag{5}$$

where BCF, the biological concentration factor, expresses the ratio between the concentration in soil and in the plant (or the microorganisms); f_{lipid} is the lipid fraction in the plant; f is, as shown above, the fraction of organic carbon in the soil; and a, b, and h are constants. The denominator expresses the fraction of the organic matter that is dissolved in the soil water. h is therefore the constant determined by Equations 1 and 2. If we use Equation 1, Equation 5 may be reformulated to the following equation:

$$\mathrm{BCF} = 1.01 f_{\mathrm{lipid}} K_{\mathrm{ow}}{}^{0.063} / f \tag{6}$$

As it is seen, BCF becomes almost independent of K_{ow} and mainly dependent on the ratio between f_{lipid} and f. A high BCF means that the concentration in the plants (eventual microorganisms) is high and a significant amount of the toxic compound is removed; the lipid fraction in the plants has to be high and the carbon content of the soil has to be low.

Physical makeup of the soil (*pore size distribution*) is of importance for bioavailability. Bacteria may be excluded from the microporous domain since most bacteria range from 0.5 to 2 μm. If such an exclusion occurs, biodegradation cannot take place in the microporous domain. The degradation rate is therefore limited by the diffusion of solute from the microporous to the macroporous domain. This is obviously of particular importance for organic contaminants with a high molecular weight. In a field situation, it is difficult to separate the effects of sorption and micropore exclusion, as some residues are protected from biodegradation by both mechanisms.

Microbial adaptations: Microorganisms have developed several strategies to increase the bioavailability of organic contaminant. One strategy is the development of increased cell affinity for hydrophobic surfaces. It allows the microorganisms to attach to the hydrophobic substrate and directly adsorb it. A second strategy is the production and release of surface active agents or biosurfactants.[12,13]

The biological adaptation is a current change of the properties of the microbial population by a selection of the microorganisms that are best fitted to survive and grow under the prevailing conditions. They are determined by the properties of the environment including the concentrations and characteristics of the contaminants. A biological adaptation is widely used to prepare a microorganism population for bioremediation. It is often possible, although not general, that a 10 times faster decomposition can be achieved by the use of adapted microorganisms.[2]

Biodegradation

See the entry entitled "Biodegradation," where a general presentation of this process included methods for estimation of the biodegradation rate from the chemical structure of the toxic organic compounds.

The usual applied procedure to follow for the utilization of microbiological biodegradation to reduce the concentration of a toxic organic matter in contaminated soil has six steps:

1. Spatial mapping of quantitative distribution: the contaminant is developed by analytical chemistry intensively. Analyses of pore water are often applied to evaluate the extent of environmental risks.
2. Laboratory test/treatability studies to verify the applicability of bioremediation.
3. Calculation (often by development of a model) to assess the feasibility of the method in situ.
4. Production of an adapted strain of the microorganisms in sufficient amount.
5. Implementation in situ. If the groundwater table is high, it is usually lowered. Injection pits are introduced into the soil, and air is blown into the soil to reinforce the decomposition of organic matter. In case of chlorinated compounds, a mixture of methane and air may be applied.
6. The results are followed by use of a wide spectrum of analytical methods including radioactive tracers, detection of intermediary metabolites, and respiration rate.

Uptake of Heavy Metals by Plants

Plants are contaminated by heavy metals originating from deposition of heavy metals (waste sites), air pollution, the application of sludge from municipal wastewater plant as a soil conditioner, and the use of fertilizers. The uptake of heavy metals from municipal sludge by plants has previously been modeled.[4] The model is based on a mass balance for cadmium in a typical Danish soil (see Figure 1).

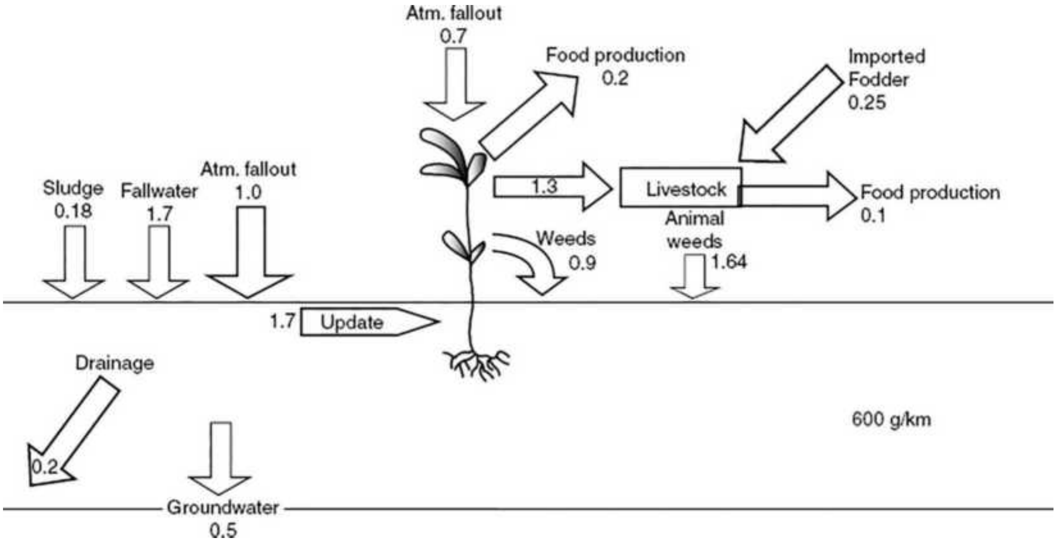


FIGURE 1 Cadmium balance of 1 ha of an average Danish agriculture land. All rates are expressed as grams of Cd per hectare per day.

The model can briefly be described as follows: Depending on the soil composition, it is possible to find for various heavy metal ions a distribution coefficient, i.e., the fraction of the heavy metal that is dissolved in the soil water relative to the total amount. The distribution coefficient was found by examination of the dissolved heavy metals relative to the total amount for several different types of soil. Correlation between pH, the concentration of humic substances, clay, and sand in the soil on the one hand, and the distribution coefficient on the other, was also determined. The uptake of heavy metals was considered a first-order reaction of the dissolved heavy metal. It is, how-ever, also possible to test acid volatile sulfide and organic carbon to describe the metal binding capacity of sediment in constructed wetlands. This will give approximately the same ratio "bound" to "bioavailable" heavy metals as the above-mentioned correlation. The basic idea is the same, namely, to find easily measurable soil properties that determine the metal binding capacity, which is crucial for the uptake of heavy metals by plants.

In addition to the uptake from soil water, the model presented below considers the following:

1. Direct uptake from atmospheric fallout onto the plants.
2. Other sources of contamination such as fertilizers and the long-term release of heavy metal bound to the soil and the non-harvested parts of the plants.

Published data on lead and cadmium contamination in agriculture are used to calibrate and validate the model that is intended to be used for the following:

1. Risk assessment for the use of fertilizers and sludge that contain heavy metals as contaminants.
2. A risk involved in the use of plants harvested from a waste site.
3. Determining the possibilities of removal of heavy metals by plants that have a particular ability to take up heavy metals. This last intended application of the model makes it useful for determination of the result of application of bioremediation.

Figure 2 shows a conceptual diagram of the Cd version of the model. The STELLA software was applied. As can be seen, it has four state variables: Cd-bound, Cd-soil, Cd- detritus, and Cd-plant. An attempt was made to use one or two state variables for cadmium in the soil, but to get acceptable accordance between data and model output, three state variables were needed. This can be explained by the presence of several soil components that bind the heavy metal differently.

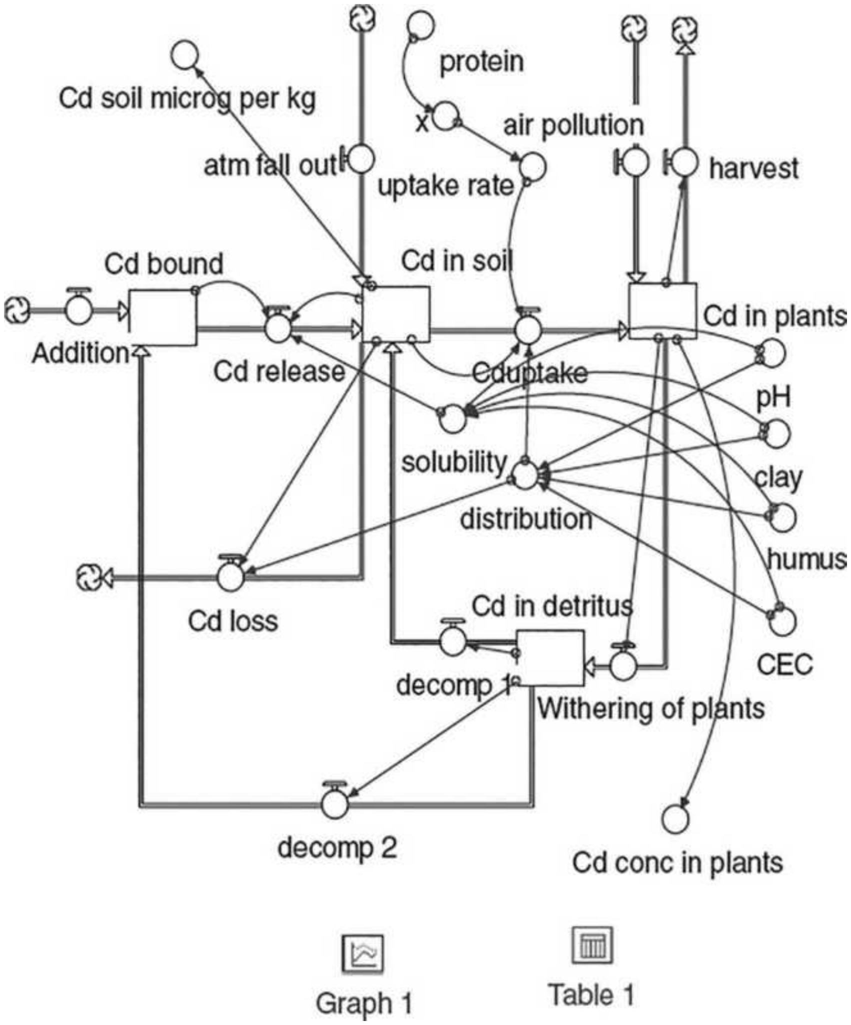


FIGURE 2 Conceptual diagram of the model. Boxes show state variables, double-line arrows denote flows, circles show functions, and single-line arrows show feedback mechanisms.

The loss covers transfer to the soil and groundwater below the root zone. It is expressed as a first-order reaction with a rate coefficient dependent on the distribution coefficient that is found from the soil composition and pH, according to the correlation found by Jørgensen.[14] The transfer from Cd-bound to Cd-soil indicates the slow release of cadmium due to a slow decomposition of the more or less refractory material to which cadmium is bound. The cadmium uptake by plants is expressed as a first-order reaction, where the rate is dependent on the distribution coefficient, as only dissolved cadmium can be taken up. It is furthermore dependent on the plant species. As will be seen, the uptake is a step function that, here (grass), is 0.0005 during the growing season, and, of course, zero after the harvest and until the next growing season starts. Cd-waste covers the transfer of plant residues to detritus after harvest. It is therefore a pulse function, which here is 60% of the plant biomass, as the remaining 40% has been harvested. Cd-detritus covers a wide range of biodegradable matter and the mineralization and is therefore accounted for in the model by use of two mineralization processes: one to Cd-soil and one to Cd-bound. The first one is rapid and is given a higher rate for the first 180 days as the addition of municipal sludge in this case is at day 0. The second one is at about the same rate, but as the cadmium is transferred to the Cd-bound, the slow release rate is considered by the very slow transfer from Cd-bound to Cd-soil. Similar models can be erected for other heavy metals, but the distribution coefficient is of course different for the other heavy metals.[14]

Plants with a high protein content will generally take up heavy metals with a higher efficiency. Several applications of plants with a high uptake efficiency have been reported in the literature; see Tongbin et al.,[15] where a very effective uptake of arsenic is reported, and Feng et al.,[16] where simultaneous removal of arsenic and antimony is reported. The use of algae for removal of heavy metals has also been successfully tested.[17]

Models are used increasingly to solve the problems of contaminated soil (see UNEP-EITC and Copenhagen University[18]). Spatial models have been developed to consider the distribution of the contaminants in eco-spatial time scale. These models are very useful in setting up and optimizing time-bound action plans of bioremediating the contaminated soil.

References

1. Roszak, D.B.; Colwell, R. Survival strategies of bacteria in natural environment. Microbiol. Rev. **1987**, *51*, 365–379.
2. Maier, R.M. Biovailability and its importance for bioremediation. In *Bioremediation*; Valdes, J.J., Ed.; Kluwer: Dordrecht, the Netherlands, 2000; 58–79.
3. Jørgensen, S.E.; Halling-Sørensen, B.; Mahler, H. *Handbook of Estimation Methods in Ecotoxicology and Environmental Chemistry*; Taylor and Francis Publ.: Boca Raton, FL, 1997; 230 pp.
4. Jørgensen, S.E.; Fath, B. *Fundamentals of Ecological Modelling*, 4th Ed.; Elsevier: Amsterdam, 2011; 398 pp.
5. Fogel et al. EPA Report 560/5–82-015 Washington, DC, 1981.
6. Zhang, Y.; Miller, R.M. Enhanced octadecane dispersion and biodegration by *Pseudomonas rhamnolipid* surfactant. Appl. Environ. Microbiol. **1992**, *58*, 3276–3282.
7. Miller, R.M. Surfactant-enhanced bioavailability of slightly soluble organic compounds. In *Bioremediation: Science and Applications*; Skipper, H., Turco, R., Eds.; Specila Publications. Soil Science Society of America: Madison, WI, 1995; 33–54.
8. Greer, L.E.; Shelton, D.R. Surfactant-enhanced bioavailability of slightly soluble organic compounds. In *Bioremediation*; Skipper, H., Turco, R., Eds.; Soil Science Society of America: Madison, WI, 1992; 33–54.
9. Miller, M.E.; Alexander, M. Kinetics of bacterial degradation of benzylamine in montmorillonic suspension. Environ. Sci. Technol. **1991**, *25*, 240–245.
10. Robinson K.G. et al. Availability of sorbed toluene in solids for biodegradation by acclimatized bacteria. Water Res. **1990**, *24*, 345–350.
11. Jørgensen, S.E. *Principles of Pollution Abatement*; Elsevier: Amsterdam, 2000; 520 pp.
12. Rosenberg, E. Microbial surfactants. Crit. Rev. Biotechnol. **1986**, *3*, 109–132.
13. Fiechter, A. Biosurfactant: moving towards industrial application. Trends Biotechnol. **1992**, *10*, 208–217.
14. Jørgensen, S.E. Do heavy metals prevent the agricultural use of municipal sludge? Water Res. **1975**, *9*, 163–170.
15. Tongbin, C. et al. Arsenic hyperaccumulator *Pteris vittata* L. and its arsenic accumulation. Chin. Sci. Bull. **2002**, *47*, 902–905.
16. Feng R. et al. Simultaneous hyper-accumulation of arsenic and antimony in Cretan brake fern: Evidence of plant uptake and sub-cellular distributions. Microchem. J. **2011**, *97*, 38–43.
17. Mitsch, W.J.; Jørgensen, S.E. *Ecological Engineering and Ecosystem Restoration*; John Wiley: New York, 2004; 410 pp.
18. UNEP-IETC and Copenhagen University. *Handbook of Phytotechnology for Water Quality Improvement and Wetland Management through Modelling Applications*, 2005.

22

Biotechnology: Pest Management

Maurizio G. Paoletti

Benefits of Genetic Engineering in Pest Control

Since 1987, many crops have been genetically modified for features such as resistance to insects, resistance to pathogens (including viruses) and herbicides, and for improved features such as longer-lasting ripening, higher nutritional status, protein content, seedless fruit, and sweetness. Up to 34 new genetically engineered crops have been approved to enter into the market.

In 1998, 27.8 million ha of engineered crops were planted in countries such as the United States, Argentina, Canada, and Australia. The United States alone contains 74% of the modified crop land-planted. Globally, 19.8% of this area has been planted with herbicide-tolerant crops, 7.7% with insect-resistant crops, and 0.3% with insect and HRCs. Five crops—soybean, corn, cotton, canola, and potato—cover the largest acreage of engineered crops.[1,2]

Disease Resistance in Crops

The crops currently on the market that have been engineered for resistance to plant pathogens are listed in Table 1. Disease-resistant engineered crops have some potential advantages because few current pesticides can control bacterial and viral diseases of crops. In addition, these engineered plants help reduce problems from pesticides.

The large-scale cultivation of plants expressing viral and bacterial genes might lead to adverse ecological consequences. The most significant risk is the potential for gene transfer of disease resistance from cultivated crops to weed relatives. For example, it has been postulated that a virus-resistant squash could transfer its newly acquired virus-resistant genes to wild squash *(Cucurbita pepo),* which is native to the southern United States. If the virus- resistant genes spread, newly disease-resistant weed squash could become a hardier, more abundant weed. Moreover, because the United States is the origin for squash, changes in the genetic make-up of wild squash could conceivably lessen its value to squash breeders.

TABLE 1 Plants Genetically-Engineered for Virus Resistance That Have Been Approved for Field Tests in the United States from 1987 to July 1995

Crop	Disease(s)	Research Organization
Alfalfa	Alfalfa mosaic virus, Tobacco mosaic virus (TMV), Cucumber mosaic virus (CMV)	Pioneer Hi-Bred
Barley	Barley yellow dwarf virus (BYDV)	USDA
Beets	Beet necrotic yellow vein virus	Betaseed
Cantelope and/or squash	CMV, papaya ringspot virus (PRV) Zucchini yellow mosaic virus (ZYMV), Watermelon mosaic virus II (WMVII)	Upjohn
	CMV	Harris Moran Seed
	ZYMV	Michigan State University
	ZYMV	Rogers NK Seed
	Soybean mosaic virus (SMV)	Cornell University
	SMV, CMV	New York State Experiment Station
Corn	Maize dwarf mosaic virus (MDMV) Maize chlorotic mottle virus (MCMV), Maize chlorotic dwarf virus (MCDV)	Pioneer Hi-Bred
	MDMV	Northup King
	MDMV	DeKalb
	MDMV	Rogers NK Seed
Cucumbers	CMV	New York State Experiment Station
Lettuce	Tomato spotted wilt virus (TSWV)	Upjohn
Papayas	PRV	University of Hawaii
Peanuts	TSWV	Agracetus
Plum Trees	PRV, plum pox virus	USDA
Potatoes	Potato leaf roll virus (PLRV), Potato virus X (PVX), Potato virus Y (PVY)	Monsanto
	PLRV, PVY, late blight of potatoes	Frito-Lay
	PLRV	Calgene
	PLRV, PRY	University of Idaho
Potatoes	PLRV, PVY	USDA
	PYV	Oregon State University
Soybeans	SMV	Pioneer Hi-Bred
Tobacco	ALMV, tobacco etch virus (TEV), Tobacco vein mottling virus	
	TEV, PVY	University of Florida
	TEV, PVY	North Carolina State University
	TMV	Oklahoma State University
	TEV	USDA
Tomatoes	TMV, tomato mosaic virus (TMV) CMV, tomato yellow leafcurl virus	Monsanto
	TMV, ToMV	Upjohn
	ToMV	Rogers NK Seed
	CMV	PetoSeed
	CMV	Asgrow
	CMV	Harris Moran Seeds
	CMV	New York State Experiment Station
	CMV	USDA

Source: Krimsky and Wrubel[4] and McCullum et al.[5]

Some plant pathologists have also suggested that development of virus-resistant crops could allow viruses to infect new hosts through transencapsidation. This may be especially important for certain viruses, e.g., luteoviruses, where possible heterologus encapsidation of other viral RNAs with the expressed coat protein is known to occur naturally. With other viruses, such as the PRV that infects papaya, the risk of heteroencapsidation is thought to be minimal because the papaya crop itself is infected by very few viruses.

Virus-resistant crops may also lead to the creation of new viruses through an exchange of genetic material or recombination between RNA virus genomes. Recombination between RNA virus genomes requires infection of the same host cell with two or more viruses. Several authors have pointed out that recombination could also occur in genetically engineered plants expressing viral sequences of infection with a single virus, and that large-scale cultivation of such crops could lead to increased possibilities of combinations. It has recently been shown that RNA transcribed from a transgene can recombine with an infecting virus to produce highly virulent new viruses.

A strategy for reduced risk would include: 1) identification of potential hazards; 2) determination of frequency of recombination between homologous, but nonidentical sequences in crops and weeds; and 3) determination of whether or not such recombinants can have selective advantage.

Assessment of Transgenic Virus-Resistant Potatoes in Mexico

An in-depth assessment of potential socioeconomic implications related to the introduction of some genetically modified varieties of virus-resistant potatoes (PVY, PVX, PIRV) in Mexico underscores the importance of this technology. This type of genetic modification could prove especially beneficial to large-scale farmers, but only marginally beneficial to small-scale farmers, because most small farmers use red potato varieties that are not considered suitable for transformation. In addition, 77% of the seeds that small farmers use come from informal sources, not from the seed providers that could sell the new resistant varieties.

The mycoplasma and virus diseases in Mexico are not currently controlled with pesticides, and rank second and third in economic damages. The major pest, the fungus *Phytophtora infestans,* ranks first in economic damages and requires, in some cases, up to 30 fungicide applications. Thus, the interesting new genetically altered varieties of potatoes are of little benefit to crop production for small farmers.

HRCS

Several engineered crops that include herbicide resistance are commercially available; 13 other key crops in the world are ready for field trials (Table 2). In addition, some crops (e.g., corn) are being engineered to contain both herbicide (glyphosate) and biotic insecticide resistance (BT α-endotoxin).

Herbicides adopted for herbicide-resistant crops employ lower doses when compared with atrazine, 2,4-D, and alachlor. However, the resistance of the crop to the target herbicide would, in practice, suggest to the farmer to apply dosages higher than recommended. In addition, costs for this new technology of HRCs are about 2-times higher in corn than the recommended herbicide use and cultivation weed control program.

Integrated pest management (IPM) could benefit from some HRCs, if alternative non-chemical methods can be applied first to control weeds and the target herbicide could be used later, only when and where the economic threshold of weeds is surpassed. Generally, though, the use of herbicide resistant crops will lead to increased use of herbicides and environmental and economic problems. Most HRCs were developed for Western agriculture. For example, in Northern African countries, most crops, such as sorghum, wheat, and canola (oilseed rape), have wild weed relatives, thereby increasing the risk that genes from the herbicide- resistant crop varieties could be transferred to wild weed relatives.

The risk of herbicide-resistant genes from a transgenic crop variety being transferred to weed relatives has been demonstrated for canola (oilseed rape) and sugar beet.

TABLE 2 Herbicide-Resistant Crops (HRCs) Approved for Field Tests in the United States from 1987 to July 1995

Crop	Herbicide	Research Organization
Alfalfa	Glyphosate	Northrup King
Barley	Glufosinate/Bialaphos	USDA
Canola (oilseed rape)	Glufosinate/Bialaphos	University of Idaho
	Glyphosate	Hoechst-Roussel/AgrEvo
		InterMountain Canola
		Monsanto
Corn	Glufosinate/Bialaphos	Hoechst-Roussel/AgrEvo
		ICI
		UpJohn
		Cargill
		DeKalb
		Holdens
		Pioneer Hi-Bred
		Asgrow
		Great Lakes Hybrids
		Ciba-Geigy
		Genetic Enterprises
	Glyphosate	Monsanto
		DeKalb
	Sulfonylurea	Pioneer Hi-Bred
		Du Pont
	Imidazolinone	American Cyanamid
Cotton	Glyphosate	Monsanto
		Dairyland Seeds
		Northrup King
	Bromoxynil	Calgene
		Monsanto
		Rhone Poulenc
	Sulfonylurea	Du Pont
		Delta and Pine Land
	Imidazolinone	Phytogen
Peanuts	Glufosinate/Bialaphos	University of Florida
Potatoes	Bromoxynil	University of Idaho
		USDA
	2,4-D	USDA
	Glyphosate	Monsanto
	Imidazolinone	American Cyanamid
Rice	Glufosinate/Bialaphos	Louisiana State University
Soybeans	Glyphosate	Monsanto
	Glyphosate	UpJohn
		Pioneer Hi-Bred
		Northrup King
		Agri-Pro
	Glufosinate/Bialaphos	UpJohn
	Sulfonylurea	Hoechst/AgrEvo
		Du Pont

(*Continued*)

TABLE 2 (*Continued*) Herbicide-Resistant Crops (HRCs) Approved for Field Tests in the United States from 1987 to July 1995

Crop	Herbicide	Research Organization
Sugar Beets	Glufosinate/Bialaphos	Hoechst-Roussel
	Glyphosate	American Crystal Sugar
Tobacco	Sulfonylurea	American Cyanmid
Tomatoes	Glyphosate	Monsanto
	Glufosinate/Bialaphos	Canners Seed
Wheat	Glufosinate/Bialaphos	AgrEvo

Source: Krimsky and Wrubel[4] McCullum et al.,[5] and Agribusiness[8]

Repeated use of herbicides in the same area creates problems of weed herbicide resistance. For instance, if glyphosate is used with HRCs crops on about 70 million ha, this might accelerate pressure on weeds to evolve herbicide resistant biotypes. Sulfonylureas and imidazolinones in HRCs are particularly prone to rapid evolution of resistant weeds. Extensive adoption of HRCs will increase the hectarage and surface treated, thereby exacerbating the resistance problems and environmental pollution problems.

Bromoxynil has been targeted in herbicide resistant cotton by Calgene and Monsanto (Table 2). This herbicide has been used on winter cereals, cotton, corn, sugarbeets, and onions to control broad leaf weeds. Drift of bromoxynil has been observed to damage nearby grapes, cherries, alfalfa, and roses. In addition, legumious plants can be sensitive to this herbicide, and potatoes can be damaged by it. Herbicide residues above the accepted standards have been detected in soil and groundwater, and as drift fallout. Rodents demonstrate some mutagenic responses to bro- moxynil. Beneficial *Stafilinid* beetles show reduced survival and egg production, even at recommended dosages of bromoxynil. Crustaceans *(Daphnia magna)* have also been severely affected by this herbicide.

Toxicity of Herbicides and HRCs

Toxic effects of herbicides to humans and animals also have been reported. For example, the Basta surfactant (sodium polyoxyethylene alklether sulfate) has been shown to have strong vasodialatative effects in humans and cardio- stimulative effects in rats. Treated mice embryos exhibited specific morphological defects.

Most HRCs have been engineered for glyphosate resistance. Although adverse effects of herbicide-resistant soybeans have not been observed when fed to animals such as cows, chickens, and catfish, genotoxic effects have been demonstrated on other non-target organisms. Earthworms have been shown to be severely injured by the glyphosate herbicide at 2.5–10.1/ha. For example, *Allolobophora caliginosa*, the most common earthworm in European, North American, and New Zealand fields, is killed by this herbicide. In addition, aquatic organisms, including fish, can be severely injured or killed when exposed to glyphosate. The beneficial nematode, *Steinerema feltiae*, a useful biological control organism, is reduced by 19%–30% by the use of glyphosate.

There are also unknown health risks associated with the use of low doses of herbicides. Due to the common research focus on cancer risk, little research has been focused on neurological, immunological, developmental, and reproductive effects of herbicide exposures. Much of this problem is due to the fact that scientists may lack the methodologies and/or the diagnostic tests necessary to properly evaluate the risks caused by exposure to many toxic chemicals, including herbicides.

While industry often stresses the desirable characteristics of their HRCS, environmental and agricultural groups, and other scientists, have indicated the risks. For example, research has shown that the application of glyphosate can increase the level of plant estrogens in the bean, *Vicia faba*. Feeding experiments have shown that cows fed transgenic glyphosate-resistant soybeans had a statistically significant difference in daily milk-fat production as compared to control groups. Some scientists are concerned that the increased milk-fat production by cows fed these transgenic soybeans may be a direct consequence of higher estrogen levels in these transgenic soybeans.

Economic Impacts of HRCs

Some analysts project that switching to bromoxynil for broadleaf weed control in cotton could result in savings of 37 million dollars each year. Furthermore, recent problems with use of glyphosate-resistant cotton in the Mississippi Delta region—crop losses resulting in up to $500,000 of this year's cotton crop—suggest that this technology needs to be further developed before some farmers will reap economic benefits. In addition, a recent study of herbicide- resistant corn suggests that the costs of weed control might be about two times more expensive than normal herbicide and cultivation weed control in corn.

While some scientists suggest that use of HRCs will cause a shift to fewer broad spectrum herbicides, most scientists conclude that the use of HRCs will actually increase herbicide use.

BT for Insect Control

More than 40 BT crystal protein genes have been sequenced, and 14 distinct genes have been identified and classified into six major groups based on amino acids and insecticidal activity. Many crop plants have been engineered with the BT α-endotoxin, including alfalfa, corn, cotton, potatoes, rice, tomatoes, and tobacco (Table 3). The amount of toxic protein expressed in the modified plant is 0.01%–0.02% of the total soluble proteins.

TABLE 3 Transgenic Insect Resistant Crops Containing BT δ-Endotoxins. Approved Field Tests in United States from 1987 to July 1995

Crop	Research Organization
Alfalfa	Mycogen
Apples	Dry Creek
	University of California
Corn	Asgrow
	Cargill
	Ciba-Geigy
	Dow
	Genetic Enterprises
	Holdens
	Hunt-Wesson
	Monsanto
	Mycogen
	NC+Hybrids
	Nortrup King
	Pioneer Hi-Bred
	Rogers NK Seed
Cotton	Calgene
	Delta and Pineland
	Jacob Hartz
	Monsanto
	Mycogen
	Northrup King

(*Continued*)

TABLE 3 (*Continued*) Transgenic Insect Resistant Crops Containing BT δ-Endotoxins. Approved Field Tests in United States from 1987 to July 1995

Crop	Research Organization
Cranberry	University of Wisconsin
Eggplant	Rutgers University
Poplar	University of Wisconsin
Potatoes	USDA
	Calgene
	Frito-Lay
	Michigan State University
	Monsanto
	Montana State University
	New Mexico State University
	University of Idaho
Rice	Louisiana State University
Spruce	University of Wisconsin
Tobacco	Auburn University
	Calgene
	Ciba-Geigy
	EPA
	Mycogen
	North Carolina State University
	Roham and Haas
Tomatoes	Campbell
	EPA
	Monsanto
	Ohio State University
	PetoSeeds
	Rogers NK Seeds
Walnuts	University of California, Davis
	USDA

Source: Krimsky and Wrubel[4] and Agribusiness.[7]

Some trials with corn demonstrate a high level of efficacy in controlling corn borers. Corn engineered with BT endotoxin has the potential to reduce corn borer damage by 5%–15% over 28 million ha in the US, with a potential economic benefit of $50 million annually. Some suggest that corn engineered with BT toxin will increase yields by 7% over similar varieties. However, it is too early to tell if all these benefits will be realized consistently. Potential negative environmental effects also exist because the pollen of engineered plants contains BT, which is toxic to bees, beneficial predators, and endangered butterflies like the Karka Blue and Monarch Butterflies.

Cotton was the first crop plant engineered with the BT α-endotoxin. Caterpillar pests, including the cotton boll- worm and budworm, cost U.S. farmers about $171 million/ yr as measured in yield losses and insecticide costs. Benedict et al.[3] predict that the widespread use of BT cotton could reduce insecticide use and thereby reduce costs by as much as 50% to 90%, saving farmers $86 to $186 million/yr.

The development of insect resistance to transgenic crop varieties is one highly possible risk associated with the use of BT D-endotoxin in genetically engineered crop varieties. Resistance to BT has already been demonstrated in the cotton budworm and bollworm. If BT- engineered plants become resistant, a key insecticide that has been utilized successfully in IPM programs could be lost. Therefore, proper resistance management strategies with use of this new technology are imperative. Another potential risk is that the BT a-endotoxin could be harmful to non-target organisms. For example, it is not clear what potential effect the BT D-endotoxin residues that are incorporated into soils will have against an array of non-target useful invertebrates living in the rural landscape. It has also been demonstrated that predators, such as the lacewing larvae (*Crysoperla carnea*) that feed on corn borers (*Ostrinia nubilalis*), grown on engineered BT corn have consistently higher mortality rates when compared to specimens fed with non-engineered corn borers. In addition, the treated larvae need three more days to reach adulthood than lacewings fed on prey from non-BT corn.

Discussion

Both pesticides and biotechnology have definite advantages in reducing crop losses to pests. At present, pesticides are used more widely than biotechnology, and thus are playing a greater role in protecting world food supplies. In terms of environmental and public health impacts, pesticides probably have a greater negative impact at present because of this more widespread use.

Genetically engineered crops for resistance to insect pests and plant pathogens could, in most cases, be environmentally beneficial, because these more resistant crops could allow a reduction in the use of hazardous insecticides and fungicides in crop production. In time, there may also be economic benefits to farmers who use genetically engineered crops; this will depend, however, on the prices charged by the biotechnology firms for these modified, transgenic crops.

There are, however, some environmental problems associated with the use of genetically engineered crops in agriculture. For example, adding BT to crops like corn for insect control can result in any of the following negative environmental consequences: 1) development of resistance to BT by pests species in corn and other crops; 2) health risks from exposure to the BT toxin to humans in their food and to livestock in feed; 3) the toxicity of the pollen from the BT-treated corn to honey bees, beneficial natural enemies, and endangered species of insects that feed on the modified corn plants or come into contact with the drifting pollen; engineered plant residues incorporated into soil can produce undesirable effects on soil micros and mesofauna.

A major environmental and economic concern associated with genetically engineered crops is the development of HRCs. Although in rare instances HRCs may result in a beneficial reduction of toxic herbicide use, it is more likely that the use of HRCs will increase herbicide use and environmental pollution. In addition, farmers will suffer because of the high costs of employing HRCs—in some instances, weed control with HRCs may increase weed control costs for the farmer threefold.

More than 40% of the research by biotechnology firms is focused on the development of HRCs. This is not surprising, because most of the biotechnology firms are also chemical companies who stand to profit if herbicide resistance in crops result in greater pesticide sales. Theoretically, the acceptance and use of engineered plants in sustainable and integrated agriculture should consistently reduce current use of pesticides, but this is not the current trend. In addition, most products and new technologies are designed for Western agriculture systems, not for poor or developing countries. For instance, if terminator genes enter into the seed market, there will be no possibility of traditional and small farmers using their plants to produce their seeds. Thus, genetic engineering could promote improvements for the environment; however, the current products—especially the herbicide-resistant plants and the BT-resistant crops—do have serious environmental impacts, similar to the consequences of pesticide use.

References

1. James, C. Global Review of Commercialized Transgenic Crops: 1998. In *ISAAA Briefs*; Cornell University: Ithaca, New York, 1998; 8–1998.
2. Moff, A.S. Toting up the early harvest of transgenic plants. Science **1998**, *282,* 2176–2178.

3. Benedict, J.H.; Ring, D.R.; Sachs, E.S.; Altman, D.W.; DeSpain, R.R.; Stone, T.B.; Sims, J.R. Influence of Transgenic BT Cottons on Tobacco Budworm and Bollworm Behavior Survival, and Plant Injury. In *Proceedings Beltwide Cotton Council*; Herber, J., Richter, D.A., Eds.; National Cotton Council: Memphis, Tennessee, 1992; 891–895.
4. Krimsky, S.; Wrubel, R.P. *Agricultural Biotechnology and the Environment*; University of Illinois Press: Urbana, Illinois, 1996.
5. McCullum, C.; Pimentel, D.; Paoletti, M.G. Genetic Engineering in Agriculture and the Environment: Risks and Benefits. In *Biotechnology and Safety Assessment*; Thomas, J.A., Ed.; Taylor and Francis: Washington, D.C., 1998; 177–217.
6. Agribusiness. The Gene Exchange, Fall 1997; http://www.uc- susa.org/Gene/F97.agribusiness.html (accessed July 5, 2001).
7. Agribusiness. The Gene Exchange, Winter 1996; http://www. ucsusa.org/Gene/W96.agribusiness.html (accessed July 5, 2001).

23

Plant Pathogens (Fungi): Biological Control

Timothy Paulitz

Introduction

In the classical definition of biological control, certain fungi, termed biocontrol agents (BCAs), can reduce the amount of inoculum or disease-producing activity of a plant pathogen, usually another fungus.[1] The net result is a reduction of plant disease and crop loss. This section will cover mechanisms of how these fungi antagonize the pathogen, what part of the pathogen life cycle can be targeted, how the BCAs can be applied, and examples of commercially available products. Within the past 20 years, there has been a tremendous increase in interest and research on the subject, spurred by a search for more environmentally benign methods of disease control. But fungal BCAs have limitations that have restricted the number of products that are currently on the market. Table 1 shows some products used against soilborne pathogens on the market as of January 1999.

Mechanisms of Biological Control by Fungi

The strategy behind managing pathogens is to target or interrupt part of the pathogen life cycle.[1,2] Like any microbe, pathogens start from inoculum in the environment, which can be spores, mycelia, or other dormant survival structures. These germinate on the plant surface, penetrate and infect the plant, and reproduce and sporulate on the plant to produce new inoculum. Many pathogens can also grow saprophytically on dead organic matter and plant debris. In biological control, the pathogen can be targeted in three ways.[1,2] First, the inoculum of the pathogen can be reduced or destroyed. This is most effective for soilborne pathogens, where the inoculum is dormant in the soil and the monocyclic disease is determined by the initial inoculum present in the field. BCAs can also interfere with inoculum formation by pathogens growing saprophytically on organic matter and plant debris. However, this strategy is not very effective for foliar polycyclic diseases, where inoculum comes from outside the field and the initial inoculum has little effect on the final outcome of the disease. Another strategy is one of protection, where a population of the BCA is established on the infection site of the plant before the pathogen attacks, thus preventing the pathogen's entry. These infection sites can be on seeds, bulbs, roots, leaves, fruit, flowers, or wounds. Finally, nonpathogenic or avirulent fungi can stimulate the plant to a higher level of resistance to a later-attacking pathogen, a concept termed induced resistance.

TABLE 1 Some Commercial Biocontrol Products for Use against Soilbome Crop Diseases

Biocontrol Fungus	Trade Name	Target Pathogen/Disease	Crop	Manufacturer
Ampelomyces quisqualis M-10	AQ10 Biofungicide	Powdery mildew	Cucurbits, grapes, ornamentals, strawberries, tomatoes	Ecogen Inc., Langhorne, Pennsylvania
Candida oleophila I-182	Aspire	Botrytis, Penicillium	Citrus, pome fruit	Ecogen Inc., Langhorne, Pennsylvania
Fusarium oxysporum (nonpathogenic)	Biofox C	Fusarium oxysporum	Basil, carnation, cyclamen, tomato	S.I.A.P.A., Galliera, Bologna, Italy
Trichoderma harzianum and *T. polysporum*	Binab T	Wilt and root rot pathogens, wood decay pathogens	Fruit, flowers, ornamentals, turf, vegetables	Bio-innovation, Algaras, Sweden
Coniothyrium minitans	Contans	Sclerotinia sclerotiorum and S. minor	Canola, sunflower, peanut, soybean, lettuce, bean, tomato	Prophyta Biologischer Pflanzenschutz, Malchow/Poel, Germany
Fusarium Oxysporum (nonpathogenic)	Fusaclean	Fusarium oxysporum	Basil, carnation, cyclamen, gerbera, tomato	Natural Plant Protection, Nogueres, France
Pythium oligandrum	Polygandron	Pythium ultimum	Sugar beet	Plant Protection Institute, Bratislavsk, Slovak Republic
Trichoderma harzianum and *T. viride*	Promote	Pythium, Rhizoctonia, Fusarium	Greenhouse, nursery transplants, seedlings	JH Biotech, Ventura, California
Trichoderma harzianum	RootShield, Bio-Trek T-22G, Planter Box	Pythium, Rhizoctonia, Fusarium, Sclerotinia homeocarpa	Trees, shrubs, transplants, ornamentals, cabbage, tomato, cucumber, bean, corn, cotton, potato, soybean, turf	Bioworks, Geneva, New York
Phlebia gigantea	Rotstop	Heterobasidium annosum	Trees	Kemira Agro Oy, Helsinki, Finland
Gliocladium virens GL-21	SoilGard (formerly GlioGard)	Damping-off and root pathogens, Pythium, Rhizoctonia	Ornamentals and food crops grown in greenhouses, nurseries, homes, interiorscapes	Thermo Triology, Columbia, Maryland
Trichoderma harzianum	Trichodex	Botrytis cinerea, Colletotrichum, Monilinia laxa, Plasmopara viticola, Rhizopus stolonifer, Sclerotinia sclerotiorum	Cucumber, grape, nectarine, soybean, strawberry, sunflower, tomato	Makhteshim Chemical Works, Beer Sheva, Israel
Trichoderma Harzianum and *T. viride*	Trichopel, Trichoject	Armillaria, Botryosphaeria, Fusarium, Nectria, Phytophthora, Pythium, Rhizoctonia		Agrimm Technologies, Christchurch, New Zealand

Source: Information provided by the U.S. Department of Agriculture, Agriculture Research Service, Beltsville, Maryland, and was compiled by D. Fravel (http://www.barc.usda.gov/psi/bpdl/bodlpood/bioprod.htm).

The most direct way a fungus can attack a fungal pathogen is by mycoparasitism, where the BCA uses the pathogen as a source of food.[3] The hyphae of the mycoparasite contact, penetrate, and colonize the hyphae, spores, or survival structures of the host fungus. Many of these mycoparasites produce enzymes that degrade the cell walls of the fungal host, including β-1–3 glucanase and chitinase. Most mycoparasites are necrotrophic and eventually kill their fungal host. Much of this research has focused on reducing the inoculum of soilborne pathogens. Classic examples include *Trichoderma* and *Gliocladium* spp. parasitizing *Rhizoctonia solani* and *Pythium* spp., which cause seed, seedling, and root rots.[4–6] *Pythium* spp. such as *Pythium oligandrum* and *P. nunn* parasitize pathogenic species of *Pythium. Coniothyrium minitans* and *Sporodesmium sclerotivorum* parasitize sclerotia of *Sclerotinia* spp., such as the white mold pathogen *S. sclerotiorum*, which attacks hundreds of plant species (57). *Ampelomyces quisqualis* parasitizes cleistothecia of powdery mildews. Major limitations to this strategy are that mycoparasites are slow acting and large amounts of mycoparasite inoculum must be added to the soil to ensure it will encounter the propagules of the pathogen. However, a promising strategy demonstrated with *S. sclerotivorum* is to render a soil suppressive to the pathogen by an inoculative release at a lower inoculum density, and allowing the mycoparasite to build up over successive seasons, using the pathogen as a food source. This is similar to the classic predator–prey relationship found in insect biocontrol.

Some fungi can produce antibiotic compounds that are toxic to other microbes, including plant pathogens. *Trichoderma* spp. produce volatile and nonvolatile antifungal compounds, including peptiabols, pyrones, and terpenoid antibiotics.[4] *Gliocladium virens* produces glioviren and gliotoxin that inhibit *R. solani* and *Pythium ultimum*. This mechanism is most effective when the BCA can grow to high populations and has an energy source to produce the antibiotic. An example would be *Trichoderma* or *Gliocladium* spp. applied to seeds or where a food base is added to the inoculum.[4,5]

Plant pathogens require carbon, nitrogen, iron, and other nutrients to grow. Many spores have an exogenous requirement for these nutrients, supplied by the plant rhizosphere or phyllosphere, in order to germinate. BCAs can compete with the pathogen for these limiting nutrients. For example, nonpathogenic species of *Fusarium oxysporum* can compete with pathogenic forma speciales for these limiting nutrients, resulting in control of wilt diseases. Competition by yeasts or hyphal fungi may protect flowers and foliage against nectotrophic pathogens such as *Botrytis cinerea* by colonizing the senescent tissue or nutrient-rich flower petals.[8,9] This mechanism, although difficult to prove experimentally, is probably one of the primary ways BCAs can protect a plant surface through preemptive exclusion of the pathogen.

Fungal biocontrol agents can also affect the pathogen by acting indirectly on the plant to make it more resistant to pathogen attack. Nonpathogenic microbes can induce a systemic resistance in plants (79). When the plant recognizes the inducing BCA, a signal is transduced systemically to the entire plant, bringing the defenses to a "high state of alert," so that a subsequent challenge by a pathogen is reduced. Nonpathogenic isolates of *F. oxysporum* induce a defense reaction against pathogenic isolates of *F. oxysporum*. This mechanism has several advantages. Once induced, the resistance is systemic, the entire plant becomes more resistant, and high populations of the BCA do not need to be maintained. It can also protect parts of the plant that cannot be protected directly by the BCA, including new growth of shoots and roots. However, more research is needed to investigate the applicability of this technology under greenhouse and field conditions.

Application of Fungal Biocontrol Agents

How are fungal BCAs applied? Most are applied in an inundative strategy in large amounts to build up the population of the BCA high enough to overwhelm and have an effect on the pathogen.[1,5] Most are also targeted toward soilborne pathogens. However, one limitation of this strategy is the large amount of inoculum that must be applied and the high cost of production of spores, conidia, biomass, or chlamydospores.[5,9] Another problem is the erratic performance of many biocontrol agents under field conditions, due to unfavorable environmental conditions for the BCA, and the problem of establishing the BCA in a niche already occupied by competing microflora. Therefore most of the commercially available products have targeted applications that avoid these problems. For example, the greenhouse and nursery markets are prime targets because of the controlled environmental conditions and the high economic value of the crops. Another method is to use a protective strategy and apply the BCA directly to the infection court when it is small. For example, high populations of *Trichoderma* or *Gliocladium* conidia can be coated onto seeds to protect against damping-off pathogens such as *R. solani* and *P. ultimum*. Transplant cuttings and bulbs can be treated with liquid suspensions of products before planting in the greenhouse or field. Products such as formulations of *G. virens* can be mixed directly into the soil or soilless mixes in the greenhouse. Granules of *Trichoderma* spp. can be added to seed furrows, mixed with seeds in a planter box, or added to a commercial seed slurry.[4]

Since many pathogens gain access to plants through wounds, biocontrol agents can also be applied to transplant or pruning wounds. A classic example, and one of the first commercially used fungal biocontrol agents, is the application of *Phlebia* (= *Peniophora*) *gigantea* to cut pine stumps to prevent the stumps from being colonized by the pathogen *Heterobasidium annosum*. The pathogen can spread from these stumps to the entire plantation via the root system.

Postharvest pathogens are weak pathogens that require wounds on fruit to gain access. Yeast-like organisms such as *Candida* spp. can be applied to fruit during processing to exclude rot pathogens such as *Penicillium* spp. from colonizing wounds.[8] However, these applications require more stringent testing for registration, since they are applied directly to a food product. Competition is preferable to antibiosis for this application, since antifungal compounds would also have to be tested for animal and human toxicity.

Roots are a difficult infection court to protect, since the susceptible tips are constantly growing and moving through space encountering new inoculum. One strategy is to treat the entire rooting medium in the greenhouse or nursery. Another approach is to use a fungus that can colonize the root system from a seed or furrow treatment and protect the expanding root surface. This characteristic, called rhizosphere competence, has been demonstrated in some strains of *Trichoderma* spp.

Foliar applications of fungi are the least common, although this is the most common method of fungicide application. One example is *Pseudozyma* (= *Sporothrix*) *flocculosa,* a yeast-like fungus which is being developed for control of powdery mildews on greenhouse roses and cucumbers in Canada. *Trichoderma* and *Gliocladium* spp. can be applied to foliage and flowers and can prevent infection by necrotrophic fungi such as *B. cinerea* and *S. sclerotiorum*.

Future of Biological Control by Fungi

In conclusion, the full potential of controlling plant diseases with fungi still has not been realized. Only a small number of products are on the market, but this is a vast improvement compared to only five years ago. There are still many economic constraints in terms of the cost of development and registration of products and the low cost production of organisms in liquid fermentation or solid on substrates.[5,9] Like chemicals, the risks of fungal BCAs need to be addressed, including the displacement of nontarget microbes, allergenicity to humans and other animals, and toxigenicity and pathogenicity to nontarget organisms.[10] However, there are no existing chemical controls for many diseases, because of deregistration of pesticides, pathogen resistance to pesticides, and environmental concerns. These diseases may be the niches for fungal biocontrol agents. It is unlikely that biological control will succeed alone, but it needs to be integrated with other disease management strategies, including cultural control and genetic disease resistance.

References

1. Cook, R.J.; Baker, K.F. *The Nature and Practice of Biological Control of Plant Pathogens*; American Phytopathologi-cal Society Press: St. Paul, MN, 1983; 539.
2. *Principles and Practice of Managing Soilborne Plant Pathogens*; Hall, R., Ed.; American Phytopathological Society Press: St. Paul, MN, 1996; 442.

3. *Fungi in Biological Control Systems*; Burge, M.N., Ed.; Manchester University Press: Manchester, 1988; 269.
4. *Trichoderma and Gliocladium*; Harman, G.E., Kubicek, C.R., Eds.; Taylor and Francis: London, 1998; 1 and 2, 278–393.
5. *Pest Management: Biologically Based Technologies*; Vaughn, J.L., Lumsden, R.D., Eds.; American Chemical Society: Washington, DC, 1993; 435.
6. *Biological Control of Plant Diseases. Progress and Challenges for the Future*; Papavizas, G.C., Cook, R.J., Tjamos, E.C., Eds.; Plenum Press: New York, 1992; 462.
7. Whipps, J.M. Biological control of soil-borne plant pathogens. Advances in Botanical Res. **1997**, *26*, 1–134.
8. *Plant-Microbe Interactions and Biological Control*; Kuykendallm, L.D., Boland, G.L., Eds.; Marcel Dekker, Inc.: New York, 1998; 442.
9. *Integrated Pest and Disease Management in Greenhouse Crops;* Lodovica Gullino, M., van Lenteren, J.C., Elad, Y., Albajes, R., Eds.; Kluwer Academic Publishers: Dordrecht, the Netherlands, 1999; 545.
10. *Biological Control: Benefits and Risks*; Hokkanen, H.M.T., Lynch, J.M., Eds.; Cambridge University Press: Cambridge, 1995; 304.

24

Plant Pathogens (Viruses): Biological Control

Hei-Ti Hsu

Economic Loss

Damage to crop plants due to virus and viroid infections is difficult to assess. The actual figures for global crop loss are not available. Plant disease losses are estimated at $60 billion annually. Losses due to virus and viroids have been considered second to those caused by fungi. Unlike diseases caused by fungi, bacteria, and nematodes, where control measures using chemical, biological, and integrated pest management approaches have been effective, diseases caused by viruses or viroids are far more difficult to manage.

Economic crop loss resulting from virus and viroid disease is due to the reduced growth and vigor of infected plants which, in turn, causes a reduction in yield. In some instances, a virus infection may kill a plant. Apart from yield reduction, the quality and market value of commercial end products may be affected. There are also costs of attempting to maintain crop health such as vector control, production of pathogen-free propagation materials, and quarantine and eradication programs. In addition, resources are being diverted to research, extension, and education as well as toward breeding for resistance to virus or viroid infection.

World Impact

No single country is exempt from crop losses. Production of food, fiber, and horticultural crops are seriously affected worldwide by virus or viroid infection of plants.[1] This is even more so in developing countries that depend on one or a few major crops; for example, *Cassava mosaic virus* in cassava plants in Kenya, *Citrus tristeza virus* in citrus trees in Africa and South America, and *Cacao swollen shoot virus* in cacao trees in Ghana. Recently, *Papaya ringspot virus* (PRV) infection has affected every region where papaya plants are grown. The virus induces a lethal disease in papaya. The widespread aphid-transmitted PRV has changed the way papayas are grown in many parts of the world. Normally, papayas are produced annually for a number of years over the life of the papaya plant. For proper management of the disease due to PRV infection, papaya has now become an annual crop in which healthy seedlings are planted each year. Even so, productivity is still below the average yield obtained before PRV became a problem. Viroids infect a limited number of crops when compared with viruses. However, they can cause severe problems in specific crops, for instance, cadang-cadang disease of coconuts, potato spindle tuber disease, and chrysanthemum stunt disease.

Control Measures

No direct chemical control means are available to combat virus infections in plants. Control of viral diseases is achieved primarily by sanitary practices that involve reducing sources of inoculum from outside, preventing spreading within the crop, and limiting the population of insects, mites, nematodes, and fungi that may serve as vectors for many plant viruses.[1] Virus disease testing programs are now common in many parts of the world where the economic importance of growing virus-free plants is recognized. Although seeds and seedlings certified as virus-free are more expensive than those that have not been tested for certain viruses, testing provides assurance of virus-free production materials. Early detection of virus in a field and removal of the infected plants minimizes spread of the virus.

Plants may be protected from development of severe disease symptoms by first introducing a mild strain of virus into a healthy plant. A plant systemically infected with a mild strain of virus is protected from infection by a severe strain of the same virus. This phenomenon in called “cross protection” and has been observed for many plant viruses.[2] It is also observed to occur between viroids or plant virus satellites. In practice, cross protection is of great interest since it has been utilized to protect plants against severe virus strains (*Citrus tristeza virus, Papaya ringspot virus, Zucchini yellow mosaic virus, Tomato mosaic virus,* etc.), in the field.

Another approach toward controlling plant virus diseases is to develop resistant or tolerant plants.[3] Historically, long-term manipulation of crop plants through breeding has produced many valuable commercial varieties resistant to plant viruses. Breeding plants resistant to vectors may also offer control of the virus they transmit. Conventional breeding of crossing and back crossing commercial varieties with plants bearing virus resistance traits takes years to develop. In order for a new variety to be commercially acceptable, undesirable traits from the resistant parent breeding line must be selected out. The process is labor intensive and time consuming. Advances in science have allowed new technology to precisely manipulate resistance genes at the molecular level.[4] Biotechnology represents the fastest growing area of biological research. The application of biotechnology in breeding for resistance to virus infection is a major area of research. Successful control of viral disease through resistance breeding will undoubtedly reduce the use of synthetic pesticides for vector control.[5]

Introducing virus resistance and vector resistance into a cultivar by gene transfer technology (genetic engineering) has been successful in combating plant viruses.[6] The technology has several major advantages over conventional cross breeding. It is a relatively fast procedure. Desirable genes can be introduced without disturbing the balanced genome of target plants. Furthermore, there is no restriction on the source of the transgenes allowing the use of genes from other plant species or even from outside the plant kingdom (Table 1).[1,7]

Several approaches for producing transgenic virus- resistant plants have been explored. Among these, plants expressing virus coat protein genes, parts of other viral genes, or virus satellite ribonucleic acids (RNAs) have been shown to offer the best control.[2,8,9] Plants expressing antisense viral RNAs, ribozymes, pathogen-related proteins, or virus-specific antibody genes may also confer resistance to virus infection. Control of virus vectors by introducing insect toxins such as trypsin inhibitor, lectin, and *Bacillus thuringiensis* (Bt) toxin genes into plants would undoubtedly contribute toward achieving the goal of controlling plant virus diseases.

Prospects

Use of resistant cultivars is considered the best approach to combat virus infection in plants. Biotechnology, no doubt, will play a significant role in the economic growth of many countries. Molecular breeding, however, will not replace but complement the efforts of conventional cross breeding. Much attention has been given to engineering resistance to plant viruses. Recently, genetic engineering of crop plants has been closely scrutinized and criticized due to increasing public concerns regarding human health and environmental impact. Careful assessment of the benefits and potential risks involving the release of

TABLE 1 Genes That Contribute or May Contribute Toward Control of Virus Diseases in Plants

Virus-Derived Gene Sequences
Coat proteins
Replicases
Movement proteins
Polyprotein proteases
Sense RNAs
Antisense RNAs
Plant Host-Derived Transgenes
Pathogen-related proteins
Anti-viral proteins
Proteinase inhibitors
Natural resistance genes
Lectins
Other Transgenes and Sequences
Satellite RNAs
Virus-specific antibodies
Interferon-induced mammalian oligoadenylate synthetase
Insect toxins
Anti-viral ribozymes (catalytic RNA)

Source: Khetarpal et al.[1]

genetically modified plants into the environment and their consumption is necessary before these crops become widely accepted by the public.[10,11]

References

1. Khetarpal, R.K., Koganezawa, H., Hadidi, A., Eds. *Plant Virus Disease Control*; APS Press: St. Paul, MN, 1998, 1–684.
2. Beachy, R.N. Coat-protein-mediated resistance to tobacco mosaic virus: discovery mechanisms and exploitation. Phil. Trans. R. Soc. Lond. B. **1999**, *354*, 659–664.
3. Salomon, R. The evolutionary advantage of breeding for tolerance over resistance against viral plant disease. Israel J. Plant Sci. **1999**, *47*, 135–139.
4. Kawchuk, L.M.; Prufer, D. Molecular strategies for engineering resistance to potato viruses. Can. J. Plant Pathol. **1999**, *21*, 231–247.
5. Barker, I.; Henry, C.M.; Thomas, M.R.; Stratford, R. Potential Benefits of the Transgenic Control of Plant Viruses in the United Kingdom. In *Plant Virology Protocols: From Virus Isolation to Transgenic Resistance;* Foster, G.D., Taylor, S.C., Eds.; Humana Press, Inc.: Totowa, NJ, 1998; 81, 557–566.
6. Dempsey, D.A.; Silva, H.; Klessig, D.F. Engineering disease and pest resistance in plants. Trends Microbiol. **1998**, *6*, 54–61.
7. Gutierrez-Campos, R.; Torres-Acosta, J.A.; Saucedo-Arias, L.J.; Gomez-Lim, M.A. The use of cysteine proteinase inhibitors to engineer resistance against potyviruses in transgenic tobacco plants. Nat. Biotechnol. **1999**, *17*, 1223–1226.
8. Maiti, I.B.; Von Lanken, C.; Hong, Y.; Dey, N.; Hunt, A.G. Expression of multiple virus-derived resistance determinants in transgenic plants does not lead to additive resistance properties. J. Plant Biochem. Biotech. **1999**, 8, 67–73.

9. Prins, M.; Goldbach, R. RNA-mediated virus resistance in transgenic plants. Arch. Virol. **1996**, *141*, 2259–2276.
10. Hammond, J.; Lecoq, H.; Raccah, B. Epidemiological risks from mixed virus infections and transgenic plants expressing viral genes. Adv. Virus Res. **1999**, *54*, 189–314.
11. Kaniewski, W.K.; Thomas, P.E. Field testing for virus resistance and agronomic performance in transgenic plants. Mol. Biotechnol. **1999**, *12*, 101–115.

25

Stored-Product Pests: Biological Control

Lise Stengård
Hansen

Introduction

Stored grain and other durables of plant origin, raw and processed, can become infested by insect and mite pests. As the environment in storage facilities in general is conducive to rapid pest development, this often results in substantial quantitative and qualitative losses. Many stored-product pests are internal feeders developing inside whole-grain cereals and legumes in storage. Another group of pests are external feeders developing on broken kernels, flour, etc., and are mainly found in processing facilities, stores of processed cereals and bulk grain, and packaged foodstuffs (Table 1).

TABLE 1 Examples of Pests and Natural Enemies in Stored Products

Pest Type	Commodity	Examples of Pest Species	Examples of Natural Enemies
Internal feeders or primary pests	Whole-grain cereals	*Sitophilus spp.* (C) *Rhyzopertha dominica* (C) *Sitotroga cerealella* (L) *Prostephanus truncatus* (C)	*Lariophagus distinguendus* (l) *Theocolax elegans* (l) *Anisopteromalus calandrae* (l) *Teretrius nigrescens* (p)
	Whole legumes	*Bruchus* spp. (C) *Callosobruchus* spp. (C)	*Dinarmus* spp. (1)
External feeders or secondary pests	Broken kernels, flour, milled rice, dried fruit, spices, nuts	*Tribolium* spp. (C) *Cryptolestes* spp. (C) *Oryzaephilus* spp. (C)	*Trichogramma spp.* (e) *Holepyris silvanidis* (l) *Cephalonomia spp.* (1) *Xylocoris flavipes* (p)
		Ephestia kuehniella (L) *Cadra cautella* (L) *Plodia interpunctella* (L)	*Trichogramma spp.* (e) *Venturia canescens* (1) *Habrobracon hebetor* (l) *Xylocoris flavipes* (p)
		Storage mites (Acarina)	*Cheyletus eruditus* (p)

Source: Schöller,[2] Subramanyam and Hagstrum.[3,4]
C, Coleoptera; e, egg parasitoid; L, Lepidoptera; l, larval parasitoid; p, predator.

Beneficial insects are often found in storage facilities and may prevent or delay pest development. The internal feeders are attacked by a range of parasitoids specialized in detecting infested kernels and placing their progeny on the larvae within the kernel. The external feeders are more freely exposed to the activities of natural enemies and are attacked by both predators and parasitoids (Table 1).

Although the concept is not new, biological control of stored-product pests is still not widely used. The following description gives some examples of the recent development within biological control of stored-product pests using beneficial insects and mites.

For general reviews on stored-product pest species and biological control of stored-product pests.[1–4]

Factors that Promote Successful Biocontrol in Storage

Storage facilities are ideally suited for development and reproduction of pests. However, these factors are also conducive to successful biocontrol: 1) Climatic conditions are relatively stable and, at least during part of the year, favorable for insect development; 2) the storage structure provides protection from temperature extremes and precipitation as well as a physical barrier to dispersal; and 3) in many storage situations, time is not a limiting factor and thus often sufficient for natural enemies to establish and exert their control. These factors can be manipulated to be more favorable to the natural enemies, e.g., grain temperature can be reduced by aeration to a level that offers a greater advantage to the activities of the beneficial insect than to the pest.[5]

"Classical" Biological Control

In "classical" biological control, a natural enemy is imported and released for establishment to control an introduced (exotic) pest. An example of this among stored-product pests is the larger grain borer, *Prostephanus truncatus,* which was accidentally introduced from Central America into Africa, where it has become a serious pest of many products, e.g., stored maize and cassava chips. A predator, *Teretrius nigrescens,* was introduced from Central America and, after large scale releases, established successfully in many locations in Africa. *T. nigrescens* has been credited with reducing losses of stored maize to *P. truncatus.* However, despite the presence of the predator, *P. truncatus* densities often reach outbreak levels in many countries. An analysis in Holst and Meikle[6] led to the conclusion that *T. nigrescens* alone is unable to exert control to an acceptable level inside a store due to the predator's intraspecific density-dependence and low growth rate compared with its prey. In this case, biocontrol must be supplemented with other integrated control measures.[7]

Biocontrol of Internally Feeding Pests

The larvae of internally feeding insect pests develop within whole-grain cereals and legumes. These species are important primary pests in stores of bulk grain. Several species of parasitoids are specialized to live on these pests. The adult parasitoids enter the bulk of grain and find infested kernels, probably by means of acoustic or olfactory cues. They then drill into the kernel, paralyze the host, and deposit an egg from which the parasitoid larva emerges. After consuming the host, the parasitoid exits the kernel as an adult. The ability of the parasitoid to locate its host within grain varies among species: *Anisopteromalus calandrae* is primarily active at the grain surface whereas *Theocolax elegans* and *Lariophagus distinguendus* are able to find infested kernels down to a depth of 2.2 and 4 m, respectively.[8,9] Great differences in life history parameters occur among different parasitoid strains.[8] These factors as well as grain characteristics such as kernel size and grain variety all affect the ability of the natural enemy to exert effective control of the pest.

A constraint against biocontrol in stored products is a reluctance to increase the amount of insects by releasing natural enemies. However, these minute parasitoids (<2 mm) can easily be removed by grain cleaning procedures. In a study using *T. elegans* against *Rhyzopertha dominica* in wheat, the number of insect fragments in the resulting flour was reduced by 89%.[10]

Biocontrol of Externally Feeding Pests

Both beetles and moths are represented among external pests that feed on broken kernels and debris as well as flour. They are important pests in grain stores, in cereal processing facilities such as flourmills, and in warehouses storing cereal products.

Many species of parasitoids, particularly Trichogram-matidae, attack the egg stage of external pests. The impact of these species is increased by their host feeding behavior, which can account for half of the mortality of host eggs.[11] These egg parasitoids are able to parasitize a wide range of species, but in nature they generally show affinity to a specific habitat; it is thus important to select strains that are adapted to the stored-product environment. *Trichogramma* species have been successfully released against pyralid moth pests in experimental peanut storages in the U.S.A.[12] and in wholesale stores and industrial bakeries in Germany.[13] These egg parasitoids do not establish within the premises and must be released on a regular basis.

Larvae of moth pests are attacked by both ectoparasitoids, e.g., *Habrobracon hebetor* and endoparasitoids, e.g., *Venturia canescens*. Both species are cosmopolitan, often occurring together in flourmills and both species show potential for biocontrol of moth pests.[12] However, *H. hebetor* may affect populations of *V. canescens* negatively by feeding on hosts parasitized by this species.

The predatory bug *Xylocoris flavipes* is a generalist living on eggs and larvae of a wide range of beetles and moths. Almost 30 species of stored-product pests have been reported as prey of *X. flavipes*.[12] In residues in empty maize stores in the U.S.A., many of these species occur together in the same store. A single introduction of *X. flavipes* led to population reductions of 70%-100% of externally feeding beetle pests, whereas internal feeders and late instar moth larvae were less affected. It was suggested that releases of predatory bugs combined with parasitoids for the moths and internal feeders might eliminate or greatly reduce residual pests before the next storage season.[12]

This strategy of introducing predators in empty stores was widely practiced to control storage mites (Acarina) in the Czech Republic. The predatory mite *Cheyletus eruditus* led to reductions in storage mite populations of 88%, compared with the 18% reduction obtained with a pesticide treatment.[14]

Conclusions

The above-mentioned examples show the great potential of biological control of stored-product pests. This research field benefits from extensive international collaboration and applicability, as the pest species as well as their natural enemies have a cosmopolitan distribution as a result of international grain trade. A great amount of faunistic surveys and laboratory research on the natural enemies has been carried out, but very few field trials have been conducted. Application of natural enemies to control pests in bulk grain as well as in empty grain stores prior to introduction of newly harvested grain is considered to hold potential in the near future. However, widespread use depends on crucial experience to be obtained from field trails. The next step is ensuring reliable supplies of natural enemies, designing introduction strategies, and establishing quality control during mass rearing and shipment of beneficials. From these starting points, other pest species and other storage situations can be covered. These activities can contribute to satisfying public demands of more environmentally friendly pest control methods and food production that is focused on consumer safety.

References

1. *Ecology and Management of Food-Industry Pests*; Gorham, J.R., Ed.; Food and Drug Administration Technical Bulletin 4; Association of Official Analytical Chemists: Arlington, 1991.
2. Schöller, M. Biologische Bekämpfung vorratschä-dlicher Arthropoden mit Räubern und Parasitoiden—Samelbericht und Bibliografie (Biological control of arthropod pests in stored product protection with predators and parasitoids— review and bibliography). Mitt. Biol. Bundesanst. Land-Forstwirtsch. **1998**, *342,* 85–189.

3. *Integrated Management of Insects in Stored Products;* Subramanyam, B., Hagstrum, D., Eds.; Marcel Dekker, Inc.: New York, 1996.
4. *Alternatives to Pesticides in Stored-Product IPM*; Subramanyam, B., Hagstrum, D., Eds.; Kluwer Academic Publishers: Boston, 2000.
5. Flinn, P.W. Temperature effects on efficacy of *Choetospila elegans* (Hymenoptera: Pteromalidae) to suppress *Rhyzopertha dominica* (Coleoptera: Bostrichidae) in stored wheat. J. Econ. Entomol. **1998**, *91,* 320–323.
6. Holst, N.; Meikle, W.G. *Teretrius nigrescens* against larger grain borer *Prostephanus truncates* in African maize stores: biological control at work? J. Appl. Ecol. **2003**, *40,* 307–319.
7. Markham, R.; Borgemeister, C.; Meikle, W.G. Can biological control resolve the larger grain borer crisis? In *Stored Product Protection*, Proceedings of the 6th International Working Conference on Stored-product Protection, Canberra, Australia, Apr 17–23, 1994; Highly, E., Wright, E.J., Banks, H.J., Champ, B.R., Eds.; CAB: Wallingford, U.K., 1994; 1087–1097.
8. Steidle, J.L.M.; Schöller,M. Fecundity and ability of the parasitoid *Lariophagus distinguendus* (Hymenoptera: Pteromalidae) to find larvae of the granary weevil *Sitophilus granarius* (Coleoptera: Cucurlionidae) in bulk grain. J. Stored Prod. Res. **2002**, *38,* 43–53.
9. Press, J.W. Comparative penetration efficacy in wheat between the weevil parasitoids *Anisopteromalus calandrae* and *Choetospila elegans.* J. Entomol. Sci. **1992**, *27,* 154–157.
10. Flinn, P.W.; Hagstrum, D.W. Augmentative releases of parasitoid wasps in stored wheat reduces insect fragments in flour. J. Stored Prod. Res. **2001**, *37,* 179–186.
11. Hansen, L.S.; Jensen, K.-M.V. Effect of temperature on parasitism and host-feeding of *Trichogramma turkestanica* (Hymenoptera: Trichogrammatidae) on *Ephestia kuehniella* (Lepidoptera: Pyralidae). J. Econ. Entomol. **2002**, *95,* 50–56.
12. Brower, J.H.; Smith, L.; Vail, P.V.; Flinn, P.W. Biological control. In *Integrated Management of Insects in Stored Products;* Subramanyam, B., Hagstrum, D., Eds.; Marcel Dekker, Inc.: New York, 1996; 223–286.
13. Schöller, M.; Flinn, P.W. Parasitoids and predators. In *Alternatives to Pesticides in Stored-Product IPM*; Subramanyam, B., Hagstrum, D., Eds.; Kluwer Academic Publishers: Boston, 2000; 229–271.
14. Zdarkova, E. Personal experience with biological control of stored food mites. IOBC=WPRS Bull. **1998**, *21* (3), 89–93.

26

Weeds (Insects and Mites): Biological Control

Peter Harris

Introduction

The usual targets for classical biocontrol are introduced weeds of uncultivated land that lack enemies and have spread to dominate large areas. Such weeds cause huge habitat and forage losses, reduce diversity of the native flora and fauna, and displace rare species.[1] The concern about establishing their natural enemies is that desirable plants might also be attacked. Testing to ensure that this does not happen costs about two scientist years per agent ($900,000). About a third of the agents fail to establish, a third remain scarce, and nearly half of those that become abundant do not achieve control. On the other hand, costs of biocontrol increase little with infestation size and the same agent may achieve control in climates as different as Australia and Canada. Thus, the cost per ha can be small and biocontrol is less damaging to nontarget plants than most alternatives. Release approval for the agents in North America rests with Animal and Plant Health Inspections Service-United States Drug Administration and the Canadian Food Inspection Agency who consult advisory groups.

History and Impact of Classical Weed Biocontrol

Classical weed biocontrol was the serendipitous result of introducing the mealybug (*Dactylopus ceylonicus* thinking it was *D. coccus*), to India in 1795.[2] The purpose was to start a dye industry on the impenetrable stands of the South American prickly pear cactus, *Opuntia vulgaris.* The result was rapid cactus kill and little dye production. After 68 years and several eradication attempts the government distributed the insect for *O. vulgaris* control.

Hawaii took the ball in 1902 by importing 23 insect species to control the introduced shrub *Lantana camara*.[2] Their intention was to establish a complex to attack all parts of the plant. Only 12 species arrived alive in sufficient numbers for release and eight established. Success was claimed for a fly that destroyed 86% of the seed, but none of the criteria for success were demonstrated: that the weed is reduced to a low density in several locations, it remains at a low density, or it returns to a high density when protected from the agent.[3]

The mealy bug *D. ceylonicus* practically eliminated *O. vulgaris* in Australia but did not affect two more abundant cacti, which by 1925 had infested 60 million acres. Concerns that introduced insects might attack cultivated plants were met with the "no-choice" test in which only species that starve on representative crop plants were released. They tested 49 insect species, released 24, of which 12 established. Control was attributed to just two: the mealy bug *D. opuntia* and the cladode-boring moth *Cactoblastis cactorum*, which, two years after increase, caused a crash of cactus from 5000 to 11 plants/acre.[4]

Australia then tackled St. John's wort *(Hypericum perforatum)*.[1] Eight insect species were released, four established and, up to seven years after release, the leaf-feeding beetle *Chrysolina quadrigemina* increased to control the weed. The four established insects were then released in California, where 40,500 ha of open range that were infested in northern California in 1929 had increased to almost 2 million ha in western North America by 1940. Increases of *C. quadrigemina* returned most of the infested area to a native bunch grass community with fair to good productivity and an increased diversity of plant species, and the weed stabilized at about 1% of its former density.[5] The two beetle species were released in Canada in 1952. Up to 13 years later, *C. quadrigemina* increased to control the weed on summer-dry sites and *C. hyperici* on summer moist sites, but both failed above 900 m elevation.[6] This program terminated chemical control of the weed in British Columbia and removed it from the Ontario noxious weed list.

Both *Chrysolina* spp. developed on Canadian *Hypericum* spp. in no-choice tests, but in the field they have not been attacked.[6] The plant genus contains photosensitizing compounds lethal to sun exposed larvae. In nature the larvae feed at dawn and dusk on a winter mat of foliage, which is absent in the Canadian species. Two nontarget species are attacked, but are poor hosts: the introduced evergreen ornamental *H. calycinum,* which has hard mature foliage, and in California the native *H. concinnum*, which has diffuse foliage (presumably shading is reduced and the need for climbing by inactive larvae is increased).

It is more predictive to use the no-choice tests to show that the larval host range is restricted to a taxonomic group of plants rather than that individual crop plants are unacceptable. Some reviewers felt that tests of crop plants were still necessary for public confidence, but gradually these tests have been eliminated. The laboratory larval host range limit exaggerates the field host range, since congeners not attacked in nature usually support development in the test. The larva role is to stay on the plant on which it hatched and eat, which may involve distinguishing the host from intermingled vegetation, but not host finding. Host finding is done by the adult female using habitat and plant cues.[7] The larval host range limit was a satisfactory measure of risk as long as native congeners were not an issue. This changed in 1997 when the attack of native *Cirsium* spp. by the seed-head weevil *Rhinocyllus conicus* was deplored.[8] The 1967 petition for release of the weevil against the introduced thistle *Carduus nutans* stated that its host range included *Cirsium*, but this caused no concern. Today the view is mixed. Some deplore the weevil, but to many a reduction of native thistles is not a concern and the control of most stands of *C. nutans* with increased forage production, makes it an eminent success.

Three Current Issues

Host-Specificity Tests

Most introduced weeds in North America have native congeners. A total of 23 of 27 insect species established in Canada developed on congeners in laboratory no-choice tests, but only one species in Canada and three in the United States have attacked the North American relatives of the weed in the field. Exclusion of species with larvae that develop on native congeners in the laboratory will practically end weed biocontrol. A scientifically sound alternative is to base release approval on the adult host range. The larval no-choice test does not distinguish between species with adults that oviposit on the weed's relatives in the field from those that exclude them. However, adult tests in large field cages or open releases allow assessment of this choice. Thus, although the beetle *Altica carduorum* develops on all *Cirsium* spp., the adult only "sees" *C. arvense* in large arenas. This combined with a low suitability of other *Cirsium* spp. for the immature stages means there is selection pressure against host plant shifts and accounts for the field specificity of the beetle in Eurasia.[9] However, the development of new host screening procedures are currently at an impasse because some reviewers still feel larval ability to complete development in the laboratory is an unacceptable risk even if the adult does not recognize the plant as a host.

Costs and Agent Success

Past cost increases of weed biocontrol have been partly compensated for by using fewer agents and by improving establishment success. The first Hawaiian and first two Australian projects tried to release 55 insect species of which 44% established and three reduced weed density. The current world establishment rate is 65% with the release of 4.1 species against each successfully controlled weed.[10] The idea that all parts of the plant need to be attacked is not supported by the fact that 81% of the successes result from single species. Where several agents contributed, they tend to do so in different habitats such as *C. quadrigemina* and *C. hyperici.*

New host tests are likely to increase costs. An obvious solution is to eliminate the approximately 80% of agents released that contribute little. Present practice is to be more discriminating with expensive agents than those that are cheap simply because testing was done by another country (Table 1). However, establishing any agent may have ecological effects. Thus, it is prudent to reduce the number released to the necessary minimum. No agent trait associated with effectiveness has been found, although establishment is broadly correlated with a high rate of increase, long adult life, the number of generations a year, and small size. However, effectiveness may be plant related. For example, root feeders have been more effective on tansy ragwort, *Senecio jacobaea*, and leafy spurge, *Euphorbia esula*, than defoliators and seed feeders, which have been effective on other plants. If this were determined in preliminary studies, host specificity tests could be restricted to insects in the relevant feeding guild. This combined with studies of the insect's ecological requirements could halve the number of failures.

Legislation

The enabling legislation used for classical weed biocontrol is the Federal Plant Pest Act of 1957 in the United States and the Plant Protection Act of 1990 in Canada. The purpose of both is to prevent pest establishment and spread, so release of nonpests is allowed by reverse logic. However, with no explicit mention of biocontrol, other acts, with various purposes, come into play. For example, the U.S. Endangered Species Act of 1973 requires federal agencies to ensure their actions are not likely to jeopardize the continued existence of endangered or threatened species. The ability of larvae to develop on an endangered plant may indicate jeopardy, but the issue is complicated. The act can save habitat from human destruction, but it cannot stop invading weeds, such as leafy spurge, which are displacing the

TABLE 1 Success of Weed Biocontrol Agents Released in Canada (1981–1990)

	Sponsor of Pre-Release Studies	
	Canada	United States
No. released	18	6
% Established	73	33
Species contributing to control	8	1
Canadian costs/agent	2 SY[a]	0.04 SY
Canadian costs/success	6.6 SY	0.24 SY

[a] SY = scientist year, currently about $450,000.

threatened northern prairie skink in Canada and the endangered western prairie fringed orchid in both countries. Biocontrol is the best hope for their survival; however, there are rare Florida spurges that will support larval development of leafy spurge agents. It is not possible under the act to weigh the benefits against the risks, although the risks are small if the agent has an obligatory winter diapause that prevents survival in Florida. Nevertheless, the approval of agents for the major North American leafy spurge problem is currently stalled. The Australian Biocontrol Act of 1984 solves the issue by instructing reviewers to approve the release of agents when the expected benefits outweigh the risks. A further need is to judge risk on a holistic assessment of adult and immature needs and habits. It would also be helpful to make weed biocontrol more open to public input and have procedures for resolving conflicts of interest.

References

1. Pimentel, D.L.; Lach, L.; Zuniga, R.; Morrison, D. Environmental and Economic Costs Associated with Non-Indigenous Species in the United States. Available at http://www.news. cornell.edu/releases/Jan99/species_costs.htm (accessed Feb 25, 1999).
2. Goeden, R.D. Biological Control of Weeds. In *Introduced Parasites and Predators of Arthropod Pests and Weeds: A World Review;* Clausen, C.P., Ed.; Part II, Agricultural Handbook 480, USDA: Washington, DC, 1978; 357–414.
3. Smith, H.S.; DeBach, P.; The measurement of the effect of entomophagous insects on population densities of their hosts. J. Econ. Entomol. **1942**, *3,* 845–849.
4. Crawley, M.J.; The success and failures of weed biocontrol using insects. Biocontrol. News Inf. **1989**, *10,* 2213–2223.
5. Huffaker, C.B.; Kennet, C.E.; A ten-year study of vegetation changes associated with biological control of klamath weed. J. Range Manage. **1959**, *12,* 69–82.
6. Harris, P. Status of Introduced and Main Indigenous Organisms on Weeds Targeted for Biocontrol in Canada. Available at http://res.agr.ca/leth/weedbio.htm (accessed Dec 20, 1999).
7. Fox, C.W.; Lalonde, R.G.; Host confusion and the evolution of insects' diet breadths. Oikos **1993**, *67,* 577–581.
8. Louda, S.M.; Kindal, D.; Connor, J.; Simberloff, D. Ecological effects of an insect introduced for the biological control of weeds. Science **1997**, *277,* 1088–1090.
9. Harris, P. Evolution of classical weed biocontrol: meeting survival challenges. Bull. Ent. Soc. Can. **1998**, *30,* 134–143.
10. Julien, M.J. Biological control of weeds worldwide: trends, rates of success and the future. Biocontrol. News Inf. **1989**, *10,* 299–306.

VI

NEC: Natural Elements and Chemicals Found in Nature

27

Antagonistic Plants

Philip Oduor-Owino

Introduction

Plant parasitic nematodes cause significant crop losses in Africa and other parts of the world. Infected plants suffer from water deficiency and low yields, and have necrotic and/or galled roots. Control of nematodes has been mainly through the use of chemicals and host resistance. However, the existence of physiological races in the pathogen's population has complicated efforts to breed for resistant cultivars. Chemical control is effective but difficult to sustain for long-term benefits. The high cost of nematicides and their toxic effects also make them less attractive. Some nematicides such as Nemagon and Fumazon have now been banned from the world market and this has placed severe constraints on strategies for nematode control. Interest in developing alternative control measures that are safe and economically attractive has now intensified worldwide. The use of antagonistic plants is viewed as a viable nematode management option.[1–3]

Nature of Antagonistic Plants

Antagonistic plants are defined as plants that produce chemicals in their roots that are toxic and/or repellant to phytonematodes in the soil ecosytem.[4] These plants include *Tagetes erecta* L; *Tagetes patula* L; *Datura stramonium* L; *Ricinus communis* L; and *Asparagus officinalis*. Fresh roots of asparagus produce asparaguric acid glycoside that is toxic, even when diluted, to most plant parasitic nematodes. Root exudates from *Tagetes, Datura,* and *Ricinus* spp. induces premature hatching of nematode eggs, blocks the processes of mitosis and meoisis, and reduces galling intensity on roots of susceptible plants. This has been attributed, in part, to the toxic effects of the alkaloids terthienyl, hyosine, and ricinine present in *Tagetes, Datura,* and *Ricinis* spp., respectively. These compounds may also disrupt female taxis to roots or male taxis to female.[4] Other plants with antagonistic properties include some crucifers and citrus. Root diffusates from crucifers reduce the pathogenicity of nematodes on potato, while a compound in citrus roots is toxic to *Tylechulus semipenetrans*.[1]

Antagonistic Plants in Cultural Pest Control

Antagonistic plants may have a great nematode-control potential in agriculture if properly utilized in crop rotation and intercropping systems.[2] For example, intercropping food crops with nematicidal plants is now a nematode management strategy in Tanzania, India, and Zimbabwe and has also been

TABLE 1 Effect of Tomato Intercropping with *Datura stramonium*, *Ricinus communis*, and *Tagetes minuta*, and Soil Treatment with Captafol and Aldicarb, on Gall Index, Number of Juveniles, Tomato Growth and Fungal Parasitism of *Meloidogyne javanica* Eggs by the Fungus *Paecilomyces lilacinus*, 50 Days after Inoculation

Soil Treatment[a]	Egg Parasitism (%)	No. of Juveniles/ 300 cm^3 soil	Shoot Dry Weight (g)	Shoot Height (cm)	Gall[b] Index (0–4)
Ne "only" untreated	1.0de	670a	1.5f	26.6f	4.0a
Soil "only" untreated	0.0e	0d	3.5b	43.1ab	0.0e
F + Ne	23.2c	660a	1.6f	30.1e	3.0ab
F + Cap + Ne	1.3de	635a	1.9f	31.2e	3.9a
F + Ald + Ne	27.6b	12de	4.6a	45.3a	1.4d
F + Tag + Ne	29.8ab	161c	2.4e	38.4b	2.1c
F + Dat + Ne	28.3ab	173c	2.5de	36.4cd	2.2c
F + Ric + Ne	30.9a	210c	2.9cd	35.4cde	2.4c
Tag + Ne	2.1de	209c	2.8d	33.4cde	2.9b
Dat + Ne	3.0d	183c	3.0cd	36.6c	2.8b
Ric + Ne	2.8de	204c	3.2bc	37.7bc	3.0ab
Ald + Ne	1.2de	14d	4.5a	46.1a	3.8a
Cap + Ne	0.0e	460b	1.9f	36.1d	4.0a

Source: Oduor-Owino.[7]

Note: Numbers are means of 10 replications. Means followed by different letters within a column are significantly different ($P = 0.05$) according to Duncan's Multiple Range Test.

[a] Ne, nematode; F, P. *lilacinus*; Cap, Captafol; Ald, Aldicarb; Tag, *T. minuta*; Ric, *R. communis.*

[b] Gall index was based on a 0–4 rating scale, where 0 = no galls and 4 = 76–100% of the root system galled.

recommended for Pakistan.[5] Field trials with *T. minuta*, *D. stramonium*, and *R. communis* are promising.[2,5,6] These plants reduce galling intensity and enhance tomato performance significantly. In India, a rotation of *D. stramonium*, maize, tomato, and pepper reduced the population of root-knot nematodes by 30% but the level of nutrient depletion by the antagonistic plant was 15%.[4] Integration of these plants with the biological agent, *Paecilomyces lilacinus* Thom (Sam), gave better results in Kenya.[3,7,8] Tomato plants grown in soils planted with the various antagonistic plants in combination with *P. lilacinus* develop significantly heavy shoots and roots and relatively fever root galls compared to controls (Table 1). Cases where antagonistic plants are used in crop rotation or intercropping systems are now increasing.[4] For instance, in Indonesia, *Tagetes* sp. *Crotalaria usaramoensis,* corn, and sweet potato (*Ipomea batatas*) are used to reduce *Meloidogyne* spp. density in the soil. For cereal-based cropping systems, the following crop sequences for root-knot nematode control are recommended in the Philippines: rice-mung bean *(Phaseolus* aureus)-corn-cabbage-rice, rice-tobacco *(Nicotiana tobacum)-rice* and rice-tobacco and *Tagetes* spp. There is also considerably less galling by *Meloidogyne* spp. on potato *(Solanum tuberosum)* roots intercropped with onion (*Allium* sp.), corn, and marigold compared with galling found on potatoes alone. Although antagonistic plants are gaining popularity as pest management tools, their benefits and risks must be understood thoroughly before one can exploit their potential in pest control.

Benefits and Risks of Antagonistic Plants

Benefits and risks associated with the utilization of antagonistic plants in agriculture are varied. Phytonematoxic plants such as *R. communis, D. stramonium, Tagetes* spp. *Crotalaria* spp. *A. najus,* and *Datura metel* L. are traditionally gaining popularity due to their medicinal significance.[9] The flowers of *D. metel* are used against asthma, while *Crotalaria* spp. is a nitrogen-fixing legume. Castor oil from *R. communis* is used for making soaps and waxes: rinolecic acid from castor seeds is a valuable laxative.[9] Despite these attributes, antagonistic plants may pose a serious threat to food production if not well utilized. They may compete with economically important crops for space and nutrients.

In addition they are slow in action, an attribute that makes them less attractive for use in a commercial setting. Because of this scenario, it is important that scientific disciplines work together in order to develop a viable pest control system. What is good for the nematologist may not be good for either the agronomist or economist.

Future Concerns

There is increasing internal awareness of the value of natural plants and their products in the development of new drugs and formulation of materials that can be used for pest control. Since some of the antagonistic plants can also be used to treat human ailments,[9] they may attract intensive scientific evaluation, recognition, and funding. However, more work should be done to reexamine the future of antagonistic plants in nematode control and in the drug industry. Efficacy of these plants against nematodes and their utilization in the pharmaceutical industry will depend highly on the concentrations of the active ingredients in their tissues.[10] It will also depend on whether they can stimulate activity of most biocontrol fungi and plant growth consistently. They have so far enhanced tomato growth in the greenhouse and in the field significantly[2,7] (Tables 2 and 3), but more trials are needed in order to understand the relationship between antagonistic plants, natural enemies, and crop performance.

TABLE 2 Effect of Soil Treatment with Aldicarb, *Tagetes minuta, Datura metel,* and *Datura stramonium* on Root-Knot Nematodes in Tomatoes (Greenhouse Test)

Soil Treatment[a]	Shoot Height (cm)	Shoot Dry Weight (g)	Gall Index (0–5)	Galls (no.g^{-1}) root Weight	Nematodes, No. $(300\ mL)^{-1}$ Soil
Soil + Ne[b], untreated (control)	42.4c[c]	1.3e	4.4a	510.0a	564.1a
Soil only, untreated	97.8a	4.9b	0.0c	0.0d	0.0e
Soil + Ne + Aldicarb	116.3a	6.1a	1.0c	23.3d	18.4d
Soil + Ne + *D. metel*	73.6d	3.7c	2.1b	77.4c	176.3c
Soil + Ne + *T. minuta*	65.4b	2.6d	2.0b	134.9b	310.4b
Soil + Ne + *D. stramonium*	73.4b	3.1c	2.2b	88.4c	170.4c
Soil + Ne + *R. communis*	69.0b	3.3c	2.4b	90.0c	173.0c

Source: Oduor-Owino, P. Effects of Aldicarb and Selected Medicinal Plants of Kenya on Tomato Growth and Root-Knot Severity, unpublished data, 1992.

[a] Autoclaved soil used.

[b] Ne, nematode eggs added to soil.

[c] Means followed by the same letter within each column are not significantly different at the 5% level (Duncan's Multiple Range Test).

TABLE 3 Effect of Soil Treatment with Aldicarb, *Tagetes minuta*, *Datura metel*, and *Datura stramonium* on Infection of Tomato by Root-Knot Nematodes (Field Test)

Treatment	Shoot Height (cm)	Shoot Dry Weight (g)	Fruit Yield (g)	Galls (No. g^{-1} Root Weight)	Nematodes, No. $(300\ mL)^{-1}$ Soil
Untreated (control)	80.3d[a]	40.5d	380.3e	69.1a	150.4a
Aldicarb	187.3a	135.1a	3800.4a	4.50d	6.4d
D. metel	157.1b	89.3b	2590.1b	6.4c	17.3bc
T. minuta	107.1c	45.1d	761.1c	11.4b	21.1b
D. stramonium	150.1b	69.4c	2030.4b	9.6b	18.4c

Source: Oduor-Owino.[2]

[a] Means followed by the same letter within each column are not significantly different at the 5% level (Duncan's Multiple Range Test).

See also *Biological Control of Nematodes*, pages 6163; *Risks of Biological Control*, pages 720–722; *Toxins in Plants*, pages 840–842; *Pest-Host Plant Relationships*, pages 593–594; *Intercropping for Pest Management*, pages 423–425.

References

1. Caswell, E.P.; Tan, C.S.; De Frank, J.; Apt, W.J. The influence of root exudates of *Chloris gayana* and *Tagetes patula on Rotylenchulus reniformis.* Revue de Nematologie **1991**, *14* (2), 581–587.
2. Oduor-Owino, P. Effects of Aldicarb, *Datura stramonium, Datura metel* and *Tagetes minuta* on the pathogenicity of root-knot nematodes in Kenya crop protection. Organic Soil Amendment **1993**, *12* (4), 315–317.
3. Oduor-Owino, P.; Sikora, R.A.; Waudo, S.W.; Schuster, R.P. Effects of aldicarb and mixed cropping with *Datura stramonium, Ricinus communis* and *Tagetes minuta* on the biological control and integrated management of *Meloidogyne javanica.* Nematologica **1996**, *42,* 127–130.
4. Yeates, G.W. How plants affect nematodes. Advances in Ecological Research **1987**, *17* (2), 61–137.
5. Oduor-Owino, P.; Waudo, S.W. Effects of antagonistic plants and chicken manure on the biological control and fungal parasitism of root-knot nematode eggs in naturally infested field soil. Pakistan Journal of Nematology **1995**, *13* (2), 109–117.
6. Oduor-Owino, P.; Waudo, S.W. Comparative efficacy of nematicides and nematicidal plants on root-knot nematodes. Trop. Agric. **1994**, *71* (4), 272–274.
7. Oduor-Owino, P. *Fungal Parasitism of Root-knot Nematode Eggs and Effects of Organic Matter, Selected Agrochemicals and Intercropping on the Biological Control of Meloidogyne javanica on Tomato*; Ph.D. Thesis, Kenyatta University: Nairobi, Kenya, 1996; 132.
8. Oduor-Owino, P.; Sikora, R.A.; Waudo, S.W.; Schuster, R.P. Tomato growth and fungal parasitism of *Meloidogyne javanica* eggs as affected by nematicides, time of harvest and intercropping. East African Agricultural and Forestry Journal **1995**, *61* (1), 23–30.
9. Oduor-Owino, P.; Waudo, S.W. Medicinal plants of Kenya: effects on *Meloigogyne incognita* and the growth of okra. Afro-Asian Journal of Nematology **1992**, *2* (1), 64–66.
10. Oduor-Owino, P. Effects of marigold leaf extract and captafol on fungal parasitism of root-knot nematode eggskenyan isolates. Nematologia Mediteranea **1992**, *20,* 211–213.
11. Oduor-Owino, P.; Waudo, S.W.; Sikora, R.A. Biological control of *Meloidogyne javanica* in Kenya: effect of plant residues, benomyl, and decomposition products of mustard *(Brassica campestris).* Nematologica **1993**, *39,* 127–134.

28

Arthropod Host-Plant Resistant Crops

Gerald E. Wilde

Introduction

Because of the many advantages plant resistance offers, virtually every cultivated crop has been evaluated for this trait and one or more resistant sources have been identified. The challenge has been to incorporate these resistant sources into agronomically adapted and consumer acceptable, high-yielding cultivars. In addition to traditional breeding methods, the use of modern breeding techniques and genetic transformation of crops has opened the door to other ways of identifying, incorporating, and employing pest-resistance genes to effectively and economically manage arthropod pest populations. The use of resistant cultivars contributes significant economic and social benefits and sustainable agricultural systems to the world's farmers. The positive effects of resistant cultivars have been demonstrated repeatedly in crops as diverse as wheat, alfalfa, grape, sorghum, maize, rice, apple, and cotton.

Percentage of Crops That Have Some Degree of Pest Resistance

Plant resistance has been employed to a greater or lesser degree in practically all of the major food, feed, and fiber crops. Table 1 lists a number of major crops grown in the world and the number of pests for which resistance has been employed to at least some extent in the field. Hectarage planted to resistant cultivars varies for each pest and crop and over time as new varieties and hybrids (both susceptible and resistant) are grown and, in some instances, as new pest biotypes (pest populations that are capable of damaging previously resistant sources) develop. For example, most of the modern rice varieties and hybrids grown in China, India, and other countries are resistant to one or more major pests. Resistant American grape rootstocks have been used extensively over the world to control *Phylloxera vittifolae* (Fitch). A large percentage of the alfalfa planted in the United States is comprised of varieties resistant to aphid species. Sorghum hybrids with resistance to the greenbug have occupied up to 80% of the hectarage in the United States. Significant hectarages of wheat and barley in the United States, Canada, and North and South Africa have resistance to at least one pest. Most commercial soybean

TABLE 1 List of Some Major Crops Grown in the World and Number of Arthropod Pests for Which Resistant Cultivars Have Been Used in the Field by Growers for Pest Management

Crop	No. of Pests
Alfalfa	6
Apple	1
Asparagus	1
Barley	3
Bean	1
Cassava	2
Chickpea	0
Cotton	6
Grape	1
Lettuce	1
Maize	10
Millet	1
Oat	1
Pea	1
Peanut	4
Potato	1
Raspberry	1
Rice	14
Rye	1
Sorghum	6
Soybean	1
St. Augustine grass	1
Sugar beet	1
Sugarcane	3
Sunflower	1
Sweet clover	1
Sweet potato	1
Wheat	7

varieties are resistant to the potato leafhopper. Several cotton varieties carrying genes for resistance to jassids (*Empoasca* sp) are grown widely in Africa, India, and the Philippines. In the United States, more than 65% of commercial maize hybrids have some resistance to corn leaf aphid, >90% have some resistance to first generation European corn borer, and >75% have some resistance to second generation corn borer.

However, many more resistance genes have been identified in all crops than have been used in modern commercial varieties and hybrids, because incorporating them into high yielding cultivars acceptable to growers has been difficult. Recently, transgenic crops have been utilized to combat major insect pests. Hybrids or varieties with insect-resistance genes have been developed in cotton, maize, and potato. An estimated 6.7 million hectares of transgenic corn resistant to the European corn borer, 2.5 million hectares of transgenic cotton resistant to several pests, and 20,000 hectares of transgenic potato resistant to Colorado potato beetle were grown in the world in 1998. The hectares planted to transgenic crops are likely to increase as additional countries register these products and this technology is used on additional crops. For example, specific biotechnology applications are being field tested for rice and wheat, which together occupy 400 million hectares globally.

Effect of Plant Resistance on Pest Populations

The growing of pest-resistant cultivars can be used as a major control tactic or adjunct to other measures. Historically, the use of resistant cultivars combined with other tactics has resulted in a reduction of many pest species to subeconomic levels. Even small increases in resistance enhance the effectiveness of cultural, biological, and insecticidal controls. The extent to which growing resistant plants affects pest populations is dependent upon the level of resistance expressed, the mechanisms of resistance involved, and the number of hectares grown. The growing of resistant wheat on 50% of the hectarage in Kansas has been shown to reduce Hessian fly populations to extremely low levels. Resistance in wheat to wheat curl mite (ca. 25% of the hectarage) was effective in limiting the spread of wheat streak mosaic virus, which the mite transmits. The incorporation of leaf and stem pubescence into most commercial soybean varieties has resulted in population suppression of the potato leafhopper over the past 60 years. As the hectarage of sorghum resistant to the greenbug increased to >50%, the area of sorghum treated with insecticide was reduced by 50%. Tenfold reductions in pest populations have been observed where insect-resistant rice cultivars have been grown widely.

Economic and Social Benefits

Assessing the economic benefits of plant resistance is difficult in the context of integrated pest management programs and is likely to be underestimated frequently and substantially. Even determining the obvious advantages (yield benefits and reduced production costs) may be difficult over a large area where pest populations vary from locality to locality and year to year. Other environmental benefits, such as cleaner water and food, reduced risks to farmers, more flexibility in planting and cropping systems, reduced disease transmission, and reduced secondary pest outbreaks, also are difficult to quantify. Nevertheless, some specific estimates are available. In the United States alone the estimated valued of using arthropod-resistant alfalfa, barley, corn, sorghum, and wheat cultivars is more than $1.4 billion each year. The net economic benefit of greenbug resistance in U.S. sorghum production is estimated at close to $400 million annually. The global economic value of arthropod-resistant wheat has been estimated at $250 million annually. The value of resistance to aphids in alfalfa in the major alfalfa-producing states of the United States is estimated at more than $100 million annually. Breeding for pest resistance in rice has been estimated to be responsible for one-third of recent yield increases and $1 billion of additional annual income to rice producers. The net return of insect-resistant Bt maize in the United States and Canada has been estimated in some studies at $42.00–$67.30 per hectare, but other studies have indicated less of an economic return. The average net economic return of insect-resistant Bt cotton in 1997 was $133 per hectare.

Bibliography

1. Antle, J.M.; Pingali, P.L. Pesticides, productivity and farmer health: A philippine case study. Am. J. Agric. **1994**, *76,* 418–430.
2. *Global Plant Genetic Resources for Insect-Resistant Crops;* Clement, S.L., Quisinberry, S.S., Eds.; CRC Press: New York, 295.
3. Harvey, T.L.;Martin,T.J.; Seifers, D.L. Importanceof plant resistance to insect and mite vectors in controlling virus diseases of plants: resistance to the wheat curl mite (Acari: Eriophyidae). J. Agric. Entomol. **1994**, *11,* 271–277.
4. Hyde, J.; Martin, M.A.; Preckel, P.V.; Edwards, L.R. The economics of Bt corn: valuing protection from the European corn borer. Rev. Agric. Econ. **1999**, *21,* 442–454.
5. James, C. *Global Review of Commercialized Transgenic Crops;* ISAAA Briefs No. 8, ISAA: Ithaca, NY, 1998, 43.

6. In *Insect Resistant Maize: Recent Advances and Utilization*, Proceedings of an International Symposium, International.
7. Maize and Wheat Improvement Center (CIMMYT), Nov 27–Dec 3, 1994; Mihm, J.A., Ed.; CIMMYT: Mexico, D.F., 1997; 302.
8. Painter, R.H. *Insect Resistance in Crop Plants;* University of Kansas Press: Lawrence, KS, 1968, 520.
9. Smith, C.M. *Plant Resistance to Insects. A Fundamental Approach*; John Wiley and Sons: New York, 1989, 286.
10. Smith, C.M.; Quisinberry, S.S. Value and use of plant resistance to insects in integrated pest management. J. Agric. Entomol. **1994**, *11,* 189–190.
11. van Emden, H.F. Host-Plant Resistance to Insect Pests. In *Techniques for Reducing Pesticides: Environmental and Economic Benefits*; Pimentel, D., Ed.; John Wiley: Chichester, England, 1997, 124–132.
12. In *Economic, Environmental, and Social Benefits of Resistance in Field Crops*, Proceedings, Thomas Say Publications in Entomology, Wiseman, B.R., Webster, J.A., Eds.; Entomological Society of America: Lanham, MD, 1999, 189.
13. Contribution No. 00–252-B of the Kansas Agricultural Experiment Station.

29

Biomass

Alberto Traverso
and David Tucker

Introduction

This work is organized into seven main sections. The first section provides the reader with a general overview on biomass, including definition, environmental benefits, energetic properties, and a short list of biomass types that can be used as energy sources. The second section illustrates the mechanical processes to produce standardized solid biomass fuels. The third section describes one of the major technologies for converting biomass into energy, combustion. The fourth section analyzes pyrolysis and gasification as promising techniques for efficient exploitation of biomass, still in a precommercial phase. The fifth section is concerned with biochemical processes for producing biogas and biofuels for transportation. The sixth section outlines the major benefits from biomass exploitation for energy purposes. The seventh section reports about the potential carbon dioxide (CO_2) emission reduction due to extensive use of biomass as a renewable energy resource. The eighth section concludes this entry.

Generalities about Biomass

In general, biomass is whatever substance produced or by-produced by biological processes. Commonly, biomass refers to the organic matter derived from plants and generated through photosynthesis. Biomass provides not only food but also construction materials, textiles, paper, medicines, and energy. In particular, biomass can be regarded as solar energy stored in the chemical bonds of the organic material or as a reduced state of carbon. CO_2 from the atmosphere and water absorbed by the plant roots are combined in the photosynthetic process to produce carbohydrates (or sugars) that form the biomass.

TABLE 1 Typical Types of Biomass for Energy Use

Supply Sector	Type	Example
Forestry	Dedicated forestry	Short rotation plantations (e.g., willow, poplar, eucalyptus)
	Forestry by-products	Wood blocks, wood chips from thinnings
Agriculture	Dry lignocellulosic energy crops	Herbaceous crops (e.g., miscanthus, reed canary-grass, giant reed)
	Oil, sugar and starch energy crops	Oil seeds for methylesters (e.g., rape seed, sunflower)
		Sugar crops for ethanol (e.g., sugar cane, sweet sorghum)
		Starch crops for ethanol (e.g., maize, wheat)
	Agricultural residues	Straw, prunings from vineyards and fruit trees
	Livestock waste	Wet and dry manure
Industry	Industrial residues	Industrial waste wood, sawdust from sawmills
		Fibrous vegetable waste from paper industries
Waste	Dry lignocellulosic	Residues from parks and gardens (e.g., prunings, grass)
	Contaminated waste	Demolition wood
		Organic fraction of municipal solid waste
		Biodegradable landfilled waste, landfill gas
		Sewage sludge

Source: Adapted from European Biomass Industry Association[1] and DOE Biomass Research and Development Initiative.[2]

The solar energy that drives photosynthesis is stored in the chemical bonds of the biomass structural components. During biomass combustion, oxygen from the atmosphere combines with the carbon and hydrogen in biomass to produce CO_2 and water. The process is therefore cyclic because the carbon dioxide is then available to produce new biomass. This is also the reason why bioenergy is potentially considered as carbon- neutral, although some non-recoverable CO_2 emissions occur due to the use of fossil fuels during the production and transport of biofuels.

Biomass resources can be classified according to the supply sector, as shown in Table 1.

The chemical composition of plant biomass varies among species. Yet, in general terms, plants are made of approximately 25% lignin and 75% carbohydrates or sugars. The carbohydrate fraction consists of many sugar molecules linked together in long chains or polymers. Two categories are distinguished: cellulose and hemi- cellulose. The lignin fraction consists of non-sugar-type molecules that act as a glue holding together the cellulose fibers.

Energy Content of Biomass

Bioenergy is energy of biological and renewable origin, normally derived from purpose-grown energy crops or by-products of agriculture and forestry. Examples of bioenergy resources are wood, straw, bagasse, and organic waste. The term *bioenergy* encompasses the overall technical means through which biomass is produced, converted, and used. Figure 1 summarizes the variety of processes for energy production from biomass.

The calorific value of a fuel is usually expressed as higher heating value (HHV) and/or lower heating value (LHV). The difference results from the vaporization of water formed from the combustion of hydrogen in the material and the original moisture.

The most important property of biomass feedstocks with regard to combustion—and to the other thermochemical processes—is the moisture content, which influences the energy content of the fuel. Wood, just after falling, has a typical 55% water content and an LHV of approximately 7.1 MJ/kg; logwood after 2–3 years of air-drying may present a 20% water content and an LHV of 14.4 MJ/kg; pellets show a quite constant humidity content of about 8% with an LHV equal to 17 MJ/kg.

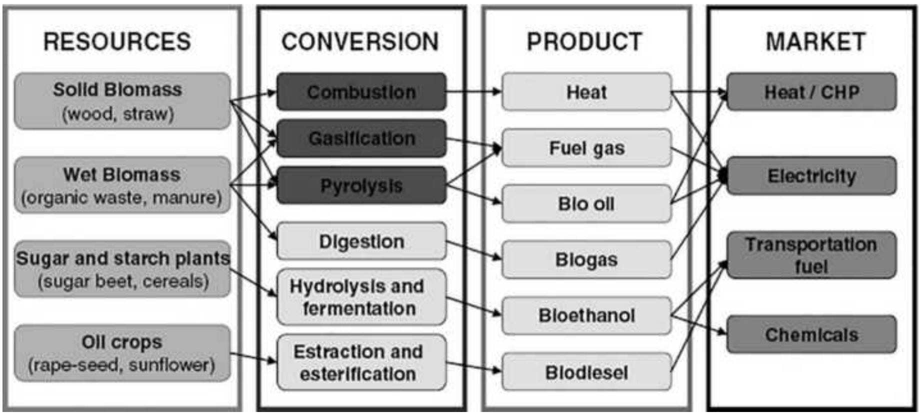


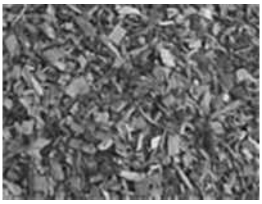
FIGURE 1 Processes to convert biomass into useful energy, i.e., bioenergy.
Source: European Biomass Industry Association,[1] Overend et al.[3] and Risoe National Laboratory.[6]

Mechanical Processes for Energy Densification

Some practical problems are associated with the use of biomass material (sawdust, wood chips, or agricultural residues) as fuel. Those problems are mainly related to the high bulk volume, which results in high transportation costs and requires large storage capacities, and to the high moisture content, which can result in biological degradation as well as in freezing and blocking the in-plant transportation systems. In addition, variations in moisture content make difficult the optimal plant operation and process control. All those problems may be overcome by standardization and densification. The former consists in processing the original biomass in order to obtain fuels with standard size and heating properties, while the latter consists in compressing the material, which needs to be available in the sawdust size, to give it more uniform properties.

Table 2 reports the main features of pellets, briquettes, and chips.

TABLE 2 Comparison of Different Solid Wood Fuels

	Pellets	Briquettes	Chips
Appearance			
Raw material	Dry and ground wood or agricultural residues	Dry and ground wood or agricultural residues. Raw material can be more coarse than for pelleting, due to the larger dimensions of final product	Dry wood logs
Shape	Cylindrical (generally Ø 6 to 12 mm, with a length 4 to 5 times the Ø).	Cylindrical (generally Ø 80 to 90 mm) or parallelepiped (150 × 70 × 60 mm)	Irregularly parallelepiped (70 × 30 × 3 mm)

Biomass Combustion

The burning of wood and other solid biomass is the oldest energy technology used by man. Combustion is a well- established commercial technology with applications in most industrialized and developing countries, and development is concentrated on resolving environmental problems, improving the overall performance with multifuel operation, and increasing the efficiency of the power and heat cycles (combined heat and power, CHP).

The devices used for direct combustion of solid biomass fuels range from small domestic stoves (1 to 10 kW) to the large boilers used in power and CHP plants (>5 MW). Intermediate devices cover small boilers (10 to 50 kW) used in single family houses heating, medium-sized boilers (50 to 150 kW) used for multifamily house or building heating, and large boilers (150 kW to more than 1 MW) used for district heating. Cofiring in fossil-fired power stations enables the advantages of large-sized plants (>100 MWe) that are not applicable for dedicated biomass combustion due to limited local biomass availability.

To achieve complete burnout and high efficiencies in small-scale combustion, downdraft boilers with inverse flow have been introduced, applying the two-stage combustion principle. An operation at very low load should be avoided as it can lead to high emissions. Hence, it is recommended to couple log wood boilers to a heat storage tank. Since wood pellets are well suited for automatic heating at small heat outputs, as needed for buildings nowadays, pellet furnaces are an interesting application with increasing propagation. Thanks to the well-defined fuel at low water content, pellet furnaces can easily achieve high combustion quality. They are applied both as stoves and as boilers and find increased acceptance in urban areas, due to the high efficiency of modern pellet stoves now used for home heating. While a conventional fireplace is less than 10% efficient at delivering heat to a house, an average modern pellet stove achieves 80%–90% efficiency. Technology development has led to the application of strongly improved heating systems, which are automated and have catalytic gas cleaning equipment. Such systems significantly reduce the emissions from fireplaces and older systems while at the same time significantly improving the efficiency.

Understoker furnaces are mostly used for wood chips and similar fuel with relatively low ash content, while grate furnaces can also be applied for high ash and water content. Special types of furnaces have been developed for straw, a very low density material that is usually stored in bales. Beside conventional grate furnaces operated with whole bales, cigar burners and other specific furnaces are in operation. Stationary or bubbling fluidized bed (SFB) as well as circulating fluidized bed (CFB) boilers are applied for large-scale applications and often used for waste wood or mixtures of wood and industrial wastes, e.g., from the pulp and paper industry.

Co-Combustion

Bioenergy production might be hampered by limitations in the supply and/or fuel quality. In those cases, cofiring of several types of biomass or of biomass with coal ensures flexibility in operation, both technically and economically. Several concepts have been developed:

- Co-combustion or direct cofiring. The biomass is directly fed to the boiler furnace, if needed after physical preprocessing of the biomass such as drying, grinding, torrefaction, or metal removal is applied. This typically takes place in SFB or CFB combustors. Such technologies can be applied to a wide range of fuels, even for very wet fuels like bark or sludge. Multifuel fluidized bed boilers achieve efficiencies of more than 90%, while flue gas emissions are lower than for conventional grate combustion due to lower combustion temperatures.
- Indirect cofiring. Biomass is first gasified and the fuel gas is then cofired in the main boiler. Sometimes, the gas has to be cooled and cleaned, which is more challenging and implies higher operation costs.
- Parallel combustion. The biomass is burnt in a separate boiler for steam generation. The steam is used in a power plant together with the main fuel.

Problems in Biomass Combustion

Biomass has a number of characteristics that makes it more difficult to handle and combust than fossil fuels. The low energy density, the high water content, and the tendency to "bridge" in tanks or pipes are the main problems in handling and transport of the biomass, while the difficulty in using biomass as fuel relates to its content of inorganic constituents. Some types of biomass used contain significant amounts of chlorine, sulfur, and potassium. The salts, KCl and K_2SO_4, are quite volatile, and the release of these components may lead to heavy deposition on heat transfer surfaces, resulting in reduced heat transfer and enhanced corrosion rates. Severe deposits may interfere with operation and cause unscheduled shutdowns.

In order to minimize these problems, various fuel pretreatment processes have been considered, including washing the biomass with hot water or using a combination of pyrolysis and char treatment.

Thermochemical Conversion of Biomass

Pyrolysis and gasification are the two most typical thermochemical processes because they convert the original bioenergy feedstock into more convenient energy carriers such as producer gas, oil, methanol, and char,[3] instead of producing useful energy directly.

Pyrolysis

Pyrolysis is a process for thermal conversion of solid fuels, like biomass or wastes, in the complete absence of oxidizing agent (air/oxygen) or with such limited supply that gasification does not occur to any appreciable extent. Commercial applications are focused on either the production of charcoal or the production of a liquid product, the bio-oil, and pyro-gas. Charcoal is a very ancient product. Traditional processes (partial combustion of wood covered by a layer of earth) were very inefficient and polluting, but modern processes presently used in industry such as rotary kiln carbonization have been optimized to maximize efficiency and minimize environmental impact. Bio-oil production (or wood liquefaction) is potentially very interesting as a substitute for fuel oil and as a feedstock for production of synthetic gasoline or diesel fuel. Pyro-gas has higher energy density than gasification gas (syngas) because it has been created without oxygen (and nitrogen, if air is employed); hence, it does not contain the gaseous products of partial combustion.

Pyrolysis takes place at temperatures in the range 400–800°C, and during this process, most of the cellulose and hemicellulose and part of the lignin will disintegrate to form smaller and lighter molecules, which are gases at the pyrolysis temperature. As these gases cool, some of the vapors condense to form a liquid, which is the bio-oil and tar. The remaining part of the biomass, mainly parts of the lignin, is left as a solid, i.e., the charcoal. It is possible to influence the product mix through a control of heating rate, residence time, pressure, and maximum reaction temperature, so that either gases, condensable vapors, or the solid charcoal is promoted.

Gasification

Modern gasification technology has been developed since the 18th century (first historical hints about gasification date from the Chinese Han Dynasty, between 206 BC and AD 220), but it is still at a development phase.[4,5] Gasification is a conversion process that involves partial oxidation at elevated temperature. It is intermediate between combustion and pyrolysis: in fact, oxygen (or air) is present but it is not enough for complete combustion. This process can start from carbonaceous feedstock such as biomass or coal and convert them into a gaseous energy carrier. The overall gasification process may be split into two main stages: the first is the pyrolysis stage, i.e., where oxygen is not present but temperature is high; typical pyrolysis reactions take place here; the second stage is the partial combustion, where

TABLE 3 Qualitative Comparison of Technology for Energy Conversion of Biomass

Process	Technology	Economics	Environment	Market Potential	Present Deployment
Combustion—heat	+++	€	+++	+++	+++
Combustion—electricity	++(+)	€€	++(+)	+++	++
Gasification	+(+)	€€€	+(++)	+++	(+)
Pyrolysis	(+)	€€€	(+++)	+++	(+)

Source: Adapted from European Biomass Industry Association[1] and Risoe National Laboratory.[6]
Note: +, low; +++, high; €, cheap; €€€, expensive.

oxygen is present and reacts with the pyrolysed biomass to release heat necessary for the process. In the latter stage, the actual gasification reactions take place, which consist of almost complete charcoal conversion into lighter gaseous products (e.g., carbon monoxide and hydrogen), through the chemical oxidizing action of oxygen, steam, and carbon dioxide: such gases are injected into the reactor near the partial combustion zone (normally, steam and carbon dioxide are mutually exclusive). Gasification reactions require temperature in excess of 800°C to minimize tar and maximize gas production. The gasification output gas, called "producer gas," is composed of hydrogen (18%–20%), carbon monoxide (18%–20%), carbon dioxide (8%–10%), methane (2%–3%), trace amounts of higher hydrocarbons like ethane and ethene, water, nitrogen (if air is used as oxidant agent), and various contaminants such as small char particles, ash, tars, and oils. The incondensable part of producer gas is called "syngas," and it represents the useful product of gasification. If air is used, syngas has a high heating value in the order of 4–7 MJ/m^3, which is exploitable for boiler, engine, and turbine operation, but, due to its low energy density, it is not suitable for pipeline transportation. If pure oxygen is used, the syngas high heating value almost doubles (approximately 10–18 MJ/m^3 high heating value). Such a syngas is suitable for limited pipeline distribution as well as for conversion to liquid fuels (e.g., methanol and gasoline). However, the most common technology is air gasification because it avoids the costs and the hazards of oxygen production and usage. With air gasification, the syngas efficiency, describing the energy content of the cold gas stream in relation to that of the input biomass stream, is on the order of 55%–85%, and typically 70%.

Comparison of Thermal Conversion Methods of Biomass

Table 3 reports a general overview on specific features of the conversion technologies analyzed here, showing the related advantages and drawbacks.

Biochemical Conversion of Biomass

Biochemical conversion of biomass refers to processes that decompose the original biomass into useful products. Commonly, the energy product is either in the liquid or in the gaseous form; hence, it is called "biofuel" or "biogas," respectively. Biofuels are very promising for the transportation sector, while biogas is used for electricity and heat production. Normally, biofuels are obtained from dedicated crops (e.g., biodiesel from seed oil), while biogas production results from concerns over environmental issues such as elimination of pollution, treatment of waste, and control of landfill greenhouse gas (GHG) emissions.

Biogas from Anaerobic Digestion

Biogas is produced most commonly by anaerobic digestion of biomass. Anaerobic digestion refers to the bacterial breakdown of organic materials in the absence of oxygen. This biochemical process produces a gas called biogas, principally composed of methane (30%–60% in volume) and carbon dioxide. Such a biogas can be converted to energy in the following ways:

- Biogas converted by conventional boilers for heating purposes at the production plant (house heating, district heating, industrial purposes).
- Biogas for CHP generation.
- Biogas and natural gas combinations and integration in the natural gas grid.
- Biogas upgraded and used as vehicle fuel in the transportation sector.
- Biogas utilization for hydrogen production and fuel cells.

An important production of biogas comes from landfills. Anaerobic digestion in landfills is brought about by the microbial decomposition of the organic matter in refuse. Landfill gas is, on average, 55% methane and 45% carbon dioxide. With waste generation increasing at a faster rate than economic growth, it makes sense to recover the energy from that stream, through thermal or fermentation processes.

Biofuels for Transport

A wide range of chemical processes may be employed to produce liquid fuels from biomass. Such fuels can find a very high level of acceptance by the market, thanks to the relatively easy adaptation to existing technologies (i.e., gasoline and diesel engines). The main potential biofuels are outlined below.

- Biodiesel is a methyl-ester produced from vegetable or animal oil to be used as alternative to conventional petroleum-derived diesel fuel. Compared to pure vegetable or animal oil, which can be used in adapted diesel engines as well, biodiesel presents lower viscosity and slightly HHV.
- Pure vegetable oil is produced from oil plants through pressing, extraction, or comparable procedures, crude or refined but chemically unmodified. Usually, it is compatible with existing diesel engines only if blended with conventional diesel fuel, at rates not higher than 5%–10% in volume. Higher rates may lead to emission and engine durability problems.
- Bioethanol is ethanol produced from biomass and/or the biodegradable fraction of waste. Bioethanol can be produced from any biological feedstock that contains appreciable amounts of sugar or other matter that can be converted into sugar, such as starch or cellulose. Also, lignocellulosic materials (wood and straw) can be used, but their processing into bioethanol is more expensive. Application to modified spark ignition engines is possible.
- Bio-ETBE (ethyl-tertio-butyl-ether) is ETBE produced on the basis of bioethanol. Bio-ETBE may be effectively used for enhancing the octane number of gasoline (blends with petrol gasoline).
- Biomethanol is methanol produced from biomass. Methanol can be produced from gasification syngas (a mixture of carbon monoxide and hydrogen) or wood dry distillation (old method with low methanol yields). Virtually all syngas for conventional methanol production is produced by steam reforming of natural gas into syngas. In the case of biomethanol, a biomass is gasified first to produce a syngas from which the biomethanol is produced. Application to spark ignition engines and fuel cells is possible. Compared to ethanol, methanol presents more serious handling issues, because it is corrosive and poisonous for human beings.
- Bio-MTBE (methyl-tertio-butyl-ether) is a fuel produced on the basis of biomethanol. It is suitable for blends with petrol gasoline.
- Biodimethylether (DME) is dimethylether produced from biomass. Bio-DME can be formed from syngas by means of oxygenate synthesis. It has emerged only recently as an automotive fuel option. Storage capabilities are similar to those of LPG. Application to spark ignition engines is possible.

Benefits from Biomass Energy

There is quite a wide consensus that, over the coming decades, modern biofuels will provide a substantial source of alternative energy. Nowadays, biomass already provides approximately 11%–14% of the world's primary energy consumption (data vary according to sources).

TABLE 4 Benefits in Reduction of GHG Emissions

+	Avoided mining of fossil resources
–	Emission from biomass production
+	Avoided fossil fuel transport (from producer to user)
–	Emission from biomass fuel transport (from producer to user)
+	Avoided fossil fuel utilization

Source: Risoe National Laboratory.[6]
Note: +, positive; –, neutral.

There are significant differences between industrialized and developing countries; in particular, in many developing countries, bioenergy is the main energy source, even if it is used in very low efficient applications (e.g., cooking stoves have an efficiency of about 5%–15%). Furthermore, inefficient biomass utilization is often associated with the increasing scarcity of hand-gathered wood, nutrient depletion, and the problems of deforestation and desertification.

One of the key drivers to bioenergy deployment is its positive environmental benefit regarding the global balance of GHG emissions. This is not a trivial matter, because biomass production and use are not entirely GHG neutral. In general terms, the GHG emission reduction as a result of employing biomass for energy is as reported in Table 4.

Since the energy cost associated with collection and transport of biomass is a significant portion, bioenergy is a decentralized energy option whose implementation presents positive impacts on rural development by creating business and employment opportunities. Jobs are created all along the bioenergy chain, from biomass production or procurement, to its transport, conversion, distribution, and marketing.

Bioenergy is a key factor for the transition to a more sustainable development.

Potential for CO_2 Emission Reduction

When biomass is used for energy production, the carbon contained in it is ultimately transformed into CO_2. In fact, such a biomass-derived CO_2 does not contribute to global warming, as it equals the CO_2 absorbed by the biomass during its growth; the relatively short time of such carbon cycle makes the biomass a carbon-neutral energy resource.

Abundant resources and favorable policies[7] enable bio-power to expand in Northern Europe (mostly cogeneration from wood residues), in the United States, and in countries producing sugar cane bagasse (e.g., Brazil).

In the short term, cofiring remains the most cost- effective use of biomass for power generation, along with small-scale, off-grid use. In the mid-long term, gasification plants and biorefineries for biofuel production could expand significantly (mainly ethanol, lignocellulosic ethanol, biodiesel). International Energy Agency projections suggest that the biomass share in electricity production may increase from the current 1.3% to some 3%–5% by 2050, depending on assumptions.[8] This is a small contribution compared to the estimated total biomass potential, but biomass are also used for heat generation and to produce fuels for transport.

Today, biomass supplies some 50 EJ/yr (1 EJ = 10^{18} joules [J] = 10^{15} kilojoules [kJ] = 24 million tons of oil equivalent [Mtoe]) globally, which represents 10% of global annual primary energy consumption (Figure 2). This is mostly traditional biomass used for cooking and heating.

Based on this diverse range of feedstocks, the technical potential for biomass is estimated in the literature to be possibly as high as 1500 EJ/yr by 2050, although most biomass supply scenarios that take into account sustainability constraints indicate an annual potential of between 200 and 500 EJ/yr (excluding aquatic biomass). Forestry and agricultural residues and other organic wastes (including municipal solid waste) would provide between 50 and 150 EJ/yr, while the remainder would come from energy crops, surplus forest growth, and increased agricultural productivity.

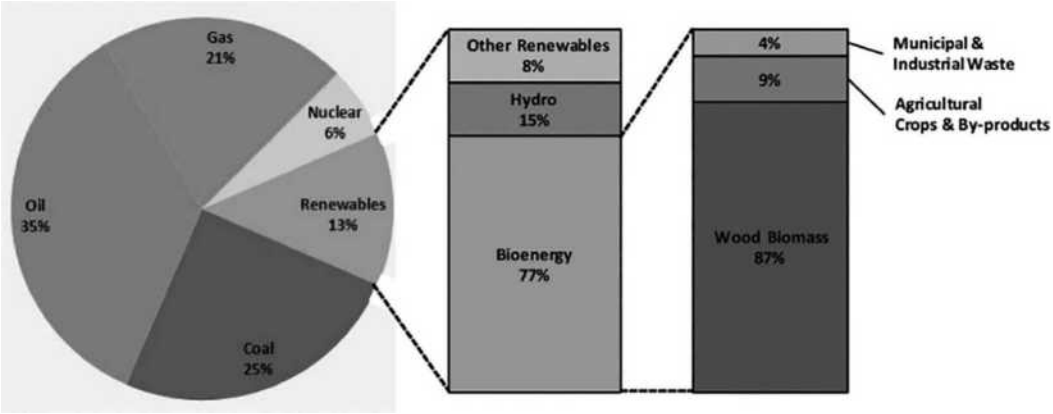


FIGURE 2 Share of bioenergy in the world primary energy mix. (*Source:* International Energy Agency.[9])

Projected world primary energy demand by 2050 is expected to be in the range of 600 to 1000 EJ, compared to about 500 EJ in 2008.[9] Scenarios looking at the penetration of different low-carbon energy sources indicate that future demand for bioenergy could be up to 250 EJ/yr. This projected demand falls well within the sustainable supply potential estimate, so it is reasonable to assume that biomass could sustainably contribute between a quarter and a third of the future global energy mix. Whatever is actually realized will depend on the cost competitiveness of bioenergy and on future policy frameworks, such as GHG emission reduction targets.

Given the CO_2-neutral nature of biomass, and assuming that biomass will primarily substitute fossil fuels, the potential for reduction in CO_2 emission in 2050 can then be estimated as the same figure (i.e., around 20%–30% of anthropogenic CO_2), compared to a business-as-usual scenario. Definitely, biomass will play a determinant role towards a CO_2-free development.

Conclusions

Biomass refers to a very wide range of substances produced by biological processes. In the energy field, special focus has been and will be placed on vegetable biomass, such as wood and agricultural by-products, because of the energy potential as well as economic and environmental benefits. Size and humidity standardization of biomass is a necessary step to make it suitable for effective domestic and industrial exploitation: chips, briquettes, and pellets are modern examples of standard solid fuels.

Biomass can be converted into energy in three pathways: combustion, thermochemical processing, and biochemical processing. The combustion of solid biomass for the production of heat or electricity and heat is the most viable technology, while pyrolysis and gasification still face economic and reliability issues. Among biochemical processes, anaerobic digestion is often used to reduce the environmental impact of hazardous waste and landfills. Biochemical processes are also concerned with the conversion of biomass into useful fuels for transportation, such as biodiesel, bioethanol, and biomethanol. All of them can effectively contribute to the transition to a more sustainable transportation system at zero GHG emissions.

Biomass represents a viable option for green energy resources of the 21st century.

References

1. European Biomass Industry Association, available at http:// www.eubia.org/.
2. DOE Biomass Research and Development Initiative, available at http://www.bioproducts-bioenergy.gov/.

3. Overend, R.P.; Milne, T.A.; Mudge, L.K. *Fundamentals of Thermochemical Biomass Conversion;* Elsevier Applied Science Publishers Ltd.: New York, 1985.
4. Bridgewater, A.V. The technical and economic feasibility of biomass gasification for power generation. Fuel J. **1995**, *74* (6–8), 557–564.
5. Franco, A.; Giannini, N. Perspectives for the use of biomass as fuel in combined cycle power plants. Int. J. Thermal Sci. **2005**, *44*, (2), 163–177.
6. Risoe National Laboratory, Denmark, available at http:// www.risoe.dk/.
7. Biomass for Power Generation and CHP, International Energy Agency (IEA), Energy Technology Essentials, ETE03, Jan 2007.
8. World Energy Outlook; International Energy Agency (IEA), 2002.
9. Bioenergy—A Sustainable and Reliable Energy Source, International Energy Agency (IEA), available at http://www.ieabioenergy.com, 2009.

30

Nematodes: Biological Control

Simon Gowen

Introduction

Many pathogens and predators of nematodes are known but few have the necessary characteristics of specificity, mobility, or speed of colonization to have a significant influence on a pest population. Attempts at their commercial exploitation as field treatments have not been successful largely because of their inconsistency. Understanding the subtleties associated with the deployment of biocontrol agents will require considerable research effort. Additionally, the recommended rates of application and the formulation on suitable carriers and nutrient sources poses a problem in practicability and in the interpretation of the biological processes involved.

Contemporary research has shown that natural control does exist and that in certain crop/nematode/pathogen situations nematode populations will decline as they are attacked by components of the soil microflora. Soils where this occurs are known as suppressive, but well-documented examples of naturally occurring suppressiveness to particular nematode pests are uncommon.

During the life of a crop the population densities of many of the serious nematode pests can increase by 1000fold. Economic damage may result from initial population densities of one nematode per gram of soil. To be effective therefore a biocontrol agent must have an impact on the numbers of nematodes that would invade a host and not simply eliminate the surplus individuals that may never locate or invade a root. This being so, those pathogens and predators that are relatively unspecific (trapping, ingesting, or parasitizing all types of free-living nematodes in soil) may be considered less promising than those that parasitize specific pests.

Significant progress has been made in the recognition and deployment of such microorganisms parasitic on some of the species of sedentary nematodes such as the root-knot nematodes, *Meloidogyne* spp., and some of the cyst nematodes, *Heterodera* spp., and *Globodera* spp.

Root-knot and cyst nematodes produce eggs either in clusters on roots or contained within or attached to the cuticle of the female nematode. Biocontrol agents that prevent these nematodes from reproducing may have more impact from an epidemiological point of view than those that kill the free-living individuals in the soil.

Biocontrol Agents Specific to Certain Nematode Pests

Verticillium chlamydosporium is a facultative, soil-dwelling fungus that parasitizes eggs in egg masses exposed on the root surface. Under the right conditions, such fungi will have a significant effect on nematode populations. The efficacy of *V. chlamydosporium* is partly dependent on its root colonizing ability; this can vary according to the plant host. Skill is required in selecting crops that support and/or increase the root colonization by the fungus but are also less favored hosts of root-knot nematodes. *V. chlamydosporium* may be less effective when it is deployed with plants that are highly susceptible and large galls are produced in response to the nematode infection. In such cases, many egg masses may not be exposed on the root surface and so escape infection.

Paecilomyces lilacinus is another fungus commonly found infecting the eggs of sedentary nematodes such as the root-knot and the cyst nematodes, and, like *V. chlamydosporium* being relatively easy to produce on defined growth media, has good potential for commercial development.

Pasteuria penetrans, an obligate bacterial parasite of root-knot nematodes begins its life cycle on free-living juveniles in the soil. Spores attach to the juveniles as they move in search of host roots. Parasitic development begins after the nematode enters a root and continues in synchrony with that of its host. The nematode eventually is overcome by its parasite; it fails to produce eggs; and its body, filled with the spores of the bacterium, eventually ruptures releasing spores into the soil.

The efficiency of *P. penetrans* as a biocontrol agent of root-knot nematodes depends on the concentrations of spores in the soil, the chances of contact with the juvenile stage, and the specificity of the particular *P. penetrans* population. Commercial success will depend therefore on finding techniques for mass-producing the bacterium and on developing populations with a broad spectrum of pathogenicity.

Other *Pastearia* species parasitic on some sedentary *(Heterodera)* and migratory *(Pratylenchus)* species have been described.

Biological control agents such as *V. chlamydosporium, P. lilacinus*, and *P. penetrans* could provide an adequate replacement for nematicides in some cropping systems but the lack of immediate effects, such as are provided by nematicide or fumigant treatments, is a disadvantage. Protection is normally needed in the early stages of plant growth such as in nursery beds. In this situation, integration with other practices such as nematicides, rotation, solarization, and mulches is necessary.

There are several reports of the successful deployment of these biocontrol agents. Small field plots treated once with *P. penetrans* spores (produced by an in vivo system) caused a decline in numbers of root-knot nematodes and increases in yield over a series of crop cycles using root-knot nematode susceptible crops. In other locations, where treatments with *P. penetrans* were combined with *V. chlamydosporium*, organic manures and grass mulches showed similar declines in nematode populations. These two organisms acted against root-knot nematodes in a complementary fashion. As part of this strategy, root systems containing spore-filled cadavers were deliberately left to disintegrate in the soil after each crop. No field treatments were effective after only one crop indicating that some crop loss must be expected during the development of suppressiveness. *P. penetrans* was also effective when used in combination with a nematicide in permanent beds within a plastic polytunnel. Better control of root-knot nematodes was achieved if the biocontrol agent was combined with other control strategies. With such treatments, beneficial effects may develop over one crop cycle.

The chlamydospores of *V. chlamydosporium* do not have the persistence of the spores of *P. penetrans*, which can remain viable for many years.

Nonspecific Biocontrol Agents

There is a long history of interest in the fungi that trap nematodes in soil such as species of *Arthrobotrys*. These are commonly found in all soils but despite much research effort the problems of the unreliability of soil applications have not been solved and none have become established as successful commercial products.

There are several rhizosphere colonists that have potential for alleviating nematode damage. The precise mechanisms are not clear. Some produce toxins but others may affect root exudation and thus indirectly the attractiveness of roots to nematodes. Experiments have shown that strains of *Pseudomonas fluorescens* can reduce root invasion by different plant parasites but as with the trapping fungi, poor consistency hinders successful development of these microorganisms as commercial products.

Future Prospects

Recently, the nematicidal (and insecticidal) effects of the toxins produced by the bacteria associated with entomopathogenic nematodes (*Photorhabdus* spp., *Xenorhabdus* spp., and *Pseudomonas oryzihabitans*) have been demonstrated.

Success in the commercial development of biocontrol agents does appear promising with those microorganisms that can be formulated as a standard product with proven reliability; others may have a future as single treatment introductions in the more intensively managed protected cropping systems but commercialization may be difficult.

Research is still needed to develop reliable methods of production, formulation, and application. The challenge is to provide a sufficient duration of protection. Such treatments will need to be part of a package of control measures.

Bibliography

1. Aalten, P.M.; Vitour, D.; Blanvillain, D.; Gowen, S.R.; Sutra, L. Effect of rhizosphere fluorescent *Pseudomonas* strains on plant parasitic nematodes *Radopholus similis* and *meloidogyne* spp. Letters Appl. Microbiol. **1998**, *27*, 357–361.
2. Bourne, M.; Kerry, B.R.; De Leij, F.A.A.M. The importance of the host plant on the interaction between root-knot nematodes (*Meloidogyne* spp.) and the nematophagous fungus, *Verticillium chlamydosporium* Goddard. Biocon. SciTech. **1996**, *6*, 539–548.
3. Crump, D.J. A method for assessing the natural control of cyst nematode populations. Nematologica **1987**, *33*, 232–243.
4. Gowen, S.R.; Bala, G.; Madulu, J.; Mwageni, W.; Trivino,C.T. In *Field Evaluation of Pasteuria penetrans for the Management of Root-Knot Nematodes*, The 1998 Brighton Conference on Pests and Diseases, **1998**, *3*, 755–760.
5. Kerry, B.R.; Jaffee, B.A. Fungi as Biocontrol Agents for Plant Parasitic Nematodes. In *The Mycota IV Environmental and Microbial Relationships*; Wicklow, D. T., Soderstrom, B.E., Eds.; Springer-Verlag: Berlin, Heidelberg, 1997; 204–218.
6. Samaliev, H.Y.; Andreoglou, F.I.; Elawad, S.A.; Hague, N.G.M.; Gowen, S.R. The Nematicidal Effects of the Bacteria *Pseudomonas oryzihabitans* and *Xenorhabdus Nematophilus* on the Root-Knot Nematode, *Meloidogyne javanica*. Nematology, *in press.*
7. Stirling, G.R. Biological Control of Plant Parasitic Nematodes. CAB International: Wallingford, U.K., 1991; 282.
8. Tzortzakakis, E.A.; Channer, A.G. de R.; Gowen, S.R.; Ahmed, R. Studies on the potential use of *Pasteuria penetrans* as a biocontrol agent of root-knot nematodes (*Meloidogyne spp.*). Plant Pathol. **1997**, *46*, 44–55.
9. Tzortzakakis, E.A.; Gowen, S.R. Evaluation of *Pasteuria penetrans* alone and in combination with oxamyl, plant re sistance and solarization for control of *Meloidogyne* spp. on vegetables grown in greenhouses in Crete. Crop Prot. **1994**, *13*, 455–462.
10. Weibelzahl-Fulton, E.; Dickson, D.W.; Whitty, E.B. Suppression of *Meloidogyne incognita* and *M. javanica* by *Pasteuria penetrans* in field soil. J. Nematol. **1996**, *28*, 43–49.

VII

PRO: Basic Environmental Processes

31

Agroforestry: Water Use Efficiency

James R. Brandle,
Laurie Hodges,
and Xinhua Zhou

Introduction

Agroforestry is the intentional integration of trees and shrubs into agricultural systems. Windbreaks, riparian forest buffers, alley-cropping, silvopastoral grazing systems, and forest farming are the primary agroforestry practices found in temperate regions of North America.[1] Placing trees and shrubs on the landscape changes the surface energy balance, influences the surrounding microclimate, and has the potential to alter water use and productivity of adjacent crops.[2,3]

In agricultural systems, water is often the major factor limiting growth. When water availability is limited as a result of limited supply or high cost, its efficient use becomes critical to successful production systems. For example, proper irrigation at the appropriate stage of crop development minimizes pumping costs and increases yield; reducing soil tillage conserves soil water and may enhance yield, and reducing surface runoff or trapping snow improves soil water storage for future crop use. These water conservation efforts contribute to the efficient use of available water and are determined primarily by management practices. In contrast, Tanner and Sinclair[4] distinguish between the efficient use of water and water use efficiency (WUE). WUE is primarily a function of physiological responses of plants to environmental conditions. This review focuses on WUE defined as the amount of biomass (or grain) produced per unit of land area for each unit of water consumed.[4]

Soil water may be consumed by evaporation from the soil surface or by the transport of water through the plant and subsequent evaporation from the leaf surface. The rate of water consumption is determined by the microclimate of the crop. Because agroforestry practices alter the microclimate of adjacent fields, they affect WUE of plants growing in those fields.

Discussion

Windbreaks, riparian forest buffers or alley-cropping systems are the practices most likely to be integrated into crop production systems. In all three practices, trees and shrubs tend to be arranged in narrow barriers adjacent to the crop field. Microclimate responses downwind of any of these types of barriers are similar and the following discussion applies to all three types of barriers. As wind approaches these barriers, it is diverted up and over the barrier creating two zones of protection, a larger zone to the lee of the barrier (the side away from the wind) and a smaller zone on the windward side of the barrier. In these zones, wind speed is reduced and turbulence and eddy structure in the vicinity of the barrier are altered. As a result of these changes, the transfer coefficients for heat and mass between the crop and the atmosphere are altered; the gradients of temperature, humidity, and carbon dioxide concentration above the soil and canopy are changed;[5] and the plant processes of transpiration and photosynthesis are altered.[6]

McNaughton[5] defined two regions within the leeward zone of protection: the *quiet zone,* extending from the top of the barrier down to a point in the field located approximately 8*H* leeward (*H* is the height of the barrier) and a *wake zone,* lying beyond the quiet zone and extending from approximately 8*H* to a distance of 20*H* to 25*H* from the barrier. Within the quiet zone where turbulence is reduced, we expect conditions to be such that the canopy is "uncoupled" from the atmospheric conditions above the sheltered zone, while in the wake zone where turbulence is increased, we expect the canopy to become more strongly "coupled" to the atmosphere above. In both locations we would expect the rates of photosynthesis and transpiration to be altered and WUE to change.

The magnitude of change in wind speed, as well as the extent of microclimate modifications within the quiet and wake zones, are largely determined by the structure of the windbreak or barrier and the underlying meteorological conditions. Structure refers to the amounts of solid material and open space and their arrangement within the barrier. Dense barriers, for example, multiple rows of conifers, generally result in greater wind speed reduction but more turbulence. More porous barriers, for example, single rows of deciduous species, result in less wind speed reduction but also less turbulence. The downwind extent of the protected area is generally greater for more porous barriers. As a result, narrow, less dense barriers (40%–60% density) are typically used to protect crop fields.

The overall influence of wind protection on plant water relations is complex and linked to temperature, humidity, wind speed, and other meteorological conditions found in the protected zone, the amount of available soil water, crop size, and stage of development.[2,3,7] Until recently, the major effect of wind protection and its influence on crop growth and yield were assumed to be due primarily to soil water conservation and reduced water stress of sheltered plants.[8,9] There is little question that the evaporation rate from bare soil is reduced in the protected zone.[3] However, the effect of reduced wind speed on transpiration, evaporation from the plant canopy, and overall plant water status is less clear.[2,3,7]

According to Grace,[9] transpiration rates may increase, decrease, or remain unaffected by wind protection depending on wind speed, atmospheric resistance, and saturation vapor pressure deficit. Cleugh[3] suggests that as stomatal resistance increases, evaporation from the canopy may actually be increased with a reduction in wind speed. When stomatal resistance is high and water is limited, stomatal resistance controls the rate of evaporation from the leaf surface, not the amount of turbulence. Under these conditions a decrease in wind speed and turbulent mixing may increase the potential for evaporation from the leaf surface.[3]

Evaporation from the leaf surface consists of two phases, an energy driven phase and a diffusion driven phase. Movement of water through the plant and out the stomata is driven by the water potential gradient within the plant. This gradient is influenced by the plant's energy balance. On the lee side of the buffer, reduced wind speed and turbulent mixing lead to increases in leaf temperature and transpiration to meet the increased energy load on the plant. If adequate water is available, it is moved through the plant to the leaf surface and the potential for evaporation from the leaf surface is increased. If water is limited, the stomata partially or completely close, transpiration is reduced, and evaporation from the leaf surface declines.

In contrast, movement of water vapor across the leaf boundary layer is controlled by the vapor pressure gradient and the thickness of the leaf boundary layer. As windspeed decreases, the thickness of this boundary layer increases, the vapor pressure gradient decreases, and the rate of evaporation from the leaf surface decreases. The relative magnitude of the two processes determines whether or not transpiration and subsequent evaporation from the canopy are increased, decreased, or remain unchanged.[7,9,10]

While these theoretical considerations are important in understanding the process, several studies[11–13] have demonstrated a good correlation between wind protection, conservation of soil water, and enhanced crop yield. Even so, the effect of wind protection on WUE is neither constant throughout the growing period[7] nor is it consistent over varying meteorological conditions.

Agroforestry practices impact the water relations of the crop by affecting the loss of water through damaged leaves. On soils subject to wind erosion, windbreaks or other agroforestry buffers provide significant reductions in the amount of wind blown soil and subsequent abrasion of plant parts and cuticular damage.[9,14] Loss of cuticular integrity or direct tearing of the leaves[15] reduces the ability of the plant to control water loss.

Agroforestry buffers have a direct effect on the distribution of precipitation, both rain and snow. In the case of snow, a porous barrier will result in a more uniform distribution of snow across the field, providing additional soil water for the crop.[16] In the case of rain, the barrier has minimal influence on the distribution of precipitation across the field; however, in the area immediately adjacent to the barrier a rain shadow may occur on the leeward side. On the windward side, the barrier may lead to slightly higher levels of measured precipitation at or near the base of the trees due to increased stem flow or dripping from the canopy.

Trees and shrubs used in agroforestry practices also consume a portion of the available water. In the area immediately adjacent to the barrier, competition for water between the crop and the barrier has a negative impact on yield. These same areas are also subject to some degree of shading depending on the orientation of the barrier. These changes in radiation load influence the energy balance and thus the growth and development of the crop and the utilization of water.[2]

Summary

In summary, agroforestry practices such as windbreaks, riparian forest buffers and alley-cropping systems generally improve both the efficient use of water by the agricultural system and the WUE of the individual crop. In the case of efficient water use, the evidence is clear. In the case of crop WUE, the evidence leaves some unanswered questions. How do we account for the varied crop yield responses reported in the literature? In many cases yields are increased but no clear relationship to crop water budget is shown. In other cases crop yield response is minimal. Under what meteorological conditions are the effects of agroforestry practices most valuable to water balance questions? Final crop yield is a integration of the environmental conditions over the entire growing season. Many different combinations of environmental conditions may result in similar plant responses. How do we address the numerous combinations of plant stress and plant growth to determine “a response” to wind protection? To answer many of these questions it will be necessary to intensify the numerical modeling methods developed by Wilson[17] and Wang and Takle.[18] With a better model to describe the turbulence fields and the transport of water, heat, and carbon dioxide as influenced by agroforestry practices, it should be possible to assess the numerous combinations of environmental factors influencing crop growth in these systems.

References

1. Lassoie, J.P.; Buck, L.E. Development of agroforestry as an integrated land use management strategy. In *North American Agroforestry: An Integrated Science and Practice*; Garrett, H.E., Rietveld, W.J., Fisher, R.F., Eds.; American Society of Agronomy, Inc.: Madison, WI, 2000; 1–29.
2. Brandle, J.R.; Hodges, L.; Wight, B. Windbreak practices. In *North American Agroforestry: An Integrated Science and Practice;* Garrett, H.E., Rietveld, W.J., Fisher, R.F., Eds.; American Society of Agronomy, Inc.: Madison, WI, 2000; 79–118.
3. Cleugh, H.A. Effects of windbreaks on airflow, microclimates and crop yields. Agrofor. Syst. **1998**, *41*, 55–84.
4. Tanner, C.B.; Sinclair, T.R. Efficient water use in crop production: research or re-search? In *Limitations to Efficient Water Use in Crop Production*; Taylor, H.M., Jordan, W.R., Sinclair, T.R., Eds.; American Society of Agronomy, Inc.: Madison, WI, 1983; 1–27.

5. McNaughton, K.G. Effects of windbreaks on turbulent transport and microclimate. Agric. Ecosyst. Environ. **1988**, *22/23,* 17–39.
6. Grace, J. Some effects of wind on plants. In *Plants and Their Atmospheric Environment*; Grace, J., Ford, E.D., Jarvis, P.G., Eds.; Blackwell Scientific Publications: Oxford, 1981; 31–56.
7. Nuberg, I.K. Effect of shelter on temperate crops: a review to define research for Australian conditions. Agrofor. Syst. **1998**, *41*, 3–34.
8. Caborn, J.M. *Shelterbelts and microclimate,* Forestry Commission Bulletin No. 29; Her Majesty's Stationery Office: Edinburgh, 1957; 135 pp.
9. Grace, J. Plant response to wind. Agric. Ecosyst. Environ. **1988**, *22/23,* 71–88.
10. Thornley, J.H.M.; Johnson, I.R. *Plant and Crop Modeling: A Mathematical Approach to Plant and Crop Physiology*; Clarendon Press: New York, 1990; 669 pp.
11. Song, Z.M.; Wei, L. The correlation between windbreak influenced climate and crop yield. In *Agroforestry Systems in China;* Zhu, Z.H., Cai, M.T., Wang, S.J., Jiang, Y.X., Eds.; International Development Research Centre (IDRC, Canada), Regional Office for Southeast and East Asia, published jointly with the Chinese Academy of Forestry: Singapore, 1991; 21–115.
12. Wu, Y.Y.; Dalmacio, R.V. Energy balance, water use and wheat yield in a Paulownia-wheat intercropped field. In *Agroforestry Systems in China;* Zhu, Z.H., Cai, M.T., Wang, S.J., Jiang, Y.X., Eds.; International Development Research Centre (IDRC, Canada), Regional Office for Southeast and East Asia, published jointly with the Chinese Academy of Forestry: Singapore, 1991; 54–65.
13. Huxley, P.A.; Pinney, A.; Akunda, E.; Muraya, P. A tree/crop interface orientation experiment with a *Grevillea robusta* hedgerow and maize. Agrofor. Syst. **1994**, 26, 23–45.
14. Kort, J. Benefits of windbreaks to field and forage crops. Agric. Ecosyst. Environ. **1988**, *22/23,* 165–190.
15. Miller, J.M.; Böhm, M.; Cleugh, H.A. *Direct Mechanical Effects of Wind on Selected Crops: A Review*, Technical Report Number 67; CSIRO Center for Environmental Mechanics: Canberra, Australia, 1995; 68 pp.
16. Scholten, H. Snow distribution on crop fields. Agric. Ecosyst. Environ. **1988**, *22/23,* 363–380.
17. Wilson, J.D. Numerical studies of flow through a windbreak. J. Wind Eng. Ind. Aerodyn. **1985**, *21,* 119–154.
18. Wang, H.; Takle, E.S. A numerical simulation of boundary-layer flows near shelterbelts. Boundary-Layer Meteorol. **1995**, *75*, 141–173.

32

Bacterial Pest Control

David N. Ferro

Introduction

Although many genera of bacteria are found to be associated with insects—such as *Clostridium, Strategus, Pseudomonas, Proteus, Diplococcus, Serratia, Bacillus,* and *Enterobacter*—only *Bacillus* and *Serratia* represent agents that cause suppression of insect populations, i.e., that perform as biological control agents. Bacteria are the most widely used microbial agents for controlling insect pests. Some species of *Bacillus* and *Serratia* kill by replicating within the host, while strains of *Bacillus thuringiensis* produce protein toxins that kill soon after being ingested. Bacteria that replicate within their hosts and that persist in the environment by maintaining an infection cycle are biological control agents in the traditional sense. However, products of *B. thuringiensis* that produce toxins that kill insect pests and are applied the way an insecticide would be applied are often not considered to be biological control agents. Bacteria in the genera *Photorhabdus* and *Xe-norhabdus* are symbiotic with nematodes in the families Heterorhabditidae and Steinernematidiae, respectively. The nematodes serve as vectors that mechanically penetrate into the insect hemocoel and deposit the bacteria. The bacteria then replicate and kill the host quickly by causing septicemia. The only commercially available bacterial products are from strains of *B. popillae*, *B. thuringiensis*, and *Serratia entomophila*.

Paenibacillus (Formerly Bacillus) Popilliae (Dutky)

Milky disease was first observed in Japanese beetle larvae (grubs) in New Jersey in 1933. *P. popilliae* is an obligate pathogen of larvae in the family Scarabaeidae, as it is only found associated with its host or in the soil surrounding its host. *P. popilliae* and *Paenibacillus lentimorbus* (Dutky) both cause milky disease of scarab beetles; however, most discussions of milky disease refer to strains of *P. popilliae*. *P. popilliae* produces a crystal or parasporal body, which allows it to survive for many years in the soil in the absence of its host. Although there are dozens of strains of *P. popilliae* that infect scarab hosts, only *P. popilliae* has been used commercially as a biological control agent of the Japanese beetle, *Popillia japonica* (Newman), a major pest of turf.

The term "milky disease" describes the advanced stages of infection in scarab larvae where the host is turned a milky white by the build-up of *Bacillus* spores in the hemolymph. The infection process begins with the scarab larvae ingesting spores while feeding on roots and organic matter in the soil. The spores then undergo germination and outgrowth in the cells of the lumen of the alimentary canal. The vegetative rods penetrate the epithelial cells of the midgut, and then move into the hemolymph where they multiply and sporulate. Death often occurs a month or more after ingestion. It is unclear what the role of the proteinaceous parasporal body is in the infection process.

Culture and Control

Many attempts have been made to rear *P. popillae* on an artificial diet. Even though spores and vegetative rods from field-collected larvae can be plated on agar media, the inability of the milky disease bacteria to grow and sporulate on standard microbiological media has made it extremely costly to produce for commercial purposes. Products, to date, are made from milky larvae, primarily from naturally infected larvae collected from the field.

The spores are formulated on talc and contain 10^8 spores/g of powder. The powder is applied at about 20 kg/ha using a fertilizer spreader or by punching holes in the soil and adding bacteria. Infection can occur in all three larval stages. For optimal replication to occur, soil temperatures need to exceed 20°C. Large overwintered larvae usually pupate before soil temperatures are high enough in late spring. For this reason, applications are targeted against small larvae late in the summer when the small larvae are actively feeding near the soil surface. Control seems to be greatest when larval densities exceed 300/m^2; however, economic losses in turf occur at densities above 100/m^2. Unless a more virulent strain is found or a more cost-effective way to produce spores is developed, the use of this bacterium is likely to be restricted to lawns and playing fields that can tolerate higher densities of larvae.

Serratia Entomophila (Grimont et al.)

Amber disease of the New Zealand grass grub *Costelytra zealandica* (White) is a chronic infection of the larval gut caused by *S. entomophila*. This disease was first observed in New Zealand in 1981. Following ingestion of bacterial cells while feeding on grass roots, the bacteria adhere to the foregut and multiply in the region of the cardiac valve; the larvae cease feeding after 2–5 days and become amber colored due to clearance of the gut. Death does not occur until 1–3 mo after ingestion. As the disease progresses, the larvae become shrunken due to a general degradation of the fat cells. Invasion of the hemocoel does not occur until late stages of the disease, when general septicemia is accompanied by death of the insect.

Culture and Control

S. entomophila is produced in large fermentors as nonsporeforming bacteria to be applied as a live microbial pesticide. Recently, the Industrial Processing Division of DSIR, New Zealand produced 4×10^{10} bacteria/mL, and field trials have shown that $>4 \times 10^{13}$/ha are needed for control. The problem with using live bacteria (vs. spores) is the difficulty of maintaining viability on the shelf and in the field prior to ingestion. Currently, refrigerated product can be kept for only 3 mo.

Grass grub larvae live in the soil as pests of low-value grasslands. Because *S. entomophila* is applied as live bacteria rather than as spores, it is more vulnerable to UV light and desiccation. For this reason, it is important to place the formulated material 2–5 cm below the soil surface using a subsurface applicator, such as a modified seed drill. This approach allows for 90% survival of the bacteria. Bacteria applied in this way quickly start an epizootic, which then spreads through the grass grub population.

Bacillus Thuringiensis (Berliner)

B. thuringiensis is a spore-forming bacterium that produces a parasporal crystal (protein delta-endotoxin). After the susceptible insect larva ingests the endotoxin, in the absence or presence of the spore, the crystal is solubilized and activated by alkaline (pH 10.5) gut proteases. The toxic subunits bind to receptor sites on the midgut epithelium within minutes of ingestion. This is quickly followed by lysis of these cells, causing a cessation of feeding within 10–15 min of ingestion. Although the spores pass into the hemocoel through pores in the epithelium of the midgut, it is the starvation in conjunction with infection that kills the insect. The toxins from these bacteria are formulated in much the same way as a synthetic toxin, and do not cause an epizootic.

There are several subspecies (= strains) of *B. thuringiensis* based on the serotype of flagellar antigens, and these subspecies produce different endotoxins, or at least different amounts of endotoxins that are relatively host specific. For example, *B.t. israelensis* is effective against Nematocera (Diptera) larvae such as mosquito larvae, *B.t. kurstaki* against Lepidoptera, *B.t. aizawai* against Lepidoptera, and *B.t. tenebrionis* against Chrysomelidae (Coleoptera). Notation for the gene that encodes for the toxin is in lowercase; for example, Cry3A gene regulates the production of the Cry3A toxin. Table 1 includes a list of some of the subspecies and toxins they produce. Because these bacteria are so host-specific, they can be quickly incorporated into a pest management program in which biological control agents are an integral component.

Culture and Control

B. thuringiensis can be produced in large quantities using commercial fermentors. Formulations can be applied to foliage or other larval substrates in the same manner as most insecticides. However, several operative factors affect the effectiveness of these bacterial agents.

B.t.s are most effective against early instars (Table 2). Their effectiveness is very dependent upon ambient temperatures; the protein endotoxin is not very persistent; thorough coverage of foliage is necessary; and they are host-specific. This host specificity allows for control of the target pest without killing other insect biological control agents; however, in many cropping systems, there is a complex of insect pests

TABLE 1 *B. thuringiensis* Subspecies and Crystal Protein Toxins

Crystal protein	*B.t.* Subspecies			
	B.t aizawai	*B.t. kurstaki*	*B.t. tenebrionis*	*B.t. israelensis*
CrylAa	*	*		
CrylAb	*	*		
Cry1Ac		*		
CrylC	*			
Cry ID	*			
Cry2A		*		
Cry2B		*		
Cry3A			*	*
Cry4A				*
Cry4B				*
Cry4C				*
Cry4D				*
CytA				*

TABLE 2 Stage-Specific Larval Mortality for the Colorado Potato Beetle Fed Foliage Treated with *B. thuringiensis san diego* (= *tenebrionis*)

	LC50	Larval	95% CI	
Larval stage	(mg/l)	weight (mg)	Lower	Upper
Early 1st instar	2.03	1.0	1.46	2.60
Late 1st instar	3.92	2.3	2.02	6.27
Early 2nd instar	4.35	4.0	3.30	5.56
Late 2nd instar	14.45	7.8	10.75	19.50
Early 3rd instar	14.86	15.6	9.95	20.48

and often these need to be controlled at the same time, which may require using the *B.t.* product with a synthetic insecticide, if natural controls fail. Novel ways have been developed to deliver the toxin for ingestion by the pest.

One of the genes that control the production of the toxin has been inserted into *Pseudomonas fluorescens.* After the fermentation has been completed, the broth is chemically treated and heated to kill the bacteria. During this process, the protein toxin becomes encapsulated by the bacterial cell wall. The encapsulation process appears to protect the toxin from degradation in the field, making it more persistent. Several genes have also been inserted into plants that express the toxin in its tissues. In the case of potatoes, the transgenic plants are highly resistant to the Colorado potato beetle, which has considerably reduced the insecticide load on potatoes.

Potential Biological Control AGENTS

Bacillus sphaericus (Neide) has been shown to be toxic only to larvae of culicid Diptera mosquitoes. This bacterium can be easily produced via fermentation. Insecticidal activity is due to crystalline toxins associated with the cell wall. The toxin is released by digestion after the host insect has consumed the bacteria. *B. alvei* and *B. brevis* are infectious for larvae of several mosquito species. There is no evidence that these species are significant biocontrol agents. The success of these bacteria in the field is likely to be dependent on selection of strains that are more virulent and that can persist in a range of aquatic environments.

Bibliography

1. Crickmore, N.; Zeigler, D.R.; Feitelson, J.; Schnepf, E.; Van Rie, J.; Lereclus, D.; Baum, J.; Dean, D.H. Revision of the nomenclature for the *Bacillus thuringiensis* pesticidal crystal proteins. Am. Soc. Micro. **1998**, *62,* 807–813.
2. Glare, T.R.; Jackson, R.A. *Use of Pathogens in Scarab Pest Management*; Intercept: Andover, England, 1992; 43–61, 179–198.
3. Jackson, T.A.; Huger, A.M.; Glare, T.R. Pathology of amber disease in the New Zealand grass grub *Costelytra zealandica* (Coleoptera: Scarabaeidae). J. Invert. Pathol. **1993**, *61,* 123–130.
4. Tanada, Y.; Kaya, H.K. *Insect Pathology***;** Academic Press: New York, 1993; 83–146.
5. Van Driesche, R.G.; Bellows, T.S., Jr. *Biological Control;* Chapman and Hall: New York, 1996.

33

Bioaccumulation

Tomaz Langenbach

Molecular Properties for Bioaccumulation

Bioaccumulation occurs only with molecules with low degradability and correlates with their grade of lipophilicity.[1] Organic substances with main bonds of aliphatic and aromatic C—C, C—H, and C—Cl (or other halogens) are predominantly nonpolar molecules (Lipophilic) with low water solubility and high stability. They are less susceptible to chemical reactions of hydrolysis, oxidation, and enzymatic attack.[2] On the other hand, bonds with different functional groups with O, P, N, and other elements turn molecules more polar, soluble, and degraded more easily. Bioaccumulation can occur with molecules between 100 and 600 units of molecular weight with the maximum of 350.[3] Probably this is related to membrane permeability capacity.

A common feature of bioaccumulation is the molecular stability of lipophilic organic substances and the nondegradability of heavy metals. The severity of heavy metals is due to many factors.

1. Metals with Hg, Cd, Zn, Cu, and Pb are the most toxic and most studied types followed by metals containing Ni, Al, As, Cr, and other elements.[4] Bioaccumulation can also occur also with essential metals such as Fe, Zn, Cu, Mo, Na, and Ca.
2. Speciation is the anions or other components that constitute the heavy metal molecules. This is important in defining solubility that, for example, is high for sulfate and low for sulphide.[5] Heavy metals bound to organic molecules such as methyl, ethyl or other aliphatic or aryl groups increase penetration capacity through membranes and consequently, have a poisoning effect.
3. The sensivity to the toxic effects of heavy metals and other xenobionts is dependent on the biological material being a microorganism, plant, animal, or type of tissue.

Bioaccumulation and the Environment

The pollution sources can be released by discharge of substances with uneven distribution in air, water, and soil. The movement of these substances up to bioaccumulation can occur by different routes mainly mediated by the food chain. This process can involve water, suspended particles, sediments, food, soil, and air particles (Figure 1). An important part of these substances can be concentrated in nonliving components. From these sources, persistent organic pollutants (POP) or heavy metals can be released to biota.[6] The final distribution presented by the mass balance in the environment with a group of

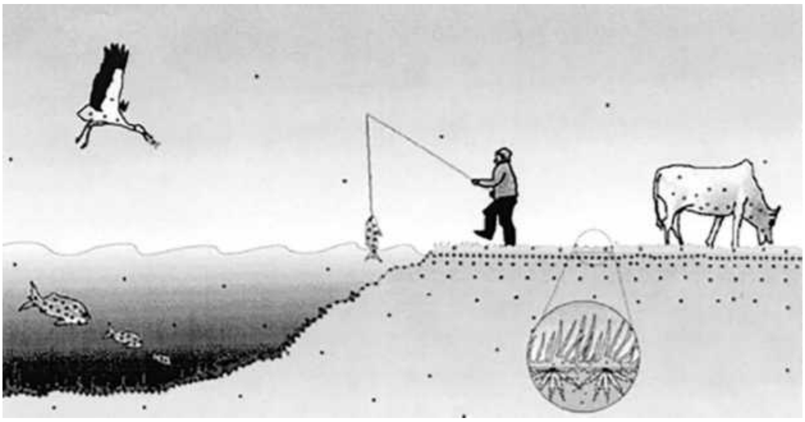


FIGURE 1 Bioaccumulation in the environment. The black points represent the pollutant molecules.

organochlorines and polyaromatics, shows that most are found in soil or sediment, whereas for highly volatile substances, most remain in the atmosphere. Less than 0.7% of the total remains in vegetation and no more than 2×10^{-3}% can be found in the aquatic biota.[7] The relationship between biota and environment shows that concentrations of bioaccumulated heavy metals in organisms are always higher than in water, but are usually lower than in sediments.[4]

Water Environment

After pollution reaches water bodies, different processes can occur to incorporate it into nonliving components as sediments and biota represented by microorganisms, plants, crustaceans, fishes, etc. The route of pollutant uptake in the biota if from waterborne, adsorption, filtration, or by food chain is an important factor in bioaccumulation.[8] Along the food chain the step-wise increase of concentration from lower to higher trophic levels, called biomagnification, can reach the bioconcentration factor (BCF) up to 100,000. In this process terrestrial animals as well as birds can be heavily contaminated by eating polluted fish.

In the global marine environment the apparent final fate of persistent organic pollutants (POP) is in the flora, fauna, and sediment of the abyss.[9] The main transport of POP follows the downstream movement of the organic flow in the water and in the long term these chemicals are incorporated in the sediment that function as final sink. It was observed that the bioaccumulation in the deep water fishes are up to 10 times higher than in surface water fishes.[9]

In the flora and fauna some heavy metals can bioaccumulate up to threshold values and others maintain a correlation with the concentration in the environment.[4] In aquatic plants, fish, and other metazoarians the distribution of the substances are quite distinct between tissues.[10] Lipophilic substances are preferentially found in adipose tissues with high lipid content.

In the Soil

Soil is polluted in large areas by pesticides application or by discharge as final disposal in landfills of industrial products. These lipophilic substances move in soil rather slowly by leaching, runoff, and volatilization. The main factor that influences bioaccumulation process in the soil is the biodisponibility. This property is conditioned by the adsorption/desorption capacity of the different soils and by the chemical nature of the pollutant. This process is driven by the stronger or weaker binding forces involved, which influence the amount that is bioavailable for plant uptake of these lipophilic pollutants.

The main flow of POP generally occurs toward organic matter from soil particles and not to biota, a process called preferential partition. A negative correlation was observed between the adsorbtion coefficient related to soil organic carbon (*K*oc) and the bioaccumulation factor by plants. This means that in organic rich soils, bioaccumulation in plants is rather small.[2] A similar situation occurs with microorganisms in which previous bioaccumulated organochorines can move out from the cell to the soil.[11] The preferential partition toward soil can be the reason why a lack of toxicity on soil microorganisms by pesticide applications was frequently observed, even in high concentrations of pesticides. Little information about bioaccumulation in soil could be observed, but nevertheless cotransport of some organochlorine accumulated in microorganisms in sand aquifers with low organic content was reported.[12]

Soil invertebrates such as earthworms, beetles, slugs, and others can bioaccumulate lipophilic pesticides. The bioaccumulation process could be seen as a soil to soil–water equilibrium followed by a soil–water to worm equilibrium.[2] Consumers of this biota in animals of higher trophic levels such as birds can biomagnificate these chemicals. Plants can adsorb and bioaccumulate products from the soil with incorporation of residues mainly in the root. The translocation from root to foliage depends on plant species and on the chemical properties of the pollutant. Several evidences show that lipophilic compounds are sorbed onto the outer surface of roots of several plants, and in this case translocation is very low.

Bioaccumulation in plants can also occur with heavy metals. As safety rules, domestic waste and sludge from wastewater treatment stations with heavy metal contamination can be applied on soil for agriculture up to limited amounts to avoid pollution with hazardous toxicological effects. Contaminated grass, grain, and fruits can be accumulated by biomagnification when consumed by mammals, insects, and birds. Terrestrial animals have a plant mediated relationship with soil contamination.

In the Air

The main sources of atmospheric pollution are pesticide spraying with the reverse process of evaporation from soil to air, poliaromatics produced by burning of fuels, and plastic incineration. The rate of entry to the atmosphere and the distance of movement are principally dependent on the vapor pressure of the pollutants and metereological conditions. In some cases movement occurs on a global scale. The dynamic nature of the atmosphere can dilute pollutants in the air to exceptionally low concentrations and in these cases no significant bioaccumultion occurs.[13] Nevertheless urban and industrial areas, as well as the margins of roads with intense traffic, can have high concentrations of pollutants. Plants exposed to xenobiotics in the form of vapor, particles, aerosol or larger droplets, can undergo a passive process of foliage adsorption with an uptake mainly in the wax cuticle.[2] Bioaccumulation in plants can result in damage and can also affect higher trophic levels that consume these vegetables.

Direct contact of animals with these chemicals can enter by the respiratory organs, in mammals, or the outer body surface, mainly in insects.

Bioaccumulation Mechanism in Biota

In terrestrial animals, pollutants can enter by dermal contact, respiration, and food consumption. Atmospheric pollutants move to the lungs, where an equilibrium is difficult to be established, while generally atmosphere dilutes pollutant concentrations, unless there is an exposure to constant pollution sources. Chemicals move from lungs to circulatory fluid (plasma) and can be metabolized with further excretion. Another route is the storage mainly in rich lipid bodies such as brain and eggs in bird's.[2] If the entrance is by food consumption, the gastrointestinal tract can eliminate[10] these substances or degrade than to more polar compounds with further excretion, or can promote adsorption by plasma following the same route described earlier.

The uptake of heavy metals in microorganisms can occur by bioadsorbtion in capsular polysaccharides and cell-wall polymers, or cross these layers and cell membranes by an active enzymatic process involving phosphatases, reaching to the cell interior.[14] Some authors define bioaccumulation as only the process in which molecules reach cell interior. Many cells from animals, plants, fungi, yeast, and bacteria have metal-binding proteins with low molecular weight called metallothioneins. These proteins bind mainly to Cd and Zn and constitute a protection mechanisms to the toxic effects of these substances. Some other cell protective mechanisms exist such as enhancement of efflux from cell to the outside. Metals bind on different macromolecules and change enzymatic activities with inhibition or stimulation effects.

Lipophilic hydrocarbons in microorganisms cross polysaccharides from capsule and cell wall polymers with adsorption mainly by the lipids of the membranes.[15] Compounds that are inserted in cell lipids are more difficult to be degraded by chemical or enzymatic processes getting higher persistence.[16] The probable mechanism of action seems to be nonspecific, this means not related to a specific target.

Applications and Future Perspectives

From the scientific point of view a better understanding of the integration between the different environment compartments and biota including modeling systems is an important approach that needs more development. Another possibility is to use bioaccumulation for environmental monitoring, based on the accumulation capacity of many pollutants in specific plants or animals, allowing chemical measurements that otherwise in water or air are below the analytic detection capacity.[17] This method has the possibility to integrate all pollutant exposure of plants or animals in a specific environment and can be in the future an important parameter for ecotoxicological evaluations.

After the disaster of the mercury pollution in the Minamata Bay in Japan in which more than 630 people died and many became physically and mentally disabled, the magnitude of poisoning effects due to bioaccumulation was recognized for the first time.[18] This was the beginning of a scientific research that produced a large amount of information. With this knowledge it became clear that bioaccumulation is a natural process that cannot be stopped by man but can be avoided with a more efficient control of pollutant release. To overcome the economic, social, and political difficulties for better pollution control together with the development of more ecological technologies are our challenge for today and for the future.

References

1. Connell, D.W. General Characteristics of Organic Compounds which Exhibit Bioaccumulation. In *Bioaccumulation of Xenobiotic Compounds*; Connell, D.W., Ed.; CRC Press: Boca Raton, Florida, 1990; 47–57.
2. Connell, D.W. Bioamagnification of Lipophilic Compounds in Terrestrial and Aquatic Systems. In *Bioaccumulation of Xenobiotics Compounds;* Connell, D.W., Ed.; CRC Press: Boca Raton, Florida, 1990; 145–185.
3. Brooke, D.N.; Dobbs, A.J.; Williams, N. Octanol: Water partition coefficients (*P*): measurement estimation and interpretation, particularly for chemicals with $P>105$. Eco-toxicol. Eviron. Saf. **1986**, *11*, 251.
4. Goodyear, K.L.; McNeill, S. Bioaccumulation of heavy metals by aquatic macro-invertebrates of different feeding guilds: a review. Sci. Total Environ. **1999**, *229*, 1–19.
5. Bourg, A.C.M. Speciation of Heavy Metals in Soils and Groundwater and Implications for Their Natural and Provoked Mobility. In *Heavy Metals: Problems and Solutions*; Salomons, N., Förstner, U., Mader, P., Eds.; Springer Verlag: Berlin, 1995; 17–31.
6. Tsezos, M.; Bell, J.P. Comparison of the biosorption and desorption of hazardous organic pollutants by life and dead bioamass. Wat. Res. **1989**, *23* (5), 561–568.

7. Connell, D.W.; Hawker, D.W. Predicting the distribution of persistent organic chemicals in the environment. Chem. Aust. **1986**, *53,* 428.
8. Carbonell, G.; Ramos, C.; Pablos, M.V.; Ortiz, J.A.; Tarazona, J.V. A system dynamic model for the assessment of different exposure routes in aquatic ecosystems. Sci. Total Environ. **2000**, *247,* 107–118.
9. Froescheis, O.; Looser, R.; Cailliet, G.M.; Jarman, W.M.; Ballschmiter, K. The deep-sea as a final global sink of semivolatile persistent organic pollutants part I: PCBs in surface and deep-sea dwelling fish of the north and south atlantic and the Monterey bay canion (California). Chemo- sphere **2000**, *40,* 651–660.
10. Lin, K.H.; Yen, J.H.; Wang, Y.S. Accumulation and elimination kinetics of herbicides butachlor, thiobencarb and Chlomethoxyfen by *Aristichthys nobilis.* Pestic. Sci. **1997**, *49,* 178–184.
11. Brunninger, B.M.; Mano, D.M.S.; Scheunert, I.; Langenbach, T. Mobility of the organochlorine compound dicofol in soil promoted by *Pseudomonas fluorescens.* Ecotoxic. Environ. Safety **1999**, *44*, 154–159.
12. Jenkins, M.B.; Lion, L.W. Mobile bacteria and transport of polynuclear aromatic hydrocarbons in porous media. Appl. Environ. Microbiol. **1993**, *59* (10), 3306–3313.
13. Connel, D.W. Environmental Routes Leading to the Bioaccumulation of Lipophilic Chemicals. In *Bioaccumulation of Xenobiotic Compounds;* Connell, D.W., Ed.; CRC Press: Boca Raton, FL, 1990; 59–73.
14. Gomes, N.C.M.; Mendonca-Hagler, L.C.S.; Savvaidis, I. Metal bioremediation by microorganisms. Rev. Microbiol. **1998**, *29,* 85–92.
15. Mano, D.M.S.; Langenbach, T. [14C]Dicofol association to cellular components of azospirillum. Pestic. Sci. **1998**, *53,* 91–95.
16. Mano, D.M.S.; Buff, K.; Clausen, E.; Langenbach, T. Bioaccumulation and enhanced persistence of the acaricide dicofol by *Azospirillum lipoferum.* Chemosphere **1996**, *33* (8), 1609–1619.
17. Maagd, P.G.J. Bioaccumulation test applied in whole effluent assessment: a review. Env. Toxic. Chem. **2000**, *19* (1), 25–35.
18. Takashi, H.; Tsubaki, T. Epidemiology: Mortality in Minamata Disease. In *Recent Advance in Minamata Disease Studies; Methylmercury Poisoning in Minamata and Niigata Japan;* Kodansha Ltd.: Tokyo, 1986; 1–23.

34

Biodegradation

Sven Erik Jørgensen

Introduction: Overview of the Properties of Toxic Chemical Compounds of Particular Importance for Environmental Management

Slightly more than 100,000 chemicals are produced in such an amount that they threaten or may threaten the environment. They cover a wide range of applications: household chemicals, detergents, cosmetics, medicines, dye stuffs, pesticides, intermediate chemicals, auxiliary chemicals in other industries, additives to a wide range of products, chemicals for water treatment, and so on. They are (almost) indispensable in modern society and cover many more or less essential needs in the industrialized world, which has increased the production of chemicals about 40-fold during the last five decades. A minor or major proportion of these chemicals are inevitably reaching the environment through their production, during their transportation from the industries to the end user, or by their application. In addition, the production or use of chemicals may cause more or less unforeseen waste or by-products, for instance, chloro-compounds from the use of chlorine for disinfection. As we would like to have the benefits of using the chemicals but cannot accept the harm they may cause, this conflict raises several urgent questions, which we already have discussed in other entries.

We cannot answer these crucial questions without knowing the properties of the chemicals. Organization for Economic Cooperation and Development (OECD) has made a review of the properties that we should know for all chemicals. We need to know the boiling point and melting point to know in which form (solid, liquid, or gas) the chemical will be found in the environment. We must know the dispersion of the chemicals in the five spheres: the hydrosphere, the atmosphere, the lithosphere, the biosphere, and the technosphere (the part of the earth that is controlled and under the influence of human technology). This will require knowledge about their solubility in water; the water/lipid partition coefficient; Henry's constant (the constant in Henry's Law, which indicates the distribution of the chemical between air and water); the vapor pressure; the rate of degradation by hydrolysis, photolysis, chemical oxidation, and microbiological processes; and the adsorption equilibrium between water and soil—all as functions of the temperature. We need to discover the interactions between living organisms and the chemicals, which implies that we should know the biological concentration factor (BCF), the magnification through the food chain, the uptake rate and the excretion rate by the organisms, and where in the organisms the chemicals will be concentrated, not only for one organism but for a wide range of organisms. We must also know the effects on a wide range of different organisms. It means that we should be able to find the Lethal Concentration causing 50% mortality oif the test animals (LC50) and Lethal Dose causing 50% mortality of the test animals (LD50) values; the Maximum Allowable Concentration (MAC) and Non-effect Concentration (NEC) values (for the abbreviations and the definitions used, see Appendix 5); the relationship between the various possible sublethal effects and concentrations; the influence of the chemical on fecundity; and the carcinogenic and teratogenic properties. We should also know the effect on the ecosystem level. How do the chemicals affect populations and their development and interactions, i.e., the entire network of the ecosystem? A reduction of one population may for instance influence the entire ecosystem because all populations are bound together in an ecological network.

Table 1 gives an overview of the most relevant physical-chemical properties of organic compounds and their interpretation with respect to the behavior of the environment. It is clear from the table that a high water solubility is not desirable as it implies that the chemical compound is very mobile. On the other hand, a low water solubility means a high solubility in fat tissue and, therefore, a high bioaccumulation and a high biomagnification. The biodegradability may, however, frequently be considered an even more important property than water solubility or K_{ow}. If a compound is biodegraded fast, it will be decomposed before it harms the environment. It means that a fastbiodegradation will, so to speak, neutralize the (harmful) effect of a high water solubility and a high solubility in fat tissue. On the other hand, if a compound is biodegraded slowly, it will stay in the environment for a very long time, which implies that a high mobility and a high risk of bioaccumulation and biomagnifications will be harmful. Therefore, it is almost possible to conclude that a compound with high biodegradability will clearly be much less harmful than a compound with a low biodegradability. Biodegradability is therefore a very crucial property for the estimation of a chemical compound's environmental effects.

The list of properties needed to give an adequate answer to the six questions mentioned above could easily be made longer (see, for instance, the list recommended by OECD). To provide all the properties corresponding to the list given here is already a huge task. More than 10 basic properties should be known for all 100,000 chemicals and organisms, which would require 1,000,000 pieces of information. In addition, we need to know at least 10 properties to describe the interactions between 100,000 chemicals and organisms. Let us say, modestly, that we use 10,000 organisms to represent the approximately 10 million species on earth. This gives a total of $1{,}000{,}000 + 100{,}000{*}10{,}000{*}10 =$ in the order of 10^{10} properties to be quantified! Today, we have determined less than 1% of these properties by measurements,

TABLE 1 Overview of the Most Relevant Environmental Properties of Organic Compounds and Their Interpretation

Property	Interpretation
Water solubility	High water solubility corresponds to high mobility.
K_{ow}	High K_{ow} means that the compound is lipophilic. It implies that it has a high tendency to bioaccumulate and be sorbed to soil sludge and sediment. BCF and Koc are correlated with K_{ow}.
Biodegradability	This is a measure of how fast the compound is decomposed to simpler molecules. A high biodegradation rate implies that the compound will not accumulate in the environment, while a lowbiodegradation rate may create environmental problems related to the increasing concentration in the environment and the possibilities of a synergistic effect with other compounds.
Volatilization, vapor	A high rate of volatilization (high vapor pressure) implies that the pressure compound will cause an air pollution problem.
Henry's constant, H	H determines the distribution between the atmosphere and the hydrosphere.
pK	If the compound is an acid or a base, pH determines whether the acid or the corresponding base is present. As the two forms have different properties, pH becomes important for the properties of the compounds.

Note: K_{ow} = Ratio solubility in octanol (represent fat tissue) divided by the solubility in water; Koc = express the adsorption ability to soil consisting of 100% organic carbon, can also with good approximation be considered as the concentration in soil with 1100% orgniac carbon divided with the concentration in water at equilibrium; H = Henry's constant in Henry's Law; pK = – log (equilibrium constant for the dissociation process of acids: $HA = A^- + H^+$).

and with the present rate of generating new data, we can be certain that during the 21st century, we shall not be able to reach 10% even with an accelerated rate of ecotoxicological measurements.

Environmental risk assessments require, among other inputs, information about the properties of the chemicals and their interactions with living organisms. It is maybe not necessary to know the properties with the very high accuracy that can be provided by measurements in a laboratory, but it would be beneficial to know the properties with sufficient accuracy to make it possible to utilize the models for management and risk assessment. Therefore, estimation methods have been developed as an urgently needed alternative to measurements. They are to a great extent based on the structure of the chemical compounds, the so-called QSAR and SAR methods (Quantitative Structure-Activity Relationship, it means estimation methods of chemical properties based on the chemical structure), but it may also be possible to use allometric principles to transfer rates of interaction processes and concentration factors between a chemical and one or a few organisms to other organisms.[1]

It may be interesting in this context to discuss the obvious question: why is it sufficient to estimate a property of a chemical in an ecotoxicological context with 20%, or sometimes 50% or higher, uncertainty? Ecotoxicological assessment usually gives an uncertainty of the same order of magnitude, which means that the indicated uncertainty may be sufficient from, for instance, the viewpoint of ecological modeling or ecological indicators, but can results with such an uncertainty be used at all? The answer in most (many) cases is “yes,” because we want in most cases to assure that we are (very) far from a harmful or very harmful level. We use a safety factor of 100–1000 in many cases. When we are concerned with very harmful effects, such as, for instance, complete collapse of an ecosystem or a health risk for a large human population, we will inevitably select a safety factor that is very high. In addition, our lack of knowledge about synergistic effects and the presence of many compounds in the environment at the same time force us to apply a very high safety factor. In such a context, we will usually go for a concentration in the environment that is magnitudes lower than that corresponding to a slightly harmful effect or considerably lower than the NEC. It is analogous to civil engineers constructing bridges. They make very sophisticated calculations (develop models) that account for wind, snow, temperature changes, and so on, and afterward, they multiply the results by a safety factor of 2–3 to ensure that the bridge will not collapse. They use safety factors because the consequences of a bridge collapse are unacceptable.

The collapse of an ecosystem or a health risk to a large human population is also completely unacceptable. Thus, we should use safety factors in ecotoxicological modeling to account for the uncertainty. Due to the complexity of the system, the simultaneous presence of many compounds, and our present knowledge, or rather, lack of knowledge, we should, as indicated above, use 10–100 or sometimes even 1000 as safety factor. If we use safety factors that are too high, the risk is only that the environment will be less contaminated at maybe a higher cost. Besides, there are no alternatives to the use of safety factors. We can, step by step, increase our ecotoxicological knowledge, but it will take decades before it may be reflected in considerably lower safety factors. A measuring program of all processes and components is impossible due to the high complexity of the ecosystems. This does not, of course, imply that we should not use the information of measured properties available today. Measured data will almost always be more accurate than the estimated data. Furthermore, the use of measured data within the network of estimation methods will improve the accuracy of estimation methods. Several handbooks on ecotoxicological parameters are, fortunately, available. References to the most important are given below. Estimation methods for the physical-chemical properties of chemical compounds were already applied 40–60 years ago, as they were urgently needed in chemical engineering. They are, to a great extent, based on contributions to a focal property by molecular groups and the molecular weight: the boiling point, the melting point, and the vapor pressure as a function of the temperature. In addition, a number of auxiliary properties result from these estimation methods, such as the critical data and the molecular volume. These properties may not have a direct application as ecotoxicological parameters in environmental risk assessment but are used as intermediate parameters, which may be used as a basis for estimation of other parameters.

The water solubility, the octanol/water partition coefficient, K_{ow}, and Henry's constant are crucial parameters in our network of estimation methods, because many other parameters are well correlated with these three parameters. The three properties can fortunately be found for a number of compounds or be estimated with reasonably high accuracy by use of knowledge of the chemical structure, i.e., the number of various elements, rings, and functional groups. In addition, there is a good relationship between water solubility and K_{ow}.[2] Particularly in the last decade, many good estimation methods for these three core properties have been developed.

During the last couple of decades, several correlation equations have been developed based upon a relationship between the water solubility, $K_{ow,}$ or Henry's constant on the one hand and physical, chemical, biological, and ecotoxicological parameters for chemical compounds on the other. The most important of these parameters are the following: the soil/water adsorption isotherms; the rate of the chemical degradation processes (hydrolysis, photolysis, and chemical oxidation); the BCF; the ecological magnification factor (EMF); the uptake rate; the excretion rate; and a number of ecotoxicological parameters. The ratio of concentrations both in the sorbed phase and in water at equilibrium Ka and BCF may often be estimated with a relatively good accuracy from expressions like Ka or BCF = $a \log K_{ow} + b$. Numerous expressions with different a and b values have been published.[3–5]

Biodegradation

It was concluded in the previous section that biodegradation is probably the most important property for the estimation of a chemical compound's environmental effect. This section is therefore devoted to the presentation of this important property.

Biodegradation rates may be expressed in several ways. Microbiological biodegradation may, with good approximation, be described as a Monod equation:[5]

$$\mathrm{d}c/\mathrm{d}t = -\mathrm{d}B/Y\mathrm{d}t = -\mu\mathrm{max}^{*} Bc/Y(\mathrm{Km}+c) \tag{1}$$

where c is the concentration of the compound considered, Y is the yield of biomass B per unit of c, B is the biomass concentration, μmax is the maximum specific growth rate, and Km is the half saturation constant. If $c \ll \mathrm{Km}$, the expression is reduced to a first-order reaction scheme:

$$\mathrm{d}c/\mathrm{d}t = -K'Bc \tag{2}$$

where $K' = \mu\mathrm{max}/(Y\,\mathrm{Km})$. B is in nature determined by the environmental conditions. In aquatic ecosystems, B is for instance highly dependent on the presence of suspended matter. B may therefore under certain conditions[5] be considered a constant, which reduces the rate expression to

$$\mathrm{d}c/\mathrm{d}t = -kc \tag{3}$$

An indication of k in the unit 1/hr, 1/24hr, 1/week, 1/mo, or 1/yr can therefore be used to describe the rate of biodegradation. If the biological half-life time is denoted t, we get the following relation:

$$\ln 2 = 0.7 = k\ t \tag{4}$$

This implies that the biological half-life time also can be used to indicate the biodegradation rate.

The biodegradation in waste treatment plants is often of particular interest, in which case the % of the Theoretical Oxygen Demand (ThOD) or the theoretical biological oxygen demand (BOD) may be used as a suitable reference. Most often, however, the 5-day BOD as percentage of the theoretical BOD is used. It may also be indicated as the BOD5fraction. For instance, a BOD5fraction of 0.7 will mean that BOD5 corresponds to 70% of the theoretical BOD. It is, however, also possible to find an indication of percentage removal in an activated sludge plant. The biodegradation is, however, in some cases very dependent on the concentration of microorganisms as expressed in the above-shown equations. Therefore *K′* indicated in the unit mg/(g dry wt 24 hr) will in many cases be more informative and correct.

In the microbiological decomposition of xenobiotic compounds, an acclimatization period from a few days to 1–2 mo should be foreseen before the optimum biodegradation rate can be achieved. We distinguish between primary and ultimate biodegradation. Primary biodegradation is any biologically induced transformation that changes the molecular integrity. Ultimate biodegradation is the biologically mediated conversion of organic compounds to inorganic compounds and products associated with complete and normal metabolic decomposition.

To conclude, the biodegradation rate is expressed by application of a wide range of units:

1. As a first-order rate constant (1/24 hr).
2. As half-life time (days or hours).
3. mg per g sludge per 24 hr[mg/(g 24 hr)].
4. mg per g bacteria per 24 hr[mg/(g 24 hr)].
5. mL of substrate per bacterial cell per 24 hr[mL/(24 hr cells)].
6. mg COD per g biomass per 24 hr[mg/(g 24 hr)].
7. mL of substrate per gram of volatile solids inclusive microorganisms [mL/(g 24 hr)].
8. BODx/BOD∞, i.e., the biological oxygen demand in x days compared with complete degradation (-), named the BODxcoefficient.
9. BODx/COD, i.e., the biological oxygen demand in x days compared with complete degradation, expressed by means of COD (-).

Estimation of Biodegradation

The biodegradation rate in water or soil is difficult to estimate because the number of microorganisms varies several orders of magnitudes from one type of aquatic ecosystem to the next and from one type of soil to the next. Artificial intelligence has been used as a promising tool to estimate this important parameter. However, a (very) rough, first estimation can be made on the basis of the molecular structure and the biodegradability. The following rules can be used to set up these estimations:

1. Polymer compounds are generally less biodegradable than monomer compounds; 1 point for a molecular weight >500 and ≤1000, 2 points for a molecular weight >1000.
2. Aliphatic compounds are more biodegradable than aromatic compounds; 1 point for each aromatic ring.
3. Substitutions, especially with halogens and nitro groups, will decrease the biodegradability; 0.5 point for each substitution, although 1 point if it is a halogen or a nitro group.
4. The introduction of a double or triple bond will generally mean an increase in the biodegradability (double bonds in aromatic rings are of course not included in this rule); -1 point for each double or triple bond.
5. Oxygen and nitrogen bridges [– O – and – N – (or=)] in a molecule will decrease the biodegradability; 1 point for each oxygen or nitrogen bridge.
6. Branches (secondary or tertiary compounds) are generally less biodegradable than the corresponding primary compounds; 0.5 point for each branch.

Find the number of points and use the following classification:

≤1.5 points: the compound is readily biodegraded. More than 90% will be biodegraded in a biological treatment plant.

2.0–3.0 points: the compound is biodegradable. Probably about 10–90% will be removed in a biological treatment plant. BOD5 is 0.1–0.9 of the theoretical oxygen demand.

3.5–4.5 points: the compound is slowly biodegradable. Less than 10% will be removed in a biological treatment plant. BOD10 ≤ 0.1 of the theoretical oxygen demand.

5.0–5.5 points: the compound is very slowly biodegradable. It will hardly be removed in a biological treatment plant, and a 90% biodegradation in water or soil will take ≥6 mo.

≥6.0 points: the compound is refractory. The half-life time in soil or water is counted in years. The structure of dichlorodiphenyltrichloroethane (DDT) corresponds, for instance, to about 7 points, and the biological half-life of DDT in soil is about 14 years.

Several useful methods for estimation of biological properties are based upon the similarity of chemical structures. The idea is that if we know the properties of one compound, it may be used to find the properties of similar compounds. If for instance we know the properties of phenol, which is named the parent compound, it may be used to give more accurate estimation of the properties of monochloro-phenol, dichloro-phenol, trichloro-phenol, and so on and for the corresponding cresol compounds. Estimation approaches based on chemical similarity give generally more accurate estimation but of course are also more cumbersome to apply, as they cannot be used generally in the sense that each estimation has a different starting point, namely, the compound, named the parent compound, with known properties.

Allometric estimation methods presume[6] that there is a relationship between the value of a biological parameter and the size of a considered organism.

The various estimation methods, including estimation methods applicable for biodegradation, may be classified into two groups:

1. General estimation methods based on an equation of general validity for all types of compounds, although some of the constants may be dependent on the type of chemical compound, or they may be calculated by adding contributions (increments) based on chemical groups and bonds.
2. Estimation methods valid for a specific type of chemical compound, for instance, aromatic amines, phenols, aliphatic hydrocarbons, and so on. The property of at least one key compound is known. Based upon the structural differences between the key compound and all other compounds of the considered type(for instance, two chlorine atoms have substituted hydrogen in phenol to get 2,3-dichloro-phenol) and the correlation between the structural differences and the differences in the considered property, the properties for all compounds of the considered type can be found. These methods are based on chemical similarity.

Methods of class 2 are generally more accurate than methods of class 1, but they are more cumbersome to use as it is necessary for each type of chemical to find for each property the right correlation. Furthermore, the requested properties should be known for at least one key component, which sometimes may be difficult when a series of properties are needed. If estimation of the properties for a series of compounds belonging to the same chemical class is required, it is tempting to use a suitable collection of class 2 methods.

Methods of class 1 form a network that facilitates possibilities of linking the estimation methods together in a computer software system, for instance,WINTOX. [1] An updated version named Estimation of Ecotoxicological Properties (EEP) is now available. EEP can estimate the biodegradability, in contrast to WINTOX. The software is easy to use and can rapidly provide estimations. Each relationship between two properties is based on the average result obtained from a number of different equations found in the literature. There is, however, a price for using such "easy-to-go" software. The accuracy of the estimations is not as good as with more sophisticated methods based upon similarity in chemical structure, but in many, particularly modeling, contexts, the results found by WINTOX and EEP can offer sufficient accuracy. In addition, it is always useful to come up with a first intermediate guess. It could, for instance, be used to estimate whether a chemical compound would be decomposed by biological treatment.

The software also makes it possible to start the estimations from the properties of the chemical compound already known. The accuracy of the estimation from use of the software can be improved considerably by having knowledge about a few key parameters, for instance, the boiling point and

Henry's constant. WINTOX and EEP are based on average values of results obtained by simultaneous use of several estimation methods for most of the parameters. It implies increased accuracy of the estimation, mainly because it gives a reasonable accuracy for a wider range of compounds. If several methods are used in parallel, a simple average of the parallel results has been used in some cases, while a weighted average is used in other cases where it has been found beneficial for the overall accuracy of the program. When parallel estimation methods are giving the highest accuracy for different classes of compounds, use of weighting factors seems to offer a clear advantage. It is generally recommended to apply as many estimation methods as possible for a given case study to increase the overall accuracy. If the estimation by WINTOX and EEP can be supported by other recommended estimation methods, it is strongly recommended to do so.

References

1. Jørgensen, S.E.; Halling-Sørensen, B.; Mahler, H. *Handbook of Estimation Methods in Ecotoxicology and Environmental Chemistry;* Taylor and Francis Publ.: Boca Raton, FL, 1998; 230 pp.
2. Jørgensen, S.E. *Principles of Pollution Abatement;* Elsevier: Amsterdam, 2000; 520 pp.
3. Jørgensen, S.E.; Jørgensen, L.A.; Nors Nielsen, S. *Handbook of Ecological and Ecotoxicological Parameters*; Elsevier: Amsterdam, 1991; 1380 pp.
4. Jørgensen, L.A.; Jørgensen, S.E.; Nors Nielsen, S. *Ecotox,* CD; Elsevier: Amsterdam, 2000; corresponding to 4000 pp.
5. Jørgensen, S.E.; Bendoricchio, G. *Fundamentals of Ecological Modelling,* 3rd Ed.; Elsevier: Amsterdam, 2001; 530 pp.
6. Peters, R.H. *The Ecological Implications of Body Size*; Cambridge University Press: Cambridge, 1983; 329 pp.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

35

Biological Control of Vertebrates

Peter Kerr and
Tanja Strive

Introduction

Biological control is the deliberate use of one organism to control another, normally a pest species, by reducing its population size, rate of increase or geographic spread and so diminishing its environmental and economic impacts. In classical biological control, usually applied to insect and plant pest species, predators, disease agents or parasites are used to diminish the impact of the pest sometimes with considerable success. Biological control of vertebrate pests has a much more limited history and very few success stories. The advantage of biological control is that while it may be expensive to search for and evaluate biological control agents, a successful biological control provides ongoing control at little cost. Against this must be balanced the risk of unforseen ecological consequences due to the introduced agent or control of the pest and the likely diminution of the impact of the control over time. These principles, successes and limitations are illustrated by examining the natural history of biological control of vertebrate pest species together with a discussion of more recent attempts to use biotechnology to develop novel biological controls.

Vertebrate Pests

Vertebrate pest species may be invasive, feral, deliberately or accidentally introduced, or native to the area; the definition of what constitutes a pest species will vary from location to location, across time, social groups and perspectives. Pest species range from jawless fish to mammals and exist on local and continental scales. Even iconic species such as Australian koalas (*Phascolarctos cinereus*) or African elephants (*Loxodonta africana*) can have local pest status and require some form of management. Similarly, options for control of vertebrate pests will inevitably be influenced by a wide range of social, ecological and economic pressures. A particular consideration, not always addressed, is that unlike plants or insect pests, vertebrate pests are capable of feeling pain and distress, so welfare must be taken into account when control measures are proposed.

Conventional control measures for vertebrate pests can basically be summarized as exclusion by some form of barrier or deterrent (physical, chemical, habitat destruction, guardian dogs, predators) or removal, which could include capture and relocation for iconic species but will more commonly be lethal, including hunting, (commercial, recreational, directed culling), trapping and poisoning. Fertility control, using surgical, hormonal, chemical or immunological means, has been applied, at least experimentally, for some species (Pickard and Holt 2007). The inability to deliver long-acting targeted reproductive controls cheaply and on large scales currently limits the application of fertility control to localized regions or high-value iconic species.

The common problem with all conventional measures is that unless the pest can be permanently eliminated, ongoing control is necessary with the recurrent costs competing with other social and political priorities. These shifting priorities frequently result in poorly sustained control.

Biological Control

Predators as Biological Control Agents for Vertebrate Pests

Biological control of vertebrate pests using introduced predators has a long if not particularly glorious history. The classic example is the use of cats (*Felis catus*) to control rodents around human settlements, which extends back into prehistory (Driscoll et al. 2007). If pest populations are low enough then predators can prevent an increase in population size (Pech et al. 1992). However, introduction of exotic predators frequently leads to unintended consequences. For example, cats have been deliberately introduced onto many islands to control rodents or European rabbits (*Oryctolagus cuniculus*) often with unforseen and devastating consequences for bird and reptile populations (Courchamp et al. 2003, Nogales et al. 2004).

Mongooses (*Herpestes auropunctatus*) were introduced into the West Indies and subsequently Hawaiian Islands to control rodents and became major pests endangering ground-dwelling birds and many other species while in Hawaii, the rats moved into the trees and threatened arboreal birds (Courchamp et al 2003).

Similarly, in an attempt to control introduced European rabbits, weasels (*Mustela nivalis*), ferrets (*Mustela furo*), stoats (*Mustela erminea*) and cats were widely released in New Zealand during the later part of the 19th century and ferret release continued into the 1920s (Gibb and Williams 1994). Ferrets, cats and stoats became well established with serious and ongoing impacts on native birds and reptiles. In addition, ferrets have provided a reservoir host for bovine tuberculosis with trade and public health consequences. It appears unlikely that these predators had a significant effect on rabbit populations at the time of release, but following intensive rabbit control measures from the 1950s onwards and the release of rabbit hemorrhagic disease virus (RHDV) in 1997, predators may be important for maintaining populations at low levels (Gibb and Williams 1994, Reddiex et al. 2002, Henning et al. 2008).

In a further example, foxes (*Dusicyon griseus*) were introduced into Tierra del Fuego in 1951 in an attempt to control an estimated 30 million European rabbits, the endemic fox species *Dusicyon culpaeus* having been driven almost to extinction. It is improbable that the 24 introduced foxes and their progeny had any impact on this rabbit population. However, the introduction of myxomatosis in 1954 dramatically reduced rabbit numbers, and it has been suggested that the introduced foxes together with the preexisting *D. culpaeus* may have helped to prevent buildup of the rabbit population, although rabbits formed less than 2% of the diet of *D. griseus* (Jaksic and Yanez 1983).

There are, however, two small-scale success stories with predators. Gray wolves (*Canis lupus*) reintroduced to Yellowstone National Park in the United States have provided a high-order predator to control elk (*Cervus elaphus*) populations. Elk browsing had prevented regeneration of aspen (*Populus tremuloides*) and other trees since the 1920s when wolves were eliminated, leading to erosion and environmental degradation (Ripple and Beschta 2007). The presence of wolves has reduced grazing pressure on lowland aspen as the elk have changed their behavior or had their numbers reduced. However, this introduction could be seen more as a restoration of the previous ecology of the system rather than as biological control.

In a more direct example of biological control, European red foxes (*Vulpes vulpes*) were deliberately introduced onto two small islands in the Aleutians and appear to have eliminated populations of previously introduced arctic foxes (*Alopex lagopus*) although whether this was by competition for limited resources, direct predation or both is not clear (Bailey 1992). To prevent breeding, only male red foxes were released on one island and on the second, five vasectomised males and five females were released.

Reintroduction of high-order predators, such as dingoes (*Canis familiaris*) in Australia and leopards (*Panthera pardus*) in South Africa, has been proposed as a means of control for multiple vertebrate pest species. This could be potentially coupled with deployment of guardian dogs to protect livestock (Minnie et al. 2015, Newsome et al. 2016, Allen et al. 2019). Introduction of predator fish has also been investigated to control European carp (*Cyprinus carpio*) in Australia and the United States (Poole and Bajer 2019).

Biological Control of Vertebrate Pests Using Parasites

Parasites in the broadest sense are essentially another form of predator, albeit often nonlethal. Parasites are traditionally divided into microparasites (bacteria, viruses, protozoa, fungi) and macroparasites (helminths, arthropods). The earliest documented experiment with a microparasite for biological control is the use of chicken cholera bacteria (*Pasteurella multocida*) by Louis Pasteur in 1887 to exterminate European rabbits on an 8 Ha walled estate in France (Pasteur 1888). However, there are only three examples of vertebrate biocontrol by a microparasite on any large scale: myxomatosis in European rabbits, rabbit hemorrhagic disease in European rabbits and feline panleukopenia in cats.

There is some evidence that macroparasites may regulate vertebrate populations in the wild (Hudson et al. 1998, Tompkins and Begon 1999, Albon et al. 2002, Redpath et al. 2006). However, the interactions of parasites including macroparasites with other biotic and abiotic factors are complex (Tompkins et al. 2011). The only successful example of biological control with macroparasites is the release of two species of flea, *Spilopsyllus cuniculi* and *Xenopsylla cunicularis*, in Australia, which, by enhancing the transmission of the virus causing myxomatosis, act as indirect biological controls for European rabbits (see Chapter 13 Myxoma virus and Rabbit hemorrhagic disease virus as biological controls for rabbits).

Other attempts to use macroparasites as biological controls for vertebrates have been less successful. In one example, field, laboratory and modeling studies on the nematode *Capillaria hepatica* suggested it could act as a biological control to prevent irruptions of house mice (*Mus musculus*) in the cereal growing areas of south-eastern Australia (Singleton and McCallum 1990). However, intensive field releases of *C. hepatica* failed to show any impact on wild mouse populations (Singleton et al. 1995, Singleton and Chambers 1996).

Lungworms (*Rhabdias* spp.) have been proposed as possible biological controls for two invasive amphibians, cane toads (*Bufo marinus*) in Australia and coquis (*Eleutherodactylus coqui*) in Hawaii (Kelehear et al. 2009, Marr et al. 2010). However, preliminary studies with coquis were unpromising, and it is difficult to see how lungworms could effectively be introduced into the invading cane toad population on a landscape scale in the Australian outback.

The Natural History of Successful Biological Controls

Biological Control of Rabbits

The European rabbit originated in the Iberian Peninsula but has been spread deliberately or accidentally into the wild throughout much of Europe, Britain, North Africa, parts of Chile and Argentina, New Zealand, over 800 islands and, perhaps most famously, Australia (Flux 1994). All domestic breeds of rabbits are derived from these European rabbits. Wild rabbits from Britain were introduced into Australia in 1859 and within 50 years had spread across most of the nontropical parts of the continent constituting the most dramatic and rapid biological invasion ever documented for a mammalian pest (Flux 1994).

Rabbits are highly adaptive generalist herbivores that substantially modify their habitat (Thompson 1994, Myers et al. 1994). Their burrows and warrens enable survival under many different climatic extremes and high reproductive rates mean that rabbit populations can rapidly expand when rainfall triggers food availability (Myers et al. 1994). In dry areas of Australia, even quite low densities of rabbits (less than 1 per hectare) can prevent regeneration of shrubs and trees leading to local species extinctions, soil erosion, weed invasion and loss of other vertebrate species which are unable to survive in the modified habitat (Williams et al. 1995, Bird et al. 2012, Cooke 2012). In Australia, rabbits are a food source for two recently introduced predators, the European red fox and the feral cat (Woinarski et al. 2015). By supporting populations of foxes and cats, rabbits contribute to the endangerment and extinction of small marsupials by these predators (Dickman 1996, Kearney et al. 2018). In addition to the major ecological damage caused by rabbits, there is also a significant cost due to lost agricultural production from grazing pressure, weed invasion, soil erosion and crop destruction (Myers et al. 1994, Williams et al. 1995) as well as the economical and ecological costs of conventional controls such as habitat destruction and widespread poisoning (Williams et al. 1995, Cooke 2012).

Biological control of rabbits was successfully undertaken with myxoma virus (MYXV), the causative agent of myxomatosis, following its release in 1950. While initially highly successful, subsequent rabbit-virus coevolution led to a diminution of the effect of myxomatosis and a rebound of rabbit numbers, particularly in the low rainfall rangelands. In 1995, the escape of RHDV from a study site, introduced a second biological control for rabbits, which again substantially reduced rabbit numbers. These viruses and their impacts are described in detail in Chapter 13 Myxoma virus and Rabbit hemorrhagic disease virus as biological controls for rabbits.

Biological Control of Cats

The domestic cat (*F. catus*) adapts quickly to a feral lifestyle and has become a serious problem following its deliberate or accidental introduction onto islands (Courchamp et al. 2003, Nogales et al. 2004). Feral cats have also been associated with species loss or endangerment in Australia and New Zealand (Dickman 1996, Fitzgerald and Gibb 2001, Woinarski et al. 2015, Murphy et al. 2019).

Marion Island, the larger of the Prince Edward Islands, is a South African possession in the southern Indian Ocean (46°54′S; 37°45′E). This sub-Antarctic island, about 1900 km south east of Cape Town, has an area of 290 km^2 and hosts major breeding populations of penguins and seabirds (Bester et al. 2002). Five domestic cats were deliberately released in 1949 to control introduced house mice around a meteorological station (van Aarde and Skinner 1981, Bester et al. 2002).

In 1975, the cat population was estimated at 2139 ± 290 (van Aarde and Skinner 1981, van Aarde 1984). These cats would have needed to consume over 450,000 burrowing petrels a year just to satisfy their minimum energy requirements (van Aarde and Skinner 1981). At least one species of petrel was believed to have been driven to local extinction by 1965 (Bester et al. 2002). The isolation, rugged landscape and the climate of Marion Island meant that conventional control of the cat population would be difficult and so biological control was an attractive option (Bester et al. 2002). Feline panleukopenia virus (FePV) was chosen as a biological control after a limited serological survey indicated that it was not present in the cat population (van Aarde 1984).

Feline Panleukopenia Virus

FePV (family *Parvoviridae*; sub-family *Parvovirinae*; genus *Protoparvovirus*; species *Carnivore protoparvovirus 1*) is a small (25 nm diameter) icosahedrally symmetrical, non-enveloped virus with a single-stranded DNA genome of 5200 nucleotides (Murphy et al. 1999, Stuetzer and Hartmann 2014). The virus is a natural pathogen of cats; it infects most if not all Felidae; closely related viruses infect dogs, mink and other carnivores (Steinel et al. 2001). Because it needs components of the host DNA polymerase to replicate its genome, FePV only replicates in the nucleus of actively dividing cells. This means it has a tropism for lymphoid tissue, bone marrow and the dividing epithelial cells in the intestinal crypts. Destruction of these cells leads to the clinical signs of gastroenteritis and panleukopenia (massive loss of white blood cells and platelets) seen in cats infected with FePV, (in neonates and fetal kittens the virus infects cells of the developing brain and heart rather than the gut) (Truyen et al. 2009).

Infected cats shed virus in feces, urine, saliva and vomit. Estimates of the duration of shedding range from 7 days to several months (Murphy et al. 1999, Steinel et al. 2001). Very high titers of virus (10^9 ID_{50} per g of feces) may be excreted and the virus can persist in the environment in an infectious state for many months. There are thus two modes of transmission: direct from an infected to a susceptible cat and indirect from the environment (Berthier et al. 2000).

Infection is mostly via the oropharynx with virus initially replicating in the pharyngeal lymphoid tissue, followed by a viremia that distributes virus to all tissues of the infected animal. Case fatality rates have been estimated at 20% in adult cats and 80% in kittens (Berthier et al. 2000). Cats that recover from infection are likely to be immune for life with immunity primarily due to neutralizing antibody. Kittens born to immune queens are protected for some weeks after birth by maternal antibody, delivered in the colostrum (Truyen et al. 2009). Kittens infected while still partially protected by maternal antibody may have an enhanced survival rate.

Feline Panleukopenia Virus on Marion Island

In March 1977, FePV was inoculated into 96 previously trapped cats which were released back into a population estimated at 3400 (Howell 1984, van Aarde 1984). This was at the end of the breeding season when kitten numbers and sub-adult numbers were likely to be high. The virus had a case fatality rate of approximately 50% in a small number of cat passages; both cat-passaged and cell culture-passaged virus were released (Howell 1984).

Following the release, cat sightings dropped from 102 to 18.9 per sampling period, suggesting that the virus had established and was having a significant impact on the cat population. In 1978, 45 of 57 cats sampled had antibodies to FePV (Van Rensburg et al. 1987), confirming that the virus had established and spread in the population. Subsequent serological testing indicated that it was persisting in the cat population (Van Rensburg et al. 1987).

Population density was estimated to have decreased by 65% 3 years after the release from 3400 to 600 (van Aarde 1984). As would be expected, the disease had its main impact on the young and sub-adult populations causing a marked shift in age distribution, with recovered and immune cats forming a long-lived breeding population and epizootics of the disease occurring in susceptible kittens each breeding season as protection from maternal antibodies waned. By 1982, it was considered that the rate of population decrease had slowed or stabilized.

Sea bird breeding had improved in areas where cats had been eliminated, but there was still sufficient predation to impede re-establishment of some species (Bester et al. 2002). Therefore, it was decided to proceed with the complete eradication of cats from Marion Island using conventional means. Cats were finally eliminated in 1991 following intensive hunting, trapping and poisoning campaigns (Bester et al. 2002). Ironically, cats were introduced to Marion Island as a form of biological control for the house mouse. Subsequent to the eradication of cats, the house mouse is now regarded as a threatening species (van Aarde et al. 1996).

Prospects for Biological Control of Cats

Because it does not have a prolonged carrier state, FePV is unlikely to be introduced onto islands by small founder populations. However, the success of FePV on Marion Island has not seen more use of this virus in cat control. The only documented example being on Jarvis Island in the Pacific (Rauzon 1985). On this 414 Ha atoll, 31 cats were inoculated with FePV and released into a population estimated to be <200 cats. No attempt was made to determine whether the virus was already present in the population. At least 10 of 19 marked cats survived the infection, and there was no good evidence that the virus established. Cats were subsequently eradicated from Jarvis Island using conventional means (Rauzon 1985).

Modeling studies (Courchamp and Sugihara 1999) suggest that two feline retroviruses which cause persistent infections, *feline leukaemia virus* and *feline immunodeficiency virus*, could be effective biological controls on island populations with feline leukemia virus capable of eradicating cats under some conditions. FePV was regarded as unsuitable, despite the apparent success on Marion Island, because of its short transmission time. However, this is at least partially compensated by the prolonged infectivity of FePV in the environment (Berthier et al. 2000).

Potential Biological Control Agents

Koi (Carp) Herpesvirus as a Potential Biological Control for the European Carp

The introduced European carp (*C. carpio*) has become the dominant fish species in inland waterways of south-eastern Australia. Their high fecundity, ecological adaptability and omnivorous diet have made carp a highly visible pest species, with impacts due to increased water turbidity, decreased vegetation and possibly increased frequency of algal blooms due to higher nutrient levels in waterways (Koehn 2004, McColl et al. 2007). European carp are also considered a pest species in North America and New Zealand but in other parts of the world are highly valued (Saunders et al. 2010).

Cyprinid herpesvirus 3 (CyHV-3; order: *Herpesvirale*; family: *Alloherpesviridae*; genus: *Cyprinivirus*) is a widely distributed emerging pathogen of European carp that is absent from Australia. The virus has been associated with massive die-offs of koi and common carp with estimates of mortality ranging from 70–100%. It appears to be species-specific for common and koi carp although it has been suggested that goldfish (*Carassius auratus*) may be asymptomatically infected (El-Matbouli and Soliman 2011). High virulence, species specificity, and apparently ready transmission in water, together with absence from Australia suggested that this virus may have potential as a biological control agent (McColl et al. 2007). CyHV-3 was imported into the high-security Australian Animal Health laboratory for evaluation (McColl et al. 2007, Saunders et al. 2010, McColl et al. 2016). Larval carp appear highly susceptible with mortality rates of 96–100% following infection. Adults had a lower mortality rate (Saunders et al. 2010). Plans have been developed for a virus release in 2019–2020; however, there are still questions being raised about the strategy (Kopf et al. 2017, Lighten and van Oosterhout 2017, Marshall et al. 2018), and it is likely that any release will be delayed. It is considered necessary that biological control would be combined with conventional control measures and possibly future biotechnological controls (Thresher et al. 2014, McColl et al. 2018).

Future Biocontrols and Biotechnology for Control of Vertebrate Pests

Biotechnology has the potential to engineer existing parasites to enhance virulence, modify antigenicity or to deliver other antigens or regulatory molecules such as interfering RNA. Any such modification for vertebrate pest control is likely to be controversial particularly when a vertebrate pest in one country, such as the rabbit in Australia, is a valued keystone species in its natural range and because of the risk of altering the host range of the parasite. More recently, genetic biocontrols such as gene drives, based around DNA editing CRISPR/Cas9-type and similar systems, which would directly modify pest species have been investigated as potential controls for vertebrate pest species.

Virally Vectored Immunocontraception

Biological controls that inhibit reproduction or development are an attractive target for pest control. Virally vectored immunocontraception (VVIC) aimed to use a genetically engineered transmissible agent to deliver an immunocontraceptive antigen to a pest species. Infection with the recombinant organism would stimulate an immune response to the immunocontraceptive antigen which would block fertility. Potential antigens include components of sperm, the zona pellucida surrounding the egg or peptide hormones such as gonadotrophin-releasing hormone. Direct injection of such antigens is already used in some wildlife population control programs (Naz and Saver 2016). However, direct injection is not feasible for control of widespread invasive species such as rabbits, mice and foxes.

In Australia, extensive studies were undertaken to develop VVIC for three key invasive species: house mice, European rabbits and European red foxes. However, none of the technologies developed were sufficiently effective for field use (Hardy et al. 2006). While mice could be rendered sterile by inoculation with recombinant murine cytomegalovirus, the virus did not transmit effectively from mouse to mouse (Redwood et al. 2007). In rabbits, >90% of females could be rendered infertile following inoculation with recombinant MYXV, but in around 50% of cases, infertility was short-lived (Mackenzie et al. 2006, van Leeuwen and Kerr 2007). Field experiments using surgical sterilization of female rabbits indicated that levels of immunocontraception efficacy >80% would be required to have an impact on the population and reduce ecological damage (Twigg et al. 2000, Williams et al. 2007). Despite considerable effort, no effective product was developed for foxes (Reubel et al. 2005, Strive et al. 2006, Strive et al. 2007). All the recombinant viruses tested expressed zona pellucida proteins as antigens. The use of possibly more effective antigens such as gonadotropin-releasing hormone was not explored.

Prospects for Virally Vectored Immunocontraception

Although potentially a means of biological control, immunocontraception vectored by a recombinant organism has efficacy and safety hurdles to overcome before it could be a reality. These include: (1) an immunocontraceptive antigen must stimulate a long-lasting immune response that cross-reacts with a self-antigen to cause sustained infertility at least in females. (2) The recombinant organism must be able to establish in a population, transmit and be maintained in the population in competition with field strains of the organism or be able to be regularly reintroduced, possibly in a bait. As demonstrated with murine cytomegalovirus, recombinant viruses may not transmit effectively. (3) The immune response should not need boosting. (4) The vector organism and preferably the antigen should be species-specific. (5) The risk of the organism jumping species by mutation or recombination should be minimal.

These scientific and technical conditions have not yet been satisfied in any species. Beyond the technical issues, there are national and international social, environmental and political issues that would need to be addressed before vector-delivered fertility control could be a reality.

Biological Controls for Cane Toads

The cane toad (*Rhinella marinus*) is native to South America and was introduced to Australia in 1935 to control beetles in sugarcane crops (Shine 2010). While the toads failed as beetle-biocontrol agents, they proved to be excellent invaders and have spread from their original release in coastal northern Queensland south to NSW and across the tropical north of Australia as far as Western Australia and are continuing to expand towards the Indian Ocean (Shine 2010).

All life stages of the cane toad are toxic to predators due to compounds collectively referred to as bufadienolides. As a result, cane toad invasions have led to declines in populations of native predator species, such as snakes, freshwater crocodiles, goannas and quolls at the invasion front although the long-term impact is more difficult to quantify (Shine 2010).

Between 1990 and 1993, an extensive search for natural pathogens was undertaken in the toad's native range in Venezuela. A number of virus isolates from cane toads and a frog were made, all belonging to the family *Iridoviridae*, genus *Ranavirus* and considered to represent a single species. While these viruses killed cane toad tadpoles, they were also lethal to at least one Australian native frog species and considered unsuitable as biocontrol agents (Shanmuganathan et al. 2010). More recently, metagenomic studies have identified a novel rhinovirus and an endogenous retrovirus in cane toads from Australia (Russo et al. 2018).

In a biotechnological strategy, it was proposed to use a recombinant virus to deliver cane toad adult-specific proteins into tadpoles. This approach was based on the hypothesis that expression of adult-specific proteins in tadpoles would create an immune response, since tadpoles are not immunologically tolerant to adult proteins. By selecting an antigen that was critical for metamorphosis, it was proposed that maturation would be inhibited. Although proof of concept had been shown many years earlier in tadpoles of the American bullfrog immunized with adult hemoglobin (Maniatis et al. 1969), this was unsuccessful in cane toads and the work has been discontinued (Shanmuganathan et al. 2010).

To date, no biocontrol is available to effectively reduce cane toad populations or halt their spread across the Australian continent.

Genetic Biocontrol of Invasive Species

The use of genetic biocontrols such as gene drives to spread an allele through a population is receiving considerable attention (Esvelt et al. 2014). The basic concept is to use a genetic construct that copies itself into homologous chromosomes, thus creating super-Mendelian inheritance whereby more than 50% of the offspring inherit the drive element. An element that disabled a key gene involved in reproduction or caused sex bias, for example, by turning all individuals into phenotypic males, could be introduced into the population causing population crashes. The advent of CRISPR/Cas9 and other bacterial nuclease/guide systems as gene editing tools has turned this from a largely theoretical approach to a potentially practical means of controlling populations. Genetic biocontrol of pest populations would have the advantages of being self-disseminating and strictly species-specific (since only transmitted by sexual reproduction). In addition, genetic biocontrol should be much more humane than conventional lethal controls or pathogens.

Proof of concept has been demonstrated in mosquito populations under containment conditions (Kyrou et al. 2018). In vertebrates, there have been proposals to use gene drives to control invasive rodent species on islands and for invasive fish (Thresher et al. 2014, Leitschuh et al. 2018). Modeling studies have examined the opportunities and constraints, particularly the selection of resistant alleles, of gene drives for rodent control (Prowse et al. 2017). A germline-expressed gene drive has been demonstrated in laboratory mice; however, super-Mendelian inheritance occurred only in female mice (Grunwald et al. 2019). Although currently in its infancy, it is likely that this technology will rapidly improve. In the meantime, there will be considerable debate on the risks associated with driving a sterilizing gene into a population and whether these can be satisfactorily managed (Esvelt et al. 2014, National Academy of Sciences 2016).

Conclusions

A number of conclusions can be drawn from the history of biological control of vertebrate pests that should inform the search for and use of biological control agents.

1. It is challenging to find effective biological controls for vertebrates and those that have been most successful, myxomatosis and rabbit hemorrhagic disease, were emerging diseases with extreme case fatality rates not pre-existing diseases of the European rabbit in its natural range.
2. Searching for biological control agents in the original home range of the vertebrate pest, as has been so successful for plant and insect pests, may be unsuccessful for vertebrate pests because it could well be that predation or resource limitation was more important in population control than parasitism particularly by a single organism (Tompkins et al. 2011). This was exemplified by the search for cane toad pathogens in Venezuela. FePV is a natural pathogen of cats but has been successfully used on only one island and may not be species-specific.
3. Laboratory trials may be poor predictors of success in the field.
4. As the early experience with myxomatosis in Australia indicates, it may be difficult to establish a novel pathogen in a population, particularly if the epidemiology of the pathogen is not well understood.
5. The initial impact of a biological control is unlikely to be sustained.

   At the very least, the pathogen may undergo local extinction when host populations fall to low numbers. For rapidly reproducing invasive pest species, this can allow local recovery of populations. Unlike insects and plants, vertebrates surviving infection develop adaptive immunity that can protect from subsequent reinfection. This has two impacts: it provides an immune pool of breeding animals, and passive antibody transfer may protect their offspring for some weeks after birth and hence alter the impact and timing of epizootics. At some level, there will likely be selection for enhanced resistance to the pathogen and, depending on the life history of the pathogen, there may be selection for more attenuated strains of the pathogen.
6. Biological controls should not be seen as “silver bullets” that remove the need for ongoing integrated pest management strategies. However, successful biological controls such as myxomatosis and rabbit hemorrhagic disease (RHD) can significantly suppress populations on continental scales.
7. The introduction of novel high-order predators to control populations has generally not been successful due to the low species specificity of the predator and the need to reduce pest numbers to very low levels before predators can keep them in check.
8. Biotechnological approaches to vertebrate biocontrol have so far been unsuccessful. These have a lot of hurdles to overcome, not least scientific and technical, but also regulatory and social. However, given the difficulty in finding natural biocontrol agents, biotechnology should not be ignored and gene biocontrol technologies offer potentially powerful new means of controlling populations if efficacy and safety can be achieved.

Bibliography

Albon, S. D., A. Stein, R. J. Irvine, R. Langvatn, E. Ropstad et al. 2002. The role of parasites in the dynamics of a reindeer population. *Proc. Roy. Soc. Lond. B*. 269:1625–1632.

Allen, B. L., L. R. Allen, G. Ballard, M. Drouilly, P. J. S. Fleming et al. Animal welfare considerations for using large carnivores and guardian dogs as vertebrate biocontrol tools against other animals. *Conserv. Biol*. 232:258–270.

Bailey, E. P. 1992. Red foxes, *Vulpes vulpes*, as biological control agents for introduced Arctic foxes, *Alopex lagopus* on Alaskan islands. *Can. Field. Nat*. 106:200–205.

Berthier, K., M. Langlais, P. Auger, and D. Pontier. 2000. Dynamics of a feline virus with two transmission modes within exponentially growing host populations. *Proc. Roy. Soc. Lond. B* 267:2049–2056.

Bester, M. N., J. P. Bloomer, R. J. van Aarde, B. H. Erasmus, P. J. J. van Rensburg et al. 2002. A review of the successful eradication of feral cats from sub-Antarctic Marion Island, Southern Indian Ocean. *S. Af. J. Wildl. Res*. 32:65–73.

Bird, P., Mutze, G., Peacock, D., Jennings, S. 2012. Damage caused by low-density exotic herbivore populations: the impact of introduced European rabbits on marsupial herbivores and *Allocasuarina* and *Bursaria* seedling survival in Australian coastal shrubland. *Biol Invasions* 14: 743–755.

Cooke, B. D. 2012. Rabbits: manageable environmental pests or participants in new Australian ecosystems? *Wildl. Res*. 39:279–289.

Courchamp, F., J. L. Chapuis, and M. Pascal. 2003. Mammal invaders on islands: impact, control and control impact. *Biol. Rev*. 78:347–383.

Courchamp, F., and G. Sugihara. 1999. Modeling the biological control of an alien predator to protect island species from extinction. *Ecol. Appl*. 9:112–123.

Dickman, C. R. 1996. Impact of exotic generalist predators on the native fauna of Australia. *Wildl. Biol*. 2:185–195.

Driscoll, C. A., M. Menotti-Raymond, A. L. Roca, K. Hupe, W. E. Johnson et al. 2007. The near eastern origin of cat domestication. *Science* 317:519–523.

El-Matbouli, M., and H. Soliman. 2011. Transmission of Cyprinid herpesvirus-3 (CyHV-3) from goldfish to naive common carp by cohabitation. *Res. Vet. Sci*. 90:536–539.

Esvelt, K. M., A. L. Smidler, F. Catteruccia, and G. M. Church. 2014. Concerning RNA-guided gene drives for the alteration of wild populations. *eLife* 3:e03401.

Fitzgerald, B. M., and J. A. Gibb. 2001. Introduced mammals in a New Zealand forest: long-term research in the Orongorongo Valley. *Biol. Cons*. 90:97–108.

Flux, J. E. C. 1994. World distribution. In The European *Rabbit*. The *History* and *Biology* of a *Successful Colonizer*, edited by H. V. Thompson and C. M. King, 8–21. Oxford: Oxford University Press.

Gibb, J. A., and J. M. Williams. 1994. The rabbit in New Zealand. In The European *Rabbit*. The *History* and *Biology* of a *Successful Colonizer*, edited by H. V. Thompson and C. M. King, 158–204. Oxford: Oxford University Press.

Grunwald, H. A., V. M. Gantz, G. Poplawski, X-R. S. Xu, E. Bier et al. 2019. Super-Mendelian inheritance mediated by CRISPR-Cas9 in the female mouse germline. *Nature* 566:105–109.

Hardy, C. M., L. A. Hinds, P. J. Kerr, A. J. Redwood, G. R. Shellam et al. 2006. Biological control of pest animals using virally-vectored immunocontraceptive vaccines. *J. Reprod. Immunol*. 71:102–111.

Henning, J., D. U. Pfeiffer, P. R. Davies, M. A. Stevenson, and J. Meers. 2008. Mortality patterns over 3 years in a sparse population of wild rabbits (*Oryctolagus cuniculus*) in New Zealand, with an emphasis on rabbit haemorrhagic disease (RHD). *Eur. J. Wild. Res*. 54:619–626.

Howell, P. G. 1984. An evaluation of the biological control of the feral cat *Felis catus* (Linnaeus, 1758). *Acta Zool. Fennica* 172:111–113.

Hudson, P. J., A. P. Dobson, and D. Newborn. 1998. Prevention of population cycles by parasite removal. *Science* 282:2256–2258.

Jaksic, F. M., and J. L. Yanez. 1983. Rabbit and fox introductions in Tierra del Fuego: history and assessment of the attempts at biological control of the rabbit infestation. *Biol. Conserv*. 26:367–374.

Kearney, S. G., J. Cawardine, A. E. Reside, D. O. Fisher, M. Maron et al. 2018. The threats to Australia's imperilled species and implications for a national conservation response. *Pac. Conserv. Biol*. https://doi.org/10.1071/PC18024.

Kelehear, C., J. K. Webb, and R. Shine. 2009. *Rhabdias pseudosphaerocephala* infection in *Bufo marinus*: lung nematodes reduce viability of metamorph cane toads. *Parasitology* 136:919–927.

Koehn, J. D. 2004. Carp (*Cyprinus carpio*) as a powerful invader in Australian waterways. *Freshwater Biol*. 49:882–894.

Kopf, R. K., D. G. Nimmo, P. Humphries, L. J. Baumgartner, M. Bode et al. 2017. Confronting the risks of large-scale invasive species control. *Nature Ecol. Evol*. 1:0172.

Kyrou, K., A. M. Hammond, R. Galizi, N. Kranjc, A. Burt et al. 2018. A CRISPR-Cas9 gene drive targeting *doublesex* causes complete population suppression in caged *Anopheles gambiae* mosquitoes. *Nat. Biotech*. 36:1062–1066.

Leitschuh, C. M., D. Kanavy, G. A. Backus, R. X. Valdez, M. Serr et al. 2018. Developing gene drive technologies to eradicate invasive rodents from islands. *Journal of Responsible Innovation* 5 (Sup 1):S121–S138.

Lighten, J., and C. van Oosterhout. 2017. Biocontrol of common carp in Australia poses risks to biosecurity. *Nat. Ecol. Evol*. 1:0087.

Mackenzie, S M., E. A. McLaughlin, H. D. Perkins, N. French, T. Sutherland, et al. 2006. The immunocontraceptive effects on female rabbits (*Oryctolagus cuniculus*) infected with recombinant myxoma virus expressing rabbit ZP2 and ZP3. *Biol Reprod* 74:511–521.

Maniatis, G. M., L. A. Steiner, and V. M. Ingram. 1969. Tadpole antibodies against frog hemoglobin and their effect on development. *Science* 165:67–69.

Marr, S. R., S. A. Johnson, A. H. Hara, and M. E. McGarrity. 2010. Preliminary evaluation of the potential of the helminth parasite *Rhabdias elegans* as a biological control agent for invasive Puerto Rican coquis (*Eleutherodactylus coqui*) in Hawaii. *Biol. Control*. 54:69–74.

Marshall, J., A. J. Davison, R. K. Kopf, M. Boutier, P. Stevenson et al. 2018. Biocontrol of invasive carp: risks abound. *Science* 359:877.

McColl, K. A., A. Sunarto, and E. C. Holmes. 2016. Cyprinid herpesvirus 3 and its evolutionary future as a biological control for carp in Australia. *Virol. J*. 13:206.

McColl, K. A., A. Sunarto, and M. J. Neave. 2018. Biocontrol of carp: more than just a herpesvirus. *Front. Microbiol*. 9:2288.

McColl, K. A., A. Sunarto, L. M. Williams, and St-J. Crane M. 2007. Koi Herpes virus: dreaded pathogen or white knight? *Aquaculture Health* 2007 (9):4–6.

Minnie, L., A. F. Boshoff, and G. I. H. Kerley. 2015. Vegetation type influences livestock predation by leopards: implications for conservation in agro-ecosystems. *Afr J Wildlife Res* 45:204–214.

Murphy, F. A., E. P. J. Gibbs, M. C. Horzinek, and M. J. Studdert. 1999. *Veterinary Virology*. 3 ed. San Diego: Academic Press.

Murphy, B. P., L-A. Woolley, H. M. Geyle, S. M. Legge, R. Palme et al. 2019. Introduced cats (*Felis catus*) eating a continental fauna: the number of mammals killed in Australia. *Biol Conservation* 237:28–40.

Myers, K., I. Parer, D. Wood, and B. D. Cooke. 1994. The rabbit in Australia. In The European *Rabbit*. The *History* and *Biology* of a *Successful Colonizer*, edited by H. V. Thompson and C. M. King, 108–157. Oxford: Oxford University Press.

National Academy of Sciences, Engineering & Medicine. 2016. Gene drives on the horizon: advancing science, navigating uncertainty and aligning research with public values. Washington, DC: National Academies Press.

Naz, R. J., and A. E. Saver. 2016. Immunocontraception for animals; current status and future perspective. *Am. J. Reprod. Immunol*. 75:426–439.

Newsome, T. M., G-A. Ballard, M. S. Crowther, J. A. Dellinger, P. J. S. Fleming et al. 2016. Resolving the value of the dingo in ecological restoration. *Restoration Ecology* 23:201–208.

Nogales, M., A. Martin, B. R. Tershy, C. J. Donlan, D. Veitch et al. 2004. A review of feral cat eradication on islands. *Conserv. Biol*. 18:310–319.

Pasteur, L. 1888. Sur la destruction des lapins en Australie et dans la Nouvelle-Zelande. *Annales de l'Institut de Pasteur (Paris)* 2:1–8.

Pech, R. P., A. R. E. Sinclair, A. E. Newsome, and P. C. Catling. 1992. Limits to predator regulation of rabbits in Australia: evidence from predator removal experiments. *Oecologia* 89:102–112.

Pickard, A. R., and W. V. Holt. 2007. Contraception in wildlife. *J. Fam. Plann. Reprod. Health. Care*. 33:48–52.

Poole, J. R., and P. G. Bajer. 2019. A small native predator reduces reproductive success of a large invasive fish as revealed by whole-lake experiments. *PLoS ONE* 14:e0214009.

Prowse, T. A. A., P. Cassey, J. V. Ross, C. Pfitzner, T. A. Wittmann et al. 2017. Dodging silver bullets: good CRISPR gene-drive design is critical for eradicating exotic vertebrates. *Proc. R. Soc. Lond. B* 284:20170799.

Rauzon, M. J. 1985. Feral cats on Jarvis Island: their effects and their eradication. *Atoll Research Bulletin* No. 282:1–30.

Reddiex, B., G. J. Hickling, G. L. Norbury, and C. M. Frampton. 2002. Effects of predation and rabbit haemorrhagic disease on population dynamics of rabbits (*Oryctolagus cuniculus*) in North Canterbury, New Zealand. *Wildl. Res.* 29:627–633.

Redpath, S. M., F. Mougeot, F. M. Leckie, D. A. Elston, and P. J. Hudson. 2006. Testing the role of parasites in driving the cyclic population dynamics of a gamebird. *Ecol. Lett.* 9:410–418.

Redwood, A. J., L. M. Smith, M. L. Lloyd, L. A. Hinds, C. M. Hardy et al. 2007. Prospects for virally-vectored immunocontraception in the control of wild house mice (*Mus domesticus*). *Wildl. Res.* 34:530–539.

Reubel, G. H., S. Beaton, D. Venables, J. Pekin, J. Wright et al. 2005. Experimental inoculation of European red foxes with recombinant vaccinia viruses expressing zona pellucida C proteins. *Vaccine* 23:4417–4426.

Ripple, W. J., and R. L. Beschta. 2007. Restoring Yellowstone's aspen with wolves. *Biol. Conserv*. 138:514–519.

Russo, A. G., J-S. Eden, D. Enosi Tuipulotu, M. Shi, D. Selechnik et al. 2018. Viral discovery in the invasive Australian cane toad (*Rhinella marina*) using metatranscriptomic and genomic approaches. *J. Virol.* 92:e00768–18.

Saunders, G., B. Cooke, K. McColl, R. Shine, and T. Peacock. 2010. Modern approaches for the biological control of vertebrate pests: an Australian perspective. *Biol. Control* 52:288–295.

Shanmuganathan, T., J. Pallister, S. Doody, H. McCallum, T. Robinson et al. 2010. Biological control of the cane toad in Australia: a review. *Animal Conserv*. 13:16–23.

Shine, R. 2010. The ecological impact of invasive cane toads (*Bufo marinus*) in Australia. *Q. Rev. Biol.* 85:253–291.

Singleton, G.R., and L. K. Chambers. 1996. A manipulative field experiment to examine the effect of *Capillaria hepatica* (Nematoda) on wild mouse populations in southern Australia. *Int. J. Parasitol.* 26:383–398.

Singleton, G.R., L. K. Chambers, and D. M. Spratt. 1995. An experimental field study to examine whether *Capillaria hepatica* (Nematoda) can limit house mouse populations in eastern Australia. *Wildl. Res.* 22:31–53.

Singleton, G.R., and H. I. McCallum. 1990. The potential of *Capillaria hepatica* to control mouse plagues. *Parasitol. Today* 6:190–193.

Steinel, A., C. R. Parrish, M. E. BLoom, and U. Truyen. 2001. Parvovirus infections in wild carnivores. *J. Wildl. Dis*. 37: 594–607.

Strive, T., C. M. Hardy, N. French, J. D. Wright, N. Nagaraja et al. 2006. Development of canine herpesvirus based antifertility vaccines for foxes using bacterial artificial chromosomes. *Vaccine* 24:980–988.

Strive, T., C. M. Hardy, and G. H. Reubel. 2007. Prospects for immunocontraception in the European red fox (*Vulpes vulpes*). *Wildl. Res.* 34:523–539.

Stuetzer, B., and K. Hartmann. 2014. Feline parvovirus infection and associated diseases. *Vet J.* 201:150–155.

Thompson, H. V. 1994. The rabbit in Britain. In The European *Rabbit*. The *History* and *Biology* of a *Successful Colonizer*, edited by H. V. Thompson and C. M. King, 64–107. Oxford: Oxford University Press.

Thresher, R., J. van de Kamp, G. Campbell, P. Grewe, M. Canning et al. 2014. Sex-ratio-biasing constructs for the control of invasive lower vertebrates. *Nat. Biotech*. 32:424–427.

Tompkins, D. M., and M. Begon. 1999. Parasites can regulate wildlife populations. *Parasitol. Today* 15:311–313.

Tompkins, D. M., A. M. Dunn, M. J. Smith, and S. Telfer. 2011. Wildlife diseases: from individuals to ecosystems. *J. Anim. Ecol*. 80:19–38.

Truyen, U., D. Addie, S. Belak, C. Boucraut-Baralon, H. Egberink et al. 2009. Feline Panleukopenia. ABCD guidelines on prevention and management. *J. Fel. Med. Surg*. 11:538–546.

Twigg, L. E., T. J. Lowe, G. R. Martin, A. G. Wheeler, G. S. Gray et al. 2000. Effects of surgically imposed sterility on free-ranging rabbit populations. *J. App. Ecol*. 37:16–39.

van Aarde, R. J. 1984. Population biology and the control of feral cats on Marion Island. *Acta Zool. Fennica* 172:107–110.

van Aarde, R. J., S. Ferreira, T. Wassenaar, and D. G. Erasmus. 1996. With the cats away the mice may play. *S. Af. J. Sci*. 92:357–358.

van Aarde, R. J., and J. D. Skinner. 1981. The feral cat population at Marion Island: characteristics, colonization and control. Colloque sur les Ecosystemes Subantarctiques. C.N.F.R.A. no. 51.

van Leeuwen, B. H., and P. J. Kerr. 2007. Prospects for fertility control in the European rabbit (*Oryctolagus cuniculus*) using myxoma virus-vectored immunocontraception. *Wildl. Res*. 34:511–522.

Van Rensburg, P. J. J., J. D. Skinner, and R. J. Van Aarde. 1987. Effects of feline panleucopaenia on the population characteristics of feral cats on Marion Island. *J. App. Ecol*. 24:63–73.

Williams, C. K., C. C. Davey, R. J. Moore, L. A. Hinds, L. E. Silvers et al. 2007. Populations responses to sterility imposed on female European rabbits. *J. Appl. Ecol*. 44:291–301.

Williams, K., I. Parer, B. Coman, J. Burley, and M. Braysher. 1995. *Managing Vertebrate Pests: Rabbits*. Canberra: Bureau of Resource Sciences. Australian Government Publishing Services.

Woinarski, J. C. Z., A. A. Burbidge, and P. L. Harrison. 2015. Ongoing unraveling of a continental fauna: decline and extinction of Australian mammals since European settlement. *Proc. Natl. Acad. Sci. USA*. 112:4531–4540.

36

Biological Controls

Heikki Hokkanen

Introduction

Biological control of pests has been actively practiced for the control of pests, weeds, and plant diseases for more than 100 years, and it has had some 150 spectacular successes,[1] which in economic terms have been just as impressive as in ecological terms: the calculated return for investment is 32:1, while for other control methods the ratio is around 2.5:1.[2,3] However, the obtained successes are only the tip of the iceberg of all the work carried out in the field. To date, more than 6000 introductions of alien natural enemies have been carried out, worldwide.[4] It is estimated that only about 35% of all introduced biocontrol agents have become ecologically established in the target ecosystem, and only 60% of these have provided any economic or biocontrol success.[3,5] Of all the individual biocontrol projects, only 16% have resulted in complete control of the target pest.[6] A major ecological and economic challenge is to improve the ratio of successes in biological control, while retaining the excellent safety record of this approach to pest control.

General Principles

While it has been shown that biological control can be effective in any climate, ecosystem, and crop, the factors determining success or failure remain largely unknown, and often are economic rather than ecological in nature.[1] Very few general principles to improve the efficacy and predictability of biological control have emerged; these include better ecological background knowledge, genetic improvement (in particular, genetic engineering) of biocontrol agents, and the utilization of new ecological associations in selecting the biocontrol agents.[7] The genetic engineering of biocontrol agents, especially insects, is still in its infancy and cannot be expected to improve the success ratios in the foreseeable future. In contrast, the new association principle has—usually unknowingly—been used for a long time, and is increasingly employed to find more effective natural enemies for current biological control programs.

New Associations

The standard biological control principle is to reestablish the ecological balance between an exotic pest and its natural enemies occurring in their country of origin (the "old association approach").[2,3] It has been argued, however, that this is an inefficient way of practicing biological control, because due to an evolved long-term equilibrium between the pest and the natural enemy, the control agent only seldom is very efficient.[6] To find more effective enemies one should search among agents that do not share an evolutionary history with the target pest (the "new association" approach). Such natural enemies can be found, for example, for the target pest in areas where the pest has been introduced only recently, or among enemies attacking related species in other geographical areas.[6]

Evidence for Improved Efficacy

Analysis of past biocontrol successes and failures have indicated that when employing the new association principle it is possible to increase the success ratio by at least 75%.[6] More detailed studies showed that some natural enemy groups may be particularly attractive as new association agents (Table 1). Such analyses are, however, often confounded by the fact that new association agents seldom have been considered as the primary choice in biological control, and consequently, usually five to seven old association agents are introduced before a new association agent is tried. In addition, on average much greater numbers (two-to fourfold) of old association agents are normally introduced (Table 1), which further increases the probability of biocontrol success, and biases the analyses against new association agents. Therefore, the estimate for improving the success ratio appears to be conservative.

Several spectacular, well-documented biocontrol successes that have employed new association control agents are known, and these include the complete control of serious pests such as the sugarcane borer *Diatraea saccharalis* in the Caribbean, coconut spike moth *Levuana iridescens* in Fiji, southern green stink bug *Nezara viridula* in Hawaii, the moth *Oxydia trychiata* in Colombia, and several scale insect species in California, Greece, and Australia.[7] Further, more detailed examples will be given below on new research with good prospects of success utilizing this approach.

TABLE 1 Comparisons of Biological Control Introductions with Old and New Association Control Agents Utilizing Tachinidae, Braconidae and Eulopidae

	Proportion (%) of Successes of All Cases (Introductions)		Total Number of Cases		Bias in the Release Numbers[a]
	Old	New	Old	New	
Tachinidae	10.9	17.1	92	41	3.8-fold
Braconidae	17.2	14.4	169	97	1.6-fold
Eulopidae	28.6	35.7	56	28	1.7-fold
Overall	17.4	18.7	317	166	

Source: Hokkanen, H. M. T., unpublished data.

[a] Indicates how many more individuals on average of old association agents were released in the introduction projects, compared with new association agents. In the case of Tachinidae and Braconidae the mean number of released new association agents was below 5000 individuals, which is considered to be the necessary number to ensure a fair chance for the natural enemies to establish themselves.

Recent Cases Employing New Associations

Eurasian Watermilfoil

The Eurasian watermilfoil (*Myriophyllum spicatum*) was introduced into North America several decades, possibly 100 years, ago. It grows rapidly, forms a dense canopy on the water surface, and often interferes with recreation, inhibits water flow, and impedes navigation. Herbicides and mechanical harvesting have been used to control infestations, costing \$150–\$2000 per acres annually in Minnesota.[8]

Sometimes naturally occurring declines of the watermilfoil have been observed. The main causal agent proved to be a native beetle *Euhrychiopsis lecontei*, the milfoil weevil, which subsequently has shown control potential in controlled field experiments. The weevil is a specialist herbivore of watermilfoils, but prefers the Eurasian to its native host, the northern watermilfoil (*M. sibiricum*). Research is in progress to use the milfoil weevil effectively as a biocontrol agent against the Eurasian watermilfoil in North America.[8]

Lantana

Lantana camara is a serious weed of Mexican or Caribbean origin, affecting cropping lands and forest areas in 47 countries. Lantana was the focus of the first weed biocontrol effort in history (1902), and there is an enormous literature on Lantana biocontrol. Several complexes of herbivores have been credited for exerting some degree of biocontrol of the weed (e.g., in Hawaii), many employing new association agents jointly with old association agents. Latest research gives data on the good efficacy and release in Australia of the moth *Ectaga garcia* originating from South America, where it feeds on the related weed *Lantana montevidensis*.[9]

Triffid (Siam) Weed

The triffid weed (*Chromolaena odorata*) is a perennial shrub native of tropical America. In recent decades it has become a serious pest of humid tropics around the world.[10] It spreads rapidly in lands used for forestry, pasture, and plantation crops and can reach a height of three meters in open situations and up to eight meters in forests. For more than two decades the triffid weed has been the subject of intensive research as a target for biological control. However, so far all attempts at biocontrol of *C. odorata* have failed. Recently the new association biological control agent, arctiid moth *Pareuchaetes aurata aurata* collected from *C. jujuensis* in South America, was considered as more promising than the related moth *P. pseudoinsulata*, an old association control agent previously thought of as one of the best biocontrol candidates.[10]

Southern Green Stink Bug

The biological control of the southern green stink bug (the green vegetable bug) (*Nezara viridula*) in Australia, New Zealand, and Hawaii has been heralded as a landmark example of classical biological control.[11] An egg parasitoid—old association agent—*Trissolcus basalis* and a tachinid fly—new association agent—*Trichopoda pennipes* have jointly provided these successes. Control by the fly has been considered as relatively more important, and indeed, in Australia where the fly has failed to establish, the control is poor and the bug remains a serious pest. Currently in Australia another new association tachinid fly, *Trichopoda giacomellii,* is being released after research showed it has excellent potential for control.[12]

Citrus Leafminer

Citrus agroecosystems have numerous potentially damaging pests often maintained under substantial to complete biological control by both old and new association agents. The citrus leafminer *Phyllocnistis citrella*, native to Asia, has spread rapidly throughout the citrus growing areas of the world in recent years.[13] It arrived in Florida in 1993 and in less than one year invaded and colonized the entire state. An old association parasitic wasp *Ageniaspis citricola* was introduced in 1994, and after establishment it has held the pest under control with significant help from native parasitoids such as *Pnigalio minio* (new association agent).[13] In some other areas native parasitoids similarly have shown significant control effect on the citrus leafminer (e.g., in Italy). This example illustrates well the fact that invading species often do not become pests, because effective local natural enemies keep them in check.

Tarnished Plant Bug

An ongoing study in the United States has identified as the most important parasitoid of the native pest *Lygus lineolaris*, the tarnished plant bug, the exotic species *Peristenus digoneutis*, originally introduced for the control of related introduced mirid plant bugs.[14] This example serves well to point out the importance of native pests, which in most if not all areas form the majority of all pest species. As old association biological control agents seldom can be utilized for the control of native pests, their biocontrol by introduced natural enemies has attracted relatively little attention and, indeed, only three decades ago was considered an impossible task. Several recent examples, usually utilizing new association control agents, show that biological control can work against native pests just as well as against exotic ones.

Future Prospects

Compared with chemical control, the success rates of biological control are outstanding. While only about one out of 15,000 tested chemicals ends up as a chemical pesticide meeting the requirements of efficacy and safety, approximately one out of seven introductions of natural enemies has been successful using old associations.[15] Using new association control agents this rate could still be increased to about one out of four, while the array of potential natural enemies is also substantially larger providing a wider choice. In addition, the potential uses for natural enemy introductions are broadened to include the control of native pests.

A major concern with respect to all biological control introductions is the question of nontarget safety. Biological control has an excellent record of safety[3,16] and it covers the new association agents as well: there have been some 1500–2000 introductions already (out of 6000) that have involved new association agents.[7] Those extremely few cases where a negative nontarget effect has been suspected as a result of biological control, all involve old association agents; therefore it is clear that new associations can safely be used to help obtain biological control successes at an increasing rate.

References

1. Hokkanen, H.M.T. Success in classical biological control. CRC Crit. Rev. Plant Sci. **1985**, *3*, 35–72.
2. Cullen, J.M.; Whitten, M.J. Economics of Classical Biological Control: A Research Perspective. In *Biological Control: Benefits and Risks*; Hokkanen, H.M.T., Lynch, J.M., Eds.; Cambridge University Press: Cambridge, U.K., 1995; 270–276.
3. *Biological Control: Benefits and Risks*; Lynch, J.M., Hokkanen, H.M.T., Eds.; Cambridge University Press: Cambridge, U.K., 1995.
4. Waage, J. In *Agendas, Aliens and Agriculture*; Global Biocontrol in the Post UNCED Era, Cornell Community Conference on Biological Control, http://www.nysaes.cornell.edu/ent/bcconf/talks/waage.html (accessed Jan 5, 1999).

5. Hokkanen, H.M.T. Pest Management, Biological Control. In *Encyclopedia of Agricultural Science*; Academic Press, Inc.: San Diego, 1994; 3, 155–167.
6. Hokkanen, H.M.T.; Pimentel, D. New associations in biological control: theory and practice. Can. Entomol. **1989**, *121,* 829–840.
7. Hokkanen, H.M.T. New Approaches in Biological Control. In *CRC Handbook of Pest Management in Agriculture*, 2nd Ed., Pimentel, D., Ed.; CRC Press: Boca Raton, FL, 1991; II, 185–198.
8. Newman, R.M. Biological Control of Eurasian Watermilfoil. http://www.fw.umn.edu/research/milfoil/milfoilbc.html (accessed Feb 5, 1999).
9. Day, M.D.; Wilson, B.W.; Latimer, K.J. The life history and host range of *Ectaga garcia*, a biological control agent for *Lantana camara* and *L. montevidensis* in Australia. BioControl **1998**, *43,* 325–338.
10. Kluge, R.L.; Caldwell, P.M. The biology and host specificity of *Pareuchaetes aurata aurata* (Lepidoptera: Arctiidae), a "New Association" biological control agent for *Chromolaena odorata* (Compositae). Bull. Entomol. Res. **1993**, *83*, 87–94.
11. Caltagirone, L.E. Landmark examples in classical biological control. Annu. Rev. Entomol. **1981**, *26*, 213–232.
12. Coombs, M. *Biological Control of Green Vegetable Bug in Australia and PNG*; Pest Management Current Programs and Projects. http://www.ento.csiro.au/research/pestmgmt/pmp16.htm (accessed Oct 5, 1999).
13. Timmer, L.W. *Citrus Leafminer Proves to be an IPM Success*; IPM Florida, 1996 *Winter.* http://www.ias.ufl.edu/~FAIRSWEB/IPM/IPMFL/v2n4/leafminer.htm (accessed Feb 5, 1999).
14. Day, W.H. Host preferences of introduced and native parasites (Hymenoptera: Braconidae) of phytophagous plant bugs (Hemiptera: Miridae) in Alfalfa-Grass Fields in the Northeastern USA. BioControl **1999**, *44*, 249–261.
15. Hokkanen, H.M.T. Role of Biological Control and Transgenic Crops in Reducing Use of Chemical Pesticides for Crop Protection. In *Techniques for Reducing Pesticide Use*; Pimentel, D., Ed.; John Wiley and Sons: New York, 1997; 103–127.
16. *Evaluating Indirect Ecological Effects of Biological Control*; Scott, J.K., Quimby, P.C., Wajnberg, E., Eds.; CABI Publishing: Wallingford, U.K., 2001; 261.

37

Biological Factors Impeding Recovery of Predatory Fish Populations

Catalina Chaparro Pedraza

Introduction

Historically, global fisheries have relied significantly on large predatory fish species (Myers and Worm 2003). The previous seemingly inexhaustible fish abundance contributed to the public's perception that fish stocks such as tuna, billfish, cod, salmon, and pike are almost extinction-proof. However, this perception proved to be wrong. Currently, 33% of global fish populations are exploited at biologically unsustainable levels (FAO 2018). In the second half of the 20th century, the fraction of large predatory fishes in global catches, that have increased by fivefold, have declined (Worm et al. 2009). By the turn of the millennia, large predatory fish biomasses were estimated to be only about one tenth of pre-industrial abundance (Myers and Worm 2003).

There has been much public concern over declining fish stocks on a global scale, spurring multiple United Nations resolutions on the conservation and management of fish stocks (UN 2019). Most management measurements have focused on reducing exploitation rates, which generally result in an improvement for fully exploited and overexploited stocks (Worm et al. 2009). However, reductions in fishing pressure, although necessary, are not always sufficient for recovery of depleted populations, which are estimated to be 13% of global exploited fish populations (Garcia et al. 2018). Indeed, a long-term dataset from more than 230 fish populations shows that most collapsed populations exhibit little or no recovery 15 years after collapse (Hutchings and Reynolds 2004). Only 12% of collapsed marine populations had fully recovered in this period, all of them clupeids (Hutchings and Reynolds 2004), which typically have a low trophic level in marine food webs. In contrast, even when fishing mortality has been reduced, collapsed fish populations of species in higher trophic levels exhibit little or no recovery as much as 15 years later (Hutchings and Reynolds 2004). An emblematic example is the Canadian cod collapse, which despite a prolonged moratorium in effect, exhibits no recovery after two decades (Maroto and Moran 2014; Pedersen et al. 2017).

Factors impeding recovery of depleted fish populations can be numerous including biological, environmental and socioeconomic factors (Garcia et al. 2018). Some biological factors preventing a successful recovery of depleted fish stocks are altered ecological interactions and intrinsic characteristics of the populations, i.e., life history traits. Environmental factors include loss of habitat, in particular, coastal and coral reef degradation, as well as natural oscillations in climatic and productivity variables that condition the recruitment into the depleted stock (Friedland et al. 2009). Several socioeconomic factors have been identified to cause the failure of recovery of depleted fish stocks, for instance, the inability to find alternative livelihoods for fishermen after the depletion preventing fishing mortality to be reduced, social organization and cohesion that may facilitate compliance or illegal behavior (Wakeford et al. 2009). In this chapter, I will focus on the biological factors that impair recovery of depleted predatory fish populations in more detail.

There is general consensus that life history traits affect the vulnerability and capacity to recover of fish populations. In particular, traits associated with maximum population growth include fecundity, body growth rate, age and size at maturity (Hutchings and Reynolds 2004). Although low fecundity is characteristic of fish species with low population growth rate (Musick 1999), high fecundity does not ensure recovery (Hutchings and Reynolds 2004; Dulvy et al. 2003; Hutchings and Kuperinen 2014). In fact, highly fecund species, such as Atlantic cod (DFO 2003), exhibit lack of recovery following a decrease in fishing pressure. Rapid body growth rate is associated with high potential to recover (Hutchings and Reynolds 2004). Age and size at maturity are negatively associated with maximum population growth and, thus, with recovery potential (Roff 2002; Denney et al. 2002; Dulvy and Reynolds 2002; Dulvy et al. 2003).

Biological Characteristics of Predatory Populations

Although some life history traits are specific of species or even populations, most predatory fish species are characterized by a significant increase in body size through ontogeny that can span four orders of magnitude (Werner and Gilliam 1984). This increase in body size results in changes in ecological interactions throughout an individual's life. In fact, the majority of predatory fishes do not have a predatory diet as small juveniles because its body size at birth is often similar to that of their prey (de Roos and Persson 2013). Therefore, individuals of predatory fish species prey on smaller fish species only when they reach a large body size, while small individuals necessarily feed upon other type of food that might be the same of their future prey. As a result, individuals of a predator and prey species may be engaged in a competitive or a predatory interaction depending on the life stage of the predator individual interacting. Furthermore, throughout life history, predatory interactions can be reversed, for instance, while individuals of the predator species in the adult stage feed upon large juveniles or adults of the prey species, the latter may prey upon small juveniles of the predator species (Gårdmark et al. 2015).

Depensation

As a consequence of the life history specifics, large predatory fish species described above may exhibit a reduced population growth capacity at low densities (i.e., below a threshold), a phenomenon known as depensation or Allee effect (Figure 1) (de Roos and Persson 2002).

Strong depensation effects are caused by the existence of alternative stable states. In this view, a high and a low population density constitute alternative states stabilized by internal feedbacks, and therefore, a population depleted to its low density by overexploitation cannot return to its high-density state when the exploitation rate decreases to the point at which the population has collapsed. Instead, the exploitation rate should be reduced to lower levels for the population to recover (Figure 2). The lag between the exploitation rate at which collapse and recovery occurs is known as hysteresis.

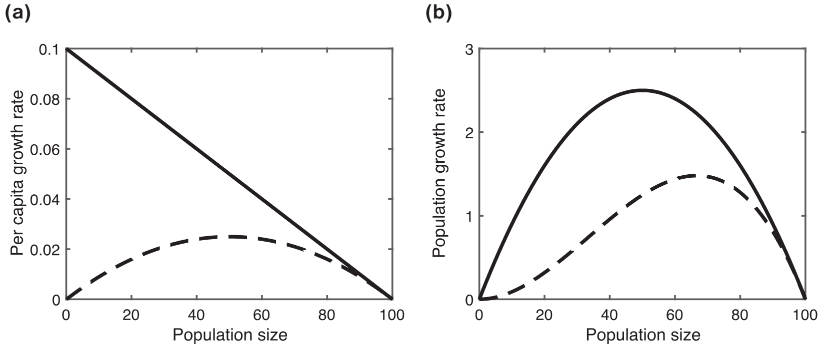


FIGURE 1 Depensation or Allee effects occur when population growth rate is reduced at small population size. (a) The per capita growth rate decreases linearly with population size when there are no Allee effects (solid line), while it is reduced at low and at high population size when there are Allee effects (dashed line). (b) As a consequence, in the absence of Allee effects (solid line), the population grows at its maximal rate at intermediate population size, while in the presence of Allee effects (dashed line), the population grows at its maximal rate when the population is large. Both populations have a carrying capacity of 100.

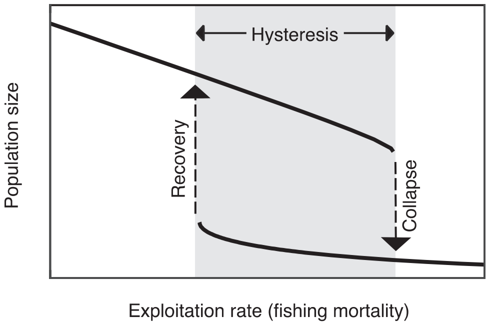


FIGURE 2 Alternative stable population states cause depensation. A population that is unexploited or exploited at low rate has a large size. As exploitation rate gradually increases, the population size gradually decreases until the exploitation rate is near the value at which the population collapses. Owing to hysteresis caused by alternative stable states, after the collapse, it is not enough for the population to recover to reduce the exploitation rate to the value where the collapse occurred. A further reduction in exploitation rate is required for the population to recover.

Several mechanisms have been hypothesized to cause depensation in fish populations that can be classified into three categories: reproduction difficulties, impaired social relations and altered ecological interactions, which are addressed in turn below.

Reproduction difficulties may emerge from: (1) sex ratio abnormalities, for instance, when fishing removes faster one sex than the other (Garcia et al. 2018); (2) insufficient spawner densities for effective egg fertilization and successful mating encounters and (3) removal of larger individuals because of their increased value and fishing tactics (i.e., minimum mesh and fish landing size (Hutchings and Reynolds 2004)). It is important to bear in mind that fecundity increases exponentially with female fish size (McIntyre and Hutchings 2003). Hence, larger fish can contribute disproportionately to population birth rate through greater weight-specific fertility and larval quality (Sogard 1997; Venturelli et al. 2012; Trippel 2004). As a consequence, populations composed of small individuals will have reduced reproductive potential (Scott et al. 2011) (migratory fish populations may be an exception when facing high costs of breeding migration (Chaparro-Pedraza and de Roos 2019)).

Depensation may be caused by impaired social relations at low population densities. For instance, by altering social behaviors such as schooling or migration, fishing may affect the reproductive process or survival of the remaining population as well as transmission of knowledge about suitable feeding and breeding areas and migration pathways from old-year classes to young ones (Petitgas et al. 2010).

Ecological interactions can cause depensation through various mechanisms. One proposed mechanism arises from the interaction between a size-selective predator and preys with density-dependent growth (de Roos and Persson 2002). Another mechanism emerges from competition between a predator and its prey in early life stages. As mentioned above, adult predatory fish individuals feed on prey fish, whereas small predatory juveniles and prey fish may compete for the same food resources (de Roos and Persson 2013). At high population density, adult predatory individuals, by controlling the prey population at low or intermediate densities through predation, reduce competition pressure on the predatory juveniles (Gårdmark et al. 2015). A release of the top-down control results in high density of prey population that controls the predatory population recruitment through competition. Alternatively, a third mechanism emerges when adult prey individuals feed upon early life stages of the predatory fish population, the recruitment in the latter is controlled by predation of abundant adult prey individuals when the predatory population is at low density.

Although theoretical work indicates that depensation through these mechanisms is likely (de Roos and Persson 2002; Maroto and Moran 2014), it has been difficult to demonstrate empirically (Garcia et al. 2018). However, recent empirical studies have revealed evidence supporting these depensatory mechanisms.

Depensation in Wild Predatory Fish

The northwest population of Atlantic cod (*Gadus morhua*) is an emblematic case of collapse and little recovery in fisheries. This collapse, due to overfishing, had devastating socioeconomic impacts. After dramatic declines of the population during the 1980s and 1990s, Canada instituted a moratorium on fishing in 1992 that left 40,000 workers unemployed, thousands left fishery to work in other trades or professions, in many cases, displaced to other parts of the country (Garcia et al. 2018). Despite this long moratorium, cod population have shown only slight increases (Pedersen et al. 2017). An analysis of the population dynamics of depleted fish stocks concluded that the lack of recovery in the Canadian cod is consistent with a depensatory phenomenon at low population densities (Maroto and Moran 2014). Likewise, the depensatory phenomenon was confirmed by a population model, which concluded that the lack of recovery of the cod population was mainly caused by predation by gray seals driving an Allee effect (Neuenhoff et al. 2018).

At the beginning of 1990s, the cod population in the Baltic Sea also collapsed as a result of overfishing concurrent with poor environmental conditions (Gårdmark et al. 2015). Following this collapse, the population did not recover during the late 1990s and early 2000s. The lack of recovery has been postulated to result from increased competition between cod larvae and sprat (*Sprattus sprattus*), a main prey of adult cod (Möllmann et al. 2009). As a consequence of the collapse of the cod population, the sprat population increased due to release from predation pressure. This caused stronger competition pressure on cod larvae for zooplankton prey. A shift in the regulation of zooplankton dynamics, from a control by limiting resources (bottom up) before the collapse of the cod population to a control by predation of sprat population (top down) after the collapse, supports this hypothesis (Casini et al. 2008). Gårdmark et al. (2015) demonstrate that the lack of top-down regulation in the sprat population after the cod collapse caused a change in the distribution of size classes in this population. As a consequence of this change, the food available for the cod in larval and adult stages was diminished and, therefore, the viability of this population reduced.

In the mid-1990s, pike (*Esox lucius*) population, a large top predator, declined substantially in the Baltic Sea (Bergström et al. 2015; Eriksson et al. 2011; Ljunggren et al. 2010; Nilsson et al. 2004). Pike is a major predator on adult threespine stickleback fish (*Gasterosteus aculeatus*) (Donadi et al. 2017).

Therefore, the decline of pike has favored the stickleback population by predator release, causing the stickleback population to grow exponentially about a decade after the collapse of the pike population (Bergström et al. 2015; Bergström et al. 2016; Donadi et al. 2017; Sieben et al. 2011). Following the collapse of the pike population in the Baltic Sea, various conservation measurements have been adopted with the aim of enabling the recovery of the pike population, including partial closure of the pike fishery and habitat restoration (Engstedt et al. 2017; Nilsson et al. 2014). In spite of this, the recovery of the pike population has been very slow, if at all (Bergström et al. 2016). A recent experimental study suggests that the cause of the impaired capacity of the pike population to recover is a depensatory effect. This study demonstrates that adult stickleback predation on pike larvae heavily reduces the recruitment of the pike population, suggesting this as the mechanism causing depensation and impeding recovery of the pike population (Nilsson et al. 2019).

The possibility of depensatory effects in severely depleted fish predatory populations may be high. However, the evidence remains scarce beyond few examples, perhaps because few populations have been depleted to densities where the population growth rate is heavily reduced (the depensatory effect or Allee effect is conspicuous) (Garcia et al. 2018). In fact, pike population in the Baltic Sea and the Canadian Cod population were depleted below 10% (Nilsson et al. 2019) and 1% (Myers and Worm 2005), respectively, of their original abundance. The collapse of these populations is, in fact, beyond the precautionary biomass limit usually imposed by fisheries that is 20%; at this limit, populations are considered heavily depleted (Garcia et al. 2018).

Evolutionary Responses to Exploitation

Most fisheries target specific age or size classes within a population. Typically, larger and older, individuals are more likely to be caught. There is compelling evidence supporting claims that large fish were caught more frequently in the early years of industrial fisheries (Allendorf and Hard 2009). For instance, cod individuals weighting more than 30 kg were frequently caught in New England waters before 1900 and even a gigantic individual of nearly 100 kg was taken off Massachusetts in 1895 (Jordan and Evermann 1902). Such selective mortality imposed on the targeted phenotypes brings about evolutionary change if the characteristics that make these phenotypes more vulnerable are heritable. Accordingly, rapid changes in phenotypic traits of fish populations have been documented as a result of fishing mortality (Haugen and Vøllestad 2001; Law 2001; Conover et al. 2007; Allendorf and Hard 2009).

Besides the immediate negative impact in the population birth rate due to loss of high fecundity potential, increased mortality of large individuals results in selection against fast body growth rate (Biro and Post 2008; Swain et al. 2007); as opposed to natural selection that usually favors this trait in unexploited populations. As a consequence of this evolutionary change in response to selective pressures induced by fishing, depleted populations are composed mainly of phenotypes with slow body growth rate, a trait associated with low potential to recover (Hutchings and Reynolds 2004). In addition to changes in growth rate, fishing pressure can alter size and age at maturity (Heino and Dieckmann 2008; Heino 1998). These and other evolutionary responses tend to improve resilience of the exploited populations to fishing mortality as slow-growth and long-lived species accelerate their turnover (Pauly 1995). However, simultaneously, these evolutionary responses may reduce trait variation and, thus, resilience of these populations to the environmental oscillations to which they were adapted through natural selection.

Since mortality caused by fishing is usually considerably higher than natural mortality, selection pressures are higher in the former. As a consequence, evolutionary changes induced by fishing pressure may emerge in few decades; however, once fishing pressure is released, it may take centuries to be reversed (Enberg et al. 2009). There is growing concern that undermined genetic variation caused by fisheries-induced evolutionary change might increase the range of hysteresis between collapse and recovery trajectories and thus enhance the effect of depensation through the mechanisms mentioned in the previous section. Therefore, evolutionary processes that may improve resilience of exploited populations, may lessen the capacity to recover once depleted.

Population Size, Environmental and Demographic Stochasticity

In addition to the negative effects on genetic variation, reduced population size makes depleted populations more vulnerable to extinction as a consequence of demographic and environmental stochasticity. Small populations are largely governed by the misfortunes of each individual. Hence, the persistence of small populations is largely dependent on realizations of random events that determine individual survival and reproduction. These random events are the result of unpredictable factors that influence the survival and fecundity of some (demographic stochasticity) or all (environmental stochasticity) individuals in a population (Lande 1993).

Along with the decline of a population, its geographical range tends to contract as individuals remain in the most favorable habitats (Swain and Wade 2010). Such geographical range contraction can further exacerbate the risk of extinction in small populations as a consequence of stochastic events. Northern cod provides an example of how such range contractions increase the risk of exposing collapsed populations to environmentally stochastic events. In April 2003, about 1500 tones of mature cod were frozen to death in a very small area of Trinity Bay along the northeast coast of Newfoundland (Hoag 2003; Hutchings and Reynolds 2004). An unpredictable oceanographic event that trapped cod in supercooled water provides the most likely explanation for this cull mass (Hutchings and Reynolds 2004). Relative to the small geographical area where the event took place, it had a disproportionally large negative effect on the population recovery because the spawning population of northern cod in 2003 was estimated to be only 14,000 tones (DFO 2003). Hence, this environmentally stochastic event in a very small geographical area probably resulted in the death of more than 10% of the breeding population.

Rebuilding Predatory Fish Populations after Depletion

Most plans for rebuilding fisheries focused on populations that have been driven below acceptable levels of productivity. These plans advocate measures that are adequate to avoid depletion, and it is often assumed that if the population has been driven to depletion, the same measures, more effectively applied, are also those that will enable the rebuilding of the population (Garcia et al. 2018). However, this assumption may be misleading. When a population has been depleted to very low level, a set of additional measures are required in order to avoid extinction and, if possible, increase the chances of recovery (Garcia et al. 2018).

Measures to reduce fishing mortality are necessary for the rebuilding of a depleted population. Rebuilding plans relying on a substantial reduction at the onset rather than on incremental minor reductions over time are more successful (Murawski 2017). However, socioeconomic and governance factors can prevent fishing mortality to be substantially reduced in the onset. For instance, the inability to find alternative livelihoods for fishermen following the depletion causes exploitation to continue, and the lack of social control facilitates illegal catch (Wakeford et al. 2009). Therefore, a successful rebuilding plan should ensure environmental and social sustainability pursuing the rebuilding of the natural resource and the people's livelihoods in concert (FAO 1995).

In addition to reductions in fishing mortality, protection of critical habitats and enhancement of stock resilience are common measures in rebuilding plans for populations that have been only lightly overfished. These measures are certainly necessary also for rebuilding populations that have been deeply overfished. However, if depensation effects govern the dynamics of the depleted population, the recovery may be slower than expected (Murawski 2017), which increases socioeconomic costs associated to the rebuilding process. Hence, implementing additional measures aim at destabilizing the present poor (small population size) stable state may speed up the rebuilding process and, thus, increase its chances of success.

Depensatory, evolutionary and stochastic effects have hampered the recovery of predatory fish populations in spite of efforts to reduce the fishing pressure and restore the habitat of these species. This indicates that, perhaps, a small population size (i.e., low density) is a stable state for these populations, and thus, additional actions that destabilize this state may be required. For instance, Persson et al. (2007)

demonstrated experimentally in a freshwater lake that the interaction between a size-selective predator and preys with density-dependent growth that has been proposed as one of the mechanisms causing depensation in predatory fish populations (de Roos and Persson 2002) can be destabilized by culling of the prey population. In this experiment, the removal of prey individuals proved to be an effective measure to trigger the recovery of the predator population.

Similarly, counterintuitive measures to reverse the negative evolutionary effects caused by size-selective fishing mortality may improve the capacity to recover of the predatory population. As mentioned before, natural selection is usually weaker than selection imposed by fishing mortality, therefore increasing the strength of selection in the same direction than natural selection may speed up the process to reverse the evolutionary effects of selection induced by fishing. For instance, while increased mortality of large size classes due to fishing results in selection against fast growing phenotypes, increased mortality of small size classes will favor these phenotypes. Indeed, experimental evidence shows that juvenile growth rates increases with high mortality of small size classes as a consequence of genetic change in somatic growth rate (Conover and Munch 2002). Furthermore, this study found that populations in which small individuals are harvested have higher yield in the long term than populations in which large individuals are harvested. Therefore, selection in favor of fast growing phenotypes through removal of small individuals in a depleted population may rapidly raise the frequency of fast growing phenotypes, increasing the potential to recover.

In this chapter, I have presented biological factors that affect the recovery of predatory fish populations. It is important to bear in mind that fisheries are social-ecological systems in which the natural and human components tightly interact and coevolve (Berkes et al. 2000). The biological factors described in this chapter interact with economic, social and governance factors in a complex manner. Therefore, the measures targeting biological factors that condition the rebuilding of predatory fish populations must account for the interactions of these with socioeconomic factors.

References

Allendorf, F. W., and J. J. Hard. 2009. “Human-Induced Evolution Caused by Unnatural Selection through Harvest of Wild Animals.” *Proceedings of the National Academy of Sciences* 106 (Supplement_1): 9987–9994. doi:10.1073/pnas.0901069106.

Bergström, L., O. Heikinheimo, R. Svirgsden, E. Kruze, L. Ložys, A. Lappalainen, L. Saks, et al. 2016. “Long Term Changes in the Status of Coastal Fish in the Baltic Sea.” *Estuarine, Coastal and Shelf Science* 169: 74–84. doi:10.1016/j.ecss.2015.12.013.

Bergström, Ulf, Jens Olsson, Michele Casini, Britas Klemens Eriksson, Ronny Fredriksson, Håkan Wennhage, and Magnus Appelberg. 2015. “Stickleback Increase in the Baltic Sea – A Thorny Issue for Coastal Predatory Fish.” *Estuarine, Coastal and Shelf Science* 163 (PB): 134–142. doi:10.1016/j.ecss.2015.06.017.

Berkes, Fikret, Carl Folke, and Johan Colding. 2000. Linking Social and Ecological Systems: Management Practices and Social Mechanisms for Building Resilience. Cambridge: Cambridge University Press, 2000.

Biro, P. A., and J. R. Post. 2008. “Rapid Depletion of Genotypes with Fast Growth and Bold Personality Traits from Harvested Fish Populations.” *Proceedings of the National Academy of Sciences* 105 (8): 2919–2922. doi:10.1073/pnas.0708159105.

Casini, M., J. Hjelm, J.-C. Molinero, J. Lovgren, M. Cardinale, V. Bartolino, A. Belgrano, and G. Kornilovs. 2008. “Trophic Cascades Promote Threshold-like Shifts in Pelagic Marine Ecosystems.” *Proceedings of the National Academy of Sciences* 106 (1): 197–202. doi:10.1073/pnas.0806649105.

Chaparro-Pedraza, P. Catalina, and André M. de Roos. 2019. “Low Marine Food Levels Mitigate High Migration Costs in Anadromous Populations.” *BioRxiv*, January, 577353. doi:10.1101/577353.

Conover, David O., and Stephan B. Munch. 2002. “Sustaining Fisheries Yields Over Evolutionary Time Scales.” *Science* 297 (5578): 94–96. doi:10.1126/science.1074085.

Conover, David O., Stephan B. Munch, and David Conover. 2007. “Sustaining Fisheries Yields Over Evolutionary Time Scales.” *Time* 297 (5578): 94–96. doi:10.1126/science.1074085.

de Roos, A. M., and L. Persson. 2002. “Size-Dependent Life-History Traits Promote Catastrophic Collapses of Top Predators.” *Proceedings of the National Academy of Sciences* 99 (20): 12907–12912. doi:10.1073/pnas.192174199.

de Roos, André M., and Lennart Persson. 2013. Population and Community Ecology of Ontogenetic Development. Princeton: Princeton University Press.

Denney, Nicola H., Simon Jennings, and John D. Reynolds. 2002. “Life-History Correlates of Maximum Population Growth Rates in Marine Fishes.” *Proceedings of the Royal Society B: Biological Sciences* 269 (1506): 2229–2237. doi:10.1098/rspb.2002.2138.

DFO, Deparment of Fisheries and Oceans. 2003. “Northern (2J+3KL) Cod.” Ottawa (Canada).

Donadi, S., B. K. Eriksson, P. Jacobson, J. P. Hansen, Å. N. Austin, M. van Regteren, U. Bergström, J. S. Eklöf, and G. Sundblad. 2017. “A Cross-Scale Trophic Cascade from Large Predatory Fish to Algae in Coastal Ecosystems.” *Proceedings of the Royal Society B: Biological Sciences* 284 (1859): 20170045. doi:10.1098/rspb.2017.0045.

Dulvy, Nicholas K., and John D. Reynolds. 2002. “Predicting Extinction Vulnerability in Skates.” *Conservation Biology* 16 (2): 440–450. doi:10.1046/j.1523-1739.2002.00416.x.

Dulvy, Nicholas K., Yvonne Sadovy, and John D. Reynolds. 2003. “Extinction Vulnerability in Marine Populations.” *Fish and Fisheries* 4 (1): 25–64. doi:10.1046/j.1467-2979.2003.00105.x.

Enberg, Katja, Christian Jørgensen, Erin S. Dunlop, Mikko Heino, and Ulf Dieckmann. 2009. “Implications of Fisheries-Induced Evolution for Stock Rebuilding and Recovery.” *Evolutionary Applications* 2 (3): 394–414. doi:10.1111/j.1752-4571.2009.00077.x.

Engstedt, O., J. Nilsson, and P. Larsson. 2017. “Habitat Restoration – A Sustainable Key to Management.” In Biology and Ecology of Pike, edited by C. Skov and A. Nilsson, 248–268. Boca Ratón: CRC Press, 2018, 1. p. 248-268.

Eriksson, Britas Klemens, Katrin Sieben, Johan Eklf, Lars Ljunggren, Jens Olsson, Michele Casini, and Ulf Bergstrm. 2011. “Effects of Altered Offshore Food Webs on Coastal Ecosystems Emphasize the Need for Cross-Ecosystem Management.” *Ambio* 40 (7): 786–797. doi:10.1007/s13280-011-0158-0.

FAO. 1995. “Code of Conduct for Responsible Fisheries.” Food & Agriculture Organization.

FAO. 2018. The State of World Fisheries and Aquaculture. doi:10.1126/science.aaw5824.

Friedland, Kevin D., Julian C. MacLean, Lars P. Hansen, Arnaud J. Peyronnet, Lars Karlsson, David G. Reddin, Niall Ó. Maoiléidigh, and Jennifer L. McCarthy. 2009. “The Recruitment of Atlantic Salmon in Europe.” *ICES Journal of Marine Science* 66 (2): 289–304. doi:10.1093/icesjms/fsn210.

Garcia, S. M., Yimin Ye, Rice Jake, and A. Charles. 2018. Rebuilding of Marine Fisheries. Part 1: Global Review. FAO Fisheries and Aquaculture Technical Paper 630/1. FAO.

Gårdmark, Anna, Michele Casini, Magnus Huss, Anieke Van Leeuwen, Joakim Hjelm, Lennart Persson, and André M. de Roos. 2015. “Regime Shifts in Exploited Marine Food Webs: Detecting Mechanisms Underlying Alternative Stable States Using Sizestructured Community Dynamics Theory.” *Philosophical Transactions of the Royal Society B: Biological Sciences* 370 (1659): 1–10. doi:10.1098/rstb.2013.0262.

Haugen, Thrond O., and Leif Asbjørn Vøllestad. 2001. “A Century of Life-History Evolution in Grayling.” *Genetica* 112–113: 475–491. doi:10.1023/A:1013315116795.

Heino, Mikko. 1998. “Management of Evolving Fish Stocks.” *Canadian Journal of Fisheries and Aquatic Sciences* 55 (8): 1971–1982. doi:10.1139/f98-089.

Heino, Mikko, and Ulf Dieckmann. 2008. “Detecting Fisheries-Induced Life-History Evolution: An Overview of the Reaction-Norm Approach.” *Bulletin of Marine Science* 83 (1): 69–93.

Hoag, H. 2003. “Atlantic Cod Meet Icy Death.” *Nature* 422 (6934): 792. doi:10.1038/422792a.

Hutchings, J. A., and A. Kuperinen. 2014. “Ghosts of Fisheries-Induced Depletions: Do They Haunt Us Still?” *ICES Journal of Marine Science* 71 (6): 1467–1473. doi:10.1093/icesjms/fst048.

Hutchings, J. A., and J. D. Reynolds. 2004. "Marine Fish Population Collapses: Consequences for Recovery and Extinction Risk." *BioScience* 54 (4): 297–309. doi:10.1641/0006-3568(2004)054[0297:mfpccf]2.0.co;2.

Jordan, D. S., and B. W. Evermann. 1902. "American Food and Game Fishes: A Popular Account of All the Species Found in America North of the Equator, with Keys for Ready Identification, Life Histories and Methods of Capture." Doubleday, New York.

Lande, Russell. 1993. "Risks of Population Extinction from Demographic and Environmental Stochasticity and Random Catastrophes." *The American Naturalist* 142 (6): 911–927.

Law, R. 2001. "No Phenotypic and Genetic Changes Due to Exploitation." In Conservation of Exploited Species, edited by J. D. Reynolds, G. M. Mace, K. H. Redford, and J. G. Robinson, 323–342. Cambridge: Cambridge University Press.

Ljunggren, Lars, Lars Ljunggren, Alfred Sandstrom, Ulf Bergstrom, Johanna Mattila, Antti Lappalainen, Gustav Johansson, et al. 2010. "Recruitment Failure of Coastal Predatory Fish in the Baltic Sea Coincident with an Offshore Ecosystem Regime Shift." *ICES Journal of Marine Science* 67 (8): 1587–1595. doi:10.1093/icesjms/fsq109.

Maroto, Jose M., and Manuel Moran. 2014. "Detecting the Presence of Depensation in Collapsed Fisheries: The Case of the Northern Cod Stock." *Ecological Economics* 97: 101–109. doi:10.1016/j.ecolecon.2013.11.006.

McIntyre, Tara M., and Jeffrey A. Hutchings. 2003. "Small-Scale Temporal and Spatial Variation in Atlantic Cod (Gadus Morhua) Life History." *Canadian Journal of Fisheries and Aquatic Sciences* 60 (9): 1111–1121. doi:10.1139/f03-090.

Möllmann, Christian, Rabea Diekmann, Bürbel Müller-karulis, Georgs Kornilovs, Maris Plikshs, and Philip Axe. 2009. "Reorganization of a Large Marine Ecosystem Due to Atmospheric and Anthropogenic Pressure: A Discontinuous Regime Shift in the Central Baltic Sea." Global *Change Biology* 15 (6): 1377–1393. doi:10.1111/j.1365-2486.2008.01814.x.

Murawski, Steven A. 2017. "Keynote Rebuilding Depleted Fish Stocks : The Good, the Bad, and, Mostly, the Ugly." *ICES Journal of Marine Science*, 67(9), 1830–1840.

Musick, J A. 1999. "Life in the Slow Lane: Ecology Andconservation of Long-Lived Marine Animals." In American Fisheries Society Symposium, Vol. 23, American Fisheries Society, Bethesda, Maryland, USA.

Myers, R. A., and Boris Worm. 2003. "Rapid Worldwide Depleation of Predatory Fish Communities." *Nature* 423 (May): 280–283.

Myers, Ransom A., and Boris Worm. 2005. "Extinction, Survival or Recovery of Large Predatory Fishes." *Philosophical Transactions of the Royal Society B: Biological Sciences* 360 (1453): 13–20. doi:10.1098/rstb.2004.1573.

Neuenhoff, Rachel D., Douglas P. Swain, Sean P. Cox, Murdoch K. McAllister, Andrew W. Trites, Carl J. Walters, and Mike O. Hammill. 2018. "Continued Decline of a Collapsed Population of Atlantic Cod (Gadus Morhua) Due to Predation-Driven Allee Effects." *Canadian Journal of Fisheries and Aquatic Sciences* 76 (1): 168–184. doi:10.1139/cjfas-2017-0190.

Nilsson, Jonas, Jan Andersson, Peter Karås, and Olof Sandström. 2004. "Recruitment Failure and Decreasing Catches of Perch (Perca Fluviatilis L.) and Pike (Esox Lucius L.) in the Coastal Waters of Southeast Sweden." *Boreal Environment Research* 9 (4): 295–306.

Nilsson, Jonas, Olof Engstedt, and Per Larsson. 2014. "Wetlands for Northern Pike (Esox Lucius L.) Recruitment in the Baltic Sea." *Hydrobiologia* 721 (1): 145–154. doi:10.1007/s10750-013-1656-9.

Nilsson, Jonas, Henrik Flink, and Petter Tibblin. 2019. "Predator-prey Role Reversal May Impair the Recovery of Declining Pike Populations." *Journal of Animal Ecology* 88: 927–939. doi:10.1111/1365-2656.12981.

Pauly, Daniel. 1995. "Anecdotes and the Shifting Baseline Syndrome of Fisheries." *Trends in Ecology & Evolution* 10 (10): 430. doi:10.1016/S0169-5347(00)89171-5.

Pedersen, Eric J., Heike Link, Patrick L. Thompson, Zofia E. Taranu, Andrew Gonzalez, R. Aaron Ball, Charlotte Moritz, et al. 2017. "Signatures of the Collapse and Incipient Recovery of an Overexploited Marine Ecosystem." *Royal Society Open Science* 4 (7): 170215. doi:10.1098/rsos.170215.

Persson, L., P.-A. Amundsen, A. M. De Roos, A. Klemetsen, R. Knudsen, and R. Primicerio. 2007. "Culling Prey Promotes Predator Recovery--Alternative States in a Whole-Lake Experiment." *Science* 316 (5832): 1743–1746. doi:10.1126/science.1141412.

Petitgas, P., D. H. Secor, I. McQuinn, G. Huse, and N. Lo. 2010. "Stock Collapses and Their Recovery: Mechanisms That Establish and Maintain Life-Cycle Closure in Space and Time." *ICES Journal of Marine Science* 67: 1841–1848. doi:10.1093/icesjms/fsq082.

Roff, Derek A. 2002. "Life History Evolution". Sunderland, Massachusetts: Sinauer Associattes.

Scott, Beth, Gudrun Marteinsdottir, and Peter Wright. 2011. "Potential Effects of Maternal Factors on Spawning Stock–Recruitment Relationships under Varying Fishing Pressure." *Canadian Journal of Fisheries and Aquatic Sciences* 56 (10): 1882–1890. doi:10.1139/f99-125.

Sieben, Katrin, Lars Ljunggren, Ulf Bergström, and Britas Klemens Eriksson. 2011. "A Meso-Predator Release of Stickleback Promotes Recruitment of Macroalgae in the Baltic Sea." *Journal of Experimental Marine Biology and Ecology* 397 (2): 79–84. doi:10.1016/j.jembe.2010.11.020.

Sogard, Susan M. 1997. "Size Selective Mortality in the Juvenile Stages of Teleost Fishes: A Review." *Bulletin of Marine Science* 60 (3): 1129–1157.

Swain, Douglas P., Alan F. Sinclair, and J. Mark Hanson. 2007. "Evolutionary Response to Size-Selective Mortality in an Exploited Fish Population." *Proceedings of the Royal Society B: Biological Sciences* 274 (1613): 1015–1022. doi:10.1098/rspb.2006.0275.

Swain, D. P., and E. J. Wade. 2010. "Density-Dependent Geographic Distribution of Atlantic Cod (Gadus Morhua) in the Southern Gulf of St. Lawrence." *Canadian Journal of Fisheries and Aquatic Sciences* 50 (4): 725–733. doi:10.1139/f93-083.

Trippel, Edward A. 2004. "Egg Size and Viability and Seasonal Offspring Production of Young Atlantic Cod." *Transactions of the American Fisheries Society* 127 (3): 339–359. doi:10.1577/1548-8659(1998)127<0339:esavas>2.0.co;2.

UN. 2019. "General Assembly Resolutions and Decisions." *Oceans and the Law of the Sea in the General Assembly of the United Nations*. https://www.un.org/Depts/los/general_assembly/general_assembly_resolutions.htm.

Venturelli, Paul A., Brian J. Shuter, and Cheryl A. Murphy. 2012. "Evidence for Harvest-Induced Maternal Influences on the Reproductive Rates of Fish Populations." *Proceedings of the Royal Society B: Biological Sciences* 276 (1658): 919–924. doi:10.1098/rspb.2008.1507.

Wakeford, Robert C., David J. Agnew, and Christopher C. Mees. 2009. "Review of Institutional Arrangements and Evaluation of Factors Associated with Successful Stock Recovery Plans." *Reviews in Fisheries Science* 17 (2): 190–222. doi:10.1080/10641260802667075.

Werner, E. E., and J. F. Gilliam. 1984. "The Ontogenetic Niche and Species Interactions in Size-Structured Populations." *Ecology* 15: 393–425.

Worm, Boris, Ray Hilborn, Julia K. Baum, Trevor A. Branch, Jeremy S. Collie, Christopher Costello, Michael J. Fogarty, et al. 2009. "Rebuilding Global Fisheries." *Science* 325 (5940): 578–585. doi:10.1126/science.1173146.

38

Bioremediation

Ragini Gothalwal

Introduction

After the industrial revolution, a significant number of industries were set up, which released organic compounds in the environment. These chemical compounds (pesticides, fertilizers, plastic, dyes, and heavy metals) are present in the form of metal, non-metal, metalloid, inorganic compounds, and organic compounds. Organic compounds are divided into aliphatic, alicyclic, aromatic, polyaromatic, and mixed types. These compounds have a tendency to transform into another compound and reach the ecosystem. Owing to the inadequacy and increased cost of physical and chemical methods available, the need for ecofriendly and cheaper methods has been strongly realized. Bioremediation has the potential to lower the cost of remediation by orders of magnitude over alternative technologies, and governmental agencies and the public as taxpayers are interested in cost-effective cleanup technologies.

Biodegradation and bioremediation are matching processes, to the extent that they are both based on the conversion or metabolism of pollutants by microorganisms. The difference between these two is that biodegradation is a natural process whereas bioremediation is a technology. Bioremediation requires an efficient bacterial strain that can degrade the largest pollutant to a minimum level. Microbial diversity offers an immense field of environmental friendly options for mineralization of contaminants or their transformations into less harmful or non-hazardous compounds. Molecular biology methods are now being employed to study bioremediation. Conjugative gene transfer occurs, through plasmid-borne catabolic genes, between bacteria in oil and the competitive, indigenous bacterial population. Therefore, it is important to understand the role of catabolic genes, by molecular cloning and characterization, in the degradation of a particular organic compound, so that it can be applied for bioremediation. This review emphasizes the distribution and extent of environmental contaminants from the ecosystem. It also accentuates the current status of research in the area of biodegradation and bioremediation and presents the new approaches available for microbial and phyto tracking in the environment.

Principle of Bioremediation

The understanding of the bioremediation process is derived from the combination of biochemical and microbiological processes,[1,2] summarized as follows:

1. The relationship of comparative biochemistry applies equally to the axenic culture of microorganisms and those in field soil.
2. Microbial growth requirements are the same whether in laboratory culture or in the field.
3. Limitations resulting from ecological interactions include the need to accommodate the presence of other microbes as well as adaptation to the physical and chemical properties of the microsite wherein the microbes live.
4. A microbe amended in a soil ecosystem not only should pass the requisite genetic information and be capable of expressing that capability *in situ,* but it must also have the capacity to become a part of the overall soil microbial community.[1] Such microbes are referred to as competent rhizospheric bacteria with intrinsic bioremediation potential.

Targets of bioremediation must be based not only in terms of structure but also in terms of the matrix containing the target. Thus, the applicability of bioremediation can be considered for each of the environmental status of matter: 1) solid (soil, sediment, sludge); 2) liquid (ground water, industrial wastewater); 3) gas (industrial air emission); and 4) subsurface environment (saturated and vadose zones). The general approaches to bioremediation (Figure 1) are as follows:

1. Bioaugmentation;
2. Biostimulation (environmental modification, through the nutrient application process);
3. Phytoremediation (addition of microbes).

The biological community for bioremediation generally consists of the natural soil microflora. However, higher plants can also be manipulated to enhance toxicant removal (phytoremediation), especially for remediation of metal-contaminated soils.[3,4]

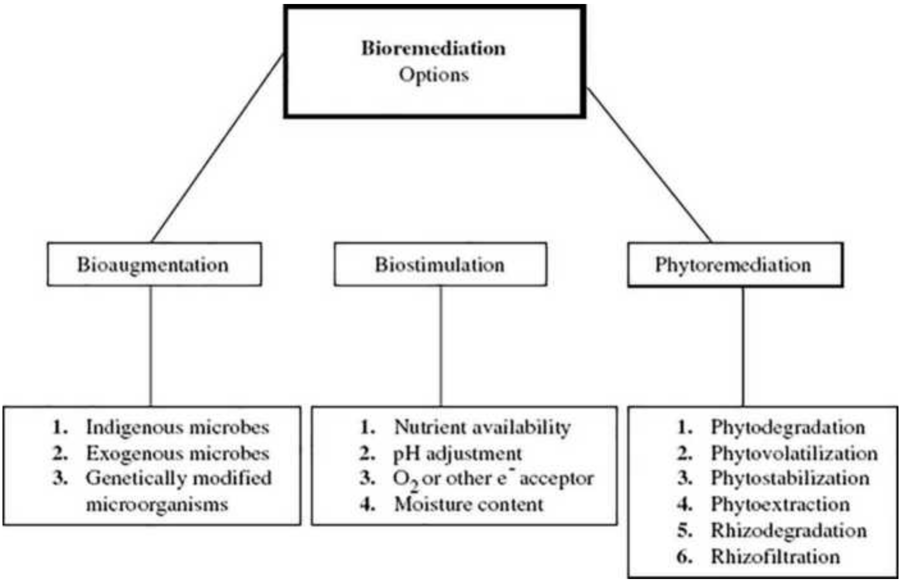


FIGURE 1 Bioremediation approaches for environmental cleanup.
Source: Musarrat and Zaidi.[1]

Microbial Degradation of Organic Pollutants

The soil microbial diversity is regarded as a major factor responsible for many biochemical transformations, including degradation of diverse organic compounds.[5–7] The chemical structures of organic pollutants exert profound influence on the metabolic abilities of microbes. Thus, harvesting the microbiological activities for biodegradation of recalcitrant hazardous chemicals and its implementation forms the basis for the increased use of biotreatment systems (Table 1). Predominantly, noxious chemicals, such as polychlorinated biphenyls (PCBs), trichloroethylene (TCE), polycyclic aromatic hydrocarbons (PAHs), and pesticide residue in soils and water are the targets. Removal of unwanted residues, as well as pesticide efficacy, is ultimately dependent on the presence, number, and enzymatic capability

TABLE 1 Examples of Bacteria, Actinomycetes, Fungi, and Cyanobacteria Engaged in Pesticide Degradation

Organisms	Organic Compounds or Pesticides
Bacteria and actinomycetes	
Alcaligenes denitrificans	Fluoranthene (PAH)
Alcaligenes faecalis	Arylacetonitriles
Archromobacter	Carbofuran
Arthrobacter	EPTC, glyphosate, pentachlorophenol (PCP)
Bacillus sphaericus	Urea herbicides
Brevibacterium oxydans IH35A	Cyclohexylamine
Burkholderia sp. P514	1,2,4,5-TeCB
Clostridium	Quinoline, glyphosate
Comamonas testosteroni	Arylacetonitriles
Corynebacterium nitrophilous	Acetonitrile, carboxylic acid, ketones
Dehalococcoides ethenogenes 195	Trichloroethylene (TCE)
Desulfitobacterium dehalogenans	Hydroxylated PCBs
Desulfovibrio sp.	Nitroaromatic compound
Flavobacterium	PCP
Geobacter sp.	Aromatic compound
Klebsiella pneumoniae	3 and 4 Hydroxybenzoate
Methylococcus capsulatus (Bath)	Trichloroethylene
Methylosinus trichosporium OB 3b	1,1,1-Trichloroethane (TCA)
Moraxella	Quinoline, glyphosate
Nitrosomonas europaea	TCA
Nocardia	Quinoline, glyphosate
Pseudomonas aeruginosa	Nitriles, biphenyl, parathion
Pseudomonas sp.	Quinoline, glyphosate
Pseudomonas stutzeri	Parathion
Pseudomonas cepacia	2,4,5-T
Pseudomonas paucimobilis	PCP
Pseudomonas putida 6786	Propane
Pseudomonas striata	Propham, chloropham
Rhodococcus chlorophenolicus	PCP
Rhodococcus corallinus	*S*-triazines
Rhodococcus rhodochrous	Propane
Rhodococcus sp.	Propane, TCA
Rhodococcus UM1	Pyrene

(*Continued*)

TABLE 1 (*Continued*) Examples of Bacteria, Actinomycetes, Fungi, and Cyanobacteria Engaged in Pesticide Degradation

Organisms	Organic Compounds or Pesticides
	Fungi
Aspergillus flavus	DDT
Aspergillus parasiticus	DDT
Aspergillus niger	2,4-D
Candida tropicalis	Phenol
Chrysosporium lignorum	3,4-Dichloroaniline
Fusarium solani	Acylamilide
Fusarium oxysporum	DDT
Hendersonula toruloidea	2,4-D
Hydrogenomonas + *Fusarium* sp.	DDM, nitrile
Mucor alternans	DDT
Penicillium	Acylamilide
Penicillium megasporum	2,4-D
Phallinus weirii	DDT
Phanerochaete chrysosporium	PAH, 2,4,6-trinitrotoluene, PCP, DDT, 2,4,5-T, lindane
Pleurotus ostreatus	DDT
Polyporus versicolor	DDT
Pullularia	Acylamilide
Rhodotorula	Benzaldehyde
Stereum hirsutum	Phenanthrene
Trametes versicolor	Dieldrin
Trichoderma sp.	Nitrile
Trichoderma viride	DDT
Trichosporon cutaneum	Phenol
Yeast	Paraquat
	Ectomycorrhizal fungi
Tylospora fibrillosa	Mefluidide
Thelephora terrestris	
Suillus variegatus	
Suillus granulatus	
Suillus luteus	
Hymenoscyphus ericae	
Paxillus involutus	
	Cyanobacteria
Cylindrospermum sp.	BHC
Aulosira fertilissima	lindane
Plectonema boryanum	diazinon, endrin
Nostoc muscorum	Carbofuran
Wollea bhardlvajae	hexachlorocyclohexane (HCH)
Nostoc musorum	HCH
Mastigocladus laminosus	Tolkan
Tolypothrix tenuis	fluchloralin
Anabaena doliolum,	Butachlor

(*Continued*)

TABLE 1 (*Continued*) Examples of Bacteria, Actinomycetes, Fungi, and Cyanobacteria Engaged in Pesticide Degradation

Organisms	Organic Compounds or Pesticides
Nostoc muscorum	
Anabaena ARM 286	BHC
Anabaena ARM 310	Ekalux
Anabaena variabilis	Bavistin
Aulosira fertilissima	eenlate
Scytonema chiastum	captan
Scytonema stuposum	dithane
	cyathion
Anabaena khannae	Butachlor
Calothrix marchica	benthiocarb
Nostoc calcicola	pandimethalin
Tolypothrix limbata	oxadiazon
Aulosira fertilissima ARM 68	Monocrotophos
Nostoc muscorum ARM 221	malathion, dichlorovos, phosphomidon, quinolphos

Source: Data from Adhikary[9] and Gothalwal and Bisen.[10]

of soil microbes. Mineralization of these compounds takes place only when environmental condition, water activity, presence of O_2, temperature, and pH are suitable for the growth and survival of the organism (Table 2). Degradation fails when the target compound is either very concentrated or much diluted. Successful biological cleanup of soil and water contaminants takes place in the presence of aerobic microorganisms. A dichlorinating anaerobic microbial population can also be helpful in cleanup processes. Genes for complete mineralization of some of the haloaromatic compounds are also reported and effectively utilized for recruitment of microbes and environmental applications. Degradation of naphthalene is more difficult than that of bionuclear compounds such as biphenyl, dibenzofuran, and dibenzo-*p*-dioxin, and of mononuclear aromatic compounds such as aniline, benzene, salicylate, phenoxyacetate, and toluene. Peripheral or funneling, central degradative and oxoadipate pathway sequences are necessary for the complete degradation of the compounds.[8] Pentachlorophenol (PCP) and PCB are major recalcitrant compounds. The major problems of PCP degradation are the formation of toxic end products during the metabolism and substitutions of chlorine.

Organochlorine pesticides (DDT, endosulfan, hexachlorobenzene, and hexachlorocyclohexane) are degraded by microbes by means of reductive dechlorination, dehydrochlorination, oxidation, and isomerization of the parent molecule. The principal reactions involved in the breakdown of phosphotriesters (organophosphate insecticides) are hydrolysis, oxidation, alkylation, and dealkylation.[11] Microbial degradation through hydrolysis of p-O-alkyl and p-O-aryl bonds is considered to be the most significant step in the degradation of parathions, methyl parathion, and p-nitrophenol. Reductive dechlorination of organochlorine is an important microbial reaction. A classic example of this is the conversion of DDT to DDD; this reaction occurs in several species of bacteria such as *Pseudomonas, Bacillus, Arthrobacter, Clostridium;* in soil actinomycetes; in yeasts; and in fungi such as *Trichoderma viridae, Mucor alterans,* white rot fungi, *Pleurotus australis, Phellinus weirii,* and *Polyporus versicolor* (Table 3). The formation of DDE from DDT through dehydrochlorination is commonly observed in algae, diatoms, and phytoplankton (*Cylindrotrea dentorium, Cyclotella nana, Isochryeii gabana, Nitzschia* spp.). Microbial isomerization reactions involve the conversion of γ-BHC to α-BHC, diedrin to photodieldrin, and D-keto andrin to endrin by *Pseudomonas putida*.[12]

Lindane is very persistent in the environment and resistant to microbial degradation.[13] It is degraded aerobically as well as anaerobically. *Pseudomonas paucimobilis* UT26 is a unique microbe

TABLE 2 Major Factors Affecting Bioremediation

Microbial
- Growth until critical biomass is reached
- Mutation and horizontal gene transfer
- Enrichment of the capable microbial populations
- Production of toxic metabolites

Environmental
- Depletion of preferential substrates
- Lack of nutrients
- Inhibitory environmental conditions

Substrate
- Too low concentration of contaminants
- Chemical structure of contaminants
- Toxicity of contaminants
- Solubility of contaminants

Biological aerobic vs. anaerobic process
- Oxidation/reduction potential
- Availability of electron acceptors
- Microbial population present in the site

Growth substrate vs. co-metabolism
- Type of contaminants
- Concentration
- Alternate carbon source present
- Microbial interaction (competition, succession, and predation)

Physico-chemical bioavailability of pollutants
- Equilibrium sorption
- Irreversible sorption
- Incorporation into humic matter

Mass transfer limitations
- Oxygen diffusion and solubility
- Diffusion of nutrients
- Solubility/miscibility in/with water

Source: Data from Boopathy.[15]

TABLE 3 Mechanisms of Radionuclide Bioremediation

S.N.	Mechanism	Microorganisms	Remediated Radio
1.	Biosorption	*Rhizopus arrhizus, S. cerevisae, Penicillium americium, Aspergillus, Aeromonas hydrophila, Candida utilis*	Uptake
2.	Engineering biosorption	Eukaryotic metallothien, *E. coli* Lan B	
3.	Bioaccumulation	*Citrobacter* sp.	UO_2^{2+}
		Rhodococcus erythropali CS98	Cesium
		Rhodococcus sp. strain CS 402	Cesium
		Micrococcus luteus	Neptunium
		Radioresistant bacteria	
		Deinococcus radiodurans	U (Vi), Tc (Vii)
		D. geothermalis	Radioactive waste

that utilizes hexachlorocyclohexane (HCH) as its sole source of carbon and energy under aerobic conditions.[14] Five structural genes (*lin A, lin B, lin C, lin D*, and *lin E*) and one regulatory gene (*lin R*) are involved in degradation of α-HCH in UT 26. *lin A*, *lin B*, *lin C*, and *lin D* codes for dehydrochlorinase, halidohydralone, dehydrogenase, and reductive dehalogenase, respectively, and *lin E* encodes ring cleavage oxygenase. Microbial degradation of γ-HCH has also been reported in *Anabaena* sp. PCC 7120 and *Nostoc ellipsosorum*.[16] The enzyme responsible for catalyzing the hydrolysis step in parathion is organophosphate hydrolase (OPH), which is encoded by the *opd* gene. This gene has been isolated from several bacteria. A naturally occurring variant of OPH enzyme, designated as opdA, capable of degrading a broad range of organophosphates was isolated from an *Agrobacterium radiobacter* strain.[17] The slow growth and low culture yields of native OPH-producing strains make them uneconomical for practical use. Therefore, efforts have been made by several researches to improve the applicability of OPH for pesticide bioremediation. A consortium comprising two bacteria that were genetically engineered (*E. coli* and *P. putida* KT 2440) efficiently worked together to break down parathion and prevent accumulation of p-nitrophenol [PNP].[18]

Catechol is a terminal metabolite formed during the degradation pathways of various compounds, and a variety of potential degraders of catechol have been reported by Kim et al.[19] Biodegradation of phenol and toluene by *Pseudomonas* sp., *Bacillus* sp., and *Staphylococcus* sp. was studied by Prasanna et al.[20] These strains were isolated from pharmaceutical industrial effluents. Mixed cultures showed more efficient degradation than pure strains within 5–7 hr at lower concentrations. Polycyclic aromatic hydrocarbon-degrading bacteria and ligninolytic and non-ligninolytic fungi are ubiquitously distributed in the natural environment such as soils and woody materials. The principal mechanism for aerobic bacterial metabolism of PAH is the initial oxidation of the benzene ring by the action of dioxygenase enzyme to form *cis*-dihydrodiol intermediates, which can then be further metabolized via catechols to CO_2 and H_2O.[21] Many bacterial, fungal, and algal strains have been shown to degrade a wide variety of PAHs (Table 1). There are limited reports on degradation of high molecular weight PAHs with more than four benzene rings. In general, high molecular weight PAHs are degraded slowly by indigenous or augmented microorganisms, as the persistence of PAHs increases with their molecular size.

The recalcitrant nature of chloroaromatic haloalkanes is due to the low electron density at the aromatic ring which makes the enzyme oxygenase unable to attack this compound. Many soil microorganisms (*Pseudomonas* and *Alcaligenes)* which synthesize the halogenase can utilize Halogenated alkonic acids [HAA].[22] Anaerobic methane-oxidizing bacteria can degrade TCE in pure *Pseudomonas* culture through a co-metabolic process.[23] The PCP-degrading *Pseudomonas* sp. strain IST 103 has been isolated, which was found to be capable of utilizing PCP as a carbon source. The enzyme PCP-4 monooxygenase was found to be responsible for the dechlorination of PCP.[24] The gene for the degradation may be plasmid encoded or present on the chromo-some.[25,26] Attempts have been made to enhance PCB biodegradation by modifying oxygenase.[27] One of the most efficient methods of biodegradation consists of sequential anaerobic and aerobic treatments for highly chlorinated compounds. Biochemical and genetic engineering approaches for dehalogenase and oxygenase could lead to “super bugs” that could be used for the bioremediation of chlorinated compounds.[28] Modified degradative genes could be introduced into the original strain and/or major indigenous strains isolated from contaminated sites, and it is hoped that these super bugs could have application in bioremediation in the near future, confirming their usefulness and safety. Raji et al.[29] have isolated a bacterial culture able to grow on benzoate and useful for remediation of PCB-contaminated sites. *Arthrobacter* sp. IFL YN 10 demonstrated mineralization of C^{14} ringlabeled atrazine. This isolates can be used to develop a consortium for bioremediation of pesticides.[30]

To improve the biodegradation efficiency and implementation, integrating various components such a microbial strain in consortium, solid O_2 source, and appropriate role of nutrients with controlled release pattern into a granule formulation with an oleophilic matrix, may provide an ideal approach to improve remediation of crude oil pollutants.[31] Abed et al.[32] reported that salinity and temperature are important environmental parameters that influence the degradation process of petroleum compounds. The inhibitory effect of salinity was shown to be more pronounced for aromatic than for aliphatic compounds.[33] Higher temperatures also reduce the viscosity of crude oil, which increases its diffusion through sediments, a process that render oil components accessible to bacteria. The possibility of the involvement of catabolic plasmid in the degradation of anthracene by *Pseudomonas* sp. isolated from an oil filling station was investigated by Kumar et al.[34] Many γ and β proteobacterial groups (*Halophaga, Geo-thrix*, *Acidobacterium*) and green non-sulfur bacteria with a strong potential to degrade hydrocarbons were present in benthic cyanobacterial mats.[35] The aliphatic fraction of petroleum hydrocarbon[36] is degraded by *Arthrobacter, Alcaligenes, Flavobacterium,* and *Bacillus*. Kniemeyer et al.[37] came across a green *Methanospirillium* that is able to degrade aromatic hydrocarbons. Singh and Lin[38] isolated 10 indigenous microorganisms from oil-containing soil; five isolates achieved 86.94% diesel degradation in 2 weeks. The results strongly indicate that the environmental condition of the contaminated site plays a crucial role in the degradation. Cohen[39] reported the development of cyanobacterial mats in oil-contaminated courts. Cyanobacterial polysaccharides play a major role in the emulsification of oil, actually breaking the oil into small droplets, which are subsequently attacked by the heterotrophs. Bioremediations of high fat and oil wastewater by selected lipase-producing bacteria such as *Bacillus subtilis*, *B. lichenformis*, *B. amyloliquifaciens, Serratia marcescens, P. aeruginosa,* and *Staphylococcus aureus* were carried out in wastewater from palm oil mill, dairy, slaughterhouse, soap industry, and domestic wastewater. After 12 days of bioremediation, the least biological oxygen demand and lipid content was observed in consortia, and the lipid degradation capacity of *P. aeruginosa* was higher than that of other bacteria.[40] Verma et al.[41] proved the biotechnological importance and advantage of using *P. aeruginosa* SL72 and *Acinetobacter* sp. SL-3 individually or as a consortium for waste treatment, resulting in substantial removal of the crude oil within a week, using a low-cost, efficient, and environment-friendly technique.

Spent wash is dark brown due to the recalcitrant melanoidin pigment. *Pseudomonas* sp. was selected for degradation of the pigment by Chavan et al.[42] Chuphal and Thakur[43] have characterized an alkalophilic bacterial consortium (*Micrococcus luteus*, *Deinococcus radiothilus*, *Micrococcus diversue, P. syringae P. myricure)* for *ex situ* bioremediation of color and adsorbable organic halogens in pulp and paper mill effluent. Nanda et al.[44] employed *Nostoc* sp. for bioremediation of tannery effluents; the main economic advantage of this system is the lack of a serious sludge disposal problem, consequently resulting in a much cheaper operating cost. No microorganism has been found to degrade polythene without an additive such as starch. The discovery of new enzymes and the cloning of genes for synthetic polymer-degrading enzyme from *Pseudomonas* sp. were reviewed by Premraj and Doble.[45]

Chemotaxis has been postulated to play an important role in enhancing biodegradation as it increases the bioavailability of pollutants to bacteria. Some toxic organic compounds are chemoattractants for different bacterial species, which can lead to improved biodegradation of these compounds. A *Ralstonia* sp. was chemotactic toward different Nitro aeromatic compounds (NACs), i.e., p-nitrophenol (PNP), p-nitrobenzoate (PNB), and o-nitrobenzene (ONB).[46]

Mycoremediation

The key to mycoremediation is determining the right fungal species to target specific pollutants. Certain strains have been reported to successfully degrade the nerve gases VX and sarin. Battelle in a plot of soil contaminated with diesel oil was inoculated with mycelia of oyster mushrooms; within weeks, more than 95% of PAH had been reduced to non-toxic components. Mycofiltration is a similar process using fungal mycelia to filter toxic waste and microorganisms from waste in soil. Breakdown (70%–100%) of anthracene oil found in PAH was reported in 27 days in an N_2-limited culture of *Phaenerochaete chrysosporium*. Pentachlorophenol is an important constituent of paper mill effluents, and *Phaenochaste chrysorperiucm* immobilized on rotating biological contactor disk efficiently degrades 2,4-dichlorophenol, 2,4,6-trichlorophenol (TCP), polychlorinated quiacole, and several chlorinated vanillins.[47]

Bioremediation of Inorganic Contaminants

Microbes encounter metals such as Cr, Mn, Fe, Co, Ni, Cu, Zn, Ni, Ag, Cd, Pb, and Au, and metalloids such as As, Se, and Sb, having a diverse nature, in the environment. Microbes can detoxify metals by valence transformation, extracellular chemical precipitation, or volatilization. Such microbes combat high concentrations of heavy metals by the following processes:

1. Inactivation of metals;
2. Alteration of the site of inhibition;
3. Enhancement in impermeability of metals; and
4. Other by-pass mechanisms.[48]

Bacterial biomass can also bioaccumulate heavy metals both in live and dead states through intracellular accumulation and extracellular absorption, respectively, giving an effective alternative for small-scale remediation purposes.[49] Bacteria can remediate heavy metals by a variety of mechanisms, including bioaccumulation, biosorption, and bioremediation (Figure 2). Bioaccumulation is the retention and concentration of a substance by an organism through the cell membrane into the cytoplasm, where the metal is separated and immobilized. However, in biosorption, the negatively charged metal ions are separated through adsorption to the negative ionic groups on the cell surface (carboxyl residue, phosphate residue, SH groups, or hydroxyl group) such as capsule or slim layers. The charged functional groups[50] serve as nucleation sites for the deposition of various metalbearing precipitatesi[51] *Bacillus* SJ-101 exhibits a much higher capacity of intracellular Ni accumulation, which is attributed to the anionic nature of its cell surface. In bioremediation, biologically catalyzed redox reactions lead to immobilization of metals. Microorganisms are known to reduce a wide variety of multivalent metals that pose environmental problems.[52,53] The reduced species are highly insoluble and precipitate out from solution.

Bioremediation is one such promising option that harnesses the impressive capabilities of microbes associated with roots to degrade organic pollutants and transform toxic metals. Since it is a plant-based *in situ* phytorestoration technique, it is proven to be economicaly efficient and easy to implement under field conditions.[54] All plant growth-promoting rhizobacterial strains (*Azotobacter chroococcum, Bacillus megatorium, B. mucilaginosus, B. subtilis* SJ-101, *Pseudomonas* sp., *P. fluorescens, Rhizobium leguminosarum, Kluyvera ascorbata* SUD165)[55] can be used for bioremediation of metals. Rhizobacteria associated with hyperaccumulators (*B. subtilis*, *B. pumilus*, *P. pseudoalcaligens*, *and Brevibacterium halotolerans*) are also widely used in bio and rhizoremediation of multimetal-contaminated sites.[56] The rhizobacteria are used or manipulated with three main objectives for bioremediation of metal-contaminated soils: 1) hyperaccumulation of metals in plants; 2) reduction of the uptake of metals; and 3) *in situ* stabilization of

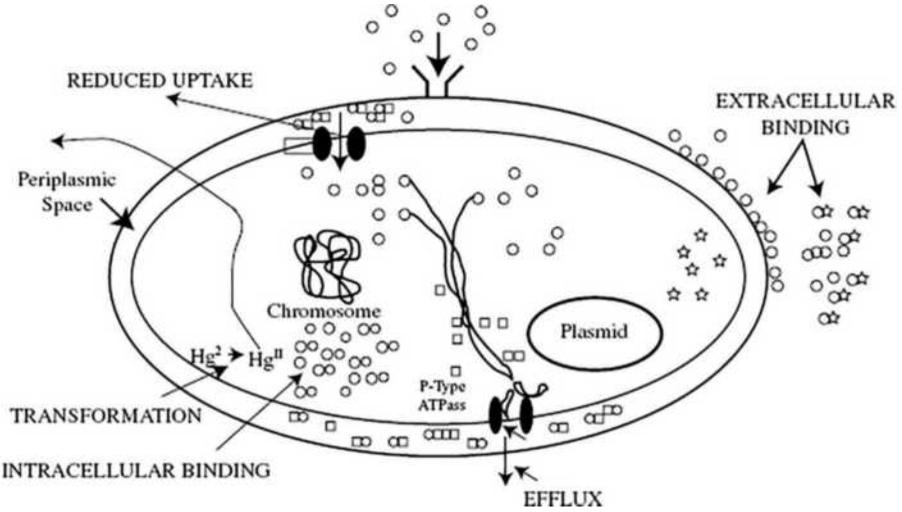


FIGURE 2 Different metal resistance mechanisms in bacteria.
Source: Singh and Srivastava.[66]

the metals as organocomplexes (Figure 2). The chemical conditions of the rhizosphere differ from bulk soil as a consequence of various processes induced by plant roots as well as by rhizobacteria,[57] such as secretion of organic acids followed by reduction in pH and production of siderophores, phytochelains, amino acid, and ACC deaminase. *Pseudomonas maltophilia* was shown to reduce the mobile and toxic Cr (VI) to non-toxic and immobile Cr (III) and also minimize the environmental mobility of other toxic ions (Hg, Pb, Cd).[58] Al Ayely et al[59] studied the effect of increasing levels of As and P on fern injected with mycorrhiza. The greater diversity of plant species may be responsible in part for the greater bacterial diversity in the bulk soils.

Redox Reaction Leading to Immobilization

The direct enzymatic reduction of soluble uranium (Vi), Cr (Vi), and Tc (Vii) to insoluble species is well documented.[60,61] A number of Cr (Vi)-reducing microbial strains, including *Oscillatoria* sp., *Arthrobacter* sp., *Agrobacter* sp., *Pseudomonas ambigua*, *Chlamydomonas* sp., *Chlorella vulgaris, Zoogloea ramigera, P. aeruginosa*,[62] and anaerobic sulfate-reducing bacteria have been isolated from chromate contaminated soil water and sediment.[63]

Redox Reaction Leading to Solubilization

Solubilization of adsorbed and co-precipitated metals may occur by direct or indirect microbial processes. The solubilization of toxic heavy metals and radionuclide from coprecipitates requires at least partial solubilization of oxide minerals.[64] Presumably, the organic acids formed by the metabolic activity of microbes lowers the pH of the system so that it interferes with the electrostatic forces that hold heavy metals and radionuclides on the surface of iron or Mn oxide minerals. Extracellular polymeric substances serve as biosorbing agents by accumulating nutrients from the surrounding environment, and also play a crucial role in biosorption of heavy metals. Being polyanionic in nature, exopolysaccharide forms complexes with metal cations, resulting in metal immobilization with the exopolymeric matrix.[65] Biofilm formation is a strategy that microorganisms might use to survive a toxic flux in these inorganic compounds; biofilm populations are protected from toxic metals by the combined action of chemical, physical, and physiological phenomena, which are in some instances linked to phenotypic variations among the constituent biofilm cells. Harrison et al.[67] have prepared a multifunctional model by which a biofilm population can withstand metal toxicity by the process of cellular diversification.

An effective ecofriendly approach for the removal of Mn from e-waste by the fungus *Helminthosporium solani* was shown by Savitha et al.[68] through the process of nonmetabolism-dependent biosorption under different environmental conditions of metal concentration, pH, and dry biomass concentration. “Metallothionein” has become a generic term applied to low molecular weight proteins or polypeptides that bind metal ions in metal thiolate clusters, and whose synthesis increases in response to elevated concentrations of certain metals.[69] The maximum cobalt removal efficacy (1 μg of ^{60}Co/g, dry wt) of bacterial mass (*B. megaterium, P. putida, Flavobacteriuam devorans, Salmonella typhimurium, Streptomyces gresius, Rhizopus* sp., *Rhodococcus* sp., *E. coli*) could be achieved within 6 hr[70] compared with the 8–500 ng/g attained after 24–48 hr. The gum kondagogu, a natural carbohydrate polymer, was investigated[71] for its adsorptive removal of the toxic metal ions Cd^{2+}, Cu^{2+}, Fe^{2+}, Pb^{2+}, Ni^{2+}, Zn^{2+}, and Hg^{2+} present in industrial effluents. Kondogogu has a potential to be used as an effective, non-toxic, economical, and efficient biosorbent cleanup matrix for the removal of toxic ions with re-adsorption capacity at 90% level even after three cycles of desorption. *Klebsiella oxytoca* was able to biodegrade cyanide to a non-toxic end product (ammonia) using cyanide as the sole N2 source, which might preceed using ammonia as an assimilatory substrate.[72]

Microorganisms interact with radionuclides via several mechanisms (Table 3), some of which are used as the basis of potential bioremediation strategies. Based on the different types of effluents generated, a single technology cannot be suitable to address the problems. Cyanobacteria have the capacity to utilize nitrogenous compounds as well as phosphates; in addition, they pick up metal ions such as Cr, Co, Cu, and Zn very effectively. It has been observed that immobilized cyanobacteria have greater potential than their free cells. Various natural polymers such as alginate agar and carrageenan, and synthetic polymers such polyacrylamide and polyurethane have already been tried.[73] Novel methods of immobilization, including coimmobilizations of various species, are required for the symbiotic interaction among themselves, which will result in synergistic enhancement or removal capabilities. A gene cluster composed of nine open reading frames involved in Ni^{2+}, Co^{2+}, and Zn^{2+} sensing and tolerance in *Synechocystis* sp. PCC 6803 has been identified by Mario Garcia et al.[74] The biosorption of Hg^{2+} by *Spirulina platensis* and *Aphanothece flocculosa* was studied under a batch stirred reaction system,[75] and more than 90% from *A. flocculosa* and 100% mercury recovery from *S. platen-sis* can be achieved for 4 and 1 cycle, respectively. Protein and total non-protein thiols were measured as stressresponsive metabolites in response to Ni in *Anabaena doliolum*.[76] As *A. doliolum* is a high-biomass-producing strain, it can be conveniently separated from the solution by filtration and can be used at pilot-scale removal of Ni from wastewater.

Phytoremediation: A Beneficial Alternative

Phytoremediation is a promising green technology for accelerated decontamination of soil and water. It is a natural process in which plants are used to remove pollutants (pesticides, solvents, explosives, crude oil, PAHs) from soil and water. Plants can degrade or transform both the organic and metal (landfill leachates and radionuclide) contaminants by acting as filters or traps. Plants and associated microbes can degrade the pollutants or at least limit their spread in the environment. There are several ways in which plants can be used for the phytoremediation of organic contaminants viz. phytoextraction, phytodegradation, rhizodegradation, and phytovolatization (Figure 3). They can be compared to solar driven pumps for extraction and concentration of elements from the environment.[4] The plants can also be used for "phytomining." Rugh et al.[77] inserted an altered mercuric ion reductive gene (*mer A*)

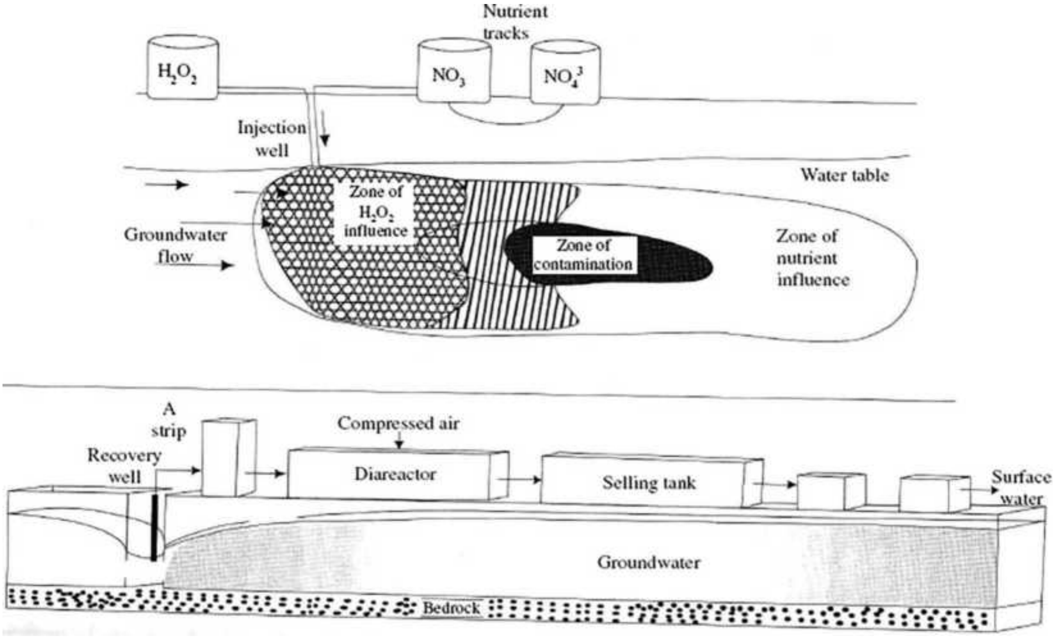


FIGURE 3 ***In situ*** bioremediation processes of bioaugmentation and biostimulation.
Source: Thakur.[81]

into *Arabidopsis thaliana,* creating a transgenic plant that volatilizes mercury to the atmosphere. Phytoremediation depends on a variety of factors, including physical conditions, chemical properties of the contaminant, and relative tolerance of the plant to the contaminant. Morphological characteristics of plant uptake are controlled by soil factors such as clay content, organic matter, soil moisture, and pH. The plant root can alter the pH in the rhizosphere by secreting organic acids, thus effecting the bioavailability of certain compounds, i.e., ionic pollutants whose solubility and desorption from soil colloids are pH dependent. Certain plants also produce biosurfactants that may increase the solubility of more lipophilic compounds.

Rugh[78] reported laboratory model plants such as *Thale cress* and tobacco to enhance phytoremediation of organomercurials, TCE, and nitroaromatic explosives that have been engineered with a non-plant transgene. Hyperaccumulator plants are also grown along with non-hyperaccumulators to enhance the heavy metal uptake by intermingling the roots and induce the colonization of efficient rhizobacteria. Recombinant *Mesorhizobium huakii*, by incorporating the phytochelatinase gene from *Arabidopsis thaliana* into *M. huakii* subsp. *rengei* 133, increased Cd accumulation by 1.5-fold in *Astragalus sinicus.*[79] The phytoremediation potential of water hyacinth (*Eichhornia crassipes*) in the treatment of tannery effluents was evaluated by Athaullah et al.[80] who found a promising risk reduction, cost-effective technology for water sanitation and conservation.

Bioremediation Technology

Bioremediation technology using microorganisms was invented by George M. Robinson in 1960. Bioremediation technologies can be generally classified as *in situ* or (microbial ecological approach) or *ex situ* (microbial approach) (Figure 3); some examples of these technologies include bioventing, land farming, bioreactor, composting, bioaugmentation, rhizofiltration, biostimulation, and solid phase bioremediation. In the development of the technology, the following points should be considered:

1. Heterogeneity of the contaminant
2. Concentration of the contaminant and its effect on the biodegradative microbe
3. Persistence and toxicity of the contaminant
4. Behaviour of the contaminant in soil
5. Conditions favorable for the biodegradative microbe or microbial population

Bioremediation is one of the most promising technologies for treating military sites, industrial wastes, municipal/urban wastes, mining waste, chemical spills and hazardous wastes, etc. *In situ* bioremediation can be implemented in many treatment modes, including aerobic, anoxic, anaerobic, and co-metabolic. The aerobic mode has proved to be most effective in reducing contaminant levels of aliphatic and aromatic compounds. Biofiltration is best suited for airstream containing volatile organic compounds.

Pollution remediation of tannery effluents is very complex. A multiprong treatment is thus required for a combination of nanotechnology and microbial technology, with a prior proceeding such as cycloning, flotation, microflotation, or electroflotation.

Bioremediation can be done on site, is often cheap, with minimal site disruption, eliminates waste permanently, eliminates long-term liability, has greater public acceptance, and can be coupled with other physical or chemical methods.[15] Table 2 shows the major factors affecting bioremediation. Another technique employed for bioremediation involves bioaugmentation, i.e., pumping genetically engineered microorganisms or microorganisms with enhanced degradation ability into the subsurface. Timely bioremediation of petroleum-contaminated soils is possible with innovative engineering and environmental manipulation to enhance microbial activity beyond the natural effective season.[82]

The first field slurry process used to remediate the nitroaromatic herbicide dinoterb used SABRE (sequential anaerobic bioremediation process), which was renamed as the FAST (fermentative anaerobic soil treatment) process.[83] Hence, the optimized remediation strategies exploring microbial diversity are being executed *in situ* successfully.

The aim is to translate research findings from the laboratory into viable technologies for remediation in the field. In the injection method, bacteria and nutrients are injected directly into the contaminated aquifer, or nutrient or enzymes are often referred fertilizer, that stimulate the activity of the bacteria are added. Bioreactors using immobilized cells have several advantages over conventional effluent-treatment technologies. Degradation of 4-chlorophenol by anaerobes attached to granular activated carbon in a biofilm reactor was evaluated during both open and closed modes of operation. Continuous flow fluidized bed reactor and bench scale continuous flow activated sludge reactor were used to study the removal of TCP, TeCP and PCP[84] which are used as sole source of carbon and energy. The ability of *Arthrobacter* cells to degrade PCP was evaluated for immobilized, non-immobilized, and co-immobilized cells. Fixed film bioreactor has been used with mixed bacterial communities for the treatment of pulp and paper mill effluent. The effluent after treatment showed a removal of color (80%), Chemical oxygen demand (71%), and chlorinated organic compound (68%).[85] The fungal slurry was successfully applied for compost preparation and biomediation of the Cr-contaminated tannery soil. The treated effluent was used for seed germination of crops without any phytotropic influence.[86]

Biocolloid formation methods have been adapted for bioremediation of metals by bacteria and fungi without the need *ex situ* treatment. Electrokinetically enhanced *in situ* soil decontamination and dispersing by chemical reaction together with microbes. Chlorophenol-contaminated saw mill soil used composting without bioaugmentation in a cheap and feasible method.[89] The use of biphenyl as an *in situ* co-substrate is expensive and environmentally problematic,[90] thus there was a need to investigate the ability of alternative cosubstrates to support the co-metabolic degradation to PCBs.

Biocapsules have been tested for various applications and can be produced for site-specific applications (Bioremediation Applied Bioscience http://www.bioprocess.com). *In situ* groundwater biodegradation in the United States has been carried out through various processes by numerous companies, as shown in Table 4. The removal effectiveness can reach 100% (Figure 4).

TABLE 4 In Situ Ground Water Biodegradation Examples of Process and Companies

S.N.	Name of Company	Technology Used	Types of Environmental Pollutant Remediated
1.	Bio-genesis Technology	Custom-blended microbial culture (GT-1000 series)	Oil, BTEX, diesel fuel
2.	Biopim	Biological sand filter	BTEX, TPH, phenol monochlorobenzene, metals
3.	ENSR Consulting and Engineering Technology (remediation beneath building foundation)	Steam injection, soil vapor extraction, ground water extraction, and air stripping	TCA, TCE, DWAPLS
4.	Petro Clean Bioremediation System	Indigenous microorganisms	Gasoline, diesel fuel, aviation fuel, solvents, PNAs, VOCs, and other organic compounds
5.	Kemron Environmental Services	Engineered, site-specific, groundwater recirculation system	Petroleum products solvents, halogenated volatiles and semivolatiles, BTEX, polynuclear aromatics, and organic acids
6.	Remediation Technologies Inc.	Treating groundwater in saturated zone injection, recovery well, monitoring wells	Dissolved contaminants
7.	SBP Technologies	Encapsulated cell inserted into a well	PAH (high mol. wt.), chlorinated aromatics (PCP), and pesticides
8.	OHM Remediation Services Corporation	Aquifer	Petroleum hydrocarbons, BTEX, chlorinated and non-chlorinated solvents

(*Continued*)

TABLE 4 (*Continued*) In Situ Ground Water Biodegradation Examples of Process and Companies

S.N.	Name of Company	Technology Used	Types of Environmental Pollutant Remediated
9.	Electrokinetics Inc. (no need to add microorganism)	Electro osmosis or electro chemical migration	TCE, BTEX, PAH
10.	Geo Microbial Technologies Inc.	Removal of H_2S anaerobic process	—
11.	EODT Services Inc.	Use of biodispersant (also same as high-energy nutrients for microorganisms)	—
12.	Ecology Technologies International Inc.	Use of FyreZyme (multifactorial liquid agents)	Organic contaminants
13.	Gaia Resource Inc.	—	Hazardous and radioactive waste
14.	IT Corporation	Pump and treat system	Industrial effluents (hazardous organic compounds
15.	Ground Water Tech Inc.	Destructive technology	Hazardous compounds
16.	Yellow Stone Environmental Science Inc.	Pump and treat, aerobic processes, denitrification, sulfate reduction	Aromatic hydrocarbon, halogenated hydrocarbon, VOC, BTEX, phenol, cresol, CCl_4, PCE, vinyl chloride
17.	Waste Stream Tech. Inc.	Bioaugmentation	Organic compounds
18.	Micro Bac International Inc.	Batch and continuous feed treatment using M-1000 microbial consortium	Host specific
19.	Kuzanci Environmental Techniques (Figure 4)[87]	Microlife DCB series	Petroleum derivatives, stops foul odor

Source: Data from Bioremediation 1999.[88]

FIGURE 4 Treatment of soil hydrocarbon-polluted soils by Microlife DCB series bioremediation products. The progress is clearly seen in the photos.

Source: http://www.microbial-products.com/microbial-bacterial products bioremedation.asp/geli (accessed December 2011).[87]

Molecular Probes in Bioremediation

The traditional method of bacterial enumeration is often insufficient for monitoring the specific microbe's biochemical reaction in mixed microbial communities. It has became apparent that a significant number of microbes in these systems are viable but non-culturable. The catabolic enzymes, genes, and proteins expressed in the microbes can be exploited for the detection of the fate and effect of microbes in the bioremediation process.[91] Antibody- and fatty acid-based probes, nucleic acid sequences, and DNA probes can be used to detect genes in the bacterial genome or on plasmid, or to detect mRNA or tRNA. Other relevant techniques employed are PCR, repetitive sequence-based PCR, 16S rDNA, random amplified polymorphic DNA, and fluorescence *in situ* hybridization.[81] In a previous study, a specific synthetic PCR-amplifiable DNA fragment was introduced into a *Pseudomonas* chromosome to allow genetically engineered microbes to be identified easily.

The development of a new field of metabolic engineering involves the improvement of cellular activities by manipulation of enzymatic, transport, and regulatory function of the cell by using recombinant DNA technology. Advances in the field of genetic engineering, sequencing of the whole genome of several organisms, and developments in bioinformatics have speed up the process of gene cloning and transformation. Furthermore, many powerful analytical techniques have been developed for metabolic pathway analysis and analysis of cellular functions, such as gas chromatography (GC), gas chromatography-mass spectrometry (GC-MS), nuclear magnetic resonance, 2D gel electrophoresis, matrix-assisted laser desorption/ionization time of flight (MALDI-TOF), liquid chromatography-mass spectrometry (LC-MS), and DNA chips. Metabolic engineering is, therefore, an effort to improve the ability of microorganisms. Bioremediation require the integration of huge amount of data from different sources. Pazos et al.[92] developed "Meta Router," a system for maintaining heterogeneous information related to bioremediation in a framework that allows its query, administration, and mining. The system can be accessed and administered through a web interface for studying and representing the global properties of the bioremediation network. Bioinformatics require the study of microbial genomics, proteomics, systems biology, computational biology, phylogenetic trees, and data mining, and the application of major bioinformatics tools for determining the structure and biodegradation pathway of xenobiotic compounds. Bioinformatics has taken on a new glittering by entering the field of bioremediation.[93] The limitations of bioremediation has paved the way for the development of Genetically engineered microorganisms (GEMs), or designer biocatalysts harboring artificially designed catabolic pathways.[94] Database such as the University of Minnesota Biocatalysts/Biodegradation database provide a scope for *in silico* designing of biocatalysts for in vivo construction followed by *in situ* application. In the era of functional genomics, it is easy to construct GEMs by reshuffling the gene(s), promoter, etc., to enhance their performance in situ.

Conclusion

The popularity of bioremediation is further enhanced because it is perceived as being more "green" than other remediation technologies. As a result, bioremediation companies have a viable future regardless of the long-term effectiveness of the process. Special emphasis is required on the exploitation of biotechnological innovations to improve presently available biocatalysts, and for the evaluation of future effects of microorganisms and their proper application in the optimization of *in situ* bioremediation. The use of enzymes for degradation of pesticides can be developed as a technology for bioremediation. A super strain can be created to achieve the required result in a short time frame. One important characteristic of this technology is that it is carried out in a non-sterile open environment, which contains a host of microbes. Therefore, a strategy should be tailored in such a manner that due consideration be given to the various environmental constraints (type and amount of pollutant, climatic condition, hydrogeodynamics) that affect a particular location. Feasibility studies are essential and can have an enormous impact on the cost of full-scale remediation. Rhizoremediation can contribute to the restoration of polluted sites. Phytoremediation will require an integration of activities by plant scientists, microbiologists, chemists, and engineers, so that these systems that can be used to prevent and remediate pollution can become a reality. Environmental friendly processed need to be developed to clean up the environment without creating harmful waste products. For the development of economically usable technologies, scientists and technologists would have to offer creative solutions for either introducing new capabilities or enhancing current efficiencies.

References

1. Musarrat, J.; Zaidi, S. Bioremediation of agrochemicals and heavy metals in soil. In *Biotechnological Applications of Microorganisms. A Techno-Commercial Approach;* Mahesh-wari, D.K., Dubey, R.C., Eds.; I.K. International Publishing House, Pvt. Ltd.: New Delhi, India, 2006; 311–331.
2. Tate, R.L. *Soil Microbiology*; John Wiley & Sons: New York, 2000; 464–494.
3. Salt, D.E.; Smith, R.D.; Raskin, I. Phytoremediation. Annu. Rev. Plant Physiol. Plant Mol. Biol. **1998**, *49,* 643–668.
4. Alkorta, I.; Garbisu, C. Biores. Technol. **2001**, *77,* 229–236.
5. Spain, J.C. *Biodegradation of Nitro Aromatics Compounds*; Plenum Press: New York, 1995.
6. Stoner, D.L. *Biotechnology for the Treatment of Hazardous Waste*; Lewis Publisher: Boca Raton, FL, 1994.
7. Unterman, R. A history of PCB biodegradation. In *Bioremediation Principle and Application;* Crawford, R.L., Crawford, D.L., Eds.; University Press, Cambridge, U.K., 1996; 209–253.
8. Timmis, K.N.; Steffen, R.J.; Unterman, R. Degrading microorganisms for the treatment of toxic waste. Annu. Rev. Microbiol. **1994**, *48,* 527–557.
9. Adhikary, S.P. General introduction. In *Blue Green Algae Survival Strategies in Extreme Environment*; Printer Publication: Jaipur, India, 2006; 7–17.
10. Gothalwal, R.; Bisen, P.S. Bioremediation. In *Encyclopedia of Pest Management;* Pimentel, D., Ed.; Marcel Dekker Pub. Ltd.: New York, 2002; 89–93.
11. Singh, N.; Asthana, R.K.; Kayastha, A.M.; Pandey, S.; Chaudhary, A.K.; Singh, S.P. Thiol: An exopolysaccharide production in a cyano bacterium under heavy metal stress. Process. Biochem. **1999**, 35**,** 63–68.
12. Benezet, H.J.; Matsumura, F. Isomerization of γ-BHC to α-BHC in the environment. Nature, **1973**, *243,* 480–481.
13. Agnihotri, N.P.; Gajbhiye, V.T.; Kumar. M.; Mahapatra, S.P. Environ. Manage. Assess. **1994**, *30,* 105–112.
14. Imai, R.; Nagata, Y.; Fukuda, M.; Takagi, M.; Yano, K. Molecular cloning of *Pseudomonas paucimobilis* gene encoding a 17-kilodalton polypeptide that eliminates HCL molecules from γ-hexachlorodihexane. J. Bacteriol. **1989**, *173,* 6811–6819.
15. Boopathy, R. Factor limiting bioremediation technologies. Bioresour. Technol. **2000**, *74,* 63–67.
16. Kuritz, T.; Wolk, P. Use of filamentous cyanobacteria for biodegradation of organic pollutants. Appl. Environ. Microbiol. **1995,** *61,* 234–238.
17. Horne, I.; Sutherland, T.D.; Harcourt, R.L.; Rusell, R.J.; Oakes Hott, J.G. Identification of an *opd* (organophosphate degradation) gene in an *Agrobacterium* isolate. Appl. Environ. Microbiol. **2002**, *68,* 3371–3376.
18. Gilbert, E.S.; Walker, A.W.; Keasling, J.D. A constructed microbial consortium for biodegradation of the organophosphorus insecticide parathion. Appl. Microbiol. Biotechnol. **2003,** *61,* 77–81.
19. Kim, K.P.; Lee, J.S.; Park, S.I.; Rhee, M.S.; Kim, C.K. Isolation and identification of *Klebsiella oxytoca* C302 and its degradation of aromatic hydrocarbons. Korean J Microbiol. **2000**, *36*, 58–63.
20. Prasanna, N.; Sarvanan, N.; Geetha, P.; Shanmugaprakash, M.; Rajasekardan, P. Biodegradation of phenol and toluene by *Bacillus* sp., *Pseudomonas* sp. and *Staphylococcus* sp. isolated from pharmaceutical industrial effluent. Adv. Biotech. **2008**, *7,* 20–24.

21. Mueller, J.G.; Cerniglia, C.E.; Pritchard, P.H. Bioremediation of environments contaminated by polycyclic aromatic hydrocarbons. In *Bioremediation: Principle and Applications;* Crawford, R.L., Crawford, D.L., Eds.; Cambridge Univ. Press: Idaho, 1996; 125–194.
22. Hardman, D.J.; Gowland, P.C.; Slater, J.H. Large plasmids from soil bacteria enriched on halogenated alkonic acids. Appl. Environ. Microbiol. **1986**, *51,* 44–51.
23. Little, C.D.; Palumbo, A.U.; Herbes, S.E.; Lidstrom, M.E.; Tyndall, R.L.; Gilmer, P.J. Trichloroethylene biodegradation by a methane-oxidizing bacterium. Appl. Environ. Microbiol. **1988**, *54,* 951–956.
24. Thakur, I.S.; Verma, P.K.; Upadhyaya. K.C. Molecular cloning and characterization of pentachlorophenol-degrading monooxygenase genes of *Pseudomonas* sp. from the chemostat. Biochem. Biophys. Res. Commun. **2002**, *290,* 770–774.
25. Khan, A.; Tewari, R.; Walia, A. Molecular cloning of 3- phenyl catechol dioxygenase involved in the catabolic pathway of chlorinated biphenyl from *Pseudomonas putida* and its expression in ***E. coli.*** Appl. Environ. Microbiol. **1988**, *54,* 2664–2671.
26. Pritchard, P.H. Fate of pollutants. J. Water Pollut. Control Fed. **1986**, *58,* 635–645.
27. Furukawa, K. Engineering dioxygenase for efficient degradation of environmental pollutants. Curr. Opin. Biotechnol. **2000**, *11*, 244–249.
28. Furakawa, K. 'Superbugs' for bioremediation. Trends Biotechnol. **2003**, *21,* 187–190.
29. Raji, S.; Mitra, S.; Sumathi, S. Dechlorination of chlorobenzoates by an isolated bacterial culture. Curr. Sci. **2007**, *93,* 1126–1129.
30. Sagarkar, S.; Nouriainen, A.; Bijorklot, K.; Purohit, H.J.; Jargensen, R.S.; Kapley, A. Bioremediation of atrazine in agricultural soil. In *52nd Annual Conference of AMI.* International Conference of Microbial Biotechnology for Sustainable Development, Nov 3–6, 2011, Chandigarh, 140.
31. Wang, Q.; Zhang, S.; Li, Y.; Klassen, W. Potential approaches to improving biodegradation of hydrocarbon for bioremediation of crude oil pollution. J. Environ. Protect. **2011**, *2*, 47–55.
32. Abed, R.M.M.; Thukain, A.A.; de Beer, D. Bacterial diversity of a cyanobacterial mat degrading petroleum compounds at elevated salinities and temperatures. FEMS Microbiol. Ecol. **2006**, *57,* 290–301.
33. Milli, G.; Almallah, M.; Bianchi, M.; Wambeke, F.V.; Bertrand, J.C. Effect of salinity on petroleum biodegradation. Fresenius J. Anal. Chem. **1991**, *339,* 788–791.
34. Kumar, G.; Singla, R.; Kumar, R. Plasmid associated anthracene degradation by *Pseudomonas* sp. isolated from filling station site. Nat. Sci. **2010**, *8*, 89–94.
35. Margesin, R.; Labbe, D.; Schinner, F.; Green, C.W.; Whyte, L.G. Characterization of hydrocarbon degrading microbial population in contaminated and pristine alpine soils. Appl. Environ. Microbiol. **2003**, *69,* 3985–3092.
36. Mishra, S.; Jyot, J.; Kuhad, R.C.; Lal, B. Evaluation of inoculum addition to stimulate in situ bioremediation of oily sludge contaminated soil. Appl. Environ. Microbiol. **2001**, *67,* 1675–1681.
37. Kniemeyer, O.; Fischer, T.; Wilkes, H.; Glockner, F.O.; Widdel, F. Anaerobic degradation of ethylbenzene by a new type of marine sulfate—Reducing bacterium. Appl. Environ. Microbiol. **2003**, *69*, 760–768.
38. Singh, C.; Lin, J. Isolation and characterization of diesel oil degrading indigenous microorganisms in Kwazulu-Natal, South Africa. Afr. J. Biotechnol. **2008**, *7,* 1927–1932.
39. Cohen, Y. Bioremediation of oil by marine microbial mats. Int. Microbiol. **2002**, *5*, 189–193.
40. Prasad, M.P.; Manjunath, K. Comparative study on biodegradation of lipid rich waste water using lipase producing bacterial species. Indian J. Biotechnol. **2011**, *10,* 121–124.
41. Verma, S.; Lata; Saxena, J.; Sharma, V. Bioremediation of crude oil contaminated waste using mixed consortium of biosurfactant and lipase producing strains**.** In *52 Annual Conference of AMI*; Int. Conf. on Microbial Biotechnology for Sustainable Development, Nov 3–6, 2011, Chandigarh, 48.
42. Chavan, M.N.; Kulkarni, M.V.; Zope, V.P.; Mahulikar, P.P. Microbial degradation of melanoidine in distillery spent wash by an indigenous isolate. Indian J. Biotechnol. **2006**, *5*, 416–421.

43. Chuphal, Y.; Thakur, I.S. Characterization of alkalophilic bacterial consortium for *ex-situ* bioremediation of color and adsorbable organic halogen in pulp and paper mill effluent. In *Microbial Diversity Current Perspective and Potential Application*; Satyanarayan, T., Johri, B.N., Eds.; I.K. International Pvt. Ltd.: New Delhi, India, 2005; 573–584.
44. Nanda, S.; Sarangi, P.K.; Abraham, J. Cyanobacterial remediation of industrial effluents. N. Y. Sci. J. **2010**, *3*, 32–36.
45. Premraj, R.; Doble, M. Biodegradation of polymers. Indian J. Biotechnol. **2005**, *4*, 186–193.
46. Pandey, G.; Chauhan, A.; Samanta, S.K.; Jain; R.K. Che-motoxin of a *Ralstonia* sp. SJ98 toward co-metabolizable nitroaromatic compounds. Biochem. Biophys. Res. Commun. **2002**, *299*, 404–409.
47. Arora, D.S.; Chander, M. Biotechnological application of white rot fungi in biodegradation of various pollutants. In *The Environment in Biotechnological Applications of Microbes*; Verma, A., Podila, K.I.K., Eds.; International Pvt. Ltd.: New Delhi, India, 2007; 262–280.
48. Belliveau, B.H.; Trevors, J.T. Mercury resistance and detoxification in bacteria. Appl. Organometal. Chem. **1989**, ***3***, 283–294.
49. Chang, J.S.; Hang, J. Biotechnol. Bioeng. **1944**, *44*, 9991006.
50. Beveridge, T.J. Biotechnol. Bioeng. Sym. N. Y. **1986**, *52*, 127–140.
51. Volseky, B.; Chang, K.H. Biotechnol. Bioeng. **1995**, *42*, 451–460.
52. Wildung, R.E.; Gorby, Y.A.; Krupka, K.M.; Hess, N.J.; Li, S.W.; Plymale, A.E.; McKinloj, J.P.; Fredrickson, J.K. Effect of electron donor and solution chemistry on products of dissimilarity reduction of technetium by ***Shewanella putrefaciens***. Appl. Environ. Microbiol. **2000**, *66*, 2451–2460.
53. Weilingo, B.; Mizuba, M.M.; Hansel, C.M.; Fendrof, S. Environ Sci. Technol. **2006**, *34,* 522–527.
54. Kamaludeen, S.P.B.; Ramasamy, K. Rhizoremediation of metals: Harnessing microbial communities. Indian J. Microbiol. **2008**, *40*, 80–88.
55. Zhuang, X.; Chen, J.; Shim, H.; Bai, Z. New advances in plant growth promoting rhizobacteria for bioremediation. Environ. Int. **2007**, *33*, 406–413.
56. Abou-Shanab, R.A.; Ghanem, K.; Ghanem, N.; Al-Kalaibe, A. The role of bacteria on heavy metal extraction and uptake by plants growing on multi-metal-contaminated soil. World J. Microbiol. Biotechnol. **2008,** *24,* 253–262.
57. Marschner, H. *Mineral Nutrition of Higher Plants*; Academic Press: London, 1995; 889.
58. Blake, R.C.; Choate, D.M.; Bardhan, S.; Revis. N.; Barton, L.L; Zocco, T.G. Chemical transformation of toxic metals by a *Pseudomonas* strain from a toxic waste site. Environ. Toxicol. Chem. **1993**, *12,* 1365–1376.
59. Al Ayely, A.; Sylvia, D.M.; Ma, L. Mycorrhiza increase arsenic uptake by the hyper accumulator Chinese brake fern (***Pleris vittata*** L.). J. Environ. Qual. **2005**, *34,* 2181–2186.
60. Henrot, J. Health Phys. **1989**, *57*, 239–245.
61. Sabaty, M.; Avazeri, C.; Pignol, D.; Vermeiglio, A. Characterisation of the reduction of selenate and tellurite by nitrate reductases. Appl. Eviron. Microbiol. **2001**, *67,* 5122–5126.
62. Chatterjee, S.; Ghosh, I.; Mukherjea, K.K. Uptake and removal of toxic Cr (V1) by *Pseudomonas aeruginosa*: Physico-chemical and biological evaluation. Curr. Sci. **2011**, *101,* 645–652.
63. Mabbett, A.N.; Lloyd, J.R.; Macaskie, L.E. Effect of complexing agents on reduction of Cr(VI) by ***Desulfovibrio vulgaris*** ATCC 29579. Biotechnol. Bioengg. **2002**, *79*, 389–397.
64. Lovely, D.R.; Phillips, E.J.P. Bioremediation of uranium contamination with enzymatic uranium reduction. Environ. Sci. Technol. **1992**, *26*, 2228–2234.
65. Pal, A.; Paul, A.K. Microbial extracellular polymeric substances; central elements in heavy metal bioremediation. Indian J. Microbiol. **2008**, *48,* 49–64.
66. Singh, S; Srivastava, S. Rhizobia and its role in rhizoremediation. In *Microbial Diversity: Current Perspective and Potential Applications*. Satynarayana, T., Johri, B.N., Eds.; I.K. International Pvt. Ltd., New Delhi, India, 2005; 655–676.

67. Harrison, J.J.; Ceri, H.; Turner, R.J. Multimetal resistance and tolerance in microbial biolfilms. Nature **2007**, ***5***, 928–938.
68. Savitha, J.; Sahana, N.; Prakash, V.K. Metal biosorption by *Helminthosporium solani*—A simple microbiological technique to remove metal from e-waste. Curr. Sci. **2010**, ***98,*** 903–904.
69. Kojima, Y. Definitions and nomenclature of metallothioneins. Methods Enzymol. **1991**, ***205,*** 8–10.
70. Rashmi, K.; Haritha, A.; Balaji, V.; Tripathi, V.S.; Venkateswaran, G.; Maruthi Mohan, P. Bioremediation of ^{60}Co from simulated spent decontamination solution of nuclear power reactors by bacteria. Curr. Sci. **2007**, ***92,*** 1407–1409.
71. Vinod, V.T.P.; Sashidhar, R.B. Bioremediation of industrial toxic metals with gum kondagogu (*Cochlospermum gassypium*): A natural carbohydrate biopolymer. Indian J. Bio-technol. **2011**, ***10,*** 113–120.
72. Kao, C.M.; Liu, J.K.; Lou, H.R.; Lin, C.S.; Chen, S.C. Biotransformation of cyanide to methane and ammonia by *Klebsiella oxytoca*. Chemosphere **2003**, ***50,*** 1055–1061.
73. Prakasham, R.S.; Ramakrishna, S.V. The role of cyanobacterium in effluent treatment. J. Sci. Ind. Res. **1998**, ***57,*** 258–265.
74. Mario Garcia, D.; Luis, L.M.; Francisco, J.F.; Jose, C.R. A gene cluster involved in metal homeostasis in the cyanobacterium *Synechocystis* sp. strain PCC 6803. J. Bacteriol. **2000**, ***182,*** 1507–1514.
75. Cain, A.; Vannela, R.; Keith Woo, L. Cyanobacteria as a biosorbent for mercuric ion. Bioresour. Technol. **2008**, ***99***, 6578–6586.
76. Shukla, M.K.; Tripathi, R.D.; Sharma, N.; Dwivedi, S.; Mishra, S.; Singh, R.; Shukla, O.P.; Rai, U.N. Responses of cyanobacteria *Anabaena doliolum* during nickel stress. J. Environ. Biol. **2009**, ***30,*** 871–876.
77. Rugh, C.L.; Senecoff, J.F.; Meagher, R.B.; Merkle, S.A. Development of transgenetic yellow poplar for mercury phytoremediation. Nat. Biotechnol. **1998**, ***16,*** 925–928.
78. Rugh, C.L. Genetically engineered phytoremediation: One man's trash in another man's transgene. Trends Biotechnol. **2004**, ***22***, 496–498.
79. Sriprang, R.; Hayashi, M.; Ono, H.; Takayi, M.; Hirata, K.; Murooka, Y. Enhanced accumulation of Cd (2) by a ***Me-sorhizobium*** sp. transformed with a gene from *Arabidopsis thaliana* coding for phytochelation synthase. Appl. Environ. Microbiol. **2003**, ***69,*** 1791–1796.
80. Athaullah, A.; Asrarsheriff, M.; Sultan Mohideen, A.K. Phytoremediation of tannery effluent using *Eichhornia crassipes* (Mart.) Solms. Adv. Bio Tech. **2011**, ***3***, 10–12.
81. Thakur, I.S. Microbial bioremediation of pollutant chlorinated in the environment. In *Biotechnological Application of Microbes*; Verma, A., Podila, K.I.K., Eds.; International Pvt. Ltd.: New Delhi, India, 2007; 239–261.
82. Filler, D.M.; Lindstrom, J.E.; Braddock, J.F.; Johnson, R.A.; Nickalashi, R. Integral biopile components for successful bioremediation in the Arctic. Cold Reg. Sci. Tech-nol. **2001**, ***32,*** 143–156.
83. Crawford, R.L. The microbiology and treatment of Nitro aromatic compounds. Curr. Opin. Biotechnol. **1995**, ***6***, 329–336.
84. Thakur, I.S.; Verma, P.; Upadhyaya, K. Involvement of plasmid in degradation of pentachlorophenol by *Pseudomonas* sp. from a chemostat. Biochem. Biophys. Res. Commun **2001**, ***286,*** 109–113.
85. Thakur, I.S. Screening and identification of microbial strains for removal of colour and adsorbable organic halogen in pulp and paper mill effluent. Process Biochem. **2004**, ***39,*** 1693–1699.
86. Shah, S.; Thakur, I.S. Enrichment and characterization of a microbial community from tannery effluent for degradation of pentachlorophenol. World J. Microbiol. Biotechnol. **2002**, ***18***, 693–698.
87. Available at http://www.microbial-products.com/microbial-bacterial-products-bioremedation.asp (accessed December 2011).
88. Bioremediation 1999, available at http://www.inweh.unu.edu/447/lectures/bioremediation.htm (accessed December 2011).

89. Hamada, M.F; Haddad, A.I.; Abd-E-L-Bury, M. Treatment of phenolic wastes in an aerated submerged fixed-film (ASFF) bioreactor. J. Biotechnol. **1987**, ***5***, 279–292.
90. Lajoie, C.A.; Zylstra, G.J.; Deflaun, M.F.; Strom, P.F. Development of field application vector for bioremediation of soil contaminated with polychlorinated biphenyls. Appl. Environ. Microbiol. **1993**, ***59***, 1735–1741.
91. Thakur, I.S. Structural and functional characterization of a stable, 4-chlorosalicyclic acid degrading bacterial community in a chemostat. World J. Microbiol. Biotechnol. **1995**, ***119,*** 643–645.
92. Pazos, F.; Gurjas, D.; Valencia, A.; Lorenzo, V.D. Meta Router: Bioinformatics for bioremediation. Nucleic Acid Res. **2005**, ***33***, D.588–D592.
93. Fluekar, M.H.; Sharma, J. Bioinformatics applied in bioremediation. Innov. Rom. Food Biotechnol. **2008**, ***2***, 28–36.
94. Urbance, J.W.; Coli, J.; Saxena, P.; Tiedje, J.M. Nucleic Acid Res. **2003**, ***31,*** 152–155.

39

Composting

Nídia Sá Caetano

Introduction: Biological Waste (Biowaste)

Solid waste composition has varied since ancient times depending on the activity that originates it. One of the most important fractions is municipal solid waste (MSW), which represents a heterogeneous collection of wastes produced in urban areas, the nature of which varies from region to region.

Biowastes (biodegradable wastes) arise from living or once-living sources from several human, agricultural, horticultural, and industrial sources, in three groups—waste of directly animal origin (manures), plant materials (grass clippings and vegetable peelings), and processed material (food industry and slaughterhouse wastes and paper and paperboard wastes)—and include any waste that is capable of undergoing anaerobic or aerobic decomposition. Different terminologies have been applied to this kind of waste, such as *putrescible, green, food, yard, biosolids, garden,* or simply *organic wastes,* but chemically speaking, biowaste is characterized by high carbon content in the form of cellulose, hemicelluloses, and lignin, or even proteins and fat, that can be biologically degraded into carbon dioxide, methane, and water.[1]

Disposal Problems Associated with Biowaste

Disposal of biowaste either through uncontrolled landfill or if abandoned, presents some health and pollution issues that should be addressed. In fact, as biowaste is biologically degradable, it is a free food source for every kind of microorganism, including pathogens that could endanger an entire population.

Leachate

As water percolates through biowaste, it leaches out inorganic and organic compounds, with the risk of soil and groundwater contamination. Also, some persistent pathogens that can be found in long-term deposits of biowaste could endanger population. Landfill leachate is an organicrich liquor that is an excellent food source for heterotrophic microorganisms, but being so concentrated, its treatment is hardly achieved. Also, the existence of heavy metal contamination is toxic to microorganisms that could otherwise be successful in performing the leachate treatment.

Methane

The second pollution issue from disposal of biowaste is that methane (CH_4) is produced under anaerobic conditions that naturally occur in landfills. The problems with methane are that it is a greenhouse gas, with more than 20 times the damaging effect of carbon dioxide (CO_2), and that it remains in the atmosphere for approximately 9–15 years.

Regulatory Issues of Waste Management

Europe is committed to recycling as one of the main objectives of the waste management policy. Through recycling, materials contained in solid waste are reintroduced in the production cycle, leading to raw materials and energy savings and reducing the cost of landfill disposal.[2] Current European Directive on solid waste management[3] demands that countries adopt appropriate waste treatment methods, aiming to reduce the amount of waste sent to its final destination—landfill. Taking this into consideration, waste valorization through reusing and/or recycling, or by using other processes (energetic, organic), is also intended. The organic fraction of solid waste can be valorized by composting or anaerobic digestion and, according to the established in the Council Directive on the landfill of waste,[4] should be diverted from the flux of wastes to landfill. Taking the year 1995 as baseline, member states should reduce landfilling of biodegradable MSW to 75% by 2010, to 50% by 2013, and to 35% by 2020.

The organic matter (OM) in solid waste (currently constitutes about 40% of the MSW in Portugal, 25% in the United States,[5] almost 50% in Abu Dhabi City,[6] 50% in France,[7] and almost 77% in Brazil[8]) can be recycled to useful products (compost, methane gas, etc.,) through biological treatment processes. MSW valorization is mainly achieved by recycling constituents such as glass, metals, plastics, paper and cardboard, and OM, which is only possible when these residues are mostly collected selectively, although construction of mechanical biological treatment (MBT) facilities, comprising screening and other physical separation units, can contribute to the achievement of the established targets.

Composting is seen as a valuable recycling process for the organic fraction of MSW (OFMSW) and, thus, is of particular importance given the already existing systems and the potential to grow. Nevertheless, in the European Union (EU), it was not yet possible to come to an agreement on biological treatment of biowaste, in spite of the long work that has been done and that resulted in the publication of a Working Document on Biological Treatment of Biowaste in 2001.[9] This is not only due to the enormous differences in the degree of development of waste management of the various member states but also, in part, due to the existing lobbies.

Biological Aerobic Waste Treatment: Composting

Composting is the biological process used most often for the controlled aerobic conversion of OFMSW and any kind of solid and semisolid organic waste to a humus-like material, known as compost. Overall, the composting process can be represented by the reaction in Figure 1.

This exothermic process is realized in the presence of oxygen by a biological consortia of microorganisms and takes place in two distinct phases: a first phase, in which predominantly thermophilic biochemical degrading reactions occur (temperature rises as a result of the heat produced biologically), and a second phase, in which the humification/stabilization processes occur.[10,11] (Hogan and collaborators[12] suggested that the temperature rise results from the low thermal conductivity of waste.) Compost resulting from this process is a stable product, free of pathogens and plant seeds that can be applied, with benefits to the soil. This definition is intended to distinguish the composting process from the ordinary decomposition that occurs in nature.[13] Moreover, Bertoldi[14] clarified that the stabilization phase corresponds to a humification process that can be prevented under conditions of oxygen scarcity and substrate inadequacy.

Depending on the feedstock nature, the nitrogen, phosphorus, and potassium content of the compost may be insufficient for its classification as an organic fertilizer, allowing instead for its usage as a soil improver. This means that compost properties allow for soil pH amendment, acting as a source of OM that can enhance the water retention and cation exchange capacity of the soil and improve soil aeration.[15]

These are the reasons why composting is currently known as a process of recycling the OM in the solid waste and why using compost in soil represents the reintroduction of OM in soils, reducing erosion and thus desertification that is increased due to intensive land use.

Composting can be successfully applied to garden waste, separated MSW, mixed MSW, co-composting with sludge from urban and industrial wastewater treatment plants (WWTPs), and agricultural and livestock residues.

Composting Backgrounds

Organic soil correction with agricultural and livestock waste dates back to the utilization of soil for crop production, having been the principal means of restoring the nutrient balance in soil.[16] Composting is known, for a long time, by farmers as a method that allows obtaining an organic fertilizer from domestic waste. There are records of composting in piles in China for more than 2000 years, and there are even biblical references on the practice of soil correction. About 1000 years ago, Abu Zacharia described these procedures that have been practiced 3000 years earlier in the manuscript of *El Doctor Excellente Abu Zacharia Iahia de Sevilla,* translated from Arabic into Spanish by the order of King Carlos V, and published in 1802 as *El Libro de Agricultura.* In this book, Abu Zacharia insisted that animal manure should not be applied too fresh or directly to the soil, but only after mixing with 5 to 10 times its weight of vegetable and animal bedding waste.[17]

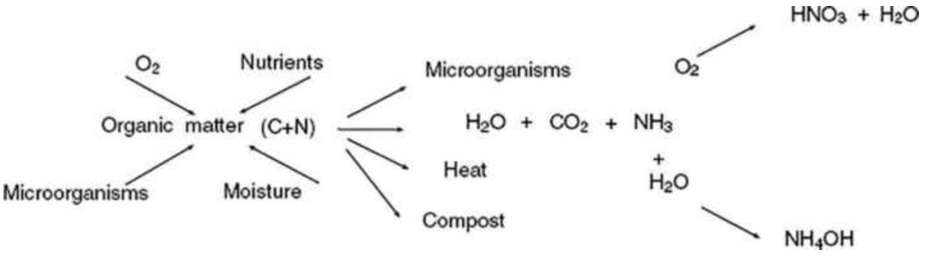


FIGURE 1 Schematics representation of the composting process.

In the growing cities in Europe, during the 18th and 19th centuries, farmers exchanged their products by MSW, using them as soil improver. Until the mid-20th century, MSWs were almost completely recycled through agriculture and did not pose an accumulation problem.[17] Composting of organic wastes and residues was envisaged as more of an art than a science until about the 1930s. By then, several developments of mechanical or intensive systems were achieved in Europe (Itano process in 1928, Beccari in 1931, and VAM in 1932). The Europeans continued to develop and install composting systems in Europe, South America, and Asia, and it was only in 1974 that the U.S. Department of Agriculture at Beltsville, Maryland, developed the "static pile" method that was currently used until the 1990s in the United States.[18]

By the end of 1960, composting was considered an attractive process for stabilization of the OFMSW with the final product being sold profitably as soil improver.[19] However, by the end of the same decade, composting was no longer that interesting for MSW management, not only due to the lower quality of solid waste but also due to the lack of market for compost. Recent stress on usage of less environmental impacting methods has redirected interest into the composting process, particularly concerning the recycling of MSW and urban and industrial WWTPs.

Despite the fact that implementing this process to agricultural wastes is ancient, scientific support was only established in the early 20th century, mainly due to the work done by an English agronomist (Sir A. Howard, 1924–1931) in Indore, India. This agronomist established the fundamental principles for the maintenance of a microbial population in optimal conditions of activity: the need to mix vegetable and animal wastes, the need for neutralization of the fermenting biomass, and the need for provision of adequate amounts of air and water. Thus, materials were stacked in piles (windrows) that have a dimension of 9×4.2 m and a height of 60 cm, which allows for maintaining adequate levels of heat and humidity.[20] Construction of these windrows was made by successive layers of waste (manure, soil, and straw) and moistening them conveniently. Windrows were revolved 16, 30, and 60 days after the start of the procedure that needed 90 days before completion and incorporation of the product in soil could be done.

Although windrow composting was the most common practice, during 1950–1960, there was a huge amount of publicity for projects involving composting in reactors.[21,22] However, these projects had almost universally poor results and had an abrupt end due to bad performance and high economic costs. Bad performance was generally the result of an inadequate project or operation and not of the process or technology itself. In the 1950s, there were already more than 20 patented composting systems.[23] In this period, mechanical separation for removal of any non-compostable from the waste stream was initiated.[14]

Co-composting of MSW and biosolids (composting simultaneously MSW and sludge) has attracted increasing attention.[24] MSWs are used as bulking agent and the biosolids act as readily available nitrogen and humidity source. This technique was investigated and applied fully and systematically in the 1950s,[25,26] but moisture content (96%) of the digested solids was a limiting factor as biosolids dehydration was not frequent by then.[27]

In the EU, production of agricultural and food wastes exceeds 1 billion tons/yr, which is 3 times larger than the production of sludge and 6 times higher than the production of MSW.[28] Intensive livestock farming incrementation worsened the problem through production of large amounts of animal waste, often in specific locations. Despite this, the relatively reduced application of composting in the management of this waste should be noticed, especially when compared with the application of this process in the management of MSW.[29] This is probably due to the fact that the agricultural wastes are often applied directly to soil, without any previous treatment, and that composting is generally considered a process for pollution control rather than a beneficial and efficient process for nutrient recycling.

Taking into account the requirements of EU legislation regarding waste management and environmental protection, and because there is nowadays a greater awareness of the importance of controlling the loss of nutrients in the waste treatment processes, recently, there has been a greater research effort to develop strategies that can control gaseous emissions, stabilize OM, and ensure nutrient retention and the absence of toxic products or pathogens in the compost. There are currently several groups of European researchers working on specific issues of composting of livestock and agricultural waste (such as kinetics of composting of MSW, effect of the composition of agricultural waste on compost quality and composting kinetics, effect of contaminants on composting process and on compost quality, etc.), aiming to produce good quality compost.[24,29–34] It is expected and highly desirable an increase in the world-wide application of composting for the treatment of these wastes and the use of its compost for agricultural purposes.

Composting has been successfully applied, including for the waste treatment of animal slaughter and carcasses.[35] Mesquita[36] enumerated some composting projects, where different raw materials were successfully used, and Williams[37] identified several large-scale separate composting schemes implemented in Europe.

Advantages and Drawbacks of Composting

If correctly used, the composting process for MSW treatment presents several advantages. When it is a part of a MSW integrated management system, these advantages are even more important; thus, valorization of the OFMSW by the integrated management can be achieved, and consequently, the final product quality can be improved.

A relevant issue is that besides being a process of effective organic waste treatment, composting is also an excellent recycling process, retaining in the final product (compost) macro- and micronutrients that may return to the soil and supplement other energy cycles in nature.[38]

The main advantages of the composting process are as follows:

- Economy and natural resource preservation and reuse of OM and of macronutrients (N, P, K, Ca, and Mg) and micronutrients (Fe, Mn, Cu, Zn, etc.,) owing to composting being a recycling process.
- Environmental benefits as a result of elimination of air, soil, and water pollution due to proper disposal of organic wastes or their appropriate confinement in controlled landfills. Also, there are additional benefits from using organic compost (instead of artificial fertilizers and soil improvers): in the recovery of degraded, contaminated sites and salty soil, in reforestation, in soil erosion control, etc.
- Public health benefits, due to the elimination of pathogens. In developing countries, this issue contributes to child mortality control.
- As an offshoot of health benefits, e.g., prevention and elimination of diseases, economic benefits arise, contributing to the reduction of costs of treatment and increased productivity. Another benefit is the elimination of costs for land remediation of uncontrolled dump sites.
- Social benefits arise from the fact that in addition to eliminating the practice of using organic waste as food for animals, composting promotes employment either in the treatment units (selective collection, sorting, processing) or in compost utilization and application.

The main disadvantages arising from composting are as follows:

- Being a labor-intensive process, composting gives rise to operational costs (compensated with positive social impact).
- Producing compost from selective organic collection entails high costs.
- The possibility of compost contamination (if the adopted process is not technologically appropriate and poorly conducted).

Methods of Composting

Composting is a simple technology that requires a fairly low intervention and has modest initial, operational, and support costs. It can be very attractive to authorities charged with biowaste management but with reduced budgets.

There are many factors that can influence the decision on the specific details of composting methods that should be adopted, but composting of OM in MSW or similar materials will most naturally be based in one of two options, either source-separated biowaste or biowaste recovered from MBT plants. Either of these options will ultimately be put in practice in home composting systems or in centralized composting facilities.

While home composting takes advantage of many distributed and local production units, centralized or large-scale facilities comprise two large groups of systems, both open and closed.[39] Among open systems of composting, three types can be distinguished: windrow (or revolved piles), the static aerated piles, and vermicomposting.[30] Closed systems of composting include reactors (either horizontal or vertical) of different sizes and shapes equipped with different feeding and aeration systems, aimed at accelerating the startup of the oxidation process, especially in cold climate countries, thus allowing for a more efficient control of odor emissions and minimizing the land area requirements.[37]

The choice of a specific type of system relies essentially on socioeconomic factors, on the amount of biowaste to compost, and on its final destination. Open systems are generally more suitable for developing countries[40,41] and for agro/livestock facilities[42] due to their easier operation and lower mechanization, building, maintenance, and operational costs.

Processing of agricultural and livestock waste using any of these technologies not only presents substantial benefits from the health, economic, and environmental points of view but also allows for a safe and potentially useful final product that can be used as a soil conditioner, valorized as a fertilizer, or for the floriculture and horticulture industries. Its use for agricultural purposes is important, both in countries where soil is extremely poor in OM (such as those of southern Europe) and in countries where extensive use of inorganic soil fertilizers has endangered its structure.[40]

The most common technologies, particularly for composting the OFMSW and similar biowaste, are presented below.

Home Composting

Home composting is one of the most interesting ways of managing biowaste. In fact, while people compost at home, they are guaranteeing the removal of OM from MSW flux, thus facilitating the achievement of the goals established in the EU Landfill and EU Waste Management Directives, as well as contributing to the reduction of waste management costs by reducing the total amount of waste that needs treatment. Also, by using home-produced compost, households do not need to use other fertilizers or soil conditioners; hence, they also have lower garden maintenance costs. Home composting demands the direct involvement of households, leading to higher separation rates of other recyclables as well.

One of the biggest issues in home composting is the choice of the compost bin. Although there are several configurations available in the market (Figure 2), some of them are far more efficient than others

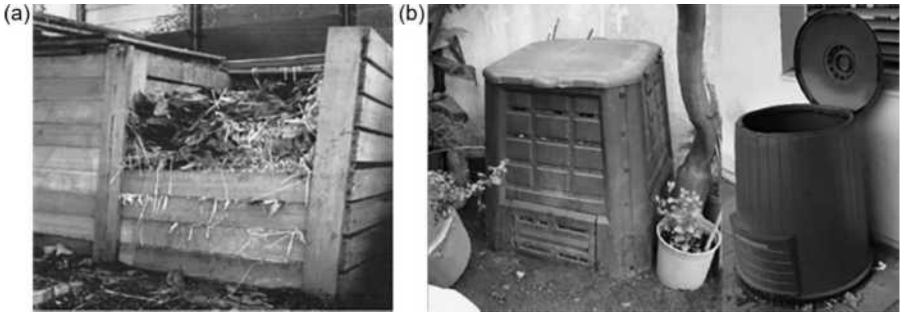


FIGURE 2 Home composters: (a) Wood, homemade, with two composting chambers, 1 m^3 each. (b) Commercial plastic, with openings for air circulation (0.4 m^3), offered by the Terra-à-Terra project.

and their cost is not always according to quality. In some cities, there are special programs that try to address this problem, reducing costs to the consumer. In Porto, Portugal, Lipor has been developing a project—Terra-à-Terra—that aims to promote organic waste reduction at the households of Lipor's municipalities. Householders with a garden or that work in companies with a garden in the project area can receive a free compost bin after attending a free composting training session. Through this project that started in 2007, more than 4500 composters have been delivered in 3 years, and it is estimated that there is a potential biowaste reduction of about 3000 tons/yr that will prevent the emission of 528 tons of CO_2 per year if this waste is treated at the energy recovery plant.[43]

Centralized Composting

Seasonal variation in the composition of the MSW requires a flexible and competent management of the composting system.[1,44] For this reason, careful choice of the system and a thorough control of the facilities are of primary importance.

Windrows

This technology is based on the Indore composting method. Rows of parallel piles of material with a height of 1.5–2 m and a width of 3–4 m are constructed until they reach 100 m or more. The shape of the piles is such that sufficient heat is generated to maintain temperatures while allowing for oxygen diffusion into the center of the pile, through natural convection. Batteries of piles must be constructed on an impermeable surface (e.g., cement) so as to facilitate the collection of leachate that eventually will form and so that piles can be easily turned. Windrows must be turned periodically (initially at the end of 5 days and then every 2 or 3 days), as shown in Figure 3. The convex shape of the windrows allows rainwater to drip along the surface and not infiltrate the pile of material. This system has some drawbacks such as being prone to anaerobiosis due to layer compaction, reduction of free space for aeration, and production of leachate that fills the remaining voids between particles, which further contributes to reducing aeration.

Aerated Static Piles

Piles are similar to those in windrow composting. However, the material is not mechanically aerated. Temperature control is achieved by natural air convection or forced aeration, either by introduction of compressed air or by vacuum induction (Figure 4).

The material for composting remains stacked during a very long period, which might result to collapsing and thus pore clogging. To prevent this, it is a common practice to mix biowaste with some bulking material (such as wood chips or sawdust) that should be more stable and have a higher particle size, thus providing a structure that will help prevent the material in the pile from collapsing. Aerated piles should be covered with a layer of compost or wood chips that confer some insulation, act as a biofilter, and reduce odor release.[42]

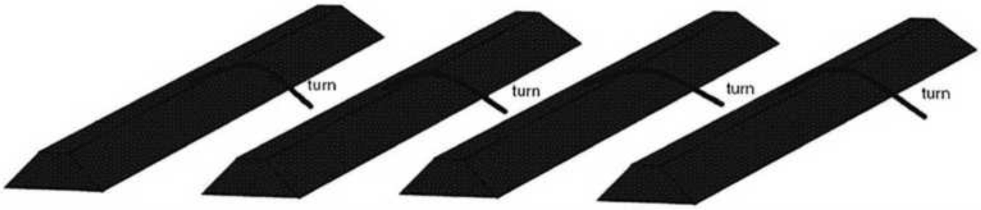


FIGURE 3 Schema of windrow composting and how to turn composting piles; the upper material will stay in the bottom after turning.

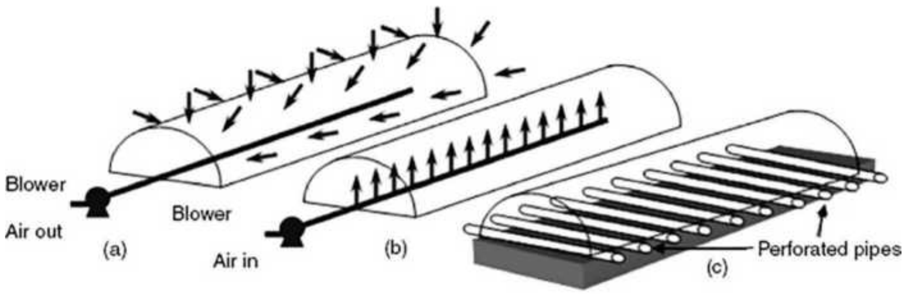


FIGURE 4 Static pile composting: (a) with suction of air from the outer layers to the base of the piles; (b) with compressed air injection through the base of the pile; and (c) natural convection.
Sources: (a) from Risse and Faucette;[45] (b) and (c) adapted from Graves and Hattemer.[46]

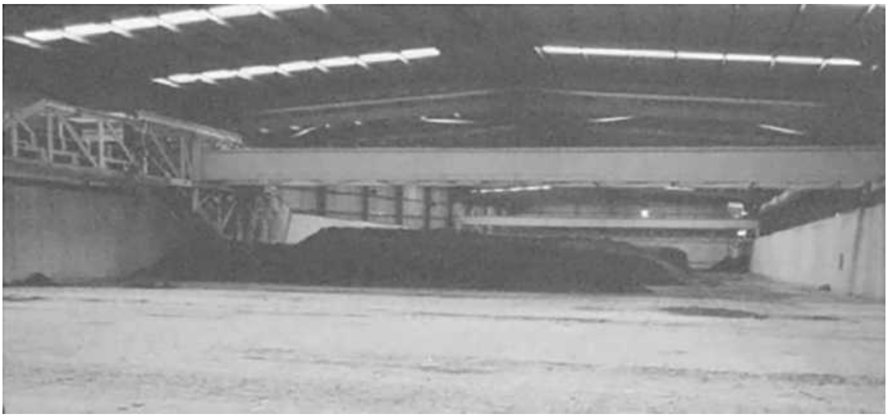


FIGURE 5 Indoor aerated static pile system with aeration through vacuum induction.

These composting technologies were traditionally implemented in open courtyards. Currently, and in regions where environmental conditions are particularly unsuited to this type of processing, these technologies are implemented in covered spaces or indoors (Figure 5), in which case, it is possible to perform air treatment in biofilters, achieving a dramatic reduction in odorous emissions.

In-vessel

Due to being enclosed, in-vessel systems allow for tighter control of temperature, moisture, aeration, and biowaste mixing. In-vessel systems include different configurations such as tunnels, rotating drums, reactor tanks, silos or towers, agitated bays and beds, enclosed halls, or even containers.[37] Although the operating systems are essentially the same as for windrows or aerated piles, these reactors are much more efficient, thus needing much less area.[1]

Silos or Towers

This technology is based on the use of silos or other types of structures with a height exceeding 4 m. These reactors are fed at the top, via a distribution mechanism, and the material progresses towards the bottom by way of gravity, usually after 14 days of composting. After this period, a curing time of ca. 2 months is needed to stabilize the compost.[42] Process control is commonly performed by air injection at the base of the reactor, moving upwards countercurrent to the biowaste (Figure 6).[47] Due to the great height of the reactor, it is necessary to use large airflow per unit surface area of distribution, which

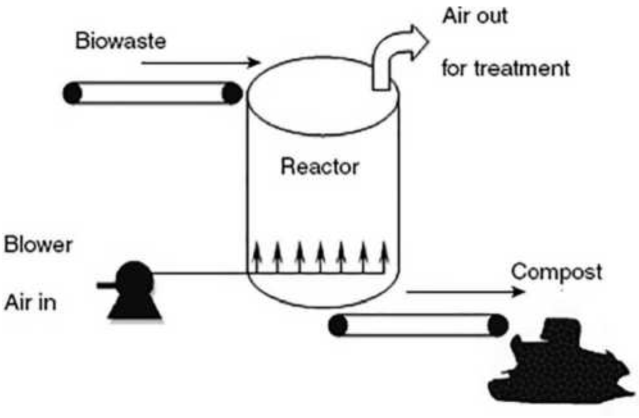


FIGURE 6 Schematic drawing of vertical reactors system for composting.

makes it difficult to control the process.[48] It is not easy to keep the temperature and oxygen at optimal levels, due to material compaction and to inexistence of mixing in the silo.[42] To minimize these problems, material should be thoroughly mixed prior to feeding the silo and air distribution and collecting systems must be improved, by changing the direction of airflow from vertical to horizontal between alternating sets of inlet and exhaust pipes.[44]

This type of reactor has been successfully used for composting of sludge from WWTPs (where the addition of bulking materials for porosity enhancement and a uniform feed facilitates process control) but is rarely used for processing heterogeneous materials, such as the OFMSW.[44]

Horizontal Reactors (Tunnels and Agitated Beds)

These reactors allow controlling the occurrence of high temperature, moisture, and oxygen gradients, which are frequent in vertical reactors, because air needs to travel lower distances through the material. There are a variety of configurations, comprising static or agitated reactors, with aeration by air injection or vacuum induction. Agitated systems usually take advantage of the revolving process to displace the material along the reactor continuously, while the static systems usually need loading and unloading mechanisms. Handling equipment can also grind the material, exposing new areas for decomposition, but excessive grinding can also reduce porosity. Typically, aeration systems are installed at the base of the reactor and can use temperature and/or oxygen concentration as control variables. The agitated systems with less than 2 to 3 m thick beds have been successfully used in heterogeneous MSW treatment (Figure 7, left).

FIGURE 7 Horizontal composting systems: (left) tunnel composting system composed of horizontal parallel beds; (right) turning machine for a tunnel composting system.

Tunnels are formed in long parallel channels (Figure 7, left), made of concrete walls with a movable porous floor, like a grid to allow for greater airflow (usually forced aeration) under the compost, and covered by a roof. Biosolids are revolved by a turning machine (Figure 7, right) running on tracks along the concrete walls between channels. The turner movement along the bed displaces the compost until it is ejected at the end of the bed.[49] The duration of the composting process is determined by the length of the bed and the turning frequency, ranging from 6 to 20 days, after which the compost must be further processed in windrows or aerated static piles for 1 to 2 months. Agitated bed systems operation ranges from 2 to 4 weeks; also, an extended curing period is needed to stabilize compost.[18]

Rotary Drums

In these reactors, biomaterial stays for only a few hours or days. Their effect is essentially homogenization and grinding of materials, only allowing start of composting by temperature control. As the reactors (diameter, 2 to 4 m; length, up to 45 m) have a slight inclination and have a rotating movement (0.5 to 2 rpm), the flow of material is continuous, countercurrent to the air supply; thus, the material leaving the reactor is cooled by fresh air and the material entering the drum makes contact with warm air, which favors bacterial growth (Figure 8).[42] These reactors are particularly interesting in processing the OFMSW for composting[46] as they allow for a faster startup of the process while reducing the malodorous emissions to the atmosphere. The DANO drum is a commercial rotating cylinder that has retention times of about 3 days with eventual interest in modern MBT plants. Most of the biological process has to be realized after leaving the reactor.

All of the systems described above can be used for composting different types of materials. Obviously, capital and operational costs of reactor processes are significantly higher than those of a windrow or aerated static pile process; thus, the residence time in the reactors is rarely enough to obtain mature compost. For this reason, a reactor is used in the early stages of composting to allow an easier control of the process, which facilitates odor emission reduction. After this processing, the material that comes out of the reactor must be stabilized in windrows or aerated static piles.[44] As the OFMSW treatment is difficult due to high cellulosic carbon content and sometimes low moisture and low porous structure, it may take as long as 6 months for compost maturation and stabilization, unlike compost from WWTPs that usually only takes 2–3 mo to be stabilized.

Vermicomposting

Vermicomposting is an alternative technology to conventional composting in which selected species of earthworms (*Lumbricus terrestris* and *Dendrobaena veneta*) and red worms (*Eisenia foetida,* Figure 9) eat organic material, absorb the nutrients they need, and excrete the rest, producing a humus-like material, known as vermicompost.[1,50] Vermicomposting should be applied preferentially to residues of fruit and vegetables, tea leaves, tea bags, coffee grounds, paper, and shredded green garden waste and has large application in agricultural wastes, including manure wastes.[51] Vermicompost produced from

FIGURE 8 Composting rotary drum: (left) side view and (right) view from the waste admission side.

FIGURE 9 *Eisenia foetida* in a home/commercial vermicomposter.

OFMSW is a "nutritive biofertilizer" 4–5 times more powerful than conventional composts and even better than chemical fertilizers for better crop growth and safer food production.[52]

Vermicompost systems can range from inexpensive wood or plastic boxes to sophisticated modular units (Figure 9) that are self-contained and fully automated to keep controlled environmental conditions while processing waste into earthworm castings. Open systems range from windrows or beds of variable scales to open field operations. Usually, windrows are built no more than 30 cm deep for easy aeration of the beds.[51] Vermicompost has the advantage of being a material that growers can produce "on-farm" and use as a biofertilizer, using own feedstock and implementing either a midscale vermicomposting unit technology or a modular unit.[1]

Worm beds can be as long as 50 m; feedstock can be the resulting product of a MBT plant, whereas OFMSW is processed in a rotary drum. Worms should be kept at temperatures of 10–35°C. Moisture is fundamental to worms; they breathe oxygen through their moist skin.[51] Resulting worm casts can be used as a biofertilizer as they contain 5 times more N, 7 times more P, 1.5 times more Ca, 11 times more K, and 3 times more exchangeable Mg than the soil. These casts are also rich in humic acids (which condition the soil), have a perfect pH balance, and have plant growth factors similar to those found in seaweed.[51]

Under the optimum temperature (20–30°C) and moisture (60%–70%) conditions for worm breeding, about 5 kg of worms (ca. 10,000) can process 1 ton of waste into vermicompost in 30 days.[53]

Worms can be easily recovered for further processing because they do not like light (a few hours of direct exposure to sunlight can cause paralysis or death), dryness, or even some specific odors (onions), migrating naturally for places where fresh wastes have been added. Worms can also be harvested by using a trommel screen.[54]

Vermicompost produced from organic wastes, such as food and yard wastes, have enormous economic potential for increasing crop yields, suppressing attacks from pests, and controlling the spread of diseases.[55–57]

Evaluation of a Composting System

Among the various criteria used for assessing the system efficiency, the most important for MSW composting are product quality, rejected ratio, and recycling rate. Compost quality can be evaluated by its appearance and by functional characteristics and the nature of the contaminants, characteristics that are critical for its commercialization. Rejected ratio and recycling rate affect the quality of the product, as they affect the contaminant's concentration and the amount of final waste that must be disposed off in landfill.

Compost quality should constitute the main objective of the composting facility. Some of the aspects of quality, such as the degree of maturation and size of particles, can be corrected at the end of the process by increasing the curing time and by using physical processes for particle size reduction. However, other issues may be far more difficult to remedy, as when there is chemical contamination.

A composting factor should be not only thoroughly thought about when conceiving the biowaste management system but also observed during the initial phases of processing. Chemical contamination can be caused by heavy metals or other chemicals, often from domestic hazardous waste. Contamination can also be of physical type; that is, the product may be contaminated with inert materials (glass, plastics and metals, brick and concrete, etc.,) that arise from domestic or commercial waste either voluntarily or by mistake. Removal of physical contaminants may not be very difficult from a technological point of view but represents a high fraction of the economic costs of a composting facility, which cannot be completely eliminated because even source-separated waste can be contaminated. However, the use of a source-separated raw material is of primary importance for high-quality assurance of the compost.

Phases of Composting

Throughout the composting process, the microbial population will vary and different types of microorganisms will play distinct roles during the various stages of the process.[38] Thus, in the first step, there will be predominantly mesophilic bacteria that hydrolyze the easily fermentable organic material. The reactions are exothermic, leading to heat release and temperature rise. Thermophilic microorganisms (bacteria, fungi, and actinomycetes) begin to develop on a large scale—from 40°C.[58] The thermophilic microorganisms multiply, and as soon as the temperature reaches 55–60°C, the attack on complex molecules (carbohydrates, proteins, etc.,) starts, which will result in their transformation into simpler products (simple sugars, amino acids, etc.,) that are used by other microorganisms (Figure 10). If there is no external control, the temperature of the composting material can reach 80°C, which represents the limit for the thermophilic population, resulting in microbiological activity inhibition and survival only of spores of bacteria. As degradation is reduced, and temperature lowers, mesophilic bacteria, actinomycetes, and fungi (mostly those that were in the outer areas of the pile) gain further activity, attacking the most resistant compounds (such as lignin and cellulose, the less readily biodegradable components of biowaste).[1] Complex enzymatic reactions are responsible for humus production, mainly by lignin and protein combination. At this stage, protozoa and some higher organisms (worms, nematodes, and millipedes) can be found in compost.

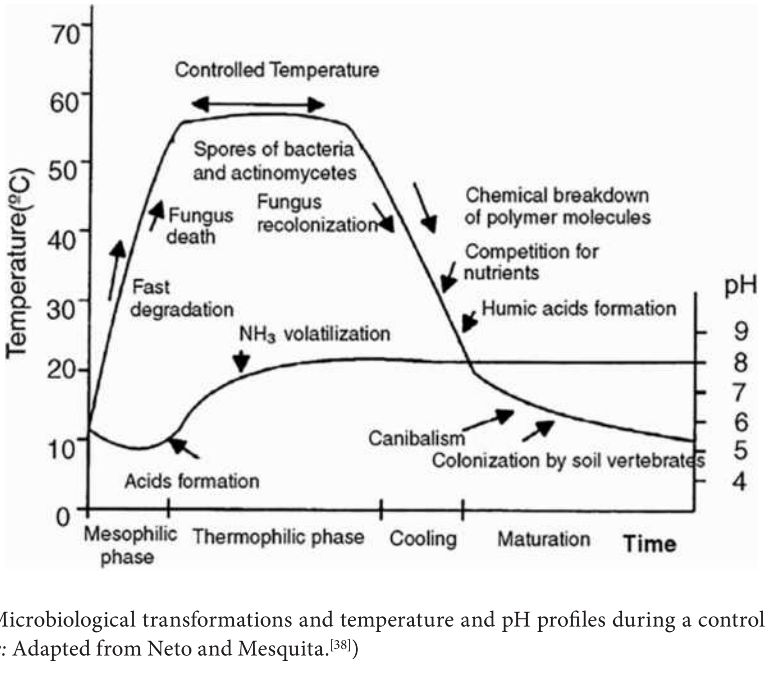


FIGURE 10 Microbiological transformations and temperature and pH profiles during a controlled composting process. (*Source:* Adapted from Neto and Mesquita.[38])

Process Parameters in Composting

Preparation of a composting process is not a simple task, especially for achieving optimal results. For this reason, in commercial applications, mechanization is a key factor, allowing for efficiently controlling the most important project factors (temperature, pH, moisture, C/N ratio, aeration rate, particle size, and mixing/turning).[58]

As any biological process, composting is conditioned by factors that affect the activity of the microorganisms involved in the process.

Temperature

Temperature is the process parameter that best describes the biological equilibrium and shows the efficiency of the process. Aerobic composting systems operate under mesophilic (20–40°C) and thermophilic (40–60°C) conditions that change along the process as a result of the respiration and metabolic activities of the organisms involved in the composting process.[1]

Temperature control of the composting material under static pile or in-vessel processes can be achieved through temperature monitoring and control of the air flow rate. In these processes, temperature can be controlled in the thermophilic range, around 60°C, which allows for the development of a more complex microbiological consortia, responsible for increasing the rate of decomposition of the OM (higher process efficiency), better compost sanitation or pathogen elimination, and elimination of weed seeds, insect larvae, and parasite eggs, among other advantages of the process.[38]

In windrow composting systems, temperature can only be controlled indirectly, by changing the turning frequency of the composting material. After turning, temperature lowers 5–10°C, rising again after only a few hours, as a result of the increase in oxygen availability. On the other hand, if temperature rises above 70°C, most of the microorganisms will die or enter the dormant phase, slowing the composting process and resulting to lower-quality compost. Also, high temperature can lead to humidity and nitrogen loss (through ammonia volatilization at a pH of 7.5).

In a system adequately controlled, temperature in the composting pile will rise up to 40–60°C from the second to the fourth day and will lower down to 35–38°C after 10 to 15 days because easily biodegradable organic material has already been converted. The increase in temperature also depends on factors such as nutrient availability, moisture content, particle size, aeration, turning of the pile, and the thermal insulation of the system (either exposed systems or confined systems).

Typical temperature profiles in non-controlled composting systems (with and without turning) are shown in Figure 11.

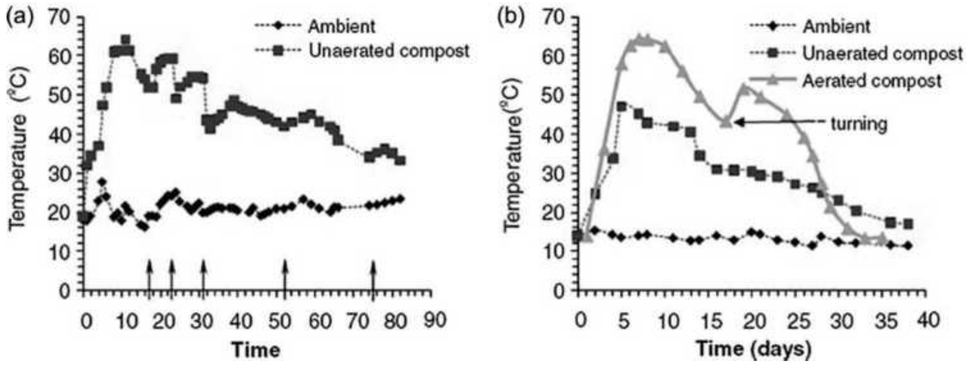


FIGURE 11 Typical temperature history observed in (a) windrow composting systems with turning[59] and (b) static piles with and without forced aeration.[59]

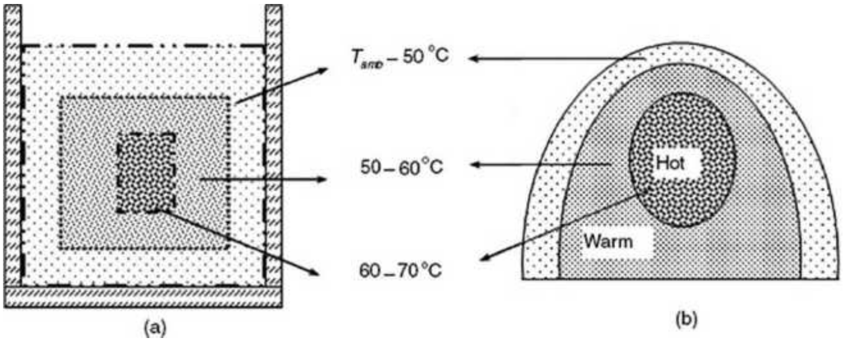


FIGURE 12 Temperature profile in a composting pile: (a) in-vessel/tunnel; (b) windrow.

It should also be noted that temperature is usually not uniform along the composting pile, the inner material having a higher temperature than the outer layers, as shown in Figure 12.

Moisture

Decomposition of the OM mainly depends on moisture content to enable microbiological activity. Microorganisms have, in their structure, about 90% water that is needed not only for new cell production but also and specially for the dissolution of the nutrients needed for cellular metabolism.

Moisture content can be adjusted either through mixing different materials or through water addition to an optimum between 40% and 70%. Below 30% humidity, composting rate dramatically decreases, but more than 70% moisture content may lead to pore filling with water, thus preventing oxygen from reaching the OM and inducing anaerobic conditions. This is why the most recommended value is 60%.[1,42]

Also turning of the composting material can help control moisture content. If the organic material has 55%–60% humidity and the composting period is 15 days, the first turning should be done on the third day and then every other day.[58]

Aeration

Aeration is very important in composting and holds two main functions: the first, and most important one, is to supply aerobic microorganisms with the oxygen they need for their respiration and metabolic activity, and the second function is to control temperature and, consequently, the rate of oxidation of the OM as well as odor emission.

Theoretically, the optimal rate of aeration would be the one that would allow for biological oxygen demand (BOD) supply during all the different composting steps. Nevertheless, parameters such as the nature of the composting material, particle size, aeration technology, moisture content, and porosity of the material in the pile may prevent full satisfaction of this requirement.[23] On the other hand, excessive aeration may cause excessive loss of humidity and heat leading to the misidentification of the end of the composting process.

In static pile systems, initial oxygen concentration in the pores is similar to that in the surrounding air (ca. 17%), because this is the air that was trapped in these spaces when the pile was built, whereas CO_2 concentration is significantly lower (ca. 0.5%–5%). As aerobic degradation occurs, oxygen concentration will decrease and CO_2 concentration will increase. Anaerobic conditions will develop at O_2 concentration lower than 5%, which can be prevented through introduction of air, vacuum induction, or turning of the material in the composting pile (which will also allow for a more efficient distribution of nutrients and microorganisms).

In systems with forced aeration, air flowrate and total air needed in the process are fundamental project parameters that help maintain the level of available O_2 very close to the required level throughout the whole process. The total amount of oxygen required and the total amount of CO_2 produced in the process can be estimated through the mass balance in Equation 1, where the composition of the OM for degradation and of the compost can be represented as $C_aH_bO_cN_d$ and $C_wH_xO_yN_z$, respectively.[58]

$$C_aH_bO_cN_d + 0.5(ny + 2s + e - c)O_2 \rightarrow$$

$$nC_wH_xO_yN_z + sCO_2 + rH_2O + (d - nz)NH_3 \tag{1}$$

with $r = 0.5\,[b - nx - 3\,(d - nz)]$ and $s = a - nw$.

If ammonia produced in the process is converted into nitrate (nitrification), some more oxygen will be needed, as estimated using Equation 2.

$$NH_3 + 2O_2 \rightarrow HNO_3 + H_2O \tag{2}$$

Particle Size

It could be expected that the lower the particle size of the organic material, the higher would be the rate of biochemical reactions, due to the increase in surface area exposed to microorganisms and to oxygen, which translates to the composting period being reduced. However, if particle size is too small, porosity will be reduced along with oxygen and CO_2 diffusion, which is fundamental during the thermophilic phase of the process, when oxygen consumption is higher.

An optimum particle size of 25–50 mm has been accepted for most of the materials, but this can vary between 13 and 75 mm depending on the nature of the composting material. A shredding operation may be needed prior to composting, which will entail an increased operational and financial cost.

Porosity of the composting material can be enhanced through introduction of bulking material that will allow for easier aeration while keeping structural characteristics that are essential for the construction of composting piles. Also, the height of the composting piles is based on particle size, in order to avoid excessive compaction during composting.

Feedstock Composition

The biomaterials used in composting have such a diversified composition that can usually supply all the elements that are fundamental for microbial metabolism (carbon, nitrogen, phosphorus, potassium, calcium, iron, copper, etc.).

Of these, the most important is undoubtedly the C/N ratio; an optimal ratio of 30/1 is recommended. Carbon is the source of energy and an essential factor for multiplication of cell material, and nitrogen is present (and available) in such an amount that it can be used for protein, amino acid, and nucleic acid formation.

Mixing of wastes with different compositions is usual since hardly ever a substrate has all the characteristics required for efficient and effective composting, leading to a compost with the recommended values of carbon/nitrogen ratio (C/N), moisture content, particle size, density, etc. Usually, these mixtures are prepared using carbon-rich vegetable waste and separate solid material (newsprint paper, manure, slurry or sludge, yard wastes) characterized by containing high levels of nitrogen, in order to obtain a C/N ratio of about 25/1 to 30/1[30] (Table 1). If the C/N ratio is lower than 30/1, there will be a loss of nitrogen as ammonia, causing malodorous emissions, but if C/N is higher than 80/1, nitrogen concentration will be so low that it will become a limiting factor. It must also be taken into consideration that in the OM, not all carbon will be in the readily available form, i.e., biodegradable, in contrast to nitrogen

TABLE 1 Nitrogen Content and C/N Ratio of Common Materials Used in Composting (Dry Basis)

Material	N (%, dry weight)	C/N Ratio[a]
	Food processing wastes	
Fruit wastes	1.52	34.8
Mixed slaughterhouse wastes	7.0–10.0	2.0
Potato tops	1.5	25.0
Coffee grounds		20.0
Vegetable wastes	2.5–4.0	11–13
	Manure	
Cow manure	1.7	18.0
Horse manure	2.3	25.0
Pig manure	3.75	20.0
Poultry manure	6.3	15.0
Sheep manure	3.75	22.0
	Sludge	
Digested activated sludge	1.88	15.7
Raw activated sludge	5.6	6.3
Wood and straw		
Lumber mill wastes	0.13	170.0
Oat straw	1.05	48.0
Sawdust	0.10	200.0–500.0
Wheat straw	0.3	128.0
Wood (pine)	0.07	723.0
	Paper	
Mixed paper	0.25	173
Newsprint	0.05	983
Brown paper	0.01	4490
Trade magazines	0.07	470
Junk mail	0.17	223
	Yard wastes	
Grass clippings	2.15	20.1
Leaves (fleshly fallen)	0.5–1.0	40.0–80.0
Tree trimmings	3.1	16
	Biomass	
Water hyacinth	1.96	20.0
Bermuda grass	1.96	24

Source: Rynk et al.[42] and Tchobanoglous et al.[58]

[a] C/N ratio based on total dry weight.

that is readily available. Thus, lignocellulosic material that is rich in carbon is very hardly biodegradable, so when it is present, the most adequate value for the C/N ratio would be 35/1 to 40/1.[60]

During the composting period, C/N ratio slowly decreases because when OM is degraded, about 65% of the carbon is released as CO_2 and the remainder is used with nitrogen in cell, which is released only when there is cellular death. The final compost should have a C/N ratio of 15/1.

pH Control

Although authors agree that pH affects biological processes, composting studies using MSW and wastewater sludge where the initial pH of the composting material was varied have shown that a self-regulation of pH occurs;[10] thus, pH is not a critical factor in the composting process. Most of the times, there is no need for pH correction, although in few particular situations, lime can be added for pH correction.

Most of the bacteria involved in the process operate better in the pH range of 6 to 7.5, but fungi prefer a pH of 7.5 to 9. In the composting process, there is a pH variation across time. Thus, in the first phase of the process, simple organic acids are produced, lowering pH to 5 or even less. After only 3 days, pH starts rising up to 8–8.5, as organic acids are being further transformed. In the cooling phase, the pH of the compost will be slightly reduced until 7–8 (Figure 13). Ideally, pH must be less than 8.5 so that nitrogen loss (as ammonia) is minimized.[58]

Odor Control

Odor problems arise from development of anaerobic conditions that allow for formation of organic acids, or ammonia release, most of which are extremely malodorous. Nevertheless, material in the pile can act as a biofilter that retains and degrades part of these compounds.

Pathogen Control

Pathogen destruction is of primary importance in composting systems as it is affected by temperature profile and aeration process. Pathogens can be eliminated under different conditions (*Salmonella* is destroyed in 15–20 min at 60°C, but at 55°C, it will take 1 hr). As a recommendation, if biowaste is kept at 70°C for 1–2 hr, all pathogens will be eliminated. It should be noted that in the second stage of the composting process (humification), the *Penicillium* fungus that destroys pathogens, acting as an antibiotic, appears.

Most European countries recommend a temperature of 55–65°C for 4–15 days to guarantee compost sanitization.[1,58]

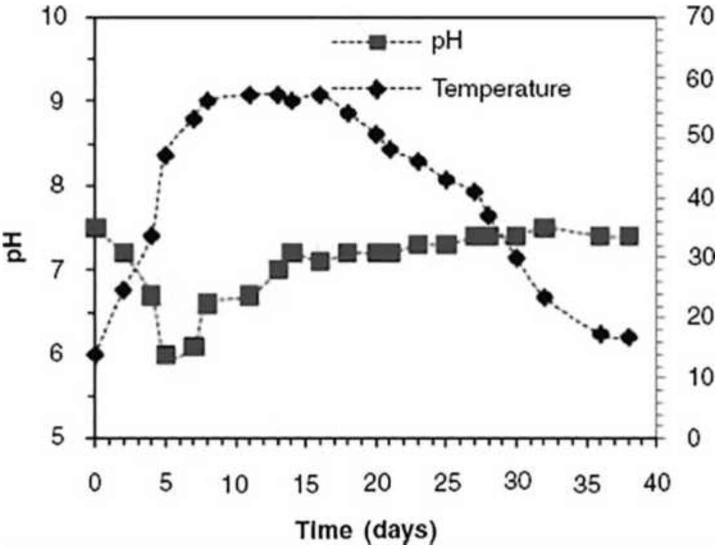


FIGURE 13 Typical temperature and pH range in windrow composting. (*Source:* Fonseca and Amorim.[59])

Evaluation of Maturation and Compost Maturity

As stated before, compost cannot be produced in record time; the length of the composting process depends on several factors, such as the feedstock material, particle size, nutrient balance, moisture content, and the composting technique. Although some authors and technologies claim to compost wastes in 7 or 20 days, this is not possible because compost produced in this time is not stabilized as total humification of the OM can only be achieved after about 110 days.[23] Thus, composting of garden or food waste materials can be successfully achieved in 3 mo under aerated, in-vessel, or even turned windrow systems but will take as long as 1 year in an unaerated static pile.

Compost maturation is fundamental in order to prevent ammonia liberation into the roots of cultures (at very low C/N ratios), biochemical reduction of soil nitrogen (when residual carbon in high C/N ratio compost is used by microorganisms), production of toxins that inhibit plant metabolism and seed germination, and microbiological activity that may lead to oxygen demand as well as other toxic effects to plants.[10]

The degree of maturation and stabilization of compost can be assessed using different methodologies: 1) final drop in temperature and self-heating capacity (the absence of temperature increase after turning and humidifying the biowaste up to 50% moisture content means that the process has ended); 2) amount of decomposable and resistant OM in the compost (lignin content higher than 30% means that compost is stable);[58] 3) chemical oxygen demand (COD) (a COD value below 350 mg O_2/g compost shows that the compost is stable); 4) oxygen uptake rate [it should be less than 40 mg of O_2 per kilogram of dry matter per hour for compost to be considered stable; a specific oxygen uptake rate (SOUR) test[61] can be performed]; 5) CO_2 production (it should be small, meaning that there is no significant microorganism activity); 6) C/N ratio (it usually decreases along the process, but it should be interpreted taking into account the initial characteristics of the OM and composting conditions. If OM has an initial C/N ratio of 35–40, the stabilized compost should have a C/N ratio of 18–20; however, if the initial C/N ratio is 10 or even lower as in manure wastes, an increase of the C/N ratio should be expected during the composting process);[17] 7) growth of the fungus *Chaetomium gracilis*;[58] and 8) the absence of ammonia in the compost and the presence of nitrates.[38]

Commercialization Issues

To be marketed, compost must have a uniform size; must be free of contaminants such as glass, plastics, and metals; and must be free from unpleasant odors. Typically, compost is a dark color product, has a wet soil smell, has a nice texture, and does not show the original raw material (Figure 14). The most common treatments before compost packaging are crushing and screening, although sometimes compost is commercialized as pellets.

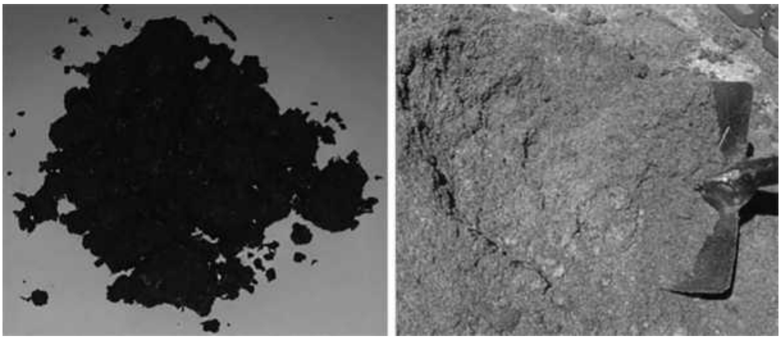


FIGURE 14 How compost looks like: (left) from OFMSW; (right) from co-composting of slaughterhouse with sawdust and agricultural waste.

Conclusion

Methods that can be used to perform composting over biowastes of different nature have been presented, from the simpler and less expensive ones, to some more sophisticated and expensive technologies. Most of the composting processes are based on reactions involving microorganisms, but worms can also be used to perform the same task of degrading OM and producing a stabilized and mature product (that is sterilized and of good quality) that can be used as a soil conditioner or fertilizer. With CO_2 and H_2O being the only by-products, composting can be considered environmentally benign and can contribute to OM recycling. The influence of the main operational parameters has been discussed, and it was shown that under appropriate conditions, almost any OM can be composted.

References

1. Evans, G. *Biowaste and Biological Waste Treatment*; James and James (Science Publishers) Ltd: London, 2001.
2. Gama, P.S. *Recolha Selectiva e Reciclagem de Resíduos Sólidos Urbanos.* Curso sobre Valorização e Tratamento de Resíduos. Prevenção, Recolha Selectiva, Compostagem e Confinamento em Aterro. LNEC/APESB, Lisbon, 10th–12th December, 1996.
3. Directive 2008/98/EC of the European Parliament and of the Council, of 19 November 2008 on waste and repealing certain Directives, Official Journal of the European Union, Brussels, 2008; L312/3-L312/30.
4. Directive 99/31/EC of the Council, of 26 April 1999 on the landfill of waste, Official Journal of the European Union, 1999; L182/1-L182/19.
5. Davis, M.L.; Masten, S.J. *Principles of Environmental Engineering and Science*, 1st Ed.; McGraw-Hill Companies, Inc.: New York, 2004.
6. Abu Qdais, H.A.; Hamoda, M.F.; Newham, J. Analysis of residential solid waste at generation sites. Waste Manage Res. **1997,** *15* (4), 395–405.
7. Francou, C.; Lineres, M.; Derenne, S.; Le Villio-Poitrenaud, M.; Houot, S. Influence of green waste, biowaste and paper–cardboard initial ratios on organic matter transformations during composting. Bioresour. Technol. **2008,** *99* (18), 8926–8934.
8. Lino, F.A.M.; Ismail, K.A.R. Energy and environmental potential of solid waste in Brazil. Energy Policy **2011,** *39* (6), 3496–3502.
9. Directorate-General Environment A.2, CEC. Working document on biological treatment of biowaste—2nd draft. European Commission, Brussels, 2001; 1–22, available at http://www.ymparisto.fi/download.asp?contentid=5765 (accessed June 8, 2011).
10. Neto, J.T.P. On the Treatment of Municipal Refuse and Sewage Sludge Using Aerated Static Pile Composting—A Low Technology Approach. Ph.D. Thesis: University of Leeds, 1987.
11. Haug, R.T. *The Practical Handbook of Compost Engineering*; Lewis Publishers: Boca Raton, FL, 1993.
12. Hogan, J.A.; Miller, F.C.; Finstein, M.S. Physical modelling of composting ecosystem. Appl. Environ. Microbiol. **1989,** *55* (5), 1082–1092.
13. Golueke, C.G. *Biological Reclamation of Solid Wastes;* Rodale Press: Emmaus, PA, 1977; 249 pp.
14. Bertoldi, M. Compost quality and standard specifications: European perspective. In *Science and Engineering of Composting: Design, Environmental, Microbiological and Utilization Aspects*; Hoitink, H.A.J., Keener, H.M., Eds.; The Ohio State University, Renaissance Publications: Ohio, 1993; 523–535.
15. Epstein, E.; Taylor, J.M.; Chaney, R.L. Effects of sewage sludge and sludge compost applied to soil on some physical and chemical properties. J. Environ. Qual. **1976,** *5*, 422–426.
16. Avnimelech, Y. Organic residues in modern agriculture. In *The Role of Organic Matter in Modern Agriculture*; Chen, Y., Avnimelech, Y., Eds.; Martinus Nijhoff Publishers: Netherlands, 1986; 1–10.

17. Brito, L.M. Taxas de Mineralização da Matéria Orgânica nos Resíduos Sólidos Urbanos: Efeitos Agronómicos e Ambientais. In Seminário sobre Produção de Correctivos Orgânicos a Partir de Resíduos Sólidos Urbanos—Sua Importância para a Agricultura Nacional, Exponor (Matosinhos), April 8, 1997.
18. Corbitt, R.A. *Standard Handbook of Environmental Engineering,* 2nd Ed.; McGraw-Hill Book Company: New York, 1998; 1248 pp.
19. Mays, D.A.; Giordano, P.M. Landspreading municipal waste compost. BioCycle **1989,** *30* (3), 37–39.
20. Howard, A.; Wad, Y.D. *The Waste Products of Agriculture: Their Utilisation as Humus*; Oxford University Press: London, 1931.
21. Golueke, C.G. Composting refuse at Sacramento, California. Compost Sci. **1960,** *1* (3), 12–15.
22. McGauhey, P.H. Refuse composting plant at Norman, Oklahoma. Compost Sci. **1960,** *1* (3), 5–8.
23. Neto, J.T.P. Aspectos Tecnológicos da produção e Quali-dade do Composto Orgânico a partir da Compostagem de Lixo Urbano. In Seminário sobre Produção de Correctivos Orgânicos a Partir de Resíduos Sólidos Urbanos—Sua Importância para a Agricultura Nacional, Exponor (Matosinhos), April 8, 1997.
24. Lu, Y.; Wu, X; Guo, J. Characteristics of municipal solid waste and sewage sludge co-composting. Waste Manage. **2009,** *29* (3), 1152–1157.
25. Black, R.J. The solid waste problem in metropolitan areas. California Vector Views **1964,** *11* (9), 51.
26. Diaz, L.F. Combining experience with common sense. BioCycle **1982,** *30* (10), 48–49.
27. Golueke, C.G.; Diaz, L.F. Historical review of composting and its role in municipal waste management. In Proceedings of the Science of Composting; Bertoldi, M., Sequi, P., Lemmes, B., Papi, T., Eds.; Chapman & Hall: Glasgow, 1996; 3–14.
28. Ferrero, G., L'Hermite, P. Composting—Progress to date in the European Economic Community and prospects for the future. In *Composting of Agriculture and other Wastes;* Gasser, J.K., Ed.; Elsevier Applied Science Pub.: Amsterdam, 1985; 3–10.
29. Lopez-Real, J.M. Composting of agricultural wastes. In Proceedings of the Science of Composting; Bertoldi, M., Sequi, P., Lemmes, B., Papi, T., Eds.; Chapman & Hall: Glasgow, 1996; 542–550.
30. Mesquita, M.M.F. Compostagem de material sólido—Principais vantagens e aspectos mais problemáticos. In Seminário para apresentação e discussão do Plano de adaptação à legislação ambiental pelo sector da suinicultura. LNEC e Federação Portuguesa das Associações de Suinicultores: LNEC, Lisbon, Jan. 22–23, 1996.
31. Hamoda, M.F.; Abu Qdais, H.A.; Newham, J. Evaluation of municipal solid waste composting kinetics. Resour., Conserv. Recycl. **1998,** *23* (4), 209–223.
32. John Paul, J.A.; Karmegam, N.; Daniel, T. Municipal solid waste (MSW) vermicomposting with an epigeic earthworm, *Perionyx ceylanensis* Mich. Bioresour. Technol. **2008,** *102* (12), 6769–6773.
33. Elango, D.; Thinakaran, N.; Panneerselvam, P.; Sivanesan, S. Thermophilic composting of municipal solid waste. Appl. Energy **2009,** *86* (5), 663–668.
34. Tintner, J.; Smidt, E.; Böhm, K.; Binner, E. Investigations of biological processes in Austrian MBT plants. Waste Manage. **2010,** *30* (10), 1903–1907.
35. Sims, J.T.; Murphy, D.W.; Handwerker, T.S. Composting of poultry wastes—Implications for dead poultry disposal and manure management. J. Sustainable Agric. **1992,** *2* (4), 67–82.
36. Mesquita, M.M.F. Aplicações Inovadoras da Compostagem e Especificações Técnicas para Avaliação da Qualidade do Composto. In Curso sobre Valorização e Tratamento de Resíduos. Prevenção, Recolha Selectiva, Compostagem e Confinamento em Aterro. LNEC/APESB, Lisbon, Dec 10–12, 1996.
37. Williams, P.T. *Waste Treatment and Disposal,* 2nd Ed.; John Wiley & Sons, Ltd.: Chichester, West Sussex, 2005.
38. Neto, J.T.P.; Mesquita, M.M.F. *Compostagem de Resíduos Sólidos Urbanos. Aspectos Teóricos, Operacionais e Epidemiológicos*; LNEC: Lisbon, 1992.
39. de Bertoldi, M.; Zucconi, F. Composting of organic residues. In *Bioenvironmental Systems*; Wise, D.L., Ed.; CRC Press: Boca Raton, Florida, 1987; Vol. III.

40. Loehr, R.C. *Pollution Control for Agriculture,* 2nd Ed.; Academic Press: Orlando, Florida, 1984.
41. Neto, J.P.; Stentiford, E.I. The main process constraints in composting. In Proceedings of the First Italian-Brazilian Symposium of Sanitary Engineering, Rio de Janeiro, Brazil, Mar 29–Apr 3, 1992.
42. Rynk, R.; van de Kamp, M.; Willson, G.B; Singley, M.E.; Richard, T.L.; Kolega, J.J.; Gouin, F.R.; Laliberty, Jr., L.; Kay, D.; Murphy, D.W.; Hoitink, H.A.J.; Brinton, W.F. *On Farm Composting.* Northeast Regional Agricultural Engineering Service. Cooperative extension. NRAES-54, Rynk, R., Ed., NRAES: Ithaca, New York, 1992, available at http://infohouse.p2ric.org/ref/24/23702.pdf (accessed May 20, 2012).
43. Lipor, Terra-à-Terra Project, 2007–2011, available at http://www.hortadaformiga.com/gb/conteudos.cfm?ss=8 (accessed August 2011).
44. Richard, T.L. MSW Composting: Biological Processing. Cornell Department of Agricultural and Biological Engineering, New York State College of Agriculture and Life Sciences: Cornell University, Ithaca, 1998, available at http://www.cals.cornell.edu/dept/compost/MSW.FactSheets/msw.fs2.html (accessed August 2011).
45. Risse, M.; Faucette, B. Food Waste Composting. Institutional and Industrial Applications. University of Georgia College of Agricultural and Environmental Sciences, 2000, available at http://www.caes.uga.edu/applications/publications/files/pdf/B%201189_2.PDF (accessed August 2011).
46. Graves, R.E.; Hattemer, G.M. Chapter 2—Composting, In *National Engineering Handbook,* Part 637—Environmental Engineering, United States Department of Agriculture, Natural Resources Conservation Service, 2010.
47. Misra, R.V.; Roy, R.N.; Hiraoka, H. *On-Farm Composting Methods*; Food and Agriculture Organization of the United Nations: Rome, 2003.
48. Diaz, L.F., Savage, G.M., Eggerth, L.L.; Golueke, C.G. *Composting and Recycling Municipal Solid Waste*; Lewis Publishers, CRC Press: Boca Raton, Florida, 1993.
49. Diaz, L.F.; Savage, G.M.; Eggerth, L.L. Solid Waste Management, United Nations Environment Programme, UNEP IETC and CalRecovery Inc., 2005, available at http://www.unep.or.jp/ietc/Publications/spc/Solid_Waste_Management/ (accessed August 2011).
50. Arancon, N.Q.; Edwards, C.A. The use of earthworms in the breakdown of organic wastes to produce vermicomposts and animal feed protein. In *Earthworm Ecology*; Edwards, C.A. Ed.; CRC Press: Boca Raton, Florida, 2004; 345–379.
51. Munroe, G. *Manual of On-Farm Vermicomposting and Vermiculture;* Organic Agriculture Centre of Canada, 2007.
52. Sinha, R.K.; Agarwal, S.; Chauhan, K.; Soni, B.K. Vermiculture technology: Reviving the dreams of Sir Charles Darwin for scientific use of earthworms in sustainable development programs. J. Technol. Investment **2010,** *1* (3), 155–172.
53. Sinha, R.K, Herat, S., Valani, D.; Chauhan, K. Earth-worms—The environmental engineers: Review of vermiculture technologies for environmental management and resource development. Int. J. Global Environ. Issues **2010,** *10* (3/4), 265–292.
54. Bogdanov, P. Commercial Vermiculture: How to Build a Thriving Business in Redworms; VermiCoPress: Oregon, 1996.
55. Atiyeh, R.M.; Subler, S.; Edwards, C.A.; Bachman, G.; Metzger, J.D.; Shuster, W. Effects of vermicomposts and composts on plant growth in horticulture container media and soil. Pedobiologia **2000,** *44,* 579–590.
56. Arancon, N.Q.; Edwards, C.A. Effects of vermicomposts on plant growth. In Proceedings of the Vermi-Technologies Symposium for Developing Countries, Department of Science and Technology—Philippine Council for Aquatic and Marine Research and Development, Los Banos, Philippines, 2006, http://www.slocountyworms.com/wp-content/uploads/2010/12/EFFECTS-OF-ERMICOMPOSTS-ON-PLANT-GROWTH.pdf (accessed August 2011).

57. Arancon, N.A.; Edwards, C.A.; Babenko, A.; Cannon, J.; Galvis, P.; Metzger, J.D. Influences of vermicomposts, produced by earthworms and microorganisms from cattle manure, food waste and paper waste, on the germination, growth and flowering of petunias in the greenhouse. Appl. Soil Ecol. **2008,** *39* (1), 91–99.
58. Tchobanoglous, G.; Theisen, H.; Vigil, S. *Integrated Solid Waste Management. Engineering Principles and Management Issues*; McGraw-Hill Book Company: Singapore, 1993.
59. Fonseca, E.; Amorim, J. Quantificação dos Resíduos Sólidos Gerados no ISEP: Projecto e Operação de Uma Unidade de Compostagem Para Tratamento da Fracção Biodegradável, *CESE em Engenharia Química—Tecnologias de Protecção Ambiental/Projecto*, ISEP/IPP: Porto, 1999.
60. Cross, F.; O'Leary, P.; Walsh, O. Operation and maintenance considerations for waste-to-energy systems. Waste Age **1987,** *18* (8).
61. Lasaradi, K.E.; Stentiford, E.I. A simple respirometric technique for assessing compost stability. Water Res. **1998,** *32* (12), 3717–3723.

40

Insects and Mites: Biological Control

Ann E. Hajek

Introduction

Biological control is defined as the use of natural enemies to suppress a pest population, making the pests and their associated damage less abundant. Natural enemies were first used to control insect pests when farmers in ancient China and Yemen moved colonies of predaceous ants to control pests of tree crops. Today, the natural enemies used to control insect and mite pests include a diversity of predators, parasitoids, and pathogens. Specific strategies have been developed for release of natural enemies or enhancement of their persistence and activity. Biological control has been used very successfully for permanent suppression of introduced pests. Among natural enemies applied for shorter-term control in 1990, even the most widely used biological control agent, *Bacillus thuringiensis*, accounted for <1% of the insecticide market. However, biological control agents are widely used for control in environmentally sensitive areas or controlled environments and constitute important components of integrated pest management programs.

Strategies for Using Biological Control

Natural enemies can be used in a variety of very different ways. The first major uses of natural enemies for pest control were directed at control of introduced insect pests. Natural enemies from the land of origin of introduced pests were released in areas of pest introduction. This strategy, called classical biological control, now also includes introduction of exotic natural enemies to control native pests. In all cases, a high degree of host specificity is required in the natural enemies to be introduced. After the exotic natural enemy is established in the new location, its effectiveness is based on population increases in response to increasing densities of pest populations. Classical biological control can be dramatically effective, with 34% of insect natural enemies that are released becoming established and 17% completely controlling devastating pests. Classical biological control is known to be extremely cost effective with cost benefit estimates of up to 200:1, if a program is successful at establishing an effective natural enemy.

A second strategy, augmentation, involves releasing natural enemies for pest control, usually in instances where natural enemies can be effective but are not sustained in the environment at high enough densities to provide control. Inundative augmentation is used when only the natural enemies that are released in high numbers are expected to exert control. Under inoculative augmentation, control effects are more delayed and are predominantly exerted by the progeny of the released organisms. Natural enemies used for augmentation are often mass reared, so understanding requirements for mass production of high quality natural enemies that are healthy and active after shipment and release is critical for the use of this strategy.

The third major strategy, conservation, involves manipulations to enhance the persistence and activity of natural enemies already occurring in the environment. This strategy takes on a diversity of forms based on requirements of the individual natural enemies. To cause less mortality of natural enemies, use of synthetic chemical pesticides that kill natural enemies can be altered in different ways ranging from eliminating their use to selecting pesticides with less impact on natural enemies to timing pesticide applications to minimize the effect on natural enemies. Alternatively, natural enemy populations can be increased by maintaining or improving the environment to provide ideal conditions. For example, irrigating, strip-harvesting, intercropping, retaining vegetation adjacent to crops, and planting cover crops all have been shown to provide favorable habitats and food to maintain or increase populations of natural enemies. In a program to control the brown plant- hopper on rice in south and southeast Asia, the activity of a suite of native natural enemies, aided by host plant resistance and application of insecticides only when absolutely necessary, provided better control than pesticides alone.

Types of Natural Enemies

Predators

Predators are generally larger than their prey and each usually consumes several prey individuals either for growth of immatures or for subsistence and reproduction of adults. The predatory life style is very common among insects and mites but predators with the most importance for biological control belong to four insect orders (Hemiptera, Coleoptera, Diptera, and Hymenoptera) and eight mite families (Figure 1). Predators feed on a diversity of prey life stages, from eggs to adults, and display a range of host specificity, from feeding only on one prey species to generalized feeding on many prey species. One of the most famous examples of classical biological control is the introduction of the highly host-specific Vedalia beetle that was imported from Australia and released against outbreak populations of the introduced cottony cushion scale threatening the southern California citrus industry. After the 1888–1889 releases of this predator, cottony cushion scale populations decreased dramatically and, by 1890, scale populations had been decimated. The immense success of this early program was instrumental in building interest in use of natural enemies for biological control. As a second example, in more recent years, phytoseiid mites attacking tetranychid spider mites have been developed for mass release in greenhouses or on some outdoor crops. Pesticide-resistant mite strains have also been developed for use against spider mites attacking tree crops.

Parasitoids

Parasitoids develop at the expense of a single host and usually kill their hosts. Parasitoids have been used extensively for biological control because, in contrast to predators, the impressive degree of host specificity often characteristic of parasitoids leads to sensitive responses to changes in host density. Parasitoids used for biological control are predominantly in the Order Hymenoptera (Figure 2) with less common use of Diptera. The immature parasitoid is usually a featureless larva associated with the host while winged adults disperse to mate and find new hosts. To enable their close association with hosts, parasitoids have adopted amazing and diverse life cycles. Different species of parasitoids attack different life stages of hosts (egg through adult) and can develop either externally on hosts or internally within

FIGURE 1 An adult of the multicolored Asian lady beetle (Harmonia axyridis), which was introduced from Asia for control of aphids and scales (length ca. 1 cm). Both larvae and adults are predatory. (Photo by J. Ogrodnick.)

FIGURE 2 An adult of the parasitoid Muscidifurax raptor (length ca. 2 mm) parasitizing house flies. Females lay eggs in fly puparia, larvae grow while consuming the fly pupae and winged adults then emerge to mate and find more hosts. These flies can be purchased and released to augment naturally occurring populations. (Photo by S. Long.)

hosts. One to many parasitoid individuals of one or more species can develop within a host. Parasitoids have been more widely used for classical biological control than either predators or pathogens. In recent years, the tiny wasp Epidinicarsis lopezi was released by land and air in 34 countries in Africa to control the introduced cassava mealybug. Due to the activity of this wasp, cassava mealybug is no longer considered a problem, saving African farmers hundreds of millions of dollars in reduced crop losses. Some parasitoids widely used for augmentative biological control are tiny species of Trichogramma attacking eggs of Lepidoptera and Encarsia formosa, a member of the Aphelinidae that attacks whiteflies. Both of these tiny wasps are mass produced in insectaries and shipped to users for release against pest populations threatening crops.

FIGURE 3 Gypsy moth (Lymantria dispar) larvae killed by the entomopathogenic fungus Paecilomyces farinosus (each larva ca. 3 cm in length). When infecting, this fungus penetrates externally through the larval cuticle, then increases within the host and, after host death, grows out through the integument to produce spores that will infect healthy hosts. (Photo by T. Ebaugh.)

Pathogens

Microorganisms that are parasitic, referred to as pathogens, are masters at exploiting insect and mite hosts (Figure 3). Pathogens important for biological control include a diversity of viruses, bacteria, fungi, protozoa, and nematodes. This range of types of pathogens exhibits a comparable medley of diverse interactions with their hosts. Of primary importance, viruses, bacteria, and most protozoa must be ingested by hosts in order to infect, while fungi and some protozoa can penetrate directly through the host cuticle. The nematodes of greatest importance to biological control, Steinernema and Heterorhabditis, can enter hosts through body openings although some possess the ability to penetrate directly through the cuticle. While some pathogens have mechanisms for active dispersal to find new hosts, host finding is generally not directed and these pathogens rely principally on their production of huge numbers of progeny in order to be assured of locating healthy hosts. Associations between pathogens and hosts range from facultative to obligate but pathogens important for biological control are all specialized for infecting only insects and mites. Pathogens have been used for classical biological control relatively infrequently although some programs have provided complete control. Much of the development of pathogens has been directed toward inundative augmentation of mass- produced microbes. The bacterium *Bacillus thuringiensis* is applied more than any other biological control agent. Strains of this bacterium predominantly kill Lepidoptera, Diptera, and Coleoptera through the activity of a toxin destroying the integrity of the gut. For many years, this bacterium was applied principally as a spray but recently several crop plants have been engineered to express genes encoding the toxin.

Biological Control in Practice

There is great demand for use of biological control programs to eliminate insect and mite pests, especially in environmentally sensitive areas and areas where humans live. Use of natural enemies to control pests can be highly effective due to the diversity of types of natural enemies and approaches. However, because biological control involves management of living organisms, it can be somewhat unpredictable. Therefore, biological control programs generally are tailored to specific pest systems and to optimize control, often require knowledge of the biology and ecology of the insect or mite host. Biological control has proven to be most effective under certain conditions (Table 1), although these generalities should not prevent investigations of use of biological control for alternative situations.

TABLE 1 Characteristics of Systems and Conditions More Commonly Associated with Successful Biological Control

Highly efficient natural enemy
Less mobile pest living in an exposed location
Perennial crop, natural habitat, or controlled environment, for example, a greenhouse
Crop or environment where some pest damage is tolerated
Controls for other pests do not interfere with the activity of natural enemies

Introduced pests are not insignificant, comprising 39% of the 600 major arthropods pests in the United States. By 1990, more than 4300 introductions of exotic parasitoids and predators had been made to control insect pests, many of which were introduced. Due to its low cost and permanent effectiveness when successful, classical biological control continues as one of the first control strategies to be investigated after a new pest has been introduced.

Augmentation and conservation are now often employed as important parts of integrated pest management programs. Although since the late 1940s (the start of the DDT era), for most pests synthetic chemical pesticides have been the first control strategy considered, use of natural enemies for control is increasing, especially for specific applications and systems.

Bibliography

1. Barbosa, P., Ed.; *Conservation Biological Control*; Academic Press: San Diego, 1998; 396.
2. Bellows, T.S.; Fisher, T.W.; Caltagirone, L.E.; Dahlsten, D.L.; Gordh, G.; Huffaker, C.B., Eds.; *Handbook of Biological Control: Principles and Applications of Biological Control*; Academic Press: San Diego, 1999; 1046.
3. Evans, H.F., Ed.; *Microbial Insecticides: Novelty or Necessity?*; British Crop Protection Council: Surrey, U.K., 1997; 301.
4. Flint, M.L.; Dreistadt, S.H. *Natural Enemies Handbook: An Illustrated Guide to Biological Pest Control*; University of California Press: Berkeley, CA, 1998; 154.
5. Greathead, D.; Waage, J., Eds.; *Insect Parasitoids*; Academic Press: London, 1986; 389.
6. Lacey, L.A.; Goettel, M.S. Current developments in microbial control of insect pests and prospects for the early 21st century. Entomophaga **1995**, 40 (1), 3–27.
7. Lynch, J.M.; Hokkanen, H.M.T., Eds.; *Biological Control. Benefits and Risks*, Cambridge University Press: Cambridge, U.K., 1995; 304.
8. New, T.R. *Insects As Predators*; New South Wales University Press: Kensington, Australia, 1991; 178.
9. Van Driesche, R.G.; Bellows, T.S., Jr. *Biological Control*; Chapman and Hall: New York, 1996; 539.

41

Invasion Biology

Jennifer Ruesink

Introduction

Biological invasions occur when species or distinct populations breach biogeographical barriers and extend their ranges to areas where they were not historically present.[1–3] Invasion biology concerns the causes and consequences of these new species, which are also referred to as invasive, introduced, alien, exotic, nonnative, or nonindigenous species.

Biological invasions occur in five steps: arrival, establishment, population growth, population spread, and impact (Table 1). Only a small proportion of species that arrive actually establish, and so forth; thus, each step acts as a filter for the invasion process. Successful invasion depends on characteristics of the invading species, recipient environment, and process by which the two are brought together.[4] However, no widely accepted method currently exists for identifying the characteristics that promote invasion a priori.[5]

TABLE 1 Reasons for and Responses to Five Steps of an Invasion

Step	Why?	Control
Arrival	Intentional releases, intentional imports, unintended hitchhiking (by product)	Risk assessment—choice of species, treatment, or quarantine of vectors
Establishment	Suitable abiotic conditions, available resources, propagule pressure	Reduce pathways, containment
Population growth	Intrinsic rate of reproduction, apomixis/vegetative reproduction	Early eradication
Population spread	Mobility—dispersal, home range size, transport by humans	Eradicate new populations, reduce human transport
Impact	Density (abundant resources), few enemies, per capita effect (new role), alteration of resource base	Effective screening prior to introduction; mechanical, chemical, or biological control; make environment less suitable

Why Do New Species Invade?

Species have always expanded their ranges, but the pace of invasion has accelerated recently due to increased human travel and trade.[5–7] Humans transport species in three ways: (1) on purpose, with the intention that they will grow in outdoor environments (fish and game, plantation trees, and biological control agents), (2) on purpose but with no intention that they will establish (pets, horticultural and agricultural plants, aquaculture, and sterile releases), and (3) accidentally (hitchhiking on packages, live imports, and people). The contribution of each of these main pathways varies among taxa. Of South Africa's weeds, 89% were intentionally introduced[8] but only 11% of insect invaders in North America were intentional.[9] Ducks, pheasants, pigeons, finches, and parrots have more introduced species worldwide than would be expected by chance because these bird families include many pet or game species.[10]

Which Species Invade?

Propagule Pressure

Species can invade a new area if abiotic (especially climatic) conditions are suitable and an exploitable resource exists. Those species that can invade, will invade if given sufficient opportunity.[11–13] High rates of arrival have been termed "propagule pressure" and can occur either through numerous releases or releases of many individuals. Some of the best evidence that propagule pressure affects invasion come from compilations of biocontrol introductions: the successful establishment of insect predators rose seven times when the number of introductions doubled, and releases >31,200 individuals were eight times more successful than those of <5000 individuals.[14]

Species Traits

For a given propagule pressure, some species may be more likely to invade than others. Traits promoting invasion could include the ability to increase rapidly from low density, a generalist diet, and broad climatic tolerance. Although statistical relationships between species traits and invasibility are often weak,[15] analyses of certain taxa introduced to particular environments have been successful.[16,17] For instance, invasive species of pines in South Africa tend to have small seeds, short intervals between reproductive bouts, and short times to maturity, whereas noninvasive species show the opposite traits.[18] For woody plants, those that have become invasive in North America often reproduce vegetatively, germinate easily, and have a long fruiting period.[19]

Where Do Species Invade?

In addition to species traits, characteristics of the recipient community may also influence the ease with which new species invade. Indeed, particular conditions in native environments (e.g., disturbance and species interactions) may select for species that are effective invaders of other areas, thus resulting in asymmetric patterns of invasion. For instance, European insects have invaded forests in North America but not vice versa,[20] and there has been a unidirectional appearance of tillering grasses in bunchgrass habitats worldwide.[21] Assemblages may be more resistant to invasion if they are undisturbed[22,23] or contain many natural enemies.[24,25] For many years, it was a rule of thumb that species-poor islands were invasible and species-rich tropics were not.[1] However, surveys of plant assemblages indicate that areas of naturally high species richness tend also to have numerous invaders, perhaps because soil and climate conditions are generally conducive to plant growth.[26] Although species richness per se may not influence invasion, loss of species could make systems easier to invade.

Habitat alteration (flooding, drought, fire, wind, eutrophication, and channeling) creates new conditions that are suitable for a new suite of species. These species are often introduced. For instance, western North America hosts many conspicuous invasive fishes in part because once fast-flowing rivers are now lakes separated by dams.[27] Roadsides in North America contain a high proportion of European plant species, probably because European plants have had millenia to evolve to take advantage of disturbance, whereas disturbance in North American habitats has risen recently.[28] Based on an Australian study of two invaders, both physical disturbance and nutrient addition improve the performance of invasive plants.[29]

Which Species Have Effects and Where?

Impacts of invaders relevant to pest management include the ecological and/or economic damage from pests and the effectiveness of biocontrol agents.[30] Impacts are expected to be particularly pronounced when species reach high abundances or have high per capita effects.[31,32] High abundances can occur when species escape limits to population growth (abundant resources or few enemies). High per capita effects may arise when a species plays a new role, especially by altering the resource base. For instance, many of the worst plant invaders of natural areas are nitrogen fixers (which alter nutrients) or climbing vines (which alter light).[33]

A disproportionate number of invasive species have harmful effects, more so if introduced accidentally rather than intentionally. In Japan, for instance, 8% of native insects are considered pests, but 72% of introduced insects are pests.[34] Of agricultural weeds in North America, 50–75% are nonindigenous.[6] On the other hand, only 1–6% of nonindigenous plant species in Great Britain have become weedy or widespread.[8] Some introduced species may be problematic due to an absence of natural enemies, but this escape from control does not appear to be entirely general. Based on a compilation of life tables for 124 holometabolous insects, mortality rates due to parasitoids, predators, and diseases do not differ for native and introduced species.[35]

Species introduced accidentally often have harmful effects, and, conversely, species introduced intentionally have beneficial effects less often than desired. Of 463 grasses and legumes introduced to Australia to improve pasture, only 5% have raised productivity.[36] About a third of established biocontrol insects actually reduce the target organism.[37] Although impact in the native environment is not necessarily a useful indicator of what a species will do once introduced, one indicator of potential impact is the fate of prior introductions.[19]

Pest status tends to be based on economic considerations, but introduced species also cause ecological damage. Invasive species contribute to endangerment of nearly 50% of species listed under the United States Endangered Species Act,[38,39] and they have dramatically altered the structure and function of ecosystems.[40,41] Ecological and economic effects have the same root causes—abundance and high per capita effects of invaders—but the affected habitats can be quite distinct. Only 25% of plants that cause problems in natural areas are also agricultural weeds.[33] Thus, screening procedures that keep out economic pests would fail to restrict many species that cause ecological harm.

Time Lags and Surprises

Introduced species have occasionally surprised researchers by expanding to previously intolerable places or irrupting after remaining localized and rare for many years. “Boom and bust” patterns have also been observed, in which an invader initially reaches high abundance and then declines, sometimes even going extinct. Tolerance of new conditions (e.g., temperature or host plants) may require genetic adaptation, which could result in time lags before invasion.[42] Of 184 woody species currently considered invasive near Brandenburg, Germany, 51% did not appear to be invasive until >200 years after their initial introduction.[43] However, only 7% of 627 cases involving biocontrol introductions showed time lags before population increase, whereas 28% increased and 27% went extinct immediately.[44]

Regardless of frequency, cases in which invaders have unexpected impacts have become well known, especially for biocontrol agents with nontarget effects such as feeding on endangered species (e.g., the weevil *Rhinocyllus conicus* on thistles and the moth *Cactoblastis* on *Opuntia* cacti)[45–47] or competing with natives (e.g., the ladybird beetle *Coccinella septempunctata*)[5]

Can Invasions Be Controlled?

The first steps of an invasion can be controlled by limiting entry or vigilantly eliminating newly established populations of invaders. For instance, an assessment of species associated with raw logs indicated a potential loss of billions of dollars due to forest pests if Siberian larch was not treated prior to import into the United States.[48] For many years, medflies (*Ceratitis capitata*) have epitomized the notion that "an ounce of prevention is worth a pound of cure." California spent $100 million to eradicate an incipient invasion in 1981, thereby preventing nine times that amount of crop damage. By 1996, however, medflies were apparently established and spreading. The extent of the invasion makes eradication unlikely.[49] Humans simply have to learn to live with these naturalized species.[50]

Control during the last steps of an invasion usually involves chemical, biological, or mechanical reduction of unwanted species in areas where effects are most serious. Control efforts at this stage would benefit from considerations of demography and behavior of invaders. For instance, seedling competition among annual plants is often fierce, so efforts to reduce seed production will not reduce plant numbers. Instead, control efforts should be directed at reducing seedling growth and survival.[51] Knowing how insects move among microhabitats could aid in trap placement or crafting habitats that promote desirable species and discourage undesirable ones.[52] Knowing encounter and feeding rates could aid in calculating the number of consumers necessary for effective biological control.

Species invasions are a form of ecological gambling in which the consequences of any particular introduction are uncertain, despite an emerging framework of factors contributing to high risk invasions. The influx of new species can be slowed by reducing pathways for introduction and intentionally introducing species only when beneficial effects will be large and native alternatives do not exist.[53] Distinct biotas are valuable and intriguing but increasingly difficult to maintain under pressures of globalization.

References

1. Elton, C.S. *The Ecology of Invasions by Animals and Plants*; Chapman and Hall: London, 1985.
2. Williamson, M. *Biological Invasions*; Chapman and Hall: London, 1996.
3. Mack, R.N.; Simberloff, D.; Lonsdale, W.M.; Evans, H.; Clout, M.; Bazzaz, F.A. Biotic invasions: causes, epidemiology, global consequences, and control. *Ecol. Appl.* **2000**, *10* (3), 689–710.
4. Lodge, D.M. Biological invasions: lessons for ecology. *Tr. Ecol. Evol.* **1993**, *8*, 133–137.
5. Ruesink, J.L.; Parker, I.M.; Groom, M.J.; Kareiva, P.M. Reducing the risks of nonindigenous species introductions: guilty until proven innocent. *Bioscience* **1995**, *45* (7), 465–477.
6. Office of Technology Assessment. *Harmful Non-Indigenous Species in the United States*; U.S. Government Printing Office: Washington, DC, 1993.
7. Cohen, A.N.; Carlton, J.T. Accelerating invasion rate in a highly invaded estuary. *Science* **1998**, *279*, 555–558.
8. Crawley, M.J.; Harvey, P.H.; Purvis, A. Comparative ecology of the native and alien floras of the British isles. *Phil. Trans. R. Soc. Lond. B.* **1996**, *351*, 1251–1259.
9. Sailer, R.I. History of insect introductions. In *Exotic Plant Pests and North American Agriculture*; Wilson, C.L., Graham, C.L., Eds.; Academic Press: New York, 1983; 15–38.
10. Lockwood, J.L. Using taxonomy to predict success among introduced avifauna: relative importance of transport and establishment. *Conserv. Biol.* **1999**, *13* (3), 560–567.

11. Veltman, C.J.; Nee, S.; Crawley, M.J. Correlates of introduction success in exotic New Zealand birds. *Amer. Nat.* **1996**, *147* (3), 542–557.
12. Green, R.E. The influence of numbers released on the outcome of attempts to introduce exotic bird species to New Zealand. *J. Anim. Ecol.* **1997**, *66*, 25–35.
13. Duncan, D.P. The role of competition and introduction effort in the success of passeriform birds introduced to New Zealand. *Amer. Nat.* **1997**, *149* (5), 903–915.
14. Beirne, B. Biological control attempts by introductions against pest insects in the field in Canada. *Can. Entomol.* **1975**, *107*, 225–236.
15. Goodwin, B.J.; McAllister, A.J.; Fahrig, L. Predicting invasiveness of plant species based on biological information. *Conserv. Biol.* **1999**, *13* (2), 422–426.
16. Panetta, F.D. A system of assessing proposed plant introductions for weed potential. *Plant Prot. Quarterly* **1993**, *8*, 10–14.
17. White, P.S.; Schwarz, A.E. Where do we go from here? The challenges of risk assessment for invasive plants. *Weed Technol.* **1998**, *12* (4), 744–751.
18. Rejmanek, M.; Richardson, D.M. What attributes make some plant species more invasive. *Ecology* **1996**, *77* (6), 1655–1661.
19. Reichard, S.H.; Hamilton, C.W. Predicting invasions of woody plants introduced to North America. *Conserv. Biol.* **1997**, *11* (1), 193–203.
20. Niemela, P.; Mattson, W.J. Invasions of North American forests by European phytophagous insects. *Bioscience* **1996**, *46* (10), 741–753.
21. Mack, R.N. Alien plant invasion into the intermountain west. A case history. In *Ecology of Biological Invasions of North America and Hawaii*; Mooney, H.A., Drake, J.A., Eds.; Springer-Verlag: New York, 1986; 191–213.
22. Orians, G.H. Site characteristics favoring invasions. In *Ecology of Biological Invasions of North America and Hawaii*; Mooney, H.A., Drake, J.A., Eds.; Springer-Verlag: New York, 1986; 133–148.
23. Hobbs, R.J.; Huenneke, L.F. Disturbance, diversity, and invasion: implications for conservation. *Conserv. Biol.* **1992**, *6*, 324–337.
24. Goeden, R.D.; Louda, S.M. Biotic interference with insects imported for weed control. *Annu. Rev. Entomol.* **1976**, *21*, 325–342.
25. Mack, R.N. Predicting the identity and fate of plant invaders: emergent and emerging approaches. *Biol. Conserv.* **1996**, *78*, 107–121.
26. Stohlgren, T.J.; Binkley, D.; Chong, G.W.; Kalkhan, M.A.; Schell, L.D.; Bull, K.A.; Otsuki, Y.; Newman, G.; Bashkin, M.; Son, Y. Exotic plant species invade hot spots of native plant diversity. *Ecol. Monogr.* **1999**, *69* (1), 25–46.
27. Moyle, P.B.; Light, T. Fish invasions in California: do abiotic factors determine success? *Ecology* **1996**, *77* (6), 1666–1670.
28. Pysek, P. Is there a taxonomic pattern to plant invasions? *Oikos* **1998**, *82*, 282–294.
29. Hobbs, R.J. The nature and effects of disturbance relative to invasions. In *Biological Invasions: A Global Perspective*; Drake, J.A., Mooney, H.A., di Castri, R., Kruger, F., Groves, R., Rejmanek, M., Williamson, M., Eds.; John Wiley and Sons: Chichester, 1986; 389–405.
30. Pimentel, D.; Lach, L.; Zuniga, R.; Morrison, D. Environmental and economic costs of nonindigenous species in the United States. *Bioscience* **2000**, *50*, 53–65.
31. Parker, I.M.; Simberloff, D.; Lonsdale, W.M.; Goodell, K.; Wonham, M.; Kareiva, P.M.; Williamson, M.H.; Von Holle, B.; Moyle, P.B.; Byers, J.E.; Goldwasser, L. Impact: toward a framework for understanding the ecological effects of invaders. *Biological Invasions* **1999**, *1* (1), 3–19.
32. Ricciardi, A.; Hoopes, M.F.; Marchetti, M.P.; Lockwood, J.L. Progress toward understanding the ecological impacts of nonnative species. *Ecol. Monogr.* **2013**, *83*, 263–282.
33. Daehler, C.C. The taxonomic distribution of invasive angiosperm plants: ecological insights and comparisons to agricultural weeds. *Biol. Conserv.* **1998**, *84*, 167–180.

34. Morimoto, N.; Kiritani, K. Fauna of exotic insects in Japan. *Bull. Nat. Inst. AgroEnvironmental Sci.* **1995**, *12*, 87–120.
35. Hawkins, B.A.; Cornell, H.V.; Hochberg, M.E. Predators, parasitoids, and pathogens as mortality agents in phytophagous insect populations. *Ecology* **1997**, *78* (7), 2145–2152.
36. Lonsdale, W.M. Inviting trouble: introduced pasture species in Northern Australia. *Australian J. Ecol.* **1994**, *19* (3), 345–354.
37. Williamson, M.; Fitter, A. The varying success of invaders. *Ecology* **1996**, *77* (6), 1661–1666.
38. Foin, T.C.; Riley, S.P.D.; Pawley, A.L.; Ayres, D.R.; Carlsen, T.M.; Hodum, P.J.; Switzer, P.V. Improving recovery planning for threatened and endangered species. *Bioscience* **1998**, *48*, 177–184.
39. Wilcove, D.S.; Rothstein, D.; Dubow, J.; Phillips, A.; Losos, E. Quantifying threats to imperiled species in the United States. *Bioscience* **1998**, *48*, 607–615.
40. Vitousek, P.M.; D'Antonio, C.M.; Loope, L.L.; Westbrooks, R. Biological invasions as global environmental change. *Am. Scientist* **1996**, *84* (5), 468–478.
41. Mack, M.C.; D'Antonio, C.M. Impacts of biological invasions on disturbance regimes. *Tr. Ecol. Evol.* **1998**, *13* (5), 195–198.
42. Secord, D.; Kareiva, P. Perils and pitfalls in the host specificity paradigm. *Bioscience* **1996**, *46* (5), 448–453.
43. Kowarik, I. Time Lags in Biological Invasions with Regard to the Success and Failure of Alien Species. In *Plant Invasions—General Aspects and Special Problems*; Pysek, P., Prach, K., Rejmanek, M., Wade, M., Eds.; SPB Academic Publishing: Amsterdam, 1995; 15–38.
44. Crawley, M.J. The population biology of invaders. *Phil. Trans. R. Soc. Lond. B.* **1986**, *314*, 711–731.
45. Louda, S.M.; Kendall, D.; Connor, J.; Simberloff, D. Ecological effects of an insect introduced for the biological control of weeds. *Science* **1997**, *277* (5329), 1088–1090.
46. Johnson, D.M.; Stiling, P.D. Distribution and dispersal of Cactoblastis cactorum (Lepidoptera: Pyralidae), an exotic opuntia-feeding moth, in Florida. *Florida Entomol.* **1998**, *81,* 12–22.
47. Cory, J.S.; Myers, J.H. Direct and indirect ecological effects of biological control. *Tr. Ecol. Evol.* **2000**, *15* (4), 137–139.
48. U.S. Department of Agriculture. Pest Risk Assessment of the Importation of Larch from Siberia and the Soviet Far East; USDA Forest Service, Misc. Publ. No. 1495. U.S. Government Printing Office: Washington, DC, 1991.
49. Carey, J.R. The future of the Mediterranean fruit fly *Ceratitus capitata* invasion of California: a predictive framework. *Biol. Conserv.* **1996**, *78* (1–2), 35–50.
50. Myers, J.H.; Savoie, A.; Van Randen, E. The irradication and pest management. *Annu. Rev. Entomol.* **1998**, *43*, 471–491.
51. McEvoy, P.B.; Rudd, N.T. Effects of vegetation disturbance on insect biological control of tansy ragwort, *Senecio jacobea*. *Ecol. Appl.* **1993**, *3* (4), 682–698.
52. Holway, D.A.; Suarez, A.V. Animal behavior: an essential component of invasion biology. *Tr. Ecol. Evol.* **1999**, *14* (8), 328–330.
53. Ewel, J.J.; O'Dowd, D.J.; Bergelson, J.; Daehler, C.C.; D'Antonio, C.M.; Gomez, L.D.; Gordon, D.R.; Hobbs, R.J.; Holt, A.; Hopper, K.R.; Hughes, C.E.; LaHart, M.; Leakey, R.B.; Lee, W.G.; Loope, L.L.; Lorence, D.H.; Louda, S.M.; Lugo, A.E.; McEvoy, P.B.; Richardson, D.M.; Vitousek, P.M. Deliberate introductions of species: research needs. *Bioscience* **1999**, *49* (8), 619–630.

Index

SECOND EDITION

ENVIRONMENTAL MANAGEMENT HANDBOOK

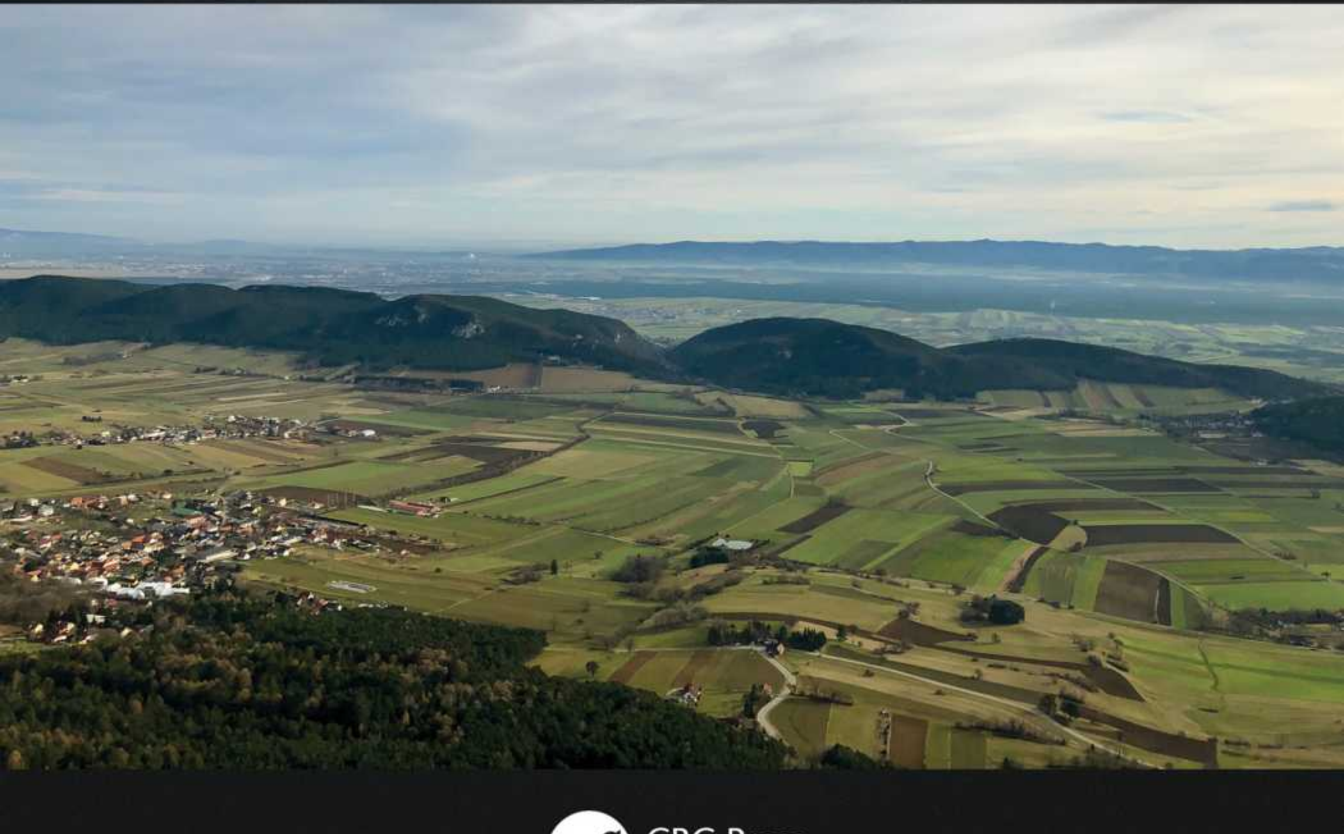
VOLUME III

Managing Soils and Terrestrial Systems

edited by

Brian D. Fath

Sven E. Jørgensen



CRC Press
Taylor & Francis Group

Managing Soils and Terrestrial Systems

Environmental Management Handbook, Second Edition

Edited by
Brian D. Fath and Sven E. Jørgensen

Volume 1
Managing Global Resources and Universal Processes

Volume 2
Managing Biological and Ecological Systems

Volume 3
Managing Soils and Terrestrial Systems

Volume 4
Managing Water Resources and Hydrological Systems

Volume 5
Managing Air Quality and Energy Systems

Volume 6
Managing Human and Social Systems

Managing Soils and Terrestrial Systems

Second Edition

Edited by
Brian D. Fath and Sven E. Jørgensen

Assisted by
Megan Cole

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

Cover photo: Hohe Wand Nature Park, Austria, N. Fath

Second edition published 2021
by CRC Press
6000 Broken Sound Parkway NW, Suite 300, Boca Raton, FL 33487-2742

and by CRC Press
2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

First edition published by CRC Press 2013

CRC Press is an imprint of Taylor & Francis Group, LLC

ISBN: 978-1-138-34265-1 (hbk)
ISBN: 978-0-429-34625-5 (ebk)

Typeset in Minion
by codeMantra

Contents

SECTION I APC: Anthropogenic Chemicals

SECTION II COV: Comparative Overviews of Important Topics for Environmental Management

SECTION III CSS: Case Studies of Environmental Management

SECTION IV DIA: Diagnostic Tools: Monitoring, Ecological Modeling, Ecological Indicators, and Ecological Services

SECTION V ELE: Focuses on the Use of Legislation or Policy to Address Environmental Problems

SECTION VI ENT: Environmental Management Using Environmental Technologies

SECTION VII NEC : Natural Elements and Chemicals Found in Nature

SECTION VIII PRO: Basic Environmental Processes

Preface

Given the current state of the world as compiled in the massive Millennium Ecosystem Assessment Report, humans have changed ecosystems more rapidly and extensively during the past 50 years than in any other time in human history. These are unprecedented changes that need certain action. As a result, it is imperative that we have a good scientific understanding of how these systems function and good strategies on how to manage them.

In a very practical way, this multi-volume *Environmental Management Handbook* provides a comprehensive reference to demonstrate the key processes and provisions for enhancing environmental management. The experience, evidence, methods, and models relevant for studying environmental management are presented here in six stand-alone thematic volumes as follows:

VOLUME 1 – Managing Global Resources and Universal Processes
VOLUME 2 – Managing Biological and Ecological Systems
VOLUME 3 – Managing Soils and Terrestrial Systems
VOLUME 4 – Managing Water Resources and Hydrological Systems
VOLUME 5 – Managing Air Quality and Energy Systems
VOLUME 6 – Managing Human and Social Systems

In this manner, this handbook introduces in the first volume the general concepts and processes used in environmental management. The next four volumes deal with each of the four spheres of nature (biosphere, geosphere, hydrosphere, and atmosphere). The last volume ties the material together in its application to human and social systems. These are very important chapters for a wide spectrum of students and professionals to understand and implement environmental management. In particular, the features include the following:

- The first handbook that demonstrates the key processes and provisions for enhancing environmental management.
- Addresses new and cutting-edge topics on ecosystem services, resilience, sustainability, food–energy–water nexus, socio-ecological systems, etc.
- Provides an excellent basic knowledge on environmental systems, explains how these systems function, and gives strategies on how to manage them.
- Written by an outstanding group of environmental experts.

Since this handbook covers such a wide range of materials from basic processes, to tools, technologies, case studies, and legislative actions, each handbook entry is further classified into the following categories:

APC: Anthropogenic chemicals: The chapters cover human-manufactured chemicals and activities
COV: Indicates that the chapters give comparative overviews of important topics for environmental management

- **CSS:** The chapters give a case study of a particular environmental management example
- **DIA:** Means that the chapters are about diagnostic tools: monitoring, ecological modeling, ecological indicators, and ecological services
- **ELE:** focuses on the use of legislation or policy to address environmental problems
- **ENT:** Addresses environmental management using environmental technologies
- **NEC:** Natural elements and chemicals: The chapters cover basic elements and chemicals found in nature
- **PRO:** The chapters cover basic environmental processes.

The third volume, with over 70 entries, hones in on soils and terrestrial systems, with chapters covering many aspects of agriculture and soil management and rehabilitation. The topic of erosion and desertification is covered by over 20 specific entries. New chapters on impacts of drought, pasture grazing, and climate impacts on soils are included.

Brian D. Fath
Brno, Czech Republic
December 2019

Editors

Brian D. Fath is Professor in the Department of Biological Sciences at Towson University (Maryland, USA) and Senior Research Scholar at the International Institute for Applied Systems Analysis (Laxenburg, Austria). He has published over 180 research papers, reports, and book chapters on environmental systems modeling, specifically in the areas of network analysis, urban metabolism, and sustainability. He has co-authored the books *A New Ecology: Systems Perspective* (2020), *Foundations for Sustainability: A Coherent Framework of Life–Environment Relations* (2019), and *Flourishing within Limits to Growth: Following Nature's Way* (2015). He is also Editor-in-Chief for the journal Ecological Modelling and Co-Editor-in-Chief for *Current Research in Environmental Sustainability*. Dr. Fath was the 2016 recipient of the Prigogine Medal for outstanding work in systems ecology and twice a Fulbright Distinguished Chair (Parthenope University, Naples, Italy, in 2012 and Masaryk University, Czech Republic, in 2019). In addition, he has served as Secretary General of the International Society for Ecological Modelling, Co-Chair of the Ecosystem Dynamics Focus Research Group in the Community Surface Modeling Dynamics System, and member and past Chair of Baltimore County Commission on Environmental Quality.

Sven E. Jørgensen (1934–2016) was Professor of environmental chemistry at Copenhagen University. He received a doctorate of engineering in environmental technology and a doctorate of science in ecological modeling. He was an honorable doctor of science at Coimbra University (Portugal) and at Dar es Salaam (Tanzania). He was Editor-in-Chief of *Ecological Modelling* from the journal inception in 1975 until 2009. He was Editor-in-Chief for the *Encyclopedia of Environmental Management* (2013) and *Encyclopedia of Ecology* (2008). In 2004, Dr. Jørgensen was awarded the Stockholm Water Prize and the Prigogine Medal. He was awarded the Einstein Professorship by the Chinese Academy of Sciences in 2005. In 2007, he received the Pascal Medal and was elected a member of the European Academy of Sciences. He had published over 350 papers, and has edited or written over 70 books. Dr. Jørgensen gave popular and well-received lectures and courses in ecological modeling, ecosystem theory, and ecological engineering worldwide.

Contributors

Erhan Akça
Technical Programs
Adiyaman University
Adiyaman, Turkey

Samson D. Angima
Oregon State University
Corvallis, Oregon

Kenneth A. Barbarick
Colorado State University
Fort Collins, Colorado

Susana Bautista
Department of Ecology
University of Alicante
Alicante, Spain

Suha Berberoğlu
Department of Landscape Architecture
University of Çukurova
Adana, Turkey

Ronny Berndtsson
Division of Water Resources Engineering
and
Center for Middle Eastern Studies
Lund University
Lund, Sweden

Marek Biziuk
Department of Analytical Chemistry
Gdansk University of Technology
Gdansk, Poland

David L. Bjorneberg
U.S. Department of Agriculture
Northwest Irrigation and Soils Research Lab
Agricultural Research Service (USDA-ARS)
Kimberly, Idaho

Julia Boike
Water and Environmental Research Center
University of Alaska
Fairbanks, Alaska

Nadia Bernardi Bonumá
Federal University of Santa Maria
Santa Maria, Brazil

John T. Brake
Department of Soil Science
North Carolina State University
Raleigh, North Carolina

D.R. Bray
Department of Animal Sciences
University of Florida
Gainesville, Florida

Zachary T. Broome
Bowen Radson Schroth, P.A.
Eustis, Florida

Steven N. Burch
University of Maryland
College Park, Maryland

David Burrow
Agriculture Victoria (Tatura)
Tatura, Victoria, Australia

Claudia Canedo-Rosso
Instituto de Hidráulica e Hidrología
Universidad Mayor de San Andrés
La Paz, Bolivia

and

Division of Water Resources Engineering
Lund University
Lund, Sweden

Carlos C. Cerri
University of São Paulo
São Paulo, Brazil

Guangnan Chen
Faculty of Engineering and Surveying
University of Southern Queensland
Toowoomba, Queensland, Australia

Jock Churchman
Land and Water
Commonwealth Scientific and Industrial Research Organization (CSIRO)
Adelaide, South Australia, Australia

Ahmet Çilek
Department of Landscape Architecture
University of Çukurova
Adana, Turkey

Alexandra Robin Collins
Department of Biology and Ecology
University of Fribourg/Perolles
Fribourg, Switzerland

Eric T. Craswell
Fenner School of Environment and Society
College of Medicine, Biology, and Environment
Australian National University
Canberra, Australian Capital Territory, Australia

Richard Cruse
Iowa State University
Ames, Iowa

Seth M. Dabney
U.S. Department of Agriculture
National Sedimentation Laboratory
Agricultural Research Service (USDA-ARS)
Oxford, Mississippi

Kathleen Delate
Departments of Agronomy and Horticulture
Iowa State University
Ames, Iowa

Detlef Deumlich
Leibniz Center for Agricultural Landscape Research (ZALF)
Institute of Soil Landscape Research
Muncheberg, Germany

Barbara Dinham
Eurolink Center
Pesticide Action Network UK
London, United Kingdom

Craig Ditzler
National Leader for Soil Classification and Standards
Lincoln, Nebraska

Douglas J. Dollhopf
Department of Land Resources and Environmental Sciences
Montana State University
Bozeman, Montana

Cenk Dönmez
Department of Landscape Architecture
University of Çukurova
Adana, Turkey

John W. Doran
U.S. Department of Agriculture
Agricultural Research Service (USDA-ARS)
University of Nebraska
Lincoln, Nebraska

Bahman Eghball
U.S. Department of Agriculture
Agricultural Research Service (USDA-ARS)
University of Nebraska
Lincoln, Nebraska

Reza Ehsani
Ohio State University
Columbus, Ohio

Krisztina Eleki
Department of Agronomy
Iowa State University
Ames, Iowa

Mehmet Akif Erdoğan
Technical Programs
Geographic Information Systems
University of Çukurova
Adana, Turkey

Gunay Erpul
Department of Soil Science
Ankara University
Ankara, Turkey

Anna Eynard
Ohio State University
Brookings, South Dakota

Delvin S. Fanning
Department of Environmental Science and Technology
University of Maryland
College Park, Maryland

David Favis-Mortlock
Environmental Change Institute
University of Oxford
Oxford, United Kingdom

Jolanta Fenik
Department of Analytical Chemistry
Gdansk University of Technology
Gdansk, Poland

Dennis C. Flanagan
U.S. Department of Agriculture
National Soil Erosion Research Laboratory
Agricultural Research Service (USDA-ARS)
West Lafayette, Indiana

Alan J. Franzluebbers
U.S. Department of Agriculture
Agricultural Research Service (USDA-ARS)
Watkinsville, Georgia

John R. Freney
Commonwealth Scientific and Industrial Research Organization (CSIRO)
Campbell, Australian Capital Territory, Australia

Brenda Frick
Bluebur Fluent Organics
Saskatoon, Saskatchewan, Canada

Monika Frielinghaus
Institute of Soil Landscape Research
Leibniz Center for Agricultural Landscape Research (ZALF)
Muncheberg, Germany

Roger Funk
Institute of Soil Landscape Research
Leibniz Center for Agricultural Landscape
Research (ZALF)
Muncheberg, Germany

Donald Gabriels
Department of Soil Management and Soil Care
Ghent University
Ghent, Belgium

Ján Gallo
Department of Plant Protection
Slovak University of Agriculture
Nitra, Slovak Republic

Rachael Garrett
Departments of Environmental System Science and Humanities
Political and Social Science
ETH Zurich
Zurich, Switzerland

Anja Gassner
Institute of Science and Technology
University of Malaysia—Sabah
Kota Kinabalu, Malaysia

David K. Gattie
Biological and Agricultural Engineering Department
University of Georgia
Athens, Georgia

Andrew S. Goudie
St. Cross College
Oxford, United Kingdom

Ed G. Gregorich
Agriculture and Agri-Food Canada
Easter Cereal and Oilseed Research Center
Ottawa, Ontario, Canada

Silvio J. Gumiere
Department of Soil Science
Laval University
Quebec City, Quebec, Canada

Andrew Paul Gutierrez
Center for the Analysis of Sustainable Agricultural Systems
University of California—Berkeley
Berkeley, California

John Holland
Head of Entomology
Game Conservancy Trust
Hants, United Kingdom

David J. Horn
Department of Entomology
Ohio State University
Columbus, Ohio

R. Cesar Izaurralde
Battelle Pacific Northwest National Laboratory
Washington, District of Columbia

Pierre A. Jacinthe
Carbon Management and Sequestration Center
The Ohio State University
Columbus, Ohio

and

Department of Earth Sciences
Indiana University- Purdue University, Indianapolis (IUPU)
Indianapolis, Indiana

Douglas Kane
Water and Environmental Research Center
University of Alaska
Fairbanks, Alaska

Burçak Kapur
Department of Agricultural Engineering
University of Çukurova
Adana, Turkey

Selim Kapur
Department of Soil Science
University of Çukurova
Adana, Turkey

Peter I.A. Kinnell
Institute of Applied Ecology, Faculty of Science and Technology
University of Canberra
Belconnen, Australian Capital Territory, Australia

Monika Kosikowska
Department of Analytical Chemistry
Gdansk University of Technology
Gdansk, Poland

John Kost
Department of Agronomy
Iowa State University
Ames, Iowa

John M. Laflen
U.S. Department of Agriculture
Agricultural Research Service (USDA-ARS), Buffalo Center
Iowa State University
Ames, Iowa

Rattan Lal
School of Environment and Natural Resources
Ohio State University
Columbus, Ohio

E.F. Legner
Department of Entomology
University of California
Riverside, California

Reynald Lemke
Agriculture and Agri-Food Canada
Swift Current
Saskatchewan, Canada

Rocky Lemus
Mississippi State University
Starkville, Mississippi

John Ludwig
Ecosystem Sciences
Commonwealth Scientific and Industrial Research Organization (CSIRO)
Winnellie, Northern Territory, Australia

Rory O. Maguire
Virginia Tech
Blacksburg, Virginia

Tek Narayan Maraseni
Australian Center for Sustainable Catchments
School of Accounting, Economics and Finance
University of Southern Queensland
Toowoomba, Queensland, Australia

Sean McGinn
Agriculture and Agri-Food Canada
Lethbridge, Alberta, Canada

Michael D. Melville
School of Biological, Earth, and Environmental Sciences
University of New South Wales
Sydney, New South Wales, Australia

Gustavo Enrique Merten
Federal University of Rio Grande do Sul
Porto Alegre, Brazil

Jean Paolo Gomes Minella
Federal University of Santa Maria
Santa Maria, Brazil

Heinz Müller-Schärer
Department of Biology and Ecology
University of Fribourg/Perolles
Fribourg, Switzerland

Mark A. Nearing
U.S. Department of Agriculture
Southwest Watershed Research Center
Agricultural Research Service (USDA-ARS)
Tucson, Arizona

Jerry Neppel
Department of Agronomy
Iowa State University
Ames, Iowa

L. Darrell Norton
U.S. Department of Agriculture
National Soil Erosion Research Laboratory
Agricultural Research Service (USDA-ARS)
West Lafayette, Indiana

Philip Oduor-Owino
Department of Botany
University of Kenyatta
Nairobi, Kenya

Michael T. Olexa
Center for Agricultural and Natural Resource Law
Institute of Food and Agricultural Sciences
University of Florida
Gainesville, Florida

Barron Orr
School of Natural Resources and the Environment
University of Arizona
Tucson, Arizona

Peter W. Plumstead
North Carolina State University
Raleigh, North Carolina

Wilfred M. Post
Oak Ridge National Laboratory
Oak Ridge, Tennessee

Franco Previtali
Department of Environmental Science
University of Milan—Bicocca
Milan, Italy

X. S. Qin
School of Civil and Environmental Engineering
Nanyang Technological University
Singapore

Martin C. Rabenhorst
University of Maryland
College Park, Maryland

Shafiqur Rahman
Department of Agricultural and Biosystems Engineering
North Dakota State University
Fargo, North Dakota

José Miguel Reichert
Federal University of Santa Maria
Santa Maria, Brazil

Pichu Rengasamy
Department of Soil and Water Science
University of Adelaide
Waite, South Australia, Australia

Philippe Rochette
Soils and Crops Research Center
Agriculture and Agri-Food Canada
Saint-Foy, Quebec, Canada

Ewald Schnug
Institute for Crop and Soil Science
Julius Kuhn Institute (JKI)
Braunschweig, Germany

Brenton S. Sharratt
U.S. Department of Agriculture
Land Management and Water Conservation Research Unit
Agricultural Research Service (USDA-ARS)
Pullman, Washington

Bal Ram Singh
Department of Plant and Environmental Sciences
Norwegian University of Life Sciences
Aas, Norway

Matt C. Smith
Biological and Agricultural Engineering Department
University of Georgia
Athens, Georgia

Hwat Bing So
Center for Environmental Systems Research
Griffith University
Nathan, Queensland, Australia

Robert E. Sojka
U.S. Department of Agriculture
Northwest Irrigation and Soils Research Lab
Agricultural Research Service (USDA-ARS)
Kimberly, Idaho

Graham P. Sparling
Landcare Research
Hamilton, New Zealand

Radost Stanimirova
Department of Earth and Environment
Boston University
Boston, Massachusetts

Larry D. Stetler
Department of Geology and Geological Engineering
South Dakota School of Mines and Technology
Rapid City, South Dakota

Matthew O. Sullivan
Department of Food, Agricultural, and Biological Engineering
Ohio State University
Columbus, Ohio

Maciej Tankiewicz
Department of Analytical Chemistry
Gdansk University of Technology
Gdansk, Poland

Daniel L. Thomas
Louisiana State University
Baton Rouge, Louisiana

David Tongway
Ecosystem Sciences
Commonwealth Scientific and Industrial Research Organization (CSIRO)
Canberra, Australian Capital Territory, Australia

Cintia B. Uvo
Division of Water Resources Engineering
Lund University
Lund, Sweden

H.F. van Emden
Department of Agriculture
University of Reading
Reading, United Kingdom

H.H. Van Horn
Department of Animal Sciences
University of Florida
Gainesville, Florida

R. Scott Van Pelt
U.S. Department of Agriculture
Wind Erosion and Water Conservation Research Unit
Agricultural Research Service (USDA-ARS)
Big Spring, Texas

Dalmo A.N. Vieira
U.S. Department of Agriculture
National Sedimentation Laboratory
and
Arkansas State University
Agricultural Research Service (USDA-ARS)
Jonesboro, Arkansas

Joel T. Walker
Department of Food, Agricultural, and Biological Engineering
Ohio State University, Columbus, Ohio

Xingxiang Wang
Institute of Soil Science
Chinese Academy of Sciences
Nanjing, China

Andrew Warren
Department of Geography
University College London
London, United Kingdom

Thomas R. Way
U.S. Department of Agriculture
National Soil Dynamics Laboratory
Agricultural Research Service (USDA-ARS)
Auburn, Alabama

Ian White
College of Science
Fenner School of Environment and Society College of Medicine, Biology and Environment
Australian National University
Canberra, Australian Capital Territory, Australia

Keith D. Wiebe
Resource Economics Division
U.S. Department of Agriculture (USDA)
Washington, District of Columbia

Y. Xu
MOE Key Laboratory of Regional Energy Systems Optimization
S-C Energy and Environmental Research Academy
North China Electric Power University
Beijing, China

Bofu Yu
Griffith School of Engineering
Griffith University
Nathan, Queensland, Australia

Taolin Zhang
Institute of Soil Science
Chinese Academy of Sciences
Nanjing, China

X.-C. (John) Zhang
U.S. Department of Agriculture
Grazinglands Research Laboratory
Agricultural Research Service (USDA-ARS)
El Reno, Oklahoma

Claudio Zucca
Department of Agriculture, Food, and Environment
University of Perugia
Perugia, Italy

I

APC: Anthropogenic Chemicals

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

1

Agricultural Soils: Nitrous Oxide Emissions

John R. Freney

Introduction

Nitrous oxide is a gas that is produced naturally by many different micro-organisms in soils and waters, and as a result of human activity associated with agriculture, biomass burning, stationary combustion, automobiles, and the production of nitric and adipic acids for industrial purposes. According to the Intergovernmental Panel on Climate Change (IPCC),[1] ~23.1 million metric tons (Mt) of nitrous oxide is emitted each year, 14.1 Mt as a result of natural processes (~4.7Mt from the oceans, ~6.3 Mt from tropical soils, and ~3.1 Mt from temperate soils); and ~9 Mt as a result of human activities (5.5 Mt from agricultural soils, 0.6 Mt from cattle and feedlots, 0.8 Mt from biomass burning, and 2.1 Mt from mobile sources and industry). While there is considerable uncertainty associated with each of these estimates, it is apparent that most nitrous oxide is derived from soils.

Because of the intimate connection between the Earth and the atmosphere, much of the nitrous oxide produced enters the atmosphere and affects its chemical and physical properties. Nitrous oxide contributes to the destruction of the stratospheric ozone layer that protects the Earth from harmful ultraviolet radiation, and is one of the more potent greenhouse gases that trap part of the thermal radiation from the Earth's surface. The atmospheric concentration of nitrous oxide is ~313 parts per billion. It is increasing at the rate of ~0.7 parts per billion each year, and its lifetime is ~166 years.[2] It seems that the increased atmospheric concentration results from the increased use of synthetic fertilizer nitrogen, biologically fixed nitrogen, animal manure, crop residues, and human sewage sludge in agriculture to produce food and fiber for the rapidly increasing world population.[3]

Nitrous Oxide Emission from Agriculture

All soils are deficient in nitrogen for the growth of plants, but the deficiency can be overcome by adding fertilizer nitrogen. When the fertilizer (e.g., urea or ammonia-based compounds) is applied to soil, it is transformed by microorganisms as follows:

$$\begin{aligned}\text{Fertilizer nitrogen} \xrightarrow{1} \text{Ammonium} \xrightarrow{2} \text{Nitrite} \\ \xrightarrow{3} \text{Nitrate} \xrightarrow{4} \text{Nitrate} \\ \xrightarrow{5} \text{Nitric oxide} \xrightarrow{6} \text{Nitrous oxide} \\ \xrightarrow{7} \text{Dinitrogen}\end{aligned} \tag{1}$$

When the soil is aerobic (i.e., when oxygen is present) ammonium is oxidized to nitrite and nitrate (Steps 2 and 3). This process is called nitrification. After addition of irrigation water or rain, the soil may become anaerobic (devoid of oxygen). The nitrate is then reduced by soil organisms to nitrite and the gases nitric oxide, nitrous oxide, and dinitrogen (Steps 4–7) in a process termed denitrification.[4]

When atmospheric scientists first expressed concern that nitrous oxide emission into the atmosphere, as a result of fertilizer use, would lead to destruction of the ozone layer, it was thought that nitrous oxide was produced mainly from the microbiological reduction of nitrate in poorly aerated soils. However, research in the latter part of the 1970s showed that significant nitrous oxide was emitted from aerobic soils during nitrification of ammonium, and subsequent work has shown that nitrification is a major source of nitrous oxide.[4]

Nitrous Oxide from Denitrification

Certain micro-organisms in the absence of oxygen have the capacity to reduce nitrate (or other nitrogen oxides). Most denitrifying bacteria are heterotrophs—that is, they require a source of organic matter for energy—but denitrifying organisms that obtain their energy from light or inorganic compounds also occur in soils. The capacity to denitrify has been reported in more than 20 genera of bacteria, and almost all are aerobic organisms that can only grow anaerobically in the presence of nitrogen oxides. The dominant denitrifying organisms in soil are *Pseudomonas* and *Alcaligenes.* In addition to the free-living denitrifiers, Rhizobia living symbiotically in root nodules of legumes are able to denitrify nitrate and produce nitrous oxide.[4]

The general requirements for biological denitrification include the presence of micro-organisms with denitrifying capacity, nitrate (or other nitrogen oxides) and available organic matter, the absence of oxygen, and a suitable pH and temperature environment. In aerobic soils, denitrification can occur in anaerobic microsites in soil aggregates or in areas of high carbon content, where active microbial activity rapidly consumes all of the available oxygen.[4]

Nitrous Oxide from Nitrification

The process of nitrification is normally defined as the biological oxidation of ammonium to nitrate with nitrite as an intermediate.[4] The first step in the reaction, the oxidation of ammonium to nitrite, is carried out mainly by the microorganism *Nitrosomonas.* The second step, oxidation of nitrite to nitrate, is carried out by *Nitrobacter.* It has been shown in a number of publications that *Nitrosomonas europaea* produces nitrous oxide during the oxidation of ammonium.[4]

The possibility that significant nitrous oxide can be produced in soils by nitrifying organisms was indicated by studies that showed that soils incubated under aerobic conditions with ammonium produced more nitrous oxide than soils amended with nitrate.[4] In addition, treatment of aerobic soils with nitrapyrin, which inhibits nitrification of ammonium but has little effect on denitrification, markedly reduced the emission of nitrous oxide.[4] Production of nitrous oxide by nitrification in soils is increased by increasing temperature, pH, water content, available carbon, and the addition of ammonium-based fertilizers, plant residues, and animal manure.

Flooded Soils

In the past few years, increased attention has been given to nitrous oxide emission from paddy soils. The concern is that the introduction of management practices to reduce methane emissions from flooded soils may result in increased emissions of nitrous oxide.

Flooded soils are characterized by an oxygenated water column overlying an oxidized layer at the soil-water interface, an aerobic zone around each root, and anaerobic conditions in the remainder of the soil. This differentiation of the flooded soil into oxidized and reduced zones has a marked effect on the transformation of nitrogen.[5] The resulting reactions are as follows:

1. Ammonium in the reduced zone diffuses to the oxidized zone;
2. Ammonium is oxidized to nitrate by nitrifying organisms;
3. The nitrate formed diffuses to the anaerobic zone;
4. Denitrification occurs with the production of nitrous oxide and dinitrogen;
5. The gaseous products diffuse through the soil and water layers to the atmosphere.[6]

It is apparent that the rate of diffusion of ammonium to an oxidized layer and the rate of nitrification in the oxidized layer are factors controlling the production of nitrous oxide in flooded soils. The rate of diffusion of nitrous oxide through the soil and water layers will control its rate of emission to the atmosphere, or its further reduction to dinitrogen.[5]

A number of mechanisms have been identified for the transfer of nitrous oxide from the soil to the atmosphere.[3] Nitrous oxide may diffuse from the zone of production through the saturated soil and water layer to the atmosphere. It may also enter the roots of the rice plant and move by diffusion through the plant to the atmosphere in the same way as methane. Bhadrachalam et al.[6] studied the importance of the two pathways in intermittently flooded rice in the field in India and found that nitrogen gas fluxes were ~30% greater when transfer through the plants was included.

In the tropics, rice is usually transplanted and fertilized some time after flooding. Because of the anaerobic conditions that develop before fertilization and the slow rate of diffusion of nitrous oxide in flooded soils, most of the nitrous oxide is reduced to dinitrogen and very little escapes to the atmosphere. Nitrous oxide emission from intermittently flooded rice was relatively large compared with that from permanently flooded rice, reflecting the different oxidation states of intermittently and continuously flooded soils.[6] Studies of nitrous oxide emission from rice fields from the time the soils were drained for harvest, through to flooding the soil in preparation for planting the next crop, showed that nitrous oxide was emitted continuously while the soil was not flooded. Overall, the rate of emission of nitrous oxide from flooded soils was less than that from upland soils after application of nitrogen fertilizer.[3]

Biomass Burning

During combustion the nitrogen in the fuel can be converted into gaseous forms such as ammonia, nitric oxide, nitrous oxide, dinitrogen, and hydrogen cyanide. It is estimated that biomass burning contributes between 0.3 and 1.6 Mt nitrous oxide per year globally to the atmosphere.[3] Most of the biomass burning (~90%) takes place in the tropics as a result of forest clearing, savanna and sugar cane fires, and burning of agricultural wastes and firewood.[7]

Biomass burning is not only an instantaneous source of nitrous oxide, but it results in a longer-term enhancement of the production of this gas. Measurements of nitrous oxide emissions from soils, before and after burning showed that significantly more nitrous oxide was exhaled after the burn through alteration of the chemical, biological, and physical processes in soil.[7]

Fertilizer Consumption and Nitrous Oxide Production

Nitrous oxide emissions from agricultural soils are generally greater and more variable than those from uncultivated land. Application of fertilizer nitrogen, animal manure, and sewage sludge usually results in enhanced emissions of nitrous oxide.[7] Generally, there is a large emission of nitrous oxide immediately after the application of fertilizer. After about 6 weeks, the emission rate falls and fluctuates around a low value. Mosier[8] concluded that interactions between the physical, chemical, and biological variables are complex, that nitrous oxide fluxes are variable in time and space, and that soil management, cropping systems, and variable rainfall appear to have a greater effect on nitrous oxide emission than the type of nitrogen fertilizer. Consequently, Mosier et al.[9] recommend the use of one factor only for calculating the emission of nitrous oxide from different fertilizer types:

$$N_2O \text{ emitted} = 125\% \text{ of N applied } (\text{kg/ha}) \tag{2}$$

This equation is based on data from long-term experiments with a variety of mineral and organic fertilizers, and encompasses 90% of the direct contributions of nitrogen fertilizers to nitrous oxide emissions.

Mosier et al.[3] developed a methodology to estimate agricultural emissions of nitrous oxide, taking into account all of the nitrogen inputs into crop production. They included direct emissions from agricultural soils as a result of synthetic fertilizer addition, animal wastes, increased biological nitrogen fixation, cultivation of mineral and organic soils through enhanced organic matter mineralization, and mineralization of crop residues returned to the field. Indirect nitrous oxide emissions resulting from deposition of ammonia and oxides of nitrogen, leaching of nitrate, and introduction of nitrogen into sewage systems were also included. They concluded that in 1989, 9.9 Mt of nitrous oxide was emitted into the atmosphere directly or indirectly, as a result of agriculture (Table 1).

TABLE 1 Calculated Emission of Nitrous Oxide from Agricultural Activities

	Mt Nitrous Oxide per Year
Direct soil emissions	
Synthetic fertilizer	1.4 (0.28–2.5)
Animal waste	0.9 (0.19–1.7)
Biological nitrogen fixation	0.16 (0.03–0.3)
Crop residue	0.6 (0.11–1.1)
Cultivation of Histosols	0.16 (0.03–0.3)
Total	3.3 (0.6–5.9)
Animal production	
Waste management systems	3.3 (0.9–4.9)
Indirect emissions	
Atmospheric deposition	0.47 (0.09–0.9)
Nitrogen leaching and runoff	2.5 (0.2–12.1)
Human sewage	0.3 (0.06–4.1)
Total	3.3 (0.35–17.1)
Total	9.9 (1.9–27.9)

Source: Modified from Mosier et al.[3]

Management Practices to Decrease Nitrous Oxide Emission

The low efficiency of fertilizer nitrogen in agricultural systems is primarily caused by the large losses of mineral nitrogen from those systems by gaseous loss: nitrous oxide emission is directly linked to the loss processes. It is axiomatic that any strategy that increases the efficiency of nitrogen fertilizer use will reduce emissions of nitrous oxide, and this has been directly demonstrated for a number of strategies.[3]

The IPCC[1] reported that some combination of the following management practices, if adopted worldwide, would improve the efficiency of the use of synthetic fertilizer and manure nitrogen, and significantly reduce nitrous oxide emission into the atmosphere:

1. Match nitrogen supply with crop demand.
2. Tighten nitrogen flow cycles by returning animal wastes to the field and conserving residues instead of burning them.
3. Use controlled-release fertilizers, incorporate fertilizer to reduce volatilization, use urease and nitrification inhibitors, and match fertilizer type to precipitation.
4. Optimize tillage, irrigation, and drainage.

The potential decrease in nitrous oxide emissions from synthetic fertilizer, as a result of the mitigation techniques, could amount to 20%.[1]

References

1. IPCC (Intergovernmental Panel on Climate Change). *Climate Change 1995. Impacts, Adaptations and Mitigation of Climate Change*: Scientific-Technical Analyses; Watson, R. T., Zinyowera, M.C., Moss, R.H., Eds.; Cambridge University Press: Cambridge, England, 1996; 1–878.
2. Hengeveld, H.; Edwards, P. 1998. An assessment of new research developments relevant to the science of climate change. Climate Change Newsletter **2000**, *12,* 1–52.
3. Mosier, A.; Kroeze, C.; Nevison, C.; Oenema, O.; Seitzinger, S.; van Cleemput, O. Closing the global N2O budget: nitrous oxide emissions through the agricultural nitrogen cycle. Nutri. Cycling Agroecosys. **1998**, *52*, 225–248.
4. Bremner, J.M. Sources of nitrous oxide in soils. Nutri. Cycling Agroecosys. **1997**, *49,* 7–16.
5. Patrick, W.H., Jr. Nitrogen transformation in submerged soils. In *Nitrogen in Agricultural Soils;* Stevenson, F.J., Ed.; American Society of Agronomy: Madison, WI, 1982; 449–465.
6. Bhadrachalam, A.; Chakravorti, S.P.; Banerjee, N.K.; Mohanty, S.K.; Mosier, A.R. Denitrification in intermittently flooded rice fields and N-gas transport through rice plants. Ecol. Bull. **1992**, *42*, 183–187.
7. Granli, T.; Bøckman, O.C. Nitrous oxide from agriculture. Norwegian J. Agric. Sci. **1994**, Supplement No. 12, 1–128.
8. Mosier, A.R. Chamber and isotope techniques. In *Exchange of Trace Gas between Terrestrial Ecosystems and the Atmosphere;* Andreae, M.O., Schimel, D.S., Eds.; John Wiley and Sons: Chichester, England, 1989; 175–187.
9. Mosier, A.R.; Duxbury, J.M.; Freney, J.R.; Heinemeyer, O.; Minami, K. Nitrous oxide emissions from agricultural fields: assessment, measurement and mitigation. Plant Soil **1995**, *181,* 95–108.

2

Agriculture: Energy Use and Conservation

Guangnan Chen
and Tek Narayan
Maraseni

Introduction

Farming is often an energy-intensive operation. With the rising energy costs and increasing concern over greenhouse gas (GHG) emissions, effective energy use and conservation are becoming increasingly important.[1,2] In many parts of the world, energy inputs represent a major and rapidly growing cost input for growers, with on-farm fuel requirement in cereal cropping systems often exceeding 100 L/ha for tillage, seeding, fertilizing, spraying, and harvesting.[3–5] In the United States, it has been found that the operations of food systems, including agricultural production, food processing, packaging, and distribution, accounted for approximately 19% of America's national fossil fuel energy use.[6]

Methods of Energy Audit

Energy audits and assessments are a first step and crucial part of the energy and environmental management process. Energy audits refer to the systematic examination of an entity, such as a firm, organization, facility, or site, to determine whether, and to what extent, it has used energy efficiently.[7] They may also assess opportunities of potential energy savings through fuel switching, tariff negotiation, and demand-side management.

Energy audits may be broadly classified into the following three levels:[7,8]

- Energy audit level 1, or preliminary audit. The main purpose of this level of energy audits is to overview the total energy consumption of a farm. This is the simplest and cheapest form of energy audits. It usually involves collating all the energy use data from the farm, including the total fuel usage from diesel, petrol, and other fuels and the total electricity energy consumed, in order to derive basic energy performance indices such as GJ/ha and GJ/t.

- Energy audit level 2, or standard/general audit. Level 2 energy audits generally involve breaking down the total energy usage into each individual major processes. A level 2 audit is usually process-based and involves some specific measurements for the key processes. It may also involve considerable farmers' interviews to identify the major energy usage.
- Energy audit level 3, or detailed specific operation investigation. This is the highest and the most expensive level of energy audits. It may involve investigating ways to improve the efficiency of a specific operation so that the investment return can be accurately predicted. This will usually involve the uses of a range of different sensors to measure the performance of a variety of machines. Data loggers may be used to record the data for a considerable period of time.

Underpinning the energy audit assessment process is the identification and development of a set of toolkits (software, hardware, and other supporting information and resources) that can be applied to different farm production systems. A number of farm energy calculators[4,7–9] have already been developed to evaluate the energy uses from agricultural systems. Various hardware and technologies such as electricity power meters, fuel flow meters, data logging, and monitoring equipment may also be used for undertaking field measurements.[9–11] By itemizing farm energy usage from each operation, farmers will be able to identify where energy is mostly consumed and therefore explore ways to reduce energy use.

Brown and Elliot[12] found that although the currently available data of agricultural energy use may be sufficient for general policy development, the quality of existing energy end-use data is often unsatisfactory. They suggested that further research be conducted to achieve a clear and consistent definition of farm types and energy end uses. Energy uses of individual operations will also need to be more accurately measured.

Effect of Farming Systems on Energy Use

Considerable research has been conducted on energy use and conservation both in agriculture[13–16] and in other industries.[17] In the cropping sector, it has been identified that a number of practice changes and technology developments may be adopted to reduce fuel/energy use or energy use intensity. Examples of these may include more effective machinery operation, conservation farming practices, improved irrigation methods and water use efficiency measures, precision agriculture, and, where appropriate, the use of renewable energy and planting of alternative crops.

Machinery Operation

Farm machinery is integral to many aspects of modern farming systems. Without it, we may not be able to effectively implement the farming systems. It was estimated that since the Second World War, more than 25% of the increase in grain productivity may be attributed to machinery innovations. When the ownership cost and timeline loss are taken into consideration, it is estimated that machinery may consist of up to 40%–50% of total farm input costs.[18]

Farm machinery often consumes a large amount of energy and therefore produces considerable emissions, especially carbon dioxide. Figure 1 shows the average percentage contribution of the direct energy input for different farming processes in European tillage-based systems. Chen et al.[19] and Smith[20] showed that by changing gear selection and engine speed, up to 30% of energy can often be saved for the same tractor power output. Careful selection of energy-efficient farming equipment and proper matching of tractor and implement are also important.[21] At the systems level, it is possible that much larger improvements can be achieved by changing the cropping practices and systems. Table 1 shows the estimated average fuel use for different tillage operations,[22] where a ratio of up to 3:1 or 4:1 from the highest to the lowest energy use can be found.

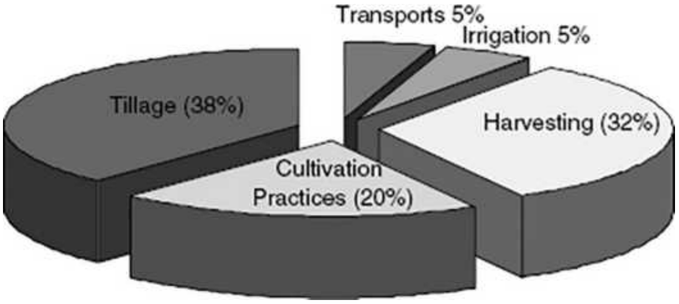


FIGURE 1 Direct on-farm energy inputs in Europe.

TABLE 1 Average Fuel Use for Different Tillage Methods

Soil Tillage Methods	Average Fuel (Diesel) Use (L/ha)
Subsoiling	18
Discing	12
Chisel ploughing	7
Power harrowing	8
Light harrowing/rolling	4

Source: Adapted from Chen and Baillie[4] and Downs and Hansen.[22]

Conservation Agriculture

Conservation agriculture is the generally accepted label for farming systems developed with the specific objective of improving productivity and sustainability, by reducing input waste and resource degradation. Conservation tillage involves reduced tillage, in which a crop is planted in the residue from a previous crop with either minimal (minimum tillage) or zero (zero tillage) soil disturbance. The weed control function may be achieved by using a combination of herbicides and agronomic measures (cover crops, rotation, seed bank depletion) and soil damage minimized by better management (soil health, traffic intensity, controlled traffic).

Based on 2 years experiments in Croatia, Kosutic et al.[23] demonstrated that compared with the conventional "intensive" tillage, up to 82.6% of energy could be saved by the adoption of no-till conservation farming system, and without the significant yield reduction. Gulden and Entz[24] found that this was 36% in Canada. Bailey et al.[25] showed that in terms of total energy used, when compared with conventional farming systems on a per-hectare basis, the integrated arable systems had the potential to reduce overall energy consumption by about 8% in the United Kingdom. However, in terms of energy use per kilogram of output, their results were less conclusive. A comparative study of conventional tillage and conservation farming was also carried out by Smith[26] to compare the impact of these practices on soil characteristics, crop performance, and economic outcomes.

While many no-till seeders have been marketed in recent years, there is still no one universally applicable machine that guarantees success under all conditions. In fact, better seed and fertilizer placement and good depth control and stubble handling have been identified by farmers as the highest priority needs in the design of planting equipment, particularly on "unprepared" soil with heavy stubble in direct drilling/conservation farming.[27] Difficulties of no-till seeding are therefore the major stumbling block to the adoption of full conservation agriculture. Despite these issues, effective machine/system combinations for no-till seeding have been evolved for many cropping environments, particularly in the subtropical grain production, where conventional tillage-based systems were causing major erosion damage and loss of organic matter. Where soil degradation problems are less obvious, for instance, in the long-established plough-based systems of Europe, adoption of conservation agriculture has been much slower,[28] and the same is true in most horticultural and intensive cropping areas, worldwide. This situation is expected to change slowly as cropping inputs, particularly fuel, fertilizer, and water resources, become more expensive, improving the attractiveness of system change.

Controlled traffic is a farming system in which the wheel tracks of all operations are confined to fixed paths so that re-compaction of soil by traffic (traction or transport) does not occur outside the selected paths. It has been shown that in addition to benefits of improved productivity and soil and water conservation, this method could also provide up to 50% reduction in power and fuel requirements of field operations.[29,30] Tullberg[31] demonstrated the effects of system change from the conventional tillage-based to no-till and to controlled traffic no-till in Australian grain production, clearly illustrating the significance of the reduction of energy and GHG emissions. It was shown that the total GHG emissions per ton of grain production can be reduced from 314 to 306 and to 175 kg, respectively. One survey indicated that in 2003, some form of controlled traffic was in use over 1 million ha in Australia.[32] To achieve further widespread adaptation, it is identified that the lack of compatibility between different machinery would have to be overcome.[31]

Irrigation Methods

Irrigation can be broadly defined as the practice of applying additional water (beyond what is available from rainfall) to soil to enable or enhance plant growth and yield. The water source could be groundwater pumped to the surface or surface water diverted from one location to another. On a global scale, it was estimated that 17% of irrigated cropland produces 40% of the total production.[33]

Farm irrigation systems may be broadly classified as surface irrigation, pressured sprinkler systems, trickle, and subsurface systems. In Australia, surface irrigation systems, such as border check and furrow irrigation, are the most widely used irrigation systems, covering some 70% of the total irrigated area.[34] This is then followed by the pressurized spray irrigation methods (22%) and trickle and subsurface systems (8%).

Irrigation is however often a very energy-intensive operation. It was estimated that 23% of the on-farm energy use for crop production in the United States was for on-farm pumping.[35] Figure 2 shows the estimated total lifecycle energy consumption for different irrigation methods.[34] It can be seen that

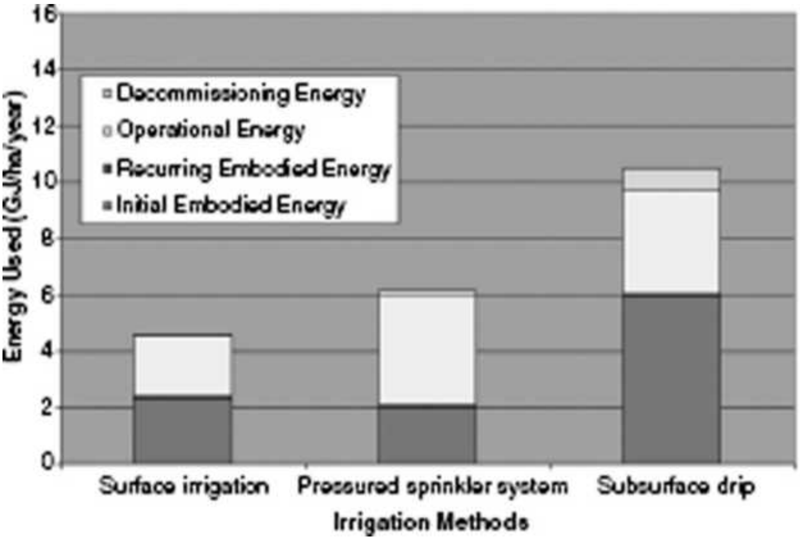


FIGURE 2 Life-cycle energy consumption for different irrigation systems.

the pressurized irrigation system can use significantly more energy over its life cycle than a gravity-fed system, with a ratio of about 1.5 or 2. The increased wateruse efficiency of pressured irrigation systems will therefore need to be balanced against the higher cost of the energy needed.[36] At the current market condition, 1 GJ of energy would typically cost around \$10–\$40. The corresponding GHG emissions range from 0 (using renewable energy) to around 78 kg CO_2/GJ (using diesel fuel).

Brown and Elliot[12] found that the largest on-farm energy savings might be available in motorized systems, especially in irrigation pumping. Barron[37] also identified that poor system design contributed significantly to low pumping efficiency. Pathak and Bining[38] showed that for irrigation, fuel saving of more than 50% was feasible through improvement in the irrigation equipment and improved water management practices. Chen et al.[19] demonstrated that with suitable operation adjustments, up to 55% of pumping energy can be saved. It has also been reported in the United States that energy efficiency audits on irrigation systems have, on average, identified savings of at least 10% of the energy bill—and in many instances up to 40%.[38] Very often, the irrigators who owned these inefficient systems were unaware of any problems.[39]

Overall Systems Comparison

Table 2 summarizes some of the published energy use data for different crops in different countries.[8,39–44] It can be seen that significant variation in energy input costs occurs between different crops and between different growers due to the adopted production system and a grower's performance within that system. Depending on the farming systems adopted, the direct on-farm energy input may range from 2.5 GJ/ha for wheat production in Europe to 53.4 GJ/ha for greenhouse tomato production in Turkey. This is a ratio of approximately 20.

It is noted that on-farm energy uses of crop production is only a proportion of total energy usage. The others include the post-farm processing, transport, distribution, and energy used to produce agricultural inputs such as fertilizers and pesticides.[44–46] Table 3 shows the estimated embodied energy content for manufacturing various chemical fertilizers and herbicides. It can be seen that with the current manufacturing technology, the production of 1 kg of nitrogen fertilizer would typically require the energy input equivalent to 1.5–2 kg of fuel, while 1 kg of pesticides would require the energy input equivalent to up to 5 kg of fuel.[44] It is therefore important to reduce not only the on-farm energy uses but also the embodied energy and post-harvest energy uses. This is especially important, as the embodied energy of farm chemicals may account for up to 50%–70% of the total energy input in agricultural production. In particular, the cost of nitrogen fertilizer is mostly a reflection of energy costs, so maximizing the

TABLE 2 Energy Input for Different Crops in Different Countries

Crops	Direct Energy Input (GJ/ha)	Indirect Energy Input (GJ/ha)	Total Energy Input (GJ/ha)	Country/ Region
Wheat	2.5–4.3			Europe
Wheat			16–32	Greece
Cotton	21.14	28.59	49.73	Turkey
Cotton			82.6	Greece
Cotton	3.7–15.2			Australia
Maize	4.75.0			Europe
Rice			64.89	United States
Pea			2.5–5.4	Canada
Dairy pasture	14.6	3.6	18.2	New Zealand
Greenhouse Tomato	53.4	53.3	106.7	Turkey

Source: Chen and Baillie.[8]

TABLE 3 Embodied Energy Content for Various Agricultural Fertilizers and Herbicides

Chemicals	Energy Content (MJ/kg Element or Active Ingredient)
N	65
P	15
K	10
S	5
Generic herbicide	270

efficiency of fertilizers or using legumes to produce nitrogen can help reduce energy costs on farms. It was demonstrated that energy input in the integrated forage-grain rotation was approximately 40% lower than that in the grain-based rotation.[47]

Yaldiz et al.[48] reported that fertilizers and irrigation energy dominated the total energy consumption in Turkish cotton production. Yilmaz et al.[41] showed that the energy intensity in agricultural production was closely related with production techniques. They estimated that cotton production in Turkey consumed a total of 49.73 GJ/ha energy, consisting of 21.14 GJ/ha (42.5%) direct energy input and 28.59 GJ/ha (57.5%) indirect energy input. Modern rice production in the United States also requires a total energy input of 64.89 GJ/ha for a yield of 5.8 tons/ha or 11.19 GJ energy input per ton of grain produced.[43]

To save energy and cost, the precision-agriculture technologies such as controlled traffic farming, high spatial and temporal precision in input placement, sensor-controlled spraying, and variable-rate fertilizer applications may offer significant benefits. Pimentel et al.[3] also showed that by appropriate technology changes in food production, processing, packaging, transportation, and consumption, fossil energy use in the U.S. food system could be reduced by about 50%. Using corn production as a model crop, it was estimated that total energy in corn production could be reduced by more than 50% by combining the following changes: 1) using smaller machinery and less fuel; 2) replacing commercial nitrogen applications with legume cover crops and livestock manure; and 3) reducing soil erosion through alternative tillage and conservation techniques. Hoeppner et al.[47] and Ziesemer[49] also showed that energy input in the organic management system was approximately 50% lower than that in the conventional management system.

Development and Applications of New Technology

With the advance of biotechnology, new crop varieties with higher yield and improved performance such as reduced herbicide and pesticide usage are being continuously introduced.[50] This will reduce not only the amount of herbicide and pesticide used but also the fuel usage associated with the crop spraying operations. For example, it was reported that with the adoption of genetically modified cotton, pesticide use has been reduced by up to 70%–90% in the Australian cotton industry.[51] To address the public's concern, the introduction of biotechnology to food crops may require a more careful approach.

Where the opportunities are appropriate, renewable energy such as solar, wind, and biofuel may also be integrated into the farming operations to save energy costs and reduce GHG emissions.[52–55] Examples of direct applications include solar crop drying, solar space and water heating, and using biomass for heating purpose. Other applications include generation of off-grid electricity for electric fences, lighting, irrigation, livestock water supply, wastewater treatment, pond aeration, communication, and remote equipment operation. Table 4 shows the estimated costs for various electricity generation technologies. Overall, renewable energy at present may not be viable to serve in large scale as a primary energy source, and in many cases, additional government support and financial incentives would still be required to compete with more polluting fossil fuels. However, because of the increasing concern over the GHG emissions and declining reserves of fossil fuel, use of renewable energy is on the rapid rise globally.[56] Various user-friendly design and cost estimation tools have also been developed to promote specific applications.[57]

TABLE 4 Estimated Costs of Various Electricity Generation Methods

Energy Type	Cost (¢/kWh)
Hydropower	4–10
Coal-fired power	2–4
Gas-fired power	5
Wind energy	5–8
Solar PV electricity	15–25

Conclusion

Energy use efficiency and conservation are of significant interest to the agricultural industry. Within highly mechanized agricultural productions systems, operational energy inputs (i.e., diesel, petrol, gas, and electricity) often represent a major cost to the growers. In the United States, it has been found that the operations of food systems, including agricultural production, food processing, packaging, and distribution, accounted for approximately 19% of America's national fossil fuel energy use. A total of up to 64.89 GJ/ha of energy is spent in the rice production in the United States.

It has been shown that significant variation in energy input costs can occur between growers due to the adopted production system and a grower's performance within that system. In the cropping sector, it has been identified that a number of practice changes and technology developments may be adopted to reduce fuel/energy use in agriculture. These may include more effective machinery operation, conservation farming practices, improved irrigation methods and water use efficiency measures, precision agriculture, and, where appropriate, the use of renewable energy and planting of alternative crops or crop rotation. It has been demonstrated that by changing the farming systems (e.g., changing to conservation tillage system, controlled traffic, or organic management system) and undertaking the same operation more efficiently (e.g., more precise targeting and use of fertilizers and irrigation), up to 50% of overall energy use may be realistically saved. It has also been found that from the highest to the lowest energy use, the estimated average fuel use for different tillage operations may be up to a ratio of 3:1 or 4:1. In many cases, the largest on-farm energy savings might be available in motorized systems, especially in irrigation pumping. The energy efficiency audits on irrigation systems in the United States have, on average, identified savings of at least 10% of the energy bill—and in many instances up to 40%. Because the embodied energy of various chemicals and machinery may make up 50%–70% of the total energy input and GHG emissions in agricultural production, the applications of variable-rate fertilizer technology and replacement of commercial nitrogen applications with legume cover crops and livestock manure would bring significant benefits.

With the development and applications of new technology and the increasing cost of fuel, fertilizer, and water resources, it is expected that various energy-saving measures will be increasingly adopted by the farmers. In particular, with the increasing concern over the GHG emissions and the depleting of fossil fuel, renewable energy such as solar, wind, and biofuel may be integrated into the different farming operations, including solar crop drying and heating, off-grid electricity supply for fences, lighting, irrigation, livestock water supply, wastewater treatment, pond aeration, communication, and remote equipment operation. With the adoption of genetically modified cotton, pesticide use has been reduced by up to 70%-90% in the Australian cotton industry.

Acknowledgments

The authors wish to extend special thanks to Dr. Jeff N. Tullberg for his helpful advice and comments to this entry.

References

1. Fischer, J.R.; Johnson, S.R.; Finnell, J.A. Energy, agriculture and the food—Cutting costs, improving efficiency, and stewarding natural resources. Resour. Mag. **2009**, *16* (2), 16–19.
2. Zegada-Lizarazu, W.; Matteucci, D.; Monti, A. Critical review on energy balance of agricultural systems. Biofuels, Bioprod. Biorefin. **2010**, *4* (4), 423–446.
3. Pimentel, D.; Williamson, S.; Alexander, C.; Gonzalez-Pagan, O.; Kontak, C.; Mulkey, S. Reducing energy inputs in the US food system. Hum. Ecol. **2008**, *36,* 459–471.
4. Chen, G.; Baillie, C. Development of EnergyCalc—A tool to assess cotton on-farm energy uses. Energy Convers. Manage. **2009**, *50* (5), 1256–1263.
5. Safa, M.; Tabatabaeefar, A. Fuel consumption in wheat production in irrigated and dry land farming. World J. Agric. Sci. **2008**, *4* (1), 86–90.
6. *Pimentel, D. Impacts of Organic Farming on the Efficiency of Energy Use in Agriculture; The Organic Center, Cornell University, 2006.*
7. *Standards Australia. Australian/New Zealand Standards. AS/NZS 3598:2000. Energy Audits.*
8. Chen, G.; Baillie, C. Agricultural applications: Energy uses and audits. In *Encyclopedia of Energy Engineering and Technology*; Capehart, B., Ed.; Taylor & Francis Books: London, 2009; 1: 1, 1–5.
9. Hill, H. *Farm Energy Calculators: Tools for Saving Money on the Farm;* National Center for Appropriate Technology: USA. Available at http://attra.ncat.org/attra-pub/PDF/farmenergycalc.pdf.
10. Svejkovsky, C. *Conserving Fuel on the Farm;* National Center for Appropriate Technology: USA. Available at http://attra.ncat.org/attra-pub/PDF/consfuelfarm.pdf.
11. *A Guide to Energy Efficiency Innovation in Australian Wineries.* Available at http://www.ret.gov.au/energy/Documents/ best%20practice%20guides/energy_bpg_wineries.pdf.
12. Brown, E.; Elliot, R.N. *Potential Energy Efficiency Savings in the Agriculture Sector*; The American Council for an Energy-Efficient Economy: Washington, D.C., 2005. Available at http://www.aceee.org/pubs/ie053.htm.
13. Pellizzi, G.; Cavalchini, A.G.; Lazzari, M. *Energy Savings in Agricultural Machinery and Mechanization*; Elsevier Science Publishing Co.: New York, USA, 1988.
14. Stout, B.A. *Handbook of Energy for World Agriculture;* Elsevier Science Publications Ltd: London, 1990.
15. Adams, S. *Direct Energy Use in Agriculture—Opportunities for Reducing Fossil Fuel Inputs,* 2007. Available at http://randd.defra.gov.uk/Document.aspx?Document=AC0401_ 6343_FRP.pdf.
16. Wilhem, L.R.; Suter, D.A.; Brusewitz, G.H. Energy use in food processing (Ch. 11). In *Food and Process Engineering Technology;* ASAE: St. Joseph, Michigan, 2004; 285–291.
17. Eastop, T.D.; Croft, D.R. *Energy Efficiency for Engineers and Technologists*; Longman Publishing Group, 1990.
18. Kastens, T. *Farm Machinery Cost Calculations*; Kansas State University. Available at http://fieldcrop.msu.edu/documents/mf2244%20Farm%20Machinery%20Operation%20Costs.pdf.
19. Chen, G.; Kupke, P.; Baillie, C. Evaluating on-farm energy performance in agriculture. Aust. J. Multi-Discip. Eng. **2009**, 7 (1), 55–61.
20. Smith, L.A. Energy requirements for selected crop production implements. Soil Tillage Res. **1993**, 25, 281–299.
21. McLaughlin, N.B.; Drury, C.F.; Reynolds, W.D.; Yang, X.M.; Li, Y.X.; Welacky, T.W.; Stewart, G. Energy inputs for conservation and conventional primary tillage implements in a clay loam soil. Trans. ASABE. **2008**, *51* (4), 1153–1163.
22. Downs, H.W.; Hansen, R.W. *Estimating Farm Fuel Requirements,* Colorado State University, 2007. Available at http://www.ext.colostate.edu/PUBS/FARMMGT/05006. html#top.
23. Kosutic, S.; Filipovic, D.; Gospodaric, Z. Maize and winter wheat production with different soil tillage systems on silty loam. Agric. Food Sci. Finl. **2001**, *10* (2), 81–90.

24. Gulden, R.H.; Entz, H.E. *A Comparison of Two Manitoba Farms with Contrasting Tillage Systems*; University of Manitoba: Canada, 2005. Available at http://umanitoba.ca/outreach/naturalagriculture/articles/energy.html.
25. Bailey, A.P.; Basford, W.D.; Penlington, N.; Park, J.; Keatinge, J.D.H.; Rehman, T.; Tranter, R.B.; Yates, C. A comparison of energy use in conventional and integrated arable farming systems in the U.K. Agric. Ecosyst. Environ. **2003**, 97, 241–253.
26. Smith, I. Conservation or conventional tillage? Nat. Resour. Manage. **2001**, *4* (2), 9–15.
27. *White, B.; Ryan, W. A Scoping Study—Potential Investment in Agricultural Engineering Research and Development for the Australian Grains Industry; Kondinin Group: Australia, 2004.*
28. Holland, J.M. The environmental consequences of adopting conservation tillage in Europe: Reviewing the evidence. Agric. Ecosyst. Environ. **2004**, *103*, 1–125.
29. Tullberg, J.N.; Yule, D.F.; McGarry, D. Controlled traffic farming—From research to adoption in Australia. Soil Tillage Res. **2007**, 97, 272–281.
30. Tullberg, J.N. Traffic effects on tillage energy. J. Agric. Eng. Res. **2000**, *75* (4), 375–382.
31. Tullberg, J.N. Tillage, traffic and sustainability—A challenge for ISTRO. Soil Tillage Res. *in press.*
32. Tullberg, J.N.; Yule, D.F.; McGarry, D. *'On Track' for sustainable cropping in Australia*, International Soil Tillage Research Organisation Conference (Invited plenary paper). ISTRO Proceedings CD, University of Queensland, 2003.
33. Available at http://www.pollutionissues.com/A-Bo/Agriculture.html.
34. Jacobs, S. *Comparison of Life Cycle Energy Consumption of Alternative Irrigation Systems*; University of Southern Queensland, 2006. Available at http://eprints.usq.edu.au/2471/.
35. Lal, R. Carbon emission from farm operations. Environ. Int. **2004**, *30,* 981–990.
36. Jackson, T.M.; Khan, S.; Hafeez, M. A comparative analysis of water application and energy consumption at the irrigated field level. Agric. Water Manage. **2010**, *97,* 1477–1485.
37. Barron, G. *Centre Pivot Irrigation in the Riverina*; Prime-fact 98, NSW DPI: Orange, NSW, Australia. Available at http://www.dpi.nsw.gov.au/agriculture/resources/water/irrigation/systems/pressurised/centre-pivot-irrigation.
38. Pathak, B.S.; Bining, A.S. Energy use pattern and potential for energy saving in rice-wheat cultivation. Energy Agric. **1985**, 4, 271–278.
39. Morris, M.; Lynne, V. *Energy Saving Tips for Irrigators*; ATTRA Publication #IP278, 2006. Available at http://attra.ncat.org/attra-pub/energytips_irrig.html.
40. Tsatsarelis, C.A. Energy inputs and outputs for soft winter wheat production in Greece. Agric. Ecosyst. Environ. **1993**, *43* (2), 109–118.
41. Yilmaz, I; Akcaoz, H.; Ozkan, B. An analysis of energy use and input costs for cotton production in Turkey. Renewable Energy **2005**, *30,* 145–155.
42. Tsatsarelis, C.A. Energy requirements for cotton production in central Greece. J. Agric. Eng. Res. **1991**, *50,* 239–246.
43. *Pretty, J.N. Regenerating Agriculture: Policies and Practice for Sustainability and Self-reliance; Earthscan Publications Ltd.: London, 1995.*
44. Wells, C. *Total Energy Indicators of Agricultural Sustainability: Dairy Farming Case Study,* Final Report to MAF Policy; University of Otago: New Zealand, 2001.
45. Hatirli, S.A.; Ozkan, B.; Fert, C. Energy inputs and crop yield relationship in greenhouse tomato production. Renewable Energy **2006**, *31* (4), 427–438.
46. Chen, G.; Maraseni, T.; Y ang, Z. Life-cycle energy and carbon footprint assessments: Agricultural and food products. In *Encyclopedia of Energy Engineering and Technology*; Capehart, B., Ed.; Taylor & Francis Books: London, 2010; 1: 1, 1–5.
47. Hoeppner, J.W.; Entz, M.H.; McConkey, B.G.; Zentner, R.P.; Nagy, C.N. Energy use and efficiency in two Canadian organic and conventional crop production systems. Renewable Agric. Food Syst. **2006**, *21* (1), 60–67.

48. Yaldiz, O.; Ozturk, H.H.; Zeren, Y.; Bascetincelik, A. *Energy usage in production of field crops in Turkey,* 5th Int. Congress on Mechanization and Energy Use in Agriculture, Kusadasi, Turkey, 1993.
49. Ziesemer, J. *Energy use in organic food systems;* Food and Agriculture Organization of the United Nations. available at http://www.fao.org/docs/eims/upload/233069/energy-use-oa.pdf.
50. *Tribe, D. Sustaining the food supply of a developing world: Genetically modified crops enter their second decade; University of Melbourne.*
51. Available at http://www.hexima.com.au/pdf_files/articles/Support_grows_for_GM-Australian.pdf.
52. Available at http://www.usaid.gov/our_work/economic_growth_and_trade/energy/publications/empowering_agriculture.pdf.
53. Svejkovsky, C. *Renewable Energy Opportunities on the Farm, National Center for Appropriate Technology*; ATTRA Publication #IP304, USA. Available at http://attra.ncat.org/attra-pub/energyopp.html.
54. Fischer, J.R.; Johnson, S.R.; Finnell, J.A.; Price, R.P. Renewable energy technologies in agriculture: Solar, wind, geothermal, and anaerobic digestion. Resour. Mag. **2009**, *16* (3), 4–9.
55. Hansen, A.C.; He, B. Biodiesel fuels for off-road vehicles. In *Encyclopedia of Agricultural, Food, and Biological Engineering;* Heldman, D.R., Ed.; 2009; 1: 1, 1–3.
56. Fischer, J.R.; Buchanan, G.A.; Orbach, R.; Harnish III, R.L.; Jena, P. Renewable energy gains global momentum. Resour. Mag. **2008**, *15* (7), 9–11.
57. Available http://www.solar-estimate.org/index.php7verifycookie=1&page=estimatoroverview&subpage=.

3

Agriculture: Organic

Kathleen Delate

Introduction

The USDA National Organic Standards Board (NOSB) defines organic agriculture as "an ecological production management system that promotes and enhances biodiversity, biological cycles, and soil biological activity.[1] It is based on minimal use of off-farm inputs and on management practices that restore, maintain, or enhance ecological harmony. The primary goal of organic agriculture is to optimize the health and productivity of interdependent communities of soil life, plants, animals and people."[1] The term "organic" is defined by law. The labels "natural," "eco-friendly," and similar statements do not guarantee complete adherence to organic practices as defined by law.

History

In 1990, the U.S. Congress passed the Organic Food Production Act (OFPA). This law was heralded as the first U.S. law established to regulate a system of farming. The OFPA requires that anyone selling products as "organic" must follow a set of prescribed practices that include avoidance of synthetic chemicals in crop and livestock production, and in the manufacturing of processed products. "Certified organic" crops must be raised on land to which no synthetic chemical (any fertilizers, herbicides, insecticides, or fungicides) inputs were applied for three years prior to the crops' sale. Organic certification agencies became established in the United States to deal with a required "third-party certification." There are at least 20 private certification agencies and 15 state agencies certifying organic production and processing in the United States. Proposed rules implementing the federal OFPA law were promulgated in 1997, after seven years of revisions. Unfortunately, these rules did not meet private certification agencies' standards, and a record number of complaints (275,000) were issued in the public comment period. Now that the federal rules are established (released in 2001), all certifiers must utilize the federal standards as the minimum standard for the "certified organic" label in the United States. European regulation is under the auspices of the International Federation of Organic Agriculture Movements (IFOAM) with national certification agencies in each country.[2] Japan currently certifies under the Ministry of Agriculture and Forestry. Certification for the European Union and Japan is extended to several U.S. certifiers that meet international standards.

Organic agriculture is the oldest form of agriculture on earth. Farming without the use of petroleum-based chemicals (fertilizers and pesticides) was the norm for farmers in the developed world until post–World War II. The war era led to technologies that were adapted for agricultural production. Ammonium nitrate used for munitions during WWII evolved into ammonium nitrate fertilizer; organophosphate nerve gas production led to the development of powerful insecticides. These technical advances since WWII have resulted in significant economic benefits, as well as unwanted environmental and social effects. Organic farmers seek to utilize those advances that yield benefits (e.g., new varieties of crops, more efficient machinery) while discarding those methods that have led to negative impacts on society and the environment, such as pesticide pollution and insect pest resistance.[3] Instead of using synthetic fertilizers and pesticides, organic farmers utilize crop rotations, cover crops, and naturally based products to maintain or enhance soil fertility.[4,5] These farmers also rely on biological, cultural, and physical methods to limit pest expansion and increase populations of beneficial insects on their farms. By managing their ecological capital through efficient use of on-farm natural resources, organic farmers produce for diverse and specialized markets that provide premium prices.

Because genetically modified organisms (GMOs) constitute synthetic inputs and pose unknown risks, GMOs, such as herbicide-resistant seeds, plants, and product ingredients are disallowed in organic agriculture. Organic livestock, like organic crops, must be fed 100% organic food or feed in their production. Synthetic hormones and antibiotics are dis- allowed in organic livestock production. Traditional farmers throughout the world have relied on natural production methods for centuries, maintaining consistent yields within their local environment. While "green revolution" technologies have led to increased yields in many less developed countries, many farmers have seen an increase in pest problems with new varieties and high input–based systems.

Motivations for organic production include economic, food safety, and environmental concerns. All organic farmers avoid the use of synthetic chemicals in their farming systems, but philosophies differ among organic farmers regarding methods to achieve the ideal system. Organic farmers span the spectrum from those who completely eschew external inputs, create on-farm sources of compost for fertilization, and encourage the activity of beneficial insects through conservation of food and natural habitats, to those farmers who import their fertility and pest management inputs. A truly sustainable method of organic farming would seek to eliminate, as much as possible, reliance on external inputs.

Worldwide Statistics

USDA does not publish systematic reports on organic production in the United States. The most recent census in 1994 identified 1.5 million acres of organic production in the United States with 4050 farmers reporting organic acreage.[6] This figure underrepresents current production because many organic farmers opt to sell their products as organic without undergoing certification. The U.S. organic industry continues to grow at a rate of 20% annually. The industry was listed as a $4.5 billion industry in 1998, with predicted future growth to $10 billion by 2003. The organic industry is a consumer–driven market. According to industry surveys, the largest purchasers of organic products are young people and college–educated consumers. Worldwide consumption of organic products has experienced tremendous growth, often surpassing U.S. figures of 20% annual gain. Much of the increase in consumption worldwide has been fueled by consumers' demand for GMO–free products. Because GMOs are disallowed in organic production and processing, organic products are automatically segregated as GMO–free at the marketplace. European consumers have led the demand for organic products, particularly in countries such as the Netherlands and Scandinavia. Two percent of all German farmland, 4% of Italian farmland, and 10% of Austrian farmland is managed organically.[2] Prince Charles of England has developed a model organic farm and established a system of government support for transitioning organic farmers. Major supermarket chains and restaurants in Europe offer a wide variety of organic products in their aisles and on their menus. Industry experts predict that the establishment of federal rules will advance organic sales in the United States. Although the organic industry began as a niche market, steady growth has led to its place in a "segment" market since 1997. The organic dairy industry, for example, expanded by 73% from 1996 to 1997, and continues to grow today. Organic markets can be divided into indirect and direct markets. Indirect or wholesale markets include cooperatives, wholesale produce operations, brokers, and local milling operations. Many supermarket chains buy directly from farmers or from wholesalers of organic products. Because meat can now be labeled as "organic," as of 1999, the marketing of organic beef, pork, chicken, and lamb has been significantly simplified. Roadside stands, farmers' markets and community supported agriculture (CSA) farms constitute the direct marketing end of the organic industry. Most consumers relate their willingness to pay premium prices for food that has been raised without synthetic chemicals because of their concern for food safety and the environment. Supporting local family farmers also enters into their purchasing decisions.

Crop and Pest Performance in Organic Systems

The basis for all organic farming systems is the health of the soil.[4,7] In addition to maintaining adequate fertility, organic farmers strive for biologically active soil, containing microbial populations required for nutrient cycling.[8,9] Crop rotations (required for all organic operations) provide nutrients such as nitrogen in the case of legume crops (alfalfa and clover) and carbonaceous biomass upon which beneficial soil microorganisms depend for survival.[10,11] A crop rotation plan is required as protection against pest problems and soil deterioration.[12–15] Ideally, no more than four out of six years should be in agronomic crops, and the same row crop cannot be grown in consecutive years on the same land. Legumes (alfalfa, clovers, and vetches) alone, or in combination with small grains (wheat, oats, and barley), must be rotated with row crops (corn, soybeans, amaranth, vegetables, and herbs) to ensure a healthy system. A typical six–year rotation in the Midwestern United States would be corn (with a cover of winter rye)–soybeans–oats (with an underseeding of alfalfa)–alfalfa–corn–soybeans.[16,17] Horticultural crops must be rotated with a leguminous cover crop at least once every five years.

Pest management in organic farming systems is based on a healthy plant able to withstand some pest injury and on the inherent equilibrium in nature, as most insect pests have natural enemies that regulate their populations in unperturbed environments.[18,19] Because only naturally occurring materials are allowed in organic production, insect predators, parasites, and pathogens exist without intervention from highly toxic insecticides.[20] Most organic farmers rely on naturally occurring beneficial insects on their farms, but some farmers purchase and release lacewings and other natural enemies every season, for example. There are also commercial preparations of natural insect pathogens, such as *Bacillus thuringiensis* (Bt), which are used to manage pestiferous larvae, such as corn borers. Botanical insecticides, such as neem and ryania, are also allowable in organic production, but as with all insecticides, sprays should be used only as a last resort. Although these materials are naturally based, some materials may affect natural enemies. Prevention is a cornerstone of organic farming.[21] Pest–free seeds and transplants, along with physical and cultural methods, are used to prevent pest infestations. Physical methods include the use of row covers for protection against insects such as cabbage butterflies and aphids. Cultural methods include sanitation and resistant varieties. Plant varieties are used that have been bred traditionally (i.e., no manipulated gene insertion or engineering involved) for insect, disease, and nematode resistance or tolerance.

Most organic farmers rely on multiple tactics for their weed management.[22,23] Allelopathic crops, cultivation, mulching, and flame burning are all methods available for organic farmers. Allelopathic crops, such as rye and oats, produce an exudate that mitigates against small weed seed germination. Depending on the crop, cultivation offers the least labor–intensive method of organic weed management. Timely cultivation is key; without specific schedules, weeds proliferate. Propane flame burning is generally used in conjunction with cultivation, particularly during times of high field moisture. Mulching with straw or wood chips is commonly used in many organic horticultural operations.

Yields comparable to conventional crops have been shown for organic crops in three university long–term experiments in the United States (South Dakota State University,[24] Iowa State University,[16] and the University of California–Davis.[25] and in many European studies.[26] Factoring in an organic premium (ranging from 50 to 400%, depending upon crop and season), organic systems consistently out–performed conventional systems in terms of economics.[17,27,28] Pest problems were not a critical factor in these organic systems. Other studies have shown the benefits of organic practices, such as composting, in mitigating root–borne diseases.[8]

Key Issues Requiring Additional Research

Continued verification of the long–term benefits of organic versus conventional farming in terms of soil quality,[29] pest management, and nutritional benefits.[30] is needed. Key issues include the development of management practices to increase nutrient cycling for maintenance of crop yields and optimize biological control of plant pests and diseases[31] Economic analysis, including risks of the three–year transition required for organic certification, will provide useful information for growers interested in alternative systems.[26,32] Appropriate tillage systems, which protect soil quality and provide adequate soil preparation, remain as important issues for organic producers. The improvement of natural parasiticide formulations, such as diatomaceous earth, is required for optimum organic livestock production. Marketing and support needs include the availability of reliable statistics for organic operations and prices. Although many European countries support their farmers in their organic production practices through environmental subsidies,[33] the United States has made small gains in this area. Some state agencies (Minnesota Department of Agriculture) and the USDA Natural Resources Conservation Services (NRCS) through the Environmental Quality Indicators Program (EQIP) offer financial incentives to organic farmers during their transitioning years. More of these support services are needed to encourage farmers interested in the conversion to alternative production.[34]

References

1. National Organic Program USDA Agricultural Marketing Service Washington, DC, 1. http://www.ams.usda.gov/nop/ (accessed June 5, 2000).
2. Lampkin N.H.; *The Policy and Regulatory Environment for Organic Farming in Europe*; Universität Hohenheim, Institut für Landwirtschaftliche Betriebslehre: Stuttgart, Germany, 1999; 1–379.
3. Altieri, M. *Agroecology*; Second Ed., Westview Press: Boulder, CO, 1995; 1–433.
4. Lampkin, N.H.; Measures, M. *1999 Organic Farm Management Handbook*, 3rd Ed. Welsh Institute of Rural Studies, University of Wales, Hamstead Marshall, Berkshire, 1999; 1–163.
5. Lockeretz, W.; Shearer, G.; Kohl, D. Organic farming in the corn belt. *Science,* **1981**, *211*, 540–547.
6. *Lipson, M. Searching for the "O–Word": analyzing the USDA current research information system for pertinence to organic farming; Organic Farming Research Foundation: Santa Cruz, CA, 1997; 1–85.*
7. Macey, A.; Kramer, D. *Organic Field Crop Handbook*; Canadian Organic Growers, Inc.: Ottawa, Canada, 1992; 1–256.
8. Drinkwater, L.E.; Letourneau, D.K.; Shennan, C. Fundamental differences between conventional and organic tomato agroecosystems in California. Ecological Appl. **1995**, 5 (4), 1098–1112.
9. Wander, M.M.; Traina, S.J.; Stinner, B.R.; Peters, S.E. Organic and conventional management effects on biologically active soil organic matter pools. Soil Sci. Soc. Am. J. **1994**, *58*, 1130–1139.
10. Drinkwater, L.E.; Wagoner, P.; Sarrantonio M. Legume–based cropping systems have reduced carbon and nitrogen losses. Nature **1998**, *396*, 262–265.
11. Yeates, G.W.; Bardgett, R.D.; Cook, R.; Hobbs, P.J.; Bowling, P.J.; Potter, J.F. Faunal and microbial diversity in three welsh grassland soils under conventional and organic management regimes. J. Appl. Ecol. **1997**, *34* (3) 453–470.
12. Adee, E.A.; Oplinger, E.S.; Grau, C.R. Tillage, rotation sequence, and cultivar influences on brown stem rot and soybean yield. J. Prod. Agric. **1994**, *7* (3), 341–347.
13. Karlen, D.L.; Varvel, G.E.; Bullock, D.G.; Cruse, R.M. Crop rotations for the 21st century. Adv. in Agron. **1994**, *53*, 1–45.

14. Lipps, P.E.; Deep, I.W. Influence of tillage and crop rotation on yield, stalk rot, and recovery of Fusarium and Tricho- derma spp. from corn. Plant Dis. **1991**, *75* (8), 828–833.
15. Stinner, D.H.; Stinner B.R.; Zaborski, E.R.; Favretto M.R.; McCartney, D.A. *Ecological Analyses of Ohio Farms under Long–term Sustainable Management*; Ecological Society of America, August: Knoxville, TN, 1994; 7–11.
16. Delate, K. *Comparison of Organic and Conventional Rotations at the Neely–Kinyon Long–Term Agroecological Research (LTAR) Site*; Leopold Center for Sustainable Agriculture Annual Report, Iowa State University: Ames, IA, 1999; 1–12.
17. Welsh, R. *The Economics of Organic Grain and Soybean Production on the Midwestern United States*; Henry, A., Ed.; Wallace Institute for Alternative Agriculture: Green- belt, MD, 1999.
18. Neher, D.A. Nematode communities in organically and conventionally managed agricultural soils. J. Nematol. **1999**, *31*, 142–154.
19. Pfiffner, L.; Niggli, U. Effects of bio-dynamic, organic and conventional farming on ground beetles and other epigaeic arthropods in winter wheat. Biol. Agri. Hort. **1996**, *12*, 353–364.
20. Kromp, B.; Meindel, P.; Harris, P.J.C. *Entomological Research in Organic Agriculture*; Academic Publishers: Bicester, England, 1999, 1–386.
21. *Lockeretz, W. Environmentally Sound Agriculture: Selected Papers from the Fourth International Conference of the International Federation of Organic Agriculture Movements (IFOAM)**;** Praeger: New York, 1983; 1–426, Cambridge, MA, Aug 18–20, 1982.*
22. Lanini, W.T.; Zalom, F.; Marois, J.; Ferris, H. Researchers find short–term insect problems, long–term weed problems. Ca. Agric. **1994**, *48*, 27–33.
23. Liebman, M.; Ohno, T. Crop Rotation and Legume Residue Effects on Weed Emergence and Growth: Applications for Weed Management. In *Integrated Weed and Soil Management*; Hatfield, J.L.; Buhler, D.D.; Stewart, B.A., Eds.; Ann Arbor Press: Chelsea, MI, 1997; 181–221.
24. Dobbs, T. Report on Organic and Conventional Grain Trials at South Dakota State University. *USDA–ERS Conference on The Economics of Organic Farming Systems,* USDA: Washington, DC, 1999; 1–4.
25. Klonsky, K.; Livingston, P. Alternative systems aim to reduce inputs, maintain profits. Ca. Agric. **1994**, *48* (5), 34–42.
26. Lampkin, N.H.; Padel, S. *The Economics of Organic Farming: An International Perspective*; CAB International: Wallingford, Oxon, U.K., 1994; 1–468.
27. Hanson, J.C.; Lichtenberg, E.; Peters, S.E. Organic versus conventional grain production in the Mid-Atlantic: an economic and farming system overview. Am. J. Alt. Agric. **1997**, *12*, 2–9.
28. Stanhill, G. The comparative productivity of organic agriculture. Agric. Ecosys. Env. **1990**, *30*, 1–26.
29. Lockeretz, W.; Shearer, G.; Klepper, R.; Sweeney, S. Field crop production on organic farms in the midwest. J. Soil Water Conserv. **1978**, 33, 130–134.
30. Woese, K.; Lange, D. Boess, C.; Bögl, K.W. A comparison of organically and conventionally grown foods—results of a review of the relevant literature. J. Sci. Food Agric. **1997**, *74*, 281–293.
31. Walz, E. *Final Results of the Third Biennial National Organic Farmers' Survey*; Organic Farming Research Foundation: Santa Cruz, CA, 1999; 1–75.
32. Chase, C.; Duffy, M. An economic comparison of conventional and reduced-chemical farming systems in Iowa. Amer. J. Alt. Agric. **1991**, *6* (4), 168–73.
33. Zygmont J. International Issues Pertaining to Organic Agriculture, Proceedings of the Workshop: Organic Farming and Marketing Research—New Partnerships and Priorities, Lipson, M., Hammer, T., Eds., Organic Farming Research Foundation: Santa Cruz, CA, 1999; 317–379.
34. D'Souza G.; Cyphers, D.; Phipps, T. Factors affecting the adoption of sustainable agricultural practices. Agric. and Resource Econ. Rev. **1993**, *22* (2), 159–205.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

4

Erosion by Water: Accelerated

David Favis-Mortlock

Introduction

Soil erosion by water occurs when rain falls onto, or runoff flows over, the surface of the earth under the influence of gravity. The kinetic energy possessed by this water may be sufficient to give it the ability to remove and redistribute the "cream of the soil," its nutrient-rich upper layers. The first section of this entry focuses on the physical processes of water erosion. There are several, and they all interact in a complex way. The results are erosional features that span a wide range of spatial and temporal scales: these are the focus of the second section.

Erosion by water is a natural phenomenon: this is "background" or "geological" erosion. However, if the soil's protective cover of vegetation is removed owing to unwise agricultural practices, overgrazing, or poorly executed construction activity, then erosion rates can increase greatly. The result is accelerated water erosion, which is a present-day problem over much of the earth's surface. Mankind's modification of natural water erosion is the focus of the third section. The impacts of accelerated water erosion may be both on-site (at the place where the erosion occurs) and off-site (downstream from this): the fourth section describes these impacts. The fifth and final section discusses future soil erosion by water. This will be affected by anthropogenic climate change: both directly, owing to changes in the amount, timing, and intensity of rainfall, and the amount and timing of snowmelt-generated runoff; and indirectly, owing to climate-driven shifts in vegetation growth and land usage.

Processes of Erosion by Water

Water erosion is driven by the energy of rainfall and flowing water. When rain falls upon unprotected soil, its kinetic energy may be sufficient to detach soil particles: this is one of the subprocesses[1] of water erosion (RD-ST in Figure 1), and is often referred to as "rainsplash erosion" or "splash erosion."

A more accurate term is "rainsplash redistribution," since although considerable quantities of soil can be moved by splash, almost all of it is merely redistributed, moved only a few millimeters or centimeters on the soil's surface. There will be a net downslope movement of splashed soil, but this is generally rather small. Rainsplash redistribution is most effective where rainfall intensities are high,[2] e.g., as a result of convective rainstorms in the world's equatorial regions. Low-intensity rainfall is often of frontal origin: where such rainfall is common (e.g., temperate northwest Europe), rainsplash is a relatively ineffective subprocess of erosion.

Some of the rain may infiltrate into (i.e., be absorbed by) the body of the soil. In dry conditions, all of the rain may infiltrate: the result is then rain with no runoff (Figure 1). Conversely, if the soil is fully saturated, none of the rainfall will infiltrate: all of it becomes either runoff (which is also known as "overland flow") or detention storage (see below). If the soil is fully saturated, this will occur irrespective of the rain's intensity (i.e., the rate at which it falls). Runoff that results from the soil being fully saturated is known as "saturation excess overland flow." It is preferentially generated from the lower parts of a catchment, since the soil in these areas is likely to become saturated first.

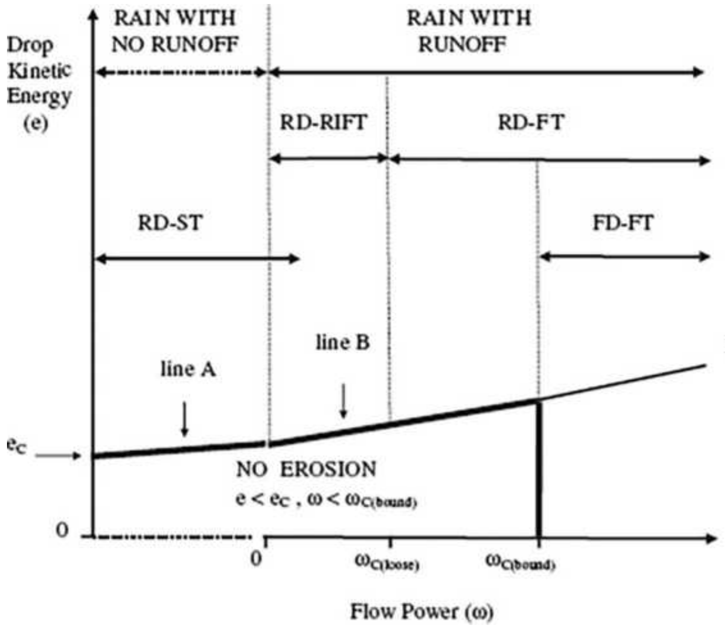


FIGURE 1 Detachment and transport processes associated with variations in raindrop and overland flow energies. The *Y* axis represents raindrop kinetic energy *e*, which is a function of rainfall intensity. The *X* axis represents overland flow power *ω*. The leftmost zero on the *X* axis (at the beginning of the dotted section) indicates the start of rainfall, and the zero to its right (at the rightmost end of the dotted section of the *X* axis) represents the beginning of overland flow. During a rainstorm, flow power generally increases with time provided sufficient rain, with a sufficiently high intensity, has fallen and produced ponding and runoff. Thus, the *X* axis is a rough surrogate for time during a rainstorm. Horizontal left-to-right line traverses of the diagram indicate the changes of detachment and transportation processes that occur during a rainstorm, for a given constant rainfall intensity. Non-horizontal left-to-right traverses incorporate some changes in rainfall intensity. e_c, critical raindrop energy to cause erosion; line A, raindrop energy before flow (this often increases slightly over time, as a soil crust develops); line B, raindrop energy when flow occurs (increases over time, as more drop energy is used in penetrating the deepening overland flow); $\omega_{\mathrm{c(loose)}}$, critical flow power for transporting loose soil material; $\omega_{\mathrm{c(bound)}}$, critical flow power for detaching soil from the surface of the soil matrix; RD-ST, raindrop detachment, splash transport; RD-RIFT, raindrop detachment, raindrop-induced flow transport; RD-FT, raindrop detachment, flow transport; FD-FT, flow detachment, flow transport.
Source: Kinnell.[1]

High-intensity rainfall (i.e., a great deal of rain arriving during a short period) can give rise to runoff of a different kind. This is "infiltration excess overland flow" (or "Hortonian overland flow"), and is the result of rain arriving too quickly to infiltrate into the soil. The pre-existing moisture content of the soil is of lesser importance in determining the occurrence of infiltration excess overland flow, compared with saturation excess overland flow. As a result, infiltration excess overland flow tends to be more spatially uniform, often being generated over a whole catchment (or at least that portion of it which is receiving rain from a single storm). However, a major factor that influences the occurrence of infiltration excess overland flow is the state of the soil's surface, in particular any crusting or capping that may have occurred. Crusting and capping can be the result of physical, chemical, or biological agencies, and is not necessarily spatially uniform.

While saturation excess overland flow and infiltration excess overland flow are distinct, they commonly interact spatially and temporally. For example, during a single rainstorm, saturation excess overland flow may be the first to occur, beginning in the topographic lows of a catchment, with the contributing area gradually increasing as a great area of soil within the catchment becomes saturated. A subsequent increase in rainfall intensity during the storm may result in infiltration excess overland flow being generated over the whole area of the catchment. A later decrease in intensity may result in a return to predominantly saturation excess overland flow, originating from only part of the catchment's surface area.

In certain circumstances, topographically lower areas within a catchment may experience subsurface return flow. This is water that has previously infiltrated in some upslope location, moved laterally downslope within the soil, and then exfiltrated to once again become (or join) hillslope runoff.

However it is generated, runoff tends to flow downhill under the action of gravity. This thin diffuse film of water has lost virtually all the kinetic energy that it possessed as falling rain; thus, it moves only slowly, has a low flow power, and is generally incapable of detaching or transporting soil particles ("no erosion" in Figure 1).

The microtopography (i.e., small-scale pattern of irregularities) of the soil's surface tends to cause this overland flow to concentrate in closed depressions, which slowly fill: this is known as "detention storage" or "ponding." Both the flowing water and the water in detention storage protect the soil from raindrop impact, so that rainsplash redistribution usually decreases over time within a storm, as the depth of surface water increases. There are, however, complex interactions between rainsplash and overland flow (Figure 1).

If the rain continues, the increasing depth of water will eventually overtop the depressions. Overland flow that is released in this way is likely to flow downhill more quickly and in greater quantities (i.e., will possess more flow power as a result of its kinetic energy). Thus, it will begin to be able to transport soil particles that have already been detached (RD-RIFT, RD-FT in Figure 1). If its velocity continues to increase, eventually the overland flow will be able to detach and transport soil particles from the soil matrix (FD-FT in Figure 1). Where it does so, the soil's surface will be lowered slightly. Lowered areas form preferential flow paths for subsequent flow, and these flow paths are in turn eroded further. Eventually, this positive feedback[3] results in small, well-defined linear concentrations of overland flow ("microrills" or "traces").

In many cases, individual microrills become ineffective over time because of sedimentation. A subset, however, grow further to become rills,[4] and a smaller subset may go on to develop into gullies. This process of "competition" between microrills and rills leads to the self-organized formation[13] of networks of erosional channels (dendritic on natural soil surfaces[5]; constrained by the direction of tillage on agricultural soils[6]), which form efficient pathways for the removal of water from hillslopes. It is in such erosional channels that water erosion also operates most effectively to detach and remove soil by its kinetic energy.[7] In most situations, erosion by concentrated flow is the main agent of erosion by water. The flow-dominated erosional channels are separated by interrill areas[8] where the dominant processes are rainsplash and diffuse overland flow; however, boundaries between rill and interrill areas are both ill defined and constantly shifting.

In some circumstances, subsurface flow may be important in determining where channel erosion will begin and develop (e.g., at the base of slopes,[9] and in areas of very deep soils such as tropical saprolites[10]). Meltwater from thawing snow operates in a broadly similar way to rain-derived overland flow,[11] detaching and transporting unfrozen soil in areas of concentrated flow; however, snowmelt erosion is less well studied and less well understood.

As erosional channels increase in size (i.e., become large rills and gullies[12]), processes such as gravitational collapse of channel walls and heads increase in importance.[13] Runoff and sediment from rills and gullies may be moved into ditches, stream and rivers, and so transported well away from the point of origin. However, sediment may also be deposited within the rill or gully, or beyond the rill or gully's confines in a depositional fan, at locations where the gradient slackens. Here, it may be stored for a variable period of time,[14] possibly being reworked by tillage activity, until a subsequent erosion event is of sufficient size to re-erode the stored sediment. It may then be redeposited further downstream, or make its way into a permanent watercourse and thence to lake or ocean.

Erosion by Water: Spatial and Temporal Scale

The complex hierarchy of subprocesses of water erosion (Figure 1) means that erosion by water operates (and is studied) over a wide range of spatial scales.[15] Rainsplash redistribution and the initiation of microrills and rills occur at a scale of millimeters (Figure 2). Rill erosion on agricultural hillslopes operates at a scale of meters to tens of meters (Figure 3), while gully erosion can occur on a scale of hundreds of meters, or even kilometers. The offsite impacts of erosion can affect very large areas, sometimes hundreds or even thousands of square kilometers (Figure 4).

At every spatial scale, however, erosion is highly patchy. Even in areas of severe erosion, rates of soil loss can vary greatly from point to point on the landscape as the vagaries of topography and land use influence erosive flows over a wide range of spatial scales.[16] Obvious erosion in one field can be found side by side with virtually untouched areas, and within an eroded field, the severity of erosion can vary markedly.

Often, erosion by water also varies across a range of temporal scales. Soil loss from water erosion occurs both incrementally, as a result of many small rainfall events, and more dramatically, as a result of large but relatively rare storms. Dramatic erosion events can produce large gullies, and create flooding and property damage that hits the news headlines. However, a significant proportion of total soil loss may be due to small but frequent erosion events, which nonetheless have a notable cumulative impact.[17]

FIGURE 2 Impacts of erosion by water at the microscale: overland flow between millimeter-scale soil aggregates. The finger is pointing at an area where concentrated flow is just beginning to incise a microrill.

Source: A rainfall simulation experiment by A.J.T. Guerra and D.T. Favis-Mortlock, 1997; unpublished (by permission).

FIGURE 3 Impacts of erosion by water at the field scale: a large rill on agricultural land in Germany.
Source: Katharina Helming, 2001 (by permission).

FIGURE 4 Impacts of erosion by water at the global scale: a sediment plume covering tens of thousands of square kilometers at the mouth of the Yellow River, China. This results from erosion by water on the Chinese Loess Plateau.
Source: NASA, 2000.

Global Problem of Accelerated Erosion by Water

While soil erosion by water has been occurring naturally for some 450 million years (since the first land plants formed the first soil), accelerated erosion is of much more recent origin. Yet, on a human timescale, accelerated erosion is old. There is considerable archaeological evidence from many parts of the world that accelerated erosion by water is often associated with early agriculture.[18,19] The association of water erosion with unwise agricultural practices was first noted within a scientific context during the second and third decades of the 20th century by pioneers of soil conservation such as Hugh Hammond Bennett in the United States, and subsequently by workers in other parts of the globe. During the period of colonialism, the imposed adoption of European agricultural methods frequently led to accelerated erosion in developing countries, a problem that continues to the present day.[20,21] During the last few decades of the 20th century, a move toward intensive agricultural technologies that leave the soil bare during times of heavy rainfall meant that previously problem-free areas of the world, such as northwest Europe, began to experience notable increases in water erosion.[22]

Yet despite the global nature of the problem of erosion by water, even today we do not have good information regarding the global extent of erosion by water.[23,24] Data on the severity of erosion are also often limited.[25]

Impacts of Erosion by Water

These can be categorized into on-site and off-site problems.

The removal of the upper horizons of the soil by erosion results in an on-site reduction in soil quality,[26] i.e., a diminution of the soil's suitability for agriculture or other vegetation. This is because the eroded upper horizons are usually the most nutrient rich. Also, because the finest constituents of eroded soil tend to be transported furthest, eroded soils become preferentially depleted of their finer fraction over time, which often reduces their waterholding capacity.[26] Increased use of artificial fertilizers may to an extent, and for a time, compensate for erosion- induced loss of soil quality where economic circumstances are favorable. This is not usually feasible in developing countries, however.[20,21] Loss of soil quality is a long-term problem; globally, it is the most serious impact of water erosion.[27,28]

Movement of sediment and associated agricultural pollutants into watercourses is the major off-site impact resulting from erosion. This leads to silting up of dams, alteration of the water and sediment conveyance capacity of streams and rivers, disruption of the ecosystems of lakes, and contamination of drinking water.[22] Rates of erosion do not have to be high for significant quantities of agricultural pollutants to be transported off-site.[29] This is a shorter-term impact than loss of soil quality; in the more affluent areas of the world, it can be the main driver for present-day soil conservation policy initiatives.[24] Another off-site effect can occur in situations where eroded soil has a decreased capacity to absorb water: increased runoff may lead to downstream flooding and local damage to property.[30]

Future of Soil Erosion by Water

Anthropogenic modification of global climate[31] will modify the future severity and impacts of water erosion. One way of categorizing the effects of future climate change on soil erosion by water is as follows.

- Direct effects: changes in rainfall amount, timing, and intensity, and changes in the amount and timing of snowmelt-generated runoff
- Indirect effects: climate-driven changes in vegetation and land usage

Direct Effects of Future Climate Change

The most direct route by which future climate change will influence future rates and extent of water erosion is by modifying the erosive power of rainfall and runoff.[32,33] The moisture-holding capacity of the atmosphere increases as global temperature increases. This leads to more water vapor in the atmosphere, and thus a more vigorous hydrological cycle, promoting a trend toward more intense precipitation events.[34] This trend is distinct from change in future rainfall amount: it is possible for rainstorm intensities to increase even if total annual precipitation decreases. Climate models predict a continued increase in intense precipitation events during the remainder of the 21st century:[31] more intense rainfall will result in an increase in runoff due to the infiltration excess mechanism, as described earlier. The result will be an increase in erosion and problems associated with erosion.

Amounts and timing of rainfall and snowfall are also predicted to change in many locations globally, but this is less amenable to generalization. Any increased rainfall during times when tillage leaves the soil bare will certainly lead to more severe erosion problems.

Indirect Effects of Future Climate Change

Future climate change will also indirectly affect soil erosion by water by modifying the protection that plants give to the soil. Changes in temperature and rainfall will influence the rate at which existing vegetation (both natural and agricultural) will grow. The complex changes in plant biomass resulting from changes in climate have the potential both to increase erosion rates through faster residue decomposition from increased microbial activity,[35] and also to decrease erosion rates through an increase in soil surface canopy cover and ground cover.[36]

Changes in climate will provide relatively affluent farmers with opportunities to modify tillage practices and timing for crops that are currently grown. It will also provide them with opportunities to introduce new crops. Significant changes in plant biomass, and the protection which it offers to the soil, will occur as a result of such shifts in agricultural practice and land use.[35,37] For example, the introduction to northwest Europe of new crops suited to warmer conditions, such as maize and sunflowers, increases the risk of erosion since both take a significant amount of time to provide adequate crop cover in early summer.[38] Also, the thermal altitudinal limit for agriculture increases as air temperature increases.[39] This could enable farmers to replace permanent grasslands in upland areas with grass leys or cereals; however, such practices have the potential to significantly increase erosion rates, as upland areas often have high rainfall and areas of steep slopes.[38] Farmers in less affluent situations, however, will have fewer options.

Conclusion

Rainfall and runoff can give rise to soil erosion by water: this is an easily learned truism of school geography lessons. However, it is an oversimplification. A closer look shows that water erosion is much more complicated. The many interacting subprocesses of erosion mean that water erosion is still only imperfectly understood, despite almost a century of scientific research. This is unsurprising, since "There are three great frontiers in science: the very big, the very small, and the very complex."[40] Water erosion is very complex, and we still have much to learn about it. This knowledge deficit has many practical implications: for example, computer models that aim to predict rates and extent of water erosion are, after 50 years of development, still decidedly imperfect.[41]

Nonetheless, accelerated soil erosion by water, viewed from still another perspective, is very simple. It is always the result of human actions such as unwise agricultural practices, overgrazing, or poorly executed construction activity. Removal of protective vegetation at times of heavy rainfall inevitably increases the risk of water erosion. The challenge for farmers, for example, is to implement agricultural practices that do not leave the soil vulnerable, while retaining acceptable levels of productivity.

This is not at all simple, and is likely to become an even more challenging task under the influence of future anthropogenic climate change.

Acknowledgments

I would like to thank an anonymous reviewer of an earlier draft for greatly improving this entry.

References

1. Kinnell, P.I.A. Raindrop-impact-induced erosion processes and prediction: A review. Hydrol. Process. **2005**, *19,* 2815–2844.
2. Nearing, M.A.; Bradford, J.M.; Holtz, R.D. Measurement of waterdrop impact pressures on soil surfaces. Soil Sci. Soc. Am. J. **1987**, *51* (5), 1302–1306.
3. Favis-Mortlock, D.T.; Boardman, J.; Parsons, A.J.; Las- celles, B. Emergence and erosion: A model for rill initiation and development. Hydrol. Process. **2000**, *14* (11–12), 2173–2205.
4. Slattery, M.C.; Bryan, R.B. Hydraulic conditions for rill incision under simulated rainfall: A laboratory experiment. Earth Surf. Process. Landforms **1992**, *17,* 127–146.
5. Helming, K.; Roth, C.H.; Wolf, R.; Diestel, H. Characterization of rainfall-microrelief interactions with runoff using parameters derived from Digital Elevation Models (DEMs). Soil Technol. **1993**, *6,* 273–286.
6. Planchon, O.; Esteves, M.; Silvera, N.; Lapetite, J.-M. Raindrop erosion of tillage-induced microrelief: Possible use of the diffusion equation. Soil Tillage Res. **2000**, *15–16,* 1–14.
7. Nearing, M.A.; Norton, L.D.; Bulgakov, D.A.; Larionov, G.A.; West, L.T.; Dontsova, K. Hydraulics and erosion in eroding rills. Water Resour. Res. **1997**, *33* (4), 865–876.
8. Govers, G.; Poesen, J. Assessment of the interrill and rill contributions to total soil loss from an upland field plot. Geomorphology **1988**, *1,* 343–354.
9. Huang, C.; Laflen, J.M. Seepage and soil erosion for a clay loam soil. Soil Sci. Soc. Am. J. **1996**, *60* (2), 408–416.
10. Fernandes, N.F.; Coelho Netto, A.L. Subsurface hydrology of layered colluvium mantles in unchannelled valleys— South-eastern Brazil. Earth Surf. Process. Landforms **1994**, *19,* 609–626.
11. Sharratt, B.S.; Lindstrom, M.J.; Benoit, G.R.; Young, R.A.; Wilts, A. Runoff and soil erosion during spring thaw in the northern U.S. corn belt. J. Soil Water Conserv. **2000**, *55* (4), 487–494.
12. Poesen, J.; Vandaele, K.; van Wesemael, B. Gully erosion: Importance and model implications. In *Modelling Soil Erosion by Water;* Boardman, J., Favis-Mortlock, D.T., Eds.; NATO-ASI Series I-55; Springer-Verlag: Berlin, 1998; 285–312.
13. De Ploey, J. A model for headcut retreat in rills and gullies. Catena Suppl. **1989**, *14,* 81–86.
14. Trimble, S.W. A sediment budget for Coon Creek basin in the driftless area, Wisconsin, 1853–1977. Am. J. Sci. **1983**, *283,* 454–474.
15. Kirkby, M.J.; Imeson, A.C.; Bergkamp, G.; Cammerat, L.H. Scaling up processes and models from the field plot to the watershed and regional areas. J. Soil Water Conserv. **1996**, *51* (5), 391–396.
16. Favis-Mortlock, D.T.; Boardman, J.; MacMillan, V.J. The limits of erosion modeling: Why we should proceed with care. In *Landscape Erosion and Evolution Modeling*; Harmon, R., Doe, W.W., III, Eds; Kluwer: New York, 2001; 477–516.
17. Boardman, J.; Favis-Mortlock, D.T. Frequency-magnitude distributions for soil erosion, runoff and rainfall—A comparative analysis. Z. Geomorphol. Suppl. **1999**, *115,* 51–70.
18. Bell, M.G.; Boardman, J., Eds. *Past and Present Soil Erosion*; Oxbow Monograph 22; Oxbow Books: Oxford, U.K., 1992.
19. O'Hara, S.L.; Street-Perrott, F.A.; Burt, T.P. Accelerated soil erosion around a Mexican highland lake caused by preHispanic agriculture. Nature **1993**, *362,* 48–51.

20. Hallsworth, E.G. *Anatomy, Physiology and Psychology of Soil Erosion*; Wiley-Interscience: New York, 1987.
21. Stocking, M. Soil erosion in developing countries: Where geomorphology fears to tread! Catena **1995**, *25* (1–4), 253–267.
22. Evans, R. *Soil Erosion and its Impacts in England and Wales*; Friends of the Earth: London, 1996.
23. United Nations Environment Program; *Global Land Degradation* map; available at http://maps.grida.no/go/graphic/degraded-soils (accessed March 2011).
24. European Environment Agency. *Down to Earth: Soil Degradation and Sustainable Development in Europe*; Office for Official Publications of the European Communities: Luxembourg, 2000; available at http://www.eea.europa.eu/publications/Environmental_issue_series_16; (accessed March 2011).
25. Boardman, J. An average soil erosion rate for Europe: Myth or reality? J. Soil Water Conserv. 1998, *53* (1), 46–50.
26. Lal, R. Monitoring soil erosion's impact on crop productivity. In *Soil Erosion Research Methods;* Lal, R., Ed.; Soil and Water Conservation Society: Ankeny, IA, 1989.
27. Larson, W.E.; Pierce, F.J.; Dowdy, R.H. The threat of soil erosion to long-term crop production. Science **1983**, *219*, 458–465.
28. Pimentel, D.; Allen, J.; Beers, A.; Guinand, L.; Hawkins, A.; Linder, R.; McLaughlin, P.; Meer, B.; Musonda, D.; Perdue, D.; Poisson, S.; Salazar, R.; Siebert, S.; Stoner, K. Soil erosion and agricultural productivity. In *World Soil Erosion and Conservation*; Pimentel, D., Ed.; Cambridge University Press: Cambridge, U.K., 1993; 277–292.
29. Harrod, T.R. Runoff, soil erosion and pesticide pollution in Cornwall. In *Conserving Soil Resources: European Perspectives;* Rickson, R.J., Ed.; CAB International: Wallingford, U.K., 1994; 105–115.
30. Boardman, J.; Ligneau, L.; de Roo, A.; Vandaele, K. Flooding of property by runoff from agricultural land in northwestern Europe. Geomorphology 1994, *10,* 183–196.
31. Solomon, S.; Qin, D.; Manning, M.; Chen, Z.; Marquis, M.; Averyt, K.B.; Tignor, M.; Miller, H.L., Eds; *Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, 2007*; Cambridge University Press: Cambridge, U.K., 2007.
32. Pruski, F.F.; Nearing, M.A. Climate-induced changes in erosion during the 21st century for eight U.S. locations. Water Resour. Res. **2002**, *38* (12), article no. 1298.
33. Soil and Water Conservation Society. *Conservation Implications of Climate Change: Soil Erosion and Runoff from Cropland;* SWCS: Ankeny, IA, 2003; available at http://www.swcs.org/en/publications/conservation_implications_of_climate_change/ (accessed March 2011).
34. Trenberth, K.E.; Dai, A.; Rasmussen, R.M; Parsons, D.B. The changing character of precipitation. Bull. Am. Meteorol. Soc. **2003**, *84* (9), 1205–1217.
35. Nearing, M.A.; Jetten, V.; Baffaut, C.; Cerdan, O.; Couturier, A.; Hernandez, M.; Le Bissonnais, Y.; Nichols, M.H.; Nunes, J.P.; Renschler, C.S.; Souchère, V.; van Oost, K. Modelling response of soil erosion and runoff to changes in precipitation and cover. Catena **2005**, *61,* 131–154.
36. Rosenzweig, C.; Hillel, D. *Climate Change and the Global Harvest. Potential Impacts of the Greenhouse Effect on Agriculture.* Oxford University Press: New York, 1998.
37. Williams, J.; Nearing, M.A.; Nicks, A.; Skidmore, E.; Valentine, C.; King, K.; Savabi, R. Using soil erosion models for global change studies. J. Soil Water Conserv. **1996**, *51* (5), 381–385.
38. Boardman, J.; Favis-Mortlock, D.T. Climate change and soil erosion in Britain. Geogr. J. **1993**, *159* (2), 179–183.
39. Hulme, M.; Hossell, J.E.; Parry, M.L. Climate change and land use in the U.K. Geogr. J. **1993**, *159,* 131–147.
40. Rees, M. *Our Cosmic Habitat*; Weidenfeld and Nicolson: London, 2002.
41. Morgan, R.C.P.; Nearing, M.A., Eds. *Handbook of Erosion Modelling;* Wiley: Chichester, U.K., 2011.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

5

Erosion: Irrigation-Induced

Robert E. Sojka and
David L. Bjorneberg

Introduction

Erosion is the greatest threat to agricultural sustainability. Most irrigation is on fragile arid soils that have enormous crop yield potential when irrigated. However, the yield potential is easily lost if the thin veneer of "topsoil" is eroded.[1] Erosion prevention on irrigated land is arguably more important than on rainfed land. Yields from irrigated land are more than double than those from nonirrigated land, with nearly triple the crop value per hectare.[2] In addition, runoff and irrigation return flows (necessary in many surface irrigation schemes) deliver sediment; human, animal and plant pathogens; nutrients and pesticides to downstream fields and riparian waters. These pollutants accumulate in runoff primarily as a result of erosion.

Irrigation's Unique Erosion Characteristics

Irrigation-induced erosion and rainfall-induced erosion result from the same physical and chemical processes. However, the processes come together and interact very differently in each case.[2–4] The magnitude of the differences depends upon the type of irrigation system and on soil and water properties. Briefly, the most important differences stem from soil and water chemistry, wetting rate, water application and infiltration patterns, and, for surface irrigation, absence of water drop impact. These factors are the basis for many erosion control practices unique to irrigation.[4–7] Since 1990, advances in irrigation erosion control have resulted from improved understanding of water quality and antecedent soil condition effects on erosion and from development of polyacrylamide (PAM) use.

The key to controlling erosion is controlling runoff. Runoff is controlled in two ways. It is minimized by scheduling irrigation to meet, but not exceed, crop water and salt leaching requirements (i.e. avoid over-irrigation), and it is managed by using application rates during each scheduled irrigation that minimize runoff and erosion for that event. Systems should be designed and operated to minimize over-irrigating of some areas in order to adequately irrigate others. In furrow irrigation this is accomplished by reducing the length of furrows; managing inflow rates and advance times, and where possible, cutting back inflow rates once runoff begins; or through use of surge irrigation (surge irrigation sometimes erodes near the inlet because of higher flows during initial pulsing of water). Sprinkler systems can reduce runoff with variable rate emitters that match application rates to soil infiltration rates at specific field locations.

Erosion reduction from improved scheduling and application management is usually proportional to runoff reduction. Reducing over application also reduces pumping costs and losses of applied nutrients and agri-chemicals. In surface irrigation systems, where 20%–40% runoff is often required to achieve field application uniformity, erosion remediation can be integrated into water supply enhancement by pumping sediment-laden drain water back onto fields. This does not prevent erosion, but does replace most of the eroded soil along with the saved water for the pump-back cost.

Methods of Control

Conversion to Sprinklers

One effective way to prevent irrigation-induced erosion is conversion from surface to sprinkler irrigation. Again, the soil conservation benefit from conversion to sprinklers derives from and is proportional to the reduction of runoff. Sprinkler irrigation has higher technical, capital, energy, and infrastructure requirements than surface irrigation. Therefore, sprinklers are used only on a small fraction of global irrigated area, whereas, nearly 60% of US irrigated land uses sprinklers. Properly designed and managed sprinkler systems eliminate 100% of off-site sediment losses. However, with sprinklers, there is a tendency to extend irrigation to steeper slopes or otherwise more erosive lands. On steep land, when sprinkler systems are poorly designed or managed, erosion can occur.

Center pivots can cause erosion problems because of water running in wheel ruts, down steep slopes, or due to high application rates at outer reaches of the pivot,[8] especially when using extendable booms and high volume end-guns to reach corners. Erosion from high application areas, or where runoff concentrates, can be reduced using tillage, pitting, and mulching between rows to increase surface roughness storage and reduce runoff.[9–11]

Soil Protection and Tillage

Many approaches developed to control rainfall-induced erosion can prevent irrigation-induced erosion, particularly under sprinklers, e.g., no-till and conservation tillage, which rely on crop residue to protect the soil surface. Despite typical erosion reductions >90%, often with increased yields,[12] no-till and conservation tillage are rarely practiced by surface irrigators. Floating residue often migrates along and clogs irrigation furrows, washing out adjacent beds and furrows, while under-irrigating the blocked furrow. In basin flood irrigation, floating debris can interfere with water spreading, sometimes concentrating initial flows, eroding some areas, and burying emerging plants with sediment or debris. No-till farming with furrow irrigation is further complicated by crop rotations that require different row (and furrow) spacings each season.

Sojka et al.[13] demonstrated 60% reduction in field sediment loss from furrow-irrigated potatoes that were paratilled (subsoiled) following planting. Slight yield increases and significant tuber grade improvements raised profitability under both furrow and sprinkler irrigation with paratilling.[14] Because irrigation assures crop water availability, yield benefits from improved root development are not consistently seen with subsoiling in irrigated crops.[15] Subsoiling is commonly practiced with sprinkler irrigation to enhance infiltration and decrease runoff, thereby reducing erosion. However, farmers are cautious about subsoiling furrow-irrigated crops because of the potential for irregular water flows in subsurface cracks to interfere with irrigation uniformity. Field preparation or land forming practices that reduce water application uniformity, or increase runoff, are avoided by irrigators.

Placing mulch or growing sod in irrigated furrows reduces erosion. Sod nearly eliminated runoff sediment.[16] Straw mulching reduced sediment loss from 52% to 71%.[17–20] Drawbacks of these techniques relate to the management of sodded furrows, the added operations and equipment needed to place straw, and debris migrating along and clogging mulched furrows.

Site Modification

Various "engineering" approaches have been used to reduce field sediment losses from surface irrigation. The most common is use of settling basins. Large quiescent pools to facilitate particulate settling from runoff collected from fields up to 20 hectares are fairly typical. Settling pond size depends upon the area served, rate and volume of runoff, sediment concentrations expected, and particle size distribution. Small settling basins along the bottom of surface irrigated fields, serving a few rows per basin, are sometimes easier to manage at season's end, when trapped sediment spread back onto the field using farm equipment. Big ponds require large scale equipment for construction, cleaning, and soil redistribution. For medium-textured soils, about 60% of suspended mass entering settling ponds is retained. The nonretained soil is of the clay size range.[21] Since clay carries most of sediment's adsorbed nutrient and chemical load, failure of ponds to retain clay impedes retention of agricultural chemical pollutants, despite the high percentage of sediment mass captured. Furthermore, effectiveness declines as ponds fill with sediment, reducing water residence time. Another variation on ponds is installation of buried drains and stand pipes to regulate water level in tail ditches.[22] The stand pipes force ponding and prevent gradual concaving of field tail ends. They do not, however, prevent loss into the drain of sediment entrained in runoff from upper field reaches.

Altering canopy configuration can reduce erosion. Sojka et al.[23] halved field sediment loss using narrow or twin row plantings. Water ran between closely placed furrows, reducing irrigation duration (and runoff) and allowing root systems and canopy debris to reduce soil detachment in the furrow. Filter strip crops drilled at right angles into the final 3–6 m of furrow-irrigated row crops also remove entrained sediments from runoff,[6] but do not prevent sediment migration from field inlet to tail end. Because filter strip management is a compromise between two crops, yield from the strips is typically half that expected for either crop alone.

Water Properties

Both the physical and chemical properties of irrigation water affect erosion. Erosion is greatly reduced by reducing sprinkler droplet size or energy[24,25] or by reducing stream flow in furrows.[26] These physical changes require adjustments in application timing, furrow lengths, and irrigation durations to properly match water application constraints with crop water needs.

Water electrolyte chemistry greatly affects the erosiveness of irrigation water.[27–30] High sodium adsorption ratio (SAR) and low electrical conductivity (EC) contribute to soil aggregate detachment, disruption, and dispersion of fine primary soil particles in runoff. The effect of low EC and high SAR are synergistic. Increasing electrolyte concentration with a calcium source lowers SAR, shrinks the ionic diffuse double layer around charged soil particles, and prevents dispersion, thereby maintaining aggregate stability and resisting erosion. The conjunctive use of waters from different sources or the addition of calcium can raise EC and/or lower SAR to reduce erosion potential and improve infiltration by stabilizing surface- soil structure.

Adding large polymeric compounds to irrigation water is an effective erosion prevention technology.[31–33] These compounds, when delivered in dilute concentrations (typically 1–10 ppm) by the irrigation stream, increase aggregate stability and inter-aggregate cohesion as water infiltrates. Erosion reduction of 95% is typical for application of 1 to 2 kg /ha per treated irrigation. Adoption has been greatest for furrow irrigation erosion reduction, but interest in extending the technology to sprinklers is growing as much to improve infiltration uniformity as to reduce erosion.[34–36] The most successful class of polymers has been anionic polyacrylamide, allowing safe, easy, and effective erosion prevention for seasonal application rates of 3–5 kg/ha^{-1}, or under \$35/ha^{-1} per season.[37]

References

1. Carter, D.L.; Berg, R.D.; Sanders, B.J. The effect of furrow irrigated erosion on crop productivity. Soil Sci. Soc. Am. J. **1985**, *49* (1), 207–211.
2. Bjorneberg, D.L.; Kincaid, D.C.; Lentz, R.D.; Sojka, R.E.; Trout, T.J. Unique aspects of modeling irrigationinduced soil erosion. Intern. J. Sed. Res. **2000**, *15* (2), 245–252.
3. Trout, T.J.; Neibling, W.H. Erosion and sedimentation processes on irrigated fields. J. Irr. Drain. Eng. ASCE. 1993, *119* (6), 947–963.
4. Sojka, R.E. Understanding and managing irrigationinduced erosion. In *Advances in Soil and Water Conservation;* Pierce, F.J., Frye, W.W., Eds.; Ann Arbor Press: Chelsea, MI, 1998; 21–37.
5. Carter, D.L. Soil erosion on irrigated lands. In *Irrigation of Agricultural Crops,* Agronomy Monograph No. 30; Stewart, B.A., Nielson, D.R., Eds.; Am. Soc. Agron.: Madison, WI, 1990; 1143–1171.
6. Carter, D.L.; Brockway, C.E.; Tanji, K.K. Controlling erosion and sediment loss from furrow-irrigated cropland. J. Irrig. Drain. Engr. ASCE. **1993**, *119* (6), 975–988.
7. Sojka, R.E.; Carter, D.L. Constraints on conservation tillage under dryland and irrigated agriculture in the United States Pacific Northwest. In *Conservation Tillage in Temperate Agroecosystems;* Carter, M.R., Ed.; Lewis Publishers: Boca Raton, FL, 1994; 285–304.
8. Gilley, J.R.; Mielke, L.N. Conserving energy with low- pressure center pivots. J. Irrig. Drain. Div. ASCE. **1980**, *106* (1), 49–59.
9. Aarstad, J.S.; Miller, D.E. Soil management to reduce runoff under center-pivot sprinkler systems. J. Soil Water Cons. **1973**, *28* (1), 171–173.
10. Kranz, W.L.; Eisenhauer, D.E. Sprinkler irrigation runoff and erosion control using interrow tillage techniques. Applied Eng. in Ag. **1990**, *6* (6), 739–744.
11. Oliveira, C.A.S.; Hanks, R.J.; Shani, U. Infiltration and runoff as affected by pitting, mulching and sprinkler irrigation. Irr. Sci. **1987**, *8* (1), 49–64.
12. Carter, D.L.; Berg, R.D. Crop sequences and conservation tillage to control irrigation furrow erosion and increase farmer income. J. Soil Water Conserv. **1991**, *46* (12), 139–142.
13. Sojka, R.E.; Westermann, D.T.; Brown, M.J.; Meek, B.D. Zone-subsoiling effects on infiltration, runoff, erosion, and yields of furrow-irrigated potatoes. Soil and Till. Res. **1993**, *25* (4), 351–368.
14. Sojka, R.E.; Westermann, D.T.; Kincaid, D.C.; McCann, I.R.; Halderson, J.L.; Thornton, M. Zone-subsoiling effects on potato yield and grade. Am. Pot. J. **1993**, *70* (6), 475–484.
15. Aase, J.K.; Bjorneberg, D.L.; Sojka, R.E. Zone-subsoiling relationships to bulk density, cone index, infiltration, runoff and erosion on a furrow irrigated soil. Trans. ASAE. **2001**, *44* (3), 577–583.
16. Cary, J.W. Irrigating row crops from sod furrows to reduce erosion. Soil Sci. Soc. Am. J. **1986**, *50* (5), 1299– 1302.
17. Aarstad, J.S.; Miller, D.E. Effects of small amounts of residue on furrow erosion. Soil Sci. Soc. Am. J. **1981**, *45* (1), 116–118.
18. Brown, M.J. Effect of grain straw and furrow irrigation stream size on soil erosion and infiltration. J. Soil Water Conserv. **1985**, *40* (4), 389–391.
19. Miller, D.E.; Aarstad, J.S.; Evans, R.G. Control of furrow erosion with crop residues and surge flow irrigation. Soil Sci. Soc. Am. J. **1987**, *51* (2), 421–425.
20. Brown, M.J.; Kemper, W.D. Using straw in steep furrows to reduce soil erosion and increase dry bean yields. J. Soil Water Conserv. **1987**, *42* (4), 187–191.
21. Brown, M.J.; Bondurant, J.A.; Brockway, C.E. Ponding surface drainage water for sediment and phosphorus removal. Trans. ASAE. **1981**, *24* (6), 1478–1481.
22. Carter, D.L.; Berg, R.D. A buried pipe system for controlling erosion and sediment loss on irrigated land. Soil Sci. Soc. Am. J. **1983**, *47* (4), 749–752.
23. Sojka, R.E.; Brown, M.J.; Kennedy-Ketcheson, E.C. Reducing erosion from surface irrigation using furrow spacing and plant position. Agron. J. **1992**, *84* (4), 668–675.

24. Bubenzer, G.D.; Jones, B.A. Drop size and impact velocity effects on the detachment of soils under simulated rainfall. Trans. of the ASAE **1971**, *14* (4), 625–628.
25. Kinnell, P.I.A. Laboratory studies on the effect of drop size on splash erosion. J. Agric. Eng. Res. **1982**, *27* (3), 431–439.
26. Trout, T.J. Furrow irrigation erosion and sedimentation: on-field distribution. Trans. of the ASAE **1996**, *39* (5), 1717–1723.
27. Le Bissonais, Y.; Singer, M.J. Seal formation, runoff, and interill erosion from seventeen california soils. Soil Sci. Soc. Am. J. **1993**, *57* (1), 224–229.
28. Lentz, R.D.; Sojka, R.E.; Carter, D.L. Furrow irrigation water quality effects on soil loss and infiltration. Soil Sci. Soc. Am. J. **1996**, *60* (1), 238–245.
29. Levy, G.J.; Levin, J.; Shainberg, I. Seal formation and interill soil erosion. Soil Sci. Soc. Am. J. **1994**, *58* (1), 203–209.
30. Shainberg, I.; Laflen, J.M.; Bradford, J.M.; Norton, L.D. Hydraulic flow and water quality characteristics in rill erosion. Soil Sci. Soc. Am. J. **1994**, *58* (4), 1007–1012.
31. Brown, M.J.; Robbins, C.W.; Freeborn, L.L. Combining cottage cheese whey and straw reduces erosion while increasing infiltration in furrow irrigation. J. Soil Water Con- serv. **1998**, *53* (2), 152–156.
32. Lentz, R.D.; Sojka, R.E. Field results using polyacrylamide to manage furrow erosion and infiltration. Soil Sci. **1994**, *158* (4), 274–282.
33. Orts, W.J.; Sojka, R.E.; Glenn, G.M. Biopolymer additives to reduce erosion-induced soil losses during irrigation. Industrial Crops and Products **2000**, *11* (1), 19–29.
34. Aase, J.K.; Bjorneberg, D.L.; Sojka, R.E. Sprinkler irrigation runoff and erosion control with polyacrylamide- laboratory tests. Soil Sci. Soc. Am. J. **1998**, *62* (6), 1681–1687.
35. Bjorneberg, D.L.; Aase, J.K. Multiple polyacrylamide applications for controlling sprinkler irrigation runoff and erosion. Applied Eng. in Ag. **2000**, *16* (5), 501–504.
36. Sojka, R.E.; Lentz, R.D.; Ross, C.W.; Trout, T.J.; Bjorneberg, D.L.; Aase, J.K. Polyacrylamide effects on infiltration in irrigated agriculture. J. Soil Water Conserv. **1998**, *53* (4), 325–331.
37. Sojka, R.E.; Lentz, R.D. Reducing furrow irrigation erosion with polyacrylamide (PAM). J. Prod. Ag. **1997**, *10* (1), 47–52.

6

Pesticide Translocation Control: Soil Erosion

Monika Frielinghaus,
Detlef Deumlich,
and Roger Funk

Soil Erosion: A Global Problem

One-third of the world's agricultural soils are reported to be in a degraded state. Water and wind erosion contribute to approximately 84% of the observed deterioration. The worldwide soil loss from soil erosion is estimated at about 75 billion tons yr^{-1}.[1] The changes caused by soils by human-induced erosion over many years are significant and have often resulted in valuable land being abandoned.[2] Given a very slow rate of soil formation, any soil loss of more than 1 ton ha^{-1} yr^{-1} within a time span of 50–100 yr can be considered as irreversible. Soil from the Sahel zone in Africa is picked up in Florida and the Brazils Amazonas region and blown across the Atlantic Ocean each year. Regularly, when the Chinese till their land in spring, whirled-up soil is detected in Hawaii. Most soil loss can be traced to inappropriate land management practices. Viewed on a large scale, these scenarios create an agricultural dilemma of global proportions.

Pesticides are used in nearly all agricultural systems and provide essential help against diseases, pests, and weeds, thereby contributing substantially to securing global food supply. However, the pesticides' translocation due to wind and water erosion causes heavy environmental damages.[3] In many regions, sediment transport by surface runoff or wind is the most important nonpoint pollutant generated by agriculture. While pollutants may be too insoluble or soil bound to be transported in runoff water, erosion can mobilize them into runoff. Erosion sediment is also an important vehicle for the translocation of pesticides that are water insoluble or strongly bound to eroding soil. Although wind erosion has been overlooked in the past as a land degradation process, it has received more attention now as an important source of atmospheric pollution.[4] Mainly, this is attributed to the removal of fine soil particles and organic material, carrying the adsorbed pesticides.

Water and Wind Erosion and Sediment Transport

The processes of soil erosion involve detachment of material by two processes: raindrop impact and drag force traction. Material is then transported either by overland water flow or by saltation through the air.

Water Erosion

Runoff caused by heavy rainstorms (high intensity or long duration) is the most important direct driver of severe soil erosion by water (Figure 1). Mean losses of 20 to 40 tons ha^{-1} in individual storms may happen once every 2 or 3 years. More than 100 tons ha^{-1} soil loss is measured after extreme storms.[5]

The water erosion rate is a complex result of high rainfall amount or intensity, soil erodibility (instability), slope steepness and length, and the type of land use. As to the latter, the main causes are inappropriate agricultural practices, deforestation, overgrazing, forest fires, and construction activities.

The formation of a water erosion system on agricultural areas without surface protection by crops or plant residues causes rainwater and sediment transport in rills or gullies. The first phase after ponding is a concentration of water at linear paths. The rill formation may be influenced by a variety of different factors like soil crusting, wheel tracks, and reduced infiltration by soil compaction. Rills and gullies often develop in close proximity to one another. The second phase is a heightened concentration of water at morphological deep lines (thalways). The hydrological power of transport water as well as the off-site risk for sediment and pesticide transport into the lakes or rivers is increased with the thalways' catchment area (Figure 2).

Often, the interaction between runoff and sediment transport complicates the process' description.

Wind Erosion

Wind erosion occurs when three conditions coincide: high wind velocity, a susceptible surface with loose particles, which can be picked up, and insufficient surface protection by plants or plant residues (Figure 3).

Wind erosion results from wind moving across a dry soil surface and dislodging soil particles by pressure and lifting forces. The process is self-perpetuating; blowing sediment disturbs additional particles that are then lifted into the airstream.

FIGURE 1 Water erosion with sediment translocation at a corn field.
Source: Photo courtesy of Deumlich.

FIGURE 2 Sediment and pesticide transport into a channel caused by water erosion.
Source: Photo courtesy of Frielinghaus.

FIGURE 3 Wind erosion at a field after seedbed preparation.
Source: Photo courtesy of Schäfer.

The potential wind erosion rate is influenced by high wind velocity, soil erodibility (size/weight of particles), the degree to which the landscape is wind exposed, and the uses of land. The real wind erosion risk depends on the actual soil moisture and the real soil cover rate.

The modes of soil particle motion are closely related to particle size, density, and shape. It is important on agriculturally used soils of each textural class with organic material and absorbed pesticides (Figure 4). Particles with a smaller terminal velocity than the turbulent motions become suspended. Particles smaller than 20 μm are subjected to long-term suspension whereby they can be carried across several hundred kilometres for several days. Particles with diameters between 20 and 70 pm remain suspended for only a few hours and cannot be transported very large distances. These kinds of small particles are an important factor regarding the adsorption of pesticides.[6]

FIGURE 4 Sediment and pesticide transport into a channel caused by wind erosion.
Source: Photo courtesy of Frielinghaus.

Assessment of Pesticide Translocation

Pesticide findings even during the non-spraying seasons indicate that the occurrence is not only due to direct contamination from accidental spills and incorrect handling but also to diffuse contamination originating from normal pesticide use.

Pesticide mobility may result in redistribution within the application site or movement of some amount of pesticide off-site.

Pesticide characteristics are relevant for determining the fate of transport by wind or water erosion. These properties include solubility in water, tendency to adsorb to the soil, and pesticide persistence in soil.[7]

Water solubility is measured in milligrams per liter. Pesticides with more than 30 mg L^{-1} solubility tend to move with surface runoff during an erosion event. Pesticides with solubility about less than 1 mg L^{-1} tend to remain on the soil surface and may move with soil sediment in surface runoff if soil erosion occurs.

Soil adsorption is measured by K_{oc}, which is the tendency of pesticides to be attached to soil particles. Higher values (>1000) indicate a pesticide that is strongly attached to soil and less likely to move unless sediment transport occurs. Lower values (<300–500) indicate pesticides that tend to move with surface runoff. Among other things, sorption is influenced by soil moisture, organic matter content, and texture. These parameters are also important to quantify soil erodibility. Soils high in clay or organic matter content, or both, have a higher potential to adsorb pesticides, because small particles have plenty of surface area and are chemically more active compared with sandy soil particles.

Pesticide persistence is measured in terms of the halflife, or the time in days required for a pesticide to degrade in soil to one-half of its original amount. A pesticide with a half-life >21 days may persist long enough to move with surface runoff or erosion before it degrades. For example, experiments with dieldrin and aldrin show half-life values of about 3 or 4 years. These two pesticides lost 90% of their original concentration in no less than 10 years.[8]

Field investigations to estimate the pesticide transport during an erosive rain were known by Gouy et al.[9] and Klöppel et al.[10]

In the first experiment, pesticides were applied onto the bar soil surface. The simulated rainfall with an intensity of 33 and 44 mm hr^{-1}, respectively, began 20 hr after pesticide application (Table 1).

Atrazine, simazine, and alachlor, for example, were mainly transported in surface runoff. In contrast, 90% of transported trifluralin was adsorbed on eroded particles. Lindane, for example, showed an intermediate status and the high dynamic of pesticide translocation.

TABLE 1 Total Amounts of Pesticide in Runoff Samples and Their Distribution between Sediment and Liquid Phases

Pesticide	Water Solubility (mg L–1)	Adsorption, Koc (mL g–1)	Total Amount (% of Applied)	Distribution On Sediment	Distribution In Solution
Atrazine	33	100	13	0.8	12.2
Simazine	6	130	9	0.6	8.4
Alachlor	240	170	14	0.8	13.2
Lindane	7	1100	17	6.0	11.0
Trifluralin	1	8000	14	12.6	1.4

The second rainfall simulation study on small field plots provided information about worst cases for pesticide translocation (rainfall intensity of about 70 mm hr^{-1} and 100 mm accumulated rainfall).[10] The pesticides of the first group had water solubility concentrations of 65 and 700 mg L^{-1}, e.g., isoproturon and dichlorprop-p, respectively. The second group had solubility concentrations <1 mg L^{-1}, e.g., bi-fenox. Two different soil cover conditions were tested.

The results demonstrate that the influence of soil cover characteristics on concentrations and total runoff losses of pesticides for the highly soluble group is restricted to the start of the rain event after pesticide application and the runoff rate. The sediment concentration of pesticides in the fairly soluble group was comparable to those of the first runoff event independent of the time lapse between the application and rainfall event. The fairly soluble pesticides are thus exiting the field sites only by being adsorbed to eroded sediment. The cumulative soil loss is the most reliable indicator that explains the decrease in sediment and pesticide concentrations during each rainstorm (Table 2).

It is not possible to compare these results with other data from literature, due to the great variety of parameters concerning the performance of experiments. However, the findings are consistent in that a rough estimation of pesticide concentration in runoff and total pesticide losses is possible when rainfall (duration, intensity) and soil erosion potential are known.

Pesticide transport by wind may occur through isolated pesticide displacement or in a sediment-bounded form. At present, only initial findings of the amount of translocated pesticides are available. It is assumed that the horizontal transport of particulate-bounded pesticides is more than 50% of the total loss resulting from extreme wind erosion events. The herbicide loss of about 1.5% of the total amount, which was applied and integrated into the upper humus horizon on steppe soils in Canada, was measured after 13 wind erosion events. Herbicide loss from a non-integrated application on soil surfaces was approximately 4.5%.[11]

Experiments in a wind tunnel with the pre-emergence herbicide flurochloridone showed <1% to <58% displacement dependent on soil erodibility and measuring height.[6]

Determination of pesticide content in the atmosphere is extremely important but very difficult. Additionally, pesticides enter the atmosphere via many different processes, particularly by volatilization (simulated with PEARL, PELMO), unconsidered in our context.[12]

The investigation of atrazine, alachlor, and acetochlor concentrations on soil surface and dissipation rates of wind-erodible sediment and larger fractions from two soil types was important.[13] Undisturbed and incorporated (5 cm deep) soil surface were analyzed. The surface (1 cm) of soil was removed by vacuum 1, 7, and 21 days after herbicide treatment. About 50% of the recovered material was classified as wind-erodible sediment. This erodible sediment contained about 65% (undisturbed soil surface) and 8% (incorporated soil) of the applied herbicides, respectively, after 1 day. The concentrations were similar after 7 and after 21 days. However, a 50% dissipation rate for each herbicide was found after 15 days for wind-erodible sediments compared with 30–55 days for greater fractions. These data indicate that wind-erodible size aggregates and particles could be a source of herbicide contamination, but there is currently no information about quantities.

TABLE 2 Average Pesticide Losses in Relation to Applied Pesticide Dependent on Time (Rainfall Intensity: 70 mm hr^{-1}, Accumulated Rainfall: 100 mm)

Name of Pesticide	Pesticide Transporting Medium	Time and Concentration Lapse between Pesticide Application and Rainfall Event				
		2 hr	1 day	3 days	5–7 days	14 days
Application on Bare Soil Surface Sites (Pre-Emergence Application) % of Applied Amount						
Isoproturon	Total	4.5–12.4	4.0–17.2	5.4–13.3	2.1–11.2	n.e.
	In runoff water	3.8–9.8	2.9–13.9	4.0–9.8	1.4–9.0	n.e.
	In eroded sediment	0.7–2.6	1.7–3.3	1.4–3.5	0.7–2.2	n.e.
Dichlorprop-P	Total	2.1–10.4	2.2–16.7	5.9–10.4	1.2–8.3	n.e.
	In runoff water	1.9–9.8	1.8–15.3	5.3–9.3	0.9–7.0	n.e.
	In eroded sediment	0.2–0.6	0.4–1.4	0.6–1.1	0.3–1.3	n.e.
Bifenox	Total	15.6–19.0	19.3–21.6	14.3–17.2	9.3–13.8	n.e
	In runoff water	<0.1–0.9	<0.1–0.3	<0.1–0.3	<0.1–0.2	n.e.
	In eroded sediment	14.7–19.0	19.0–21.6	14.0–17.2	9.3–13.6	n.e.
Application on Small Covered Soil Surface (Barley with 3–5 Leaves) % of Applied Amount						
Isoproturon	Total	2.8–13.3	1.8–16.4	4.1–11.2	1.5–6.0	1.6
	In runoff water	2.2–11.4	1.5–13.5	3.3–8.6	1.0–4.6	1.1
	In eroded sediment	0.4–1.9	0.3–2.9	0.8–2.6	0.5–1.4	0.5
Dichlorprop-P	Total	0.8–10.0	0.9–9.2	1.0–8.3	0.7–4.1	0.8
	In runoff water	0.7–9.4	0.8–8.4	0.8–7.5	0.6–3.7	0.7
	In eroded sediment	0.1–0.6	0.1–0.8	0.2–0.8	0.1–0.4	0.1
Bifenox	Total	11.2–15.0	7.8–15.5	7.8–15.9	3.4–9.8	5.2
	In runoff water	<0.1–0.9	<0.1–0.2	<0.1–0.4	<0.1–0.3	<0.1
	In eroded sediment	11.2–14.1	7.8–15.5	7.4–15.9	3.4–9.5	5.2

It is important to recognize that only one fraction of pesticides in use is very strongly soil bound as to be transported in the sediment phase of runoff principally today (insecticides paraquat and pyrethroid, and other non-ionic hydrophobic species). Such pollutants may be too insoluble or soil bound to be transported in runoff or wind stream, but erosion can mobilize them.

Monitoring and Modeling

Monitoring in landscapes with a high erosion risk (water and wind) is predicted upon the estimation of rain and wind erosivity, soil erodibility, and the morphological factors of the areas (slope steepness, field length, wind openness, thalways). In sites with an erosion risk, well-designed field studies are the best way to assess off-site transport paths. When these data are not available, estimation about modeling will be necessary.

Water Erosion Models

Pesticide transport models are rare. Therefore, currently, water and wind erosion models serve as a basis for risk assessment, as a means to an end (Table 3).[14]

Most of the models were developed to assess impacts of different agricultural management practices; they are not adapted to predict exact pesticide, nutrient, or sediment loading in an area. The curve number method is an event or field scale orientated model and limits the results of some other models (ANSWERS, GLEAMS, CREAMS, and SWAT). In most erosion models, runoff and sediment load are only computed for the catchment outlet. Most of the hydrological models predict total runoff better than

TABLE 3 Selected Erosion Models and Hydrological Models with Integrated Pesticide Transport

Erosion Models	Models with Pesticide Transport
ANSWERS	CREAMS
KINEROS	GLEAMS
EUROSEM	AnnAGNPS
LISEM	SWAT
EROSION 3D	
WEPP	

Note: Wind Erosion Models: Wind Erosion Prediction System (WEPS); Revised Wind Erosion Equation (RWEQ).
Source: http://www.soilerosion.net/doc/models_menu.html.

sediment load. The models over-or underestimate empirical results for small erosion events especially. Also, these models do not consider any attenuation or partitioning during transport and therefore fail to predict loads of soluble pesticides to surface waters.

WEPP and EROSION 3D models simulate the erosion and sediment transport continuously. The WEPP model is based on fundamentals of erosion theory, soil and plant science, channel flow hydraulics, and rainfall–runoff relationship.

It is possible to extend erosion and sediment transport models to a transport model for pesticides as well as sediment particles. The pesticide transport behavior that includes interaction processes in solutions is unknown for watershed scale.

Wind Erosion Models

A physical-based process model is the wind erosion prediction system (WEPS). This is a continuous, daily time-step model for simulation of weather, field conditions, and wind erosion. It has the capability of simulating spatial and temporal variability or soil surface parameters and soil loss or deposition within a field. To aid in the evaluation of offsite impacts, the soil loss is subdivided into components and reported as saltation creep, total suspension, and fine particulate matter components (PM 10). The transport capacity for insoluble pesticides bounded in kind of suspension or PM 10 is appreciable.[5]

Further models are the revised wind erosion equation (RWEQ) and WHEELS, which can be used for single event simulation, long-term risk assessment, and assessment of changing management strategies.

Soil Erosion and Pesticide Translocation Control

The best management strategy requires inventory and risk analysis.

A step-by-step analysis is the basis for an effective pesticide translocation control:

Step 1: Monitoring the pesticide content in lakes or rivers.
Step 2: Quantification and elimination of pesticide point sources.
Step 3: Definition of non-point pesticide sources.
Step 4: Evaluating the pesticide management in the catchment (water solubility, adsorption, persistence, time of application).
Step 5: Estimating the potential water erosion or wind erosion risk (erosivity of rainstorms and wind, erodibility of soil, landscape and field openness, composition, thalways).
Step 6: Estimating the actual water and wind erosion risk (soil cover by plants or plant residues, crop rotation, humus status, soil tillage practices).
Step 7: Development of an efficient concept to realize a sustainable control of soil erosion, runoff, and pesticide transport.

The most important soil-protection strategy is based on the principles of precaution that took into consideration the economic and ecological consequences of the high costs and the endangerment to life and biodiversity.

As a result, the risk of soil erosion and pesticide transport should be addressed by an appropriate selection and a carefully timed application of pesticides. Products with low water-soluble active substances and a very short persistence should be preferred for slopes with a high runoff potential, caused by a network of waterways nearby the lakes or rivers. For example, it is advisable to replace products with isoproturon or isopropylamine with products containing pendimethalin or bifenac. For purposes of crop protection, water-soluble products that infiltrate into the upper soil layer should be used on areas with a high risk of wind erosion. Pre-emergent application on bare soil surface should be avoided on areas with a high water or wind erosion risk.

The most effective system to prevent sediment transport and reduce runoff is a temporal and spatially closed vegetation or residue covering. This cover protects the soil surface from the initial soil detachment action caused by raindrop splash forces or wind power. The result is a lower volume of runoff and sediment transport. Plant roots retain soil particles and reduce the sediment load (Table 4).

The extent of soil cover material, green plant mass, or crop residues can be influenced by the farmer's management and soil tillage practices.

The farmers can estimate the risk of their tillage system independently (Table 5).

The rate of soil cover is a highly effective indicator for assessing the risk of water and wind erosion. This indicator addresses the following questions: 1) how much soil cover is necessary to reduce the threat of erosion, runoff, and pesticide translocation for a high-risk area; and 2) how much cover can be realized dependent on crop type, crop rotation, tillage, and management practices in different regions. Based on this analysis, appropriate preventive management practices can be required.

The greatest wind and water erosion risk occurs after seedbed preparation, which is characterized by the lowest degree of soil surface roughness and bare soil. Crop selection, improved crop rotations, and a change in soil tillage practices are established methods for increasing the soil cover for a reduced pesticide transport by wind and water erosion and runoff.

More basic management methods such as contour tillage or strip cropping practiced in some parts of the world are not as effective as conservation tillage in minimizing soil erosion.

Other technical solutions to reduce pesticide transport in areas with high erosion risk are engineering measures such as small ponds or farm-track construction. Alteration of field design, hedge planting to reduce wind openness, and waterways with permanent grass-covered banks or buffer strips nearby the lakes or rivers are effective arrangements to interrupt pesticide transport.

Therefore, precautionary measures against water and wind erosion as well as pesticide translocation have to be targeted (for example, prevention of pesticide pollution), site specific (for example, adequate for water and wind erosion risk), and applied for a longer period.

TABLE 4 Correlation between Cover Rate, Runoff, and Soil Loss Estimated from Long-Term Experiments with Conventional and Conservation Tillage Treatments.

Soil Cover with Green Plants	Soil Cover with Plant Residues	Runoff	Relative Soil Loss by Sediment Transport in Runoff	Relative Soil Loss Caused by Wind Erosion
(%)	($t\ ha^{-1}$ Dry Matter)	(% of Rain)	(%)	(%)
0	0	45	100	100
Approx. 20–30	0.5	40	25	15
Approx. 30–50	1–2	35	8	3
Approx. 50–70	2–3	ca. 30	3	<1
>70	>3–4	ca. 30	<2	<1

Source: Deumlich et al.[5]

TABLE 5 Appraisal Matrix to Estimate Soil Cover of Crops and Crop Rotations and Reduced Tillage Systems

			Evaluation Criteria			
	Period between Sowing and Efficient Soil Cover	Cover Rate Dependent on Distribution	Soil Cover during Summer Time	Soil Cover during Winter Time	High Risk Dependent on Tillage System	Total Evaluation
				Soil Cover		
		Crops				
Grass	1	1	1	1		1
Winter barley	1	1	1.5	1		1
Winter wheat						
Sowing before October 1	2	1	1.5	1.5		1.5
Sowing after October 1	3	1	3	2.5		2.5
Summer barley	1.5	1	2	3		2
Potatoes	2.5	3	3	3		3
Sugar beets	3	2.5	2.5	3		3
Maize	3	3	2.5	3		3
Sunflower	3	3	2.5	3		3
		Crop Rotations				
Grass rotation	1	1	1	1		1
Maize–winter wheat–winter barley	2.5	2	2	2.5	2.5	2.5
Sugar beet–winter wheat–winter barley	2.5	2	2	2.5	2.5	2.5
		Conservation Tillage Systems				
Winter barley–cash crops–mulch (frozen cash crops)–maize–winter wheat	1.5	1	1.5	1.5	1.5	1.5
Winter rye with undersown crops–winter barley–cash crops–mulch (frozen cash crops)–maize–winter wheat	1.5	1	1.5	1.5	1.5	1.5
Winter barley–cash crops–mulch (frozen cash crops)–sugar beet–winter wheat–winter barley	2	1	1.5	2	1.5	2
		Appraisal:				
1) Efficient soil protection						
2) Moderate soil protection						
3) Increased erosion and pesticide translocation risk						

Erosion and Runoff Control: Part of the Environmental Legislation

Precautionary measures play an important role in promoting sustainability and realizing the European Water Framework Directive (EU-Com, 2000, Water Framework Directive. Official Journal L 327). The Directive is a contribution to the progressive reduction of hazardous substances emission to water. Strategies against pollution of water are formulated in Articles 2 and 16: "Pollution means the direct or indirect introduction of substances or heats into the air, water or land as a result of human activity, which may be harmful to human health or the quality of aquatic ecosystems or terrestrial ecosystems directly depending on aquatic ecosystems …" The Commission identified priority hazardous substances like pesticides (Annex X). Quality standards as to the concentration of these substances in surface water, sediment, or biota exist.

There are diverse conditions and needs in the community members that require different specific solutions. The main principles of soil erosion control are formulated in the strategy entry of the EU-Com: "Strategy for soil protection 1995–2005."

A regional legislation is the German Soil Protection Act (1999).

The meaning of hazard prevention is defined in § 4: "(1) official directive for analysis of damage and erosion pattern after serious events; (2) legislative ordinance for conservation and land-use change to prevent future hazards."

The definition of precaution is paraphrased with the term "best management practice in agriculture" (§17 of the Act). The essentials of best practice in agricultural soil use are the monitoring of soil fertility and functional capacity as a natural resource. Best management practice is founded on the principle that "soil erosion shall be avoided wherever possible, by means of site-adapted use, especially use that takes slope, water and wind conditions and the soil cover into account; … the predominantly natural structural elements of field parcels that are needed for soil conservation, especially hedges, field shrubbery and trees, field boundaries and terracing, shall be preserved.…"

The German Federal Immission Control Act (2002) must also be considered in implementing the precautions necessary to protect the environment by wind-initialized pesticide transport in the form of PM 10 emission.

Similar acts and regional precautionary basis information exist in other European countries like Austria and Switzerland.

The Australian Pesticide Act (1999) aims to reduce the risks associated with the use of pesticides to human health, the environment, property, industry, and trade.

The U.S. Food, Conservation, and Energy Act of 2008 contains Title II, "Conservation" with the Conservation Reserve Program (Subtitle B) and the Environmental Quality Incentives Program (Subtitle F). The Environmental Protection Agency (EPA) regulates the sale and use of pesticides in the United States through registration and labeling of pesticide products. The EPA is directed to restrict the use of pesticides as necessary to prevent unreasonable adverse effects on people and the environment (Pesticide Registration Improvement Act of 2004, P.L.108–199).

Conclusion

Most soil loss caused by wind and water erosion can be traced to inappropriate land management practices. Pesticide findings even during the non-spraying season indicate that the occurrence is due not only to direct contamination from accidental spills and incorrect handling (point source). Non-point sources of pesticide water and air pollution are water and wind erosion. Pesticide solubility, soil adsorption of pesticides, and pesticide persistence are important parameters to estimate the translocation by runoff water and sediment movement. Rainfall simulation studies demonstrate the different kinds of translocation, dependent on the solubility: highly soluble pesticides were analyzed in runoff water and the concentration decreased after application (for example, atrazine, simazine, isoproturon, and dichlorprop); fairly soluble pesticides were analyzed, and they only adsorbed onto eroded sediment (for example, bifenox and trifluralin). The second group could be transported via water or wind erosion.

Only one fraction of pesticides in use is very strongly soil bounded and therefore too insoluble to be transported in runoff or wind stream, but erosion can mobilize them.

That means that soil erosion control includes non-point pollution by pesticide translocation. A step-by-step analysis is the basis for an effective, sustainable, and site-adapted land use and agriculture management. The risk for soil erosion and pesticide translocation should be addressed by an appropriate selection on a carefully timed application of pesticides. Products with low water-soluble active substances and a very short persistence should be preferred for slopes with water erosion risk. Water-soluble products that infiltrate into the upper soil layer should be used on areas with a high wind erosion risk. The most effective system to prevent sediment transport and runoff reduction is a temporal and spatial closed vegetation or residue covering. The farmers can evaluate their management self-contained to change the system.

References

1. Hurni, H. Soil conservation policies and sustainable land management: A global overview. In *Soil and Water Conservation Policies and Programs*; Napier, T.L., Napier, S.M., Tvrdon, J., Eds.; CRC Press: New York, 2000; 19–30.
2. Pimentel, D. Soil erosion: A food and environmental threat. Environ. Dev. Sustainability **2006**, *8*, 119–137.
3. Reichenberger, S.; Bach, M.; Skitschak, A.; Frede, H.G. Mitigation strategies to reduce pesticide inputs into ground- and surface water and their effectiveness. A review. Sci. Tot. Environ. **2007**, *384*, 1–35.
4. Gobin, A.; Govers, G.; Jones, R.; Kirkby, M.; Kosmas, C. *Assessment and Reporting on Soil Erosion*, Technical Report 94; European Environment Agency: Copenhagen, 2003.
5. Deumlich, D.; Funk, R.; Frielinghaus, M.; Schmidt, W.A.; Nitzsche, O. Basics of effective erosion control in German agriculture. J. Plant Nutr. Soil Sci. **2006**, *169*, 370–381.
6. Funk, R.; Reuter, H. Wind erosion. In *Soil Erosion in Europe*; Boardman, J.; Poesen, J., Eds.; John Wiley & Sons Ltd, 2006; 563–582.
7. Stevenson, D.E.; Baumann, P.; Jackmann, J.A. Pesticide properties that affect water quality. Texas Agriculture Extension Service B—6050.
8. Ghadiri, H.; Rose, C.W. Water erosion processes and the enrichment of sorbed pesticides. J. Environ. Manage. **1993**, *37*, 23–35.
9. Gouy, V.; Dur, J.-C.; Calvet, R.; Belamie, R.; Chaplain, V. Influence of adsorption-desorption phenomena on pesticide run-off from soil using simulated rainfall. Pestic. Sci. **1999**, *55*, 175–182.
10. Klöppel, H.; Haider, J.; Kordel, W. Herbicides in surface runoff: A rainfall simulation study on small plots in the field. Chemosphere **1994**, *28* (4), 649–662.
11. Larney, F.J.; Cessna, A.J.; Bullock, M.S. Herbicide transport on wind-eroded sediment. J. Environ. Qual. **1999**, *28*, 1412–1421.
12. Wolters, A.; Linnemann, V.; Smith, K.E.C.; Klingelmann, E.; Park, B.J.; Vereecken, H. Novel chamber to measure equilibrium soil-air partitioning coefficients of low-volatility organic chemicals under conditions of varying temperature and soil moisture. Environ. Sci. Technol. **2008**, *42* (13), 4870–4876.
13. Clay, S.A.; DeSutter, T.M.; Clay, D.E. Herbicide concentration and dissipation from surface wind-erodible soil. Weed Sci. **2001**, *49* (3), 431–436.
14. Schulz, M.; Matthies, M. Runoff of pesticides: Achievements and limitations of modelling agrochemical dislocation from non-point sources at various landscape related scales. Living Rev. Landscape Res. **2007**, 1.

7

Pesticides

Marek Biziuk,
Jolanta Fenik,
Monika Kosikowska, and
Maciej Tankiewicz

Introduction

A numerous and diverse group of chemical compounds with an extensive range of action,[1] pesticides are very widely used, though primarily to eliminate insect pests and weeds and to limit their negative effects in agriculture and in the household.[2–4] They enable the quantities and quality of crops and food to be controlled and help to limit the many human diseases transmitted by insect or rodent vectors. Despite their many merits, however, pesticides are some of the most toxic, environmentally stable and mobile substances in the environment, which are able to bioaccumulate. They can also participate in various physical, chemical, and biological reactions, as a result of which even more toxic substances may be produced; by accumulating in living organisms, these can lead to irreversible, deleterious changes. The non-rational application of pesticides also adversely affects the environment and humans, increasing susceptibility to diseases and poisoning.[1]

Characteristics of Pesticides

The following are the current uses of pesticides:

- To remove, control the number of, or attract various kinds of pests destroying plants and plant products
- To kill weeds
- To destroy foliage and excess numbers of flowers
- To increase the production of animal and plant biomass
- To combat human, animal, and plant pathogens
- To control the growth of plants or their parts
- To combat microorganisms causing farm produce to rot and decay
- To combat insects and other animals occurring in homes, cellars, stores, etc.
- To protect fabrics in textile mills and dry-cleaning establishments
- To prevent the growth of algae in swimming pools
- To combat fungi in paints and paper products
- To protect museum exhibits against the action of pests
- To counteract growths on boats and ships[1]

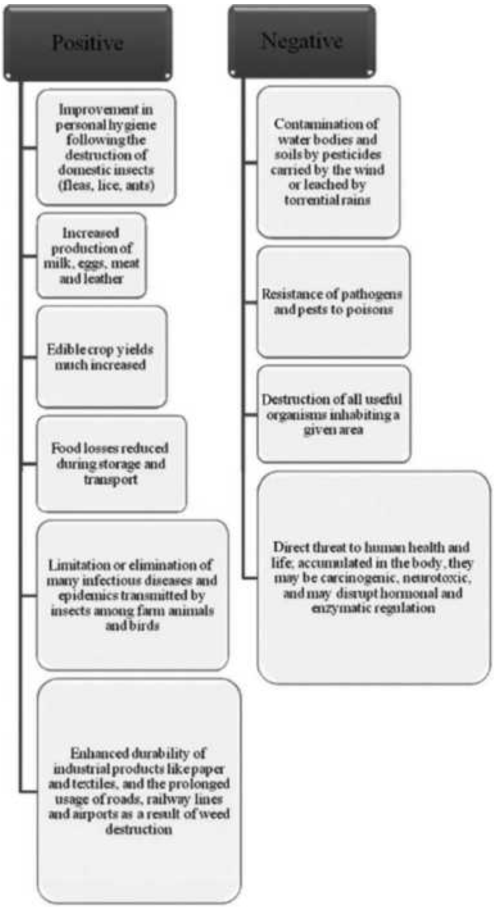


FIGURE 1 The effects of using pesticides.
Source: Biziuk et al.,[1] Moreno et al.,[5] and Hajslova and Zrostlikova.[6]

Ideally, pesticides should be applied only in accordance with their purpose, so that they do not have a negative effect on humans, flora, or fauna. In practice, however, such complete selectivity is unattainable.[1,4] The factor determining whether a particular agent can be used is its rapid biodegradability and its minimal toxicity vis-à-vis the environment.

Chemical pesticides are of enormous importance in increasing the efficiency and quality of agricultural produce. They enter the environment in various forms: powders, moistened powders, powders for preparing aqueous solutions, and concentrates for making up emulsions or sprays. The ubiquitous application of pesticides not only contaminates water, soil, and air but also causes them to accumulate in farm produce like fruit and vegetables. Figure 1 summarizes the effects of pesticide application.[1,5,6]

According to Their Chemical Structure

TABLE 1 Classification of Pesticides according to Their Chemical Structure

Class	Examples
Inorganic pesticides	Arsenic and fluoride insecticides, inorganic herbicides
Organic pesticides	Organochlorine and organophosphorus insecticides, carbamates

Source: Biziuk et al.[1]

According to Their Chemical Class

TABLE 2 Classification of Pesticides according to Their Chemical Class

Class	Examples
Organochlorine	DDT, endosulfan, methoxychlor
Organophosphorus	Malathion, dichlorvos, fonophos
Carbamates and their derivatives (urethanes)	Aldicarb, aminocarb, furathiocarb
Derivatives of phenoxycarboxylic acids	2,4-D, MCPA, dicamba
Triazines and their derivatives	Atrazine, simazine, anilazine

Source: Biziuk et al.[1]

The diversity of their chemical structure, action, and application makes any classification of pesticides difficult.[1] There are a number of criteria according to which they can be categorized: toxicity, purpose of application, chemical structure, environmental stability, and the pathways by which they penetrate target organisms. Tables 1–4 outline some of these classifications:[1]

According to Their Application

TABLE 3 Classification of Pesticides according to Their Application

Class	Subclass	Application
Zoocides—agents for combating animal pests	Insecticides	Destruction of insects
	Aphicides	Destruction of aphids
	Acaricides	Destruction of plant mites
	Attractants	For attracting pests
	Bactericides	Destruction of bacteria
	Larvicides	Destruction of larvae
	Limacides	For killing slugs

(*Continued*)

TABLE 3 (Continued) Classification of Pesticides according to Their Application

Class	Subclass	Application
	Molluscicides	For killing snails
	Nematocides	For killing nematodes
	Ovicides	Destruction of the eggs of insects and mites
	Repellents	For repelling insects
	Rodenticides	For combating rodents
Fungicides		Fungicidal and fungistatic agents
Herbicides		Weed killers
Plant growth regulators—	Deflorants	For removing excess flowers
stimulants or inhibitors of the	Defoliants	For removing excess leaves
life processes of plants	Desiccants	For drying plants
Synergetics		For potentiating the action of other substances

Source: Biziuk et al.[1]

According to Their Toxicity

TABLE 4 Classification of Pesticides according to Their Toxicity

Toxicity Class	Median Lethal Dose LD50 (mg/kg body mass) When Administered via the Digestive Tract	Stability in the Soil (Time for Degradation to Harmless Products)	Stability in the Aquatic Environment (Time for Degradation to Harmless Products), Number of Days
I—Highly toxic	≤25	More than 2 years	>30
II—Toxic	25 < LD50≤200	0.5–2 years	11–30
III—Harmful	200 < LD50 ≤2000	1–6 mo	6–10
IV—Not very harmful	>2000	Within 1 mo	<5

Source: Biziuk et al.[1]

Circulation of Pesticides in the Environment

Nowadays, pesticides are very widely applied, not just in agriculture. Their ubiquity and ever-increasing consumption pose a greater hazard to the environment to which they are transported. Pesticides can enter the environment in their primary form or as decomposition products. When in the environment, they are subject to various transformations. They may be borne by wind or rain from their points of application to neighboring areas and crops, where they are not required or may be harmful. The quantity of pesticides circulating in a given region depends to a large extent on their intensity of application or the type of crop being grown there. Both modern pesticides and older ones are present in the environment. Figure 2 illustrates the circulation of pesticides in nature.

Pesticides in the Aquatic Environment

Pesticides are the most common contaminants in surface water and groundwater. There are also many reports of pesticides being found in drinking water, well water in farming areas, rainwater, subterranean water, and ice from the polar regions. Once in the aquatic environment, pesticides have a deleterious effect on the quality of waters used as sources of drinking water for the large majority of the Earth's

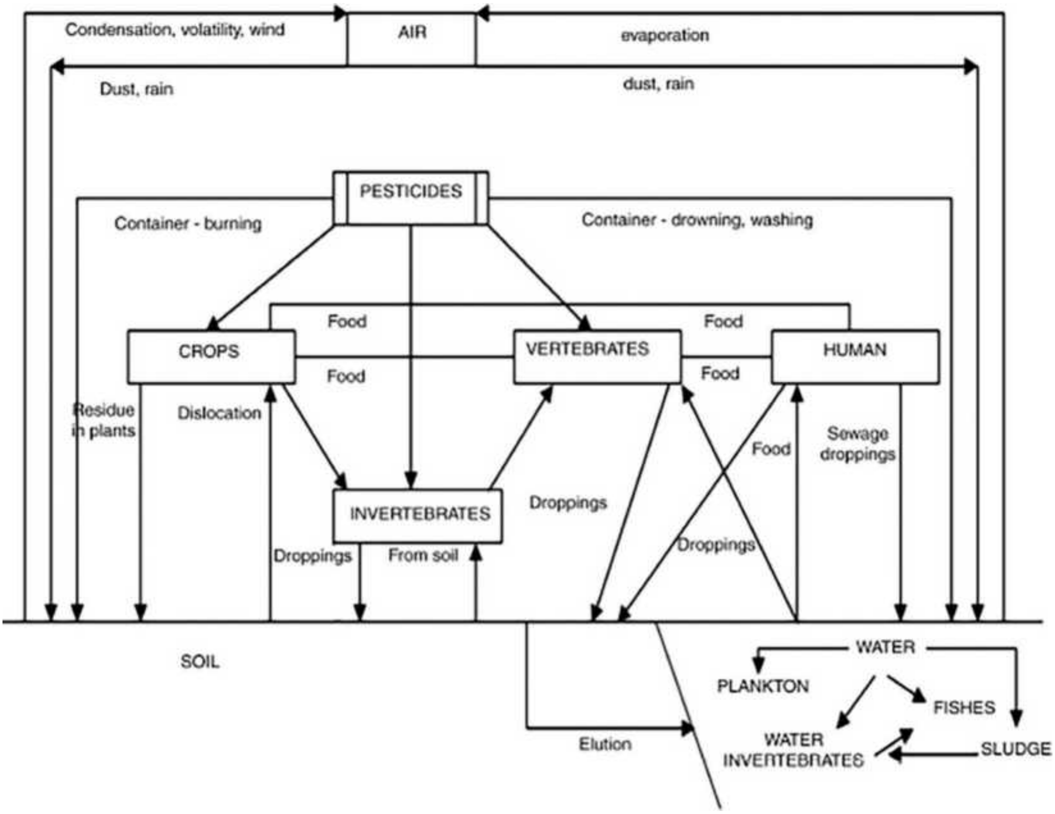


FIGURE 2 The circulation of pesticides in nature.
Source: Biziuk et al.[1]

human population. Concentrations are highest during the spring snow melt period and when pesticides are being applied to crops.[1,2] There are many sources from which pesticides get into the aquatic environment. Usually, these are area sources (e.g., precipitation, farming areas), but they may also come from point sources such as effluents of various kinds or leaking waste disposal sites. They can also be transported in the air for great distances.[1,7] When in the aquatic environment, pesticides are subject to a variety of transformations and processes:

- Physical (accumulation, deposition, dilution, diffusion)
- Chemical (hydrolysis, oxidation)
- Photochemical (photolysis, photodegradation)
- Biochemical (biodegradation, biotransformation, bioaccumulation)[3,8]

Substances of greater toxicity may be formed as a result of these processes, and when they accumulate in aquatic organisms, they can cause much irreversible damage to them.

Pesticides in the Air

The widespread application of pesticides contaminates not only waters and soils; the air is also affected. This is because pesticides rise into the air as they are being sprayed and thereafter (post-application emission).[1] An estimated 30%–50% of pesticides sprayed onto crops get into the atmosphere. By "post-application emission," we understand the evaporation of pesticides from soil or plant surfaces and due to the wind erosion of the soil. Depending on their stability, pesticides in the air may be degraded, or transported over long distances before being deposited. After their application, pesticides are present in the atmosphere in gaseous form, adsorbed onto solid particles, or dissolved in water vapor.

Pesticides in Crops

The presence of pesticides in the environment causes them to accumulate in crops like fruit and vegetables. The contamination of plant foods by pesticides is particularly dangerous because these compounds can reach every part of the plant, regardless of how they are applied. The degree of contamination of crops depends, among other things, on the dose and number of applications, the form of the pesticide preparation, the weather conditions, and the time elapsing between pesticide application and crop harvesting. Contamination is reduced by such factors as rainfall, wind, and chemical changes caused by oxygen, moisture, light, and plant enzymes. Plants with a large surface area relative to their mass retain larger amounts of pesticides.[1] Fruit and vegetables are capable of retaining larger quantities of pesticides. Pesticides can accumulate in fruit skins. The crops most exposed to the presence of pesticides are grapes, citrus fruits, and potatoes. Adsorbed pesticides can reduce the nutritious value of crops or alter their organoleptic properties. The contamination of crops with pesticides is due to their application in contradiction to good agricultural practice and the insufficient monitoring of their application. Their total elimination is, of course, often not possible, but the amounts applied can be limited to those that are harmless to human health.

Pesticides in Soil

Pesticides are transported to soil because of deliberate human activity. It is estimated that about 50% of pesticides come from washing processes of soils' surface. Transport of pesticides to plant material depends on their stability, and the main ways of absorption are via roofs and leaves. They can be also adsorbed on the molecules of clay. In this case, they do not penetrate the soil. Stability and fate of pesticides in soil depend on chemical structure, type of plant material, type of soil and their pH/temperature, weather conditions, etc.[1]

Modern Pesticides

Organochlorine pesticides were widely used in agriculture and pest control. These pesticides are insecticides composed primarily of carbon, hydrogen, and chlorine (see Figure 3). Most of them break down slowly and can remain in the environment long after application and in organisms long after exposure.

These chemicals were introduced for the first time (DDT [dichlorodiphenyltrichloroethane]) in the 1940s, and many of their uses have been nowadays cancelled or restricted because of their environmental persistence and potential adverse effects on wildlife and human health. Many organochlorines are no longer used in most countries, but other countries continue to use them (in Africa, South America, and Asia).

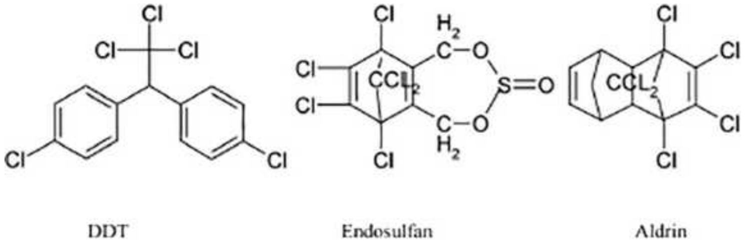

DDT Endosulfan Aldrin

FIGURE 3 The chemical structure of some organochlorine pesticides.

Organochlorine pesticides can enter the environment after pesticide applications, disposal of contaminated wastes into landfills, and releases from manufacturing plants that produce these chemicals. Some organochlorines are volatile, and some can adhere to soil or particles in the air. In aquatic systems, sediments adsorb organochlorines, which can then bioaccumulate in fish and other aquatic mammals. These chemicals are fat soluble, so they are found at higher concentrations in fatty foods. Organochlorine pesticides are hydrophobic, lipophilic, and extremely stable. Organochlorines have a wide range of both acute and chronic health effects, including cancer, neurological damage, and birth defects. Many organochlorines are also suspected endocrine disruptors.[1,9,10]

The present-day trend is to move away from persistent pesticides and to apply agents with a short decomposition time and no tendency to bioaccumulate.[1] That is why organochlorine pesticides have been withdrawn.[11–14] Because of their great stability (as long as 30 years), however, they may still be present in the environment and be transported by air or water over great distances. Organochlorine pesticides have been replaced by organonitrogen and organophosphorus pesticides. They have become very popular because they are cheap and readily available, have a wide range of efficacy, are able to combat a large number of pest species, and have a shorter environmental half-life than their organochlorine predecessors.

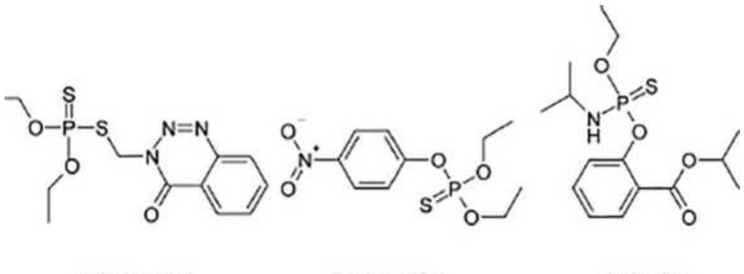
Currently used pesticides in comparison with formerly used ones are less persistent in the environment and generally are more polar. Polarity of pesticides determines their presence, transport, and stability in different compartments of the environment. Polar pesticides are more soluble and penetrate the surface water and groundwater faster. Because they are less stable in the environment, their presence can be established in less time (since the application). By contrast, non-polar pesticides are typically sparingly soluble in water, which is associated with their greater persistence in the environment, and transport mechanisms of these compounds are different. They are mostly adsorbed on solid particle and only partially dissolve in water. In addition, pesticides accumulate in sediments or suspensions and are present there even for decades.[1]

One of the principal classes of compounds used for plant protection, organophosphorus pesticides (insecticides) embrace all organic compounds containing phosphorus. Usually taking the form of esters (Figure 4) and degrading fairly easily, they are very poorly soluble in water, though better so in organic solvents and fats.

The umbrella term "organonitrogen pesticides" is a convenient way of referring to the large number of nitrogen- containing organic pesticides. In practice, however, these pesticides are known by the names of the various chemical classes. In the literature, the term "organonitrogen pesticides" usually refers to carbamates and triazines and their derivatives (see Figure 5).[15–17]

Table 5 presents the characteristics of organonitrogen pesticides.[1,18,19]

In recent years, new insecticides—neonicotinoids— have been introduced to replace pyrethroids, organophosphorus compounds, and carbamates. This was necessary as insect pests were becoming extremely resistant to the latter agents. At present, neonicotinoids are used to combat the Colorado potato beetle (Mospilan 20 SP, active substance: acetamiprid).[20]



Azinphos-ethyl Ethyl parathion Isofenphos

FIGURE 4 The chemical structure of some organophosphorus pesticides.

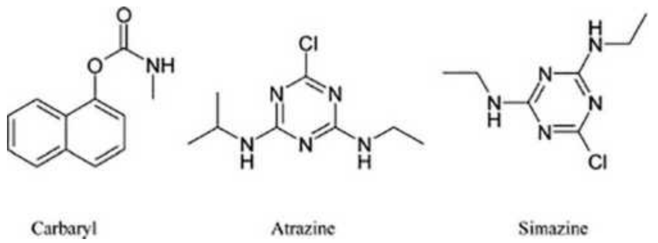

Carbaryl Atrazine Simazine

FIGURE 5 The chemical structure of some organonitrogen pesticides.

TABLE 5 The Characteristics of Organonitrogen Pesticides

Chemical class	Characteristics
Carbamates	Mostly carbamic acid esters; decompose rapidly in the soil; poorly soluble in water, better soluble in organic solvents and fats; used worldwide to combat insects, fungi, and weeds, and also as plant growth regulators
Triazines	Used to combat weeds and to control the growth of maize, soybean, grains, and other crops; highly toxic towards mono- and dicotyledons, and extremely stable in the soil; one of the triazines is atrazine, 90% of which is applied to maize crops

Source: Biziuk et al.,[1] Zhang and Lee,[18] and Sabik et al.[19]

Dangers Resulting from the Application of Pesticides

Human beings come into contact with pesticides in a variety of situations.[7,21] Ubiquitous and constituting a risk to human health, these substances can enter the body through

- The skin
- The respiratory system (through the nose)
- The digestive system (accidental swallowing of the product)
- Consumption of food contaminated with pesticides

Penetration through the skin is easier because pesticides can remain on the skin for a long time, the area of contact is large, and absorption of these compounds is rapid, especially through damaged skin. The inhalation of pesticides in contaminated air is very dangerous, as harmful substances very quickly enter the lungs and blood, and thence reaching all the organs of the body. The symptoms of pesticide poisoning include headache, a feeling of cold, giddiness/dizzi- ness, and skin rash. Organochlorine pesticides are capable of accumulating in the fatty tissues of living organisms, although their quantities in particular organs depend on the degree of fatness of the latter. When poisoning reaches a critical level, the most vulnerable organ is the brain. It was stated not long ago that some pesticides stimulate the production of microsomal hepatic enzymes responsible for the metabolism of some drugs (hypnotics, antiepileptics, analgesics), which reduces their therapeutic efficacy. Further symptoms of pesticide poisoning are as follows:

- Impairment of ion transport, leading to impaired neural conductivity
- Disturbances to neurotransmitter metabolism
- Abnormal immunological reactions
- Destruction of hepatocytes, leading to cirrhosis of the liver
- Cardiac arrhythmia

Organophosphorus and organonitrogen pesticides act as receptor inhibitors. They bind to the receptors of the enzyme acetylcholinesterase, which is essential for the correct functioning of the nervous system, preventing the decomposition of acetylcholine and acting through contact or systemically.[1,22] Blockage of cholinesterase activity causes the amount of acetylcholine at the synapses to increase, leading to a state of hyperarousal and paralysis of the muscles and central nervous system. If a xenobiotic binds to a receptor, one of the more important stages in metabolism may be blocked, elicit a harmful metabolic process, or alter the rate of transmission of nervous signals.[5,7,23,24] In addition, some carbamate insecticides, (e.g., carbaryl) can, in large doses, be teratogenic and can be nitrosated to form strongly carcinogenic nitroso compounds.

Methodologies for Determining Pesticides

It is essential to monitor the levels of pesticide residues as they can easily make humans more vulnerable to different diseases. An assessment of the state of contamination by pesticides requires a knowledge of the maximum permitted levels of individual active substances. The full determination procedure consists of several stages (see Figure 6). Because of the complex composition of samples and the different concentrations of the target analytes, the sample must be adequately prepared for analysis by the use of

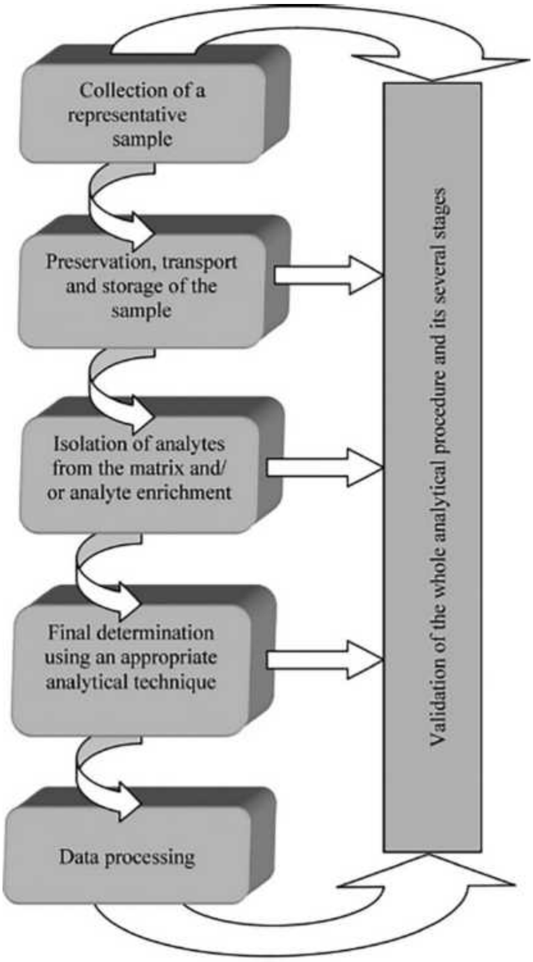


FIGURE 6 The main stages in the analytical procedure for determining modern pesticides.
Source: Biziuk et al.,[1] Beyer and Biziuk,[4] Namiesnik,[25] and Namiesnik and Górecki.[26]

techniques for efficiently extracting the target compounds and for their cleanup prior to the quantitative determination stage.

The choice of methodology for determining pesticides depends in large measure on the sample matrix and the structure and properties of the target analytes. In view of the numerous legal regulations laying down highest permissible levels of pesticides in various matrices, sensitive and selective analytical techniques are used, appropriate to the low concentrations at which the target analytes occur in them.[27] In addition, each stage in the analytical procedure, as well as this process in its entirety, should be validated.[4,26]

The analysis of environmental samples for the presence of pesticides is fraught with difficulties, because arduous and time-consuming operations and processes have to be carried out in order to prepare the samples for analysis, and these could be a source of further contamination and error.

The normally low concentrations of target analytes and the often complex matrix composition are further problems the analyst has to face. Highly efficient extraction and cleanup techniques therefore need to be applied to the target compounds prior to their quantitative determination.

Determination of Modern Pesticides in Water Samples

Collection and Preparation of Samples for Analysis

Environmental samples are diverse and complex materials to analyze because of the sampling site, type of matrix, the presence of interferents, and the concentration of target analytes. To a large extent, it depends on the appropriate preparation of the sample whether the analysis will actually provide the desired information about the sample. That is why this stage is usually a complex task, and the operations and processes involved in this stage may be a cause of analyte loss as well as a source of additional contamination. Table 6 lists the usual sample preparation processes and examples of their implementation.[1,21,25,26,28–30]

A representative sample is taken for analysis using an appropriate sampler. Which type of sampler is used depends on the type of sample, the objective of the analysis, the site and frequency of sampling, and the sample size. Mains water is usually collected in dark glass bottles; surface waters or sewage (effluent) is placed in different containers.

Techniques for Isolating and Enriching Analytes from Water Samples

Because of the low concentrations of pesticides in the various compartments of the environment, it is essential not only to isolate the organic compounds from the complex matrix but also to enrich them prior to final determination. The use of passive dosimetry at the sample collection stage, with which analytes can be isolated and preconcentrated at the same time, is worth mentioning. The free transport of mass takes place across a membrane to the sorbent as a consequence of the difference in chemical potentials of the compounds in the sorbent and the water in which the sampler has been immersed.[30,31]

TABLE 6 The Usual Sample Preparation Processes and Examples of Their Implementation

Basic Processes during the Preparation of Samples for Analysis	Examples of Implementation
Separation of suspended matter	Filtration
Chemical preservation	Reduction of the sample pH, addition of bactericides (e.g., saturated aq HgCl2, formaldehyde, sodium azide)
Physical preservation	Storage of the sample at a low temperature (2–4°C), storage in the dark, sterilization by UV irradiation

Source: Biziuk et al.,[1] Biziuk et al.,[21] Namieśnik et al.,[25] Namieśnik and Górecki[26] Namieśnik and Szefer,[28] Demkowska et al.,[29] and Łopuchin and Namieśnik.[30]

For sampling pesticides from coastal waters and rivers, one uses polyethylene dosimeters packed with iso-octane or stainless steel dosimeters filled with cyclohexane as sorbent. Three or four samplers of one type are deployed, usually for 30 days. Most passive samplers are fairly small in size, which substantially reduces the amounts of organic solvents required (the volume of the dosimeter chamber is ca. 1 mL) and makes them easy to transport and to assemble at the deployment site. With the use of passive dosimeters, there are fewer steps in the analytical procedure, which means results are more reliable and reproducible. The shortcomings of such dosimeters, however, include an insufficient sensitivity to brief fluctuations in target analyte levels and a susceptibility to environmental factors like temperature and water movement, which means that samplers have to be calibrated in order to set the rate of analyte collection.[30,31]

During extraction and/or enrichment steps, the concentration of analytes is raised in order to make their determination at all possible. Moreover, analytes are transferred from the primary matrix to a secondary one with concomitant removal of interferents. The choice of technique depends on the properties of the target analytes, their volatility, polarity, and solubility in water and organic solvents. Table 7 lists the standard techniques for extracting modern pesticide analytes.

An equally important technique is membrane extraction, which may replace classical liquid-liquid extraction (LLE) as it uses smaller quantities of solvents or even none at all.[65] As a result, a number of drawbacks are eliminated, among them, the problem of emulsion formation. Table 8 lists the membrane extraction techniques used during the determination of pesticides in water samples.[65]

TABLE 7 The Standard Techniques for Extracting Modern Pesticide Analytes

Extraction Technique	Characteristics
LLE (liquid-liquid extraction)	Based on the partition of analytes between two immiscible liquids, usually an aqueous solution and an organic solvent. The most common extraction solvents are dichloromethane[17,32–36] and mixtures of petroleum ether and dichloromethane[37] and methylene chloride and hexane·[33] Though relatively simple and cheap, this method has many shortcomings: it requires relatively large quantities of often toxic solvents, and there is a risk of an emulsion forming during agitation. To achieve the desired enrichment coefficient, the excess solvent usually has to be evaporated, and the extracts often have to be cleaned up.
DLLME (dispersive liquid–liquid microextraction)	This is one example of the miniaturization of LLE, in which a dispersing solvent, e.g., acetone or methanol (0.5–2 mL) with added extraction solvent, e.g., C2CI4, CS2 (10–50 μL), is added to the water sample (5–10 mL). This procedure yields a turbid solution, which is then centrifuged to obtain ca.5 mL of extract phase. By reducing the amounts of solvents, some of the inconveniences of LLE have been eliminated.
SPE (solid phase extraction)	This technique is based on the sorption of analytes on a sorbent—this is usually silica gel, aluminum oxide, Florisil, a porous polymer XAD-2, XAD-4, XAD-7, XAD-16, usually modified with octadecyl groups·[15,33,38–51] Desorption of adsorbed analytes is usually carried out using methanol,[15,37,40,43,46,51] ethyl acetate,[33,35,44] dichloromethane,[38,39,41,50] and mixtures of methanol and water,[45,49,51] methanol and acetonitrile,[42,47] and acetonitrile and water.[48] Research is in progress to develop selective sorbents (MISPE: *Molecularly Imprinted Solid Phase Extraction; immunosorbents*), which would be applied in the case of samples with a complex matrix composition such as sewage.
SPME (solid phase microextraction)	Simple to use and does not require solvents. Based on extraction from the liquid phase to an adsorption or absorption layer coated on an extraction fiber. It is also possible to introduce the fiber into the headspace (HS-SPME). The materials used for coating fibers include polydimethylsiloxane (PDMS),[53–58] polyacrylate (PA),[48,54–57,59,60] and also mixtures of polydimethylsiloxane and polydivinylbenzene (PDMS-DVB),[53–55,60–64] carbowax and polydivinylbenzene (CW-DVB),[53,57,64] and carbowax and molecularly imprinted resin (CW-TPR).[60] The analytes most frequently adsorbed are transferred to the GC injector where they are thermally desorbed and then determined.
SBSE (stir bar sorptive extraction)	The extractant, usually polydimethylsiloxane (PDMS), is used to coat a magnetic stir bar.[29] The adsorbed analytes are then thermally desorbed. The parameters determining the effectiveness of this type of extraction are the sample volume, the extraction time, and the stirring speed.

TABLE 8 Membrane Extraction Techniques Used during the Determination of Pesticides in Water Samples

Technique	Type of Membrane	Combination of Phases Used: Donor/Membrane/Acceptor
MASE: membrane-assisted solvent extraction; PME: polymer membrane extraction	Non-porous	Aqueous/polymer/aqueous Organic/polymer/aqueous Aqueous/polymer/organic
MMLLE: microporous membrane liquid–liquid extraction	Non-porous (microporous)	Aqueous /organic/organic Organic/organic/aqueous

Source: Jönsson and Mathiasson.[65]

Determination of Modern Pesticides in Ambient Air

Sampling Techniques

Pesticides in the air can be determined in the following types of samples:

- Atmospheric dusts
- Analytes sampled from the gaseous phase

Dynamic[66] and passive (diffusional)[67,68] sampling methods are the usual ways of collecting air for determining its pesticide content.

Dynamic sampling enables the collection of pesticides present in the gaseous phase and in suspended form by pumping air through a filter coated with a solid adsorbent. Pesticides present in suspended dust particles are retained on the filter, while those in the gaseous phase are adsorbed by the sorbent.

The standard filters are of glass with different diameters (Ø = 30 cm, 10 cm, 90 mm, and 25 mm) and of quartz filters (Ø = 102 and 150 mm).

For collecting gas samples, the sorbents (XAD-2, XAD-4, Carbopack, Carbotrap, Carboxen, Tenax TA, Chromo- sorb, silica gel) are coated onto the filters, as mentioned above, or placed in stainless steel or glass vessels containing polyurethane foam (PUF).

The conventional methods of collecting gas samples for the determination of their pesticide content involve passing a fixed volume of air through a solid sorbent, for which pumps and flowmeters are needed. The expense of the former and the need to frequently calibrate the latter mean that it is difficult to collect gas samples in a fully professional manner. Analytes retained on sorbents require not only thermal desorption but also chemical desorption with expensive and potentially toxic solvents. The sampling time depends on the method's sensitivity and the breakthrough volume of the sorbents by the target analytes (Table 9). In order to collect samples without the need to use toxic solvents and expensive pumping equipment, we need an unpowered air sampler.[69]

Universal and inexpensive passive air samplers (PASs) have been developed as an alternative to the conventional active sampling methods. Passive sampling is based on the free flow of analytes from the environment sampled to a medium on which the analytes are sorbed (Table 10).

There are several types of PASs, using the following:

- SPMD semipermeable membranes.
- PUF discs; the passive samples consist of PUF discs placed inside stainless steel containers covered with a dome-shaped lid to reduce the effect of wind speed on the rate of sampling. This arrangement is also used to protect the PUF discs from precipitation, the direct deposition of dust, and UV light.
- Resins.
- A thin layer of ethylene/vinyl acetate as the sampling medium.

TABLE 9 Comparison of Techniques for Collecting Air Samples Depending on the Sampling Time, Type of Filter, and Adsorbent

Sampling Time (hr)	Type of Filter	Adsorbent	References
5.5–9	GFF	XAD-2, s-PUF	[70]
24	GFF	XAD-2	[66–68,71,72]
24	GFF	PUF	[73–76]
48	GFF	XAD-2	[77]
18–67	GFF	PUF	[78]
84	GFF	XAD-2, s-PUF	[79–81]
168	GFF	XAD-2, s-PUF	[82–85]
12	QFF	XAD-2, s-PUF	[86]
23	QFF	PUF	[87]
24	QFF	PUF	[88–92]
24	QFF	XAD-2	[93–95]

TABLE 10 List of Techniques for the Passive Sampling of Air

Type	Sampling Time	References
PUF discs	4 months	[96–98]
PUF discs	6 weeks	[99]
POG	7 days	[100,101]
PUF discs	2 months	[102]
XAD-2	1 year	[103]
XAD-2	5–8 months	[104]
PUF discs	28 days	[105–108]
PDMS	14 days	[109]
SPMD	7 days	[98,110]

Solid-phase microextraction (SPME) is another way of sampling analytes from air. A solventless method, it is very convenient for use in the field, simple to operate, and, under optimal conditions, does not require the use of pumps.

When only suspended dust is to be sampled, the samples are often fractionated according to particle diameter, a process that is possible because the different fractions have different physical and chemical properties: the fine-grained fraction (<2.5 pm) is often acidic, whereas the coarse-grained fraction (>2.5 pm) is basic. Fractionation minimizes interactions between particles of different pH. Cascade impactors (absorbers) are also used. These operate on the principle that the dust-containing air is made to flow through nozzles of ever-decreasing diameter. At each nozzle, the linear velocity of the airflow increases and the relevant dust fractions are separated at the steps of the cascade impactor. If dust concentrations in gas samples are very high, cascade cyclones are used. These direct the dust-containing gas at a tangent to a cylinder, and the particles move under centrifugal force from the stream of gas to the inside wall of the cylinder.[111]

Methods of Extracting Analytes from Sorbents and Filters

The next step, after sampling but before the final determination, involves extracting the preconcentrated analytes from the solid sorbents or filters (Table 11). This is the most crucial step in the analytical process, since the aim of extraction is to liberate the greatest possible quantities of pesticides from the sorbent. To this end, both solvent and solventless techniques are employed.

TABLE 11 Techniques Most Frequently Used for Extracting Analytes from Air Samples

Sample	Extraction Technique	References
PM2.5 dust	Accelerated solvent extraction (ASE, PFE, PLE)	[76,112]
Dust and gaseous phase	Soxhlet apparatus	[66,67,71,7, 113–119,121]
Gaseous phase	Agitation assisted (LE)	[122]
Gaseous phase	Ultrasound assisted (UE)	[123]
PM2.5 dust	Microwave assisted (MAE)	[124]
Gaseous phase	Thermal desorption	[125]

TABLE 12 Solvents Most Commonly Used for Extracting Analytes from Air Samples

Sample	Solvent	References
Suspended dust	Petroleum ether	[74,88,96]
Suspended dust	Acetone	[76,79,80,83]
Suspended dust/air	Hexane/acetone	[113]
Suspended dust	Hexane/dichloromehane	[66,67,71,72,75,114–116]
Suspended dust/air	Dichloromethane/ petrochemical ether (diethyl ether, MTBE)	[117–119,123]
Suspended dust/air	Hexane/benzene	[121]
Suspended dust	Dichloromethane	[112]
Air	Ethyl acetate	[122]

The most commonly used solvents for extracting analytes from air samples are shown in Table 12.

The next step in sample preparation is evaporation and/or change of solvent. It is standard practice to evaporate the solvent in a vacuum evaporator, a quick and simple method. Another possibility is to evaporate the solvent in a stream of gas, usually nitrogen, again simple to carry out. These methods may be applied separately or in combination.

Extract Cleanup

Extract cleanup involves fractionating the extract, which can be done in a number of ways:

- With column LC chromatography with normal phase (NP) or reversed phase (RP), including high-performance liquid chromatography (HPLC)
- With adsorption chromatography
- Passing the extract through a filter
- With gel chromatography (GPC)

In some cases, mainly when highly selective detectors like gas chromatography-tandem mass spectrometry (GC–MS/ MS) and liquid chromatography–tandem mass spectrometry (LC–MS/MS) are used, the cleanup of extracts is not necessary. Before deciding whether the cleanup step can be omitted or not, however, matrix effects must be investigated for possible interferents and their influence on the apparatus.

Determination of Modern Pesticides in Fruit and Vegetables

Sample Collection and Preparation of Samples for Analysis

It is extremely important that samples of material intended for analysis are homogeneous and representative; they should be stored frozen and in the dark.[28] The sample preparation process consists of several steps. In the case of fruit and vegetables, one of these steps involves the removal of surface contaminants by, for example, washing the sample in distilled water, after which the sample is dried (at an elevated temperature, at ambient temperature, or with the aid of a desiccant). Then, the sample has to be broken up, crumbled, minced, or ground with a pestle and mortar. It is then homogenized in special equipment. The sampling process depends on the type of biological material under investigation, and one has always to be mindful of possible losses of analytes and contamination of the sample as a result of human agencies.

Techniques for Isolating Pesticides from Fruit and Vegetable Samples

Isolation and/or enrichment involves the transfer of analytes from the primary matrix to a secondary one with the concomitant removal of interferents and the increase of analyte concentrations to levels above the limit of detection (LOD) of the analytical technique used. Since pesticide concentrations in fruit and vegetables are low, samples often require enrichment.[126,127]

The following are the usual techniques for isolating pesticides from fruit and vegetable samples:

- Liquid-liquid extraction (LLE)
- Accelerated solvent extraction (ASE)
- Supercritical fluid extraction (SFE)
- Soxhlet or Soxtec extraction
- Microwave-assisted extraction (MAE)

The merits of solvent extraction by agitation are ease of implementation and low cost; in addition, expensive and complicated apparatus can be dispensed with. Unfortunately, however, this technique requires a large quantity of toxic solvents and its selectivity leaves much to be desired. Soxhlet extraction are easy to carry out and enables the extract to be separated from the residue; on the downside, this type of extraction takes a long time with large amounts of solvent, and only single samples can be extracted. For this reason, the more efficient Soxtec apparatus was introduced; with this albeit very expensive apparatus, the extraction time is now much shorter, and several samples can be extracted simultaneously. Solvent extraction techniques are frequently assisted with microwaves or ultrasound in order to improve extraction yields. Acetone, acetonitrile, and ethyl acetate are the solvents (or their mixtures) (Table 13) most commonly used for the extraction of fruit and vegetable samples.

Extract Cleanup

The cleanup of the extract is an essential step that should always precede its analysis. It is important, because during isolation, we obtain not only the target analytes but also in- terferents, which can distort the final result of the analysis. The usual cleanup techniques used with fruit and vegetable samples are as follows (Table 14):

- Solid-phase extraction (SPE)[131]
- Solid-phase microextraction (SPME)[132]
- Matrix solid-phase dispersion extraction (MSPDE)[133]
- Stir bar solvent extraction (SBSE)[134]
- Gel chromatography (GPC)[135–137]

TABLE 13 Solvents for the Extraction of Samples of Fruit and Vegetables

Sample	Solvent
Cabbage, grapes	Ethyl acetate, methanol, acetone[128]
Oranges, apples, grapes, pears	Ethyl acetate[126]
Apples, tomatoes	Acetone[127]
Tomatoes, pears, oranges	Acetonitrile[129]
Peppers, tomatoes, Brussels sprouts, melons, apples, lemons	Ethyl acetate[130]

TABLE 14 Extract Cleanup Techniques in the Determination of Pesticides in Fruit and Vegetables

Sample	Extract Cleanup Technique
Oranges, apples, grapes, pears	mspd[126]
Apples, tomatoes	hs-spme[127]
Oranges, apples, peaches, pears, tomatoes, lettuce, potatoes, leeks	gpc[140]
Tomatoes, pears, oranges	spe[129]
Vegetables	spme[141–144]
Oranges	MSPD, SBSE[145]
Strawberries, cherries	hs-spme[146]
Oranges, tangerines, grapefruits, lemons	mspd[147]

SPE is currently the most popular extract cleanup tech- nique.[138] It involves passing the sample through a bed of sorbent on which the target analytes are adsorbed. The retained compounds are then liberated by means of a solvent and analyzed. This method is easy to carry out and readily lends itself to automation. Its shortcomings, however, are that the sorbent bed has to be conditioned before each use and that analyte yields are low. A modification of SPE, i.e., SPME,[139] involves the adsorption of analytes on a fiber coated with a suitable solid phase that can be pushed out of a microsyringe. The analyte is then thermally desorbed and transferred to the GC injector. The benefits of this method are that solvents can be eliminated and that it is impossible to overload the column because of the limited volume of adsorbent. Optimization of the method is a problem, however. Depending on where the fiber is placed in relation to the sample, SPME can be divided into direct (Direct Immersion [DI]-SPME) or headspace (HS-SPME).

GPC is just as frequently used as a cleanup technique. It enables micromolecular pesticides to be separated from macromolecular substances present in the matrix. While the considerable longevity of the columns is an advantage, the poorer resolution in comparison with adsorption techniques, especially when gradient elution techniques are used, is a disadvantage.

Approaches are being sought to develop pesticide determination techniques that are quick, simple, cheap, effective, and safe. QuEChERS (quick, easy, cheap, effective, rugged, and safe)[148–157] is one such method. It is based on a number of stages (see Figure 7).

The consumption of sample and toxic solvents with the QuEChERS method is minimal. By applying QuEChERS to the determination of pesticides in fruit and vegetables, matrix effects are eliminated and high recoveries of target analytes are possible. The method can be modified depending on the type of sample and the target analytes. To improve the extraction of polar organophosphorus pesticides, the method is modified by the addition of acetic acid. When samples of citrus fruit are under investigation, protective wax coatings can be removed by freezing the samples for at least 1 hr. For the analysis of citrus fruits, blackcurrants, and raspberries, it is recommended to add aq NaOH to reach pH = 5 and to improve the analysis.

Determination of Recently Applicable Modern Pesticides in Soil Samples[158–161]

Sample Collection and Preparation of Samples for Analysis

In order to determine the chemical composition of soil samples, it is necessary to preserve their natural structure. They are cut by manual scoop or a similar tool from the central, typical part of the material. In the next step, the material is introduced to a special cylinder of known volume. After this operation, the drying of the sample is desirable.

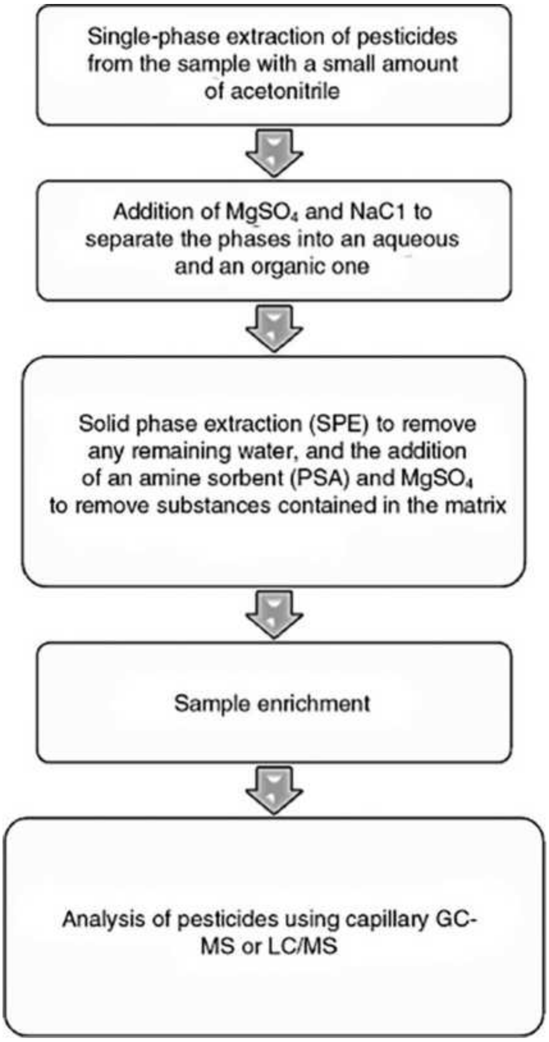


FIGURE 7 Stages in the determination of pesticides using the QuEChERS method.

The drying step consists of two stages:

- Preliminary drying (also drying with circulation air masses) in temperature from 40°C to 100°C/105°C and drying in reduced temperature (lyophilization)
- Final drying, drying by chemical binding water, and drying using water adsorption

Techniques of Isolation and/or Enrichment Analytes from Soil Samples

For extracting the analytes from soil samples, the ultrasound-assisted liquid extraction (shaking)[162] or conventional Soxhlet extraction[163,164] can be used. In these cases, a mixture of non-polar and polar solvents is used. In addition, the extraction of analytes from soil samples can be investigated:

- Using supercritical fluid in-process—critical conditions are obtained by using highly compressed gases (the most compressed being CO2) in the critical temperature range.[165,166]

- Using microwave process—microwave energy is absorbed by molecule compounds.
- As an ASE—the process runs at elevated temperature and pressure, providing appropriate solvent condition (liquid state of solvent).

Evaporation of the solvent is the next step, which provides the enrichment of the extract. Among the known methods for enriching the extract, the following are the most popular:

- Evaporation of the solvent in vacuum evaporator (vacuum distillation)
- Evaporation of the solvent in Kudern-Danish apparatus
- Evaporation of the solvent in a gas stream (nitrogen or air)

An important step in the preparation of the soil samples is removal of interfering substances. For elimination of sulfur, it is possible to use various chemical reactions, for example, reactions with sodium sulfate (IV), and aluminum oxides, or reactions with heavy metals. Aside from the use of chemical reactions, the methods of electrophoresis, chromatography, and saponification are also notable in the removal of interfering substances. To remove fats from soil samples, LLE,[167] adsorption chromatography, saponification, decomposition reaction using sulfuric acid at low temperature, precipitation, GPC,[168] and semipreparative HPLC can be applied.

Purification of the Extract

The purification process of extracts, which contain pesticides, includes fractionation of the extract. It can be carried out by different methods. Among the available methods are adsorption chromatography, GPC, and column chromatography.[169]

Derivatization of Pesticides

By converting pesticide compounds into their derivatives, selectivity, analyte enrichment, target analyte resolution in the column, and selectivity of detection are all improved. Derivatization also enables a larger number of compounds to be determined in the sample. The inference is that the isolation of target analytes from the matrix and their enrichment are not always effective without changing the chemical structure of the analytes or the matrix.[1] Pesticides to be determined by GC are usually derivatized with pentafluorobenzyl bromide (PFBB).[67,72,75]

Final Determination

The last stage in the analytical procedure is the identification of compounds and their quantitative determination using an appropriate analytical technique. The usual techniques involved include the following:

- Capillary GC. Pesticides to be determined by GC should be volatile and thermally stable. During capillary GC, the injector can be operated in split mode (when the stream of carrier gas is divided) or splitless mode (when the carrier gas stream is not divided); either mode can be used for determining pesticides in environmental samples. The on-column injector, with which samples can be injected directly into the chromatographic column, is also used for the GC determination of pesticides.
- HPLC, usually in reversed-phase mode, is used for determining pesticides that cannot be determined by GC, for example, thermally unstable polar compounds like herbicides, carbamates, triazines, and also compounds that require derivatization.

Gas chromatography can be used to determine the residues of all pesticide classes. The choice of chromatographic column is extremely important as regards the separation of analytes and their qualitative and quantitative determination. The column should be highly efficient and be resistant to changes in separation parameters. The solid (stationary) phase should be thermally stable and be highly selective with respect to the constituents of the mixture being analyzed. The following detectors are used (Table 15):[157,170–172]

- ECD (electron capture detector)—highly sensitive in relation to compounds containing electronegative atoms
- FPD (flame photometric detector)—applied in the determination of organophosphorus compounds
- NPD (nitrogen phosphorus detector)—used for the simultaneous determination of organonitrogen and or- ganophosphorus pesticides
- TSD (thermionic specific detector)—used to determine compounds containing nitrogen or phosphorus
- MS (mass spectrometry) or MS/MS (tandem mass spectrometry)
- UV–DAD (ultraviolet–diode array detector)—used in LC

Once a target analyte has been detected quantitatively, for example, by GC-NPD, the result must be confirmed by another independent method. The conditions of the process can be altered, say, by changing the temperature program or by using a different chromatographic column. It is crucial to obtain confirmation by another method as identification based solely on retention times is insufficient. MS/MS improves sensitivity and selectivity of analytical methods. In this technique, ions that were separated in the first analyzer are again fragmented and the derivative ions were analyzed in the second one. The chromatogram background is reduced, as a result of which the signal value is enhanced with respect to noise and the LOD of the target analytes is lowered.[174–176] Better chromatographic peak resolution and a

TABLE 15 Characteristics of Detectors Used for the Determination of Pesticide Analytes

Type of Detector	Merits	Drawbacks
Mass spectrometer (MS and tandem MS)	• Both quantitative and qualitative information obtainable • Highly sensitive • Used with both GC and LC • Applicable to trace determinations, in complex matrices • Can be universal or selective, as the need arises • Has a wide range of application used in multimethods for determining pesticides of different classes • Better sensitivity and greater specificity sometimes obtainable with GC coupled to tandem MS (MS/MS) • Can operate in single ion monitoring (SIM) mode, which provides for greater sensitivity than the SCAN mode • A highly specific detector; in the case of pesticides, used to determine organochlorine compounds	• Expensive
Electron capture detector (ECD)	• A detector suitable for selected classes of compounds	• Applicable only to compounds containing electronegative atoms
Thermoionic detector (NPD)	• Has a wide range of application	• Useful only for compounds containing nitrogen and phosphorus
Flame ionization detector (FID)	• Suitable for almost all organic compounds • Highly sensitive and with a wide range of linearity • Highly sensitive and with a wide range of linearity	• The sample is destroyed

Source: http://www.pg.gda.pl.[173]

smaller influence of the matrix on the final result can also be achieved using two-dimensional (2D) gas chromatography (GC×GC). This uses two columns: the partially separated constituents from the first column are further separated in the second one by a different mechanism.

Fast GC is equally frequently used to shorten the time of analysis and to obtain better peak resolution. Compared to classical GC, it requires shorter capillary columns with a smaller diameter and solid-phase films ca. 0.1 μm in thickness, as well as a faster flow rate and higher pressure of the carrier gas. These parameters yield determination results of a better precision.[174,177,178]

The trend at present is to develop analytical methods enabling a broad spectrum of analytes to be determined in a single analytical run (MRMs—multiresidue methods). However, the problem here is that the compounds to be determined simultaneously, often present at low concentrations, have different physicochemical properties depending on their chemical structure. Such a methodology, apart from being able to determine a large number of compounds in one run, should:

- Ensure maximum removal of interferents from extracts
- Give large recoveries of target compounds, high sensitivity, and good precision
- Be environmentally friendly, i.e., require the smallest possible quantities of samples and chemical reagents, especially organic solvents
- Be cheap, quick, and easy to carry out

Research continues for the improvement of existing analytical methods and the development of new ones capable of supplying reliable results for a wide range of analytes.

Summary

The increasingly widespread application of pesticides— substances with different physical and chemical proper- ties—means that ever-larger amounts of these compounds are getting into the environment. As a result of the various processes they are subjected to, they may be converted to even more toxic compounds. They are currently regarded as some of the most dangerous environmental contaminants because of their stability, mobility, and long-term effects on living organisms. Moreover, they are usually present in the environment at very low concentrations in complex matrices, which makes their analysis difficult. Their determination is therefore prolonged by the need to perform time-consuming and often arduous processes in order to prepare samples for analysis, which themselves may be sources of further contamination and error. That is why for the analysis of pesticides in the environment, procedures that enable a large number of compounds to be detected simultaneously are required. It is often the case at present that target pesticide analytes have to be isolated from the matrix and enrichment before their final determination can be undertaken. To counteract the adverse effects of pesticides on human health, the continuous monitoring of their presence in the environment is fundamental. European Union recommendations are in place for analyzing the quality of water, air, and food in this respect. Existing techniques are being improved and new ones are being developed so that different classes of pesticides can be reliably determined in a quick, simple, cheap, and environmentally friendly manner.

References

1. Biziuk, M.; Hupka, J.; Wardencki, W.; Zygmunt, B.; Siłło- wiecki, A.; Żelechowska, A.; Dąbrowski, Ł.; Wiergowski, M.; Zaleska, A.; Tyszkiewicz, H. *Pestycydy, występowanie, oznaczanie i unieszkodliwianie;* Wydawnictwo Naukowo- Techniczne WNT: Warszawa, 2001.
2. Sudo, M.; Kawachi, T.; Hida, Y.; Kunimatsu, T. Spatial distribution and seasonal changes of pesticides in lake Biwa, Japan. Limnology **2004**, *5*, 77–86.
3. Sosnowska, K.; Styszko-Grochowiak, K.; Gołlaś, J. Nowe zanieczyszczenia w środowisku wodnym-źródlła, zagroże- nia, problemy analityczne. Analityka **2009**, *4*, 44–48.

4. Beyer, A.; Biziuk, M. Przegląd metod oznaczania pozosta- łości pestycydów i polichlorowanych bifenyli w próbkach zywności. Ecol. Chem. Eng. **2007**, *14,* 35–58.
5. Moreno, J.L.; Liebanas, F.J.; Frenich, A.G.; Vidal, J.L. Evaluation of different sample treatments for determining pesticide residues in fat vegetable matrices like avocado by low-pressure gas chromatography–tandem mass spectrometry. J. Chromatogr. A. **2006**, *1111*, 97–105.
6. Hajslova, J.; Zrostlikova, J. Matrix effects in (ultra)trace analysis of pesticide residues in food and biotic matrices. J. Chromatogr. A. **2003**, *1000,* 181–197.
7. Bartoszek-Pączkowska, A.; Bontemps-Gracz, M.; Mazerska, Z.; Biziuk, M.; Czerwiński, J.; Namieśnik, J.; Zasławska, żelechowska, A.; Czarnowski, W.; Jaśkowski, J., *Zarys Ekotoksykologii;* Eko-Pharma: Gdańsk, 1995.
8. Biziuk, M.; Buszewski, B.; Chrzanowski, W.; Czerwiński, J.; Gazda, K.; G,órecki, T.; Jamrógiewicz, Z.; Janicki, W.; Konieczka, P.; Kot-Wasik, A.; Łukasiak, J.; Makuch, B.; Namieśnik, J.; Pawliszyn, J.; Polkowska, Ż.; Przyjazny, A.; Sieńkowska-Zyskowska, E.; Wardencki, W.; Wolska, L.; Zasławska, L.; Zygmunt, B. *Fizykochemiczne metody kon- troli zanieczyszczń środowiska;* Wydawnictwo Naukowo- Techniczne WNT: Warszawa, 1998.
9. Ding, X.; Wang, X.; Wang, Q.; Xie, Z.; Xiang, C.; Mai, B.; Sun, L. Atmospheric DDTs over the North Pacific Ocean and the adjacent Arctic region: Spatial distribution, congener patterns and Skurce implication. Atmos. Environ. **2009**, *43,* 4319–4326.
10. Available at http://www.minddisrupted.org/documents/ MD%20 Pesticide%20Fact.pdf (accessed May 16, 2012).
11. Papadakis, E.N.; Vryzas, Z.; Papadopoulou-Mourkidou, E. Rapid method for the determination of 16 organochlorine pesticides in sesame seeds by microwave-assisted extraction and analysis of extracts by gas chromatography–mass spectrometry. J. Chromatogr. A. **2006**, *1127,* 6–11.
12. Chen, S.; Shi, L.; Shan, Z.; Hu, Q. Determination of organo- chlorine pesticide residues in rice and human and fish fat by simplified two-dimensional gas chromatography. Food Chem. **2007**, *104,* 1315–1319.
13. Kodba, Z.C.; Koncina, D.B. A rapid method for the determination of organochlorine, pyrethroid pesticides and polychlorobiphenyls in fatty foods using GC with electron capture detection. Chromatographia **2007**, *66,* 619–624.
14. Escuderos-Morenas, M.L.; Santos-Delgado, M.J.; Rubio- Barroso, S.; Polo-Diez, L.M. Direct determination of monolinuron,linuron and chlorbromuron residues in potato samples by gas chromatography with q nitrogen–phosphorus detection. J. Chromatogr. A. **2003**, *1011,* 143–153.
15. Guardia Rubio, M.; Ruiz Medina, A.; Pascual Reguera, M. I.; Fernández de Córdova, M.L. Multiresidue analysis of three groups of pesticides in washing waters from olive processing by solid-phase extraction–gas chromatography with electron capture and thermionic specific detection. Mi- crochem. J. **2007**, *85,* 257–264.
16. Santos, F.J.; Galceran, M.T. The application of gas chromatography to environmental analysis. Trends Anal. Chem. **2002**, *21,* 672–685.
17. Lartiges, S.B.; Garrignes, P. Gas chromatographic analysis of organophosphorus and organonitrogen pesticides with different detectors. Analysis **1995**, *23,* 418–421.
18. Zhang, J.; Lee, H.K. Application of liquid-phase microextraction and on-column derivatization combined with gas chromatography–mass spectrometry to the determination of carbamate pesticides. J. Chromatogr. A **2006**, *1117,* 31–37.
19. Sabik, H.; Jeannot, R.; Rondeau, B. Multiresidue methods using solid-phase extraction techniques for monitoring priority pesticides, including triazines and degradation products, in ground and surface waters. J. Chromatogr. A **2000**, *885,* 217–236.
20. Available at http://www.farmer.pl/srodki-produkcji/ochrona roslin/jak_zwalczac_stonke_ziemniaczana_,d1f3f3d8d5452695e416.html (accessed February 2010).
21. Biziuk, M.; Przyjazny, A.; Czerwiński, J.; Wiergowski, M. Occurrence and determination of pesticides in natural and treated waters. J. Chromatogr. A. **1996**, *754,* 103–123.

22. Jokanovic, M. Medical treatment of acute poisoning with organophosphorus and carbamate pesticides. Toxicol. Lett. **2009**, *190,* 107–115.
23. Jokanovic, M. Medical treatment of acute poisoning with organophosphorus and carbamate pesticides. Toxicol. Lett. **2009**, *190,* 107–115.
24. Jokanovic, M. Biotransformation of organophosphorus compounds. Toxicology **2001**, *166,* 139–160.
25. Namiesnik, J.; Jamrógiewicz, Z.; Pilarczyk, M.; Torres, L. *Przygotowanie próbek środowiskowych do analizy*; Wy- dawnictwo Naukowo-Techniczne WNT: Warszawa, 2000.
26. Namieśnik, J.; Górecki, T. Preparation of environmental samples for the determination of trace constituents. Pol. J. Environ. Stud. **2001**, *10,* 77–84.
27. Directive of the European Union 98/83/EC of 3.XII.1998 r. on the quality of water intended for human consumption, Available at http://www.scribd.com/doc/8264629/Directive-9883EC-Drinking-water-standards (accessed February 2010).
28. Namieśnik, J.; Szefer, P. Preparing samples for analysis—The key to analytical success. Ecol. Chem. Eng. **2008**, *15,* 167–249.
29. Demkowska, I.; Polkowska, Ż.; Namieśnik, J. Ekstrakcja z wykorzystaniem ruchomego elementu sorpcyjnego typu Twister™ (SBSE). Analityka **2009**, *4*, 8–11.
30. Łopuchin, E.; Namieśnik, J. Dozymetria pasywna, dogodne narzqdzie do badania zanieczyszczń środowiska wodnego. Analityka **2009**, *1*, 29–36.
31. Namieśnik, J.; Chrzanowski, W.; Szpinek, P.; Bode, P.; Brzózka, Z.; Bulska, E.; Buszewski, B.; Gazda, K.; Górecki, T.; Jönsson, J.Å.; Kościelniak, P.; Levsen, K.; Lopez de Alda, M.J.; Matusiewicz, H.; Nawrocki, J.; Paschke, A.; Sandra, P.; Szefer, P.; Szpunar, J.; Trojanowicz, M.; Wal- kowiak, A.; Wardencki, W. *New Horizons and Challenges in Environmental Analysis and Monitoring;* CEEAM: Gdańsk, 2003.
32. Sankararamakrishnan, N.; Sharma, A.K.; Sanghi, R. Or- ganochlorine and organophosphorous pesticide residues in ground water and surface waters of Kanpur, Uttar Pradesh, India. Environ. Int. **2005**, *31,* 113–120.
33. *Standardized Analytical Methods for Environmental Restoration Following Homeland Security Events-Revision 5.0;* United States Environmental Protection Agency (EPA): Cincinnati, USA, 2009.
34. Tse, H.; Comba, M.; Alaee, M. Method for the determination of organophosphate insecticides in water, sediment and biota. Chemosphere **2004**, *54,* 41–47.
35. Sabik, H.; Jeannot, R. Determination of organonitrogen pesticides in large volumes of surface water by liquidliquid and solid-phase extraction using gas chromatography with nitrogen–phosphorus detection and liquid chromatography with atmospheric pressure chemical ionization mass spectrometry. J. Chromatogr. A **1998**, *818,* 197–207.
36. Sabik, H.; Fouquet, A.; Proulx, S. Ultratrace determination of organophosphorus and organonitrogen pesticides in surface water. Analysis **1997**, *25,* 267–273.
37. Tahboub, Y.R.; Zaater, M.F.; Al-Talla, Z.A. Determination of the limits of identification and quantitation of selected or- ganochlorine and organophosphorous pesticide residues in surface water by full-scan gas chromatography/mass spectrometry. J. Chromatogr. A **2005**, *1098,* 150–155.
38. Polkowska, Ż.; Tobiszewski, M.; Górecki, T.; Namieśnik, J. Pesticides in rain and roof runoff waters from an urban region. Urban Water J. **2009**, *6*, 441–448.
39. Polkowska, Ż.; Górecki, T.; Namieśnik, J. Quality of roof runoff waters from an urban region (Gdańsk, Poland). Chemosphere **2002**, *49*, 1275–1283.
40. Grynkiewicz, M.; Polkowska, Ż.; Górecki, T.; Namieśnik, J. Pesticides in precipitation from an urban region in Poland (Gdańsk–Sopot–Gdynia Tricity) between 1998 and 2000. Water Air Soil Pollut. **2003**, *149,* 3–16.
41. Polkowska, Ż.; Kot, A.; Wiergowski, M.; Wolska, L.; Wolowska, K.; Namieśnik, J. Organic pollutants in precipitation: Determination of pesticides and polycyclic aromatic hydrocarbons in Gdańsk, Poland. Atmos. Environ. **2000**, *34,* 1233–1245.

42. Grynkiewicz, M.; Polkowska, Ż.; Górecki, T.; Namieśnik, J. Pesticides in precipitation in the Gdansk region (Poland). Chemosphere **2001**, *43,* 303–312.
43. El-Kabbany, S.; Rashed, M.M.; Zayed, M.A. Monitoring of the pesticide levels in some water supplies and agricultural land, in El-Haram, Giza (A.R.E.). J. Hazard. Mater. A **2000**, *72*, 11–21.
44. Ballesteros, E.; Parrado, M.J. Continuous solid-phase extraction and gas chromatographic determination of or- ganophosphorus pesticides in natural and drinking waters. J. Chromatogr. A **2004**, *1029,* 267–273.
45. Baugros, J.B.; Giroud, B.; Dessalces, G.; Grenier-Lousta- lot, M.F.; Cren-Olivé, C. Multiresidue analytical methods for the ultra-trace quantification of 33 priority substances present in the list of REACH in real water samples. Anal. Chim. Acta **2008**, *607,* 191–203.
46. Lyytikäinen, M.; Kukkonen, J.V.K.; Lydy, M.J. Analysis of pesticides in water and sediment under different storage conditions using gas chromatography. Arch. Environ. Con- tam. Toxicol. **2003**, *44*, 437–444.
47. Nogueira, J.M.F.; Sandra, T.; Sandra, P. Multiresidue screening of neutral pesticides in water samples by high performance liquid chromatography–electrospray mass spectrometry. Anal. Chim. Acta **2004**, *505,* 209–215.
48. Martínez, R.C.; Hermida, C.G.; Gonzalo, E.R.; Miguel, L.R. Behaviour of carbamate pesticides in gas chromatography and their determination with solid-phase extraction and solid-phase microextraction as preconcentration steps. J. Sep. Sci. **2005**, *28,* 2130–2138.
49. Liu, W.; Lee, H.K. Quantitative analysis of pesticides by capillary column high performance liquid chromatography combined with solid-phase extraction. Talanta **1998**, *45,* 631–639.
50. Potter, T.L.; Mohamed, M.A.; Ali, H. Solid-phase extraction combined with high-performance liquid chromatography-atmospheric pressure chemical ionization-mass spectrometry analysis of pesticides in water: Method performance and application in a reconnaissance survey of residues in drinking water in Greater Cairo, Egypt. J. Agric. Food Chem. **2007**, *55,* 204–210.
51. López, F.J.; Beltran, J.; Forcada, M.; Hernández, F. Comparison of simplified methods for pesticide residue analysis use of large-volume injection in capillary gas chromatography. J. Chromatogr. A **1998**, *823,* 25–33.
52. Hu, J.Y.; Aizawa, T.; Magara, Y. Analysis of pesticides in water with liquid chromatography/atmospheric pressure chemical ionization mass spectrometry. Water Res. **1999**, *33*, 417–425.
53. Gonçalves, C.; Alpendurada, M.F. Multiresidue method for the simultaneous determination of four groups of pesticides in ground and drinking waters, using solid-phase microextraction-gas chromatography with electron-capture and thermionic specific detection. J. Chromatogr. A **2002**, *968,* 177–190.
54. Yao, Z.; Jiang, G.; Liu, J.; Cheng, W. Application of solidphase microextraction for the determination of organophosphorous pesticides in aqueous samples by gas chromatography with flame photometric detector. Talanta **2001**, *55*, 807–814.
55. Frías, S.; Rodríguez, M.A.; Conde, J.E.; Pérez-Trujillo, J.P. Optimisation of a solid-phase microextraction procedure for the determination of triazines in water with gas chromatography-mass spectrometry detection. J. Chromatogr. A **2003**, *1007,* 127–135.
56. Su, P.; Huang, S. Determination of organophosphorus pesticides in water by solid-phase microextraction. Talanta **1999**, *49,* 393–402.
57. Natangelo, M.; Tavazzi, S.; Fanelli, R.; Benfenati, E. Analysis of some pesticides in water samples using solid-phase microextraction–gas chromatography with different mass spectrometric techniques. J. Chromatogr. A **1999**, *859,* 193–201.
58. Rocha, C.; Pappas, E.A.; Huang, Ch. Determination of trace triazine and chloroacetamide herbicides in tile-fed drainage ditch water using solid-phase microextraction coupled with GC–MS. Environ. Pollut. **2008**, *152,* 239–244.
59. Magdic, S.; Boyd-Boland, A.; Jinno, K.; Pawliszyn, J.B. Analysis of organophosphorus insecticides from environmental samples using solid-phase microextraction. J. Chromatogr. A **1996**, *736,* 219–228.

60. Volmer, D.A.; Hui, J.P.M. Rapid SPME/LC/MS/MS analysis of *N*-methylcarbamate pesticides in water. Arch. Environ. Contam. Toxicol. **1998**, *35*, 1–7.
61. Tomkins, B.A.; Ilgner, R.H. Determination of atrazine and four organophosphorus pesticides in ground water using solid phase microextraction (SPME) followed by gas chromatography with selected-ion monitoring. J. Chromatogr. A **2002**, *972,* 183–194.
62. Gonçalves, C.; Alpendurada, M.F. Solid-phase microextraction-gas chromatography–(tandem) mass spectrometry as a tool for pesticide residue analysis in water samples at high sensitivity and selectivity with confirmation capabilities. J. Chromatogr. A **2004**, *1026,* 239–250.
63. Gonçalves, C.; Alpendurada, M.F. Comparison of three different poly(dimethylsiloxane)–divinylbenzene fibres for the analysis of pesticide multiresidues in water samples: Structure and efficiency. J. Chromatogr. A **2002**, *963,* 19–26.
64. Sng, M.T.; Lee, F.K.; Lakso, H.Å. Solid-phase microextraction of organophosphorus pesticides from water. J. Chromatogr. A **1997**, *759,* 225–230.
65. Jönsson, J.Å.; Mathiasson, L. Liquid membrane extraction in analytical sample preparation. Trends Anal. Chem. **1999**, *18,* 318–325.
66. Sanusi, A.; Millet, M.; Mirabela, P.; Wortham, H. Comparison of atmospheric pesticide concentrations measured at three sampling sites: Local, regional and long-range transport. Sci. Total Environ. **2000**, *263,* 263–277.
67. Scheyer, A.; Morville, S.; Mirabel, P.; Millet, M. A multiresidue method using ion-trap gas chromatography- tandem mass spectrometry with or without derivatisation with pentafluorobenzylbromide for the analysis of pesticides in the atmosphere. Anal. Bioanal. Chem. **2005**, *381,* 1226–1233.
68. Peck, A.M.; Hornbuckle, K.C. Gas-phase concentrations of current-use pesticides in Iowa. Environ. Sci. Technol. **2005**, *39*, 2952–2959.
69. Wilson, W.E.; Chow, J.C.; Claiborn, C.; Fusheng, W.; Engelbrecht, J.; Watson, J.G. Monitoring of particulate matter outdoors. Chemosphere **2002**, *49,* 1009–1043.
70. White, L.M.; Ernest, W.R.; Julen, G.; Garron, C.; Leger, M. Ambient air concentrations of pesticides used in potato cultivation in Prince Edward Island, Canada. Pest Manage. Sci. **2006**, *62*, 126–136.
71. Sanusi, A.; Millet, M.; Mirabel, P.; Wortham, H. Gas- particle partitioning of pesticides in atmospheric samples. Atmos. Environ. **1999**, *33*, 4941–4951.
72. Scheyer, A.; Morville, S.; Mirabel, P.; Millet, M. Variability of atmospheric pesticide concentrations between urban and rural areas during intensive pesticide application. Atmos. Environ. **2007**, *41,* 3604–3618.
73. Harrad, S.; Mao, H. Atmospheric PCBs and organochlorine pesticides in Birmingham, U.K.: Concentrations, sources, temporal and seasonal trends. Atmos. Environ. **2004**, *38,* 1437–1445.
74. Alegria, H.; Bidleman, T.F.; Figueroa, M.S. Organochlo- rine pesticides in the ambient air of Chiapas, Mexico. Environ. Pollut. **2006**, *140,* 483–491.
75. Scheyer, A.; Graeff, C.; Morville, S.; Mirabel, P.; Millet, M. Analysis of some organochlorine pesticides in an urban atmosphere (Strasbourg, east of France). Chemosphere **2005**, *58,* 1517–1524.
76. Coscolla, C.; Yusa, V.; Marti, P.; Pastor, A. Analysis of currently used pesticides in fine airborne particulate matter (PM 2,5) by pressurized liquid extraction and liquid chromatography–tandem mass spectrometry. J. Chromatogr. A **2008**, *1200,* 100–107.
77. Schummer, C.; Mothiron, E.; Appenzeller, B.M.R.; Rizet, A.; Wennig, R.; Millet, M. Temporal variations of concentrations of currently used pesticides in the atmosphere of Strasbourg, France. Environ. Pollut. **2010**, *158*, 576–584.
78. Alegria, H.A.; Bidleman, F.F.; Shaw, T.J. Organochlorine pesticides in the ambient air of Belize, Central America. Environ. Sci. Technol. **2000**, *34,* 1953–1958.
79. Cessna, A.J.; Waite, D.T.; Kerr, L.A.; Grover, R. Duplicate sampling reproducibility of atmospheric residues of herbicides for paired pan and high-volume air samplers. Chemo- sphere **2000**, *40*, 795–802.

80. Waite, D.T.; Cessna, A.J.; Grover, R.; Kerr, L.A.; Snihura, A.D. Environmental concentrations of agricultural herbicides: 2,4-D and triallate. J. Environ. Qual. **2002**, *31,* 129–144.
81. Waite, D.T.; Cessna, A.J.; Grover, R.; Kerr, L.A.; Snihura, A.D. Environmental concentrations of agricultural herbicides: Bromoxynil, dicamba, diclofop, MCPA and triflura- lin. J. Environ. Qual. **2004**, *33*, 1616–1628.
82. Yao, Y.; Tuduri, L.; Harner, T.; Blanchard, P.; Waite, D.; Poissantd, L.; Murphy, C.; Belzerf, W.; Aulagnierd, F.; Li, Y.; Sverko, E. Spatial and temporal distribution of pesticide air concentrations in Canadian agricultural regions. Atmos. Environ. **2006**, *40,* 4339–4351.
83. Waite, D.T.; Bailey, P.; Sproull, J.F.; Quiring, D.V.; Chau, D.F.; Bailey, J.; Cessna, A.J. Atmospheric concentrations and dry and wet deposits of some herbicides currently used on the Canadian Prairies. Chemosphere **2005**, *58,* 693–703.
84. Raina, R.; Sun, L. Trace level determination of selected organophosphorus pesticides and their degradation products in environmental air samples by liquid chromatography–positive ion electrospray tandem mass spectrometry. J. Environ. Sci. Health B **2008**, *43,* 323–332.
85. Bailey, R.; Belzer, W. Large volume cold on-column injection for gas chromatography–Negative chemical ionization–mass spectrometry analysis of selected pesticides in air samples. J. Agric. Food Chem. **2007**, *55*, 1150–1155.
86. Sofuoglu, A.; Odabasi, M.; Tasdemirc, Y.; Khalilid, N.R.; Holsen, T.M. Temperature dependence of gas-phase polycyclic aromatic hydrocarbon and organochlorine pesticide concentrations in Chicago air. Atmos. Environ. **2001**, *35*, 6503–6510.
87. Lammel, G.; Ghim, Y.; Grados, A.; Gao, H.; Huhnerfuss, H.; Lohmann, R. Levels of persistent organic pollutants in air in China and over the Yellow Sea. Atmos. Environ. **2007**, *41,* 452–464.
88. Gioia, R.; Offenberg, J.H.; Gigliottib, C.L.; Tottena, L.A.; Dua, S.; Eisenreich, S.J. Atmospheric concentrations and deposition of organochlorine pesticides in the US MidAtlantic region. Atmos. Environ. **2005**, *39,* 2309–2322.
89. Xu, D.; Dan, M.; Song, Y.; Chai, Z.; Zhuang, G. Concentration characteristics of extractable organohalogens in PM2.5 and PM10 in Beijing, China. Atmos. Environ. **2005**, *39,* 4119–4128.
90. Yang, Y.; Li, D.; Mu, D. Levels, seasonal variations and sources of organochlorine pesticides in ambient air of Guangzhou, China. Atmos. Environ. **2008**, *42,* 677–687.
91. Dvorska, A.; Lammel, G.; Klanova, J.; Holoubek, I. Koset- ice, Czech Republic—Ten years of air pollution monitoring and four years of evaluating the origin of persistent organic pollutants. Environ. Pollut. **2008**, *156,* 403–408.
92. Contamination de l'air par les produits phytosanitaires en region Centre, Rapport final, 2006, available at http://www. ligair.fr (accessed May 16, 2012).
93. Burhler, S.; Basu, I.; Hites, R.A. A Comparison of PAH, PCB, and pesticide concentrations in air at two rural sites on Lake Superior. Environ. Sci. Technol. **2001**, *35*, 2417–2422.
94. Burhler, S.; Basu, I.; Hites, R.A. Causes of variability in pesticide and PCB concentrations in air near the Great Lakes. Environ. Sci. Technol. **2004**, *38,* 414–422.
95. Baraud, L.; Tessier, D.; Aaron, J.J.; Quisefit, J.P.; Pinart, J. A multi-residue method for characterization and determination of atmospheric pesticides measured at two French urban and rural sampling sites. Anal. Bioanal. Chem. **2003**, 377, 1148–1152.
96. Motelay-Massei, A.; Harner, T.; Shoeib, M.; Diamond, M.; Stern, G.; Reosenberg, B. Using passive air samplers to assess urban–rural trends for persistent organic pollutants and polycyclic aromatic hydrocarbons. 2. Seasonal trends for PAHs, PCBs, and organochlorine pesticides. Environ. Sci. Technol. **2005**, *39*, 5763–5773.
97. Gouin, T.; Harner, T.; Blanchard, P.; Mackay, D. Passive and active air samplers as complementary methods for investigating persistent organic pollutants in the Great Lakes Basin. Environ. Sci. Technol. **2005**, *39,* 9115–9122.
98. Harner, T. Using passive air samplers to assess urban–rural trends for persistent organic pollutants. 1. Polychlorinated biphenyls and organochlorine pesticides. Environ. Sci. Technol. **2004**, *38,* 4474–4483.

99. Jaward, F.; Farrar, N.; Harner, T.; Sweetman, A.J.; Jones, A. Passive air sampling of PCBs, PBDEs, and organochlo- rine pesticides across Europe. Environ. Sci. Technol. **2004**, *38,* 34–41.
100. Farrar, J.; Harner, T.; Shoeib, M.; Sweetman, A.; Jones, C. Field deployment of thin film passive air samplers for persistent organic pollutants: A study in the urban atmospheric boundary layer. Environ. Sci. Technol. **2005**, *39,* 42–48.
101. Farrar, N.J.; Prevedouros, K.; Harner, T.; Sweetman, A.J.; Jones, K.C. Continental scale passive air sampling of persistent organic pollutants using rapidly equilibrating thin films (POGs). Environ. Pollut. **2006**, *144*, 423–433.
102. Pozo, K.; Harner, T.; Shoeib, M.; Urrutia, R.; Barra, R.; Parra, O.; Focardi, S. Passive-sampler derived air concentrations of persistent organic pollutants on a north–south transect in Chile. Environ. Sci. Technol. **2004**, *38,* 6529–6537.
103. Shen, L.; Wania, F.; Lei, Y.D.; Teixeira, C.; Bidleman, T.F. Atmospheric distribution and long-range transport behavior of organochlorine pesticides in North America. Environ. Sci. Technol. **2005**, *39*, 409–420.
104. Wania, F.; Shen, L.; Lei, Y.D.; Texeira, C.; Muir, D.C.G. Development and calibration of a resin-based passive sampling system for monitoring persistent organic pollutants in the atmosphere. Environ. Sci. Technol. **2003**, *37,* 1352–1359.
105. Klanova, J.; Kohoutek, J.; Hamplova, L.; Urbanova, P.; Holoubek, I. Passive air sampler as a tool for long-term air pollution monitoring: Part 1. Performance assessment for seasonal and spatial variations. Environ. Pollut. **2006**, *144,* 393–405.
106. Gouin, T.; Shoeib, M.; Harner, T. Atmospheric concentrations of current-use pesticides across south-central Ontario using monthly-resolved passive air samplers. Atmos. Environ. **2008**, *42,* 8096–8104.
107. Cupr, P.; Klanova, J.; Bartos, T.; Flegrova, Z.; Kohoutek, J.; Holoubek, I. Passive air sampler as a tool for long-term air pollution monitoring: Part 2. Air genotoxic potency screening assessment. Environ. Pollut. **2006**, *144,* 406–413.
108. Kot-Wasik, A.; Zabiegala, B.; Urbanowicz, M.; Dominiak, E.; Wasik, A.; Namieśnik, J. Advances in passive sampling in environmental studies. Anal. Chim. Acta **2007**, *602,* 141–163.
109. Wennrich, L.; Poppb, P.; Hafner, C. Novel integrative passive samplers for the long-term monitoring of semivolatile organic air pollutants. J. Environ. Monit. **2002**, *4*, 371–376.
110. Esteve-Turillas, F.A.; Pastor, A.; de la Guardia, M. Evaluation of working air quality by using semi-permeable membrane devices: Analysis of organophosphorus pesticides. Anal. Chim. Acta **2008**, 626, 21–27.
111. Namieśnik, J.; Łukasiak, J.; Jamrógiewicz, Z. *Pobiera- nie próbek środowiskowych do analizy;* PWN: Warszawa, 1995.
112. Wu, S.; Tao, S.; Zhang, Z.; Lan, T.; Zuo, Q. Distribution of particle-phase hydrocarbons, PAHs and OCPs in Tianjin, China. Atmos. Environ. **2005**, *39,* 7420–7432.
113. Sadiki, M.; Poissant, L. Atmospheric concentrations and gas-particle partitions of pesticides: Comparisons between measured and gas-particle partitioning models from source and receptor sites. Atmos. Environ. **2008**, *42,* 8288–8299.
114. Sauret, N.; Wortham, H.; Putaud, J.; Mirabel, P. Study of the effects of environmental parameters on the gas/particle partitioning of current-use pesticides in urban air. Atmos. Environ. **2008**, *42*, 544–553.
115. Scheyer, A.; Morville, S.; Mirabel, P.; Millet, M. Gas/par- ticle partitioning of lindane and current-used pesticides and their relationship with temperature in urban and rural air in Alsace region (east of France). Atmos. Environ. **2008**, *42,* 7695–7705.
116. Sauret, N.; Millet, M.; Herckes, P.; Mirabel, P.; Wortham, H. Analytical method using gas chromatography and ion trap tandem mass spectrometry for the determination of *S*- triazines and their metabolites in the atmosphere. Environ. Pollut. **2000**, *110,* 243–252.
117. Millet, M.; Wortham, H.; Sanusi, A.; Mirabel, P. A multiresidue method for determination of trace levels of pesticides in air and water. Arch. Environ. Contam. Toxicol. **1996**, *31,* 543–556.

118. Sofuoglu, A.; Cetin, E.; Bozacioglu, S.S.; Sener, G.D.; Odabasi, M. Short-term variation in ambient concentrations and gas/particle partitioning of organochlorine pesticides in Izmir, Turkey. Atmos. Environ. **2004**, *38,* 4483–4493.
119. Qiu, X.; Zhu, T.; Li, J.; Pan, H.; Li, Q.; Miao, G.; Gong, J. Organochlorine pesticides in the air around the Taihu Lake, China. Environ. Sci. Technol. **2004**, *38,* 1368–1374.
120. Schummer, C.; Mothiron, E.; Appenzeller, B.M.R.; Rizet, A.; Wennig, R.; Millet, M. Temporal variations of concentrations of currently used pesticides in the atmosphere of Strasbourg, France. Environ. Pollut. **2010**, *158,* 576–584.
121. Batterman, S.A.; Chernyak, S.M.; Gounden, Y.; Ma- tooanne, M.; Naidoo, R.N. Organochlorine pesticides in ambient air in Durban, South Africa. Sci. Total Environ. **2008**, 397, 119–130.
122. Seiber, J.N.; Glotfelty, D.E.; Lucas, A.D.; McChesney, M.M.; Sagebiel, J.C.; Wehner, T.A. A multiresidue method by high performance liquid chromatography-based fractionation and gas chromatographic determination of trace levels of pesticides in air and water. Arch. Environ. Con- tam. Toxicol. **1990**, *19*, 583.
123. Briand, O.; Bertrand, F.; Seux, R.; Millet, M. Comparison of different sampling techniques for the evaluation of pesticide spray drift in apple orchards. Sci. Total Environ. **2002**, *288,* 199–213.
124. Coscolla, C.; Yusa, V.; Beser, M.I.; Pastor, A. Multi-residue analysis of 30 currently used pesticides in fine airborne particulate matter (PM 2.5) by microwave-assisted extraction and liquid chromatography–tandem mass spectrometry. J. Chromatogr. A **2009**, *1216,* 8817–8827.
125. Clement, M.; Arzel, B.; Le Bot, B.; Seux, R.; Millet, M. Ad- sorption/thermal desorption-GC/MS for the analysis of pesticides in the atmosphere. Chemosphere **2000**, *40,* 49–56.
126. Ramos, J.J.; González, M.J.; Ramos, L. Comparison of gas chromatography-based approaches after fast miniaturised sample preparation for the monitoring of selected pesticide classes in fruits. J. Chromatogr. A. **2009**, *1216*, 7307–7313.
127. Cai, L.; Gong, S.; Chen, M.; Wu, C. Vinyl crown ether as a novel radical crosslinked sol–gel SPME fiber for determination of organophosphorus pesticides in food samples. Anal. Chim. Acta **2006**, *559*, 89–96.
128. Mol, H.G.J.; van Dam, R.C.J.; Streijger, O.M. Determination of polar organophosphorus pesticides in vegetables and fruits using liquid chromatography with tandem mass spectrometry: Selection of extraction solvent. J. Chromatogr. A. **2003**, *1015,* 119–127.
129. Kmellár, B.; Fodor, P.; Pareja, L.; Ferrer, C.; Martinnez- Uroz, M.A.; Valwerde, A.; Fernăndez-Alba, A.R. Validation and uncertainty study of a comprehensive list of 160 pesticide residues in multi-class vegetables by liquid chromatography–tandem mass spectrometry J. Chromatogr. A. **2008**, *1215,* 37–50.
130. Ferrer, I.; Garĭa-Reyes, J.F.; Mezcua, M.; Thurman, E.M.; Fernändez-Alba, A.R. Multi-residue pesticide analysis in fruits and vegetables by liquid chromatography–time-of- flight mass spectrometry. J. Chromatogr. A. **2005**, *1082,* 81–90.
131. Melo, L.F.C.; Collins, C.H.; Jardim, I.C.S.F. New materials for solid-phase extraction and multiclass high-performance liquid chromatographic analysis of pesticides in grapes. J. Chromatogr. A. **2004**, *1032,* 51–58.
132. Sanusi, A.; Guillet, V.; Montury, M. Advanced method using microwaves and solid-phase microextraction coupled with gas chromatography–mass spectrometry for the determination of pyrethroid residues in strawberries. J. Chromatogr. A. **2004**, *1046,* 35–40.
133. Blasco, C.; Font, G.; Pico, Y. Comparison of microextraction procedures to determine pesticides in oranges by liquid chromatography–mass spectrometry. J. Chromatogr. A. **2002**, *970,* 201–212.
134. Kende, A.; Csizmazia, Z.; Rikker, T.; Angyal, V.; Torkos, K. Combination of stir bar sorptive extraction–retention time locked gas chromatography–mass spectrometry and automated mass spectral deconvolution for pesticide identification in fruits and vegetables. Microchem. J. **2006**, *84,* 63–69.

135. Sanchez, A.G.; Martos, N.R.; Ballesteros, E. Multiresidue analysis of pesticides in olive oil by gel permeation chromatography followed by gas chromatography–tandem mass- spectrometric determination. Anal. Chim. Acta **2006**, *558*, 53–61.
136. Gelsomino, A.; Petroviowi, B.; Tiburtini, S.; Magnani, E.; Felici, M. Multiresidue analysis of pesticides in fruits and vegetables by gel permeation chromatography followed by gas chromatography with electron-capture and mass spectrometric detection. J. Chromatogr. A. **1997**, *782,* 105–122.
137. Vreuls, J.J.; Swen, R.J.J.; Goudriaan, V.P.; Kerkhoff, M.A.T.; Jongenotter, G.A.; Brinkman, U.A.Th. Automated on-line gel permeation chromatography-gas chromatography for the determination of organophosphorus pesticides in olive oil. J. Chromatogr. A. **1996**, *750,* 275–286.
138. Stajnbaher, D.; Zupancic-Kralj, L. Multiresidue method for determination of 90 pesticides in fresh fruits and vegetables using solid-phase extraction and gas chromatography-mass spectrometry. J. Chromatogr. A. **2003**, *1015,* 185–198.
139. Reinhard, H.; Sager, F.; Zoller, O. Citrus juice classification by SPME-GC–MS and electronic nose measurements. LWT—Food Sci. Technol. **2008**, *41,* 1906–1912.
140. Knezewvic, Z.; Serdar, M. Screening of fresh fruit and vegetables for pesticide residues on Croatian market. Food Control **2009**, *20,* 419–422.
141. Wu, J.; Luan, T.; Lan, Ch.; Wai Hung Lo, T.; Yuk Sing Chan, G. Removal of residual pesticides on vegetable using ozonated water. Food Control **2007**, *18,* 466–472.
142. Lambropoulou, D.A.; Albanis, T. A headspace solid-phase microextraction in combination with gas chromatography–mass spectrometry for the rapid screening of organophos- phorus insecticide residues in strawberries and cherries. J. Chromatogr. A. **2003**, *993,* 197–203.
143. Juan-Garĭa, A.; Picó, Y.; Font, G. Capillary electrophoresis for analyzing pesticides in fruits and vegetables using solid-phase extraction and stir-bar sorptive extraction. J. Chromatogr. A. **2005**, *1073,* 229–236.
144. Liu, W.; Hu, Y.; Zhao, J.; Xu, Y.; Guan, Y. Determination of organophosphorus pesticides in cucumber and potato by stir bar sorptive extraction. J. Chromatogr. A. **2005**, *1095*, 1–7.
145. Soler, C.; Mañes, J.; Picó, Y. Routine application using single quadrupole liquid chromatography–mass spectrometry to pesticides analysis in citrus fruits. J. Chromatogr. A. **2005**, *1088,* 224–233.
146. Riu-Aumatell, M.; Castellari, M.; Lopez-Tamames, E.; Galassi, S.; Buxaderas, S. Characterisation of volatile compounds of fruit juices and nectars by HS/SPME and GC/ MS. Food Chem. **2004**, *87,* 627–637.
147. Soler, C.; Mañes, J.; Picó, Y. Routine application using single quadrupole liquid chromatography–mass spectrometry to pesticides analysis in citrus fruits. J. Chromatogr. A. **2005**, *1088*, 224–233.
148. Mezcua, M.; Ferrer, C.; García-Reyes, J.F.; Martínez- Bueno, M.J.; Sigrist, M.; Fernández-Alba, A.R. Analyses of selected non-authorized insecticides in peppers by gas chromatography/mass spectrometry and gas chromatog- raphy/tandem mass spectrometry. Food Chem. **2009**, *112*, 221–225.
149. Lehotay, S.T.; Ae Son, K.; Kwon, H.; Koesukwiwat, U.; Fu, W.; Mastovska, K.; Hoh, E.; Leepipatpiboo, N. Comparison of QuEChERS sample preparation methods for the analysis of pesticide residues in fruits and vegetables. J. Chromatogr. A. **2010**, *1217,* 2548–2560.
150. Thanh Dong, N.; Ji Eun, Y.; Dae Myung, L.; Gae-Ho, L. A multiresidue method for the determination of 107 pesticides in cabbage and radish using QuEChERS sample preparation method and gas chromatography mass spectrometry. Food Chem. **2008**, *110,* 207–213.
151. Lesueur, C.; Knittl, P.; Gartner, M.; Mentler, A.; Fuer- hacker, M. Analysis of 140 pesticides from conventional farming foodstuff samples after extraction with the modified QuECheRS method. Food Control **2008**, *19,* 906–914.
152. Mayer-Helm, B. Method development for the determination of 52 pesticides in tobacco by liquid chromatography–tandem mass spectrometry. J. Chromatogr. A. **2009**, *1216,* 8953–8959.

153. Lacina, O.; Urbanova, J.; Poustka, J.; Hajslova, J. Identifi- cation/quantification of multiple pesticide residues in food plants by ultra-high-performance liquid chromatography- time-of-flight mass spectrometry. J. Chromatogr. A. **2010**, *1217*, 648–659.
154. Gilbert-López, B.; García-Reyes, J.F.; Molina-Díaz, A. Sample treatment and determination of pesticide residues in fatty vegetable matrices: A review. Talanta **2009**, *79,* 109–128.
155. Cunha, S.C.; Fernandes, J.O.; Oliveira, M.B.P.P. Fast analysis of multiple pesticide residues in apple juice using dispersive liquid–liquid microextraction and multidimensional gas chromatography–mass spectrometry. J. Chromatogr. A. **2009**, *1216,* 8835–8844.
156. Economou, A.; Botitsi, H.; Antoniou, S.; Tsipi, D. Determination of multi-class pesticides in wines by solid-phase extraction and liquid chromatography–tandem mass spectrometry. J. Chromatogr. A. **2009**, *1216,* 5856–5867.
157. Moral, A.; Sicilia, M.D.; Rubio, S. Anal. Chim. Acta **2009**, *650,* 207–213.
158. Gevao, B.; Semple, K.T.; Jones, K.C. Bound pesticide residues in soils: A review. Environ. Pollut. **2000**, *108,* 3–14.
159. Arias-Estevez, M.; Lopez-Periago, E.; Martinez-Carballo, E.; Simal-Gandara, J.; Mejuto, J.C.; Garcia-Rio, C. The mobility and degradation of pesticides in soils and the pollution of groundwater resources. Agric. Ecosyst. Environ. **2008**, *123*, 247–260.
160. Köhne, J.M.; Köhne, S.; Š imönek, J. A review of model applications for structured soils: b) Pesticide transport. J. Contam. Hydrol. **2009**, *104*, 36–60.
161. Muller, K.; Magesan, G.N.; Bolan, N.S. A critical review of the influence of effluent irrigation on the fate of pesticides in soil. Agric. Ecosyst. Environ. **2007**, *120,* 93–116.
162. Hussen, A.; Westbom, R.; Megersa, N.; Mathiasson, L.; Bjorklund, E. Selective pressurized liquid extraction for multi-residue analysis of organochlorine pesticides in soil. J. Chromatogr. A **2007**, *1152,* 247–253.
163. Harner, T.; Wideman, J.L.; Jantunen, L.M.M.; Bidleman, T.F.; Parkhurst, W.J. Residues of organochlorine pesticides in Alabama soils. Environ. Pollut. **1999**, *106,* 323–332.
164. Oldal, B.; Maloschik, E.; Uzinger, N.; Anton, A.; Székács, A. Pesticide residues in Hungarian soils. Geoderma **2006**, *135,* 163–178.
165. Goncalves, C.; Carvalho, J.J.; Azenha, M.A.; Alpendurada, M.F. Optimization of supercritical fluid extraction of pesticide residues in soil by means of central composite design and analysis by gas chromatography–tandem mass spectrometry. J. Chromatogr. A **2006**, *1110,* 6–14.
166. Koinecke, A.; Kreuzig, R.; Bahadir, M. Effects of modifiers, adsorbents and eluents in supercritical fluid extraction of selected pesticides in soil. J. Chromatogr. A **1997**, *786*, 155–161.
167. Wang, X.; Zhao, X.; Liu, X.; Li, Y.; Fu, L.; Hu, J.; Huang, Ch. Homogeneous liquid–liquid extraction combined with gas chromatography–electron capture detector for the determination of three pesticide residues in soils. Anal. Chim. Acta **2008**, *620,* 162–169.
168. Wanner, U.; Burauel, P.; Fuehr, F. Characterisation of soil- bound residue fractions of the fungicide dithianon by gel permeation chromatography and polyacrylamide gel electrophoresis. Environ. Pollut. **2000**, *108,* 53–59.
169. Bi, E.; Schmidt, T.C.; Haderlein, S.B. Practical issues relating to soil column chromatography for sorption parameter determination. Chemosphere **2010**, *80,* 787–793.
170. Dugo, G.; Di Bella, G.; La Torre, M.; Saitta, M. Rapid GC-FPD determination of organophosphorus pesticide residues in Sicilian and Apulian olive oil. Food Control **2005**, *16,* 435–438.
171. Sobel, T.; Gul, O.; Buket, A. Simultaneous determination of various pesticides in fruit juices by HPLC-DAD. Food Control **2005**, *16,* 87–72.
172. Cunha, S.C.; Fernandes, J.O.; Alves, A.; Oliveira, M.B.P.P. Fast low-pressure gas chromatography–mass spectrometry method for the determination of multiple pesticides in grapes, musts and wines. J. Chromatogr. A. **2009**, *1216,* 119–126.
173. Available at http://www.pg.gda.pl/chem/Katedry/Anality- czna/index.php?option=com_content&task=view&id=130&Itemid=86 (accessed February 2010).

174. Sadowska-Rociek, A.; Cieślik; E. Stosowane techniki i najnowsze trendy w oznaczaniu pozostałości pestycydów w żywności metodą chromatografii gazowej. Metrologia **2008**, *13,* 33–38.
175. Walorczyk, S. Różne możliwości wykorzystania chromatografii gazowej połączonej ze spektrometrią. mas w analizie pozostałości środków ochrony roślin. Prog. Plant Protect. **2007**, *47,* 111–114.
176. Alder, L.; Greulich, K.; Kempe, G.; Vieth, B. Residue analysis of 500 high priority pesticides: Better by GC–MS or LC–MS/MS? Mass Spectrom. Rev. **2006**, *25*, 838–865.
177. Namieśnik, J. Modern trends in monitoring and analysis of environmental pollutants. Pol. J. Environ. Stud. **2001**, *10,* 127–140.
178. Beyer, A.; Biziuk, M. Methods for determining pesticides and polychlorinated biphenyls in food samples—Problems and challenges. Crit. Rev. Food Sci. **2008**, *48,* 888–904.

8

Salt-Affected Soils: Sustainable Agriculture

Pichu Rengasamy

Introduction

Worldwide, more than 800 million hectares of land are estimated to be salt-affected.[1] This is based on the traditional outlook on saline soils as being dominated by sodium salts. However, this area can be larger if lands affected by the input of several other types of salts, including contaminants, are taken into account. The demand for agricultural production from salt-affected soils will increase in the future as the expected world population increases to 9 billion by the year 2050. Global agriculture is now in a state of crisis, both from the economic and ecological considerations. Salinity in agricultural soils enhances the possibility of economic downturn in the farming sector and environmental degradation.

Agricultural soils accumulate soluble salts in the root zone by various processes, including weathering of minerals, salt added through rain and agronomic practices such as fertilizer and pesticide application, saline groundwater intrusion, irrigation with saline water sourced from bore, recycled waters or wastewaters, dumping of industrial and municipal wastes into soils, and other soil processes leading to reduced leaching of salts from the root zone. Depending on the concentration and nature of salts, the growth and productivity of different crops are variously affected. Soluble ions affect the soil structure and porosity, restricting air and water movement within soil layers, and also the water-storing capacity. Salt movement through leaching will alter the quality of groundwater, and that by drainage will influence the environment in the disposal region. Thus, environmentally sustainable agriculture depends on managing many soil processes to maintain appropriate soil conditions for plant growth and avoiding contamination of the surrounding natural resources. Attention by the global community to soil health care and enhancement to achieve water conservation and efficient use in conjunction with the adoption of climate-resilient technologies has become a necessity.

Soluble Salts and Salinity Impact

Even though several types of organic and inorganic compounds are present in the soil, the concentration and type of ions in soil water derived from soluble compounds cause “salt stress” on plants, affecting crop productivity in addition to influencing soil properties (Figure 1). While the salt input into soils can vary considerably according to various resources, the contribution by irrigation water can be estimated by the salinity of irrigation water and the application rate (Table 1).

The predominant constituent cations of total soluble salts commonly found in salt-affected soils are sodium, potassium, calcium, and magnesium; the anions are chloride, sulfate, bicarbonate, and carbonate. However, the presence of ionic species of plant nutrient elements like N and P, and other microelements such as Fe, Al, B, Cd, Zn, Cu, Mn, Co, Cr, Mo, As, Pb, Hg, Se, and Sr, although in small amounts (hence called microelements), can either interfere with the plant nutrition or induce toxicity in plants. The solubility of all these compounds and the nature of ionic species depend on the pH of the soil water, and also pE (–log [electron]) when a soil is waterlogged. Crop growth responds to ionic concentrations in soil water in two phases: a continuous osmotic phase that inhibits the water uptake by plants due to osmotic pressure of soil water lowering its potential energy (water always moving from a higher to lower potential energy levels); and a slower-acting ionic phase when the accumulation of a given ion in the plant over a period leads to ion toxicity or ion imbalance.[2] However, the interactions between root zone environments and plant responses to increased osmotic pressure (due to total dissolved salts) or specific ion concentrations in the field soil are complicated by many soil processes such as soil water dynamics, soil structural stability, nutrient and water movement, and changes in soil air composition.[3]

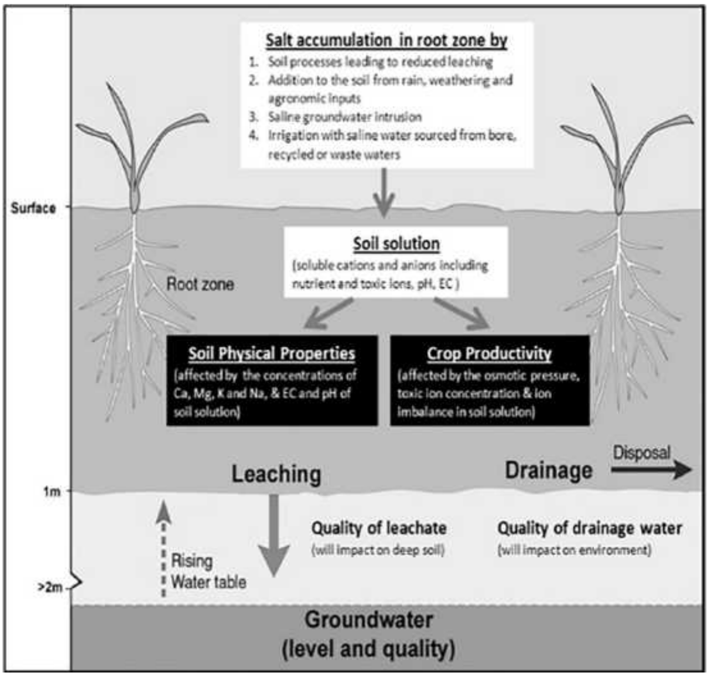


FIGURE 1 Schematic diagram of salt accumulation, soil solution affecting soil structure and crop productivity, leaching, and drainage—all influencing sustainable management of salt-affected soils.

TABLE 1 Example of Amount of Salt Applied per Hectare Based on Irrigation Water Salinity (dSm–1) and Application Rate (MLha–1)

	Irrigation Application Rate			
Irrigation Water Salinity EC (dSm–1)	2 MLha–1	4 MLha–1	8 MLha–1	16 MLha–1
	Salt input into soil (tha^{-1})			
0.5	0.64	1.28	2.56	5.12
1.0	1.28	2.56	5.12	10.24
3.0	3.84	7.68	15.36	30.72

Soil Structural Stability in Salt-Affected Soils

The integrity of soil aggregates when in contact with water, either during rainfall or irrigation, is an important feature of soil structural stability. Unstable soil structure leads to collapse of aggregates into fine particles. This results in agricultural soils becoming less porous, affecting water and air movement within root zone and with reduced water storage capacity. During high rainfall events, these soils can be waterlogged and soil particles can be easily eroded by overflowing water.

Soil microaggregates (<≈250 μm) are formed by the linking of clay, silt, and sand particles by various inorganic ions and organic compounds. The stability of microaggregates depend on the nature of interaction between water molecules and soil particles, particularly clay. Clay-to- clay bonding is mediated by cations such as sodium, potassium, magnesium, and calcium, which are adsorbed on soil clays from soil solution and known as exchangeable cations. Clay particles are negatively charged and cations are positively charged. Hence, they attract each other and form different types of bonding linking the clay particles. However, the nature of this bonding can be ionic or covalent or a combination of both. Water molecules, being polar, can interact with ionic bonds while they do not interact with covalent bonds.

Organic compounds are all made up of covalent bonds and hence they do not react with water. When organic matter is high in soils, water interaction is minimal because clays are either bound by clay-organic bonds, which are mostly covalent, or soil aggregates are enveloped by organic materials formed by covalent bonding. Hence, the soil aggregate structure is highly stable. Further, microaggregates are linked to form larger macroaggregates influenced by organic materials, including living species such as plant roots and fungal hyphae.[4] The phenomenon of "slaking" occurs when macroaggregates on wetting break into microaggregates, depending on the nature of the linkage.

Clay–sodium bonding is highly ionic, and the clays bound by sodium are easily separated by water. Clay particles with high exchangeable sodium easily separate when in contact with water. This phenomenon is known as clay swelling when water content is low and clay dispersion when water content is high, both leading to soil structural deterioration.

Soil structural deterioration increases with an increase in exchangeable sodium percent (ESP), which is defined as

$$\mathrm{ESP} = (\text{exchangeable Na} / \text{cation exchange capacity}) \times 100$$

In Australia, soil with an ESP of 6 is considered "sodic," and when ESP is >15, it is considered as "highly sodic." In the United States and other parts of the world, an ESP >15 is the criterion for a sodic soil. The value of ESP depends on the ratio of sodium to divalent ions in soil solutions or soil extracts. This ratio is known as sodium adsorption ratio (SAR), which is defined as follows:

$$\mathrm{SAR} = \mathrm{Na} / (\mathrm{Ca} + \mathrm{Mg})^{0.5},$$

where the concentrations of cations are expressed as mmolL^{-1}.

Generally, the SAR of soil saturation extract is approximately equal to the ESP of the soil. As the measurement of SAR is simpler than that of ESP, SAR is conveniently used as index of soil sodicity and the degree of structural deterioration. Clay dispersion from soils, measured in the laboratory, is highly related to SAR in soil solutions.[5]

Generally, sodium is considered to cause clay dispersion in soils. However, potassium, being a monovalent cation, can also induce similar effects. Our studies have shown that clay–cation bonding has different degrees of ionicity, depending on the cation. The degree of ionicity is in the following order: Na>K ≫ Mg>Ca. While both Na and K promote clay dispersion, they are not equal in their effects. Similarly, both Ca and Mg promote clay flocculation (aggregation of clay particles), but they have different flocculating powers.[6] Salt-affected soils are found to have different combinations of these cations, and to consider only sodium as dispersive cation can lead to erroneous conclusions. Therefore, we[7] have derived a new formula, "cation ratio of soil structural stability (CROSS)," analogous to SAR but which selectively incorporates the individual dispersive effects of Na and K, and the individual flocculating effects of Mg and Ca. This ratio was derived from the flocculating power of these cations detailed in Rengasamy and Olsson.[6]

$$\mathrm{CROSS} = (\mathrm{Na} + 0.56\ \mathrm{K}) / (\mathrm{Ca} + 0.6\ \mathrm{Mg})^{0.5}$$

The effects of cations on clay dispersion from salt- affected soils (indicative of soil structural degradation), characterized by either SAR or CROSS, are evident only when the ionic strength of the soil solution is below a "threshold level."[5]Above this level, the soil clays are flocculated by dissolved salts without any impact on soil physical behavior. Thus, Figure 2 schematically describes the relation between the indices (SAR or CROSS) of soil structural degradation and the ionic strength of the soil solution (measured as electrical conductivity, EC, of the soil solution) in salt-affected soils. As these indices increase clay dispersion is favored, while as EC increases flocculation (aggregation of clay particles) is promoted. However, above a certain level of EC, plant growth is affected by salinity stress. The relationship featured in Figure 2 is highly dependent on soil pH, which affects the net charge on clay particles.[8]

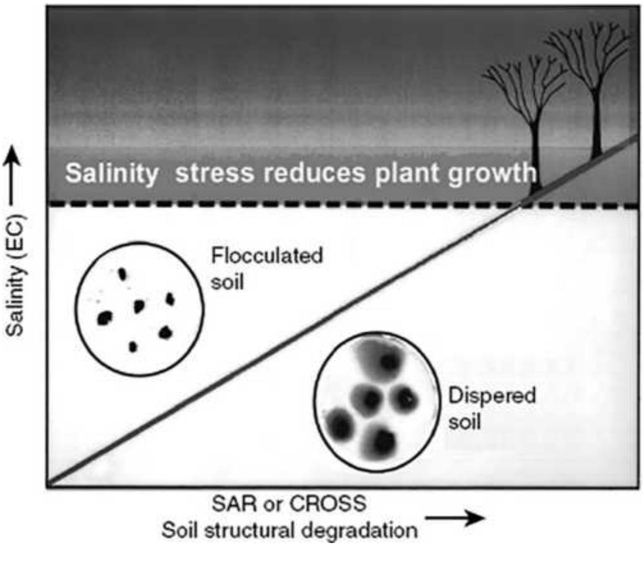


FIGURE 2 Schematic diagram relating SAR or CROSS with EC of soil solutions to indicate soil structural degradation in salt- affected soils.

Categories of Salt-Affected Soils

As sodium salts dominate in naturally occurring salt- affected soils, the soils are generally "saline–sodic," which become "sodic" when soluble salts are leached. Similarly, a sodic soil becomes saline–sodic when salts accumulate in soil layers because of restricted drainage imposed by a degraded soil structure. When sodium salts are not dominant, the soils are referred to as "saline" soils. Generally, saline, sodic, and saline–sodic soils have a spectrum of disorders and the soil solutions have a range of values of cations concentrations and electrical conductivity. Sodic soils are considered[9] to have pH >8.5 because of the predominance of carbonate ions in arid soils. However, sodic soils with neutral and acidic pH with the predominance of anions Cland SO_4^{2-} have been reported to be widely present, particularly in Australia.[10] Soil pH also controls the presence of a variety of ionic species, in addition to influencing soil structure. For example, chlorides and sulfates dominate in neutral and acidic soils, whereas in alkaline soils bicarbonate and carbonate ions are prevalent. Microelements such as Fe, Al, Zn, Mn, Mo, Cr, B, Cu, As, Cd, and Se occur in soil solutions as differently charged ionic species ranging from cationic to anionic forms dictated by equilibrium pH.[11]

Thus, the SAR, EC, and pH of soil solutions are determined by the ionic composition and hence, the impact on crop growth and yield, and as well as on soil structure. Table 2 gives different categories of salt-affected soils on the basis of SAR_e and EC_e of soil saturation paste extracts (subscript e denoting saturation paste extract) and pH measured in 1:5 soil-water suspensions. Measurements of ions in extracts from saturated soil paste are common in many parts of the world. If measurements are made in soil solutions from different soil-to-water ratios, the values will be different and appropriate conversion factors have to be used to derive SAR_e and EC_e.[3] Table 2 also gives the possible mechanisms through which plant growth is affected in each category. Measurements of all ions in soil solution are necessary to identify toxicity, deficiency, or ion imbalance of various nutrients and microelements in relation to crop productivity.

Toxic Elements in Salt-Affected Soils

The accumulation of toxic elements including heavy metal ions in the form of soluble salts in agricultural soils, in recent years, is caused by the long-term application of wastewater, industrial effluents, mining wastes, sewage sludge (or biosolids), and compost. As a result of global industry outputs, atmospheric deposition (such as acid rains) is another cause of toxic ion dynamics in productive soils. Landfill leachates also pollute nearby farming lands. Although the amounts of these elements are usually much smaller than the major elements (considered to be useful in crop production), their availability for plant uptake or being leached into groundwater is an important environmental issue. Plant uptake of toxic ions, in addition to causing decline in crop production, can affect human health by entering into the food chain. More details on the distribution and properties of heavy metals (or trace elements) in soils can be found in many publications.

Soil Water Dynamics and Salinity Stress

Salt concentration in agricultural soils is transient and varies with depth and changes throughout the growing season in response to rainfall and irrigation, surface evaporation, water use by vegetation, and the hydraulic conductivity (hence the leaching fraction) of soil layers.[12] As the soil dries because of evapotranspiration, the salt concentration increases, as does the osmotic pressure of soil water. Concomitant changes in matric and osmotic potentials (Table 3) determine plant water uptake in the field. The influence of soil texture and type of clay on plant-available water compounds the effects of matric plus osmotic potentials. As the total water potential decreases below –900 kPa, plant water uptake is greatly reduced, and at –1500 kPa plant wilts completely.[3] The two plant responses to salinity, viz. osmotic stress and ion-specific stress, can occur sequentially, giving rise to a two-phase growth response.[2] Our experiments[13]

TABLE 2 Categories of Salt-Affected Soils Based on EC_e (dSm^{-1}), SAR_e, and $pH_{1:5}$ of Soil Solutions, and Possible Mechanisms of Impact on Plants

No.	Category of Saline Soil	Criteria	Possible Mechanisms of Impact on Plants
1	Acidic–saline soil	ECe >4; SARe <6; pH <6	Osmotic effect; microelement (Fe, Al, Mn, etc.), toxicity; SO42– toxicity in very low pH
2	Neutral saline soil	ECe >4; SARe <6; pH 6–8	Osmotic effect; toxicity of dominant anion or cation other than Na+
3	Alkaline–saline soil	ECe >4; SARe <6; pH 8–9	Osmotic effect; HCO3– and CO32– toxicity
4	Highly alkaline-saline soil	ECe >4; SARe <6; pH >9	Osmotic effect; HCO3– and CO32– toxicity; microelement (Fe, Al, Mn, etc.) toxicity
5	Acidic–saline–sodic soil	ECe >4; SARe >6; pH <6	Osmotic effect; Na+ and microelement (Fe, Al, Mn, etc.), toxicity
6	Neutral saline–sodic soil	ECe >4; SARe >6; pH 6–8	Osmotic effect; Na+ toxicity; toxicity of dominant anion (Cl– or SO42–)
7	Alkaline–saline–sodic soil	ECe >4; SARe >6; pH 8–9	Osmotic effect; Na+ toxicity; HCO3– and CO32– toxicity
8	Highly alkaline–saline–sodic soil	ECe >4; SARe >6; pH >9	Osmotic effect; Na+ toxicity; HCO3– and CO32– toxicity; microelement (Fe, Al, Mn, etc.), toxicity
9	Acidic–sodic soil	ECe <4; SARe >6; pH <6	Indirect effect due to soil structural problems; seasonal waterlogging can induce microelement (Fe, Al, Mn, etc.), toxicity
10	Neutral sodic soil	ECe <4; SARe >6; pH 6–8	Indirect effect due to soil structural problems; seasonal waterlogging; Na+ toxicity at high SARe
11	Alkaline–sodic soil	ECe <4; SARe >6; pH 8–9	Indirect effect due to soil structural problems; seasonal waterlogging; Na+ toxicity at high SARe; HCO3– and CO32– toxicity
12	Highly alkaline–sodic soil	ECe <4; SARe >6; pH >9	Indirect effect due to soil structural problems; seasonal waterlogging; Na+ toxicity at high SARe; HCO3– and CO32– toxicity; microelement (Fe, Al, Mn, etc.), toxicity
13	Non-salt-affected soil	ECe <4; SARe <6; pH 6–8	Can have problems due to factors other than salts

Note: Toxicity, deficiency, or ion imbalance due to various ions will depend on the ionic composition of soil solution.
Source: Data from Rengasamy.[3]

using soils in pot experiments and also field observations have shown that when the osmotic pressure of soil water is >900 kPa (EC of soil water >25 dSm^{-1}), the osmotic effect is predominant and specific ion (such as Na or Cl) effect may be significant at lower levels of salinity (Figure 3). In moderately saline soils when soil fertility level is low and nutrient deficiency is an issue, application of fertilizer at an appropriate level alleviates the salinity stress on plants.[14] Similarly, adding a small amount of calcium has been reported to enhance the salt tolerance of plants at moderate levels of NaCl salinity.[15]

Different levels of salinity in the field during the crop growing season induced by the changing soil water content (Table 3) will affect crops differently, either by a predominant osmotic effect or an ion-specific effect. In dryland cropping, a high concentration of salt combined with a low rainfall in the start of the season severely affects the germination of seeds. Similarly, dry spells, enhancing salt concentration, during critical physiological periods of crop growth such as flowering and grain filling, can result in the significant reduction in yield.[3] Apart from vertical and temporal variations, spatial variation in salinity across a field can also be large in both dryland and irrigated cropping regions. These factors have to be considered in choosing appropriate crop species relevant to different sections in the field and seasonal weather pattern in farming systems.

TABLE 3 Changes in Matric, Osmotic, and Total Water Potentials of an Alfisol (Loam) and a Vertisol (Clay) with Changes in Soil Water Content

	Water Content	EC	Soil Water Potential (kPa)		
Soil Water Status	(g g–1)	(dS m–1)	Matric	Osmotic	Total
Alfisol (loamy soil)					
1:5 Soil water	5.00	0.64	+0.1	–23.0	–22.9
Saturation paste	0.35	9.14	–0.1	–329.1	–329.2
Field capacity (FC)	0.20	16.00	–10.0	–576.0	–586.0
Drier than FC	0.14	22.86	–680.0	–823.0	–1503.0
Permanent wilting point	0.05	64.00	–1500.0	–2304.0	–3804.0
Vertisol (clayey soil)					
1:5 Soil–water	5.00	0.64	+0.1	–23.0	–22.9
Saturation paste	0.80	4.00	–0.1	–150.0	–150.1
Field capacity (FC)	0.40	8.00	–10.0	–288.0	–298.0
Drier than FC	0.16	20.00	–780.0	–720.0	–1500.0
Permanent wilting point	0.08	40.00	–1500.0	–1440.0	–2940.0

Source: Data from Rengasamy et al.[12]

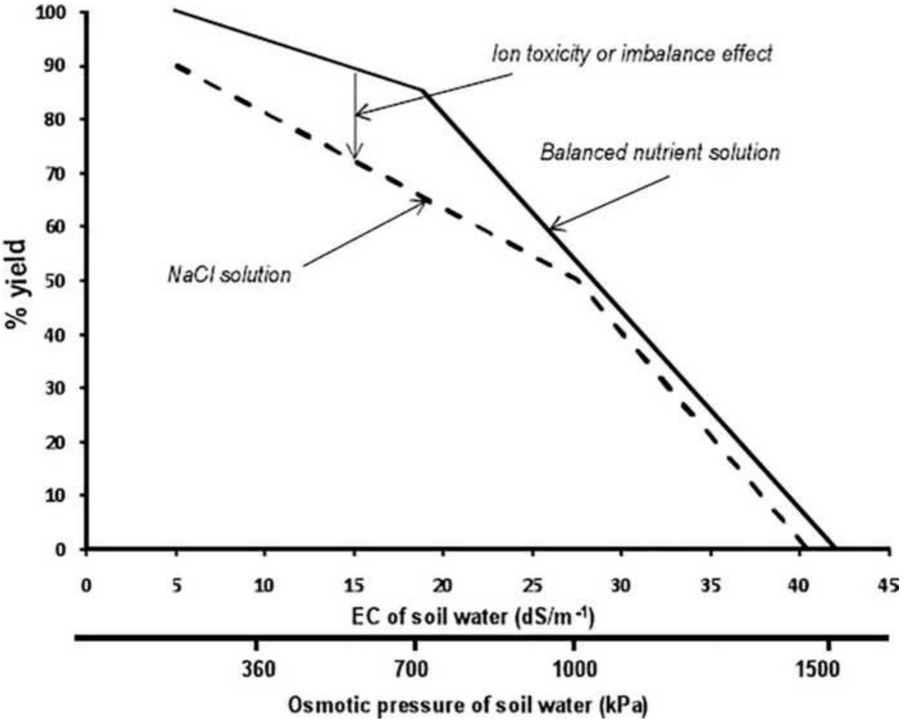


FIGURE 3 Schematic diagram showing the effect of osmotic pressure of soil solution on yield of crops and ion-specific effect at low levels of salinity.
Source: Rengasamy, P.[13]

Hypoxia and Salinity Effects

During high rainfall seasons, soils with restricted drainage become waterlogged and air-filled soil pores are eliminated with severely reduced oxygen availability (hypoxia). Sodic soils with limited porosity become saturated with water in wet conditions, and hypoxia is common in these soils. Hypoxia may lead to many problems that plants may experience such as shortage of ATP in anoxic root cells, reduced carbohydrate production, production of reactive oxygen species, toxic reduction products and raised CO_2 and organic acids, and shoot wilting from decreased hydraulic conductivity.[16] Hypoxia can worsen the growth of plants exposed to salinity.[17] Because of increased solubilities due to changes in pH and pE associated with hypoxia,[11] toxicity of Mn, Fe, Al, B, and reduction products such as H_2S and methane is common in waterlogged salt-affected soils.

Management of Salt-Affected Soils in Relation to Agricultural Production and Environmental Protection

The solution to improving agricultural production in salt-affected soils without degrading the environments is to link the plant improvement to tolerate salinity and other associated abiotic stress with appropriate soil management and cropping practices. As mentioned earlier, there are different categories of salt-affected soils, each one requiring a specific strategy depending on how root zone processes and plant interactions are affected by soil solution composition. Broadly, there are three ways with some possible actions to resolve the problems in salt-affected soils, based on the schematic presentation in Figure 1: 1) prevention of salt accumulation; 2) removal of accumulated salts; and 3) adaptation to saline environment.

Prevention of Salt Accumulation

Natural processes responsible for salt input, such as mineral weathering and deposits from rain and wind, are difficult to prevent or control. However, salt input through irrigation can be reduced by managing irrigation water quantity and quality. Adding salts and toxic elements by adding composts, industrial effluents, mining wastes, sewage sludge, biosolids, and wastewater can be prevented by environmental regulations and guidelines for safe application of these materials. Salt input through groundwater intrusion in soils with shallow water tables can be minimized by growing deep-rooted perennial plants with high transpiration rates to use groundwater and reduce recharge, thus controlling groundwater levels. However, this strategy will concentrate more salt in the root zone in soils where the water table is deep.[12]

Removal of Accumulated Salts

Accumulated salts in the root zone layers can be removed by leaching. In dryland cropping regions, good infiltration of rainfall and water movement through the soil profile by improving soil structure can effectively reduce salt load. In irrigation agriculture, the quantity and quality of irrigation water should be evaluated for providing leaching fraction to maintain salt balance. In both cases, care should be taken to avoid leachates from the soil contaminating the deeper soils or the surroundings in the drainage disposal region (Figure 1). Contamination due to As and Se from the drainage water has been reported in different parts of the world.

Adaptation to Saline Environment

Agricultural production in the saline environments can be enhanced by using plants that tolerate salt in the root zone. High genetic variations exist among plant species and crop cultivars that are thriving under different saline environments (or categories of salt-affected soils, Table 2). This can be exploited to develop better-adapted salt-land plants. Selection, breeding, and genetic modification to improve crop performance should focus on the many soil processes and soil solution compositions discussed above to find the new cultivars that are effective in the field so that farmers can adopt them.[18] Plant improvement must also be linked to appropriate soil management and agronomic practices, taking into account of temporal, vertical, and spatial variations in salinity in the field dictated by soil water dynamics.

Conclusions

The constituent cations and anions in solutions of salt- affected soils affect both soil properties and crop productivity. The interactions between root zone environments and plant response to increased osmotic pressure or specific ion concentrations are also influenced by soil processes such as soil water dynamics, soil structural stability, and soil pH and pE. Soil structural stability affects plant performance and also the environment by way of soil erosion. There are 12 categories of salt-affected soils based on the SAR_e, EC_e, and pH of soil extracts. The mechanisms by which plant growth is affected are specific to each category. Elements toxic to plant growth can occur naturally in these soils and also owing to use of wastewater and waste materials such as biosolids, sewage sludge, and compost. The solubilities of these elements are influenced by several factors, but mainly by soil pH and pE. Management of salt- affected soils in relation to agricultural production and environmental protection can be achieved by three ways, viz. prevention of salt accumulation, removal of accumulated salts, and adapting to saline environment by growing appropriate salt tolerant crops. While minimizing salt load in the root zone, care should be taken to avoid leachates from the soil contaminating the deeper soils or the surroundings in the drainage disposal region.

References

1. FAO. FAO Land and Plant Nutrition Management Service, 2008; Available at http://www.fao.org/ag/agl/agll/spush (accessed May 24, 2010).
2. Munns, R.; Tester, M. Mechanisms of salinity tolerance. Annu. Rev. Plant Biol. **2008**, *59*, 651–681.
3. Rengasamy, P. Soil processes affecting crop production in salt-affected soils. Funct. Plant Biol. **2010**, *37*, 613–620.
4. Tisdall, J.M.; Oades, J.M. Organic matter and water-stable aggregates in soils. J. Soil Sci. **1982**, *33*, 141–163.
5. Rengasamy, P.; Olsson, K.A. Sodicity and soil structure. Aust. J. Soil Res. **1991**, *29*,935–952.
6. Rengasamy, P.; Sumner, M.E. Processes involved in sodic behaviour. In *Sodic Soils: Distribution, Properties, Management and Environmental Consequences*; Oxford University Press: New York, 1998: 35–50.
7. Rengasamy, P.; Marchuk, A. Cation ratio for soil structural stability (CROSS). Soil Research **2011**, *49*, 280–285.
8. Chorom, M.; Rengasamy, P. Dispersion and zeta potential of pure clays as related to net particle charge under varying pH, electrolyte concentration and cation type. Eur. J. Soil Sci. **1995**, *46*, 657–665.
9. Brady, N.C.; Weil, R.R. Soils of dry regions: Alkalinity, salinity, and sodicity. In *The Nature and Properties of Soils*, 14th Ed.; Pearson/Prentice Hall: Columbus, OH, 2008; 401–442.
10. Northcote, K.H.; Skene, J.K.M. *Australian Soils with Saline and Sodic Properties. Soil Publication 27*; CSIRO Publication; Melbourne, Australia, 1972; 1–62.
11. Sposito, G. The soil solution. In *The Chemistry of Soils,* 2nd Ed.; Oxford University Press: New York, 2008; 94–118.
12. Rengasamy, P.; Chittleborough, D.; Helyar, K. Root-zone constraints and plant-based solutions for dryland salinity. Plant Soil. **2003**, *257*, 249–260.
13. Rengasamy, P. Osmotic and ionic effects of various electrolytes on the growth of wheat. Aust. J. Soil Res. **2010**, *48*, 120–124.

14. Elgharably, A.; Marschner, P.; Rengasamy, P. Wheat growth in a saline sandy loam soil as affected by N form and application rate. Plant Soil. **2010**, *328*, 303–312.
15. Cramer, G.R. Kinetics of maize leaf elongation. II. Response of a Na-excluding cultivar and a Na-including cultivar to varying Na/Ca salinities. J. Exp. Bot. **1992**, *43*, 857–864.
16. Colmer, T.D.; Voesenek, L.A.C.J. Review. Flooding tolerance: Suites of plant traits in variable environments. Funct. Plant Biol. **2009**, *36*, 665–681.
17. Barrett-Lennard, E.G. The interaction between waterlogging and salinity in higher plants: Causes, consequences and implications. Plant Soil **2003**, *253*, 35–54.
18. Passioura, J.B. Scaling up: The essence of effective agricultural research. Funct. Plant Biol. **2010**, *37*, 585–591.

9

Sodic Soils: Irrigation Farming

David Burrow

Properties of Sodic Soils Affecting Irrigation Farming

Hydraulic Properties

The movement of water into the soil following rainfall or surface irrigation, its redistribution within subsurface layers, and drainage below the rootzone are important for productive irrigation farming. Water entry into the soil, measured by the infiltration rate (IR), and through soil layers, measured by the hydraulic conductivity (HC), both decrease with increasing exchangeable Na and decreasing soil salinity. For strongly sodic soils, heavy rainfall or low-salinity irrigation may lead to surface waterlogging on flat land or run off on sloping ground.[1] Similarly, freshwater irrigation of sodic soils with low HC may lead to shallow depths of wetting, uneven wetting, and low plant available water capacity.[2] These hydraulic traits, together with lower air-filled porosity and excessive penetration resistance in dry soils, can restrict plant growth.[3]

The low IR of a strongly sodic soil during irrigation is usually attributed to the formation of a surface seal or crust, which in turn derives from the physiochemical dispersion of clay-sized particles from unstable surface aggregates. These dispersed clay particles migrate with water flow and eventually block conducting pores.

Early studies showed that soil HC varies up to two orders of magnitude, depending on the salt concentration of a percolating solution and soil-exchangeable sodium percentage.[4] Reductions in HC were explained by swelling and dispersion of clay particles but these factors are modulated by clay mineralogy, sesquioxide and $CaCO_3$ content, soil texture, organic matter content, pH, soil weathering, and other exchangeable cations.

The effects of soil sodicity on IR and HC, resulting from saline–sodic irrigation are exacerbated in climatic zones with a marked rainfall season because of the leaching of salts from surface soil horizons. Rainfall, with negligible salinity (EC <0.1 dS/m) infiltrates and reduces soil EC below a threshold electrolyte concentration (TEC) in these layers, resulting in clay swelling or dispersion. In the field, these processes are not necessarily detrimental to HC as they may be accommodated by pore geometry and slow wetting rates, or circumvented by preferential flow through large cracks. However, the maintenance of soil EC above TEC is an essential first step toward managing sodic soils under irrigation.

Nutrient Status

In sodic soils, high Na/Ca and Mg/Ca ratios reduce the amount of Ca available for plant uptake and can induce Ca deficiencies in crops. Data, obtained principally from greenhouse and laboratory studies, indicate that cereal species are sensitive to Ca deficiencies and particularly during seedling growth stages.[5]

Poor productivity on sodic soils has also been attributed to lowered availability and plant uptake of nitrogen or micronutrients. Nitrogen availability in these soils is restricted under anaerobic conditions,[6] while micronutrient availability is controlled by adsorption and precipitation reactions.[7]

The relative solubility of boron (B) and high Na/Ca ratios in some sodic soils also means that B and Na toxicity may contribute to crop production losses. However, the relative importance of nutrient availability and specific ion toxicity on crop production losses for sodic soils is difficult to judge, because of the compounding influence of poor soil physical properties on crop growth and yield.

Soil Management

Specific tillage operations and cropping systems have been designed to overcome some of the poor physical properties of sodic soils.[1,8] Field grading and raised-bed cropping systems are used in landscapes of low gradient to encourage runoff and limit surface waterlogging. Deep ripping is used to break up massive subsoil structures and for the incorporation of gypsum into subsurface horizons. Shallow tillage, at moisture contents close to the lower plastic limit, is used to generate seedbeds of fine tilth, rather than massive clods. Following crop establishment, zero tillage is typically advocated in order to minimize mechanical destruction of surface aggregates and subsequent dispersion of clays during irrigation.

Tillage practices, which aim to stabilize surface soil aggregates of an appropriate size distribution, are aided by the use of cover crops, mulch applications, and gypsum applications.[1] Soil cover crops, such as *Lolium rigidum,* tend to stabilize topsoil aggregates by means of entwining root systems and the incorporation of root mucilages within the rootzone. Mulches are used for the creation of a more favorable soil microclimate, which encourages turnover in microbial populations, microbial detritus, and the generation of stabilizing, organic glues. Careful applications of gypsum or other calcium products at the soil surface aim to reduce soil sodicity or maintain soil EC above TEC.

Irrigation Management

Water Quality

Water quality guidelines for crop production based on salinity (EC_{iw}) and sodicity (SAR_{iw}) have been developed for infiltration, crop salinity tolerance, and specific ion toxicity.[8] In contrast, water quality guidelines for soil permeability (HC) and potential dispersion of clay tend to be based on soil chemical properties.

According to infiltration guidelines, calculated on paired EC_{iw} and SAR_{iw} ranges, the quality of irrigation waters used throughout the world (Table 1) should rarely limit their application to land.[8] However, when the EC of infiltrating water falls below 0.2 dS/m, e.g., as occurs after rainfall, severe infiltration problems are predicted. Crop salinity tolerance guidelines predict severe Na toxicity for sensitive crops, based on $SAR_{iw}>9$, and severe salinity effects on crop productivity, based on $EC_{iw}>3$ dS/m. However, these predictions when applied to data in Table 1 only concern irrigation of crops with drainage waters, because of their typically high SAR_{iw} and EC_{iw}.

TABLE 1 Mean and Range of Values for EC_{iw}, SAR_{iw}, and adj R_{Naiw} of Irrigation Sources throughout the World

Measure	Canals	Drains	Ground Waters	Rivers	Wadis	Wells	Waste Waters	All Sources
Samples	11	5	4	88	12	115	5	250
Countries	5	3	2	29	3	42	3	54
EC	1.37	4.38	1.45	0.80	2.74	1.78	0.84	1.47
	0.20–4.15	*1.06–6.20*	*0.46–3.80*	*0.03–7.42*	*0.30–8.01*	*0.14–7.93*	*0.69–1.11*	*0.03–8.01*
SAR	4.75	9.98	1.43	1.87	8.63	5.12	2.46	4.04
	0.60–14.0	*3.50–18.0*	*1.20–1.90*	*0.10–14.0*	*0.30–36.0*	*0.20–49.0*	*0.40–4.10*	*0.10–49.0*
adj RNa	3.91	11.74	1.98	1.99	9.3	5.73	2.78	4.42
	0.60–11.2	*3.70–22.0*	*1.30–3.80*	*0.10–18.0*	*0.30–38.0*	*0.20–66.0*	*0.50–4.70*	*0.10–66.0*

Note: Units ECiw (dS = m), SARiw, and adj RNa iw ([mmolc/L]0.5).
[a] Range values are shown in italics.
Source: Ayers and Westcot.[8]

The use of an adjusted SAR_{iw} (adj $R_{Na\ iw}$) in Table 1, accounts for the effects of high bicarbonate concentrations in some irrigation waters. Bicarbonate tends to precipitate Ca, and to a lesser extent Mg, as a carbonate solid, with the net result of an increase in SAR_{iw}. Transformations of Ca (and Mg) between soluble, exchangeable, and precipitated forms are modulated by pH and the partial pressure of CO_2, hence these two factors indirectly affect SAR_{iw}.

Irrigation Practices

Specific irrigation practices complement tillage and chemical additions to improve the production potential of sodic soils. These practices include increased frequency of irrigation, preplanting irrigation, and adoption of alternative irrigation systems. Uniform applications of polyacrylamide or finely powdered gypsum to topsoils, for the purpose of stabilizing fragile aggregates, are also best achieved through an initial irrigation.

Increasing the frequency of short-duration irrigations on soils with high initial IR (e.g., cracking clays) and low final IR (due to sodicity), maximizes the periods of high IR to wet the rootzone, and decreases the risk of waterlogging. Preplanting irrigation wets the entire rooting depth and if timed correctly allows restoration of soil aeration before crop planting. This technique is especially effective on sodic soils of very low IR. A change from surface irrigation systems (e.g., flood or furrow) to sprinklers or drippers, can alleviate waterlogging on low IR soils by matching water delivery rates to the IR. The slow wetting of topsoils, under these alternative irrigation systems, also decreases aggregate slaking and clay dispersion.

Drainage of Sodic Soils

Drainage is essential for the continued practice of irrigation farming on sodic soils. Surface drainage prevents waterlogging and the development of anaerobic conditions in the topsoil, while adequate subsurface drainage allows leaching of toxic or plant-dehydrating salts from the profile.

For some irrigated regions, the presence of a shallow watertable restricts drainage through a diminished hydraulic gradient. These watertables are frequently saline–sodic and contribute to sodification of soils via capillary transmission of Na-salts. Hence, the removal in these regions of subsurface water through drains (open or piped) and pumped bores is essential for sodium management. Because installation of open or piped drains is expensive, the use of cheaper mole drains within subsurface layers has been proposed. However, in unstable sodic soils, mole drains soon collapse and therefore are not generally used.[9]

Periodic leaching of soil profiles is required for removal of salt or toxic ions. Analytical equations have been developed to predict the leaching requirement for soil sodicity under saline–sodic irrigation.[10]

However, these equations do not fully account for Na-induced changes in IR or HC, which impinge on leaching capability.

Practically, intermittent irrigation is more efficient in salt removal than continuous ponding by promoting diffusion of salts into more favorable leaching pathways. Efficient leaching of sodic soils also requires irrigation with water of low SAR_{iw} and $EC_{iw} > TEC$, together with provision of sub-surface drainage for removal of salts.

Conclusions

A range of farm management practices on irrigated, sodic soils have been briefly described. Adequate control of water quality with respect to SAR_{iw} and EC_{iw}, and provision of drainage to remove excess sodic salts from the rootzone are pivotal in controlling soil sodicity. Concurrent management of surface soils with ameliorants (e.g., gypsum, organic polymers, mulches), cover crops, and appropriate irrigation methods enhances farm operations on these sodic soils at relatively low cost. Adequate modification of drainage and sodicity in subsoils is more costly and therefore less frequently attempted, particularly for cereal and forage crops. Consequently, the control of soil sodicity in irrigated farming still remains focused on irrigation water quality, irrigation systems and scheduling, provision of surface runoff, and application of soil ameliorants to the soil surface. The continual emergence of irrigation and soil chemistry simulation software for computers should enhance further scientific research and refine farm management of soil sodicity.

References

1. Oster, J.D.; Jayawardane, N.S. Agricultural management of sodic soils. In *Sodic Soils: Distribution, Properties, Management, and Environmental Consequences;* Sumner, M.E., Naidu, R., Eds.; Oxford University Press: New York, 1998; 125–147.
2. Shaw, R.J.; Coughlan, K.J.; Bell, L.C. Root zone sodicity. In *Sodic Soils: Distribution, Properties, Management, and Environmental Consequences;* Sumner, M.E., Naidu, R., Eds.; Oxford University Press: New York, 1998; 95–106.
3. Jayawardane, N.S.; Blackwell, J. The effects of gypsumen-riched slots on moisture movement and aeration in an irrigated swelling clay. Aust. J. Soil Res. **1985**, *23,* 481–492.
4. Fireman, M. Permeability measurements on disturbed soil samples. Soil Sci. **1944**, *58,* 337–353.
5. Curtin, D.; Naidu, R. Fertility constraints to plant production. In *Sodic Soils: Distribution, Properties, Management, and Environmental Consequences*; Sumner, M.E., Naidu, R., Eds.; Oxford University Press: New York, 1998; 107–123.
6. Cairns, R.R.; Bowser, W.E.; Milne, R.A.; Chang, P.C. Effect of nitrogen fertilization of bromegrass on solonetz soil. Can. J. Soil Sci. **1967**, *47,* 1–6.
7. Naidu, R.; Rengasamy, P. Ion interactions and constraints to plant nutrition in Australian sodic soils. Aust. J. Soil. Res. **1993**, *31,* 801–819.
8. Ayers, R.S.; Westcot, D.W. Water quality for agriculture. In *FAO Irrigation and Drainage Paper (Rev. 1)*; Food and Agricultural Organization of the United Nations: Rome, 1985; Vol. 29, 174 pp.
9. Muirhead, W.A.; Christen, E.; Moll, J. Mole Drains Under Slots Reduce Waterlogging and Salinity in an Irrigated Clay Soil of Semi-Arid Australia, XII World Congr. Agric. Eng. Rep. 94-A-018.
10. Jurinak, J.J.; Suarez, D.L. The chemistry of salt-affected soils and waters. In *Agricultural Salinity Assessment and Management*; Tanji, K.K., Ed.; American Society of Civil Engineers: New York, 1990; 42–63.

II

COV: Comparative Overviews of Important Topics for Environmental Management

10

Agricultural Soils: Carbon and Nitrogen Biological Cycling

Alan J. Franzluebbers

Introduction

Carbon (C) and nitrogen (N) are two of the most important elements that affect soil productivity and environmental quality.[1] Carbon is found throughout nature in a wide variety of forms and particularly in soil as 1) complex organic compounds (e.g., carbohydrates—$C_xH_{2x}O_x$, lignin, etc.,) derived from living organisms; 2) carbonate minerals such as calcite ($CaCO_3$) and dolomite [$CaMg(CO_3)_2$]; and 3) carbon dioxide (CO_2) and methane (CH_4) as decomposition end products. Nitrogen is an essential element of plants, animals, and microorganisms—a part of chlorophyll, enzymes, amino acids, and proteins, which are necessary for growth and development of organisms. In typical unpolluted soil, quantity of N in organic matter and fixed as ammonium (NH_4^+) in clay minerals far exceeds quantities in plant-available forms of soluble nitrate (NO_3^-) and NH_4^+. Among several soils in North America, total N in 1 m depth of soil was 16.0 ± 6.9 Mg N ha^{-1} with 13% ± 15% fixed as NH_4^+ in clay minerals and <1% as soluble NO_3^-.[2]

Agriculture, i.e., the growing of plants and animals for human and livestock consumption, is a widespread land usage throughout the world. Globally, agriculture occupies approximately 38% of the total land area with 1.5 billion ha in cropland and 3.4 billion ha in perennial grassland.[3] Addition of N, phosphorus (P), potassium (K), and other nutrients to soil is often needed to satisfy the demands by high-production crops and forages. A portion of these nutrients is naturally supplied through plant residue and soil organic matter decomposition, but amendment with inorganic or organic fertilizers is often needed to achieve high production. Unfortunately, there are many pathways for nutrients to escape from the agricultural landscape into nearby streams, lakes, groundwater, and the atmosphere. Preventing these losses is one of the goals of sustainable, ecologically based approaches to agricultural production.

Carbon and Nitrogen Cycles

Both C and N are biologically fixed from inorganic atmospheric forms to organically bound plant and microbial forms. Photosynthesis converts inorganic CO_2 from the atmosphere into organic carbohydrates in plants, algae, and cyanobacteria. Biological N fixation is a unique transformation carried out by a number of bacteria, which convert N_2 gas into ammonia (NH_3) for biological utilization. N-fixing bacteria are most prevalent in symbiotic relationships with plants, such as *Rhizobium* that forms nodules on the roots of clovers where the nitrogenase enzyme catalyzes the reaction. Fertilizer manufacturing converts N_2 gas into NH_3 in a similar manner without an enzyme, but rather large quantities of energy necessary to create the pressure required for the transformation.

Under certain conditions, both inorganic C and N can be chemically fixed in the subsoil. Carbon dioxide forms carbonic acid in water, which can precipitate with the basic cations, Ca^{2+}, Mg^{2+}, and Na^{+}, to form pedogenic carbonates. Inorganic C is most abundant in soils of the semiarid and arid regions. Ammonium can be fixed as nonexchangeable components of the lattice structure of 2:1-type clay minerals, which are especially prevalent in the subsoil of many younger soils.

Carbon and N occur in various forms and undergo transformations from one form to another, primarily through biochemical manipulations involving enzymes.[4,5] Enzymes are proteins, functioning to catalyze very specific reactions either 1) intracellularly within plants, microorganisms, or soil animals; or 2) extracellularly in soil solution or attached to soil colloids. Some major enzyme categories and their reactions with C and N substrates in soil are 1) hydrolases, such as amylase and cellulase, which hydrolyze various carbohydrate and macromolecular compounds; 2) oxyreductases, which catalyze various electron transfer reactions; 3) proteinases, which convert proteins to amino acids; 4) lignocellulases, which catalyze the ecologically resistant step of lignin breakdown; and 5) lyases, which form double bonds through reactions other than hydrolysis or oxidation. Two key enzymes involved in the fixation of C and N into organic forms are 1) ribulose bisphosphate carboxylase (rubisco), which is the photosynthetic enzyme catalyzing the transformation of CO_2 from the atmosphere into carbohydrates; and 2) nitrogenase, which catalyzes the biological N fixation reaction in symbiotic bacteria associated with leguminous plant roots:

$$N_2 + 6\ H + \text{energy} \rightarrow 2\ NH_3$$

Forms and fluxes of an element are commonly illustrated in a cycle following the principles of conservation of mass (i.e., elements are transferred from one molecule to another) (Figure 1). Carbon and N cycles have global dimensions with terrestrial, aquatic, and atmospheric components of major significance.[6,7] The sun initiates a chain of energy reactions, which drive elemental cycles. The elemental cycles of C and N interact closely with the water cycle, as water is a fundamental internal component of life and a major transport mechanism of nutrients.

In natural systems without significant import of N from fertilizers, the cycling of N is largely dependent upon the cycling of C. Since growth of plants is often limited in N supply due to the strong competition for N by soil microorganisms, which have a steady supply of C-rich substrates at the surface of undisturbed soil, N losses from natural systems are typically low.[8] The need for additional N in agricultural systems can be historically derived from two major pathways: (1) high protein harvest of grain, forage, and animal products that requires supplemental N to replace the already limited N supply in natural systems (and eventual lack of recycling waste and manure by-products from harvested food products back to the land); and (2) loss of soil-surface residue cover and soil organic matter with intensive tillage that initially stimulates N release to crops, but that eventually exhausts the soil resource in its ability to supply N to crops. Loss of C-rich surface residue and soil organic matter essentially removes the C stimulus needed to conserve N in soil, thereby resulting in major losses of N from agricultural systems with time and creating a system that relies heavily on external N inputs to supply crops with not only the N removed from harvested crops but also the N lost via leaching, runoff, volatilization, and

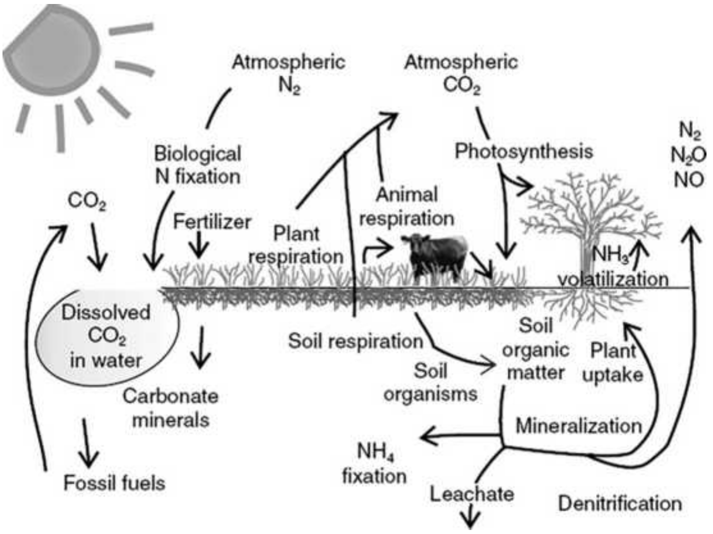


FIGURE 1 Generalized diagram of the C and N cycles in soil.

denitrification. Estimated global production of N and P fertilizers was 100 and 41 Tg (10^{12} g), respectively, in 2007,[3] which compared to about 10 Tg of production for each nutrient in 1960.[9]

Autotrophic fixation of atmospheric CO_2 by plants captures the energy of the sun within organic compounds via the process of photosynthesis (Figure 1):

$$6\ CO_2 + 6H_2O + \text{energy} \rightarrow C_6H_{12}O_6 + 6\ O_2$$

Inorganic N is taken up by plant roots and synthesized into amino acids and proteins during plant development. Plants are eventually consumed by animals or microorganisms, transferring portions of this stored energy through biochemical processes into various cellular components. Once in soil, the C cycle is dominated by the heterotrophic process of decomposition, i.e., the breakdown of complex organic compounds into simple organic constituents. Mineralization is the complete decomposition of organic compounds into mineral constituents:

$$C_6H_{12}O_6 + 6O_2 \rightarrow 6\ CO_2 + 6H_2O + \text{energy}$$

$$R - NH_2 + H_2O \rightarrow R - OH + NH_3$$

Immobilization of N occurs simultaneously with N mineralization when soil organisms require additional inorganic N to meet the high demand for new body tissue while decomposing C-rich substrates low in available N. Net N mineralization occurs when gross N mineralization exceeds that of N immobilization.

Environmental Influences on Soil Microbial Activity

Organisms predominantly responsible for decomposition of organic matter and associated mineralization of C and N are soil microorganisms, composed of bacteria, actinomy-cetes, fungi, and protozoa.[10,11] Soil fauna are larger soil organisms, such as beetles and earthworms (macrofauna, >2 mm width × >10 mm length), collembolan and mites (mesofauna, 0.1–2 mm width × 0.2–10 mm length), and protozoa and nematodes (microfauna, <0.1 mm width × <0.2 mm length), that also indirectly affect C and N cycling by 1) comminuting plant residues and exposing a greater surface area to soil microorganisms; 2) transporting plant and animal residues to new locations in the soil to facilitate decomposition, interaction with soil nutrients, or isolation from environmental conditions; 3) inoculating partially digested organic substrates with specific bacteria and enzymes; and 4) altering physical characteristics of soil by creating burrows, fecal pellets, and distribution of soil particles that influence water, air, nutrient, and energy retention and transport. With suitable environmental conditions, soil microorganisms grow rapidly in response to the availability of organic substrates rich in C and N.

Soil Temperature

Temperature controls both plant and soil microbial activity, although not at the same level (Figure 2). Plant and soil microbial activity are limited by low temperature resulting in low photosynthetic potential, as well as low decomposition potential. For many plants, net photosynthetic activity is optimized between 20 and 30°C, because at higher temperatures, plant respiration consumes energy for maintenance. In many temperate soils, microbial activity is maximized between 30 and 35°C and decreases at higher temperatures. An intermediate temperature is often ideal for maximizing C retention in soil, because optimum plant activity competes well against soil microbial activity.

Soil Water Content

Diversity of soil microorganisms is greatest under aerobic conditions, where maximum energy is obtained. However, there are a number of soil bacteria that thrive under anaerobic conditions, in which alcohols, acetic acid, lactic acid, and CH_4 become C end products via fermentation and nitrate is

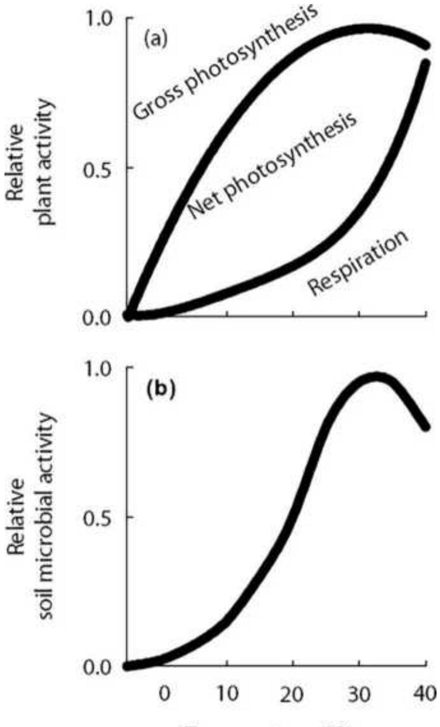


FIGURE 2 Typical responses of plant and soil microbial activities to temperature.

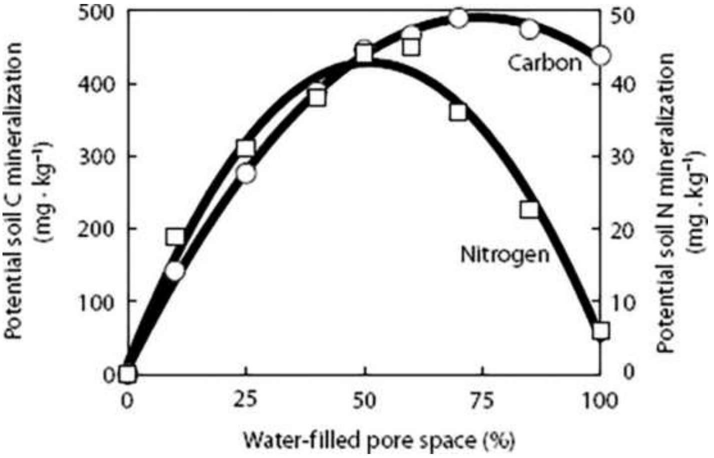


FIGURE 3 Responses of potential soil C and N mineralization to water-filled pore space in Typic Kanhapludults in Georgia, USA [air-filled pore space would be 100 − (water-filled pore space)].
Source: Franzluebbers.[12]

converted to N gases (e.g., N_2, N_2O, NO) via the process of denitrification. Soil C and net N mineralization are maximized at an optimum balance between soil moisture and oxygen availability (Figure 3). Significant denitrification occurs at water-filled pore space >70%, resulting in low availability of inorganic N to plants.

Soil Texture

Soil texture can influence both the quantity of C and N accumulation in soil and their potential mineralization. Potential C mineralization is often greater in coarse-textured soils than in fine-textured soils, due to both increased microbial predation by soil fauna and greater accessibility of organic substrates in coarse-textured soils. Organic C and N can also be protected from decomposition when bound within soil aggregates. Water-stable aggregates are a coherent assemblage of primary soil particles (i.e., sand, silt, clay) cemented through natural forces and substances derived from root exudates and soil microbial activity.

Spatial Distribution of Organic Substrates

Distribution of organic substrates in soil has a major impact on potential C and N mineralization. Potential C mineralization is often several-fold greater in the rhizosphere (i.e., 0–5 mm zone surrounding roots) than in bulk soil. However, because of the high demand for N by plant roots and the stimulated soil microflora, net N mineralization is often initially lower in the rhizosphere because of immobilization of N. Keeping soil active with roots whenever conditions are conducive for plant growth will 1) keep inorganic N at low levels (as well as keep soil covered with protective plant cover to guard against soil erosion); 2) stimulate soil biological activity; and 3) create a richly diverse soil microbial community, all of which prevent nutrients from being lost from the soil.

Surface soil often contains greater quantities of organic matter than at lower depths due to surface deposition of plant residues, as well as greatest plant root activity. Surface soil usually undergoes the most extreme drying/wetting cycles and has the greatest exchange of gases, both of which contribute to enhanced soil microbial biomass and activity. Tillage of soil with traditional agriculture redistributes organic substrates uniformly within the plow layer, often resulting in immediately stimulated soil microbial activity from disruption of organic substrates protected within stable soil aggregates.

Minimum soil disturbance with conservation tillage practices can reduce oxidation of soil organic matter and preserve more C within soil, which can have implications for potentially mitigating the greenhouse effect.[13]

Stratification of soil organic matter with depth is common in natural ecosystems and in conservation agricultural systems (Figure 4). Conservation agricultural systems are defined as those that 1) minimize soil disturbance with tillage; 2) maximize soil-surface cover with continuous plant and/or residue cover; and 3) stimulate biological activity through diverse crop rotations and integrated nutrient and pest management.

Depth stratification of soil organic matter with time occurs when soils remain undisturbed from tillage (e.g., with conservation tillage and pastures) and sufficient organic materials are supplied to the soil surface (e.g., with cover crops, sod rotations, and diversified cropping systems). Depth stratification with time can be viewed as an improvement in soil quality, because several key soil functions are enhanced, including water infiltration, conservation and cycling of nutrients, and sequestration of C from the atmosphere.[15] Depth stratification of soil organic C generally reduces water runoff volume and soil loss from agricultural fields. Grasslands often reduce water runoff volume and soil loss even further than with conservation-tilled cropland due to even greater accumulation of surface soil organic matter. Total runoff loss of nutrients is often lower with conservation tillage than with conventional tillage, because of a reduction in sediment-borne nutrients (Figure 5). Soluble (or dissolved) N and P in water runoff can be a threat to water quality with excessive nutrient applications from fertilizers and manures (even under conservation management), and therefore, further research is being conducted to identify ways of reducing nutrient loss.[31]

Stratification ratio of soil organic C has been proposed as an index of soil quality, because soil-surface enrichment of organic matter is important for improving water-stable aggregation, water infiltration and storage, nutrient cycling, and soil microbial biomass, activity, and diversity.[32] In a land-use survey in the southeastern United States, stratification ratio of soil organic C was related to the total stock of soil organic C in the surface 20 cm depth (Figure 6). This relationship indicates that the majority of C stored with conservation management in these Ultisols and Alfisols of the region occurred within the surface 5 cm.

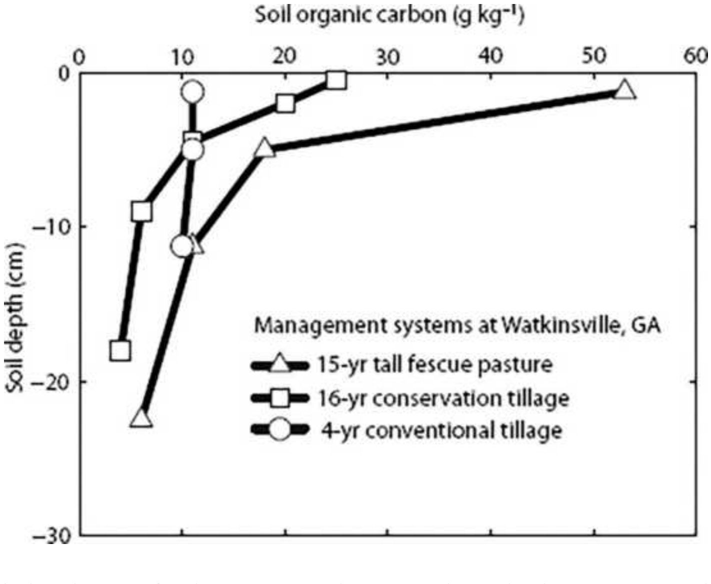


FIGURE 4 Depth distribution of soil organic C under pastured grassland, conservation-tillage cropland, and conventional-tillage cropland on a Typic Kanhapludult in Georgia, USA.
Source: Schnabel et al.[14]

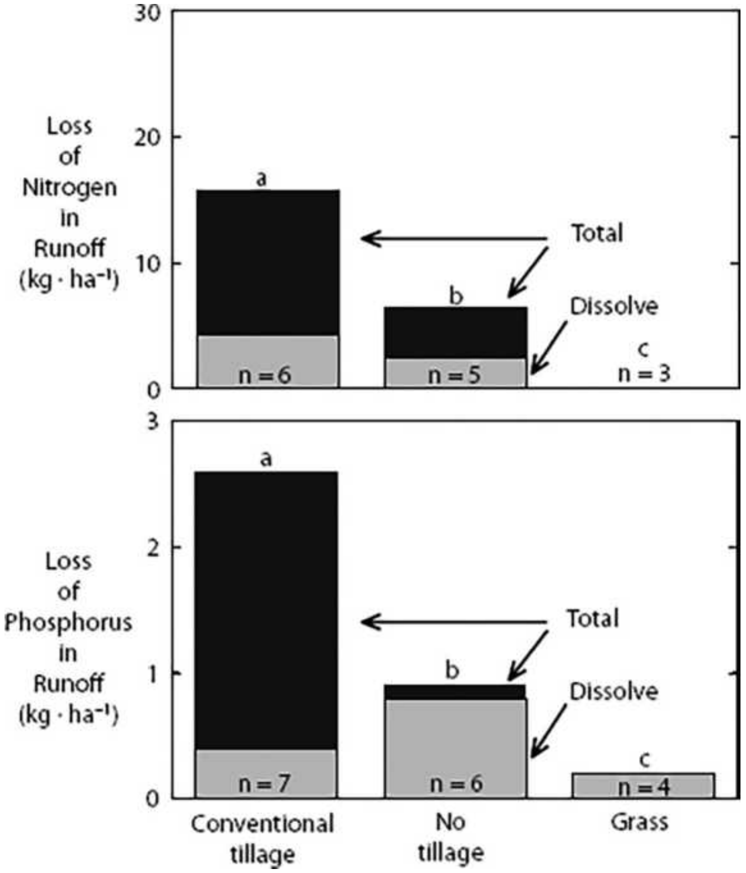


FIGURE 5 Mean loss of N and P in water runoff across several water catchment studies in the USA.
Source: Data from Van Doren et al.,[16] Langdale et al.,[17] Blevins et al.,[18] Seta et al.,[19] Sharpley and Smith,[20] Shipitalo and Edwards,[21] Endale et al.,[22] Endale et al.,[23] Endale et al.,[24] Ross et al.,[25] Rhoton et al.,[25] Rhoton et al.,[26] Sharpley and Kleinman,[27] Truman et al.,[28] Harmel et al.,[29] and Bosch.[30]

Organic Substrate Quality

Quality of organic substrates has a major influence on the rate of decomposition and the transformations that occur in soil. Plant residues do not vary greatly in total C concentration on a dry-weight basis (e.g., 37–47 mg g^{-1}), but do vary enormously in the type of C compounds, which determine its quality or conversely its resistance to degradation. The diversity of organic compounds attacked by soil microorganisms is extensive (e.g., organic acids, polysaccharides, lignins, aromatic and aliphatic hydrocarbons, sugars, alcohols, amino acids, purines, pyrimidines, proteins, lipids, and nucleic acids). Almost all naturally occurring organic compounds, and even most synthetic organic compounds, are susceptible to decomposition given the appropriate environment, microbial community, and time.[34,35] Generally, the primary components of plants can be categorized according to relative rate of decomposition: rapid (sugars, starches, fats, and proteins), intermediate (cellulose and hemicellulose), and slow (lignin and lignocellulose). Young plants are of high quality and low resistance to decomposition, whereas with aging, lignin and polyphenolic concentrations increase, resulting in greater resistance to decomposition. Low N concentration of organic amendments usually results in temporary net N immobilization into microbial biomass, which grows rapidly in response to the availability of organic C. Soil microbial biomass typically maintains a C-to-N ratio of 10 ± 5. Following a proliferation of microbial biomass

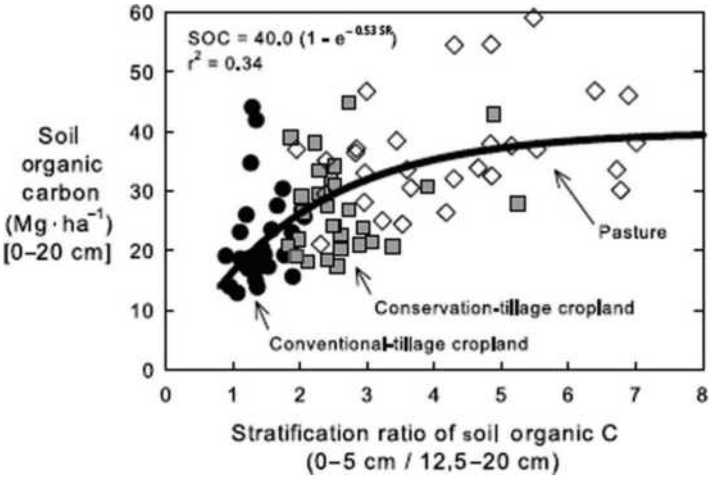


FIGURE 6 Relationship of soil organic C storage at a depth of 0–20 cm to the stratification ratio of soil organic C among conventional-tillage, conservation-tillage, and pasture land uses on different soils throughout the southeastern USA.
Source: Causarano et al.[33]

that depletes the source of readily decomposable organic C, N in excess of microbial demands becomes mineralized and available for plant uptake (Figure 7). In general, plant residues with C-to-N ratio >40 will result in longer periods of net N immobilization.

Soil Organic Matter

Soil organic matter is composed of a large diversity of organic compounds that can be characterized in many ways. A useful separation of soil organic matter for modeling is based on turnover times, whereby at least three pools can be defined: 1) active (composed of microbial biomass and light fraction material with a turnover time of <1 year); 2) passive (composed of macroorganic matter and protected organic matter with a turnover time of 3–10 years); and 3) slow (stable humus fraction with a turnover time of >100 years). Fractions of soil organic matter have also been methodologically characterized and can be tied to above-mentioned kinetic pools.[36] Active fractions have been characterized using soil microbial biomass techniques (chloroform fumigation-incubation, chloroform extraction, substrate-induced respiration, microwave irradiation, phospholipids, and flush of CO_2) and determination of labile substrates (mineralizable C and N during incubation, hot-water extractable C and N, dilute permanganate oxidizable C, carbohydrates, light-fraction C and N, amino sugars). Passive fractions have been estimated using particulate organic matter, glycoproteins, and humic materials. Slow fractions of soil organic matter have been characterized with various resistant components, such as charcoal, lignins, aliphatic macromolecules, humin, nonhydrolyzable organic matter, and silt-and clay-associated organic matter.

Losses of Nitrogen from Soil

Nitrogen cycling in soil is different from that of C because of the more numerous transformations that can occur upon mineralization to an inorganic form.[2] Mineralization of N from organic matter results in NH_4^+released into soil solution. In the presence of nitrifying bacteria, NH_4^+ is converted to NH_3^-, a process called nitrification. The fate of NH_3^- in soil depends upon environmental conditions. Active

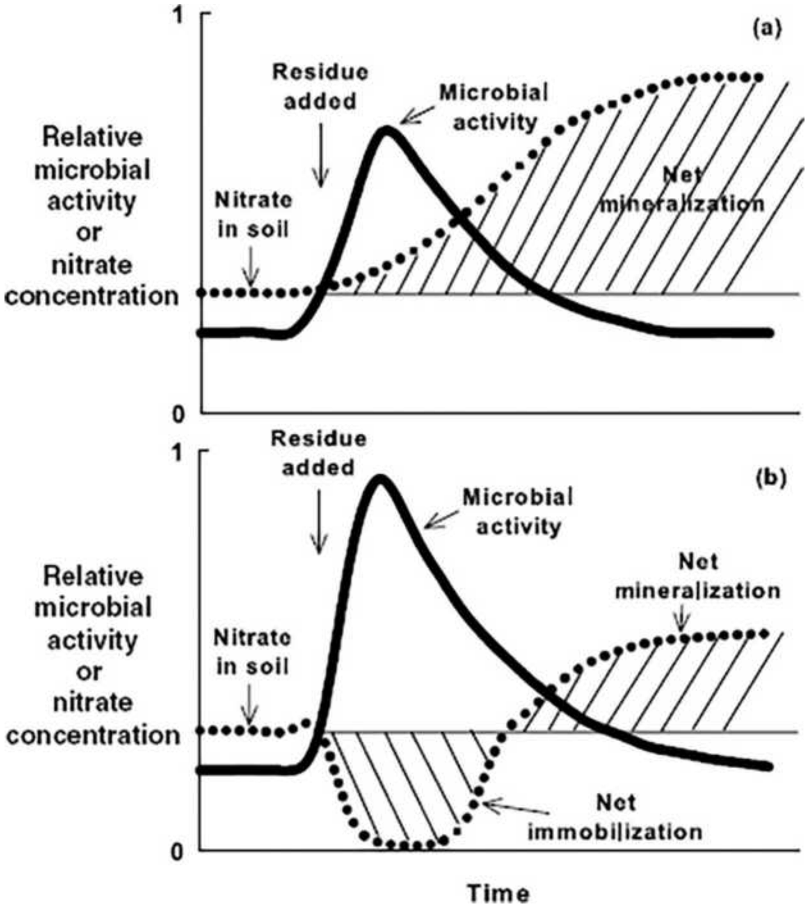


FIGURE 7 Typical responses in soil microbial activity and soil nitrate concentration with the addition of plant residue of (a) high N concentration and (b) low N concentration.

plant growth in natural and agricultural systems would provide enough demand to remobilize N into organic forms. However, NH_3^-can be used as an electron acceptor in place of O_2 under anaerobic conditions, resulting in gaseous loss of N to the atmosphere via denitrification. In temperate soils with a net negative charge on colloidal surfaces, NH_3^- can readily leach into the vadose zone and contaminate groundwater if not taken up by plants or denitrified. In tropical soils with a net positive charge, NH_3^- can be retained on anion exchange sites. Opposite behavior of the cations, NH_4^+ and NH_3^-, occurs with respect to clay mineralogy.

Humans, as well as roving animals, impose great demands on the C and N cycles. Management of agricultural and forest land for food and fiber often removes nutrients from soil for consumption and utilization elsewhere. Return of these nutrients to soil is possible when municipal and agricultural solid wastes and wastewater are applied to land. Losses of C and N from managed lands also occurs through soil erosion, which transports nutrients via 1) water from overland flow into streams, lakes, and oceans and 2) air as dust from bare land surfaces.

Volatilization of NH_3 to the atmosphere is possible when NH_4^+ is exposed to alkaline soil conditions. Significant ammonia volatilization can occur with surface application of urea fertilizer to non-acidic soils, from animal manures, and from green plant materials not incorporated into the soil.

Strategies to Mitigate Loss of Nitrogen from Agricultural Soils

Loss of nitrogen from agricultural soils occurs through harvest, runoff, leaching, volatilization, and denitrification. These losses are potential threats to the environmental quality of water and air resources, as well as to the quality of soil itself. Harvest losses of N can be accepted as practically necessary, but such losses can be partially mitigated by returning waste by-products from food processing and from animal and human consumption back to the soil as organic amendments. This recycling is an important step towards global sustainability of nutrient use and protection of the environment from nutrient loading, particularly relevant in industrialized countries, which seem to have abandoned this age-old practice.

Runoff losses of N can be controlled by protecting the soil surface from soil loss via erosion and by creating a porous surface to allow rainfall to infiltrate rather than runoff. Various conservation agricultural approaches are available to mitigate runoff losses, including reduced or no tillage (i.e., conservation tillage), intensified crop rotations to avoid bare fallow periods, winter cover cropping, crop–pasture rotation, moderate grazing of perennial pastures, and timely and deep placement of fertilizers. Loss of N in runoff can occur as part of N-rich sediment (both organic and inorganic N) and as soluble inorganic N in overland flow of water. Stopping sediment loss is most effective by protecting the soil surface with plants and surface residue, both of which mitigate the energy of rainfall impact and, therefore, avoid soil detachment. Surface application of fertilizers, whether inorganic or organic, is susceptible to runoff loss as water flows over the landscape even if soil is not detached. Therefore, timing fertilizer application to when the plant needs it the most will be an effective N-loss mitigation strategy, as well as placing fertilizer into the soil in proximity with roots with deep banding will limit access to overland flow potential.

Leaching of N through the soil profile is a concern in well-drained soils. Conditions for significant leaching often occur due to overapplication of N fertilizer and long bare-fallow periods that limit plant uptake of residual fertilizer in the soil profile. Mitigation of N leaching can be through continuous plant growth with diverse crop rotations and winter cover cropping. Perennial pastures are also often effective in mitigating N leaching, because of the extended growing season of diverse assemblages of forages. Limiting the quantity of N fertilizer applied is still a basic principle to avoid potential N leaching.

Volatilization of NH_3 can occur when sufficient NH_4^+accumulates at the soil surface from partial decomposition of animal manures and N-rich plant residues and from ammonium-based fertilizer sources. Volatilization is most prevalent in soils with high pH and when soil is rather dry. Mitigation of ammonia volatilization is possible by getting animal manure and N-rich plant residues in close proximity with moist soil and avoiding surface application of N-rich organic residues and NH_4-based fertilizers on soils with high pH.

Denitrification of $NO_3^-$to N_2 (and to N_2O or NO) is a concern in soils under the following conditions: 1) low oxygen composition of the soil atmosphere, due to either high water content or rapid consumption of oxygen by vigorous microbial activity and/or poor air exchange; 2) readily decomposable source of organic C compounds for energy; 3) abundant supply of NO_3^-; and 4) suitable temperature for microbial activity. Mitigation of N loss through denitrification is possible by keeping soil well aerated by avoiding compaction and reducing the accumulation of $NO_{3.}^-$

Chemical nitrification inhibitors have been developed to inhibit or slow down the nitrification process that converts $NH_4^+$to NO_3^-Biological approaches to control susceptibility of NO_3^- to loss via leaching and denitrification focus more on maintaining a continuously growing and diverse plant root system penetrating a soil that develops excellent tilth from a healthy and diverse soil biological community, resulting in stable aggregate structure and vigorous cycling of C and N.

Conclusion

Nitrogen is a key element for the production of high-value crops and animal products around the world. High input of N in the current industrialized model of agricultural production often leaves behind a significant amount of inorganic N, such that loss from leaching, runoff, volatilization, and denitrification can occur. Ecologically based strategies of agricultural production recognize the importance of having sufficient N for production, but rely on biological cycling among plants, animals, and soil microorganisms and fauna to synchronize the release of organically bound N into inorganic forms and avoid environmental pollution. Ecological principles of high biological diversity, continuous plant growth and soil cover, and limited soil disturbance can be used in many different climatic and ecological conditions to avoid environmental pollution by excessive N entering water bodies and air systems.

References

1. Follett, R.F.; Stewart, J.W.B.; Cole, C.V., Eds. *Soil Fertility and Organic Matter as Critical Components of Production Systems*; Spec. Publ. No. 19, Soil Science Society of America: Madison WI, 1987; 166 pp.
2. Stevenson, F.J., Ed. *Nitrogen in Agricultural Soils;* Agronomy Monograph 22; American Society of Agronomy: Madison WI, 1982; 940 pp.
3. Food and Agriculture Organization of the United Nations, available at http://faostat.fao.org (accessed March 2010).
4. Burns, R.G., Ed. *Soil Enzymes;* Academic Press: London, 1978; 380 pp.
5. Tabatabai, M.A. Soil enzymes. In *Methods of Soil Analysis, Part 2. Microbiological and Biochemical Properties;* Weaver, R.W., Angle, J.S., Bottomley, P.S., Eds.; Book Series No. 5; Soil Science Society of America: Madison WI, 1994; 775–833.
6. Stevenson, F.J.; Cole, M.A. *Cycles of Soil: Carbon, Nitrogen, Phosphorus, Sulfur, Micronutrients,* 2nd Ed.; John Wiley and Sons, Inc.: New York, 1999; 427 pp.
7. Schlesinger, W.H. *Biogeochemistry: An Analysis of Global Change,* 2nd Ed.; Academic Press: San Diego CA, 1997; 588 pp.
8. Woodmansee, R.G. Comparative nutrient cycles of natural and agricultural ecosystems: A step toward principles. In *Agricultural Ecosystems: Unifying Concepts*; Lowrance, R., Stinner, B.R., House, G.J., Eds.; John Wiley and Sons: New York, 1984; 145–156.
9. Tilman, D.; Cassman, K.G.; Matson, P.A.; Naylor, R.; Po-lasky, S. Agricultural sustainability and intensive production practices. Nature **2002**, *418,* 671–677.
10. Alexander, M.A. *Introduction to Soil Microbiology,* 2nd Ed.; Krieger Publ. Co.: Malabar, FL, 1991; 467 pp.
11. Sylvia, D.M.; Fuhrmann, J.J.; Hartel, P.G.; Zuberer, D.A., Eds. *Principles and Applications of Soil Microbiology*; Prentice-Hall: Upper Saddle River, NJ, 1998; 550 pp.
12. Franzluebbers, A.J. Microbial activity in response to water-filled pore space of variably eroded southern Piedmont soils. Appl. Soil Ecol. **1999**, *11*, 91–101.
13. Lal, R.; Kimble, J.M.; Follett, R.F.; Cole, C.V., Eds. *The Potential of U.S. Cropland to Sequester Carbon and Mitigate the Greenhouse Effect;* Ann Arbor Press: Chelsea, MI, 1998; 128 pp.
14. Schnabel, R.R.; Franzluebbers, A.J.; Stout, W.L.; Sanderson, M.A.; Stuedemann, J.A. The effects of pasture management practices. In *The Potential of U.S. Grazing Lands to Sequester Carbon and Mitigate the Greenhouse Effect*; Follett, R.F., Kimble, J.M., Lal, R., Eds.; Lewis Publishers: Boca Raton, FL, 2001, 291–322.
15. Franzluebbers, A.J. Linking soil and water quality in conservation agricultural systems. J. Integr. Biosci. **2008**, *6* (1), 15–29.
16. Van Doren, D.M., Jr.; Moldenhauer, W.C.; Triplett, G.B., Jr. Influence of long-term tillage and crop rotation on water erosion. Soil Sci. Soc. Am. J. **1984**, *48,* 636–640.
17. Langdale, G.W.; Leonard, R.A.; Thomas, A.W. Conservation practice effects on phosphorus losses from Southern Piedmont watersheds. J. Soil Water Conserv. **1985**, *40,* 157–161.
18. Blevins, R.L.; Frye, W.W.; Baldwin, P.L.; Robertson, S.D. Tillage effects on sediment and soluble nutrient losses from a Maury silt loam soil. J. Environ. Qual. **1990**, *19,* 683–686.

19. Seta, A.K.; Blevins, R.L.; Frye, W.W.; Barfield, B.J. Reducing soil erosion and agricultural chemical losses with conservation tillage. J. Environ. Qual. **1992**, *22*, 661–665.
20. Sharpley, A.N.; Smith, S.J. Wheat tillage and water quality in the Southern Plains. Soil Tillage Res. **1994**, *30,* 33–48.
21. Shipitalo, M.J.; Edwards, W.M. Runoff and erosion control with conservation tillage and reduced-input practices on cropped watersheds. Soil Tillage Res. **1998**, *46,* 1–12.
22. Endale, D.M.; Cabrera, M.L.; Radcliffe, D.E.; Steiner, J.L. Nitrogen and phosphorus losses from no-till cotton fertilized with poultry litter in the Southern Piedmont. In *Proceedings of the Georgia Water Resources Conference*, Athens GA, 26–27 March 2001; 408–411.
23. Endale, D.M.; Schomberg, H.H.; Jenkins, M.B.; Cabrera, M.L.; Radcliffe, D.R.; Hartel, P.G.; Shappell, N.W. Tillage and N-fertilizer source effects on yield and water quality in a corn-rye cropping system. In *Proceedings of the Southern Conservation Tillage Conference*, Raleigh, NC, June 8–9, 2004; 37–48.
24. Endale, D.M.; Schomberg, H.H.; Steiner, J.L. Long term sediment yield and mitigation in a small Southern Piedmont watershed. Int. J. Sediment Res. **2000**, *14,* 60–68.
25. Ross, P.H.; Davis, P.H.; Heath, V.L. *Water Quality Improvement Resulting from Continuous No-Tillage Practices: Final Report*; Virginia Polytechnic Institute and State University: Blacksburg, VA, 2001.
26. Rhoton, F.E.; Shipitalo, M.J.; Lindbo, D.L. Runoff and soil loss from midwestern and southeastern US silt loam soils as affected by tillage practice and soil organic matter content. Soil Tillage Res. **2002**, *66*, 1–11.
27. Sharpley, A.; Kleinman, P. Effect of rainfall simulator and plot scale on overland flow and phosphorus transport. J. Environ. Qual. **2003**, *32,* 2172–2179.
28. Truman, C.C.; Reeves, D.W.; Shaw, J.N.; Motta, A.C.; Bur-mester, C.H.; Raper, R.L.; Schwab, E.B. Tillage impacts on soil property, runoff, and soil loss variations from a Rhodic Paleudult under simulated rainfall. J. Soil Water Conserv. **2003**, *58,* 258–267.
29. Harmel, R.D.; Torbert, H.H.; Haggard, B.E.; Haney, R.; Dozier, M. Water quality impacts of converting to a poultry litter fertilization strategy. J. Environ. Qual. **2004**, *33*, 2229–2242.
30. Bosch, D.D.; Potter, T.L.; Truman, C.C.; Bednarz, C.W.; Strickland, T.C. Surface runoff and lateral subsurface flow as a response to conservation tillage and soil-water conditions. Trans. Am. Soc. Agric. Eng. **2005**, *48,* 2137–2144.
31. Sharpley, A.N.; Herron, S.; Daniel, T. Overcoming the challenge of phosphorus-based management in poultry farming. J. Soil Water Conserv. **2007**, *62* (6), 375–389.
32. Franzluebbers, A.J. Soil organic matter stratification as an indicator of soil quality. Soil Tillage Res. **2002**, *66*, 95–106.
33. Causarano, H.J.; Franzluebbers, A.J.; Shaw, J.N.; Reeves, D.W.; Raper, R.L.; Wood, C.W. Soil organic carbon fractions and aggregation in the Southern Piedmont and Coastal Plain. Soil Sci. Soc. Am. J. **2008**, *72*, 221–230.
34. Tate, R.L., III. *Soil Organic Matter: Biological and Ecological Effects;* John Wiley and Sons, Inc.: New York, 1987; 291 pp.
35. Alexander, M. *Biodegradation and Bioremediation;* Academic Press: San Diego, CA, 1994; 302 pp.
36. Wander, M.M. Soil organic matter fractions and their relevance to soil function. In *Soil Organic Matter in Sustainable Agriculture;* Magdoff, F., Weil, R.R., Eds.; CRC Press: Boca Raton, FL, 2004; 67–102.

11

Agricultural Soils: Phosphorus

Anja Gassner and
Ewald Schnug

Introduction

Phosphorus mirabilis, the light-bearing nutrient, was probably discovered by an Arabian alchemist named Alchid Bechil, although its discovery is usually attributed somewhat later to Henning Brandt.[1] Since the early work of Justus Liebig (1803–1873), agricultural researchers have tried to tackle the mystery of phosphorus (P) availability to plants. Phosphorus research intuitively reminds one of Heisenberg's theory: "The more you see, the less you know." Despite the substantial amount of information available about the transformation products of P fertilizer within temperate[2] as well as tropical soils[3] and numerous attempts to conduct elaborate soil test methods to predict P availability to crops and algae,[2] today the management of phosphate, a finite nonrenewable resource, is far from being in accordance with the principles of sustainability.[4]

This is partly because of the still prevailing misconception that soil is a homogenous, static entity. Phosphorus is regarded traditionally as immobile, but its chemical reactivity with environmental factors, acting at different spatial scales and over different time periods, results in the formation of P species, which not only differ in their plant availability but also in their spatial distribution throughout the field. This spatial speciation[5] is the key for a new approach to assess soil analysis methods.

Chemical Speciation of Phosphorus

With the exception of small molecular organic P fractions, plants can only utilize P in its soluble form as orthophosphate.[6] Unlike carbon (C) and nitrogen (N), which can be added to the soil system from the atmosphere, the P status of natural systems is essentially controlled by the occurrence of primary apatite minerals.[7,8] Consequently, the P enrichment of soils depends directly on P inputs by mineral fertilizers and manure. Although the fate of P when applied to soil remains something of an enigma, it is widely accepted that more than 80% is immobilized by the soil because of precipitation and sorption processes,[3] whereby the limiting step to furnish crop requirements is the dissolution of initial reaction products during the cropping season.[9]

For a more practical approach, the soil P continuum is thought of as three functional pools: a readily available pool, a reversibly available pool, and a sparingly available P pool,[10] whereby the readily available P pool is related to the so-called "intensity factor," which is a measure of the gradient in the electrochemical potential of the phosphate ions across the adsorbing surfaces of plant roots and, in its simplest form, can be regarded as the P concentration in the soil solution.[11] This pool represents P that is readily accessed by plant roots. The reversibly available P denotes the soil P reserve that can be converted into soluble (readily available) P, by either living organisms or by weathering during the growth season. This pool relates to the so-called "quantity factor" or "richness factor."[12] Whereas the sparingly available P is not available on a short time scale such as one or more crop cycles, a small fraction of this pool may become available during long-term soil transformation.

Mobility of Phosphorus in Soils

As a tetrahedral oxyanion, phosphate has a very low solubility in soils and in general does not move with solvent fluxes, apart from small distance diffusion.[6] In this sense, P is regarded as an "immobile" nutrient. Thus the physical movement of P is restricted to the movement of P associated with soil particles and large-molecular- weight organic matter (particulated P) by either bioturbation, soil tillage activities, or soil erosion during flow events.[13,14] This behavior of soil P was recognized long ago by European geographers and since then has been used in archaeology to trace back ancient settlements.[15] Conway[16] introduced the use of total P distribution patterns for the analysis of small-scale occupation deposits. For example, one building showed evidence of having been demolished and partially reincorporated into the courtyard of a subsequent structure. The floor area of the remnant original structure was protected by a layer of small stones and contained high levels of P. That portion of the floor, which was subsequently converted into a courtyard, unprotected by stones, had less total P, having lost it by exposure and erosion. In an agricultural context, this translates to the following conclusions. First, as far as its total amounts are concerned, P applied with fertilizers may not move from the place it is applied to, and, second, keeping in mind that most soils are naturally poor in P, nearly all the spatial distribution of total P in agricultural soils should be more or less random, reflecting only the spatial sum of distribution faults of past anthropogenic activities (e.g., fertilization, animal husbandry). Spatial relationships may only have developed under the influence of erosion processes.[16]

Spatial Speciation of Phosphorus

During the last 10 years, the study of spatial variation of soil fertility parameters has expanded considerably, but studies that investigated the spatial distribution of soil P generally focused solely on the distribution of so-called plant-available P. The results of these studies showed that plant-available P does not fluctuate randomly, but shows distribution patterns with well-defined lag ranges, where the ranges differ, depending on the sampling procedure and scale of investigation (Table 1). For an introduction into geostatistical terminology one may consult Gassner and Schnug[17] in this issue.

Additionally, it is generally accepted that the distribution of available P does not necessarily resemble the distribution of total P.[20] It appears that the speciation of soil P is dependent on site-specific factors and, as such, is a spatial process.[21–23] In this sense, "spatial speciation" is defined as the chemical reactivity of a nutrient with site- specific environmental factors, and the subsequent formation of geochemical species that display different spatial dependencies (Figure 1).[5]

TABLE 1 Parameters of Autocorrelation for Soil P Extracted by Different P Methods in Selected Investigations

Reference	Soil Texture	Sampling Design (m)	Extraction Method	Variogram Model	Range of Autocorrelation (m)
Trangmar[a]	—	1.5 × 1.5 m	Truog	Spherical	5.6
Boyer et al.[a]	μL	2 m transects	Bray I	Spherical	37
Doberman et al.[a]	T	5-m triangular grid	Olsen	Nested	48
Simard et al.[a]	μT—T	12 × 15 triangular	Mehlich III	Exponential	139
Karlen et al.[a]	—	15 × 15 grid	Bray I	Spherical	70
Webster and McBratney[a]	—	16 × 16 random	Morgan	Spherical	241
Gupta et al.[a]	sT	20 × 20 grid	Mehlich I	Exponential	29
Romanokov[a]	μL	20 × 20 grid	0.2 N HCI	Spherical	50–60
Nolin et al.[a]	μT—T	30 × 30 grid	Mehlich III	Exponential	39
Haneklaus et al.[a]	lS—sL	30 × 30 grid	CAL	Spherical	153
Gassner et al.[19]	μL	30 × 30 grid	AAC–EDTA	Spherical	110
Haneklaus et al.[a]	sL—μL	50 × 50 grid	DL	Spherical	115
Haneklaus et al.[a]	lS—sL	50 × 50 grid	DL	Spherical	131
Gassner et al.[5]	lS—sL	50 × 50 grid	CAL	Spherical	253
Chien et al.[a]	sL—μL	250 × 250 grid	Mehlich III	Spherical	580
Yost et al.[a]	—	1–2 km transect	Olsen	Exponential	1000

[a] Cited in Grassner[18]

Insights into the environmental processes that result in the spatial speciation of soil P and, as such, govern the behavior of applied fertilizer are necessary to predict the interconversion and equilibrium distribution of different soil P pools under specific conditions such as geomorphology, field management, and soil types.

The ratio between the readily available P pool (intensity factor) and the reversible available P pool (quantity factor), reflecting the ease of P withdrawal by the plant, is expressed as the "capacity factor."[12] The capacity factor, the P adsorption capacity of the soil, is predominantly dependent on the negative surface charges as well as the specific surface area of soil particles, and is mainly a function of the amount and nature of available adsorption sites in the soil.[24]

For a homogeneous soil, most of the potential adsorption sites will have a similar bonding energy for P. Thus the speciation of P will mainly be a function of the soil pH and the aging of initial reaction products of freshly applied fertilizer P. In this case, the spatial speciation is assumed to be low (Figure 2). For a heterogeneous soil, the nature of the chemical reactions between P and particle surface is more diverse and the chemical speciation will result in a differentiated distribution of P among soil components. In this case, the spatial speciation is assumed to be more pronounced as the distribution of different soil components with a particular bonding energy for P will result in a spatial differentiation of P species.

Apart from the site-specific adsorption capacity of a field, site-specific anthropogenic and environmental factors such as management, biological, and physical factors can result in a spatial speciation of P. Whereas the total P—and, therefore, the sparingly available P pool—is relatively inert to short-term environmental impacts, such as recent fertilizer placement or changes in tillage practices,[5,25] the reversibly available P pool is most affected by management.[26,27] Although the readily available P pool is directly influenced by crop selection and management, McDowell and Sharpley[28] showed that under similar management conditions, the specific adsorption capacity of the soil (capacity factor) controlled the rate of P release into solution.

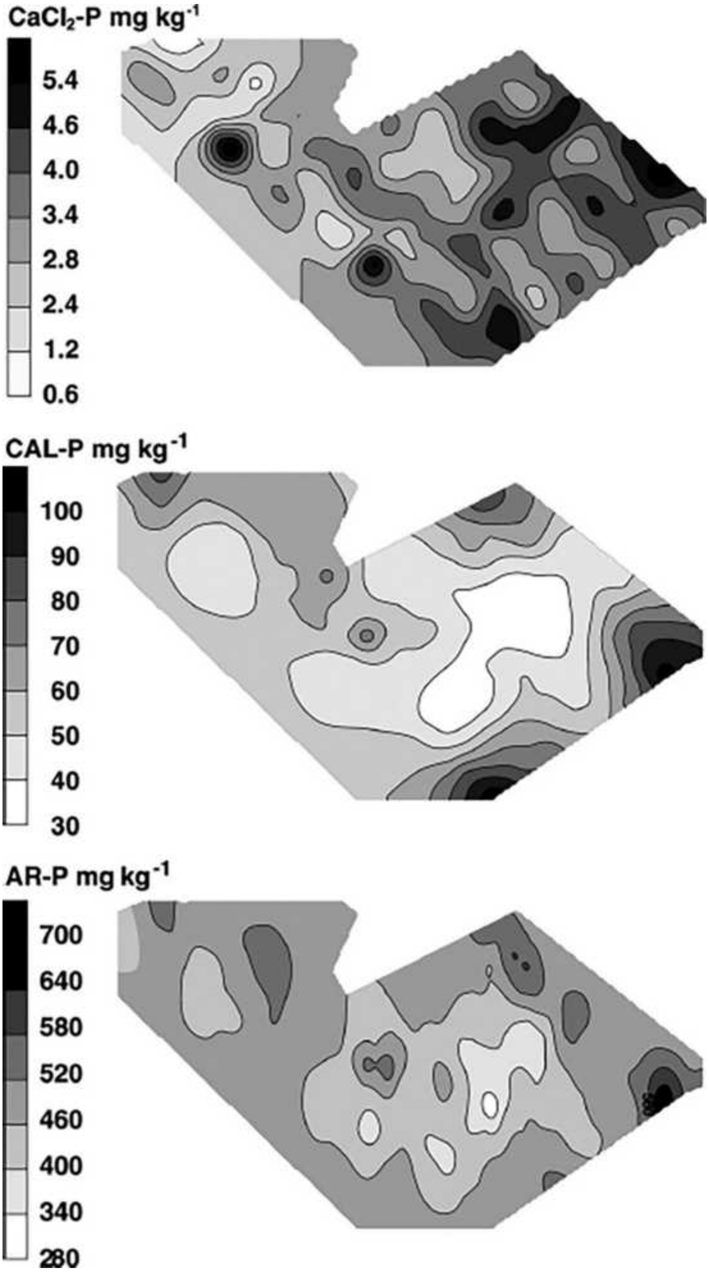


FIGURE 1 Spatial distribution of three different P fractions: $CaCl_2$ extractable P (upper map), Ca-lactate extractable P (middle map), and Aqua Regia extractable P (bottom map), at Kassow (E12°06 , N53°10), Northern Germany.

In a study investigating the spatial speciation of P at three different study sites, which differed in climate, parent material, topography, P fertilizer regime, and land management, the main environmental factors that controlled the spatial distribution of individual P pools were: soil texture for the readily available P; degree of aging of fresh soil P fractions, precipitated from application of soluble fertilizer P, for the reversibly available P; and primary P minerals and geomorphology for the sparingly available P.[18]

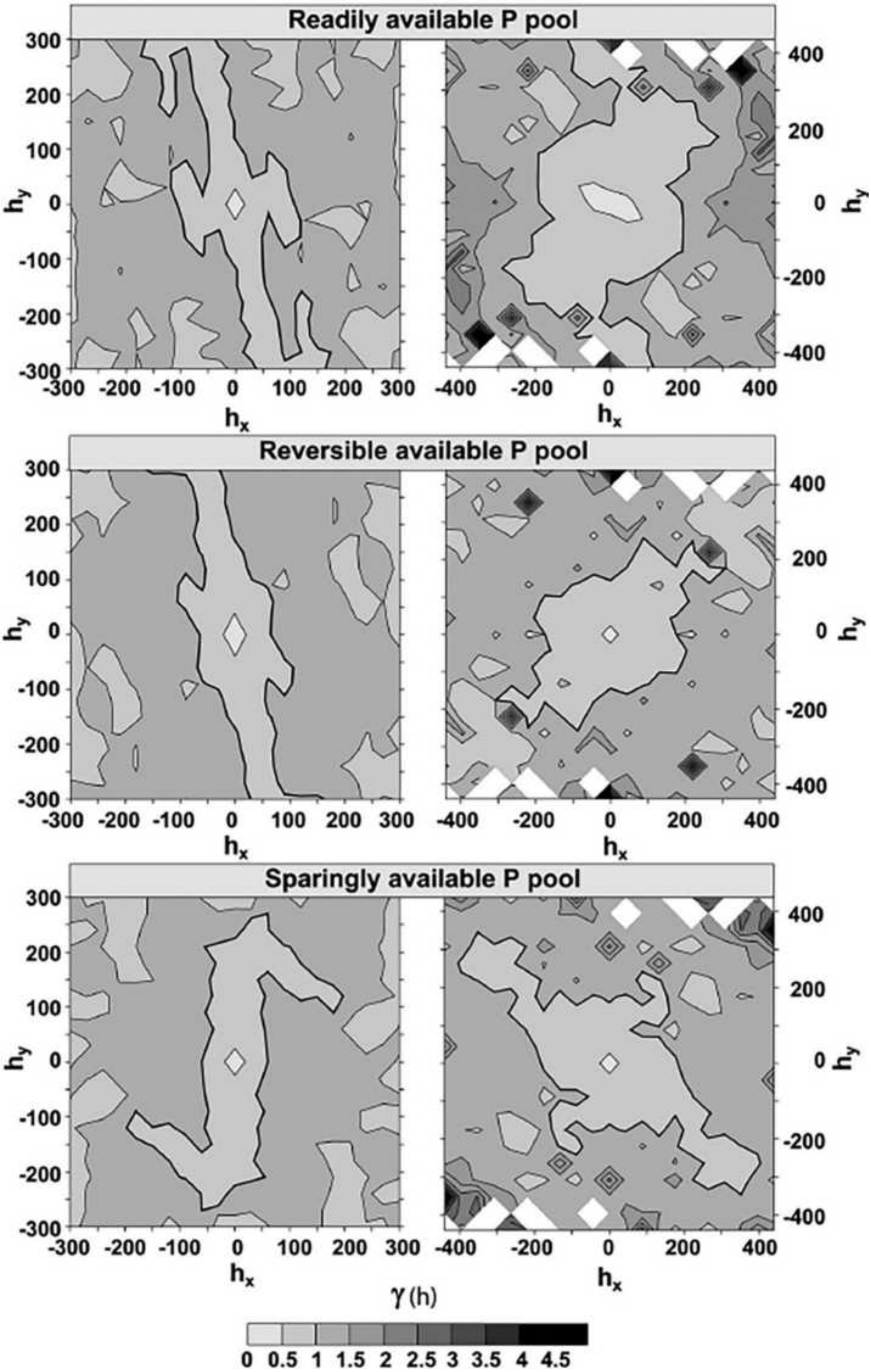


FIGURE 2 Variogram maps showing the spatial distribution of the semivariance of different P pools in a loamy sand (right) and a loamy clay (left). The semivariance is a measure of the average degree of dissimilarity between two data points.

Source: Gassner and Schnug.[17]

Conclusions

Phosphorus is present in soils not as definite and easily separated species, but as a continuum of compounds of different compositions and plant availability, in equilibrium with each other. This heterogeneous equilibrium is constantly disturbed by the uptake of the growing plant and by physical, chemical, or biochemical changes in the soil as well as fertilizer input. Within fields and across short distances, these factors can vary significantly in well- defined patterns, resulting in the spatial speciation of P. Consequently, soil P tests, evaluating the P status of a field, have to be adjusted to the site-specific soil characteristics. Furthermore, the interpretation of these results for subsequent management practices have to consider site-specific factors such as topography and past management history. Finally, the concept of spatial speciation should be applied to other nutrients.

References

1. Shapiro, J. Introductory lecture at the international symposium ‘phosphorus in freshwater ecosystems,’ Uppsala, Sweden, October 1985. Hydrobiologia **1985**, *170,* 9–17.
2. Frossard, E.; Condron, L.M.; Oberson, A.; Sinaj, S.; Fardeau, J.C. Processes governing phosphorus availability in temperate soils. J. Environ. Qual. **2000**, *29,* 15–23.
3. Buehler, S.; Oberson, A.; Rao, I.M.; Friesen, D.K.; Frossard, E. Sequential phosphorus extraction of a P-33-labeled oxisol under contrasting agricultural systems. Soil Sci. Soc. Am. J. **2002**, *66* (3), 868–877.
4. Steen, I. Phosphorus availability in the 21st century—management of a non-renewable resource. Phosphorus Potassium 1998, 217, 25–31. British Sulphur Publishing.
5. Gassner, A.; Fleckenstein, J.; Haneklaus, S.; Schnug, E. Spatial speciation—A new approach to assess soil analysis methods. Commun. Soil Sci. Plant Anal. **2002**, *33* (15–18), 3347–3357.
6. Barber, S.A. Soil-plant interactions in the phosphorus nutrition of plants. In *The Role of Phosphorus in Agriculture;* Khasawneh, F.E., Sample, E.C., Kamprath, E.J., Eds.; ASA CSSA and SSSA: Madison, WI, 1980; 591–613.
7. Walker, T.W.; Syers, J.K. The fate of phosphorus during pedogenesis. Geoderma **1976**, *15,* 1–19.
8. Bowman, R.A.; Rodriguez, J.B.; Self, J.R. Comparison of methods to estimate occluded and resistant soil phosphorus. Soil Sci. Soc. Am. J. **1998**, *62,* 338–342.
9. Schnug, E.; Rogasik, J.; Haneklaus, S. Die ausnutzung von phosphor aus düngemitteln unter besonderer berücksichtigung des ökologischen landbaus—the utilization of fertiliser with special view of organic farming. Landbauforsch. Voelken- rode (FAL Agric. Res.) **2003**, *53* (1), 1–11.
10. Guo, F.; Yost, R.S. Partitioning soil phosphorus into three discrete pools of differing availability. Soil Sci. **1998**, *163* (10), 822–833.
11. Olsen, S.R.; Khasawneh, F.E. Use and limitation of physical–chemical criteria for assessing the status of phosphorus in soils. In *The Role of Phosphorus in Agriculture;* Khasawneh, F.E., Sample, E.C., Kamprath, E.J., Eds.; ASA CSSA and SSSA: Madison, WI, 1980; 361–410.
12. *Williams, D.E.G. The Intensity and Quality Aspects of Soil Phosphate Status and Laboratory Extraction Values;* Anales de Edafología y Agrobiología, 1966.
13. Catt, J.A.; Johnston, A.E.; Quinton, J.N. Phosphate losses in the woburn erosion reference experiment. In *Phosphorus Loss from Soil to Water;* Tunney, H., Carton, O.T., Brookes, P.C., Johnson, A.E., Eds.; CAB International, 1997; 374–377.
14. Sharpley, A.; Foy, B.;Withers, P. Practical and innovative measures for the control of agricultural phosphorus losses to water: an overview. J. Environ. Qual. **2000**, *29,* 1–9.
15. Arrhenius, O. Die phosphatfrage. Z. Pflanzenernahr. Dung. Bodenkd. **1929**, *14* (A), 185–194.
16. Conway, J.S. An investigation of soil phosphorus distribution within occupation deposits from a Romano-British hut group. J. Archaeol. Sci. **1983**, *10,* 117–128.

17. Gassner, A.; Schnug, E. Geostatistics for soil science. In *Encyclopedia of Soil Science;* Lal, R., Ed.; Marcel Dekker: New York, 2003, *in press.*
18. Gassner, A. Factors controlling the spatial speciation of phosphorous in agricultural soils. Landbauforsch. Völkenrode (Braunschweig) **2003**, *244.*
19. Gassner, A.; Habib, L.; Haneklaus, S.; Schnug, E. Significance of the spatial speciation of phosphorous in agricultural soils for the interpretation of variability. Landbauforsch. Volkenrode (FAL Agric. Res.) **2003**, *53* (1), 19– 25.
20. Kamprath, E.J.; Watson, M.E. Conventional soil and tissue tests for assessing the phosphorus status of soils. In *The Role of Phosphorus in Agriculture;* Khasawneh, F.E., Sample, E.C., Kamprath, E.J., Eds.; ASA CSSA and SSSA: Madison, WI, 1980; 433– 469.
21. *Nowack, K.-H. Phosphorversorgung Biologisch Bewirtschafteter Acker und Möglichkeiten der Bioindikation;* University of Göttingen: Göttingen, 1990.
22. Strohbach, B. Gesetzmäßigkeiten der arealen verteilung des gesamtphosphorgehaltes auf landwirtschaftlich genutzten flachen. Arch. Acker-Pflanzenbau Bodenkd. (Berlin) **1986**, *30* (9), 573– 580.
23. *Jentsch, U. Kapazitäts-, Quantitäts-, Intensitäts- und Kinetikparameter des Phosphats in verschiedenen Bolden der DDR;* Humbold-Univ. Berlin: Berlin, 1986.
24. Schwertmann, U.; Taylor, R.M. Iron oxides. In *Minerals in Soil Environments;* Dixon, J.B., Weed, S.B., Eds.; SSSA: Madison, WI, 1989; 379–427.
25. Beckett, P.H.T.; Webster, R. Soil variability: a review. Soils Fert. 1971, *34,* 1– 15.
26. Schepers, J.S.; Schlemmer, M.R.; Ferguson, R.B. Site- specific considerations for managing phosphorus. J. Environ. Qual. **2000**, *29,* 125– 130.
27. Mallarino, A.P. Spatial variability patterns of phosphorus and potassium in no-tilled soils for two sampling scales. Soil Sci. Soc. Am. J. **1996**, 60, 1473– 1481.
28. McDowell, R.; Sharpley, A. Availability of residual phosphorus in high phosphorus soils. Commun. Soil Sci. Plant Anal. **2002**, *33* (7– 8), 1235– 1246.

12

Erosion and Global Change

Taolin Zhang and
Xingxiang Wang

Introduction

Soil erosion is the movement and transport of soil by various agents, which lead to a loss of soil, particularly by water, wind, and mass movement. It is the most serious and widespread form of soil degradation.[1] Global change is becoming a great concern since it was put forward by ICSU in 1986. Global change, as a result of physical, chemical, and biological processes in the Earth's system, particularly of interactions between humans and the environment, refers to changes in global environmental components that are vital to human survival, such as climate, land, water and air in the form of greenhouse effect, diminution of forest, depletion of biodiversity, degradation of land (desertification), and deficiency in water resources. Soil erosion is an important form of global change, and it severely threatens the quality of the soil, land, air and water resources, and hence agricultural development and human life. Soil erosion and global changes are related by bi-directional interactions,[2] and global change may exacerbate soil erosion problems.[3]

Contribution of Soil Erosion to Global Change

Extent of Soil Erosion in China and in the World

China is one of the nations that are most severely affected by soil erosion in the world. The land area affected by accelerated soil erosion is estimated at 367 Mha in 1980s, and 356 Mha in 1990s in China, including 165 Mha by water erosion and 191 Mha by wind erosion (Table 1). The annual soil loss due to soil erosion is estimated at 5000 million tons, including 2400 million tons from watershed of Yangtze River, and 1600 million tons from watershed of Yellow River.[4] The region most severely affected by soil erosion in China is the watershed of Yellow River, with area of 45 Mha and erosion modulus of

TABLE 1 Area Affected by Accelerated Soil Erosion in China

Region	Water Erosion (Mha)	Wind Erosion (Mha)
Slight	83	79
Moderate	55	25
Severe	18	25
Very severe	6	27
Extreme	3	35

Source: Bulletin of soil and water loss in China in 2000.[4]

6821 $\text{tkm}^{-2}\ \text{yr}^{-1}$, especially in the Loess plateau with more than 10,000 $\text{t/ km}^{-2}\ \text{yr}^{-1}$ erosion modulus[5] (Figure 1). Fortunately, about 24% of erosion area in Loess Plateau of China has been controlled since 1980s.[6] On a global scale, total land area affected by water erosion is 1094 Mha, of which 751 Mha is severely affected, and that by wind erosion is 549 Mha, of which 296 Mha is severely affected[1,7,8] (Table 2). The annual soil loss due to soil accelerated erosion is estimated at 75 billion Mg.[9]

FIGURE 1 Landscape of severe soil erosion in the Loess Plateau of China (the soil has been blown or washed away and only naked rocks remain).

TABLE 2 Global Area Affected by Accelerated Soil Erosion

	Area Affected by Severe Erosion (Moderate + Level)(Mha)	
Region	Water Erosion	Wind Erosion
Africa	169	98
Asia	317	90
Europe	93	39
Central America	45	5
North America	46	32
Oceania	4	16
South America	77	16
World	751	296

Source: Oldeman,[1] Scherr,[7] and Lal.[8]

Soil Erosion Degrades Soil Quality, Leads to Loss of Agricultural Lands and Threatens to Global Food Security

Soil is one of the most important natural resources and a major factor in global food production. Soil erosion is widely considered as the most serious form of soil degradation, posing a significant threat to world's food production capacity and global food security.[10]

Over the past 40–50 years, the arable land has been lost due to soil erosion at a rate of 0.6 Mha per year in China and nearly one-third of the world's arable land has been lost by erosion and continues to be lost at a rate of more than 10 Mha per year at a global scale. The average rates of soil loss have been estimated at 17 t ha^{-1} yr^{-1} in the United States and Europe, and 30–40 t ha^{-1} yr^{-1} in Asia, Africa and Southern America, mainly due to inadequate agricultural land use.[9]

Soil erosion not only reduces area of arable land, but also leads to depletion of nutrient, reduction in soil organic carbon and a negative alteration of the soil physical properties in terms of nutrient and water-holding capacity, soil biodiversity, and soils become more vulnerable to diseases.[11] The negative effect of erosion on crop productivity can be assessed using field runoff plots or paired watersheds, and that of future erosion using topsoil removal and addition technique. It has been extensively researched since the turn of the 20th century, although the on-site impacts of soil erosion on productivity are easily minimized through additional external agricultural input and adoption of improved agricultural technology. The actual loss may depend on weather conditions during the growing season, farming systems, soil management, and soil ameliorative input. Partial global production loss estimates suggest that, each year farmers lose about 2.2 million Mg maize, 0.07 million Mg millet, 1.1 million Mg potatoes, 0.1 million Mg soybeans and 1.1 million Mg wheat. Erosion-caused losses of food production are most severe in Asia, Sub-Saharan Africa, and elsewhere in the tropics rather than in other regions.[10] Furthermore, as a driver of land-use change, soil erosion will induce abandonment of arable land due to declining productivity, and reduction in crop yield.[12]

Soil Erosion Deteriorates Global Environment and Influences Global Change

Soil erosion deteriorates the global environment through flood disasters, environmental pollution, desertification, etc. It aggravates flood disasters through raising riverbeds, silting lakes, and shrinking water bodies. Flood disasters increased from once per 20 years 1300 years ago to once per 1.6 years in the 1990s in the Yangtze River Basin due to soil erosion, and soil erosion played an important role in the special phenomena of the lower peak charge but highest water level in the 1998 flood, which caused at least economic losses of USD 30 billion estimated by the Washington-based World watch Institute.[13] Riverbed in lower reaches of Yellow River was raised by 8–10 cm each year due to soil erosion.[4] Soil erosion is also a cause of factor of nonpoint source pollution and siltation associated with erosion in rivers and lakes was identified as the second leading cause of water quality impairment in the U.S.[14] With an enrichment of nutrients, pesticides, salts, trace elements, pathogens, and toxic substances, soil erosion leads to contamination of surface and groundwater bodies and diminishes water quality and aquatic ecosystems stability. Meanwhile, soil erosion can also have an adverse effect on transportation, communication, and human health through increasing dust in the air. In addition, soil erosion can also be a driving force to soil desertification.

Soil erosion influences global climatic change by changing the carbon, nitrogen, and water cycles. The global soil carbon pool is about 2500 gigatons and it is 3.3 times the size of the atmospheric pool and 4.5 times the size of the biotic pool.[15] Soil erosion has a profound impact on both quality and quantity of organic matter in soils, and can be a significant factor in local carbon losses and redistribution. Small changes in the pools of carbon and nitrogen in the world soils could have large effects on atmospheric concentration of CO_2, CH_4, and NO_2. It was estimated that 4.0–6.0 Pg Cyr^{-1} was translocated by water erosion, including 2.8–4.2 Pg C redistributed over the landscape and transferred to depressional sites, 0.4–0.6 Pg C transported into the ocean and coastal ecosystems, and 0.8–1.2 Pg C emitted into the atmosphere at the global scale.[7]

Soil erosion may cause a shortage of freshwater through deteriorating soil physical structure, increasing surface runoff, and changing global water cycling. For example, 20–30 billion m^3 water was used to wash the sediments into sea to fall riverbeds in the lower reaches of Yellow River. More than 19.6 Mha arable land suffer from drought in China each year, and they are mostly located in the area severely eroded.[4]

Moreover, it may also directly lead to changes in microclimate, regional and/or zonal landscapes and ecosystems, and it may even alter the global climatic pattern. In addition, soil erosion may also lead to a change in land uses and land covers through degrading soil quality and land suitability.

Global Change May Accelerate Soil Erosion

The main factors influencing soil erosion are rainfall (amount, frequency, duration, and intensity), wind speed (direction, strength, and frequency of high-intensity events), land use and management, topography, and soils and their properties. Soil erosion forecasting is relevant at a range of time and space scale, from the field scale at which it affects the individual farmer, up to national or global scales where it can make an important contribution to planning decisions. Since the 1950s, significant advances have been made in predicting erosion risks, particularly with the development of modeling technology.[3,16–19] However, the response of soil erosion to global change is very complex temporally and spatially due to uncertainties in general circulation models (GCMs) and the complex relationships between soil erosion processes and climate, land use, and land cover.

Future changes in greenhouse gases concentration and climate will change the hydrologic cycle, and hence affect the soil–plant–water interactions, which in turn affect soil erosion. Climatic erosivity is a major determinant of soil erosion and sediment transport, and is directly affected by change in climate. Much of the increase in precipitation that has been observed worldwide has been in the form of heavy precipitation events, and climate models are predicting a continued increase in intense precipitation events during the 21st century.[20,21] The responses of models of LISEM, MEFIDIS, RUSLE, STREAM, KINESOR, SWAT, and WEEP to rainfall change indicate that soil erosion is likely to increase significantly under future climate change unless offsetting amelioration measures are taken.[19] In some regions, any increase in rainfall, weather amount, intensity, or frequency, which occurs as a result of climate change, may directly exacerbate erosion. On the contrary, in other regions a decrease in rainfall may be expected. In this case, a decline in soil moisture because of the enhanced evaporation will occur. Drought soil conditions and less vegetation can make soils more vulnerable to wind erosion. Moreover, soil erosion tends to be dominated by extreme events, which might occur only rarely. An increase in the possibility of extreme rainfall events in the future, which is suggested by several climatic change models, may accelerate soil erosion. While a large emphasis has been put on climate change and how it may impact soil erosion, other relevant global change factors may have a more profound influence on soil loss. Global change may influence soil erosion through changing temporal and spatial evolution of global soil cover, such as change in soil-forming processes and soil properties. The response of soil to the erosion process is complex, and is also influenced by soil properties such as texture, structural stability, organic matter content, clay mineralogy, and chemical constituents. All these factors are likely to change with change in climate and/or land use. Soil organic matter (SOM), which is a key element of the soil properties, is susceptible to change with changing climate. Higher temperature will usually increase rates of decomposition of organic matter, and decrease its content in soils. A decline in SOM level would cause a decrease in soil aggregate stability, lower infiltration rates, increase runoff, and hence increase the likelihood of erosion. Fortunately, the impacts of SOM on climate change and food security has been concerned and some strategy related to increase SOM has been suggested.[15]

The processes of land use and culture changes are also strongly impacted by the climatic changes. With climatic change and soil degradation, land use and land cover may be changed. A decline in forest cover due to deforestation and wildfire, especially tropical rainforests, and the conversion from forest or grassland to arable land in some areas under global change will accelerate soil erosion.

Sea level change is also an important item in global change. Recent research indicates coastal erosion is triggered by sea level rise,[22] for example, annual beach sediments loss was $1.67*10^4 m^3$ due to sea level rising in the southern Shandong sandy coast in China.[23]

Conclusions

Soil erosion is an important form of global change, and it severely threatens the quality of the soil, land, air and water resources, and hence agricultural development and human life. Soils erosion and global changes are related by bi-directional interaction, and soil erosion is likely to increase significantly under global change unless amelioration measures are taken. It is critical to take action against soil erosion under global change.

Acknowledgments

This project was supported by the knowledge innovation Program of the Chinese Academy of Chinese.

References

1. Oldeman, L.R. The global extent of soil degradation. In *Soil Resilience and Sustainable Land Use*; Greenland, D.J., Szabolcs, I., Eds.; CAB International: Wallingford, U.K., 1994; 99–118.
2. Mosier, A.R. Soil processes and global change. Biol. Fert. Soils 1998, ***27***, 221–229.
3. Ingram, J.; Lee, J.; Valentin, C. The GCTE soil erosion network: a multi-participatory research program. J. Soil Water Conserv. **1996**, *51* (5), 377–380.
4. Bulletin of soil and water loss in China in 2000 (in Chinese); http://www.ewater.net.cn/public/public/bulletin.asp.
5. Chen, H.S.; Shao, M.A. Soil erosion and its control measures in China. Science **2000**, *2*, 55–57, (in Chinese).
6. Hu, X.B.; Li, Z.B.; Hao, M.D., et al. Down-sacle analysis for water scarcity in response to soil-water conservation on Loess Plateau of China. Agriculture Ecosystem and Environment **2003**, *94,* 355–361.
7. Scherr, S. Soil degradation: a threat to developing country food security by 2020. In *IFRI Food, Agric. and Environment Discussion Paper 27;* Washington, DC, 1999; 63.
8. Lal, R. Soil erosion and the global carbon budget. Environment International **2003**, *29,* 437–450.
9. Pimentel, D.; Harvey, C.; Resosudarmo, P., et al. Environmental and economic costs of soil erosion and conservation benefits. Science **1995**, *267,* 1117–1123.
10. Biggelaar, C.; Lal, R.; Wiebe, K., et al. The global impact of soil erosion on productivity II: effects on crop yields and production over time. Advances in Agronomy **2004**, *81,* 49–95.
11. Lal, R.; Mokma, D.L.; Lowery, B. *Relation between soil quality and erosion*. Soil and Water Conservation Society: Ankeny, IA, 1999.
12. Bakker, M.M.; Govers, G.; Kosmas, C., et al. Soil erosion as a driver of land-use change. Agriculture ecosystem and environment **2005**, *105,* 467–481.
13. Abramovitz, J.N.; Dunn, R. Record year for weather- related disasters. http://www.worldwatch.org/press/news/1998/11/27.
14. US environmental protection agency, Office of water, national water quality inventory, 1994. report to congress; Office of Water: Washington, DC, 1995.
15. Lal, R. Soil carbon sequestration impacts on global climate change and food security. Science **2004**, *304* (11), 1623–1627.

16. Rose, C.W. Modeling erosion by water and wind. In *Methods for Assessment of Soil Degradation*; Lal, R., Blum, W.H., Valentine, C., Stewart, B.A., Eds.; CRC Press: Boca Raton, FL, 1998; 57–88.
17. Williams, J.; Nearing, M.; Nicks, A., et al. Using soil erosion models for global change studies. J. Soil Water Conserv. **1996**, *51* (5), 381–385.
18. Matternicht, G.; Gonzalez, S. FUERO: foundations of a fuzzy exploratory model for soil erosion hazard prediction. Environment Modelling and Software **2005**, *20,* 715–728.
19. Nearing, M.A.; Jetten, V.B.; Cerdan, C., et al. Modeling response of soil erosion and runoff to changes in precipitation and cover. Catena **2005**, *61,* 131–154.
20. IPCC, working group I, 2001, Climate Change: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the IPPC; Cambridge University Press, 2001.
21. IPCC, working group II, 2001, Climate Change Impacts, Adaptation, and Vulnerability. Contribution of Working Group II to the Third Assessment Report of the IPPC; Cambridge University Press, 2001.
22. Ren, M.E. Research advances in sea level research. Journal of Nanjing University **2000**, *36* (3), 269–278 (in Chinese).
23. Zhuang, Z.Y.; Ying, P.; Wu, J.Z., et al. Coastal erosion and its influence on southern Shandong sandy coast. Marine Geology and Qualternary Geology **2000**, *20* (3), 15–21 (in Chinese).

13

Erosion and Precipitation

Bofu Yu

Erosion is a natural process to detach soil and rock fragments for subsequent removal, or transportation, of these materials to areas of lower elevation on the surface of the earth. In the context of agriculture, the primary agents for erosion are water and wind. Climate, precipitation in particular, plays a critical role in determining where and when erosion occurs and the magnitude of erosion rate. Rainfall erosivity, i.e., the ability of rain to cause erosion, is largely a function of rain amount and peak intensity. The larger the rain amount, and/or the more intensive the rain rate, the greater the amount of water erosion would result. Low rainfall, dry soil surface, and poor ground cover are the necessary conditions for wind erosion to prevail.

Rainfall erosivity is a measure of the climatic influence on water erosion. When other variables such as topography and vegetation cover are held constant, the rate of erosion is directly related to the level of rainfall erosivity. A number of rainfall erosivity indices have been proposed so that the amount of soil eroded is linearly proportional to the rainfall erosivity index *ceteris paribus*. The most commonly used rainfall erosivity index is EI_{30}, where E is the total kinetic energy per unit area for a storm (MJ ha^{1}) and I_{30} is its peak 30 min intensity (mm hr^{-1}). Wischmeier and Smith[1] found that the combination of kinetic energy and peak intensity was most closely related to the observed amount of soil loss. The R-factor in the Universal Soil Loss Equation (USLE) and Revised USLE (RUSLE) is the mean annual sum of these EI_{30} values.[2,3] Other measures of rainfall erosivity worthy of note include the modified Fournier Index,[4,5] KE>1 index,[6] and the so-called Universal Index of Onchev.[7] Numerous other attempts have been made to search for a rainfall-based estimator of the observed amount of erosion that is superior to EI_{30}. Most of these studies have relied on restricted databases that have limited their applicability. Most of these other indices or estimators are highly correlated with each other and with EI_{30}.

Although the definition of EI_{30} is straightforward, its calculation requires long-term rainfall data at short time intervals (<30 min), which are not widely available for most parts of the world. To develop a better understanding of what is exactly involved in EI_{30}, it is helpful to examine how this index is calculated. I_{30} is the maximum intensity for any 30 min interval during a rainstorm, while the storm energy depends on how rainfall intensity varies during the event:

$$E = \int_{T}^{0} e(I)\, I\, dt \qquad (1)$$

where I is rainfall intensity, T is rain duration, and $e(I)$ is a function of rain intensity called the unit energy equation. The consensus is that the unit energy as a function of rain intensity assumes the following functional form:[3,8]

$$e(I) = e_{\max}\left(1 - \alpha e^{-I/I_0}\right) \qquad (2)$$

For RUSLE, the following was recommended: $e_{\max} = 0.29$ MJ ha^{-1} mm^{-1}; $\alpha = 0.72$; $I_o = 20$ mm hr^{-1}.[3] It can be shown from Equations 1 and 2 that the storm energy is bounded:

$$0.28e_{\max}P < E < e_{\max}P \tag{3}$$

where P is the total rain (mm). The theoretical upper and lower bounds are related to zero and infinite intensity, respectively. Analyzing 6 min rain data for a number of sites around Australia shows that the ratio of storm energy to $e_{\max}P$ ranges mostly from 0.5 to 0.8, and the ratio is slightly higher in tropical/ subtropical regions than in temperate regions (Table 1). It is also clear from Table 1 that the storm energy is always highly correlated with rain amount. Empirical study using hourly rainfall intensity data likewise indicated a strong relationship between daily rainfall and storm energy for a number of sites in the U.K.[9] Given that storm energy is primarily a function of rain total, it follows that rainfall erosivity, as defined in relation to USLE/ RUSLE, depends mainly on rain total and peak intensity and, to a much lesser extent, on rain duration.

For areas where long-term high-resolution rain data are unavailable, a simpler method to estimate rainfall erosivity in the context of USLE/RUSLE is to use the fairly consistent relationship between the mean annual rainfall and the R-factor:[10,11]

$$R - \text{factor} = 0.05(\text{MAR})^{1.6} \quad R^2 = 0.82 \tag{4}$$

where MAR is the mean annual rainfall (mm). The regression equation (Equation 4) is based on a combined database for 161 sites (132 sites in the United States and 29 sites in Australia[10,11]). MAR ranges from 67 to 2060 mm for these sites. The non-linear relationship suggests that a 1% change to MAR would lead to a 1.6% change to rainfall erosivity. This highly sensitive nature of rainfall erosivity to rainfall would have important implications for the impacts of climate change on soil erosion. Reasonably good relationships between the Modified Fournier Index and the R-factor have also been noted.[5,10] The difference between the two estimates, however, is small, and little is gained by using the Modified Fournier Index.[11] If we need to estimate the seasonal distribution of rainfall erosivity, daily rain data can be used, especially in areas with a marked wet season in winter. Monthly and annual rain totals are no longer adequate because summer rain with high peak intensity can lead to higher rainfall erosivity in the relatively drier months. Rainfall erosivity can be related to rain amount using a power function in the form:

$$EI_{30} = aP^{\beta} \tag{5}$$

The calibrated values of β for a number of sites around the world are summarized in Table 2. The β value mostly varies from 1.5 to 1.8 with higher values found largely at higher latitudes. Such relationships for daily erosivity are sufficient for determining the seasonal variation of rainfall erosivity for USLE/RUSLE. Simple sinusoidal functions to allow seasonal variation in the parameter a can be used to achieve unbiased estimates of monthly EI_{30} values.[16,21,23]

TABLE 1 Linear Relationship between Rain Amount (P) and Storm Energy (E) as in $E = \alpha e_{max}P$ for Selected Sites in Australia

Location	Climate	α	n	R2
Perth	Temperate,winter rain	0.521	2354	0.96
Melbourne	Temperate, uniform rain	0.530	1800	0.93
Brisbane	Subtropical summer rain	0.626	4088	0.96
Darwin	Tropical, summer	0.742	3701	0.98

Note: n = number of storms analyzed; R^2 = coefficient of determination, representing the fraction of the total variation in the observed E values that can be explained by rain amount.

TABLE 2 The Average Exponent and Its One Standard Deviation in the Power Function Relating Daily Rain (P) to Rainfall Erosivity (EI_{30}) as in $EI_{30} = aP^{\beta}$

Country	Latitude Range	No. of Sites	$\beta \pm 1$ SD	References
Finland	60°N–66°N	8	1.77 ± 0.06	[12]
Canada	49°N–53°N	12	1.75 ± 0.13	[13]
United States	31°N–43°N	11	1.81 ± 0.16	[14]
Italy	36°N–42°N	35	1.53 ± 0.19	[15]
Equatorial (Malaysia, Indonesia, Brazil)	4°N–10°S	4	1.64 ± 0.18	[16–18]
Australia (tropical region)	10°S–25°S	41	1.49 ± 0.28	[19]
South Africa	31°S–33°S	4	1.47 ± 0.17	[20]
Australia (temperate region)	28°S–35°S	33	1.49 ± 0.25	[21,22]

Climate change has become a topic of considerable interest in recent years. In the context of precipitation and erosion, a number of attempts have been made to identify secular variation in rainfall erosivity in the recent past and the likely change to rainfall erosivity using output from global climate models. Long-term (of the order of 100 years) rainfall intensity data at short time intervals (6–10 min) can be used to detect decadal variations in rainfall erosivity.[24,25] Investigation of this nature is usually spatially restricted as data of this kind are available for a limited number of sites around the world. Annual or monthly rainfall totals have also been used to infer the likely change to rainfall erosivity at the regional or national scale.[26–28] Recorded daily rainfall totals can also be used to increase the accuracy of the estimated rainfall erosivity.[29,30] A more promising approach is to combine downscaled output from global climate models with stochastic weather generators to calculate rainfall erosivity for different climate change scenarios.[31]

Rain total and peak rainfall intensity are also key precipitation variables for a physical description of water erosion processes.[32–34] Mass balance dictates that in an area of net erosion, the amount of soil loss, SL, is given by:

$$SL = Qc \tag{6}$$

where Q is runoff amount and c is the sediment concentration. In this context, the effects of rain on erosion manifest themselves in terms of the amount of surface runoff generated and the level of sediment concentration in the runoff water. With non-climatic variables held constant, the amount of runoff is largely determined by rainfall amount and, to a lesser extent, by the rainfall intensity. Sediment concentration is related to both rainfall intensity and runoff rate. Rainfall detachment is linearly related to rainfall intensity. Shear stress or stream power commonly used to quantify flow detachment is intrinsically related to the runoff rate. Thus, in this physical framework for soil erosion, rainfall intensity plays a direct role in rain detachment. Rain amount and intensity also play an indirect role in flow detachment and transport of eroded sediments by determining the magnitude of runoff amount and runoff rate.

Precipitation is important to water erosion because soil particles and aggregates are detached by raindrops and surface runoff. A lack of precipitation, on the other hand, leads to low moisture levels near the soil surface and thus renders the soil particularly susceptible to wind erosion. Wind speed, precipitation, and potential evaporation were used to develop indices of wind erosivity.[35–37] For given wind speed and potential evaporation, wind erosivity is inversely related to precipitation. Figure 1 shows schematic relationships between precipitation and vegetation cover, rainfall and wind erosivity, and predominant erosion processes. In high rainfall areas, the rate of actual erosion is not necessarily high in spite of high rainfall erosivity unless the usually good vegetation cover is removed and the soil surface is exposed. In arid and semiarid areas with low rainfall, the combined effects of poor ground cover and dry soil surface would render the land particularly vulnerable to wind erosion.

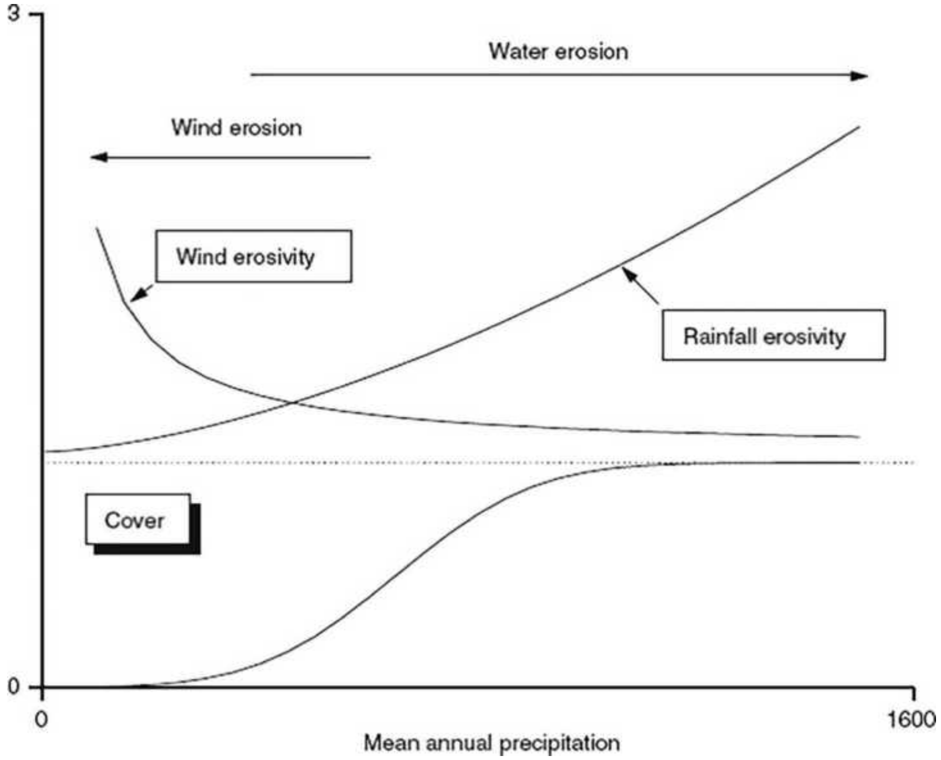


FIGURE 1 Schematic relationships between precipitation, vegetation cover, rainfall, and wind erosivity.

Precipitation is a key climatic variable that determines the type and magnitude of erosion. In the context of water erosion, rain amount and peak intensity are the most important variables in determining the erosion rate. For areas without high-resolution rainfall intensity data, the R-factor and its seasonal variation for USLE/RUSLE can be estimated from mean annual rainfall and daily rain amount. Absence of rain and concomitant dry soil surface and poor ground cover are the necessary conditions for wind to become the dominant erosion agent.

References

1. Wischmeier, W.H.; Smith, D.D. Rainfall energy and its relationship to soil loss. Trans. Am. Geophys. Union **1958**, *39,* 285–291.
2. Wischmeier, W.H.; Smith, D.D. *Predicting Rainfall Erosion Losses—A Guide to Conservation Planning;* Agriculture Handbook No. 537; U.S. Department of Agriculture. Washington DC, 1978.
3. Renard, K.G.; Foster, G.R.; Weesies, G.A.; McCool, D.K.; Yoder, D.C. coordinators. *Predicting Soil Erosion by Water: A Guide to Conservation Planning with the Revised Universal Soil Loss Equation (RUSLE);* Agriculture Handbook No. 703; U.S. Department of Agriculture, 1997.
4. Fournier, F. *Climatet erosion: la relation entre l'erosion du sol par l'eau et les precipitations atmospheriques;* Presses Universitaires de France: Paris, 1960.
5. Arnoldus, J.M.J. Methodology used to determine the maximum potential average annual soil loss due to sheet and rill erosion in Morocco. FAO Soils Bull. **1977**, *34,* 39–51.
6. Hudson, N.W. *Soil Conservation;* B.T. Batsford: London, 1976.

7. Onchev, N.G. Universal index for calculating rainfall erosivity. In *Soil Erosion and Conservation;* El-Swaify, S.A., Moldenhauer, W.C., Lo, A., Eds.; Soil and Water Conservation Society: Ankeny, Iowa, 1985; 424–423.
8. van Dijk, A.I.J.M.; Bruijnzeel, L.A.; Rosewell, C.J. Rainfall intensity-kinetic energy relationships: A critical literature appraisal. J. Hydrol. **2002**, *261,* 1–23.
9. Davison, P.; Hutchins, M.G.; Anthony, S.G.; Betson, M.; Johnson, C.; Lord, E.I. The relationship between potentially erosive storm energy and daily rainfall quantity in England and Wales. Sci. Tot. Environ. **2005**, *344,* 15–25.
10. Renard, K.G.; Freimund, J.R. Using monthly precipitation data to estimate the *R*-factor in the Revised USLE. J. Hydrol. **1994**, *157,* 287–306.
11. Yu, B.; Rosewell, C.J. A robust estimator of the R-factor for the Universal Soil Loss Equation. Trans. ASAE **1996**, *39,* 559–561.
12. Posch, M.; Rekolainen, S. Erosivity factor in the Universal Soil Loss Equation estimated from Finnish rainfall data. Agric. Sci. Finl. **1993**, 2, 271–279.
13. Bullock, P.R.; de Jong, E.; Kiss, J.J. An assessment of rainfall erosion potential in southern Saskatchewan from daily rainfall records. Can. Agric. Eng. **1989**, *32,* 17–24.
14. Richardson, C.W.; Foster, G.R.; Wright, D.A. Estimation of erosion index from daily rainfall amount. Trans. ASAE **1983**, 26, 153–157, 160.
15. Bagarello, V.; D'Asaro, F. Estimating single storm erosion index. Trans. ASAE **1994**, 37, 785–791.
16. Yu, B.; Hashim, G.M.; Eusof, Z. Estimating the R-factor using limited rainfall data: A case study from Peninsular Malaysia. J. Soil Water Conserv. **2001**, *56* (2), 101–105.
17. van der Linden, P. Soil erosion in central-Java (Indonesia): A comparative study of erosion rates obtained by erosion plots and catchment discharges. In *Rainfall Simulation, Runoff and Soil Erosion;* de Ploey, J., Ed.; Catena Supplement 4, 1983; 141–160.
18. Elsenbeer, H.; Cassel, D.K.; Tinner, W. A daily rainfall erosivity model for Western Amazonia. J. Soil Water Conserv. **1993**, *48,* 439–444.
19. Yu, B. Rainfall erosivity and its estimation for Australia's tropics. Aust. J. Soil Res. **1998**, *36,* 143–165.
20. van Breda Weaver, A. Rainfall erosivity in Ciskei: Its estimation and relationship with observed soil erosion. S. Afr. Geogr. **1989–1990**, *17* (1/2), 13–23.
21. Yu, B.; Rosewell, C.J. An assessment of a daily rainfall erosivity model for New South Wales. Aust. J. Soil Res. **1996**, *34,* 139–152.
22. Yu, B.; Rosewell, C.J. Rainfall erosivity estimation using daily rainfall amounts for South Australia. Aust. J. Soil Res. **1996**, *34,* 721–733.
23. Angulo-Martinez, M.; Begueria, S. Estimating rainfall erosivity from daily precipitation records: A comparison among methods using data from the Ebro Basin (NE Spain). J. Hydrol. **2009**, *379,* 111–121.
24. Yu, B.; Neil, D.T. Long-term variation of rainfall erosivity in Sydney. Weather Clim. **1995**, *15* (2), 57–66.
25. Verstraeten, G.; Poesen, J.; Demaree, G.; Salles, C. Long-term (105 years) variability in rain erosivity as derived from 10 min rainfall depth data for Ukkel (Brussels, Belgium): Implications for assessing soil erosion rates. J. Geophys. Res. [Atmos.] **2006**, *111* (D22); 11. Doi: 10.1029/2006JD007169.
26. Sauerborn, P.; Klein, A.; Botschek, J.; Skowronek, A. Future rainfall erosivity derived from large-scale climate models—Methods and scenarios for a humid region. Geoderma **1999**, *93* (3–4), 269–276.
27. Nearing, M.A. Potential changes in rainfall erosivity in the US with climate change during the 21st century. J. Soil Water Conserv. **2001**, *56* (3), 229–232.
28. de Luis, M.; Gonzalez-Hidalgo, J.C.; Longares, L.A. Is rainfall erosivity increasing in the Mediterranean Iberian Peninsula? Land Degrad. Dev. **2010**, *21* (2), 139–144.

29. D'Asaro, F.; D'Agostino, L.; Bagarello, V. Assessing changes in rainfall erosivity in Sicily during the twentieth century. Hydrol. Processes **2007**, *21* (21), 2862–2871.
30. Zhang, G.H.; Nearing, M.A.; Liu, B.Y. Potential effects of climate change on rainfall erosivity in the Yellow River basin of China. Trans. ASAE **2005**, *48* (2), 511–517.
31. Zhang, Y.G.; Nearing, M.A.; Zhang, X.-C.; Xie, Y.; Wei, H. Projected rainfall erosivity changes under climate change from multimodel and multiscenario projections in Northeast China. J. Hydrol. **2010**, *384* (1–2), 97–106.
32. Foster, G.R.; Flanagan, D.C.; Nearing, M.A.; Lane, L.J.; Risse, L.M.; Finkner, S.C. Hillslope erosion component (Ch. 11). In *USDA-Water Erosion Prediction Project: Hillslope Profile and Watershed Model Documentation;* Flanagan, D.C., Nearing, M.A., Eds.; NSERL Report No. 10; USDA-ARS Nat. Soil Erosion Research Lab.: West Lafayette, Ind., 1995.
33. Rose, C.W.; Coughlan, K.J.; Ciesiolka, C.A.A.; Fentie, B. Program GUEST (Griffith University Erosion System Template). In *A New Soil Conservation Methodology and Application to Cropping Systems in Tropical Steeplands;* Coughlan, K.J., Rose, C.W., Eds.; Technical Report, No. 40; Australian Centre for International Agricultural Research: Canberra, 1997; 34–58.
34. Morgan, R.P.C.; Quenton, J.N.; Smith, R.E.; Govers, G.; Poesen, J.W.A.; Auerswald, K.; Chisci, G.; Torri, D.; Styczen, M.E. The European soil erosion model (EUROSEM): A dynamic approach for predicting sediment transport from fields and small catchments. Earth Surf. Processes Land-forms **1998**, *23*, 527–544.
35. Chepil, W.S.; Siddoway, F.H.; Armbrush, D.V. Climatic factor for estimating wind erodibility of farm fields. J. Soil Water Conserv. **1962**, 9, 257–265.
36. FAO (Food and Agriculture Organization, United Nations). Soil erosion by wind and measuring for its control on agricultural lands. Development Paper No. 71. Rome, Italy, 1979.
37. Skidmore, E.L. Wind-erosion climatic erosivity. Clim. Change **1986**, 9, 195–208.

14

Erosion: History

Andrew S. Goudie

Introduction

Erosion is a process that has operated since the earth was created. Erosion of soil by water has likewise been a process that has been an entirely natural phenomenon ever since soils first appeared. However, during the last few millions of years that humans have inhabited the earth, their activities have caused spasms of accelerated soil erosion associated with land cover and land use changes. In classical times, it was noted that slopes in areas such as Greece, Turkey, and the Levant had been destabilized by deforestation and overgrazing. Undoubtedly, such actions as the deliberate setting of fire, adoption of pastoralism and agriculture, deforestation, urbanization, and use of machinery to move and disturb the soil have all contributed to accelerating rates. Appreciation of the nature, causes, and consequences of soil erosion has a long history, and during the 20th century, there were some notable studies of the phenomenon[1–3] stimulated by such events as the Dust Bowl in the United States and the menace of *donga* (gully) formation in Africa.[4]

However, it has never been easy to separate the role of climatic fluctuations in causing soil erosion from the role of human activities, and this has spawned long-running debates about the origin of phases of slope erosion, valley incision, and valley sedimentology in areas such as the Mediterranean basin[5,6] and the bottomlands of the western United States (see, for example, the discussion of *arroyos*[7]). Difficulties of determining how rates have been changed by human activities have also been bedeviled by an absence of direct long-term monitoring data. However, erosion leads to sedimentation, and so the study of rates of sediment accumulation in lakes, swamps, estuaries, reservoirs, and river floodplains provides a means of obtaining long-term data from which erosion rates can be inferred.

Deforestation

Deforestation[8] has been a crucial cause of accelerated soil erosion in many areas (Figure 1). Forests protect the underlying soil from the direct effects of rainfall, generating an environment in which erosion rates tend to be low. The canopy shortens the fall of raindrops, decreases their velocity, and thus reduces their kinetic energy. Most canopies reduce the erosion effects of rainfall. The presence of humus in forest soils[9] absorbs the impact of raindrops and gives them extremely high permeability. Thus forest soils have high infiltration capacities. Forest soils also transmit large quantities of water through their fabrics because they have many macropores produced by roots and their rich soil fauna. They are also

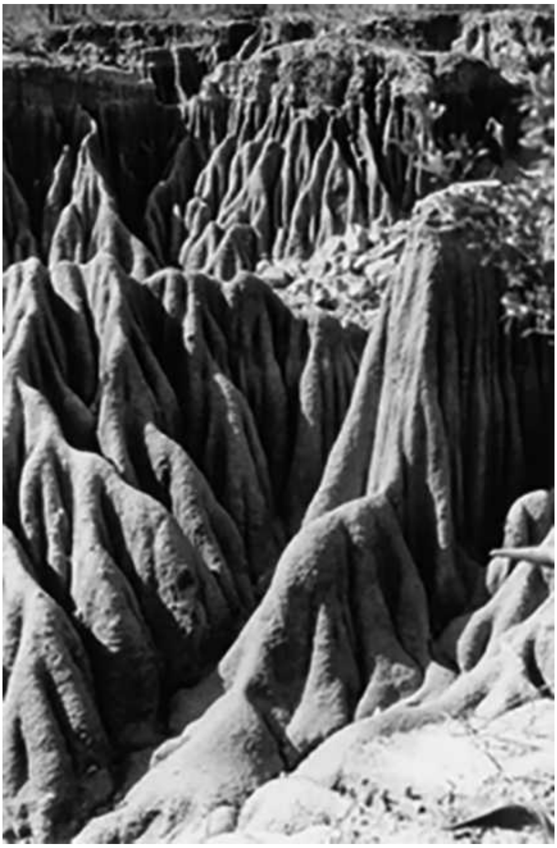


FIGURE 1 An erosional badland (donga) in Swaziland, Southern Africa. It may date back to deforestation in the Iron Age.

well aggregated, making them resistant to both wetting and water drop impact. This superior aggregation is a result of the presence of considerable organic material, which is an important cementing agent in the formation of large water-stable aggregates. Furthermore, earthworms also help to produce large aggregates.

It is therefore to be expected that with forest removal, rates of soil loss will rise and mass movements will increase in magnitude and frequency. The rates of erosion will be high if the ground is left bare; under crops, the increase will be less marked. Furthermore, the method of plowing, the time of planting, the nature of the crop, and the size of the fields will influence the severity of erosion.

Sedimentation Rates

A good example of using long-term sedimentation rates to infer long-term erosion rates is provided by a study[10] of the Kuk Swamp in Papua New Guinea. This identified low rates of erosion until 9000 BP, when, with the onset of the first phase of forest clearance, they increased from 0.15 to about 1.2 cm/1000 years. Rates remained relatively stable until the last few decades when, following European contact, the extension of anthropogenic grasslands, subsistence gardens, and coffee plantations produced a rate that is very markedly higher: 34 cm/1000 years.

A further long-term study of the response rates of erosion to land cover changes is provided by a study undertaken on the North Island of New Zealand.[11] During the last 2000 years of human settlement, catchments underwent a change from indigenous forest fern/scrub following Polynesian settlement (c. 560 years BP) and then a change to pasture following European settlement (AD 1878). Sedimentation rates under European pastoral land use were between 5 and 6 times the rates that occurred under fern/ scrub and 8–17 times the rates under indigenous forest. A broadly comparable study[12] looked at rates of infilling of an estuary fed by a steepland catchment in another part of New Zealand.

In pre-Polynesian times, rates of sedimentation were 0.1 mm year^{-1}, during Polynesian times, the rates climbed to 0.3 mm year^{-1}, while since European land clearance in the 1880s, the rates have shot up to 11 mm year^{-1}.

Major changes in rates of erosion have also taken place in Australia as a result of European settlement over the last two centuries.[13] Particularly important was the introduction of grazing stock.

There is increasing evidence to suggest that silty valley fills in Germany, France, and Britain, many of them dating back to the Bronze Age and the Iron Age, are the result of accelerated slope erosion produced by the activities of early farmers.[14] Indeed, in recent years, various studies have been undertaken with a view to assessing the importance of changes in sedimentation rate caused by humans at different times in the Holocene in Britain. Among the formative events that have been identified are initial land clearance by Mesolithic and Neolithic people; agricultural intensification and sedentarization in the late Bronze Age; the widespread adoption of the iron plow in the early Iron Age; settlement by the Vikings; and the introduction of sheep farming.

A core from Llangorse Lake (Brecon Beacons, Wales)[15] provides long-term data on changing rates of sedimentation. The 13-fold increase in rates after 5000 BP seems to have occurred rapidly and can be attributed to initial forest clearance. The second dramatic increase of more than 4-fold took place in the last 150 years and is a result of agricultural intensification.

The work on the lakes of the Peten region of northern Guatemala (Central America), an area of tropical lowland dry forest, is also instructive with respect to early agricultural colonization.[16] Combining archaeology and lake sediment stratigraphy, the diverse environmental consequences of the growth of Mayan civilization were reconstructed.

This showed a dramatic growth after 3000 years BP, but collapsed in the 9th century A.D. The hypotheses put forward to explain this collapse include warfare, disease, earthquakes, and soil degradation. The population has remained relatively low ever since, and after the first European contact (A.D. 1525), the region was virtually depopulated. The period of Mayan success saw a marked reduction in vegetation cover, an increase in lake sedimentation rates and in catchment soil erosion, an increased supply of inorganic silts and clays to the lakes, a pulse of phosphorus derived from human wastes, and a decrease in lacustrine productivity caused by high levels of turbidity.

Serious sedimentation of bays and estuaries has been caused by human activity on the eastern coast of the United States. Gottschalk[17] calculated that at the head of the Chesapeake Bay, 65 million m^3 of sediment was deposited between 1846 and 1938. The average depth of water over an area of 83 km^2 was reduced by 0.76 m. New land comprising 318 ha was added to the state of Maryland and, as Gottschalk remarked, "the Susquehanna River is repeating the history of the Tigris and Euphrates." Much of the material entrained by erosion on upper slopes as a result of agriculture in Maryland, however, was not translocated as far as the coast. Only about one-third of the eroded material left the river valley.[18] The remainder accumulated on floodplains as alluvium and colluvium at rates of up to 1.6 cm/year. Similarly, an intensive augering survey of floodplain soils in Wisconsin established that, since the development of agriculture, floodplain aggradation had proceeded at a rate of approximately 0.85 cm/year.[19] Channel and floodplain aggradation caused the flooding of low alluvial terraces to be more frequent, extensive, and deeper. The rate of sedimentation has since declined[20] because of less intensive land use and the institution of effective erosion control measures on farmland.[21]

Various attempts have also been made to establish rates of accelerated erosion on the plainlands of Russia.[22] It has been calculated that during the period 1696–1796, a total of 19.5×10^9 m^3 of soil was mobilized by sheet and rill erosion; for 1796–1887, it was $36.7\times10^9 m^3$, and for 1887–1980, it was $42.5\times10^9 m^3$. This increasing trend was a result of an increase in the area under cultivation and the assimilation of land more prone to erosion.

Conclusion

Accelerating rates of soil erosion are neither inevitable nor universal. In some parts of the world (e.g., New England or steep slopes in some of the Mediterranean countries[23]), the agricultural frontier has retreated and pressures on the soil have been reduced. Elsewhere, a whole range of soil conservation techniques has been introduced with some success. Nevertheless, accelerated soil erosion has a number of adverse consequences: loss of soil resource, sedimentation behind dams and in lakes, and a loss of water quality because of turbidity and other effects.

References

1. Marsh, G.P. *Man and Nature; Scribner:* New York, 1864.
2. Bennett, H.H. *Soil Erosion and Conservation*; McGraw Hill: 1938.
3. Shaler, N.S. *Man and the Earth; Duffield:* New York, 1912.
4. Jacks, G.V.; Whyte, R.O. *The Rape of the Earth: A World Survey of Soil Erosion*; Faber and Faber: London, 1939.
5. Vita-Finzi, C. *The Mediterranean Valleys*; Cambridge University Press: Cambridge, 1969.
6. Andel, T.H. van; Zangger, E.A.; Demitrack, A. Land use and soil erosion in prehistoric and historical Greece. J. Field Archaeol. **1990**, *17,* 379–396.
7. Cooke, R.U.; Reeves, R.W. *Arroyos and Environmental Change in the American South-West*; Clarendon Press: Oxford, 1976.
8. Williams, M. *Deforesting the Earth*; Chicago University Press: Chicago, 2002.
9. Trimble, S. The impact of organisms on overall erosion rates within catchments in temperate regions. In *Biogeomorphology;* Viles, H.A., Ed.; Basil Blackwell: Oxford, 1988; 83–142.
10. Hughes, R.J.; Sullivan, M.E.; Yok, D. Human-induced erosion in a highlands catchment in Papua New Guinea: The prehistoric and contemporary records. Z. Geomorphol., Suppl.bd **1991**, *83*, 227–239.
11. Page, M.J.; Trustrum, N.A. A late Holocene lake sediment record of the erosion response to land use change in a stee-pland catchment, New Zealand. Z. Geomorphol. **1997**, *41*, 369–392.
12. Sheffield, A.T.; Healy, T.R.; McClone, M.S. Infilling rates of a steepland catchment estuary, Whangamata, New Zealand. J. Coast. Res. **1995**, *11*, 1294–1308.
13. Olley, J.M.; Wasson, R.J. Changes in the flux of sediment in the upper murrumbidgee catchment, southeastern Australia, since European settlement. Hydrol. Process. **2003**, *17,* 3307–3320.
14. Bell, M.L. The effect of land-use and climate on valley sedimentation. In *Climatic Change in Late Prehistory*; Harding, A.F., Ed.; Edinburgh University Press: Edinburgh, 1982; 127–142.
15. Jones, R.; Benson, B.; in Evans, K.; Chambers, F.M. Human influence upon sedimentation in Llangorse lake wales. Earth Surf. Processes Landf. **1985**, *10*, 227–235.
16. Binford, M.W.; Brenner, M.; Whitmore, T.J.; Higuera-Grundy, A.; Deevey, E.S.; Leyden, B. Ecosystems, palaeo-ecology and human disturbances in subtropical and tropical America. Quat. Sci. Rev. **1987**, *6*, 115–128.
17. Gottschalk, L.C. Effects of soil erosion on navigation in Upper Chesapeake Bay. Geogr. Rev. **1945**, *35*, 219–238.
18. Costa, J.E. Effects of agriculture on erosion and sedimentation in the Piedmont province, Maryland. Bull. Geol. Soc. Am. **1975**, *86*, 1281–1286.

19. Happ, S.C. Effect of sedimentation on floods in the Kicka-poo Valley. Wisconsin. J. of Geo. **1944**, *52*, 53–68.
20. Trimble, S. Modern stream and valley sedimentation in the Driftless area, Wisconsin USA. 23rd Int. Geogr. Congr. **1976**, *1*, 228–231. Section.
21. Magilligan, F.J. 1985 Historical floodplain sedimentation, Galena River Basin, Wisconsin and Illinois. Ann. Assoc. Am. Geogr. **1985**, *75*, 583–594.
22. Sidorchuk, A.Y.; Golosov, V.N. Erosion and sedimentation on the Russian Plain: the history of erosion and sedimentation during the period of intensive agriculture. Hydrol. Process. **2003**, *17*, 3347–3358.
23. Grove, A.T.; Rackham, O. *The Nature of Mediterranean Europe*: An Ecological History; Yale University Press: New Haven, 2001.

15

Erosion by Wind: Global Hot Spots

Andrew Warren

Natural Erosion

The Sahara, the Central Asian Deserts, and the Chinese deserts are the sites of very high rates of "natural" erosion. Much more sediment leaves the Sahara in wind than in water. Contemporary outputs from the Sahara are of the order of 10^9 tons per year.[1] Dry Australia, the southwestern United States, and southern South America are less important, but still significant natural hot spots. The dust reaches far into the oceans and beyond, as from northern African to the United States, and from Australia to New Zealand (Figure 1). Marine and land sediments (loess and stabilized sand dunes) show that the hot spots were hotter in the Pleistocene, and that they got hotter as it progressed. In the central Sahara, Egypt, western

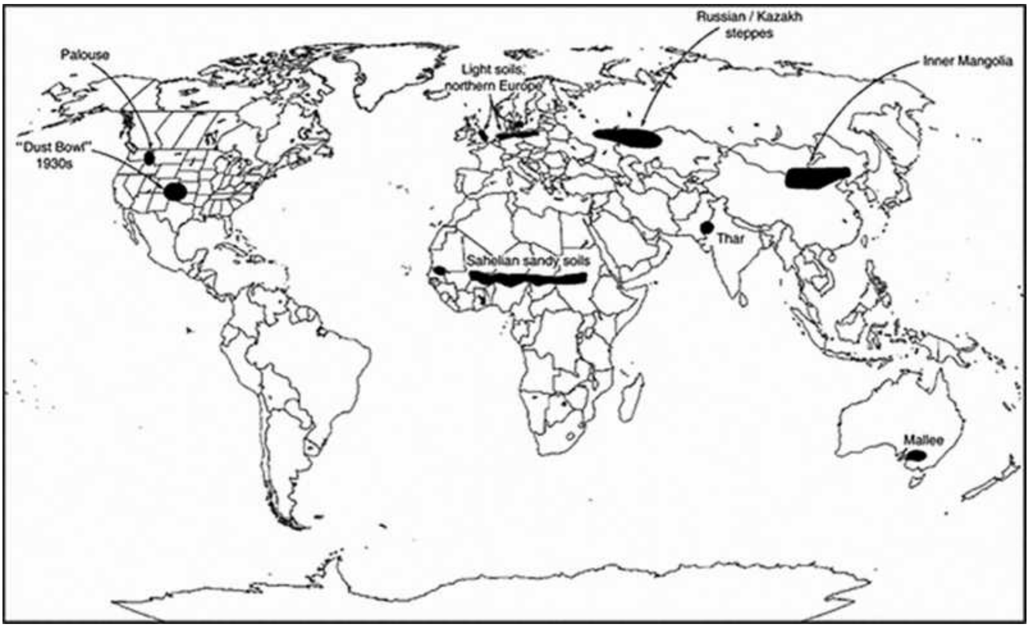


FIGURE 1 Global hot spots for wind erosion.

Argentina, eastern Iran, and parts of western China, great grooves were cut at those times by wind erosion, some in hard rock. There were also more hot spots, as in parts of northern Europe and central North America. Ice-core data from Greenland, the Tibetan Plateau, and the Antarctic corroborate the picture. The contemporary global source areas are all very dry. Explanation of the temporal change is more disputed.[2]

Seen at a finer scale, most "natural" erosion in these deserts occurs in soft sediments, particularly those in ancient, dry lake basins. The northern Chad Depression in the central Sahara and the Lake Eyre basin in central Australia are major natural hot spots.[3] In the late Pleistocene, the bare outwash plains round glaciers were the sources of loess, and of sand that formed dunes (now mostly stabilized) in northern Europe, Asia and northern Canada. People have little influence on rates of wind erosion in deserts. There are exceptions. Off-road vehicles create a minor hot spot in the Mojave of southern California; and wars and training for wars has created others in North Africa during the 1940s, and during the Gulf War.[2]

Induced Erosion

Identifying global hot spots of induced wind erosion is more problematic. Measurements of emissions are very few indeed, measurements of rates of erosion are even fewer, and measurements of economic impact are the rarest of all. This leaves some results from modeling, historical accounts built on disparate evidence, and informed guesses. Modeling and measurement suggest short-term rates of the order of 30–60 t ha^{-1} yr^{-1} on dry high plains sites, depending on winds, rainfall, and land use. Caesium-based measurements in the Sahel of West Africa, on sandy soils under fallow-cultivation cycling, and with mean annual rainfalls of about 550 mm give rates of up to 40 t ha^{-1} yr^{-1} over a 30 years period.[4]

Some historical accounts are well documented and compelling. The Dust Bowl of the American Great Plains in the 1930s was an undisputed hot spot.[5] Dust storms were more frequent and violent and many farmers went out of business, though more from drought than erosion. There were renewed hot spots here in the 1950s, 1960s, and 1970s, on the evidence of agricultural surveys and dust emissions. Probably more serious, though less well documented, at least in English, were the hot spots in the Soviet Union, especially northern Kazakhstan and neighboring parts of Russia, at about the same time[6] (although the earliest recorded paper in Russian on wind erosion control was written in 1768). More surprising, even to some experts, are histories of wind erosion in Western Europe. An even earlier paper on wind erosion, titled "A curious and exact relation of a sand cloud which hath lately overwhelmed a great tract of land in the county of Suffolk" appeared in the *Philosophical Transactions of the Royal Society of London* in 1669. In Sweden, no less a person than the great naturalist Linnaeus recorded destitution because of wind erosion in Skania in the 18th century. There is similar evidence in northern Germany, the Netherlands, Denmark, Poland and Hungary. Archaeology now shows that there were hot spots as far back as the Neolithic when ancient farms were buried in sand in southeastern England and the Netherlands. In all these areas, it was and is sandy soils, inherited from glacial outwash, that are the most susceptible. Much more recently, the period of desiccation in the Sahel of West Africa is said to have been a hot spot for wind erosion. Some evidence lies in dust-storm data,[1] and, as in the Great Plains of the 1930s, there was undoubtedly great hardship. However, here too, the exact role of wind erosion is debatable. Much of the dust may have come from quite small perimeters round dust-measuring stations, and the hardship was to do more with drought than erosion.

Informed guesses are a more dubious source. The literature is no great help, even with the wind erosion bibliography (listing works till 1995) and its 2700 odd entries.[7] It is undoubtedly biased toward work in English, toward areas where funds have been more available and to places where there has been a strong scientific tradition. For example, research in the USA, and the Great Plains in particular, where massive funds for research were released and maintained after the 1930s, overwhelms the bibliography, yet the aforementioned historical evidence suggests that the problem may have been more serious in the USSR. Indeed, Petrov's bibliography of relevant work in Russian (1768–1950) has 852 entries. Germany, with its long scientific tradition, apparently dominates research in Europe, yet it is known that the problem is as serious in neighboring counties. Southeastern Australia is also, by this evidence, a hot-spot,[8] yet in terms of damage to cropland, the Sahel of West Africa is probably more of one (the Sahel is only belatedly receiving scientific attention). Using the literature, Crosson[9] estimated that the Sahel was a major wind erosion hot spot, yet there are few data on which to base this conclusion.[4]

Only one systematic attempt at informed guessing on a worldwide scale has been made. This is GLASOD (The global assessment of soil degradation). It is most accessible in the *World Atlas of Desertification,* which is now in its second edition.[10] The assessment was for "polygons" of the order of size of southeast England, the province of Skania in Sweden, or a quarter of an average US State. Experts who knew each polygon were asked a set of carefully constructed questions about the extent and severity of soil degradation, including wind erosion. The result is the best guess we have of the severity of wind erosion globally. But it has serious flaws: in Africa and many other parts of the world, there were no measurements at the time of the questionnaire against which to compare the guesses; the distinction between induced at natural erosion is unclear; some experts have interpreted the questions differently from others; the identifications are generally of potential rather than actual erosion; the polygons are huge.

The expert opinion used in GLASOD is based on our present scientific knowledge about wind erosion. If, as a simple framework, we take the factors in the famous Wind Erosion Equation of Woodruff and Siddoway, we can follow it in predicting the hot spots.[11] Three factors are almost equally important. First is the Climatic (C) factor. Most induced wind erosion occurs in areas with semi-arid climates, as corroborated by maps of the C factor in the Sahel of West Africa, and in the United States. This is also true in the USSR, Australia, and northwestern Europe. However, wind erosion can and does take place in much wetter climates. In Wales and in the Pennine Hills of England peat is blown away after fires and in long dry spells. Erosion occurs in parts of England where the mean annual rainfall is of the order of 1500 mm. At roughly the same mean annual rainfall, there is also significant wind erosion on the coastal plain of South Carolina. Wind erosion is reported in the Austrian Tyrol, where rainfalls approach 2000 mm. Even in Iceland, where evapotranspiration is very low, there is some serious wind erosion. The C factor is also important in the temporal dimension: when there are dry years, wind erosion hot spots appear, as in the Dust Bowl and the Sahel.

The K, or soil, factor in the Wind Erosion Equation is the second important control, for wind erosion rarely affects fine-texture soils, although a few may be "pelletized" into aggregates of sand size, which may then be moved by the wind. Areas with agriculture on sandy soils are, worldwide, the common locations of hot spots, because they lack binding agents for aggregation. Peats, which also may be loose, are also eroded in Western Europe and Florida when they are drained, cleared, and cultivated. Light loess soils, as in the Palouse of Washington State, parts of the Great Plains and the Russian steppes, are other hot spots.

The last crucial factor is vegetation cover, and this is the most susceptible to management. The plowing, or harrowing of a field can raise considerable quantities of dust,[12] but it is clearance of vegetation in dry and windy seasons that is the more generally damaging. The reasons why farmers clear susceptible fields are exceedingly complex. Ignorance may sometimes produce a minute hot spot, but it is an inexperienced farmer indeed who does not realize that erosion follows clearance. In the Sahelian hot spot, for example, the laying of millet (*Pennisetum glaucum*) stalks to counter erosion is standard practice. Economic forces produce the main hot spots for induced erosion. In Europe, at least, the short-term economic impact of wind erosion does not provide a strong enough incentive to persuade farmers to manage it. Worster believed that the huge hot spot in the US Dust Bowl was mainly due to farmers being driven to cultivate by other economic forces, in this case the need to pay off loans from banks and from machinery salesmen. The massive Soviet hot spots are said to have been the result of yet another kind of economic force: centralized and ill-informed attempts to reach US standards of economic output. In our study of a village in Niger, economics takes yet another guise: wind erosion is the outcome of a gamble by those farmers who can afford to lose a crop, because they have other economic options.[13]

References

1. Middleton, N.J. Climatic controls on the frequency, magnitude and distribution of dust storms: examples from India/Pakistan, Mauritania and Mongolia. In *Palaeoclimatology and Palaeometeorology: Modern and Past Processes of Global Atmospheric Transport*; Leinen, M., Sarnthein, M., Eds.; Kluwer: Dordrecht, 1989; 97–132.
2. Livingstone, I.; Warren, A. *Aeolian Geomorphology: An Introduction;* Addison-Wesley Longman: Harlow, 1996.
3. McTainsh, G.H. Dust processes in Australia and West Africa: a comparison. Search **1985**, *16* (3–4), 104–106.
4. Chappell, A.; Warren, A.; Oliver, M.A.; Charlton, M. The utility of ^{137}Cs for measuring soil redistribution rates in South-West Niger. Geoderma **1998**, *81* (3–4), 313–338.
5. Worster, D. *Dust Bowl: The Southern High Plains in the 1930s;* Oxford University Press: Oxford, 1979.
6. Alayev, E.B.; Badenkov, Y.P.; Karavaeva, N.A. The Russian plain. In *The Earth as Transformed by Human Action;* Turner, B.L. II, Clark, B.K., Bob, W., Richards, J.F., Matthews, J.T., Meyer, W.B., Eds.; Cambridge University Press: Cambridge, 1990; 453–560.
7. Wind Erosion Bibliography: http://www.csrl.ars.usda.gov/wewc/apm.htm.
8. McTainsh, G.H., Boughton, W.C., Eds.; *Land Degradation Processes in Australia*; Longman Cheshire: Melbourne, 1993.
9. Crosson, P.R. Will erosion threaten agricultural productivity? Environment **1997**, *39* (8), 4–31.
10. United Nations Environment Programme [UNEP]. *World Atlas of Desertification,* 2nd Ed.; Arnold: London, 1997.
11. Woodruff, N.P.; Siddoway, F.H. A wind erosion equation. Proceedings of the Soil Science Society of America **1965**, *29* (5), 602–608.
12. Lee, J.A.; Tchakerian, V.T. Magnitude and frequency of bowing dust on the southern high plains of the United States, 1947–1989. Annals of the Association of American Geographers **1995**, *85* (4), 684–693.
13. Warren, A.B.; Simpon, P.J.; Osbahr, H. Soil erosion in the sahel of West Africa: a review of research issues and an assessment of new approaches. Global Environmental ChangeHuman and Policy Dimensions **2000**, *11* (1), 79–95.

16

Erosion by Wind: Principles

Larry D. Stetler

Erosion Processes

Erosion of fine-grained sediment and soil particles by wind is a natural geologic process occurring worldwide on a nearly continuous basis. Throughout the geological record evidence of wind processes eroding, transporting, and depositing sediment exists primarily as vast accumulations of eolian (wind-blown) sand—both modern sand dune fields and ancient eolian sandstone formations—and loess (wind-blown silt). Currently, about 30% of the total land surface of the planet consists of semi-arid and arid lands which give rise to most of the transported sediment and dust. Modern agriculture and industrial activities are also a significant contributor to transported sediment and dust due to induced or mechanical destabilization of the land surface.

Conceptually, mobilization of loose sediment and soil particles occurs during wind storms or wind "events," where the mean wind speed exceeds a critical threshold value based on some erodible fraction of the sediment. Individual wind events may be of short (less than one to a few hours) or long (multiple hours to days) duration. In terms of distance and total mass moved, a typical erosion event will transport the greatest mass a short distance and a significantly lesser mass a greater distance. Perception of moving sediment, however, may make this statement appear to be contradictory since there is, generally, an abundance of the smaller and lighter particles which are visually acute when airborne. Exceptional, or non-typical, wind events are able to transport fine particles across entire continents and oceans to distant lands where visual impacts to air quality may be readily observed.

Knowledge of particle entrainment and transport studies was greatly advanced through the work of Bagnold who used field observations in the deserts of Egypt and later wind tunnel studies to detail the mechanics of the erosional and transportation process. This work resulted in his well- known and widely used treatise *The Physics of Blown Sand and Desert Dunes.*[1] He observed that once a particle begins to move it will be transported by one of three modes: creep, saltation, or suspension. The primary difference in these three transport modes is a relatively simple relation between the particle size and the wind force which, unfortunately, becomes a complex mechanical problem that is today still not fully understood.

Wind Dynamics

Wind is by nature a random process and as such, the internal structure of a natural wind contains random variations known as turbulence (except in possibly very stable conditions). Turbulence develops in natural flows as inertial forces, which provide forward momentum (motion), react to viscous forces, which impede motion due to fluid properties. This results in a flow that is strongly sheared.

The unit of measure for wind speed is the mean wind velocity, $\bar{U}$, defined as an average value for a continuous time series of instantaneous velocities (Figure 1):

$$\overline{U} = \frac{\sum u_1 + u_2 + \ldots + u_i}{T} \tag{1}$$

where

$u_1 + u_2 + \ldots + u_i$ = components of instantaneous velocity
T = time interval over which velocity values are summed

As shown in Figure 1, at every instant a fluctuating component of velocity, u'_i exists defined as the difference between the instantaneous and the mean velocity values:

$$u_i^{'} = u_i - \overline{U} \tag{2}$$

It is these velocity fluctuations, whether positive or negative in regard to the mean velocity, that give rise to flow instability and thus, turbulence. As such, turbulence is not a property of the fluid but of the flow of the fluid. Turbulent flows are, therefore, unique in their ability to mix momentum and kinetic energy and to erode and transport particles.

Particle Entrainment

Bagnold described the instant of initial particle motion by a threshold fluid velocity which represents the wind speed at which the most susceptible particle becomes entrained. As a flow passes over a stationary particle, turbulent kinetic energy is imparted to the grain which builds a reservoir of stored potential energy.[2] Further increases in this energy reservoir cause the grain to oscillate and at the critical point

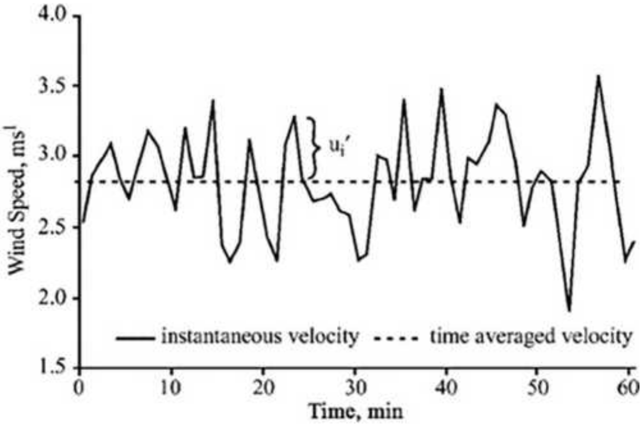


FIGURE 1 Time series of one min wind speeds illustrating principles of determining mean wind speed, instantaneous and fluctuating values. Unpublished data from author.

where potential energy exceeds the energy of adherence, the grain lifts, or more accurately is ejected in a near vertical trajectory away from the surface[2,3] initiating saltation. For a single grain of diameter d, Bagnold showed that the threshold fluid velocity, v_t, required to initiate saltation motion is related to particle diameter by:

$$v_t = 5.75A\sqrt{\frac{\rho_p - \rho_f}{\rho_p} - gd\,\log}\,\frac{z}{k} \tag{3}$$

where

A = Empirical constant = 0.1 for fully turbulent flows
ρ_p = Particle density
ρ_f = fluid density
g = gravity force
z = height for which threshold velocity is being calculated
k = Surface roughness height equal to about 1/30 the surface grain diameter.

For the case of a naturally irregular surface, such as a field or a sand dune, fluid threshold is not defined by a single wind speed or particle size but by a continuum of wind speeds over a length of time where grains of various diameter become entrained almost simultaneously.[4]

Once these grains are ejected into the overriding flow they attain a high-energy ballistic trajectory that depends upon particle size and wind velocity,[1,5,6] (Figure 2). Heavier and larger particles will travel relatively close to the surface (A) contrasted to lighter and smaller particles that initially are ejected higher into the wind profile (B). At some distant point downstream these grains (impactors) will impact on the sediment bed and either rebound upward again (C), eject new impactors (D), or lodge in the loose sediment. In either case, it is probable that upon impact sufficient kinetic energy will be transferred to one or several additional grains that are "splashed," or dislodged a short distance (E) known as reptation,[5,6] a low-energy process that includes a small number of grains traveling upstream. Reptation includes what Bagnold referred to as creep— grains that are too large to be entrained aerodynamically and move by rolling or sliding along the surface. Impact force is also a primary method leading to disaggregation of larger particle agglomerates by breaking particle bonds or the particles themselves into smaller pieces, enabling aerodynamic transport.

Saltation, which is defined as movement in a series of short hops, includes particles in the high-energy population (A, B, C, and D). These particles are initially moved by wind speeds exceeding the fluid threshold and subsequently through surface impacts that eject new particles at a point known as the impact threshold. Bagnold suggested that the impact threshold involves the same mechanics included in Equation 3 except $A = 0.08$. In other words, after saltation is initiated, additional grains are placed in motion at the threshold impact velocity that is about 80% of the threshold fluid velocity. This infers that once motion has begun, further motion can be sustained at wind speeds less than those required for initial grain motion.

The third transport mode, suspension, occurs in grains that are small enough to respond directly to turbulent fluctuations and thus, are lofted above the surface saltation layer.

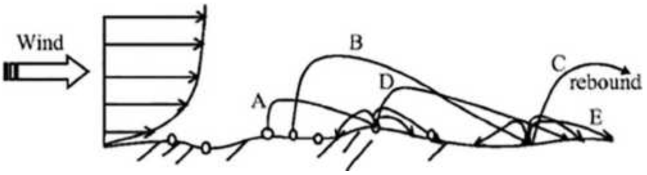


FIGURE 2 Particle transport paths in response to the wind profile (at left). A, B, C, and D are high-energy impacting grains moving in saltation. E are low-energy reptating particles including those moving by creep.

Self-Balancing Concept

Energy required to transport particles comes from extraction of turbulent momentum contained in the flow and a reduction in shearing forces near the surface.[7,8] Thus, as the flow adjusts to the presence of particles, a self-balancing condition in mass flux develops downstream of the point of initial grain motion.[9] In other words, the carrying capacity of the wind has been realized. This does not imply additional particles are not eroded but that as larger grains become deposited behind surface roughness elements, the loss of mass is balanced by erosion of new grains from the surface, primarily by impact and abrasion forces.

The self-balancing concept for maximum horizontal mass flux, $f_{\max}$, has been analyzed by Stout,[8] reducing to:

$$f_x = f_{\max}\left(1 - e^{\frac{x}{b}}\right) \tag{4}$$

where

f_x = horizontal mass flux at downstream distance, x, from a non-eroding boundary
b = length scale for stability in horizontal mass flux

and is applied to height-specific mass flux measurements. In utility, Equation 4 states that beginning from a non-eroding boundary at $x = 0$, f_x increases with x until $f_{\max}$ is attained. Downstream of this point, horizontal mass flux is relatively unchanged. Typical vertical and horizontal mass flux profiles from two wind events are shown in Figure 3.[10] Two components require discussion. First, vertical mass flux profiles, from 0.1 to 1.5 m height, show mass decreases with height. This is consistent with the discussion above, e.g., saltating particles having less mass loft higher into the flow. For any single profile, mass is greatest near the surface and decreases vertically. Second, horizontal mass flux increases downstream until the maximum transport capacity is reached. In the figure, mass continues to increase downstream not yet having attained $f_{\max}$.

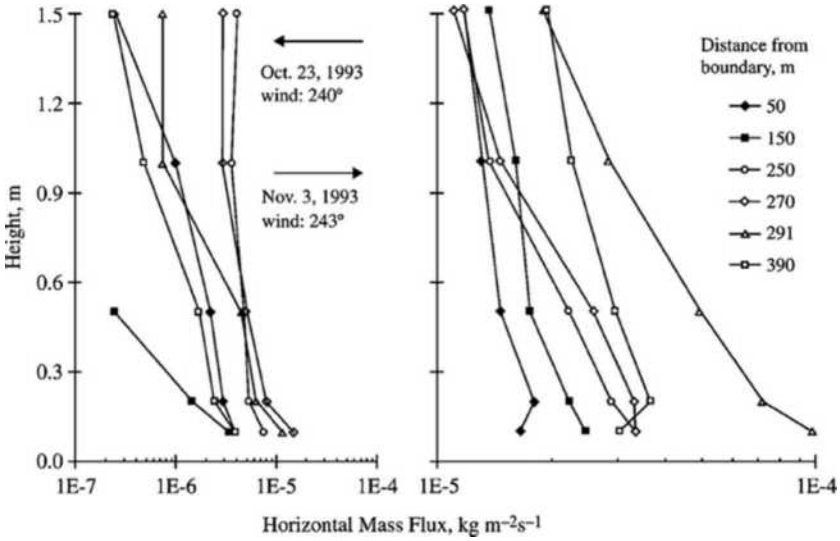


FIGURE 3 Vertical and horizontal mass flux profiles for two wind events illustrating the decrease and increase of mass with height and downstream distance, respectively.
Source: Stetler and Saxton.[10]

Conclusions

Complications to the above processes arrive due to numerous natural and human-induced conditions including: soil type, particle size distribution and fraction of erodible particles, climate, and land use. In cultivated fields, surface roughness is a primary consideration complicating the entrainment process as well as the establishment of the saltation layer. Rough surface elements have the effect of reducing wind shear forces thereby limiting erosion contrasted to a smoother surface that is favorable for a higher erosion rate.[11] Surface roughness can be either from soil clods, vegetation, or a combination of both.[12] Additionally, given all the factors discussed above, increases in soil moisture will cause an increase in the wind speed required to initiate grain motion.[13]

Prediction of wind erosion is, therefore, an inexact science, even with the advent of sophisticated computer models. The numerous and complex relations between all the critical parameters imply a degree of uncertainty, even in the best of calculations. The remaining entries will address some of these parameters more specifically.

References

1. Bagnold, R.A. *The Physics of Blown Sand and Desert Dunes*; Methuen and Co., Ltd.: London, 1941; 265 pp.
2. Reeks, M.W.; Hall, D. Deposition and resuspension of gas- borne particles in recirculating turbulent flows. J. Fluids Engr. **1988**, *110,* 165–171.
3. Greely, R.; Iverson, J.D. *Wind as a Geological Process on Earth, Mars, Venus and Titan;* Cambridge University Press: New York, 1985; 333 pp.
4. Nickling, W.G. The initiation of particle movement by wind. Sedimentology **1988**, *35,* 499–510.
5. Anderson, R.A.; Sorenson, M.; Willets, B.B. A review of recent progress in our understanding of Aeolian sediment transport. Acta Mechanica Suppl. **1991**, *1*, 1–19.
6. Unger, J.E.; Haff, P.K. Steady state saltation in air. Sedimentology **1987**, *34,* 289–299.
7. Anderson, R.S.; Haff, P.K. Simulation of Eolian saltation. Science **1988**, *241,* 820–823.
8. Stout, J.E. Wind erosion within a simple field. Trans ASAE **1990**, *33* (5), 1597–1600.
9. Owen, P.R. Saltation of uniform grains in air. Jour. Fluid Mech **1964**, *20,* 225–242.
10. Stetler, L.D.; Saxton, K.E. Wind erosion, PM_{10} emissions, and dry-land farming on the Columbia Plateau. Earth Surface Processes and Landforms **1996**, *21,* 673– 685.
11. Stout, J.E.; Zobeck, T.M. The Wolfforth field experiment: A wind erosion study. Soil Sci. **1996**, *161* (9), 616–632.
12. Horning, L.B.; Stetler, L.D.; Saxton, K.E. Surface residue and soil roughness for wind erosion protection. Trans. ASAE **1998**, *41* (4), 1061–1065.
13. Pye, K. *Aeolian Dust and Dust Deposits*; Academic Press: London, 1987; 334 pp.

17

Erosion: Snowmelt

Donald K. McCool

Introduction

In many areas of the world, winter hydrology is an important part of the annual erosion process; in some regions it is the primary cause of erosion. Sharratt et al.[1] indicate that about half of the Earth's land surface is frozen at some time during the year. In this topic, we will consider winter processes to include a range of conditions from those where the soil freezes and thaws diurnally and is subjected mainly to rainfall to those where the soil remains frozen for several months and precipitation occurs as snow or in other solid form during all or part of the winter season. In areas where daily minimum temperatures are rarely below 0°C, winter erosion processes are of relatively minor importance as compared to spring, summer and autumn erosion. At the other end of the spectrum are areas where the soil freezes in the autumn and stays frozen until spring and the ground is blanketed with snow for the entire winter.

FIGURE 1 Erosion on fall seeded winter wheat in eastern Washington caused by rain and snowmelt on thawing soil. Measured rill erosion was 200 metric tons per hectare.

Erosion is then confined to a period of spring snowmelt. Some areas such as the Palouse region of eastern Washington, northern Idaho, and northeastern Oregon are subjected to several runoff and erosion events each winter because the winter temperatures and precipitation patterns lead to multiple soil freezing and thawing occurrences each winter with accompanying rain or snowmelt events. The result of a particularly severe rain and snowmelt event when the soil had thawed at the surface on a fall-seeded field is shown in Figure 1.

Precipitation

Precipitation in areas where winter processes are important can occur as rain, or in solid form as sleet or snow. Rain can cause erosion due to splash and runoff detachment whereas sleet and snow cause detachment by runoff as they melt. In general, winter rainfall intensities are lower than summer rainfall intensities because of the lack of thunderstorm activity in the winter. However, the frequency and duration of precipitation are commonly greater. Since water loss from the soil is considerably lower during the winter due to lower mean temperatures and the absence of actively growing crops, even low intensity rainfall can increase the moisture content of soil rather quickly with a resultant dramatic increase in its susceptibility to erosion. The kinetic energy and intensity associated with rainstorms are not direct factors in snowmelt erosion. Snowmelt erosion is more closely related to volume and peak rate of runoff, which can be strongly influenced by rainfall when it occurs concurrently with the snowmelt.

Soil

The phase of the water in the surface layer of the soil is important in the winter erosion process. When soil water is frozen, erodibility, the susceptibility of the soil to erosion, is very low and erosion rates under sheet flow conditions are generally quite low. However, when runoff concentrates in small channels, the flow may cut into and through the frozen layer, leaving gullies and deeply incised channels. Soils that have been frozen and are thawing from the surface are weakened due to water expansion during the freezing process. Lee[2] cited a number of studies that indicate an increase in water content as water moves to the freezing front from deeper in the soil because the area where the soil is freezing is at very high water tension. The soil will regain strength after the frost has thawed and the soil reconsolidates as water drains from the soil.[3]

FIGURE 2 Gully erosion from snowmelt runoff in northern Idaho.

Water content of the soil at the time of freezing is important to the permeability of the soil and the opportunity for infiltration and deep percolation. Lee[2] found a linear relationship with a negative slope between the ratio of frozen soil infiltration hydraulic conductivity to the unfrozen soil infiltration hydraulic conductivity. Soil frozen at water content near saturation is frequently impervious and runoff can be nearly 100% of the rainfall or snowmelt,[2] thus leading to severe concentrated flow and gully erosion (Figure 2).

Effect of Snow and Freezing Conditions

During the nonwinter period or in climatic regions where winter erosion processes are not a factor, the only type of water erosion that occurs is rainfall on unfrozen soil. Winter erosion is more complex because there are seven basic types of erosion events:

1. Rain on frozen soil
2. Rain on thawing soil
3. Rain on unfrozen soil
4. Snowmelt on frozen soil
5. Snowmelt on unfrozen soil
6. Rain on snowmelt on frozen soil
7. Rain on snowmelt on unfrozen soil

Snowmelt on thawing soil and rain on snowmelt on thawing soil are not listed because, as discussed later, these conditions generally do not occur due to the insulating properties of snow. Further complicating these various types of events is that they typically do not stand alone but occur in some combination and how they combine will dictate the severity of any resultant erosion.

A blanket of snow on the soil surface can have positive as well as negative effects with regard to soil erodibility. Snow is an excellent insulator; the deeper the snow the better the insulation. Thus, a snow layer prior to or concurrent with freezing conditions can retard or prevent frozen soil. However, this insulating property also works extremely well at preventing frozen soil from thawing from the surface down (though thawing will occur in an upward direction due to deeper, warmer unfrozen layers if above ground temperatures are favorable). Snow cover on frozen soil typically results in the soil surface remaining frozen until bare soil begins to appear as the snow melts with warming temperatures, at which time the exposed soil begins to thaw from the surface down (Figure 3). The potential for severe

FIGURE 3 Melting snow exposes bare soil to thaw and increases erodibility.

erosion can be quite high with the presence of snow and depends on how all the contributing factors come together. If rainfall accompanies the warming temperatures required for snowmelt, runoff can be initiated over the frozen soil beneath the snowpack. If these conditions persist as patches of soil begin to be exposed and surface thawing follows, soil erosion will begin. The nature and severity of the erosion will be determined largely by the intensity and duration of the rainfall and to a lesser degree by temperature which affects the rapidity of the thawing process.

Runoff Events

It is difficult to ascribe a single set of characteristics to describe winter hydrology events on hillslopes or small watersheds. Events from rainfall on frost-impacted soil can be of very short duration whereas events resulting from extended periods of snowmelt can be of much longer duration. Events involving rainfall without snowmelt are usually brief although erosion rates can be high if the soil is thawing from the surface. Events involving snowmelt alone are extended in nature; runoff may fluctuate diurnally as air temperature dips below freezing at night. Erosion resulting from snowmelt alone is generally not severe. The highest runoff rates and flooding result from a combination of rainfall and snowmelt with high air temperatures. As long as the soil is frozen, sheet and rill erosion rates may be quite small, although concentrated flow erosion may be significant. When substantial quantities of bare soil appear, and the soil frost starts to thaw from the surface, erosion rates may increase dramatically. Sediment concentration graphs may lead or lag hydrographs under winter conditions. When there is no snow and the soil has thawed at the surface, a small amount of rainfall may lead to very high sediment concentration that occurs early in the event. When snow melts and the soil is not frozen, the sediment concentration and hydrographs may coincide. When snow melts over frozen soil, there may be little sediment until a portion of the soil is bare and melts. Then the runoff from the melting snow will detach and transport sediment, leading to a significant lag of peak sediment concentration.

Modeling

Erosion models are used to estimate location specific soil losses under various management systems within different land uses. They are also used to estimate expected amounts of sediment and chemicals that may be associated with those sediments. Models may be empirical in nature, such as the universal soil loss equation (USLE)[4] or the revised universal soil loss equation (RUSLE),[5] or they may be process based such as WEPP.[6] The inclusion of winter erosion processes even in empirical models such as USLE or RUSLE complicates the models. Erosion relationships based on runoff characteristics rather than rainfall characteristics may be more appropriate, creating the need to estimate runoff. Process-based models require inclusion of soil freezing and thawing, snowfall and snowmelt, and movement of water into frost impacted soil. Simple soil frost models are driven by air temperature, but more complex models consider radiation and other methods of energy transfer. It is difficult to strike a balance between excessive data requirement and adequate performance for applied use.

Amelioration

Preventing erosion damages from winter processes is similar to preventing erosion damages in other seasons. Cropland and rangeland are best protected by crop and surface cover. Snow collected in standing residue can prevent soil freezing and the creation of an impermeable condition. Surface cover protects from splash detachment, insulates the soil, and slows runoff. No till practices produce root channels and other pores that enable infiltration when the soil is frozen. Likewise, for areas where frost depth seldom exceeds 30–40 cm, deep ripping, chiseling, or slot mulching to expected frost depth can prevent formation of a continuous frost layer and lead to increased infiltration.

References

1. Sharratt, B.S.; Radke, J.K.; Hinzman, L.D.; Iskandar, I.K.; Groenevelt, P.H. Physics, chemistry, and ecology of frozen soils in managed ecosystems: an introduction. In *Intern. Symposium Physics, Chemistry, and Ecology of Seasonally Frozen Soils;* Iskandar, I.K., Ed.; U.S. Army Cold Regions Res. Engin. Lab Spec. Rep. 97–10, 1997; 1–7.
2. Lee, H.W. Determination of Infiltration Characteristics of a Frozen Palouse Silt Loam Soil Under Simulated Rainfall. Dissertation, Doctor of Philosophy in Agricultural Engineering: University of Idaho, Moscow, 1983.
3. Formanek, G.E.; McCool, D.K.; Papendick, R.I. Freeze- thaw and consolidation effects on strength of a wet silt loam. Trans. ASAE **1984**, *26* (6), 1749–1752.
4. Wischmeier, W.H.; Smith, D.D. *Predicting Soil Erosion by Water—a Guide to Conservation Planning;* USDA Agriculture Handbook No. 537; 1978; 58 pp.
5. Renard, K.G.; Foster, G.R.; Weesies, G.A.; McCool, D.K.; Yoder, D.C. (Coordinators). *Predicting Soil Erosion by Water: A Guide to Conservation Planning with the Revised Universal Soil Loss Equation (RUSLE)*; USDA Agriculture Handbook No. 703; 1997; 404 pp.
6. Flanagan, D.C.; Nearing, M.A., Eds.; *USDA–Water Prediction Project: Hillslope Profile and Watershed Model Documentation;* NSERL Report No. 10; USDA–ARS National Soil Erosion Research Laboratory: West Lafayette, IN, 1995.

18

Erosion: Soil Quality

Craig Ditzler

Introduction

Soil erosion involves the detachment and removal of soil material from one site and its transport to another location. Soil erosion usually degrades soil quality and a soil of poorer quality is less able to withstand further erosion, thus creating a downward spiral of soil degradation. When the surface soil is removed through erosion, organic matter and clay particles may be lost, with consequent reductions in fertility, biological activity, aggregation and rooting depth. Other potential effects of erosion on soil quality include reduced porosity and infiltration, formation of crusts on the soil surface, changes in soil texture, and compaction. These changes in turn reduce the capacity of the soil to supply and cycle nutrients, filter and degrade toxic materials, store and supply moisture and sustain plant and biological productivity. They may also result in increased runoff, less biomass production and plant cover, and greater susceptibility to further erosion. Erosion increases the variability in soil quality across a field and, on a broad scale, is associated with widespread loss of agricultural productivity and declining water quality.

Soil erosion is a natural process by which soil particles are detached and moved by water, wind, gravity, or ice. All soils have an inherent erodibility, or natural susceptibility to erosion, based on soil features, topography, and climate. However, human activities, such as logging, livestock grazing, tillage, removal of vegetation, and urban development, can greatly accelerate natural rates of erosion. Cleared and managed as they are for crop and livestock production, agricultural soils are particularly susceptible to wind and water erosion (Figures 1 and 2), and also to a recently recognized process known as tillage erosion—the loosening of soil by tillage equipment and its downslope movement under gravity.[1]

Eroded soil may move only a few meters in a field and come to rest in lower positions, resulting in increased variability of surface soil properties across the field as subsoil becomes exposed in some places and surface layers are buried and over-thickened in others. It may also move great distances, being deposited in neighboring fields, roadside ditches, and water bodies. In some cases of wind erosion, fine soil particles may travel many kilometers before being deposited.

Effects of Soil Erosion on Soil Quality

The loss of surface soil in a landscape may impair soil function, and thus soil quality, through adverse effects on many physical, chemical, and biological properties of the soil.[2] Table 1 summarizes some of the effects of erosion on the soil quality functions listed by Gregorich.[4]

FIGURE 1 Deposition of wind-blown soil material. (From USDA, Natural Resources Conservation Service—Soil Quality Institute.)

FIGURE 2 Severe erosion by water has removed all of the surface layer and much of the subsoil. (From USDA, Natural Resources Conservation Service—Soil Quality Institute.)

TABLE 1 Some General Effects of Soil Erosion on Soil-Quality Functions

Erosion Effects	Soil-Quality Functions Affected
Reduced cation-exchange capacity	Sustaining plant growth and animal life Buffering or filtering Regulating energy
Formation of surface crusts	Regulating water, gases, and energy Sustaining plant growth and animal life
Changes in rooting volume Changes in surface layer texture	Sustaining plant growth and animal life Sustaining plant growth and animal life
Loss of organic matter	Regulating water and gases Sustaining plant growth and animal life
Compaction	Regulating energy Regulating water and gases
Deterioration of soil structure	Sustaining plant growth and animal life Regulating water and gases
Exposure of subsoil material	Sustaining plant growth and animal life Sustaining plant growth and animal life Regulating water, gases, and energy

Physical Effects

Erosion selectively removes the finer, lighter particles from the soil surface, leaving coarser particles behind. Depending on the severity of erosion, an eroded soil may become very coarse in texture, sometimes with a gravelly surface. The deposition of the eroded material in lower topographic areas may result in a thickening of the topsoil and an increase in rooting volume.

Tillage of eroded soils may result in a mixing of subsoil with the surface soil, altering its composition. For example, cultivation of eroded soils having clay-enriched subsoils may increase the clay content of the surface soil. As erosion progresses, plant roots must enter progressively deeper into the subsoil layer to obtain nutrients and water. Where subsoil layers restrict root growth because of their physical and/or chemical properties, the depth of rooting is reduced, along with the capacity of the soil to supply water and nutrients to plants.

Individual soil particles are held together in aggregates, which form the structural fabric of the soil. Soils with good structural arrangement of aggregates and the pore spaces between them provide good aeration for soil roots and microbes; allow ready movement and storage of water and plant nutrients in the pore spaces; and retain their structure when exposed to stresses such as cultivation and the impact of raindrops.[3,4] Abrasion by wind, rain, and tillage can disintegrate aggregates at the soil surface, and the resulting fine particles can plug larger soil pores and form a hard, thin crust on the surface. This crust seals the surface and limits the infiltration of air and water, as well as impedes the emergence of seedlings. Finer soil particles are also more easily compacted as the pore spaces between them are reduced under the pressure of farm machinery. Deteriorating soil structure further increases the risk of soil erosion—fine soil particles created by the breakdown of aggregates at the soil surface are especially vulnerable to wind and water erosion. Furthermore, compacted, crusted soils resist the infiltration of water, increasing the volume of surface runoff and compounding the effects of erosion. Thus, erosion reduces soil quality, making the soil prone to further erosion and further degradation.

Chemical Effects

Organic matter and clay are important sites of cation exchange in the soil. Cation-exchange capacity is a measure of negatively charged sites on the soil particles that are capable of holding positively charged ions, including many plant nutrients. As organic matter and clay particles are lost from the soil surface through erosion, attached nutrients are relocated in the landscape, often to adjacent water bodies where they contribute to declining water quality. Loss of these fine soil particles also impairs the ability tostore nutrients, reducing soil fertility. This effect is more pronounced in sandy soils containing small amounts of clay, though it may be offset to some degree if erosion causes the surface layer to become more clayey. Where subsoils are more enriched in clay than the surface soil, the clay particles may form chemical bonds with phosphorus, fixing it into forms not easily available to plants and thus reducing fertility.

For some acid soils with subsoil pH <5.0, concentrations of available aluminum may be at levels toxic to plant roots and removal of surface soil by erosion can effectively reduce the rooting depth of the soil.

Subsoil layers having pH >8.5 often contain high levels of sodium. Surface soil removal and subsequent exposure of these subsoils to rainwater or irrigation waters with low ionic concentrations can lead to dispersion of clay particles, loss of soil structure, surface sealing, and greatly reduced water infiltration.

Biological Effects

Removal of organic matter and nutrients from the soil surface by erosion reduces the food and energy supply needed to sustain healthy populations of soil organisms and support plant growth. Soil organisms are the agents of organic matter decomposition and nutrient cycling in soil. In addition, they play an important role in the stabilization of soil aggregates through the production of binding agents such as roots, fungal hyphae, polysaccharides, gums, and complex molecules consisting of humic substances combined with iron, aluminum, or aluminosilicates.[5,6] So, as erosion proceeds, the food source for organisms is reduced, leading to declines in populations. Soil structure and stability then deteriorate, the soil becomes more susceptible to further erosion, and the cycle continues.

Conclusions

The interaction of accelerated soil erosion and soil quality is complex. Soil erosion usually reduces soil quality, and a soil of poorer quality is less able to withstand erosion, thus creating a downward spiral of soil degradation. However, many soil-conservation practices have been useful in mitigating the effects of erosion. Reduced tillage systems limit soil disturbance and build soil structure. Crop residue management, under seeding, cover cropping, and permanent cover (e.g., pasture) protect the soil from the action of wind and water. Contour cultivation, grassed waterways, and terracing alter the flow of surface water, curbing water erosion. Herbaceous wind barriers and woody windbreaks and shelterbelts help to control wind erosion. Nevertheless, large tracts of agricultural land throughout the world are still subject to the unsustainable loss of soil as a result of erosion, and continued adoption of preventive methods are needed to protect and restore soil quality.

References

1. Govers, G.; Lobb, D.A.; Quine, T.A. Tillage erosion and translocation: emergence of a new paradigm in soil erosion research. Soil Tillage Res **1999**, *51,* 167–174.
2. Lal, R. Ed. Soil quality and soil erosion. CRC Press: Boca Raton, FL, 1998.
3. Karlen Douglas, L.; Stott Diane, E. A framework for evaluating physical and chemical indicators of soil quality. In *Defining Soil Quality for a Sustainable Environment,* SSSA Special Publication No. 35; Doran, J.W., Coleman, D.C., Bezdicek, D.F., Stewart, B.A., Eds.; Soil Science Society of America: Madison, WI, 1994; 53–72.
4. Topp, G.C.; Wires, K.C.; Angers, D.A.; Carter, M.R.; Culley, J.L.B.; Holmstrom, D.A.; Kay, B.D.; Langille, D.R.; McBride, R.A.; Patterson, G.T.; Perfect, E.; Rasiah, V.; Rodd, A.V.; Webb, K.T. Changes in soil structure. In *The Health of Our Soil: Toward Sustainable Agriculture in Canada;* Acton, D.F., Gregorich, L.F., Eds.; Agriculture Canada: Ottawa, Ont., 1995; 51–60.
5. Tisdall, J.M.; Oades, J.M. Organic matter and waterstable aggregation. J. Soil Sci. **1982**, *33*(2), 141–163.
6. Wright, S.F.; Starr, J.L.; Paltineanu, I.C. Changes in aggregate stability and concentration of glomalin during tillage management transition. Soil Sci. Soc. Am. J. **1999**, *63*(6), 1825–1829.

19

Farming: Organic

Brenda Frick

Introduction

IFOAM, the International Federation of Organic Agriculture Movements, defines organic agriculture as follows: "Organic agriculture is a production system that sustains the health of soils, ecosystems and people. It relies on ecological processes, biodiversity and cycles adapted to local conditions, rather than the use of inputs with adverse effects. Organic agriculture combines tradition, innovation and science to benefit the shared environment and promote fair relationships and a good quality of life for all involved."[1]

IFOAM identifies four overlapping principles for which organic agriculture strives:[2]

- "The principle of ecology: the desire to work with, emulate and help sustain living ecological systems and cycles.
- The principle of care: management in a responsible manner protecting the well-being of people and environment.
- The principle of health: sustaining and enhancing the health of soil, plant, animal, human and planet.
- The principle of fairness: ensuring fairness among people and in relationship to other living beings."

Organic Farming Standards

Organic standards around the world share a vision of the farm as an ecological system. They emphasize holistic management through biological and ecological processes, soil building, suitable crop rotations, humane handling of livestock, and a thorough audit trail to assure consumers of the integrity of the process. Organic standards generally prohibit genetically modified organisms, synthetic fertilizers, pesticides, growth regulators, and antibiotics. Some organic standards include social justice standards as well.[3,4]

Organic production is a long-term commitment. Farms must meet the regulations for a minimum period of time, generally 3 years, before they qualify for organic status. They must have a farm plan that includes proactive techniques to improve the farm's environmental status, build soil fertility, and reduce the impacts of non-beneficial organisms.

In recent years, certification standards have become mandatory in many countries.

In the United States, the Organic Foods Production Act was passed in 1990, enacting a standard that was built with an unprecedented amount of input from farmers, processors, and consumers. The standard was fully implemented in 2002. Now any products sold in the United States that are labeled "organic" must be in compliance with the standard.[5]

The Canadian Organic Products Regulation was enacted in 2006 and came into force July 2009. Products labeled as organic in Canada must meet the standard.[6]

Europe has the EU standard, EEC No. 2092/91, as well as smaller national or private standards. Japan also has a national standard.[7]

Products sold across boundaries must meet the standards of the importer. Canada and the United States signed equivalency agreements to recognize each other's standards in 2009. The two sets of standards differ in only a few "variances." A product that is certified organic in one of these countries is approved as organic in the other after an inspection verifies that it also meets the variance requirements. In practice, this additional step is generally included with the annual inspection process. Equivalency is intended to ease trade between these two major trading partners. Similar agreements will likely follow for other countries.[8,9]

Soil Building

Soil is literally the foundation on which all agriculture is based. In organic agriculture, healthy soils are seen as the key to healthy plant growth, which in turn is the key to healthy livestock and healthy people.

Soil management in organic systems is a matter of fostering biological processes in the soil, rather than applying synthetic off-farm inputs. Synthetic nitrogen fertilizers are made in an energy-intensive process generally starting with natural gas. By eliminating synthetic nitrogen sources, organic farming can significantly reduce energy consumption 30%–50% in field crops (less where nitrogen fertilization is less).[10,11] Synthetic nitrogen and phosphorus fertilizers are primary pollutants of rivers and estuaries downstream from agricultural cropland. By using less leachable, biological sources of fertility, organic farms are responsible for less nutrient pollution.[12]

In organic systems, biological processes in the soil are encouraged by practices such as use of green manures, cover crops, and mulches. Tillage is used judiciously when needed to manage weeds and pests or to prepare a seedbed.[13]

Green manures are crops that are grown to provide food and habitat for soil organisms, rather than products for harvest. Legumes are especially useful green manures, as they can partner with *Rhizobia,* which capture nitrogen from the air and make it available to the legumes. When these legumes are incorporated into the soil, soil organisms release the nitrogen in a form that other plants can use. Legumes and many other plants can also form beneficial relationships with mycorrhizae. Mycorrhizae form extensive networks of filaments throughout the soil and can "capture" soil nutrients from a much larger volume of soil than the plant roots could access on their own. Again, these nutrients are released in an accessible form when the green manure plants break down in the soil. Other green manure plants, such as cereals, provide "fiber," or soil organic matter, which is crucial to soil organisms involved in nutrient cycling.

Cover crops are used to reduce erosion potential and nutrient loss between crops. For instance, a cover crop of rye might be seeded after the harvest of a summer annual crop. The rye would act to stabilize the soil during the erosion-prone fall and spring. In the case of a legume crop, the cover crop could also act to sequester the nitrogen released from the legumes, so that it would not leach away.

Mulches can be living short-statured plants that grow between the rows of the desired crop, providing shade for the soil. This reduces moisture loss, temperature fluctuation, and germination of weed seed. It increases the potential activity of beneficial insects and soil organisms. Mulches can also be killed plants.

Some cover crops can be killed and seeded into in one pass of a roller crimper (see Rodale's No-till+).[14] This technology reduces the problematic effects of tillage and uses less energy by combining operations. Mulches of straw or other plant residues can also be used in smaller-scale operations.[15]

Organic systems tend to have greater soil biological activity as a result of these practices. In a study conducted across the Canadian prairies, soils from organic rotations had greater aggregate stability, a key in reducing erosion potential, than conventional rotations. This may arise from greater mycorrhizal fungi, which produce "sticky" polysaccharides that bind soil particles together.[16]

Farm operations that include livestock can use these livestock to improve soil fertility. Grazing green manures, for instance, can improve the economic value of the green manure while releasing most of the nutrients in a form that is very effectively used by the soil organisms. This method also reduces off-farm energy consumption, by allowing the animals to harvest the material themselves. Intensive rotational grazing systems can rejuvenate soils by releasing nutrients from the standing vegetation back into a soil-available form. Grazing pigs essentially "till" soil, loosening and aerating it. They can also clean up culls or fallen grain. When natural animal activity replaces some of the tractor-based tillage, energy is saved.

Genetically Engineered Organisms

Organic systems prohibit genetically engineered products. The results of moving genetically engineered products into our food system are, as yet, unknown. In North America, governments are reluctant to label food products containing the products of engineering. Thus, for those who wish to avoid this form of contamination of their food system, organic foods are their safest course.

Pest Management Practices

Organic systems prohibit synthetic herbicides and pesticides. They thus eliminate the concern of off-target losses of these products. Significant levels of biocides are detected in the air and water globally, especially near agricultural fields.[17–19] Exposure to these biocides is linked directly to poisoning of aquatic systems and of pesticide handlers and indirectly to increasing levels of human cancer and other diseases.[17]

Organic production instead chooses a systems approach to farming. Producers strive to understand the ecological relationships that influence the abundance of the various species in their systems and to avoid outbreaks of those species that harm the crop. Many of the methods used to favor the crop in the ecological community can be summarized by good crop husbandry—avoidance of pests through sanitation and quarantine; seed cleaning; management of soil fertility; selection of locally adapted and competitive crops and crop cultivars and breeds of livestock; appropriate timing, depth, and rate of seeding; and integration of cropping and livestock and judicious use of off-farm in- puts.[20] Use of farm-grown seed and closed herd management reduce immigration of pests. Where problems arise, introduction of new genetics may be warranted. Mechanical and biological pest controls are used as necessary but only as a last resort.[16,21]

Rotation is one of the strongest tools that the producer uses. A diverse crop rotation increases microbial and mycorrhizal activity in the soil and improves crop vigor. Populations of crop-specific pests, such as many diseases and insects, are severely reduced by years when a given crop is not grown.[22] Crop rotations alter the timing and competitive relationships of crops and reduce the build-up of weed communities adapted to any given management practice. Crop and cultivar selection is also very important. Matching crops to fields improves their competitive relations with weeds, and helps them to resist both insect and pathogen attacks.

Similarly, rotation of livestock is important to maintain quality of grazing lands. Multiple species increases diversity and reduces problems. For instance, chickens following beef animals can reduce fly larvae; sheep and cattle can graze different plant species and different forage heights.

Weeds

In organic systems, weeds are considered to be a part of the ecological system. They can be beneficial in moderating the soil environment, providing habitat and food for micro- and macrofauna, moving soil nutrients to the surface, and indicating soil or management problems. Of course, large weed populations are often detrimental. Weed management begins with avoidance, includes managing the crop and weed environment to minimize weeds, and only lastly involves direct weed control.[16]

A strongly competitive crop is an excellent defense against weeds. Competitive crops, such as fall rye and sweet clover, and perennial crops, such as alfalfa, are particularly effective at reducing the weed community. Less competitive crops, such as flax and lentil, are best seeded in less weedy fields, following strongly competitive crops.[16] Crop competition can be increased by appropriate cultivar selection and by crop management techniques. One of the most effective of these is increasing the seeding rate.[16] Heavier seeding rates are more effective in weedy situations. Organic recommendations are generally 1.5 to 2 times the seeding rates recommended when herbicides are applied. When used in combination, cultural techniques such as appropriate cultivar, diverse rotation, and increased seeding rate increased crop yield and reduced weed competition.[16]

Tillage can be used to reduce weed populations. Tillage can be used after harvest, before seeding, and in fallow years. Concerns over the negative effects of tillage on soil quality, especially on erosion potential, have reduced the frequency of fall and fallow tillages, though these may be cautiously used for perennial weed control.[23] Delayed seeding after spring tillage remains an important tool. Early tillage stimulates the germination of volunteer crop and weed seeds. These are destroyed with a second tillage at or before seeding. This strategy has been especially effective at reducing the abundance of early emerging species such as wild oats, *Avena fatua,* and winter annual weeds such as stinkweed *(Thlaspi arvense)*.[16]

Harrowing after seeding or even after crop emergence can also be effective. This strategy is most effective for weeds that emerge from shallow depths, such as green foxtail, in crops that have large, deeply placed seeds. For row crops, interrow cultivation can be effective. The combination of early harrowing across the rows and interrow cultivation can offer good weed control. Other mechanical weed control techniques include flaming, burning, and mowing. Chaff collection at harvest can remove significant numbers of weed seeds. It is most effective for volunteer crop seeds, and weed seeds such as lambsquarters *(Chenopodium album),* that are largely retained on the plant at a height above the stubble.[16]

Organic farmers are increasingly seeking ways of reducing tillage, substituting machinery such as noble blade cultivators, which kill weeds but leave them standing, and roller crimpers, which crimp weeds but leave them attached to the soil, and mowing. All these methods reduce erosion potential and soil disturbance relative to cultivation. Their efficacy for weed control is still under investigation. Use of forages, silage crops, and direct stubble grazing or green manure grazing further varies the rotation, allowing better weed control with less tillage. For many organic farmers, there is a trade-off between reducing tillage and increasing weed pressures. Further research is needed in this area.[16]

Biocontrol of weeds includes the use of livestock, weed-eating insects, and weed-suppressing diseases. Livestock can be used for grazing or to consume mowed weeds. Livestock can also be used to consume chaff or seed screenings. Biocontrol insects generally target perennial weed species. For instance, the black dot spurge beetle has been released for control of leafy spurge. Few fungal biocontrol agents are available. Examples include DeVine® for strangler vine, Collego® for northern joint vetch, Biomal® for round-leaved mallow, and *Phoma* for dandelion. Beneficial organisms can be encouraged by practices such as reducing tillage, maintaining shelterbelts, and growing crops (such as legumes and some cereals) that encourage mycorrhizae. These can also be effective for weed control.[21]

Insects and Other Invertebrates

A majority of insects found in crop fields are beneficial or are of no economic importance. Diversity of habitat and wildlife encourages positive interactions among species and reduces the potential for outbreaks of insect pests.

Organically grown crops may be less attractive to insects than crops grown with abundant synthetic fertilizers. Synthetic fertilizers may result in the accumulation of excess nitrate in plant tissue; composted manure and green manures release nitrogen more slowly, reducing the potential for this accumulation. Excess nitrate makes plants more attractive to insect pests and can increase the reproductive rate of some insects, thus increasing the severity of an insect outbreak.

Crop rotations can be effective against insect pests with limited dispersal capabilities, such as the corn rootworm. For these types of insects, reducing crop residues and volunteer crop are also important.

For widely dispersed insects that are attracted to specific crops, large-scale cropping diversity is important. Monocultures favor these insects. Increasing the presence of non-attractive plant species reduces the incidence of attack. Practices such as strip cropping and intercropping can reduce the attraction of insect pests by diluting the aroma of the crop or confusing the insect's search image. For instance, underseeding canola with yellow sweet clover may reduce the incidence of flea beetles. Even weed populations may function this way. For instance, weeds in alfalfa may increase the habitat for parasitic wasps that control alfalfa caterpillars. Barrier strips of an unattractive crop around the outside of the susceptible crop may prevent insects from crossing into the susceptible crop. For instance, a border strip of peas may reduce the movement of grasshoppers into wheat. Trap strips of an attractive crop may be sown. After insects accumulate there, it can be mowed or cultivated, thus destroying many of the insects. Small-sized fields also limit the problem of insect outbreaks.

The time of seeding can sometimes be altered to avoid insect pests. For instance, late seeding of canola can reduce the severity of flea beetles, and early seeding of wheat can reduce the attack of wheat midge.

Cultivar selection may help prevent insect problems. Wheat cultivars differ substantially in their susceptibility to wheat midge, in large part due to their rate of development.

Biocontrol agents are available for some insect pests. Predators such as ladybugs and lacewings can be used to reduce aphid populations. *Bacillus thuringiensis* strains have been developed for control of several insects, including caterpillars and beetles.[24] Some of these control organisms can be purchased. Maintaining a varied habitat in and around the crop field can also help to harbor such natural organisms.

Pheromone traps may be used to lure insects away from crops. Sticky traps can also be used. These trapping methods are particularly appropriate to monitoring insect populations.[24] They are less effective at large-scale insect removal. In high-value crops, such as potatoes or strawberries, insect "vacuums" can be used for larger insects such as Colorado potato beetles or lygus bugs.

A few products can be used under organic certification standards. Soaps and oils can be used to suffocate insects such as aphids. Diatomaceous earth can be used to discourage soft-bodied insect larvae, rusty grain beetles, and slugs and snails. Natural products such as pyrethrum and rotenone are also acceptable under most certification standards.

Diseases

As with insects, many fungal and bacterial species are beneficial or benign. Practices that increase biodiversity will likely increase species that compete with or prey on disease species and thus reduce the outbreak of disease epidemics. The severity of disease can be limited by reducing the population of the pathogen or the susceptibility of the host or by changing environmental conditions that favor infection.

Rotation is an important key to reducing disease by reducing the inoculum level of the pathogen. Crop rotation is effective when pathogens are obligate and host specific, with low dormancy and poor aerial spread. Leaf blights of cereals and ascochyta blights of lentil can be reduced by rotations that include different crops. Crop rotation alone is not sufficient to eliminate diseases in perennial crops; diseases such as common root rots and seedling blights with a wide host range; diseases that persist in soil, such as *Fusarium* wilt of flax; diseases with rapid spread, such as powdery mildew of peas; or diseases that are widespread in the air or by insect vectors, such as cereal grain rusts and aster yellows.

Straw, residue, and weed management can be important. Straw is a primary inoculum for some disease species. Incorporation of residues speeds their decomposition and thus reduces the pathogen population. The risks to soil quality with excess tillage need to be balanced against the risk of leaving inoculum in the field. Other sources of infection include volunteers of the target crop and weeds that are closely related to it. These sources of inoculum can reduce the effectiveness of a rotation away from the target crop.

Where diseases spread slowly from adjacent fields, field edges can be treated separately. Barrier strips or early swathing or mowing of the severely affected area may reduce the spread of disease into the crop.

Seed quality can impact disease potential. Reducing seed damage during harvest, storage, and seeding can reduce the susceptibility to disease, especially seedling blights. This is especially important for seeds such as flax, rye, and pulses. Disease-free seed reduces the spread of seed-borne diseases into new areas.

Host susceptibility can be reduced by the selection of appropriate cultivars. Resistance or relative tolerance to a number of diseases varies greatly among cultivars. Differences among cultivars in disease susceptibility may reflect differences in their rate of growth; differences in their architecture and thus canopy humidity; or physical, biochemical, or genetic properties that restrict disease entry. Crop timing may also be important. For instance, earlier seeding may reduce the incidence of diseases such as barley yellow dwarf, powdery mildew of pea, and pasmo of flax. Late seeding of fall-seeded crops can reduce spread of disease by reducing their overlap with similar spring-seeded crops. Late seeding into warm soil can increase seedling vigor and thus reduce crop susceptibility to seedling blights. Reduced seeding rates can reduce the spread of disease by reducing contact among plants and by altering the environment in the canopy. Intercropping or strip cropping can be effective as well.

Some environmental manipulations can make disease frequency less severe. Selection of an appropriate field for a given crop can be important. Nutrient imbalance can make diseases such as take-all in cereals more severe. The incorporation of manure and green manures in rotations can reduce the severity of disease by encouraging microbes that are antagonistic to crop pathogens.

Future Concerns

Organic farming has many of the concerns that plague conventional farming: an aging cohort of farmers; a scarcity of reliable information sources; an economic system that puts little value on ecological services provided by farmers and farmland; and increasingly, commoditization of organic products.

Despite strong organic regulations, there is evidence that many consumers are not well informed. This leads to copycat items in the marketplace, claiming organic attributes without the rigor behind the organic claims.

There is concern among organic producers that the popularity of biotechnology in crop breeding will result in the abandonment of traditional breeding programs. If all or even most future genetic disease and insect resistance is incorporated into a genetically modified background, it will be unavailable to organic producers. This will greatly reduce their pest management options.

Farmers are forced to make compromises between the soil benefits of reducing tillage and the efficacy of tillage for weed control. Most organic farmers export products while attempting to cycle materials internally. Erratic weather threatens cropping; immature market infrastructure creates volatility in farm income. Farming of any sort is financially risky relative to most other professions. Poor crops and poor markets cause people to seek off-farm income. This and a local labor shortage reduce the ability of many farms to access the benefits of integrating livestock into their cropping systems.

Organic producers have pioneered a path away from high-energy-consuming and polluting synthetic fertilizers and biocides. They have chosen to emulate nature, fostering natural ecosystem processes and biological cycling. It seems they are on a path toward sustainability, but further development is needed on this path.

References

1. IFOAM. *Definition of Organic Agriculture,* available at http://www.ifoam.org/growing_organic/definitions/doa/index.html (accessed September 2010).
2. IFOAM. *Principles,* available at http://www.ifoam.org/about_ifoam/principles/index.html (accessed September 2010).
3. Anonymous. *Organic Rules and Certification,* available at http://organicrules.org/ (accessed September 2010).
4. IFOAM. *Regulations,* available at http://faolex.fao.org/docs/pdf/eur72928.pdf (accessed September 2010).
5. Anonymous. *The US Organic Standards,* available at http://ota.com/organic/us_standards.html (accessed September 2010).
6. Anonymous. *Canadian Organic Regulations,* available at http://www.cog.ca/index.php?page=organic-regulations (accessed September 2010). Suggest replacement with Anonymous. General Principles and Management Standards CAN/ CGSB-32–310–2006, available at http://www.tpsgc-pwgsc. gc.ca/ongc-cgsb/programme-program/norms -standards/ internet/bio-org/principes-principles-eng.html#a065 (accessed May 2012).
7. Anonymous. *Organic Standards,* available at http://www.ioia.net/standards.html (accessed September 2010).
8. Anonymous. Canada-European Union Organic Equivalency Arrangement – Overview. Available at http://www.inspection.gc.ca/food/organic-products/equivalence- arrangements/european-union-overview/eng/1328074578961/ 1328074648480 (accessed May 2012).
9. Anonymous. US EU Organic Equivalence Arrangement. Available at http://www.ota.com/GlobalMarkets/US-EU- Organic-Equivalence-Arrangement.html (accessed May 2012).
10. Hoeppner, J.W.; Entz, M.H.; McConkey, B.G.; Zentner, R.P.; Nagy, C.N. Energy use and efficiency in two Canadian organic and conventional crop production systems. Renew. Agric. Food Syst. **2006**, *21,* 60–67.
11. Pimentel, D.; Hepperly, P.; Hanson, J.; Douds, D.; Seidel, R. Environmental, energetic, and economic comparisons of organic and conventional farming systems. Bioscience **2005**, *55*, 573–582.
12. Lynch, D. Environmental impacts of organic agriculture: A Canadian perspective. Can. J. Plant Sci. **2009**, *89,* 621–628.
13. Mäder, P.; Fliebbach, A.; Dubois, D.; Gunst, L.; Fried, P.; Niggli, U. Soil fertility and biodiversity in organic farming. Science **2002**, *296* (5573), 1694–1697.
14. Anonymous. No-Till Revolution, available at http://www.rodaleinstitute.org/no-till_revolution (accessed May 2012).
15. Nelson, A.; Froese, J.; Beavers, R.L. *Lowering Soil Erosion Risk in Organic Cropping Systems*; Final Research Report W2006–09, available at http://www.organicagcentre.ca/Docs/OACC_bulletins06/OACC_Bulletin9_erosion_risk.pdf (accessed February 2010).
16. Frick, B., Ed. *Organic Farming on the Prairies,* 2nd ed.; Saskatchewan Organic Directorate: Muenster, SK, 2012.
17. Ibitayo, O. Agricultural pesticide contamination. *The Encyclopedia of Earth*; 2007, available at http://www.eoearth.org/article/Agricultural_pesticide_contamination (accessed October 2010).
18. Kumar, Y. *Pesticides in Ambient Air in Alberta.* ISBN 0–7785-1889-4. Report prepared for the Air Research Users Group, Alberta Environment: Edmonton, Alberta, 2001.

19. Sanborn, M.; Cole, D.; Kerr, K.; Vakil, C.; Sanin, L.; Bassil, K. *Pesticides Literature Review,* available at http://www.ocfp.on.ca/English/OCFP/Communications/CurrentIssues/ Pesticides/ default.asp?s=1 (accessed October 2010).
20. Harker, K.N.; O'Donovan, J.T.; Irvine, R.B.; Turkington, T.K.; Clayton, G.W. Integrating cropping systems with cultural techniques augments wild oat *(Avena fatua).* Manage. Barley Weed Sci. **2009**, *57* (3), 326–337.
21. Wallace, J., Ed. *Organic Field Crop Handbook*; Canadian Organic Growers, Inc.: Ottawa, Canada, 2001.
22. Bullock, D.G. Crop rotation. Crit. Rev. Plant Sci. **1992**, *11,* 309–326.
23. Snyder, C.; Spaner, D. The sustainability of organic grain production on the Canadian prairies—A review. Sustainability **2010**, 2, 1016–1034.
24. Reganold, J.P.; Glover, J.D.; Preston, K.A.; Hinman, H.R. Sustainability of three apple production systems. Nature **2001**, *410,* 926–930.

20

Global Climate Change: World Soils

Rattan Lal

Introduction

Earth's vast reservoir of carbon (C) is contained in four distinct pools (Table 1). The soil C pool comprises both organic and inorganic forms. Soil organic carbon (SOC) is estimated at 1,550 Pg (1 Pg=petagram=10^{15} g=1 billion metric tons) and soil inorganic carbon (SIC) at 940 Pg.[1] The atmospheric pool contains 820 Pg of C, and the vegetation for the biotic pool amounts to about 620 Pg of C.[2] The pedologic or soil C pool, with a total reserve of about 2,490 Pg to 1 m depth in both organic and inorganic forms, is about 3.5 times the atmospheric pool and 3.8 times the biotic pool.[1–6]

In addition to being the largest terrestrial pool, world soils also play a major role in the global C cycle (Figure 1).[3,4] The SOC pool is highly dynamic, concentrated near the soil surface, and comprises material of plant and animal origin at various stages of decay and decomposition. A large part of the SOC pool is held in the organic soils of tundra and boreal forest ecosystems.[5] It is this dynamic SOC pool

TABLE 1 Global Pools of C

Pool	Amount (Pg)
Ocean	
Intermediate and deep ocean	38,100
Surface ocean	1,020
Dissolved organic carbon	700
Surface sediment	150
Marine biota	3
Soil (1 m depth)	
SOC	1,505
SIC	940
Vegetation	620
Atmosphere	760

Source: Kennel,[8] Lal,[3] and Schlesinger.[9]

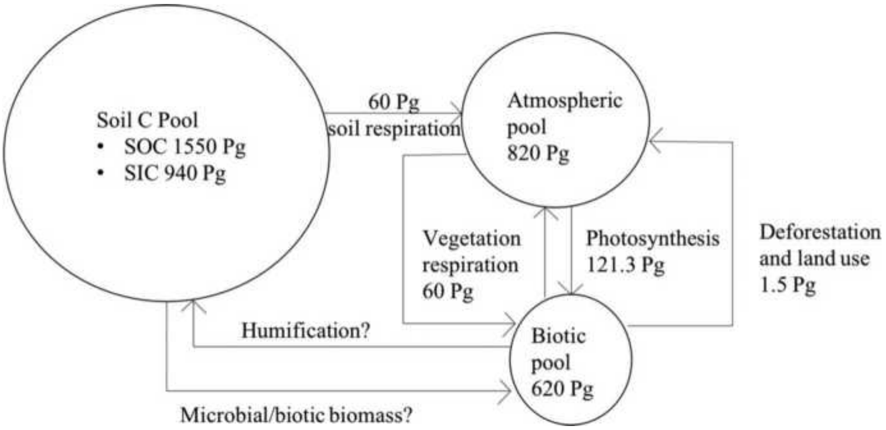


FIGURE 1 Interactive effects among the three C pools (soils, biota, and atmosphere) have led to a steady increase in the atmospheric pool at the expense of the soil and biotic pools. Values in the arrow indicate direction and magnitude of the known fluxes (Pg/yr), while the question mark indicates the lack of information.
Source: Lal,[3] Jansson et al.,[2] and Le Quéré et al.[7]

that is intimately connected with the projected global warming or the accelerated greenhouse effect. Change in soil C pool by 1 Pg is equivalent to 0.47 ppm change in atmospheric concentration of CO_2.[6]

These global pools are interconnected through fluxes of C (Figure 1). Sources and sinks of carbon for the decade of 2008–2017 are shown in Table 2. The two principal sources are fossil-fuel combustion (9.4 Pg/yr) and deforestation and soil disturbance (1.5 Pg/yr). The two proven sinks are atmosphere (4.7 Pg/yr) and ocean (2.4 Pg/yr). The unknown sink for the remaining 3.2 Pg/yr is most likely the world soils and other terrestrial ecosystems.[7] The exchanges occur between (1) world biota and the atmosphere with an annual flux of 121.3 Pg photosynthesized by plants and returned to the atmosphere in equal amounts through plant and soil respiration; and (2) oceans and the atmosphere amounting to an annual net increase of the ocean sink by about 2.4 Pg/yr.[7]

There has been a steady increase in the atmospheric concentration of greenhouse gases (GHGs—CO_2, CH_4, and N_2O) since the beginning of agriculture,[10] but more strongly with the onset of the industrial revolution in the 1850s. This increase is primarily due to human activities, e.g., industry-based fossil-fuel

TABLE 2 Sources and Sinks of Atmospheric CO_2 for the Decade of 2008–2017

Source/Sink	Flux (Pg C/yr)
Source	
Fossil-fuel combustion	9.4, 0.5
Deforestation, land use, and world soils	1.5, 0.7
Total	10.9
Sinks	
Increase in atmospheric concentration	4.7, 0.02
Ocean uptake	2.4, 0.5
Total	7.1
Land	3.2, 0.8
Imbalance	0.5

Source: Le Quéré et al.[7]

consumption, and agriculture-based deforestation and conversion to cultivable land involving intensive use of fertilizers and manures. The CO_2 concentration in the atmosphere has increased from 280 ppmv in the 1850s to about 405.5 ppmv in 2017, and is increasing at the rate of 0.55% per year.[11]

Agricultural Activities and Greenhouse Gas Emissions

GHGs are emitted by numerous agricultural activities (Figure 2). Agricultural sources of CO_2 are deforestation, land use conversion, biomass burning, soil plowing and other mechanical disturbances of the topsoil, and fossil-fuel consumption. Among principal sources of CH_4 are enteric fermentation in ruminant animals, cultivation of rice paddies, anaerobiosis in poorly drained soils, and biomass burning. Atmospheric concentration of CH_4 has increased from 722 ppb during the pre-industrial era to 1,859 ppb in 2017, and is increasing at the rate of 0.7 ppb/yr.[11] Use of nitrogenous fertilizers and manures is the principal source of N_2O emission. Concentration of N_2O in the atmosphere has increased from 270 ppb in pre-industrial era to 330 ppb in 2017, and is increasing at the rate of 0.9 ppb/yr.[11] Soil management plays a major role in the type and amount of emission of GHGs.[4]

The rate of oxidation or mineralization of the SOC pool depends on its composition and numerous other factors. Factors that increase oxidation are soil disturbance, high soil temperature, lack of protective vegetal cover, and risks of soil erosion and degradation. In contrast, factors that decrease oxidation are minimal soil disturbance, low soil temperature, presence of protective vegetal cover, and high soil resilience. The annual flux of CO_2 from soil to the atmosphere, about 60 Pg, is due to soil respiration and its rate increases with increase in soil temperature. The mean resident time of SOC in soil is about 32 years, but some resistance humic substances can stay in soil for thousands of years. Losses of SOC pool by oxidation are accentuated by soil erosion of water and wind,[12–14] and other degradative processes.

Introduction of mechanized agriculture (e.g., tractor-based plowing and other farm operations causing soil disturbance) has accentuated losses of SOC pool from cultivated soils. Some soils lose as much as 25% of their SOC pool within 1 year after deforestation and cultivation.[15] The rate of loss of SOC

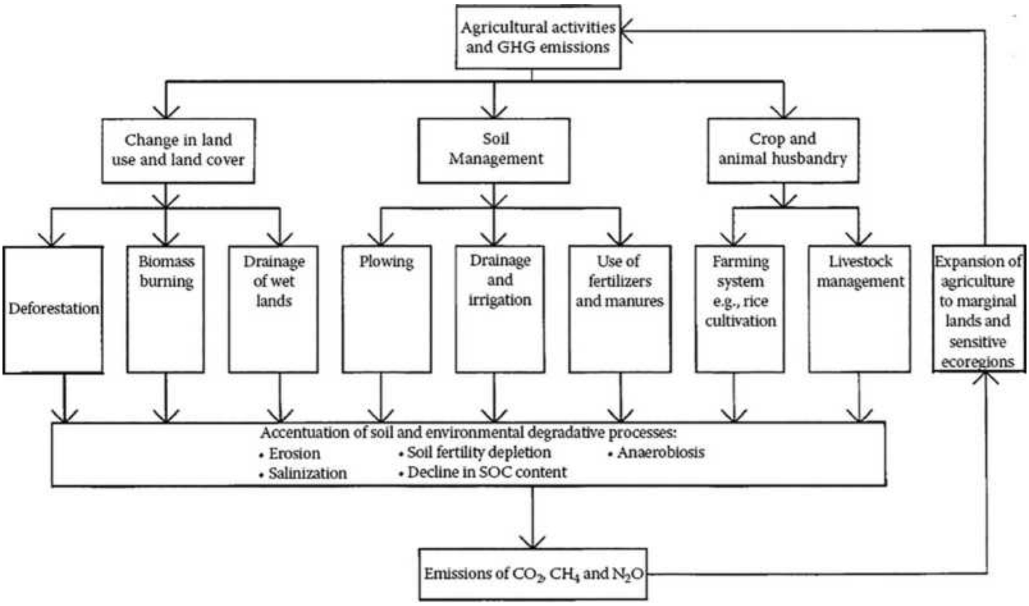


FIGURE 2 Agricultural activities and emission of GHG from soil to the atmosphere.

content is more in the tropics than that in the temperate climate. Some tropical soils, cultivated for millennia in Asia, have extremely low levels of remaining SOC content (as low as 0.2%). These soils have probably lost 75%–80% of their antecedent SOC content. It is likely that as much as 135 Pg of C in the atmospheric pool has been contributed by world soils.[3]

In addition to being a source of atmospheric C (as CO_2, CH_4, or CO), world soils are also a major source of N_2O. The relative contribution of N_2O has also increased because of the intensive use of nitrogenous fertilizers since World War II. The magnitude of N_2O emission varies among soils, type of nitrogenous fertilizer used (formulation), climate, land use, and soil and crop management.[16]

Potential of World Soils as C Sink

The SOC in most agricultural soils is below their potential capacity. Assuming a considerable part of atmospheric CO_2 has come from world soils, it is possible to reverse the trend and store C back into the soil. Increase in SOC content, through change in land use and farming practices, could slow the rate of increase of atmospheric CO_2 and provide a negative feedback to the greenhouse effect. Because of their vast pool, it is very likely that terrestrial ecosystems are the unknown sink for 3.2 Pg of C/yr (Table 2), of which soils are a major component.[3]

Principal mechanisms by which soils can be used as an effective sink for some of the atmospheric CO_2 through increase in SOC content include (1) conversion of biomass into resistant humic substances with long turnover time, (2) aggregation or formation of stable organo-mineral complexes that render SOC inaccessible to microbes, and (3) deep incorporation of SOC below the plow layer through deep root system development and translocation by leaching. These mechanisms can be set-in-motion through adoption of proper land use, choice of appropriate farming/cropping systems, and use of judicious soil and crop management practices.[3,4]

The potential of C sequestration in world soils depends on the current land use, present level of SOC content, the initial SOC content prior to soil disturbance by deforestation and conversion to other land uses, climatic or ecoregional characteristics, and farming systems. Among all land uses, the largest potential increase in SOC content is in arable lands. If the SOC content of the world's agricultural land can be increased at the low rate of 0.04% per year to 40 cm depth, which may be difficult to achieve for soils of arid and semiarid regions, the C sequestration potential of such a possibility is about 2.5 Pg/yr (Eq. 1). These calculations, based on the assumption that the mean bulk density of soils is 1.25 Mg/m^3, are mere examples of what may be possible. The estimates thus computed are not indicative of what is happening now nor what may happen in the future.

Thus, an example of the estimation of C sequestration potential of $5{,}000 \times 10^6$ ha of world agricultural land to 40 cm depth.

Managing World Soils for Mitigation of the Greenhouse Effect

There are several soil management practices that can be used to increase SOC content and reduce the risk of accelerated greenhouse effect (Figure 3). These practices can be grouped under two broad categories: (1) those that decrease losses due to soil degradative processes and (2) those that increase SOC content through soil restorative processes (Table 3). The principal strategy is of agricultural intensification through adoption of appropriate land use and recommended agricultural practices.

A tremendous potential of mitigation of the greenhouse effect exists through restoration of degraded and drastically disturbed soils. It is estimated that the area of the degraded soils of the world is about 2.0 billion ha.[23] These soils include the presently cultivated land area, pastoral land areas, and other managed and natural ecosystems. Restoration to improve soil quality can lead to an increase in ground biomass production (both above and below the ground), increase in SOC content, and strengthening of a negative feedback to the greenhouse effect. An example of calculating the sequestration potential of restoration of degraded soils is shown in Eq. (2).

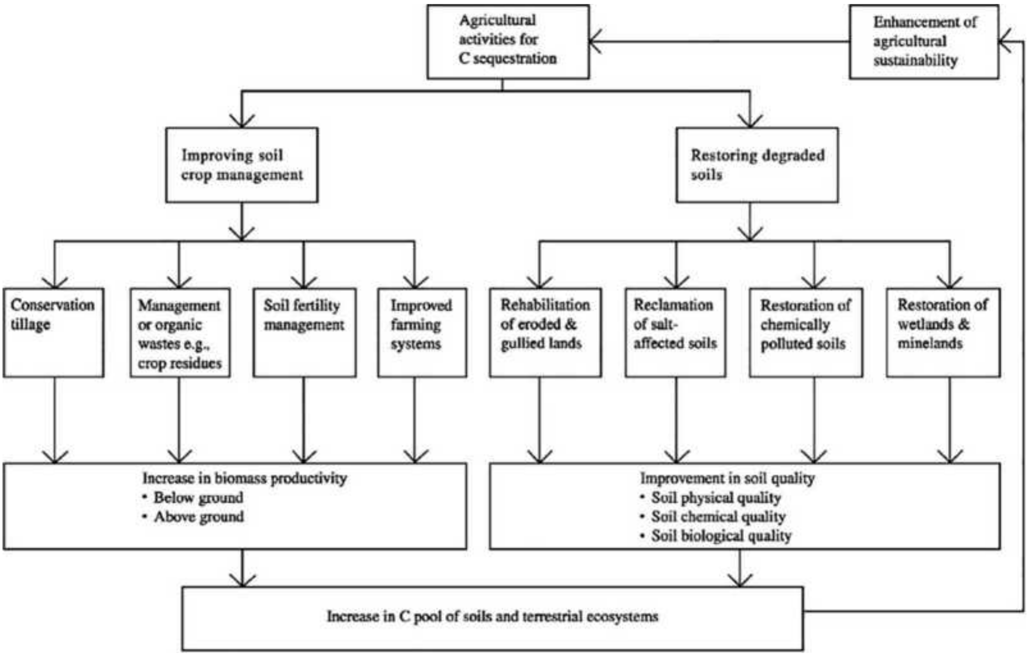


FIGURE 3 Agricultural activities with a potential of sequester C in world soils that decrease the risks of the greenhouse effect.

TABLE 3 Strategies for Increasing SOC Content for Mitigation of the Greenhouse Effect

Practices That Decrease Loss of SOC Content	Practices That Enhance SOC Content
Soil erosion control techniques	Improved soil fertility and integrated nutrient management
1. Conservation agriculture	1. Precision farming
Residue management Terraces and vegetable contour barriers Improved cropping systems Afforestation	Nutrient cycling Application of biosolids and biological nitrogen fixation
2. Decreasing losses of organic carbon through erosion, mineralization, and leaching	2. Restoration of degraded soils Eroded soils Mineland and drastically disturbed soils Salt-affected soils
	3. Water management (controlled drainage and drip subirrigation)

C sequestration potential of restoration of degraded

(2)

These calculations are based on the assumption that mean soil bulk density to 0.4 m depth is 1.25 Mg/m^3. It is also assumed that the potential, 1 Pg/yr of restoring degraded soils combined with 2.5 Pg C/yr of managing all agricultural land at 3.5 Pg C/yr, can be realized over 25–50 years. Realization of this potential requires development of strict land care ethics and global soil policy that must be strictly adhered to. While attempts at decreasing the emissions due to fossil-fuel combustion must be continued, it is important that the vast potential of world soils in C sequestration may be realizable through adoption of known and proven technologies. The attainable potential of world soils for C sequestration is 0.6–1.2 Pg C/yr.[4]

Conclusions

World soils constitute a major terrestrial C pool, and even a small change in the C pool of world soils can lead to a substantial change in the atmospheric C pool. A considerable increase in atmospheric C pool, since the industrial revolution, has come from the soil and biotic pools. Agricultural activities that lead to emissions of CO_2 and CH_4 from soil to the atmosphere include deforestation, biomass burning, and plowing. Gaseous emissions are accentuated by soil degradative processes including accelerated erosion and structural decline.

Conversion to an appropriate land use and adoption of judicious management can reverse the trend by making world soils a major sink for C. The SOC content of agricultural soils can be maintained or enhanced by use of conservation agriculture, application of crop residue mulch, and judicious use of chemical fertilizers and organic amendments which enhance soil quality. Total potential of agricultural soils (~5 billion ha) and degraded soils (~2 billion ha) of 3.5 Pg C/yr, merely an example of calculation, necessitates land use planning and the appropriate policy implications.

References

1. Monger, C., Khresat, S., Kraimer, R.A., Cole, D.R. Sequestration of inorganic carbon in soil and groundwater. *Geology* **2015**, *43*, 375–378.
2. Jansson, C., Wullschleger, S.D., Kalluri, U.C., Tuskan, G.A. Phytosequestration: carbon biosequestration by plants and the prospects of genetic engineering. *Bioscience* **2010**, *60*(9), 685–696.
3. Lal, R. Digging deeper: a wholistic perspective of factors affecting SOC sequestration. *Glob. Change Biol.* **2018**, *24*(8) doi: 10.1111/gcb.14054.
4. Lal, R. Soil carbon sequestration impacts on global climate change and food security. *Science* **2004**, *304*, 1623–1627.
5. Jungkunst, H.F., Krüger, J.P., Heitkamp, F., Erasmi, S., Glatzel, S., Fiedler, S., Lal, R. Accounting more precisely for peat and other soil carbon resources. In R. Lal et al. (Eds) *Recarbonization of the Biosphere*. Springer: Dordrecht, Holland, 2012; 127–157.
6. Lal, R., Smith, P., Jungkunst, H.F., Mitsch, W., et al. The carbon sequestration potential of terrestrial ecosystems. *J. Soil Water Conserv.* **2018**, *73*, 145A–152A.
7. Le Quéré, C., Andrew, R.M., Friedlingstein, P., Sitch, S. Hauck, J., et al. Global carbon budget 2018. *Earth Syst. Sci. Data* **2018**, *10*, 2141–2194.
8. Kennel, C. *UC Revalle Program on Climate, Science and Policy*. Scripps Institution of Oceanography: La Jolla, CA, 2000.
9. Schlesinger, W.H. *Biogeochemistry: analysis of global change*. Academic Press: San Diego, CA, 1991.
10. Ruddiman, W.F. The anthropogenic greenhouse era began thousands of years ago. *Clim. Change* **2003**, *61*, 261–283.
11. WMO. *Greenhouse Gas Bulletin*. World Meteorological Organization: Geneva, 2018.
12. Lal, R. Global soil erosion by water and carbon dynamics. In R. Lal, J. Kimble, E. Levine, B.A. Stewart (Eds) *Soils and Global Change*. CRC/Lewis Publishers: Boca Raton, FL, 1995; 131–141.
13. Lal, R. Soil erosion and the global carbon budget. *Env. Intl.* **2003**, *29*, 437–450.
14. Lal, R. Soil erosion and carbon dynamics. *Soil Tillage Res.* **2005**, *81*, 137–142.
15. Lal, R. Deforestation and land use effects on soil degradation and rehabilitation in Western Nigeria. III. Runoff, soil erosion and nutrient loss. *Land Degrad. Dev.* **1996**, *7*, 99–119.
16. Ussiri, D., Lal, R. *Soil Emissions of N_2O and Its Mitigation*. Springer Verlag: Holland, 2013.

21

Integrated Farming Systems

John Holland

Introduction

The system of agricultural production has undergone immense changes in the past three decades.[1] Crop yields have increased through a combination of plant breeding, intensive use of pesticides and inorganic fertilizer, and greater mechanization. Farming systems based on high inputs of agrochemicals and intensive tillage are now recognized as being unsustainable because they cause soil erosion, soil nutrient depletion, environmental pollution, increased pest problems, and public health hazards. In addition the profitability of cereal and oil crop production also declined at the end of the 20th century forcing farmers to examine ways of reducing their production costs.[1] As a consequence farming is facing economic, political, environmental, and social pressures that are forcing changes in the methods of crop production. In response, a range of different production systems has evolved that try to address these issues. The most extreme are the organic or biological systems in which agrochemical use is almost eliminated. Various systems have been defined that are between the extremes, such as integrated farming systems, and these aim to use lower inputs and as such are regarded as being more sustainable.[2] Many definitions exist for integrated farming but the most appropriate is all encompassed in the description by El Titi[3] "*an holistic pattern of land use, which integrates natural regulation processes into farming activities to achieve a maximum replacement of off-farm inputs and to sustain farm income.*" Integrated farming therefore incorporates the principles of integrated crop management (ICM) and integrated pest management (IPM) but takes a long-term, whole-farm approach that considers all aspects of crop production and land management.[2] The emphasis is on preserving farm profitability by optimizing inputs, although consequently there may be ecological benefits and overall greater sustainability.[4]

What Is Forcing the Changes?

The driving forces where integrated farming has been adopted vary between countries and farmers. The main influences are outlined below.

Legislation

The withdrawal of many pesticides (e.g., in Denmark) has led to some farmers having no option but to adopt integrated or organic systems. Some countries (e.g., the Netherlands) have set targets for national conversion to sustainable agriculture based on integrated farming.[5]

Economic Pressure

Recent declines in cereal and oil seed values and price support has made European farmers look closely at their variable and fixed production costs and therefore a more integrated approach. Agrienvironment payments encourage better habitat management. Cross compliance, modulation, etc., if adopted may force more widespread changes.

Green Marketing Incentive

Crop assurance schemes may include integrated practices. These may provide a premium over conventionally produced food or are a requirement for the larger supermarket retailers.[5]

Environmental

The widespread decline in farmland wildlife is now recognized by governments, farmers, environmental organizations, and the general public.[1,6] To mitigate some of these changes farmers are adopting some of the principles of integrated farming, especially with respect to pesticide use and management of their ecological infrastructure.

Agronomic

When pesticide resistance reaches such levels that chemical control is no longer possible or economical then farmers have to look for alternative solutions.[7]

Development of Integrated Farming

The integrated approach incorporates all aspects of farm management and therefore requires knowledge of the interactions between components in the system, before the concept can be promoted to farmers. In many countries, long-term farm studies have been conducted to develop an integrated system and in some cases economic and environmental comparisons were made with the conventional approach.[4,8] Overall integrated farming proved equally profitable.[4,8] Lower yields could be expected, but this loss was compensated for by using lower inputs of agrochemicals and energy. In the United Kingdom results from nine farm-scale studies in which integrated and conventional farming was compared were pooled for analysis. In the integrated system substantial reductions in inputs were achieved without losing crop quality (Table 1).

The reduction in agrochemical inputs is likely, but as yet not satisfactorily proven to reduce the risk of pollution and toxic or sublethal effects to nontarget organisms. Nevertheless, some environmental benefits were attained.[1,6] Some studies reported greater numbers and diversity of soil fauna where an

TABLE 1 Integrated Farming Inputs and Their Costs as a Percentage of Conventional Systems

Input	% Change in Quantity of Use	% Change in Cost
Fungicides	−50	−41
Herbicides	−42	−32
Insecticides	−40	−42
Fertilizers	−18	−15
Seeds	+10	+12
Operational	–	−9

Source: Adapted from Integrated Farming: Agricultural Research into Practice; MAFF Publications: London, 1998.

integrated system was used, although this was not always apparent. Beneficial insects vary naturally in their abundance between years, fields, and different crops and these fluctuations were usually greater than those between the two farming systems. In small-scale (often <5 ha) trials conventional inputs appeared to have few long-term, direct effects on nontarget arthropods, however, farm-scale monitoring in the United Kingdom for the past 20 years has indicated that many arthropods have declined since the advent of intensive farming and are showing little sign of recovering.[1] In all farming system experiments weed, disease, and insect pest infestation levels were routinely monitored to aid crop protection decisions. Most studies reported lower disease and insect pest incidence, but higher weed burdens in the integrated system, although rigorous experimental appraisals were seldom given.

Principles of Integrated Farming

Arable crops are complex ecological systems[1] and consequently a blueprint for their management using an integrated approach cannot be applied universally. Each farm, crop, and even each field must be considered separately based upon the previous history, current crop monitoring, and the farmers' individual objectives. There are, however, some broad principles that apply.[2,3,5,7,9]

Multifunctional Crop Rotation

Rotation is the core of any integrated system with a minimum of four different crops, although a longer rotation may add more opportunities to build soil fertility and prevent disease and pest carryover. Although the choice of crops will depend on locality, markets, and prices, there are some broad principles. Crops that build soil fertility (leguminous crops and those with high inputs of organic matter, e.g., grasses) must be alternated with those that reduce soil physical and chemical fertility (root crops with low input of organic matter, e.g., potatoes and sugar beet). The sequence of each crop can also be used to ensure there are opportunities to control grass weeds in broadleaved crops and that pest organisms are kept below economic thresholds by ensuring that no single type or group of crops are grown in succession. There may have to be compromises and the rotation chosen will reflect the economic, agronomic, or environmental priorities of individual farmers.

Integrated Nutrient Management

Inputs of inorganic fertilizers are essential to maintain yield and quality but must be balanced to ensure supply does not exceed demand by more than 10%. Excessive nutrient supply encourages weeds, pests, and diseases and causes pollution. Application should maximize uptake by the crop and minimize loss to the environment. Where possible inorganic fertilizer should be replaced with animal and green manures. This has many benefits:

Improved soil nutrient supply and physical structure
Encouragement of beneficial organisms such as predatory invertebrates and microbials
Lower cost
Savings in nonrenewable resources

Optimizing nutrient supply through judicious use of off-farm inputs, organic manures wherever possible, and choice of a nutrient-supplying crop rotation will generate economic benefits for the farmer and the environment, especially water quality.

Minimum Soil Cultivation

Soil cultivations require high inputs of energy and machinery, are damaging to beneficial soil fauna, and encourage erosion and nutrient leaching. Adopting a soil cultivation system to minimize these effects is an essential component of any integrated system. The type and intensity of cultivation must be chosen according to soil type and agronomic concerns (e.g., weed species composition). Conservation tillage/noninversion tillage in combination with mulching should be maximized thereby concentrating soil organic matter in the upper soil layers. This improves soil structure and workability, decreases erosion and leaching, improves drought tolerance, and creates more favorable conditions for beneficial predatory invertebrates, soil microorganisms, and earthworms. Weeds may be encouraged but these may be overcome by utilizing the benefits of rotation and targeted weed control. Considerable cost savings may be made and this is the driving force behind the recent interest in the noninversion tillage in Western Europe.

Integrated Crop Management

Integrated crop management encompasses IPM but extends the principles to the control of weeds and diseases. The approach maximizes natural regulatory mechanisms through adoption of the above components and uses plant resistance, time of planting, economic thresholds, and the most environmentally safe pesticides. This requires more intensive crop monitoring and in some cases greater tolerance of pests. Mechanical weed control combined with low doses of herbicide may be appropriate. Encouragement of pest predators through provision of suitable habitat and trap cropping is also recommended.

Ecological Infrastructure Management

This is the provision of a suitable proportion of noncrop or ecological areas that if correctly managed can

Provide a source of beneficial species for pest control;
Act as buffer zones along ecologically sensitive areas such as watercourses preventing nutrient loss and soil erosion;
Provide a refuge from which reinvasion may occur after any adverse agricultural operations;
Prevent weed ingress into fields; and
Act as a source of biodiversity.

These areas may be field margins, hedgerows, fence lines, grass, or herbaceous strips arranged to ensure the maximum connectivity while keeping individual field size to a minimum. A minimum area of 5% of cropped land is recommended[3] although the ecological gains will be dependent on the quality of these areas. Research has identified the potential of different ecological areas to provide pest control and act as buffer zones.

Implementing an Integrated System

The approach requires considerable knowledge and commitment by the farmer. A whole farm evaluation helps the transition from high to lower inputs. On-farm trials may be needed to ascertain the most effective system. Extension services are needed for the transition period and afterward if it is to remain successful.[2] New technology, such as decision support and precision farming techniques, are advocated.

References

1. *The Ecology of Temperate Cereal Fields*; Carter, N., Darbyshire, J.F., Potts, G.R., Firbank, L.G., Eds.; Blackwell Scientific Publications: Oxford, U.K., 1991; 469.
2. Vereijken, P. A methodic way to more sustainable farming systems. Neth. J. Agric. Sci. **1992**, *40*, 209–223.
3. El Titi, A.; Boller, E.F.; Gendrier, J.P. *Integrated Production—Principles and Technical Guidelines*; IOBC/WPRS Bulletin 16 (1), IOBC/WPRS: France, 1993; 96.
4. Jordan, V.W.L. The development of integrated arable production systems to meet potential economic and environmental requirements. Outlook on Agric. **1998**, *27* (3), 145–151.
5. Proost, J.; Matteson, P. Integrated farming in the Netherlands: flirtation or solid change. Outlook on Agric. **1997**, *26* (2), 87–94.
6. *Ecology and Integrated Farming Systems*; Greaves, M.P., Glen, D.M., Eds.; John Wiley and Sons: Chichester, U.K., 1995; 329.
7. Tiwari, P.N. Integrated farming research for sustaining food production. J. Nucl. Agric. Biol. **1993**, *22* (1), 1–13.
8. Holland, J.M.; Frampton, G.K.; Çilgi, T.; Wratten, S.D. Arable acronyms analysed—a review of integrated arable farming systems in Western Europe. Ann. Appl. Biol. **1994**, *125* (2), 399–448.
9. Integrated Farming Systems. Available at http://www.gct.org.uk/index.html (accessed July 2000).

22

Organic Soil Amendments

Philip Oduor-Owino

Introduction

Modern agriculture is faced with the challenge of becoming more productive and yet more sustainable. One important goal toward this end is to boost crop production through proper management of weeds, insect pests, and plant pathogens. These management tactics must be implemented without adversely affecting the ecosystem. Therefore, there is a need to change from the use of pesticides to safer pest management practices, which can be adopted in integrated pest management (IPM) programs. The use of organic soil amendments for the control of plant pathogens and/or pests may provide a viable alternative.[1–3]

Organic Soil Amendments and Their Mechanisms of Action

Amending soil with organic matter such as chitin, oil cakes, compost, animal manures, and other industrial by-products in pest management studies is well recognized.[4,5] However, effects of these materials on disease development are not clear, and have been attributed, in part, to the factors discussed below.

Impacts of Organic Soil Amendments on Plant Health and Weeds

Soil amendments improve plant growth by enhancing plant nutrition.[1] The levels of nitrogen, phosphorus, potassium, and other essential elements are increased when organic matter is added to soil and is associated with better crop performance.[1] Changes in physical characteristics of soil may also enhance plant growth and the associated weeds, an attribute that should be utilized in disease management. Healthy plants produce higher yields, compete with weeds, and tolerate fungal, nematode, and insect damage better than unthrifty plants.[1,6]

Organic Soil Amendments and Plant Resistance

Materials such as oil cakes and sawdust have high phenolic content and alter the attractiveness of host plants to nematodes.[7] For example, seed treatment with ground oil cakes boosts plant resistance to *Tylenchulus semipenetrans* Cobb and root-knot nematodes due to increased levels of phenols in treated citrus and tomato roots, respectively.[4,7] In contrast, organic materials from *Tithonia diversifolia* (Hems) and chicken manure increase the severity of dry root-rot of French bean (*Phaseolus vulgaris* L.cv. Monel) caused by *Fusarium solani* f.sp. *phaseoli* (Mart) Sacc. (Table 1). This has been attributed to the formation of stimulatory ammonium compounds during decomposition.[8]

Release of Compounds Toxic to Insects and Plant Pathogens

Some organic materials release insecticidal, nematoxic, and/or fungitoxic chemicals during decomposition, for instance, neem oil and neem cake powder from the neem tree, *Azadirachtin indica.* A. Juss contains the limonoid azadirachtin, which is nematoxic and insecticidal in nature.[6,9] The black bean aphid, *Aphis fabae* (Scop), has been successfully controlled by this product (Table 2). The nematicidal activity of marigolds (*Tagetes* spp.) and castor (*Ricinus* spp.) has also been recognized, but in this case

TABLE 1 Effect of Cowdung (Cd) and Organic Soil Amendments from *T. diversifolia* (Td) and Their Combination with Metalaxyl (Mt) on Plant Growth and Dry Root Rot of French Beans 72 Days after Planting in Soils Inoculated with *F. solani* f.sp. *phaseoli* (Fs)

Soil Treatment	Mean Shoot Dry Weight (g)	Mean Root Dry Weight (g)	Mean[a] L.D.E.T (mm)	Mean Root Rot Index[b] (1–9)	Mean Number Pods per Plant	Mean Dry Weight of 100 Seeds
Cd + Fs	4.84b[c]	1.398a	30.0bc	3.13c	22.0a	25.05a
Cd + Mt + Fs	4.26b	1.166b	50.0abc	5.41b	21.0a	19.02b
Td + Fs	3.36b	0.232b	62.1ab	6.23b	5.0b	23.57b
Td + Mt + Fs	1.03c	0.014e	90.0ab	8.00a	0.0c	0.00c
Fs alone	3.12b	0.167e	97.5a	7.25a	4.0bc	21.37a
+ Cd; No Fs	7.84a	0.733c	6.11c	1.00d	22.0a	32.17a
Td alone	2.31c	0.796c	5.43c	1.00d	4.0bc	32.85a

Source: Wagichunge.[14]

[a] Length of discoured tissue (mm) (L.D.E.T).

[b] Mean root-rot index was based on a 0–10 rating scale, where, 0 = no symptoms and 10 = whole root system decayed.

[c] Numbers are means of five replicates. Means followed by the same letter within the same column are not significantly different at $P = 0.05$ level by Duncan's Multiple Range Test (DMRT).

TABLE 2 Weekly Mean Aphid Scores on French Beans Following Treatments with the Insecticide Gaucho Neem Kernel (NKCP) and Different Neem Products

Treatment	Week 1	Week 2	Week 3	Week 4	Mean No. of Pods/Plant (Week 4)
Karate (2 mL/L)	0.1c[a]	0.1c	0.6c	1.2c	1.03c
Neem oil EC (3%)	0.1c	0.5c	0.5c	1.3c	9.3c
NKCP/WE (50 g/L)	1.1b	1.0c	0.7c	1.6c	6.2c
Gaucho (8 mL/kg)	0.9b	1.5b	2.1b	3.6b	5.0b
Control	3.4a	4.9a	6.7a	7.0a	3.3a

Source: Maundu.[6]

[a] Numbers are means of 10 replicates. Means followed by the same letter within columns do not differ significantly at $P = 0.05$ by Duncan's Multiple Range Test (DMRT).

the toxic principles are Polythienyls and ricin, respectively.[4,7] Antimicrobial chemicals such as nitrites and hydrogen sulfide are also produced during decomposition and play an important role in disease control. Unfortunately, various changes in quality and quantity of these chemicals occur over time, making it difficult to obtain more than circumstantial evidence that any one compound is responsible for disease suppression.[2]

Stimulation of Antagonistic Microorganisms

The hypothesis that organic soil amendments stimulate the activity of antagonistic microorganisms was proposed over 50 years ago.[5] When organic matter is added to soil, a sequence of microbial changes is initiated, none of which should be viewed in isolation. It is possible that the ability of nematophagous fungi such as *Paecilomyces lilacinus* Thom. (Samson) and *Verticillium chlamydosporium* (Goddard) to destroy/parasitize eggs of root-knot nematodes is stimulated by soil amendments.[2,10] Egg parasitism of up to 37% has been achieved with organic matter from castor plant or chicken manure (Table 3). Besides egg parasitism, the diverse range of microorganisms in amended soils competes with nematodes and other invertebrate pests for space and oxygen, thereby creating unfavorable anaerobic microsites in the soil. Bacteria such as *Streptomyces anulatus* (Beijerinck) Waksman, the collembolan, and *Entomobyroides dissimilis* (Moniez) are good examples.[5] *Armillaria* root rot of fruit and forest trees, caused by *Armillaria mellea* Vahl ex.fr, is minimized using coffee pulp that stimulates the antagonistic effects of *Trichoderma viride* link ex. Fries against a wide range of *Armillaria* spp.[11]

In conclusion, it is evident that various activities of soil microorganisms contribute significantly to the detrimental effects of organic matter on plant pathogens. However, it is difficult to determine whether any one activity or group of organisms is directly responsible for the suppression of specific diseases. The evidence available remains largely circumstantial.

TABLE 3 Effect of Organic Soil Amendments, and Soil Treatments with Captafol or Aldicarb on the Parasitism (%) of *Meloidogyne javanica* Eggs with *P. lilacinus* and Growth of Tomato cv Money Maker Plants

Soil Treatment[a]	Egg Parasitism (%)	Juveniles/300 mL Soil	Gall Index[b] (0–4)	Shoot Height (cm)	Shoot Dry Weight (g)
Tag + Mj	0.5f	188c	2.3c	36.0f[c]	2.8g
Dat + Mj	0.7f	189c	2.5c	35.0g	2.9g
Ric + Mj	1.2f	207c	2.4c	32.1h	3.5e
Ch.M + Mj	1.0f	217c	1.8d	46.8b	5.0b
Ald + Mj	0.8f	18e	1.3e	49.8a	5.4a
Cap + Mj	0.0f	521a	3.8a	32.5h	2.9g
F + Mj	21.2e	438b	3.5b	28.4T	1.71
F + cap + Mj	1.3f	501a	3.6ab	30.2hi	2.6h
F + Ald + Mj	26.2d	10e	0.5f	46.6b	5.1b
F + Tag + Mj	30.9b	206c	2.0d	39.2e	3.0g
F + Dat + Mj	28.4c	201c	2.0d	42.8c	3.2f
F + Ric + Mj	37.2a	187c	2.5c	39.7e	3.6e
F + Ch.M + Mj	37.3a	147d	1.8d	42.0d	4.2d
Mj. “Only”	0.5f	425b	3.4b	23.5j	1.4j
Soil “Only”	0.0f	0.0e	0.0f	44.9bc	4.5c

Source: Oduor-Owino.[1]

[a] F = fungus; Mj = *M. javanica*; Cap = Captafol; Ald = aldicarb; Tag = *Tagetus minuta*; Dat = *Datura stramonium;* Ric = *Ricinus communis*; and Ch.M = chicken manure.

[b] Gall index was based on a 0–4 rating scale, where 0 = no galls and 4 = 76%–100% of the root system galled.

[c] Numbers are means of 10 replicates. Means followed by different letters within a column are significantly different ($P = 0.05$) according to Duncan's Multiple Range Test.

Use of Organic Soil Amendments in the 21st Century

Studies on the efficacy of organic soil amendments against plant pathogens should be intensified worldwide. Organic plant materials such as chitin, compost, and oil cakes have great nematode control potential but have remained unutilized in biological control systems due to inadequate and inconsistent information on their efficacy and compatibility with antagonistic microorganisms.[10,12] It is not known if these organic materials and fungal antagonists/predators can successfully be integrated into the same pest control systems. The future challenge in this case is to determine ways of boosting the antagonistic potential of specific beneficial organisms by using locally available amendments in quantities realistic for broad-scale agricultural use. The complexity of the soil environment may thwart efforts to achieve this, but previous studies[4,5,7] and recent work on the interaction between nematodes and organic soil amendments[1–3,13] suggest that this is a promising area for further research.

Future Concerns

Organic soil amendments have a positive future in pest and disease control.[10,14] However, recent techniques used in the fields of biotechnology and molecular genetics[15] may dominate biological control research with a view of alleviating problems that are presently confronting researchers in an attempt to look for safe pest control alternatives. It is important that scientists, in their eagerness to embrace these new technologies, do not lose sight of the fact that the ultimate objective is the development of environmentally friendly pest control systems that can be applied in the field. We must strike the right balance between theoretical investigations and the more applied biological control studies aimed at developing viable pest management options.

References

1. Oduor-Owino, P. Fungal Parasitism of Root-Knot Nematode Eggs and Effects of Organic Matter, Selected Agrochemicals and Intercropping on the Biological Control of *Meloidogyne javanica* on Tomato. Ph.D. thesis, Kenyatta University: Nairobi, 1996.
2. Oduor-Owino, P.; Waudo, S.W. Effects of delay in planting after application of chicken manure on *Meloidogyne javanica* and *Paecilomyces lilacinus*. Nematol. Mediter. **1996**, *24* (3), 7–11.
3. Oduor-Owino, P.; Sikora, R.A.; Waudo, S.W.; Schuster, R.P. Effects of aldicarb and mixed cropping with *Datura stramonium*, *Ricinus communis* and *Tagetes minuta* on the biological control and integrated management of *Meloidogyne javanica*. Nematologica **1996**, *42* (2), 127–130.
4. Bhattacharya, D.; Goswani, B.K. Comparative efficacy of neem and groundnut oil-cakes with aldicarb against *Meloidogyne incognita* in tomato. Rev. Nematol. **1987**, *10* (1), 467–470.
5. Linford, M.B. Stimulated activity of natural enemies of nematodes. Science **1937**, *85* (1), 123–124.
6. Maundu, M.E. Control of the Black aphid, *Aphis fabae* Scop of Beans Using Neem-Based Pesticides in Kenya. M.Sc. thesis, Kenyatta University, Nairobi, 1999.
7. Bandra, T.; Elgindi, D.M. The relationship between Phenolic content and *Tylenchulus semipenetrans* populations in nitrogen-amended citrus plants. Rev. Nematol. **1979**, *2* (3), 161–164.
8. Waudo, S.W.; Oduor-Owino, P.; Kuria, M. Control of Fusarium wilt of tomatoes using soil amendments. East Afr. Agric. For. J. **1995**, *60* (4), 207–217.
9. Oduor-Owino, P.; Waudo, S.W. Comparative efficacy of nematicides and nematicidal plants on root-knot nematodes. Trop. Agric. **1994**, *71* (4), 272–274.
10. Oduor-Owino, P.; Waudo, S.W.; Sikora, R.A. Biological control of *Meloidogyne javanica* in Kenya. Effect of plant residues, benomyl and decomposition products of mustard (*Brassica campestris*). Nematologica **1993**, *39* (3), 127–134.
11. Onsando, J.M.; Waudo, S.W.; Magambo, M.J.S. A biological control approach to root rot of tea *Armilleria mellea* in Kenya. Tea **1989**, *10* (2), 165–173.

12. Oduor-Owino, P.; Waudo, S.W.; Makhatsa, W.L. Effect of organic amendments on fungal parasitism of *Meloidogyne incognita eggs* and growth of tomato (*Lycopersicon esculentum* Mill) cv money maker. Int. J. Pest Manage. **1993**, *39* (4), 459–461.
13. Oduor-Owino, P.; Waudo, S.W. Medicinal plants of Kenya. Effects of *Meloidogyne incognita* and the growth of Okra. Afro-Asian J. Nematol. **1992**, *2* (1), 64–66.
14. Wagichunge, A.G. Efficacy of Seed-Dressing and Organic Amendments Against Fusarium Root-Rot of French Beans (*Phaseolus vulgaris* L.cv. Monel) in Kenya. M.Sc. thesis, Kenyatta University, Nairobi, Kenya, 2000.
15. Kerr, A. Commercial release of a genetically engineered bacterium for the control of crown gall. Agric. Sci. **1989**, *2* (1), 41–44.

23

Pasturelands, Rangelands, and Other Grazing Social-Ecological Systems

Radost Stanimirova
and Rachael Garrett

Introduction

Approximately 26% of Earth's land (excluding Antarctica) is occupied[1] by grazing systems, making them the largest non-forest land use on the planet (Erb et al. 2017, FAO 2011). A defining feature of livestock production is that ruminants (e.g., bovines, sheep, and goats) spend at least some of their lifetime grazing; however, there are several terms used to refer to large grassy areas depending on the intensity of use: grasslands, rangelands, pasturelands, and grazing lands. Practitioners and researchers usually distinguish among natural grazing lands (also called rangelands) and improved grazing lands (also called pasturelands) (Allen et al. 2011, Garnett et al. 2017). While the term "grasslands" describes ecological communities that are dominated by grasses, the term "grazing lands" refers to the potential or realized use of grasslands to humans. Rangelands are dominated by natural vegetation whose species composition has not been altered to boost livestock productivity (Allen et al. 2011, Garnett et al. 2017)

---

[1] Though the land occupied for grazing may include multiple non-grazing uses as well, just as grazing may occur in areas not primarily devoted to livestock production.

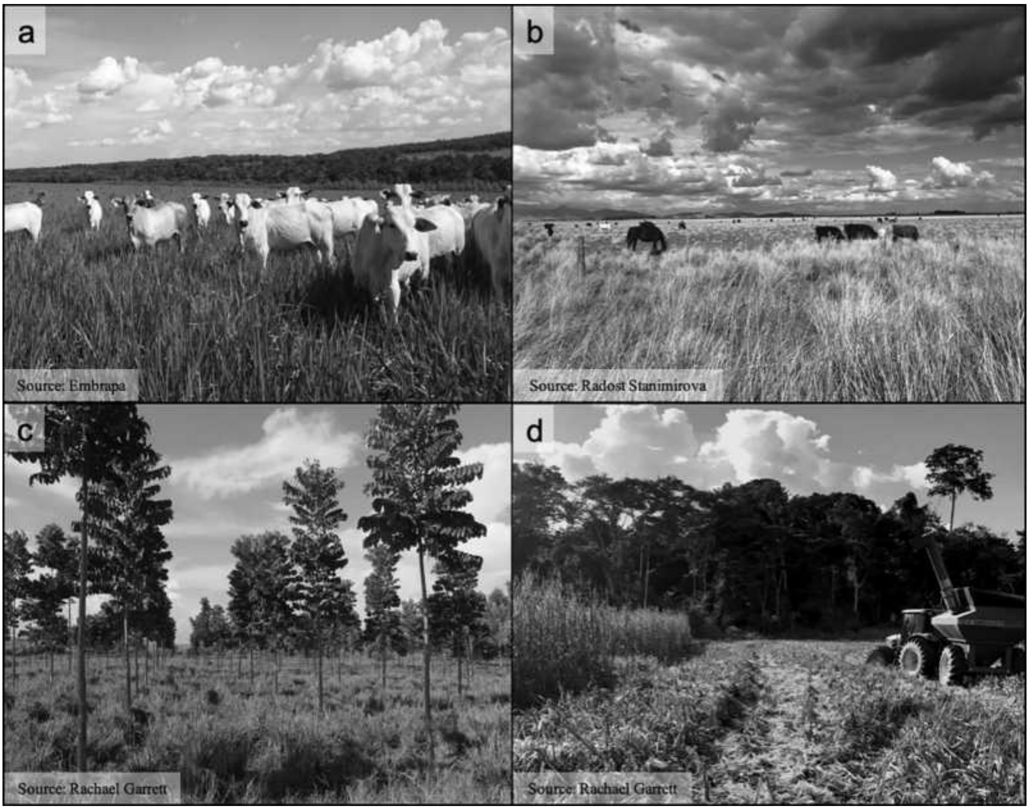


FIGURE 1 Examples of (a) pastureland (i.e., intensively managed planted grassland), (b) rangeland (i.e., extensively managed natural grassland), (c) silvo-pastoral system, and (d) integrated crop-livestock system.

(Figure 1b). Rangelands can be characterized as grasslands, shrub steppe, shrublands, savannas, and deserts that are managed to provision ecosystem services to benefit human well-being. Pasturelands, on the other hand, are more intensively managed by planting nutrient-rich grasses and legumes, or by using fertilizer and other amendments (Allen et al. 2011, Garnett et al. 2017) (Figure 23.1a). Heretofore, we refer to rangelands and pasturelands together as grazing lands or grazing systems.

Ecosystem Services Provided by Grazing Systems

There are four different categories of ecosystem services: provisioning, regulating, supporting, and cultural (MA 2005, Sala et al. 2017). Because grazing systems are heterogeneous, they provide numerous ecosystem services in addition to provisioning feed for livestock, and food and fiber for the world's pastoralists (Sala et al. 2017). In the case of grazing lands, *provisioning services*, which are the products that can be directly harvested and have market value, are freshwater and forage for livestock to produce meat, milk, wool, and leather (Sala and Paruelo 1997, Sala et al. 2017). The main ecosystem *regulating service* in grasslands is soil carbon sequestration. *Supporting services* enable the production of all other ecosystem services such as nutrient cycling, conservation of soils, habitat, and biodiversity (MA 2005, Sala and Paruelo 1997). Lastly, *cultural services*, which are the nonmaterial benefits that humans obtain from grazing lands, include social status, traditional lifestyles, and tourism (Sala et al. 2017). From local-to-global scales, ranchers, commercial entities, and the public sector all benefit from the ecosystem services that are provided by grazing lands.

Provisioning Services

Grasslands are a key feed resource, accounting for close to 50% of feed use in livestock production systems globally (Herrero et al. 2013). While in many parts of the world, grassland provisioning services are directly traded in markets, in other places pastoralists depend on provisioning services for their livelihoods. Specifically, the livestock sector provides income for approximately 1.3 billion people worldwide and food for approximately 800 million food-insecure people (Herrero et al. 2013). Globally, the livestock sector produced approximately 586 million tons of milk, 59 million tons of beef, and 11 million tons of small ruminant meat in 2000 (Herrero et al. 2013). Although most rangeland-based systems occupy less productive lands compared to pasturelands, they provide a key source of livelihood for pastoralists in regions that rely on extensive livestock production. Indeed, rangeland-based systems account for the majority of small ruminant meat production, for local production of beef in Latin America and Oceania, and for small ruminant milk production in sub-Saharan Africa (Herrero et al. 2013).

Regulating Services

Grassland soils contain one-third of the terrestrial soil organic carbon pool because they are geographically extensive and have relatively high per unit area carbon storage (Garnett et al. 2017, White et al. 2000). However, the ability of grasslands to sequester carbon is limited due to variability in many factors (e.g., soil type, nutrient availability) that determine if organic matter is converted to stable belowground carbon. Planting deep-rooted grasses and legumes, adding manure to soils, stimulating forage productivity, and changing the timing and intensity of grazing can potentially increase the rate of soil carbon sequestration (Batjes 2014, Garnett et al. 2017). Good grazing management at the right stocking rate can maintain or even increase soil carbon sequestration by stimulating plant growth, which can lead to an increase in the proportion of organic matter that's partitioned belowground (Batjes 2014, Garnett et al. 2017). Although soil carbon stocks can quickly increase after an improvement in management, especially in degraded grazing lands, the rate of increase progressively declines as the system approaches an equilibrium (Smith 2014). As a result, while managing for land-based carbon sequestration in grazing systems is useful, it only plays a minor role in offsetting greenhouse gas emissions to which the livestock sector itself is a significant contributor.

Supporting Services

Wildlife Habitat and Biodiversity

Maintaining biodiversity in grazing systems is key not only to sustaining grassland flora and fauna but also to delivering all other ecosystem services. Managing for grassland structural heterogeneity, for example, can improve forage availability, enhance ecosystem functionality, and increase habitat availability for different plant, insect, bird, and mammal species (Fuhlendorf et al. 2006, 2017, Tews et al. 2004). There are two aspects of biodiversity that appear to be critical for provision of ecosystem services and resilience to environmental change: functional-group diversity and functional-response diversity (Elmqvist et al. 2003, Folke et al. 2004). While functional-group diversity refers to groups of organisms that perform different functions in an ecosystem (i.e., graze, fix nitrogen), functional-response diversity refers to the ability of species from the same functional group to respond differently to environmental change (i.e., rainfall, grazing). Loss of both types of functional diversity can cause dramatic alterations in ecosystem functioning (Chapin 1997, Duffy 2002, Jackson 2001).

Soil Conservation and Nutrient Cycling

Globally, grassland soils are very diverse and their characteristics and functions are critical determinants of the ecosystems services that grazing systems can provide. Grassland soil structure and biogeochemical cycles are controlled by three major functional groups of organisms: vascular plants, biological soil crusts, and soil microbial communities (Evans et al. 2017). In addition, soil biogeochemical cycles (i.e., water, carbon, and nutrients) vary as a function of precipitation, phenology, disturbance regimes, and management practices (Evans et al. 2017). Rangelands, in particular, are a critical component of global biogeochemical cycles because they are globally extensive and sensitive to changes in resource availability caused by ecosystem drivers (Donohue et al. 2013, Poulter et al. 2014). In grazing lands, the quality and productivity of forage and, in turn, livestock are determined by the efficiency of these nutrient cycles. Managing these ecosystems for soil conservation and nutrient cycling involves controlling livestock density to minimize soil compaction and overgrazing, maintaining plant diversity, planting legumes, and incorporating manure and fertilizers to the soil according to a comprehensive nutrient management plan (Bellows 2001).

Cultural Services

While the provisioning, regulating, and supporting services that grazing lands provide often translate into material benefits to farmers (i.e., food and income), many land use choices are driven by the nonmaterial benefits these systems provide (Garrett et al. 2017). The degree to which different nonmaterial outcomes result in improved well-being is related to cultural preferences. Many farmers choose grazing systems as a livelihood due to the low and less seasonal labor demands relative to most systems of crop production (Hecht 1993, Muchagata and Brown 2003), a desire to pursue a lifestyle independent of markets and governments (Skaggs 2008, van der Ploeg 2010) and as a status symbol (Hoelle 2011, Walker et al. 2000). For example, the persistence of low-income cattle ranching in the Brazilian Amazon and Cerrado has been linked to its cultural importance as a source of social status and a tranquil lifestyle (Garrett et al. 2017, Wilcox 2017). Grazing lands and the biodiversity they harbor also provide aesthetic and spiritual values, including a sense of place, to both inhabitants and visitors (Knight 2002). These cultural values are reflected in the value of tourism activities and land prices, which often exceed the income provided by livestock rearing (Skaggs 2008).

It is important to note that while grassland ecosystems provide cultural services, cultural traditions also shape ecosystem services by determining land use diversity and intensity. In the Swiss Alps, for example, Romanic, German, and Walser villages exhibit very different land use configurations resulting in different levels of plant diversity within similar types of grazing systems (Maurer et al. 2006). Among Old Order Amish farmers in the United States, cultural preferences toward autonomy, social controls on the introduction of new technologies (e.g., synthetic inputs, heavy machinery), and a refusal of government assistance, such as subsidized insurance (Stinner et al. 1989), have helped protect grazing systems within increasingly homogenized cropland landscapes (Garrett In Review, Parker 2013).

Management of Grazing Systems

Climate is a major driver of ecosystem structure and function in grasslands, with near-term precipitation being the dominant climatic control and antecedent precipitation being secondary (Knapp and Smith 2001, Sala et al. 2012, Wu et al. 2015). As a result, the timing and amount of precipitation are key factors that limit soil moisture availability and annual vegetation productivity, and subsequently constrain grazing livestock density (Sala et al. 2012, Sloat et al. 2018, Wilcox et al. 2017). Changes in water availability due to drought, for example, have negative impacts on the provision of grassland ecosystem services, with especially profound response in soils where microbial activity, nutrient availability, and soil carbon sequestration depend on soil moisture (Evans et al. 2017).

In addition to climate variability, vegetation dynamics and productivity in *rangelands* (i.e., *extensively managed systems*) are also driven by livestock density (Briske et al. 2003, Ellis and Swift 1988, Gaitán et al. 2014, Sala et al. 2012, Wilcox et al. 2017). Specifically, while grazing primarily affects the long-term structure and composition of rangelands, precipitation variability primarily affects yearly plant production and forage availability (Briske et al. 2003, Ellis and Swift 1988, Fernandez-Gimenez and Allen-Diaz

1999). In many regions of the world, continuous, unconstrained grazing (i.e., overgrazing) on rangelands with already poor soils or in climates with long dry seasons has led to degradation (decreased water infiltration and surface organic material leading to a loss of soil structure, nutrients, and water availability, and encroachment by invasive species) (King and Hobbs 2006, Perkins and Thomas 1993). Recent sustainable rangeland management approaches have focused on reversing land degradation and restoring overgrazed ecosystems (Bestelmeyer et al. 2012, Briske et al. 2008). Specifically, resilience-based rangeland management aims to maintain desirable ecological states (i.e., grasslands) in order to avoid ecological thresholds that lead to less desirable ecological states (i.e., desert shrubland) (Bestelmeyer et al. 2012, Briske et al. 2008, Elmqvist et al. 2003).

In *pasturelands* (i.e., *intensively managed systems*), climate is also a major driver of productivity. Yet, human interventions such as irrigation, rotational grazing (frequently moving cattle through smaller paddocks to avoid grazing beyond an optimal threshold), applying soil correctives and fertilizer, and intercropping leguminous plant species (e.g., forage peanut) are mechanisms that can outweigh climatic effects on grass productivity while also improving animal nutrition (Condron et al. 2014, Nobilly and Bryant 2013, Valentim and Andrade 2005). Grazing duration and intensity can also be managed to optimize productivity and economic returns. Other important improved grassland systems include integrated crop-livestock systems (planted pastures that are rotated with cropland) (Figure 1d) and silvo-pastoral systems (planted or natural pastures that are intercropped with tree species) (Figure 1c). These systems rely on synergies between the cropland, tree, and grazing land uses to increase nutrient availability and improve soil structure when livestock are kept at low to moderate grazing intensities (Carvalho et al. 2014, Lemaire et al. 2014, Ryschawy et al. 2017). These synergies can reduce needs for external inputs while maintaining or increasing productivity and profitability (Garrett et al. 2017, Gil et al. 2018, Schiere et al. 2002). While mixed and integrated crop-livestock production systems are common in sub-Saharan Africa and Southeast Asia (Robinson et al. 2011), they remain very rare in the Global North and other important grazing regions (Garrett et al. In Review).

Collectively, these management decisions can substantially increase the amount of forage produced, as well as the number of cattle that can be supported, while decreasing the amount of time it takes the cattle to achieve slaughter weight. In Brazil, for example, existing ranching management practices in Brazilian Cerrado are highly extensive, and the current productivity of these pastures is only 32%–34% of its potential (Strassburg et al. 2014). Studies have shown that introducing rotational grazing can double the amount of protein produced per hectare, while integrated crop and livestock systems can increase the amount of protein produced by nearly 700% (Gil et al. 2018).

Traditional approaches to grazing system management have focused on the provision of food and fiber but have often been inadequate in guiding the provision of other ecosystem goods and services (Sala et al. 2017). For example, despite management efforts, rangelands have remained vulnerable to accelerated soil degradation, which impairs their ability to provide both regulating and supporting ecosystem services (Sala et al. 2017). At different temporal and spatial scales, climate variability, land degradation, and the feedbacks between these two processes can reduce the ability of grazing systems to maintain ecosystem services.

Challenges in Grazing Systems

Quantifying the susceptibility of grazing systems to climate variability and land degradation is critical for anticipating grassland feed shortages and for monitoring regions that are vulnerable to desertification and woody encroachment. However, modeling the response of these ecosystems to climate variation, for example, is challenging because they are actively grazed by livestock and managed by humans, which means that their dynamics are influenced by grazing, soil amendments, and precipitation at multiple spatial and temporal scales. The interacting dynamics of climate variability, land degradation, and management shape the demand and provision of ecosystems services of grazing social-ecological systems (Figure 2).

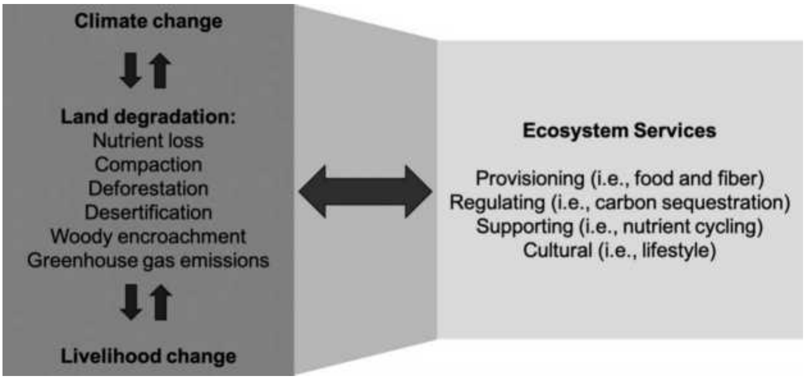


FIGURE 2 Generalized diagram of ecosystem services and challenges in grazing social-ecological systems. Humans and the environment interact to shape the demand and provision of ecosystem goods and services.

Climate Change

In the coming decades, the responses of ecosystem processes to changes in the frequency and intensity of drought are likely to be exacerbated by rising temperatures (Huang et al. 2016, Polley et al. 2017). Recent studies suggest that up to 66% of global land areas are experiencing drying, and precipitation events in arid and semi-arid regions are forecast to become shorter, less frequent, and less widespread (Huang et al. 2016, Reeves et al. 2014). When combined with overgrazing, a warmer and drier climate can ultimately result in decreased productivity and directional shifts in the cover and composition of plant communities in grazing systems (Polley et al. 2017). If realized, then these changes pose a significant threat to the sustainability of these systems, especially in arid and semi-arid regions.

Desertification

Twenty to thirty-five percent of the world's pastures are in poor condition and show signs of degradation, especially when compared to natural grasslands (non-pasture) where direct human impacts are limited (Conant 2010, Garnett et al. 2017). Dryland ecosystems are especially vulnerable to degradation due to multiple biophysical and social stressors that exert pressure at different temporal and spatial scales: climate, land use, economic and institutional factors, national policies, and population growth (Geist and Lambin 2004, Reynolds et al. 2007). When degradation is characterized by severe reductions in biological activity and by a persistent state change, a regime shift, it is also known as desertification. Desertification often happens when poor management practices such as depletion of water resources and overgrazing are compounded by exploitative and distant economic and political systems (Bestelmeyer et al. 2015, Reynolds et al. 2007). Most dryland livelihoods depend closely on highly variable biophysical drivers such as precipitation, which means that there is great uncertainty in detecting thresholds of state change and maintaining the sustainability and future of these social-ecological systems.

Woody Encroachment

In drylands, woody encroachment is often an example of desertification because continuous grass cover is replaced by patchy vegetation cover with increased proportion of bare soils, which leads to increased water and wind erosion, and nutrient loss for the soil in plant interspaces (Barger et al. 2011, D'Odorico et al. 2012, Evans et al. 2017). Encroachment is likely to result in degradation of value for cattle and sheep production, but does not necessarily result in degradation of structural and functional capacities of the ecosystem (Eldridge et al. 2011). The drivers of woody encroachment are complex and vary greatly by region and ecological zone but often reflect an interaction among climate (i.e., intensity and frequency of precipitation), land use (i.e., grazing), atmospheric chemistry (i.e., CO_2 concentrations), soil properties (i.e., texture and depth), and disturbance regimes (i.e., fire frequency and intensity) (Archer et al. 2017, Briggs et al. 2005). Robust generalizations about the importance of the different drivers remain unavailable due to context-dependent factors that co-occur and interact to promote or constrain woody cover (Archer et al. 2017, D'Odorico et al. 2012). In the past few decades, brush management of woody encroachment has focused on reversing the directional shifts in vegetation with the goal of improving livestock production (Archer et al. 2017).

Poverty and Food Insecurity

Grazing systems are the most common forms of agricultural production in the developing world, in terms of both the total land area and the number of people supported (Herrero et al. 2012). In Africa, for example, livestock production systems cover 73% of the land area (Thornton 2014). Since livestock provides a critical source of income and food in livestock-dependent communities, interactions between land degradation and climate change that reduce grazing land productivity are an existential threat to many of the world's poor (HLPE 2012). There are various adaptation strategies that can help reduce livestock-dependent communities' vulnerability to climate change, including improved grazing, soil, and nutrient management (Thornton 2014). Yet, existing institutional and household asset constraints typically limit farmers' abilities to pursue these activities and lock farmers into the extensive practices that contributed to land degradation in the first place. These social-ecological dynamics create what are known as "resource degradation poverty traps", whereby human behaviors erode natural capital and reduce land productivity, food production, and income, leading to further constraints on the range of behaviors possible, further degradation, and greater poverty (Barrett 2008).

Nutrient Loss due to Runoff, Leaching, and Erosion

In grasslands, rainwater and snowmelt can remove nutrients from soils via runoff, leaching, and erosion (Bellows 2001). Runoff and erosion transport either dissolved nutrients or nutrients attached to soil particles over the soil surface to streams, rivers, and lakes. Leaching, on the other hand, is the downward transport of dissolved nutrients through the soil profile, out of reach of plant roots, and into groundwater (Bellows 2001). Leaching is more likely to happen when the number of cattle herd exceeds the assimilative capacity of the land (Billota et al. 2007). All three processes are detrimental to livestock production and environment health because they deplete nutrients that could be used for forage production and contaminate water bodies (Bellows 2001). High nitrate levels in drinking water can cause health problems for animal and human infants, while high phosphorous levels in streams and lakes can lead to eutrophication.

Maintaining adequate livestock density, planting deep-rooted and prostrate vegetation, and keeping plant residue cover can enhance water infiltration and provide good protection against nutrient loss (Bellows 2001, Billota et al. 2007). In addition, maintaining a healthy and diverse population of soil organisms helps incorporate nutrients into soil, reducing the risk of runoff, leaching, and erosion. Finally, protection of riparian areas surrounding streams and rivers is essential to minimizing broader ecosystem pollution (Dosskey et al. 2010, Osborne and Kovacic 1993).

Compaction

Soil compaction reduces the ability of nutrients, water, and roots to move through the soil (Bellows 2001). Soil compaction is caused by animals or equipment continuously moving across the soil, pressing down on the soil, and squeezing soil pore spaces together. The likelihood of compaction increases as soil moisture, animal weight, animal numbers, and the duration of grazing increase (Billota et al. 2007). Soil compaction makes vegetation susceptible to disease, restricts root growth, reduces nutrient availability for plant uptake, increases the potential for runoff and erosion, and reduces forage yields (Bellows 2001).

The likelihood of compaction can be reduced by (i) maintaining a diverse population of soil organisms that burrow in the soil and form soil aggregates, (ii) not grazing animals on wet or poorly drained soils, and (iii) decreasing the number of animals and/or the time they spend in each paddock (Bellows 2001).

Greenhouse Gas Emissions and Deforestation

Regardless of whether they are "conventional" or grass fed, most ruminants spend a majority of their life grazing and a relatively short period at the end of their life in a forage or feed-based finishing system. Consequently, a majority of their emissions from digestion (i.e., enteric fermentation) are associated with grazing. For example, a study in Alberta, Canada, found that across the entire cattle production life cycle, the reproduction and grazing period associated with calf–cow production contributed 80% of the greenhouse gas emissions for each finished steer (Beauchemin et al. 2010). Consequently, these authors recommend that interventions to reduce the greenhouse gas footprint of beef production from enteric fermentation should focus on improving forage quality within cow–calf grazing systems. Conversely, extending the amount of time that cattle spend grazing, i.e., in fully grass-fed systems, would substantially increase the amount of greenhouse gas emissions associated with beef production (Hayek and Garrett 2018).

Grazing on lower productivity rangelands and grasslands, including degraded systems, lowers animal weight gain. The longer it takes to raise a ruminant for slaughter, the more methane it will produce over its lifetime as part of its normal digestive processes. Land degradation can also contribute to greater greenhouse gas emissions via deforestation. When lands become degraded, they may be abandoned, leading farmers to clear new land that has more fertile soil (Balbino et al. 2011, Dias-Filho 2014, Walker et al. 2000). This is a major sustainability challenge throughout much of Latin America. In the Brazilian Amazon, for example, pastures accounted for 59% of the cleared forest area in 2012 and 22% of these areas were in some degree of degradation (INPE 2012). Improved pasture management, including better grass varieties, rotational grazing, and integrated systems, can reduce deforestation and spare land for biodiversity by increasing production on existing cleared lands (Cohn et al. 2014, Gil et al. 2018, Strassburg et al. 2014). Since land cover change is one of the leading sources of greenhouse gas emissions (Houghton et al. 2012), improved pasture management can be an effective way to mitigate climate change (Strassburg et al. 2014), so long as it is coupled with stringent conservation policies (Garrett et al. 2018).

Major Knowledge Gaps in Grazing Systems

Despite the fact that livestock grazing is the largest land use activity on the planet (Erb et al. 2017, Herrero et al. 2013) and that it is important in tackling climate change and improving food security, quantitative global analyses of the spatial patterns, and dynamics of grazing are rare (Erb et al. 2017, Kuemmerle et al. 2013). While there is some knowledge on the biogeochemical and biophysical aspects of livestock management, there is a lack of robust global datasets on the extent and intensity of grazing systems and associated management practices (Erb et al. 2017). Spatially explicit, consistent, and comprehensive global grazing data are crucial for doing social-ecological research. Data gaps are largest in developing countries, which are currently experiencing rapid land use change that is likely to accelerate in the coming decades.

Aside from spatially explicit information on grazing area and management and associated biomass production and use, there are many additional knowledge gaps related to fundamental processes and outcomes in grazing social-ecological systems. Very little is known about animal outcomes, including welfare, behavior, genetics, reproduction, and health under different management and environmental conditions. While much progress has been made in measuring ecosystem services and livelihood outcomes in grazing systems, there has not been sufficient study of landscape scale outcomes across these two interrelated sets of factors to understand how grazing systems can be improved to achieve more win-wins (and fewer trade-offs). These knowledge gaps become even more prominent when considering the complex ways that climate change may alter both human management and ecological responses in grazing systems. Finally, greater social research is needed to better understand the multiple livelihood and community uses and resulting social values of grazing systems.

Conclusions

Grazing systems are extensive and heterogeneous; they provide numerous ecosystem services in addition to provisioning feed for livestock, including carbon sequestration, habitat for diverse species, regulation of biogeochemical cycles, and maintenance of traditional lifestyles. While livestock rearing has important cultural significance and provides income and nutrition in areas that cannot support any other livelihoods, the livestock production sector poses some challenges to the sustainability of our planet as a result of deforestation, habitat and land degradation, water pollution, and greenhouse gas emissions. Ensuring food security, satisfying increased demand for meat and dairy, and protecting grassland ecosystems in an era of climate change are significant scientific and societal challenges. Given the complex and multi-dimensional social-ecological processes that shape grazing systems, there is an urgent need to improve our understating of the biophysical and biogeochemical aspects, and spatially explicit dynamics of grazing in order to devise and apply sustainable management practices.

References

Allen, V.G., C. Batello, E.J. Berretta, J. Hodgson, M. Kothmann, X. Li, J. McIvor, J. Milne, C. Morris, A. Peeters, and M. Sanderson. 2011. "An International Terminology for Grazing Lands and Grazing Animals: Grazing Lands and Grazing Animals." *Grass and Forage Science* 66 (1): 2–28. doi:10.1111/j.1365-2494.2010.00780.x.

Archer S.R., E.M. Andersen, K.I. Predick, S. Schwinning, R.J. Steidl, and S.R. Woods. 2017. Woody Plant Encroachment: Causes and Consequences. In *Rangeland Systems: Processes, Management and Challenges*, ed. D.D. Briske, 25–84. Cham: Springer.

Balbino, L. C., L. A. M. Cordeiro, V. Porfírio-da-Silva, A. Moraes, G. B. Martínez, R. C. Alvarenga, A. N. Kichel, R. S. Fontaneli, H. P. Santos, J. C. Franchini, and P. R. Galerani. 2011. "Evolução tecnológica e arranjos produtivos de sistemas de integração lavoura-pecuária-floresta no Brasil [Evolution of Technology and Productive Arrangements in Integrated Crop-Livestock-Forestry Systems in Brazil]." *Pesquisa Agropecuária Brasileira* 46 (10) : i-xii. http://dx.doi.org/10.1590/S0100-204X2011001000001.

Barger, N.N., S.R. Archer, J.L. Campbell, Cho-ying Huang, Jeffery A. Morton, and Alan K. Knapp. 2011. "Woody Plant Proliferation in North American Drylands: A Synthesis of Impacts on Ecosystem Carbon Balance." *Journal of Geophysical Research* 116. doi:10.1029/2010JG001506.

Barrett, C.B. 2008. Poverty Traps and Resource Dynamics in Smallholder Agrarian Systems. *Economics of Poverty, Environment and Natural-Resource Use*. A. Ruijs, R. Dellink, eds., Springer.

Batjes, N.H. 2014. "Total Carbon and Nitrogen in the Soils of the World." *European Journal of Soil Science* 65 (1): 10–21. doi:10.1111/ejss.12114_2.

Beauchemin, K.A., H. Henry Janzen, Shannan M. Little, Tim A. McAllister, and Sean M. McGinn. 2010. "Life Cycle Assessment of Greenhouse Gas Emissions from Beef Production in Western Canada: A Case Study." *Agricultural Systems* 103 (6): 371–379. doi:10.1016/j.agsy.2010.03.008.

Bellows B. 2001. Nutrient Cycling in Pastures. Appropriate Technology Transfer for Rural Areas. http://pss.uvm.edu/pdpforage/Materials/SoilFert/nutrientcycling.pdf (accessed May 30, 2019).

Bestelmeyer, B.T., and D.D. Briske. 2012. "Grand Challenges for Resilience-Based Management of Rangelands." *Rangeland Ecology & Management* 65 (6): 654–663. doi:10.2111/REM-D-12-00072.1.

Bestelmeyer, B.T, G.S. Okin, Michael C. Duniway, Steven R Archer, Nathan F. Sayre, Jebediah C. Williamson, and Jeffrey E. Herrick. 2015. "Desertification, Land Use, and the Transformation of Global Drylands." *Frontiers in Ecology and the Environment* 13 (1): 28–36. doi:10.1890/140162.

Bilotta, G.S., Brazier, R.E., and Haygarth, P.M. 2007. "The Impacts of Grazing Animals on the Quality of Soils, Vegetation, and Surface Waters in Intensively Managed Grasslands." *Advances in Agronomy* 94: 237–280.

Briggs, J.M., A.K. Knapp, J.M. Blair, J.L. Heisler, G.A. Hoch, M.S. Lett, and J.K. McCarron. 2005. "An Ecosystem in Transition: Causes and Consequences of the Conversion of Mesic Grassland to Shrubland." *BioScience* 55 (3): 243. doi:10.1641/0006-3568(2005)055[0243:AEITCA]2.0.CO;2.

Briske, D.D., S.D. Fuhlendorf, and F.E. Smeins. 2003. "Vegetation Dynamics on Rangelands: A Critique of the Current Paradigms." *Journal of Applied Ecology* 40 (4): 601–614. doi:10.1046/j.1365-2664.2003.00837.x.

Briske, D.D., B.T. Bestelmeyer, T.K. Stringham, and P.L. Shaver. 2008. "Recommendations for Development of Resilience-Based State-and-Transition Models." *Rangeland Ecology & Management* 61 (4): 359–367. doi:10.2111/07-051.1.

Carvalho, P.C.de F., A. de Moraes, L.da S. Pontes, I. Anghinoni, R.M. Sulc, and C. Batello. 2014. "Definições e terminologias para sistema integrado de produção agropecuária." *Revista Ciencia Agronomica* 45 (5): 1040–1046.

Chapin III, F.S. 1997. "Biotic Control over the Functioning of Ecosystems." *Science* 277 (5325): 500–504. doi:10.1126/science.277.5325.500.

Cohn, A.S., A. Mosnier, P. Havlik, H. Valin, M. Herrero, E. Schmid, M. O'Hare, and M. Obersteiner. 2014. "Cattle Ranching Intensification in Brazil Can Reduce Global Greenhouse Gas Emissions by Sparing Land from Deforestation." *Proceedings of the National Academy of Sciences* 111 (20): 7236–7241. doi:10.1073/pnas.1307163111.

Conant, R.T. 2010. Challenges and Opportunities for Carbon Sequestration in Grassland Systems. A Technical Report on Grassland Management and Climate Change Mitigation, Integrated Crop Management. Food and Agriculture Organisation of the United Nations, Rome. doi:10.3329/jard.v7i1.4430 (accessed May 30, 2019).

Condron, L.M., D.W. Hopkins, E.G. Gregorich, A. Black, and S.A. Wakelin. 2014. "Long-Term Irrigation Effects on Soil Organic Matter under Temperate Grazed Pasture: Pasture Production and Soil Organic Matter." *European Journal of Soil Science* 65 (5): 741–750. doi:10.1111/ejss.12164.

D'Odorico, P., G.S. Okin, and B.T. Bestelmeyer. 2012. "A Synthetic Review of Feedbacks and Drivers of Shrub Encroachment in Arid Grasslands: Feedbacks and Drivers of Shrub Encroachment." *Ecohydrology* 5 (5): 520–530. doi:10.1002/eco.259.

Dias-Filho, M.B. 2014. *Reclaiming the Brazilian Amazon: The Restoration and Management of Pasture Lands*. Embrapa Eastern Amazon: Belém.

Donohue, R.J., M.L. Roderick, T.R. McVicar, and G.D. Farquhar. 2013. "Impact of CO_2 Fertilization on Maximum Foliage Cover across the Globe's Warm, Arid Environments: CO_2 Fertilization and Foliage Cover." *Geophysical Research Letters* 40 (12): 3031–3035. doi:10.1002/grl.50563.

Dosskey, M.G., P. Vidon, N.P. Gurwick, C.J. Allan, T.P. Duval, R. Lowrance. 2010. "The Role of Riparian Vegetation in Protecting and Improving Chemical Water Quality in Streams1." *JAWRA Journal of the American Water Resources Association* 46: 261–277.

van der Ploeg D. 2010. "The Peasantries of the Twenty-First Century: The Commoditisation Debate Revisited." *The Journal of Peasant Studies* 37 (1): 1–30. doi:10.1080/03066150903498721.

Duffy, J.E. 2002. "Biodiversity and Ecosystem Function: The Consumer Connection." *Oikos* 99 (2): 201–219. doi:10.1034/j.1600-0706.2002.990201.x.

Eldridge, David J., M.A. Bowker, F.T. Maestre, E. Roger, J.F. Reynolds, and W.G. Whitford. 2011. "Impacts of Shrub Encroachment on Ecosystem Structure and Functioning: Towards a Global Synthesis: Synthesizing Shrub Encroachment Effects." *Ecology Letters* 14 (7): 709–722. doi:10.1111/j.1461-0248.2011.01630.x.

Ellis, J.E., and D.M. Swift. 1988. "Stability of African Pastoral Ecosystems: Alternate Paradigms and Implications for Development." *Journal of Range Management* 41 (6): 450. doi:10.2307/3899515.

Elmqvist, T., C. Folke, M. Nyström, G. Peterson, J. Bengtsson, B. Walker, and J. Norberg. 2003. "Response Diversity, Ecosystem Change, and Resilience." *Frontiers in Ecology and the Environment* 1 (9): 488–94. doi:10.1890/1540-9295(2003)001[0488:RDECAR]2.0.CO;2.

Erb, K.-H, S. Luyssaert, P. Meyfroidt, J. Pongratz, A. Don, S. Kloster, T. Kuemmerle, T. Fetzel, R. Fuchs, M. Herold, H. Haberl, C.D. Jones, E. Marin-Spiotta, I. McCallum, E. Robertson, V. Seufert, S. Fritz, A. Valade, A. Wiltshire, and A.J. Dolman. 2017. "Land Management: Data Availability and Process Understanding for Global Change Studies." *Global Change Biology* 23 (2): 512–533. doi:10.1111/gcb.13443.

Evans, R.D., R.A. Gill, V.T. Eviner, and V. Bailey. 2017. Soil and Belowground Process. In *Rangeland Systems: Processes, Management and Challenges*, ed. D.D. Briske, 131–168. : Cham: Springer.

FAO. 2011. *The State of the World's Land and Water Resources for Food and Agriculture*. Food and Agriculture Organization of the United Nations (FAO). http://www.fao.org/3/a-i1688e.pdf (accessed May 30, 2019).

Fernandez-Gimenez, M.E., and B. Allen-Diaz. 1999. "Testing a Non-Equilibrium Model of Rangeland Vegetation Dynamics in Mongolia." *Journal of Applied Ecology* 36 (6): 871–885. doi:10.1046/j.1365-2664.1999.00447.x.

Folke, C., S. Carpenter, B. Walker, M. Scheffer, T. Elmqvist, L. Gunderson, and C.S. Holling. 2004. "Regime Shifts, Resilience, and Biodiversity in Ecosystem Management." *Annual Review of Ecology, Evolution, and Systematics* 35 (1): 557–81. doi:10.1146/annurev.ecolsys.35.021103.105711.

Fuhlendorf, S.D., W.C. Harrell, D.M. Engle, R.G. Hamilton, C.A. Davis, and D.M. Leslie. 2006. "Should Heterogeneity Be the Basis for Conservation? Grassland Bird Response to Fire and Grazing." *Ecological Applications* 16 (5): 1706–1716. doi:10.1890/1051-0761(2006)016[1706:SHBTBF]2.0.CO;2.

Fuhlendorf S.D., R.W.S. Fynn, D.A. McGranahan, and D. Twidwell. 2017. Heterogeneity as the Basis for Rangeland Management. In *Rangeland Systems: Processes, Management and Challenges*, ed. D.D. Briske, 169–196. Cham: Springer.

Gaitán, J.J., G.E. Oliva, D.E. Bran, F.T. Maestre, M.R. Aguiar, E.G. Jobbágy, G.G. Buono, D. Ferrante, V.B. Nakamatsu, G. Ciari, J.M. Salomone, and V. Massara. 2014. "Vegetation Structure Is as Important as Climate for Explaining Ecosystem Function across Patagonian Rangelands." Edited by Scott Wilson. *Journal of Ecology* 102 (6): 1419–1428. doi:10.1111/1365–2745.12273.

Garnett, T., C. Godde, and A. Muller. 2017. Grazed and Confused? Food Climate Research Network. https://fcrn.org.uk/sites/default/files/project-files/fcrn_gnc_report.pdf (accessed May 30, 2019).

Garrett, R. D., J. Ryschawy, L. W. Bell, O. Cortner, J. Ferreira, A. V. N. Garik, J. D. B. Gil, L. Klerkx, M. Moraine, C. A. Peterson, J. C. dos Reis, and J. F. Valentim. 2020. Drivers of decoupling and recoupling of crop and livestock systems at farm and territorial scales. *Ecology and Society* 25 (1): 24. https://doi.org/10.5751/ES-11412-250124

Garrett, R.D., T.A. Gardner, T.F. Morello, S. Marchand, J. Barlow, D.E. de Blas, J. Ferreira, A.C. Lees, and L. Parry. 2017. "Explaining the Persistence of Low Income and Environmentally Degrading Land Uses in the Brazilian Amazon." *Ecology and Society* 22 (3). doi: 10.5751/ES-09364-220327.

Garrett, R.D., I. Koh, E.F. Lambin, Y. le Polain de Waroux, J.H. Kastens, and J.C. Brown. 2018. "Intensification in Agriculture-Forest Frontiers: Land Use Responses to Development and Conservation Policies in Brazil." *Global Environmental Change* 53: 233–243. doi:10.1016/j.gloenvcha.2018.09.011.

Garrett, R.D., M.T. Niles, J.D.B. Gil, A. Gaudin, R. Chaplin-Kramer, A. Assmann, T.S. Assmann, et al. 2017. "Social and Ecological Analysis of Commercial Integrated Crop Livestock Systems: Current Knowledge and Remaining Uncertainty." *Agricultural Systems* 155: 136–146. doi:10.1016/j.agsy.2017.05.003.

Garrett, R. D., J. Ryschawy, L. W. Bell, O. Cortner, J. Ferreira, A. V. N. Garik, J. D. B. Gil, L. Klerkx, M. Moraine, C. A. Peterson, J. C. dos Reis, and J. F. Valentim. 2020. Drivers of decoupling and recoupling of crop and livestock systems at farm and territorial scales. *Ecology and Society* 25(1): 24. https://doi.org/10.5751/ES-11412-250124

Geist, H.J., and E.F. Lambin. 2004. "Dynamic Causal Patterns of Desertification." *BioScience* 54 (9): 817. doi:10.1641/0006-3568(2004)054[0817:DCPOD]2.0.CO;2.

Gil, J.D.B, R.D. Garrett, A. Rotz, V. Daioglou, J. Valentim, G.F. Pires, M.H. Costa, L. Lopes, and J.C. Reis. 2018. "Tradeoffs in the Quest for Climate Smart Agricultural Intensification in Mato Grosso, Brazil." *Environmental Research Letters* 13 (6): 064025. doi:10.1088/1748-9326/aac4d1.

Hayek, M.N., and R.D. Garrett. 2018. "Nationwide Shift to Grass-Fed Beef Requires Larger Cattle Population." *Environmental Research Letters* 13 (8): 084005. doi:10.1088/1748-9326/aad401.

Hecht, S.B. 1993. "The Logic of Livestock and Deforestation in Amazonia." *BioScience* 43 (10): 687–695. doi:10.2307/1312340.

Herrero, M., P.K. Thornton, A. Notenbaert, S. Msangi, S. Wood, R. Kruska, J. Dixon, J. Bossio, D.J. van de Steeg, H.A. Freeman, X. Li, and P. ParthasarathyRao. 2012. Drivers of Change in Crop-Livestock Systems and Their Potential Impacts on Agro-Ecosystems Services and Human Wellbeing to 2030. https://cgspace.cgiar.org/bitstream/handle/10568/3020/SLP%20drivers%20study%20final%20draft.pdf (accessed May 30, 2019).

Herrero, M., P. Havlik, H. Valin, A. Notenbaert, M.C. Rufino, P.K. Thornton, M. Blummel, F. Weiss, D. Grace, and M. Obersteiner. 2013. "Biomass Use, Production, Feed Efficiencies, and Greenhouse Gas Emissions from Global Livestock Systems." *Proceedings of the National Academy of Sciences* 110 (52): 20888–20893. doi:10.1073/pnas.1308149110.

HLPE (High Level Panel of Experts). 2012. Food Security and Climate Change. A Report by the High Level Panel of Experts on Food Security and Nutrition of the Committee on World Food Security, Rome 2012. http://www.fao.org/3/a-me422e.pdf (accessed May 30, 2019).

Hoelle, J. 2011. "Convergence on Cattle: Political Ecology, Social Group Perceptions, and Socioeconomic Relationships in Acre, Brazil: Convergence on Cattle." *Culture, Agriculture, Food and Environment* 33 (2): 95–106. doi:10.1111/j.2153-9561.2011.01053.x.

Houghton, R.A., J.I. House, J. Pongratz, G.R. van der Werf, R.S. DeFries, M.C. Hansen, C. Le Quéré, and N. Ramankutty. 2012. "Carbon Emissions from Land Use and Land-Cover Change." *Biogeosciences* 9 (12): 5125–5142. doi:10.5194/bg-9-5125–2012.

Huang, J., H. Yu, X. Guan, G. Wang, and R. Guo. 2016. "Accelerated Dryland Expansion under Climate Change." *Nature Climate Change* 6 (2): 166–171. doi:10.1038/nclimate2837.

INPE. 2012. *Projeto TerraClass 2012: Mapeamento do uso e cobertura da terra na Amazônia Legal Brasileira (Project TerraClass 2012: Mapping of the Use and Cover of Land in the Brazilian Legal Amazon).* http://www.inpe.br/noticias/arquivos/pdf/TerraClass_2012.pdf.

Jackson, J.B.C. 2001. "Historical Overfishing and the Recent Collapse of Coastal Ecosystems." *Science* 293 (5530): 629–637. doi:10.1126/science.1059199.

King, E.G., and R.J. Hobbs. 2006. "Identifying Linkages among Conceptual Models of Ecosystem Degradation and Restoration: Towards an Integrative Framework." *Restoration Ecology* 14 (3): 369–378. doi:10.1111/j.1526-100X.2006.00145.x.

Knapp, A.K., and M.D. Smith. 2001. "Variation among Biomes in Temporal Dynamics of Aboveground Primary Production." *Science, New Series* 291 (5503): 481–484.

Knight, R.L. 2002. *Ranching West of the 100th Meridian: Culture, Ecology, and Economics*. Washington, DC: Island Press.

Kuemmerle, T., K. Erb, P. Meyfroidt, D. Müller, P.H. Verburg, S. Estel, H. Haberl, P. Hostert, M.R. Jepsen, T. Kastner, C. Levers, M. Lindner, C. Plutzar, P.J. Verkerk, E.H. van der Zanden, and A. Reenberg. 2013. "Challenges and Opportunities in Mapping Land Use Intensity Globally." *Current Opinion in Environmental Sustainability* 5 (5): 484–493. doi:10.1016/j.cosust.2013.06.002.

Lemaire, G., A. Franzluebbers, P.C. de Faccio Carvalho, and B. Dedieu. 2014. "Integrated Crop-Livestock Systems: Strategies to Achieve Synergy between Agricultural Production and Environmental Quality." *Agriculture, Ecosystems & Environment* 190: 4–8. doi:10.1016/j.agee.2013.08.009.

MA. 2005. *Millennium Ecosystem Assessment (MA) Synthesis Report*. Washington, DC: Millennium Ecosystem Assessment. https://www.millenniumassessment.org/documents/document.356.aspx.pdf (accessed May 30, 2019).

Maurer, K., A. Weyand, M. Fischer, and Jürg Stöcklin. 2006. "Old Cultural Traditions, in Addition to Land Use and Topography, Are Shaping Plant Diversity of Grasslands in the Alps." *Biological Conservation* 130 (3): 438–446. doi:10.1016/j.biocon.2006.01.005.

Muchagata, M., and K. Brown. 2003. "Cows, Colonists and Trees: Rethinking Cattle and Environmental Degradation in Brazilian Amazonia." *Agricultural Systems* 76 (3): 797–816. doi:10.1016/S0308-521X(02)00015-X.

Nobilly, F., and R.H. Bryant. 2013. "Productivity of Rotationally Grazed Simple and Diverse Pasture Mixtures under Irrigation in Canterbury." *Proceedings of the New Zealand Grassland Association* 75: 165–172.

Osborne, L.L. and D.A. Kovacic. 1993. "Riparian Vegetated Buffer Strips in Water-Quality Restoration and Stream Management." *Freshwater Biology* 29: 243–258.

Parker, J.S. 2013. "Integrating Culture and Community into Environmental Policy: Community Tradition and Farm Size in Conservation Decision Making." *Agriculture and Human Values* 30: 159–178. doi:10.1007/s10460-012-9392-8.

Perkins, J.S., and D.S.G. Thomas. 1993. "Spreading Deserts or Spatially Confined Environmental Impacts? Land Degradation and Cattle Ranching in the Kalahari Desert of Botswana." *Land Degradation & Development* 4: 179–194.

Polley H.W., D.W. Bailey, R.S. Nowak, and M. Stafford-Smith. 2017. Ecological Consequences of Climate Change on Rangelands. In *Rangeland Systems: Processes, Management and Challenges*, ed. D.D. Briske, 229–260. Cham: Springer

Poulter, B., D. Frank, P. Ciais, R.B. Myneni, N. Andela, J. Bi, G. Broquet, J.G. Canadell, F. Chevallier, Y.Y. Liu, S.W. Running, S. Sitch, and G.R. van der Werf. 2014. "Contribution of Semi-Arid Ecosystems to Interannual Variability of the Global Carbon Cycle." *Nature* 509 (7502): 600–603. doi:10.1038/nature13376.

Reeves, M.C., A.L. Moreno, K.E. Bagne, and S.W. Running. 2014. "Estimating Climate Change Effects on Net Primary Production of Rangelands in the United States." *Climatic Change* 126 (3–4): 429–442. doi:10.1007/s10584-014-1235-8.

Reynolds, J.F., D.M.S. Smith, E.F. Lambin, B.L. Turner, M. Mortimore, S.P.J. Batterbury, T.E. Downing, H. Dowlatabadi, R.J. Fernandez, J.E. Herrick, E. Huber-Sannwald, H. Jiang, R. Leemans, T. Lynam, F.T. Maestre, M. Ayarza, and B. Walker. 2007. "Global Desertification: Building a Science for Dryland Development." *Science* 316 (5826): 847–851. doi:10.1126/science.1131634.

Robinson, T.P., Thornton, P.K., Franceschini, G., Kruska, R.L., Chiozza, F., Notenbaert, A., Cecchi, G., Herrero, M., Epprecht, M., Fritz, S., You, L., Conchedda, G., See, L., 2011. *Global Livestock Production Systems*. Food and Agriculture Organization of the United Nations (FAO) and International Livestock Research Institute (ILRI). http://www.fao.org/3/i2414e/i2414e.pdf (accessed May 30, 2019).

Ryschawy, J., G. Martin, M. Moraine, M. Duru, and O. Therond. 2017. "Designing Crop-Livestock Integration at Different Levels: Toward New Agroecological Models?" *Nutrient Cycling in Agroecosystems* 108 (1): 5–20. doi:10.1007/s10705-016-9815-9.

Sala, O.E., and J. Paruelo. 1997. Ecosystem Services in Grasslands. In *Nature's Services: Societal Dependence on Natural Ecosystems*, ed. G.C. Daily, 237–251. Washington, DC: Island Press.

Sala, O.E., L.A. Gherardi, L. Reichmann, E. Jobbagy, and D. Peters. 2012. "Legacies of Precipitation Fluctuations on Primary Production: Theory and Data Synthesis." *Philosophical Transactions of the Royal Society B: Biological Sciences* 367 (1606): 3135–3144. doi:10.1098/rstb.2011.0347.

Sala O.E., L. Yahdjian, K. Havstad, and M.R. Aguiar. 2017. Rangeland Ecosystem Services: Nature's Supply and Human's Demand. In *Rangeland Systems: Processes, Management and Challenges*, ed. D.D. Briske, 467–489. Cham: Springer.

Schiere, J.B, M.N.M Ibrahim, and H. van Keulen. 2002. "The Role of Livestock for Sustainability in Mixed Farming: Criteria and Scenario Studies under Varying Resource Allocation." *Agriculture, Ecosystems & Environment* 90 (2): 139–153. doi:10.1016/S0167-8809(01)00176-1.

Skaggs, R. 2008. "Ecosystem Services and Western U.S. Rangelands." *Choices* 23: 37–41.

Sloat, L.L., J.S. Gerber, L.H. Samberg, W.K. Smith, M. Herrero, L.G. Ferreira, C.M. Godde, and P.C. West. 2018. "Increasing Importance of Precipitation Variability on Global Livestock Grazing Lands." *Nature Climate Change* 8 (3): 214–218. doi:10.1038/s41558-018-0081-5.

Smith, P. 2014. “Do Grasslands Act as a Perpetual Sink for Carbon?” *Global Change Biology* 20 (9): 2708–2711. doi:10.1111/gcb.12561.

Stinner, D.H., M.G. Paoletti, B.R. Stinner. 1989. “In Search of Traditional Farm Wisdom for a More Sustainable Agriculture: A Study of Amish Farming and Society.” *Agricultural Ecology and Environment* 27: 77–90.

Strassburg, B.B.N., A.E. Latawiec, L.G. Barioni, C.A. Nobre, V.P. da Silva, J.F. Valentim, M. Vianna, and E.D. Assad. 2014. “When Enough Should Be Enough: Improving the Use of Current Agricultural Lands Could Meet Production Demands and Spare Natural Habitats in Brazil.” *Global Environmental Change* 28: 84–97. doi:10.1016/j.gloenvcha.2014.06.001.

Tews, J., U. Brose, V. Grimm, K. Tielbörger, M.C. Wichmann, M. Schwager, and F. Jeltsch. 2004. “Animal Species Diversity Driven by Habitat Heterogeneity/Diversity: The Importance of Keystone Structures: Animal Species Diversity Driven by Habitat Heterogeneity.” *Journal of Biogeography* 31 (1): 79–92. doi:10.1046/j.0305-0270.2003.00994.x.

Thornton, P.K. 2014. *Livestock and Mixed Crop-Livestock Systems*. Washington, DC: International Food Policy Research Institute (IFPRI).

Valentim, J.F., and C.M.S. Andrade. 2005. Forage Peanut (Arachis pintoi): A High Yielding and High Quality Tropical Legume for Sustainable Cattle Production Systems in the Western Brazilian Amazon. In *International Grassland Congress*, Dublin, 329.

Walker, R., E. Moran, and L. Anselin. 2000. “Deforestation and Cattle Ranching in the Brazilian Amazon: External Capital and Household Processes.” *World Development* 28 (4): 683–99. doi:10.1016/S0305-750X(99)00149-7.

White R., S. Murray, and M. Rohweder. 2000. *Pilot Analysis of Global Ecosystems: Grassland Ecosystems*. World Resources Institute. https://wriorg.s3.amazonaws.com/s3fs-public/pdf/page_grasslands.pdf (accessed May 30, 2019).

Wilcox, Kevin R., Z. Shi, L.A. Gherardi, N.P. Lemoine, S.E. Koerner, D.L. Hoover, E. Bork, K.M. Byrne, J. Cahill Jr., S.L. Collins, S. Evans, A.K. Gilgen, P. Holub, L. Jiang, A.K. Knapp, D. LeCain, J. Liang, P. Garcia-Palacios, J. Penuelas, W.T. Pockman, M.D. Smith, S. Sun, S.R. White, L. Yahdjian, K. Zhu, Y. Luo. 2017a. “Asymmetric Responses of Primary Productivity to Precipitation Extremes: A Synthesis of Grassland Precipitation Manipulation Experiments.” *Global Change Biology* 23 (10): 4376–85. https://doi.org/10.1111/gcb.13706.

Wilcox, R.W. 2017. *Cattle in the Backlands: Mato Grosso and the Evolution of Ranching in the Brazilian Tropics*. Austin: University of Texas Press.

Wu, D., X. Zhao, S. Liang, T. Zhou, K. Huang, B. Tang, and W. Zhao. 2015. “Time-Lag Effects of Global Vegetation Responses to Climate Change.” *Global Change Biology* 21 (9): 3520–3531. doi:10.1111/gcb.12945.

24

Salt-Affected Soils: Physical Properties and Behavior

Hwat Bing So

Introduction

Apart from gravel and stones, soils consist largely of a mixture of sand, silt, and clay particles, held together as aggregates by organic matter, iron-aluminum oxides, and/ or electrostatic forces between clay particles mediated by the cations adsorbed on the negative charges of the clay surfaces. The proportion of sand, silt, and clay particles is referred to as the texture of the soil, which imparts certain fundamental properties onto the soil, e.g., a sandy soil behaves quite differently from a clayey soil. These particles are generally organized as compound aggregates and the arrangement of these aggregates and particles is known as the structure of the soil, which modifies and imposes properties other than those resulting from its texture, e.g., a well-structured or -aggregated clay soil will hold water well but will also transmit water and drain readily, while a poorly aggregated clay soil of the same texture will drain slowly and tends to waterlog following rainfall as water content remains above field capacity (upper limit of the non-limiting water range) for excessively long periods.

As part of the soil structure, soil particles and aggregates are generally arranged in a hierarchical order, where the lower orders tend to have higher densities and greater internal strength than the higher orders.[1,2] At the lowest order, clay particles (<2 microns) tend to form clay domains (packets of parallel clay sheets of around 5–7 sheets). Several domains form clusters, followed by several orders of clusters, microaggregates (approximately 20–250 microns), and macroaggregates or clods as the highest order. Only the larger microaggregates and macroaggregates are visible to the naked eye. The hierarchical nature implies that destruction of a lower order will result in the destruction of all higher orders, and therefore, the stability of these clay domains are fundamental in determining the stability of the aggregates and, hence, soil structure, except where the bonding at the higher orders are extremely

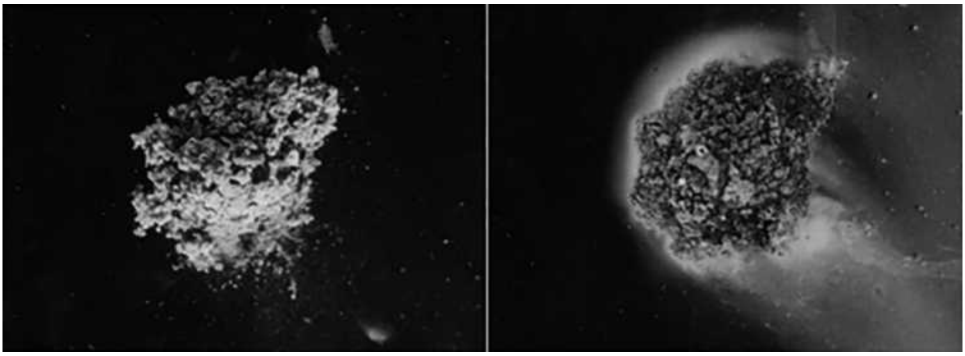


FIGURE 1 Left: Slaking of a sodic aggregate subjected to rapid wetting in a petri dish with water. Right: Spontaneous dispersion (shown as cloud around slaked aggregate) after 1 hr in water.
Source: Photo courtesy of R. Greene, Australian National University.

strong. The higher orders are physically held together by organic matter and/or iron-aluminum oxides, while the lower orders and domains are predominantly held together by electrostatic forces between clay particles (flocculation). The breakdown of the higher orders (macro- to microaggregates) is defined as slaking, while the breakdown of the lower orders (microaggregates to individual clay particles or domains) is termed dispersion (see Figure 1).

Good structural stability is essential for a biologically productive soil, to maintain adequate transmission and supply of water, adequate drainage and aeration, and adequate chemical fertility in terms of cations adsorbed on the clay surfaces. Roots must be able to grow through the soil to access water, nutrients, and oxygen, and this can occur only in soils with sufficiently large pores and low mechanical resistance, such as found in structurally stable soils.

Cations such as sodium, potassium, calcium, magnesium, aluminum, and hydrogen are commonly found in the soil as adsorbed or exchangeable cations on the surfaces of the colloidal clay particles. In alkaline soils, the first four are the predominant cations, while in acid soils, hydrogen and aluminum become increasingly dominant as the acidity increases or the pH decreases. By virtue of their small ionic radius and/or their high valency, calcium, aluminum, and hydrogen tend to impart high stability onto the aggregates as the clay domains remain stable and clay particles remain flocculated. Sodium and potassium, due to their large ionic radii and low valency, tend to cause clays to swell and disperse, which leads to various problems that can potentially limit plant growth. An exception occurs with Potassium fixation on some clay minerals, where potassium act as a very strong bonding agent. Despite being divalent, magnesium tends to have a higher ionic radius than calcium and tends to have an adverse effect on dispersion, particularly in combination with high sodium concentration.

In most soils, sodium is present only in small quantities in the soil, and these soils generally exhibit good aggregate stability. However, when the exchangeable sodium in the soil is present in sufficiently high quantities, it disperses a proportion of the clay domains, producing sufficient quantities of dispersed clay to block the soil's transmission pores to the extent that it affects other soil properties that limit plant growth, and the soil is then considered to be sodic.[3] Thus, sodic soils are also equated with dispersive soils, although dispersive soils are not necessarily sodic.

Sodic soils are found in every continent (Table 1).[3,4] High concentrations of sodium in the soil are generally associated with the presence of sodium salt in the soil or groundwater, in particular under arid or semiarid climatic conditions and sometimes under humid conditions.[5] High evaporative demand in combination with limited rainfall tends to concentrate sodium salts in the surface soil, which gives rise to high exchangeable sodium on the clay and colloidal fractions in the surface soil. Therefore, sodic soils are generally considered as a subset of saline soils.[3] It is a natural feature of particular soil types,

TABLE 1 Global Distribution of Sodic Soils

Continent	Estimated Area[a] (Million ha)	Land Area[b] (Million ha)	% of Land Area
North America	9.6	2,425.6	0.4
South America	57.9	1,781.9	3.2
Africa	27.1	3,006.5	0.9
Europe	22.9	993.8	2.3
Asia	121.9	4,457.9	2.73
Australia	340	768.7	44.2
Global total	579.4	13,434.4	4.3

[a] From Murphy.[3]
[b] From Col and Spector.[4]

but it may also be a consequence of human activities, such as irrigation practices and clearing of native vegetation on the upper slopes. Both activities give rise to increasing salinity associated with rising saline water tables and commonly referred to as secondary salinity. As irrigation practices are expanding globally to feed the growing population, and some with poor-quality water, secondary salinity and sodicity are becoming increasingly important as a production constraint. In Australia alone, approximately 2.5 million ha of arable land are affected by secondary salinity, costing an estimated $200 million in lost production annually.[6] Although sodic soils are found in every continent, they generally cover small proportions of the total land area, except on the Australian continent, where they occur over approximately 44% of the land area (Table 1).[3,7,8]

Measurement of Sodicity and Definition of Sodic Soil

The sodicity of the soil is most commonly measured as the exchangeable sodium percentage (ESP), which is given as

$$\mathrm{ESP} = 100\left(\text{exchangeable Na/CEC}\right) \tag{1}$$

where the cation exchange capacity (CEC) is generally determined at a reference pH of 7[9] or as the sum of exchangeable cations (Ca+Mg+Na+K) at a pH of 8.2.[10] These methods gave acceptable results for non-saline soils; however, for saline soils, the extraction process will result in dilution of the soil solution and change the ESP. A correction should be made by subtracting the soluble cations from the total extractable cations.[11]

Other less commonly used measures of sodicity are the exchangeable sodium ratio (ESR)[12] and exchangeable sodium content (ES)[13,14] of the soil, which are given as

$$\mathrm{ESR} = \text{exchangeable Na/exchangeable}\left(\mathrm{Ca} + \mathrm{Mg} + \mathrm{K}\right) \tag{2}$$

and

$$\mathrm{ES} = \text{exchangeable Na}\left(\text{cmole}\right)/\text{kg soil.} \tag{3}$$

As sodicity is inextricably linked to salinity, it is also important to define the sodium hazard of the irrigation water or soil solution commonly expressed as sodium adsorption ratio (SAR), which is given as

$$\mathrm{SAR} = \left[\mathrm{Na}\right]/\left(\left[\mathrm{Ca} + \mathrm{Mg}\right]/2\right)^{0.5} \tag{4}$$

where the brackets [ ] denote concentrations in mmol+/L.

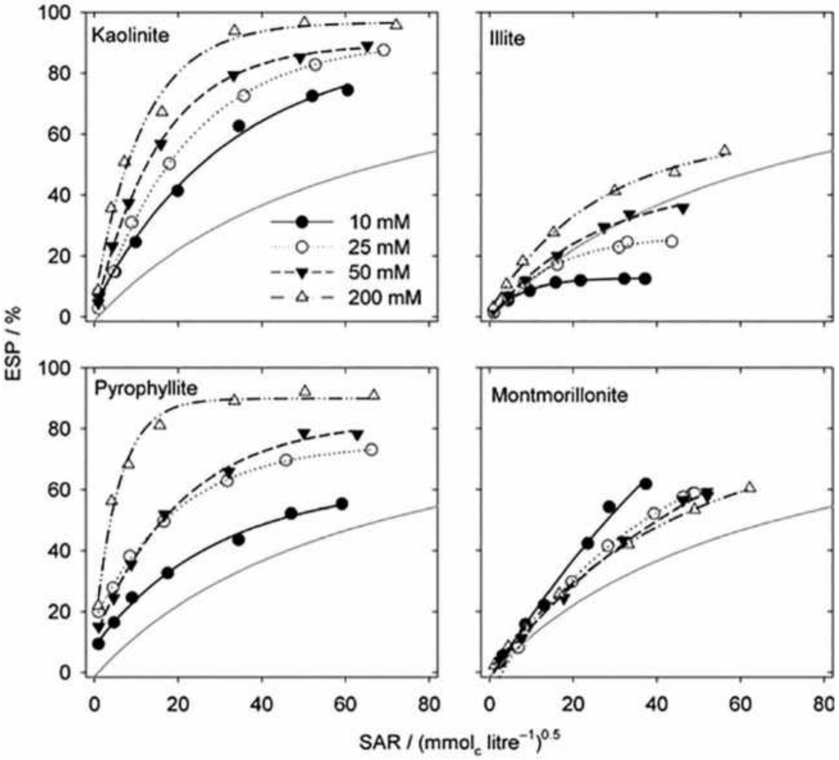


FIGURE 2 Changes in exchangeable sodium percentage (ESP) in four clay minerals (kaolinite, illite, pyrophyllite, and montmorillonite) following equilibration with solutions of various sodium adsorption ratios (SARs; as calculated by analysis of the entrained solution) at various ionic strengths. The solid line without data represents the average relationship for 59 salt-affected soils for the western United States.
Source: Adapted from Kopittke et al.[6]

As exchange reactions take place between the exchange phase and the soil solution, the ESP of the soil can be derived from the SAR of the soil solution as saturated paste, a 1:5 soil solution extract,[15] or irrigation water. The ESP–SAR relationship for all soils used to be assumed constant for all soils based on the soils of the Western United States,[12] but in reality, it is dependent on the clay mineralogy and the ionic strength, as shown in laboratory studies (Figure 2).[6]

In the selection of an appropriate parameter for sodicity, it should be noted that ESP is a parameter associated with the active clay and colloidal fraction of the soil and is useful when relating this parameter to other properties of that fraction, such as the dispersion ratio of clay (DRC), while ES is a bulk soil parameter and will be more suitable when we are concerned with bulk soil properties, such as dispersible clay content (DC) and hydraulic conductivity (K). The effect of increasing sodium on the amount of dispersed clay in 128 vertisols from the Darling Downs in Queensland, Australia, is shown in Figure 3.[13,16] Although it is common to relate sodium content as ESP to the amount of dispersed clay as DRC, a much better relationship was obtained for the same dataset when expressed as DC and ES, both expressed as a proportion of the bulk soil. Clearly, sodicity is better expressed as ES, at least for vertisols, and the regression is

$$\mathrm{DC} = 9.15 + 8.67(\mathrm{ES}) \quad \mathrm{R}^2 = 0.71^{***} \tag{5}$$

where *** indicates significance at p<0.001

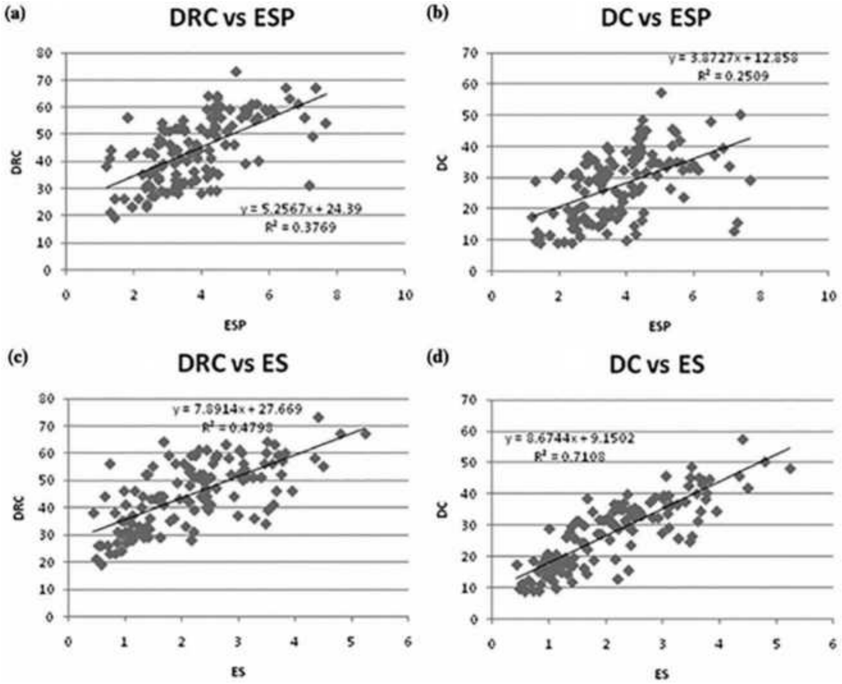


FIGURE 3 The relationship between dispersed clay and exchangeable sodium on 128 soil samples from 2 vertisols from the Darling Downs, Queensland, Australia. Dispersed clay is expressed as (a) dispersion ratio of clay (DRC) vs. exchangeable sodium percentage (ESP); (b) dispersed clay content(DC) vs. ESP; (c) DRC vs. ES; and (d) DC vs. ES.
Source: Adapted from Cook.[13]

For the same soils, DC and ES are strongly related to the soil's hydraulic conductivity, and all are bulk soil properties, while ESP would be a very poor predictor of hydraulic conductivity for these soils (Figure 4). A review of 15 published works to compare the abilities of ESP and ES to explain variation in soil structural stabilities across different soil types concluded that ES is a better index of sodium levels than ESP.[17]

How is a sodic soil defined? Generally, when the exchangeable sodium in the soil is present in sufficient quantities to influence its behavior, the soil is considered to be sodic.[18] However, defining whether a soil is sodic or not implies that there is a threshold ESP where the soil properties change dramatically. Although there is some evidence that this concept may apply in the laboratory,[19] the majority of work in the laboratory and in the field shows that there is little evidence of a threshold ESP and that sodium exerts a progressively increasing influence as ESP increases, as shown in Figure 4. However there is a need to develop a criterion to assist in the classification of soil as sodic if the probability of detrimental effects of sodicity is sufficiently high. The United States Salinity Laboratory (USSL) determined this criterion as an ESP of 15,[12] while in Australia,[20] the criterion is an ESP of 6 (highly sodic when the ESP reaches 15 or higher). Differences between the two criteria are due to the electrolyte levels of the water used at the time for deriving the ESP of the soils. The tapwater used in Riverside, California, United States, had a concentration of 3–10 mmol/L, while in Adelaide, Australia, the tapwater used at the time had a concentration of 0.7 mmol/L.[15,21,22] Furthermore, the Australian opinion was based on experience with many clay soils, whereas the U.S. experience was based on soils with textures lighter than sandy loam.[23] This again illustrates the strong interaction between clay type, electrolyte concentrations (or salinity in its extreme) and sodium levels, and this interaction is an important factor in the management of sodic soils.[24,25]

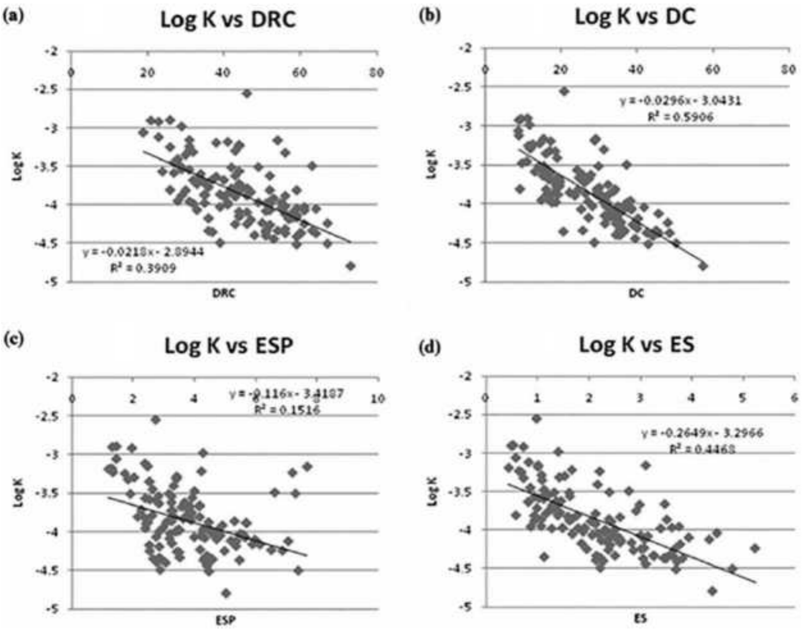


FIGURE 4 Soil hydraulic conductivity K as a function of (a) DRC, (b) DC, (c) ESP, and (d) ES for the two vertisols from Queensland, Australia, n = 128. Relationship is poor with ESP and good with ES.
Source: Adapted from Cook.[13]

Chemistry of the Clay–Cation Interaction and Its Effect on Swelling and Dispersion

In the soil, cations are attracted onto the negative clay surfaces, resulting in a diffuse double layer (DDL) where the cation concentrations near the surface are higher than in the bulk soil solution.[15,26,27] Conversely, the concentrations of anions near the clay surface are lower than in the bulk solution. The net positive charge of the DDL counteracts the net negative charge of the clay surfaces, thereby maintaining electro-neutrality of the system. The thickness of the DDL represents the hydration layer of the clay surface and is strongly affected by the valency of the cations in the DDL and the ionic concentration of the bulk solution. The DDL determines the degree of swelling and whether the clay domains disperse or remain flocculated. The DDL tends to be expanded with low-valency cations such as Na or K and/or low ionic concentration of the bulk solution. On the other hand, it tends to be compressed with high-valency cations such as Ca, Mg, and Al and/or high ionic concentrations in the bulk solution.

In homoionic montmorillonitic clay systems, Ca-clay systems exhibit limited swelling, with a maximum water content of 0.8 cm^3/g, and remain flocculated. Na-clay swells considerably more than Ca-clay (Figure 5), and this is shown by the very high water content (>1 cm^3/g) at a pF of 1 and a solution concentration of 0.01 mol NaCl/L or less, which is approximately similar to normal soil solution concentrations.[18,28] Swelling in Na-clay increases rapidly as the electrolyte concentration decreases and clay particles become detached from each other and exhibit spontaneous dispersion. Figure 5 shows clearly the combined effect of the valency of the dominant cation and the electrolyte concentration of the soil solution on the swelling and dispersion of the clay particles. In summary, high Sodium and low EC represent excessive swelling and dispersion, while high Ca and high EC represent limited swelling and flocculation.

In field soils, the clay fraction is associated with a mixture of cations, which is generally dominated by Ca and Mg, with a smaller complement of Na and K ions, and Al ions in acid soils with a pH less

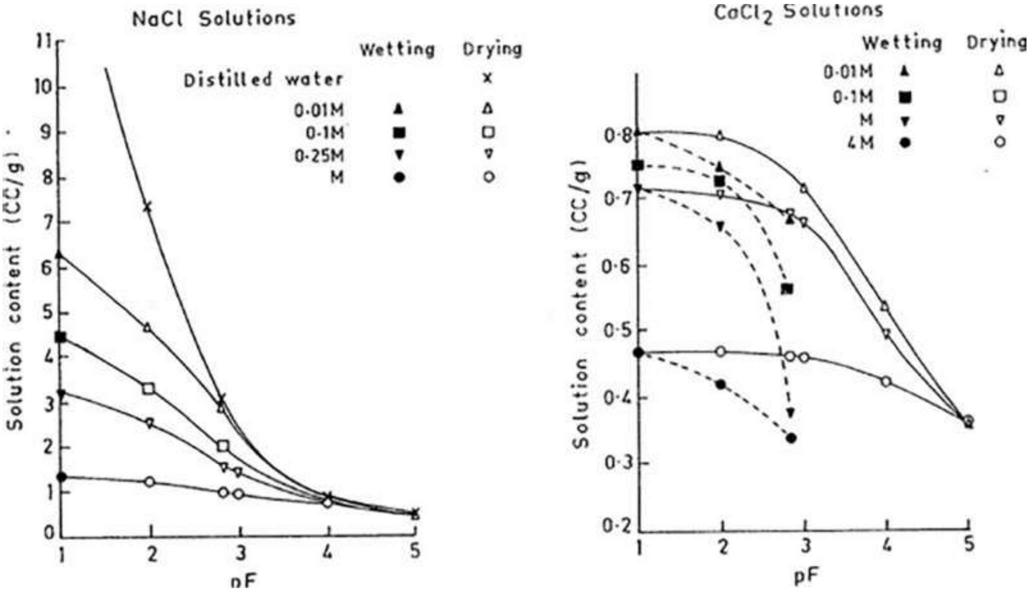


FIGURE 5 Swelling of Na and Ca montmorillonite clay, as a function of solution content and water potential expressed as pF.
Source: Adapted from Aylmore and Quirk.[28]

than 5.5. The principal effect of a high proportion of sodium in field soils is the swelling of some of the clay fraction following wetting associated with the hydration of the Sodium ions, which leads to slaking of the aggregates and dispersion of a proportion of the clay size fraction. These dispersed materials are effective in the blockage of pores, leading to a reduction in transmission pores (>15 µm corresponding to field capacity), which in turn affects the hydrologic and mechanical properties and behavior of the soil.[2] The severity of the effect of dispersion is partly dependent on the clay content of the soil; a small percentage of dispersion can result in large decreases in soil's hydraulic conductivity.[29] At lower clay contents (<30%), the soil matrix is sandy with inherently large transmission pores and high hydraulic conductivities. In contrast, soils with higher clay contents (>30%) have a clayey matrix, and the number of transmission pores is determined by the degree of structural development. The same degree of dispersion tends to have a greater effect on soils with a clayey matrix compared with soils with a sandy matrix.

The combined effect of electrolyte concentration and sodium on clay dispersion and hydraulic conductivity was first established by Quirk and Schofield[24] on illite dominated soils. They used the reduction in hydraulic conductivity to separate soils that are unstable and likely to be affected by dispersion from soils that are stable. This was later refined using actual dispersion measurements and resulted in a classification scheme for the illite dominated red-brown earths of southern Australia as shown in Figure 6.[25] This classification used the SAR of a 1:5 soil water extract as a measure of the exchangeable sodium of the soil, as this is strongly related to ESP, and the total cation concentrations (TCCs) in mmol +/ L as the measure of electrolyte concentration. It classifies soils as dispersive soils, potentially dispersive soils, and flocculated soils. The spontaneously dispersive soils have high exchangeable sodium levels and low electrolyte concentrations, except where SAR is less than 3, as these soils do not spontaneously disperse, because its TCC is always above 0.6 mmol+/L and requires a small external energy to disperse. The potentially dispersive soil requires an external energy to disperse the soil, e.g., cultivation or raindrop impact. The flocculated soil does not disperse even with the input of external energy. Similar classification can be developed for other soil types. In Figure 6, the dotted line through SAR 3 was drawn to indicate that in low-sodium soils, dispersion is less likely to be a problem.

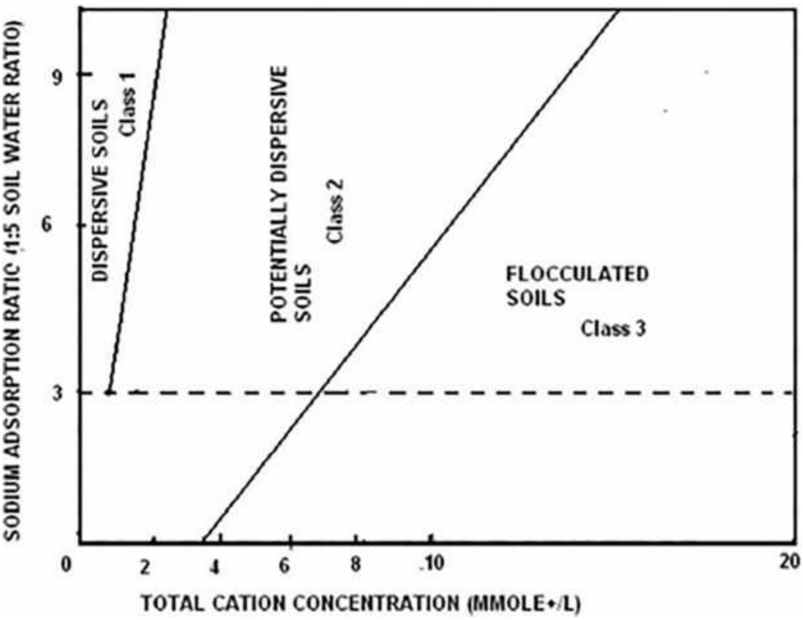


FIGURE 6 A classification scheme for the prediction of the behavior of the A-horizon of a red-brown earth.
Source: Adapted from Rengasamy.[7]

Effect of Sodicity on the Soil's Physical Property and Behavior

Sodic soils are structurally unstable, and the effect on the soil's physical property and behavior is shown conceptually in the diagram in Figure 7.

Slaking and dispersion are the fundamental processes of degradation of structurally unstable sodic soils. Slaking alone is generally not adequate to cause deleterious effects (self-mulching vertisols in the virgin state are highly productive soils, (they slake when wetted), but when followed by dispersion, it generally results in a decrease in soil's productivity. It is important to note here that dispersion for class 1 and 2 soils in Figure 6 is strongly affected by an external energy input, such as rapid wetting, cultivation under wet plastic conditions, or raindrop impact. It is also strongly affected by the electrolyte concentration and the sodium–calcium ratio of the soil solution. Rapid wetting (on surface soils under rainfall or irrigation) is a particularly destructive force associated with the rapid release of heat of wetting, which pushes the clay domains apart. However, when wetting is done slowly, the integrity of the domains and aggregates are often maintained, and the soils are able to maintain their transmission pores. Slow wetting occurs in the subsoil, on soil under a reasonably thick mulch or under low-energy/low-output spray irrigation (e.g., fine spray or mist).

In the field, sodicity is most problematic when it occurs on the surface, where the soil is almost always subjected to rapid wetting from rainfall. Sodicity in the subsoil is less of a problem and is much easier to manage, and these subsoils should not be exposed to the surface unless it is ameliorated.

Effect on the Hydrologic Properties and Behavior of Soils

The first and foremost effect of slaking and dispersion is the formation of a surface seal and the subsequent reduction of the soil's hydraulic conductivity (K) associated with blockage of transmission pores between the larger particles or stable aggregates. A small increase in dispersion may result in large

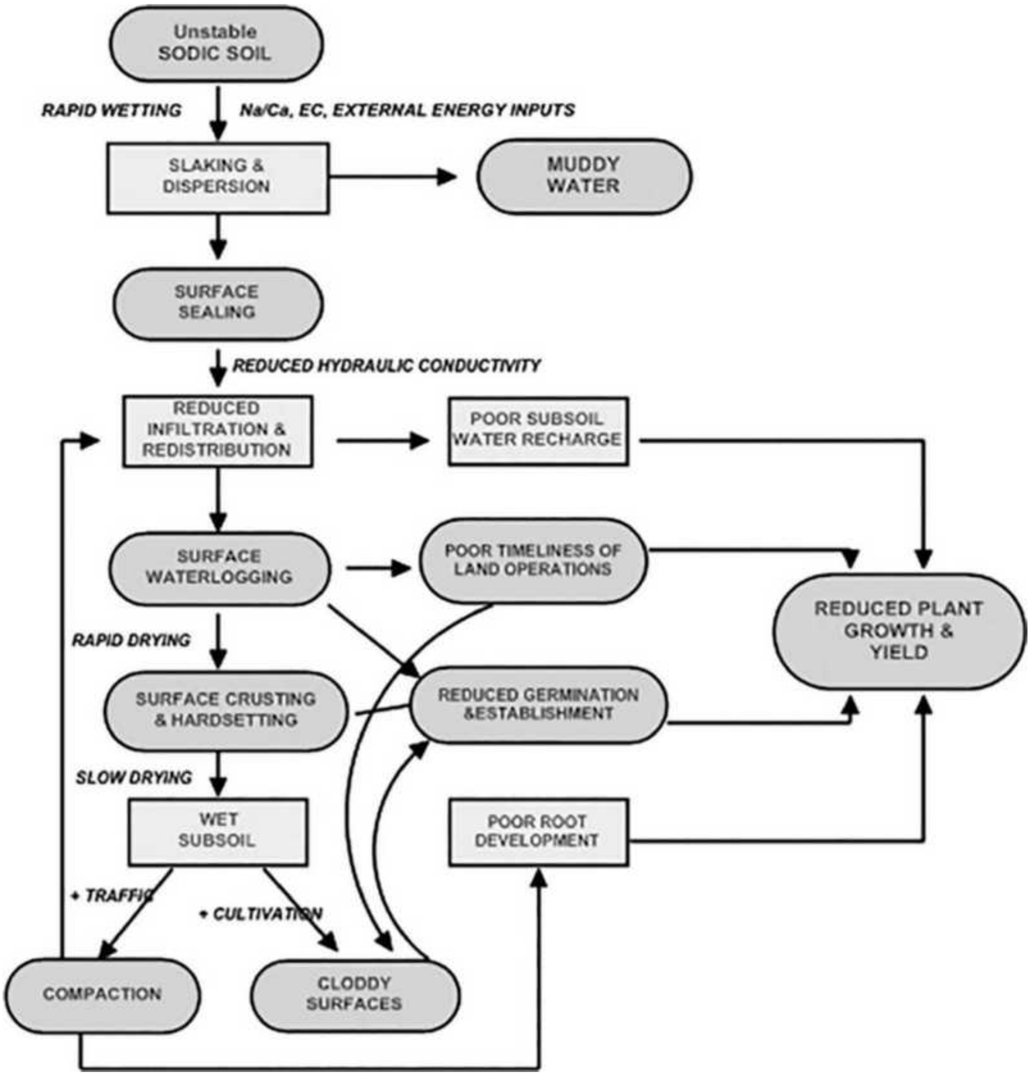


FIGURE 7 Diagram showing how slaking and dispersion associated with excessive sodium results in a range of seemingly independent soil physical problems in the field.
Source: Adapted from So.[2]

decreases in K[16,29] These processes are associated with many soil physical problems observed in the field that affect plant growth or soil behavior, shown in shaded oval shapes in Figure 7.[15,22] Waterlogging, low infiltration, crusting, hardsetting, poor timing of farming operations, and poor germination and seedling establishment are generally perceived as distinctly different problems in the field, but as shown in Figure 7, they have a common cause in slaking and dispersion and can therefore be corrected using the same treatment. This also applies to non-sodic dispersive soils. Hence, an ameliorative treatment applied to reduce or prevent dispersion and to correct one particular problem will also correct many of the other physical problems shown in Figure 7.

The degree of slaking and dispersion can be readily assessed in the field with the Emmerson dispersion test[30] or its modified versions.[31,32] These quick field tests rely on visual assessments and scoring of the degree of slaking and dispersion (Plate 1) of natural aggregates or remolded aggregates in water. Remolding is conducted at field capacity to simulate the application of external energy such as tillage under wet conditions or raindrop impact. The resulting scores from these tests have been successfully related to crop yields.[33,34]

Field assessment of the extend of degradation associated with sodium is generally based on whether there are detrimental effects on the soil's productivity, which is principally affected by the amount of plant-available soil water (ASW) over the crop growing season. The ASW is the difference between water input, controlled by the soil's infiltration rate, and the water output from the root zone, controlled by the drainage rate, and both rates are affected by the soil's hydraulic conductivity. As K decreases, the soil's infiltration and drainage rates also decrease. Hence, as sodicity increases, the recharge of the soil water storage maybe reduced but may initially remain adequate for the requirement of the plants. However, at the critical ESP or above, the reduction in K becomes sufficiently large to reduce the subsoil water recharge (Figure 7) or the amount of available soil water storage such that crop growth becomes affected. Therefore, the critical ESP should vary with the prevailing seasonal rainfall, irrigation, and soil type,[34] but a general relationship between these factors has yet to be quantified. As mentioned earlier, the most common critical level of sodicity used to indicate the probability of such reduction is ESP of 15% in the United States and 6% in Australia.

A reduction in the rate of drainage in sodic soils will increase the probability of surface waterlogging following rainfall or irrigation, which may significantly reduce germination and establishment of seedlings (see Figure 7).

Effect on the Mechanical Properties and Behavior of Soils

Surface waterlogging would reduce soil strength and its ability to support mechanical loads, and therefore, its trafficability is reduced. This increases the time when the soil is in a plastic state and not suitable for tillage; therefore, its workability is also reduced. These two conditions lead to poor timeliness of land operations (Figure 7).

Upon drying, the surface seal forms a crust or a hardsetting surface depending on the texture and structural status of the soil. At this stage, drying would occur at a high rate, and the thickness of the crust will depend on this rate of drying, being thicker the slower the rate of drying. The thickness of the crust will also depend on the size of the dry aggregates.[35] Furthermore, the higher the ESP, the harder is the crust formed, as shown in Figure 8.[18,36] The development of surface crusting leads to a reduction in the evaporation rate, leading to a subsoil below the crust that remains wet for longer periods. When the soil is subjected to traffic under this condition, the load transmitted to this wet subsoil will lead to compaction, while cultivation leads to cloddy surface conditions, which are not conducive to seed germination and establishment. This in turn leads to reduced crop growth and yield.

In contrast, fine sandy or silty soils are dominated by particles that are sufficiently large to have very low or no surface activity and therefore tend to remain as individual particles. On the other hand, they are small enough to block transmission pores. Therefore, soils with a high content of fine sand and silt in the surface horizon are inherently unstable; their aggregates readily slake and disperse throughout this horizon when wet, particularly when they are also sodic. Under wet conditions, these soils are very soft and highly susceptible to deformation and compaction. Because of the low water holding capacity associated with the predominance of coarse particles, these soils rapidly dry to create a hardsetting soil and are sometimes referred to locally as a "Sunday soil," being too wet for cultivation on Saturday and too dry on Monday. As conditions suitable for cultivation are restricted, sodic hardsetting soils tends to give rise to cloddy seedbeds. Cultivation when it is too wet will compress the soils into large clods, while cultivation when it is too dry with a high tensile strength will also result in cloddy seedbeds. Both conditions will require additional cultivation and energy to create a suitable seedbed. In practical terms, this means increased cost of production.

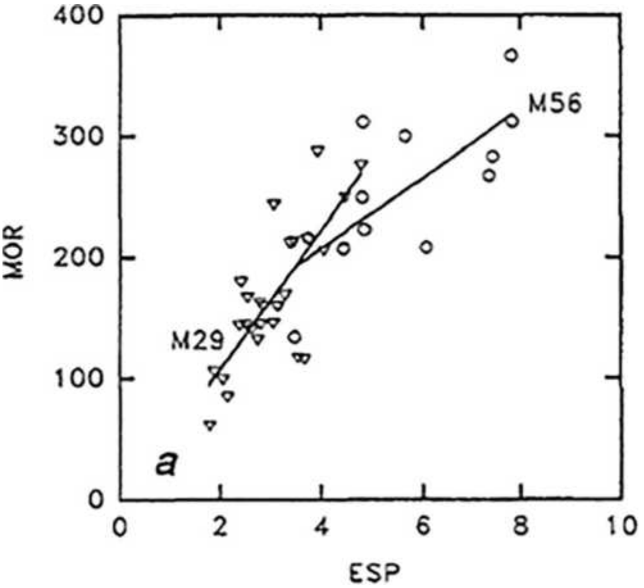


FIGURE 8 Modulus of rupture (MOR), a measure of crust strength, as a function of ESP for two clay soils from Western Australia.
Source: Adapted from So.[2]

High ESPs are associated with lower liquid and plastic limits and would require more energy to cultivate[37] and lead to the perception of a soil with a heavier texture than indicated by its particle size distribution. The use of gypsum to ameliorate the effect high ESP on clay soils increases the ease of tillage due to a reduction in soil strength.[38]

Management of Sodic Soils

The management of sodic soils is fundamentally based on the management of the slaking and dispersion processes in the soil. Table 2 lists a range of management options that have been used in the field to control these processes.

Slaking can be managed at the macroaggregate level by avoiding or minimizing the breakdown of existing bonds (organic matter) through minimizing soil disturbance (zero or minimum tillage practices) and/or through the addition of surface mulch or aggregate bonding materials such as organic matter,[39] polymers [polyvinyl alcohol (PVA), polyacrylamide (PAM)], and calcium carbonates (lime). It can also be achieved by incorporating grasses into the existing crop rotation as grass roots and associated fungal hyphae bind microaggregates into macroaggregates.[40]

On the other hand, dispersion can be managed at the clay domain and microaggregate level by using divalent cations and/or high EC in the soil solution or a combination of the two. The most popular ameliorant used is gypsum ($CaSO_4.2H_2O$), which provides Ca^{2+} and a continuous presence of adequate EC levels associated with its limited solubility (2.2dS/m in water at 25°C). The solubility of gypsum in soils is less than in water[41] and is dependent on the particle size of gypsum. Ca^2+ replaces Na^+ from the clay exchange sites (the exchangeable cation effect), which represents a permanent or residual effect of gypsum. The presence of a high electrolyte concentration where free gypsum is present provides the dominant initial effect of gypsum. It represents the temporary electrolyte effect of gypsum.[41] Lime will provide a similar effect but is a much slower acting ameliorant, while slaked lime is a fastacting soluble ameliorant.

TABLE 2 Strategies for Managing Slaking and Dispersion

Problem	Aggregate Hierarchical Level Affected	Action Required	Management Options
Slaking	Macroaggregates	Prevent breakdown of bonds	1. Zero tillage or minimum tillage 2. Surface mulch 3. Slow wetting during irrigation 4. Reduce droplet energy during spray irrigation
		Increased aggregate bonding and bonding strength	1. Incorporate organic matter 2. PVA 3. PAM 4. Lime or calcium carbonates 5. Introduce grasses in crop rotation
Dispersion	Microaggregates and clay domains	Increase flocculation by replacing monovalent cations with di- or polyvalent cations (exchangeable cation effect)	1. Gypsum application (S or H_2SO_4 application produce $CaSO_4$ in the soil) 2. Lime application (slow response) 3. Slaked lime [$Ca(OH)_2$] application (quick response) 4. $CaCl_2$ application (expensive and mostly confined to laboratory trials)
		Increase flocculation by increasing electrolyte concentrations (temporary electrolyte effect)	1. Gypsum application 2. Slaked lime [$Ca(OH)_2$] application 3. $CaCl_2$ application

Source: Adapted from So.[2]

Conclusion

Sodic soils are generally associated with the presence of high concentrations of sodium salts and considered as a subset of saline soils. The presence of high sodium concentrations may be a natural feature of the landscape but can also be associated with human activities, in particular, clearing of native vegetation on the upper slopes of the landscapes and irrigation. The need for increasing food production to feed the growing population leads to the conversion of marginal lands, including naturally sodic soils, and the introduction of more irrigation with poor-quality water.

The fundamental problem associated with sodic soils has been shown to be related to the basic processes of slaking and dispersion, which in turn may lead to a variety of field problems that affect plant growth and yield, such as crusting and hardsetting, waterlogging, lack of subsoil water, compaction, cloddy surfaces, poor germination, and seedling establishment. Sodic soils can be managed effectively at the macro scale to minimize or prevent slaking through addition of aggregate bonding materials (organic matter, PVA, PAM, lime, or $CaCO_3$) or to slow the process of wetting. Alternatively, it can be managed at the micro scale by targeting the dispersion process with the use of divalent cations in combination with high electrolyte levels (gypsum, slaked lime).

References

1. Dexter, A. Advances in characterisation of soil structure. Soil TillageRes. **1988**, *11*, 199–238.
2. So, H. Slaking, Dispersion and crust formation. In *Encyclopedia of Soil Science*; Lal, R., Ed. Marcel Dekker: New York & Basel, 2002; 1206–1212.
3. Murphy, B. Sodic soils, formation and global distribution of. In *Encyclopedia of Soil Science*; Lal, R., Ed.; Marcel Dekker: New York, Basel, 2002; 1213–1217.

4. Col, J.; Spector, M. The Continent: Land area, available at http://www.enchantedlearning.com/geography/continents/Land.shtml.
5. Indurante, S. Sodium-affected soils in humid areas. In *Encyclopedia of Soil Science*; Lal, R., Ed.; Marcel Dekker: New York, Basel, 2002; 1229–1232.
6. Kopittke, P.; So, H.; Menzies, N. Effect of ionic strength and clay mineralogy on Na–Ca exchange and the SAR-ESP relationship. Eur. J. Soil Sci. **2005**, *57,* 626–633.
7. Rengasamy, P. Sodic soils. In *Encyclopedia of Soil Science*; Lal, R., Ed.; Marcel Dekker: New York, Basel, 2002; 1210–1212.
8. Bui, E.; Krogh, L.; Lavado, R.; Nachtergaele, F.; Toth, T.; Fitzpatrick, R. Distribution of sodic soils: The world scene. In *Sodic Soils: Distribution, Properties, Management and Environmental Consequences;* Sumner, M., Naidu, R., Eds. Oxford University Press: New York, 1998; 19–33.
9. Thomas, G. Exchangeable cations. In *Methods of Soil Analysis;* Page, A., Miller, R., Keeney, D., Eds.; American Society of Agronomy: Madison, 1982; Part 2.: Chemical and Microbiological Properties. Second Edition. 159–166.
10. Rhoades, J. Cation exchange capacity. In *Methods of soil analysis,* Page, A., Miller, R., Keeney, D., Eds.; American Society of Agronomy: Madison, WI, 1982; Vol. Part 2.
11. So, H.; Menzies, N.; Bigwood, R.; Kopittke, P. Examination into the accuracy of exchangeable cation measurement in saline soils. Commun. Soil Sci. Plant Anal. **2006**, *37,* 1819–1832.
12. Richards, L. *Diagnosis and Improvement of Saline and Alkali Soils;* USDA: Washington, 1954.
13. Cook, G. *Degradation of Soil Structure under Continuous Cultivation on the Darling Downs, Queensland*; The University of Queensland, Brisbane, 1988.
14. Cook, G.; So, H.; Dalal, R. Structural degradation of two vertisols under continuous cultivation. Soil Tillage Res. **1992**, *24,* 47–64.
15. Sumner, M. Sodic soils: New perspectives. Aust. J. Exp. Agric. **1993**, *31,* 683–750.
16. So, H.; Cook, G. The effect of slaking and dispersion on the hydraulic conductivity of soils. In *Catena Supplement*; Poessen, J., Nearing, M., Eds.; 1993; Vol. 24, 55–64.
17. Cook, G.D.; Muller, W.J. Is exchangeable sodium content a better index of soil sodicity than exchangeable sodium percentage?: A reassessment of published data. *Soil Sci.* **1997**, *162* (5), 343–349.
18. So, H.B. Sodic soils: Physical properties and behavior. *Encyclopedia of Soil Science,* 2nd Ed.; **2008**; 1–4.
19. Shainberg, I.; Caiserman, A. Studies on Na/Ca montmorillonite systems. 2. The hydraulic conductivity. Soil Sci. **1971**, *111*, 277–281.
20. Northcote, K.; Skene, J. *Australian Soils with Saline and Sodic Properties.* CSIRO, 1972.
21. Shainberg, L.; Letey, J. Response of soils to sodic and saline conditions. Hilgardia **1984**, *52*, 1–57.
22. So, H.; Aylmore, L. How do sodic soils behave? The physical properties of sodic soils. Aust. J. Soil Res. **1993**, *31,* 761–777.
23. McIntyre, D. Exchangeable sodium, subplasticity and hydraulic conductivity of some Australian soils. Aust. J. Soil Res. **1970**, *17,* 115–120.
24. Quirk, J.; Schofield, R. The effect of electrolyte concentration on soil permeability. J. Soil Sci. **1955**, *6*, 163–178.
25. Rengasamy, P.; Greene, R.S.B.; Ford, G.W.; Mehanni, A.H. Identification of dispersive behaviour and the management of red-brown earths. Aust. J. Soil Res. **1984**, *22* (4), 413–431.
26. Quirk, J. Interparticle forces: A basis for the interpretation of soil physical behaviour. Adv. Agron. **1994**, 53, 121–183.
27. Churchman, G. Sodic soils, reclamation of. In *Encyclopedia of Soil Science;* Lal, R., Ed.; Marcel Dekker: New York, Basel, 2002; 1224–1228.
28. Aylmore, L.; Quirk, J.P. The structural status of clay systems. Clays Clay Miner. **1962**, *9*, 104–130.
29. So, H.; Tayler, D.; Yates, W.; McGarity, J. Amelioration of structurally unstable grey and brown clays. In *Modification of Soil Structure;* Emmerson, W., Bond, R., Dexter, A., Eds.; Wiley: Chichester, 1978.

30. Emmerson, W. Physical properties and structure. In *Soil Factors in Crop Production in a Semi-Arid Environment;* Russell, J., Graecen, E., Eds.; University of Queensland Press: St Lucia, 1977.
31. Loveday, J.; Pyle, J. *The Emmerson Dispersion Test and its Relationship to Hydraulic Conductivity;* Canberra, 1973; 1–7.
32. Field, D.J.; McKenzie, D.C.; Koppi, A.J. Development of an improved vertisol stability test for SOILpak. Aust. J. Soil Res. **1997**, *35* (4), 843–852.
33. So, H.; Onus, G. The extent of unstable grey and brown clays in the Lower Gwydir Valley, New South Wales, and its relationship to wheat yields. In *The Properties and Utilisation of Cracking Clay Soils;* McGarity, J., Hoult, E., So, H., Eds.; University of New England: Armidale, NSW, Australia, 1984; 132–135.
34. McKenzie, D.C.; Bernardi, A.L.; Chan, K.Y.; Nicol, H.I.; Banks, L.W.; Rose, K.L. Sodicity v. yield decline functions for a vertisol (Grey Vertosol) under border check and raised bed irrigation. Aust. J. Exp. Agric. **2002**, *42* (3), 363–368.
35. Collis-George, N.; Greene, R. The effect of aggregate size on the infiltration behaviour of a slaking soil and its relevance to ponded irrigation. Aust. J. Soil Res. **1979**, *17*, 65–73.
36. Cochrane, H.; Aylmore, L. Assessing management-induced changes in the structural stability of hardsetting soils. Soil Tillage Res. **1991**, *20,* 123–132.
37. Guarnieri, A.; Fabbri, A.; Molari, G. Influence of sodicity and salinity on the mechanical properties of two Italian soils. Biosyst. Eng. **2005**, *91,* 239–243.
38. McKenzie, D.; So, H. Effect of gypsum on vertisols of the Gwydir Valley, New South Wales. 2. Ease of tillage. Aust. J. Exp. Agric. **1989**, *29*, 63–67.
39. Clark, G.; Sale, P.; Tang, C. Organic amendments initiate the formation and stabilisation of macroaggregates in a high clay sodic soil. Aust. J. Soil Res. **2009**, *47,* 770–780.
40. Tisdall, J.; Oades, J. Stabilisation of soil aggregates by the root system of rye grass. Aust. J. Soil Res. **1979**, *17,* 429–441.
41. Loveday, J. Rerlative significance of electrolyte and cation exchange effects when gypsum is applied to a sodic soil. Aust. J. Soil Res. **1976**, *14*, 361–371.

25

Sodic Soils: Properties

Pichu Rengasamy

Introduction

Sodic soils are those with a high proportion of sodium adsorbed to the soil particles compared to other cations such as Ca, Mg, and K. Exchangeable sodium percentage (ESP) is a criterion used to identify sodic soils. When a soil with high ESP is wetted, the bonds between soil particles are weakened. As a result, the clay particles swell and often become detached and disperse at high water content. This leads to poor physical properties of soils. In classification system developed in U.S.A., a soil with an ESP >15 is considered as "sodic," and in Australia, the criterion is >6. These differences are owing to the influence of several soil factors including salt levels, pH, organic matter, and clay mineralogy on the adverse effects of ESP on soil properties. Sodic soils are widespread in arid and semiarid regions of the world extending up to 30% of the total land area (Table 1). Use of saline water, including waste and effluent waters containing sodium salts, for irrigation induces sodicity in soils. Sodicity is a latent problem in many salt-affected soils where deleterious effects on soil properties are evident only when salts are leached below a threshold level.[2] While soil salinity reduces plant growth, directly affecting physiological functions through osmotic and toxicity effects on plants, sodicity causes deterioration of soil physical properties indirectly affecting plant growth and survival. Sodic soils are subjected to severe structural degradation and exhibit poor soil-water and soil-air relationships; these properties (Table 2) adversely affect root growth, thereby restricting plant production and making it difficult to work in soils when they are wet or dry.

Yield Decline in Sodic Soils

Sodic soils make the paddocks prone to waterlogging, poor crop emergence and establishment, gully erosion, and in some instances tunnel erosion. Because of the heterogeneity in the accumulation of sodium by soil particles, these symptoms may be observed only in certain parts of the paddock. Generally, patchy growth and barren patches are visible in a number of spots in a paddock, while the rest of the field may look normal. However, the effects of sodicity are fully realized in the harvested yield. The actual yield obtained in sodic soils is often less than half of the potential yield expected on the basis of climate,

TABLE 1 World Distribution of Sodic Soils

Continent	Country	Area of Sodic Soils (000 ha)
North America	Canada	6,974
	U.S.A.	2,590
South America	Argentina	53,139
	Bolivia	716
	Brazil	362
	Chile	3,642
Africa	Algeria	129
	Angola	86
	Botswana	670
	Cameroon	671
	Chad	5,950
	Ethiopia	425
	Ghana	118
	Kenya	448
	Liberia	44
	Madagascar	1,287
	Namibia	1,751
	Niger	1,389
	Nigeria	5,837
	Somalia	4,033
	Sudan	2,736
	Tanzania	583
	Zimbabwe	26
South Asia	Bangladesh	538
	India	574
	Iran	686
North and Central Asia	China	437
	U.S.S.R.	119,628
Australasia	Australia	339,971

Source: Bui et al.[1]

particularly rainfall and evapotranspiration.[3,4] Relative yield of cereals grown in dryland sodic soils in Australia in relation to average root zone ESP is given in Figure 1.

Swelling and dispersion of sodic aggregates destroy soil structure, reduce the porosity and permeability of soils, and increase the soil strength even at low suction (i.e., high water content). These adverse conditions restrict water storage and transport. Soils are, therefore, either too wet immediately after rain or too dry within a few days for optimal plant growth. Thus, the range of soil water content that does not limit plant growth and function ("nonlimiting water range") is very small.[5] Dense, slowly permeable sodic subsoils reduce the supplies of water, oxygen, and nutrients needed for obtaining maximum potential yield. During the rainy season, even with prolonged ponding of water on the surface, only a small increase in water content occurs in subsoil. The low porosity leads to slow internal drainage and water redistribution within the profile.[6] This reduction in water storage causes crop water stress during prolonged dry periods. Subsoil as a source of water and nutrients becomes more important in dryland cropping regions than in irrigated soils.

TABLE 2 Physical and Chemical Properties of a Typical Sodic Soil Profile in South Australia

Properties	0–20 cm	20–40 cm	40–100 cm
Chemical properties			
$pH_{1:5}$ (water)	7.9	8.9	9.2
EC_e (dS/m)	0.4	3.8	4.9
Organic carbon (%)	1.2	0.6	0.3
Exchangeable sodium (%)	6.2	14.6	24.5
$CaC0_3$ (%)	0.1	2.8	4.5
Boron (mg/kg)	1.2	22.0	38.5
Water soluble $Al(OH)_4^-$ (mg/kg)	0.0	1.2	2.6
Physical properties			
Spontaneously dispersed clay (%)	1.2	8.6	9.4
Swelling (mm/mm)	0.04	0.18	0.20
Hydraulic conductivity at saturation (mm/day)	22.8	4.5	2.3
Penetrometer resistance at lOOkPa suction (MPa)	1.8	4.2	4.8
Aeration porosity (%)	9.7	4.8	3.9
Bulk density (Mg/m^3)	2.0	2.2	2.3
Final infiltration rate in the field (mm/hr)	0.2		

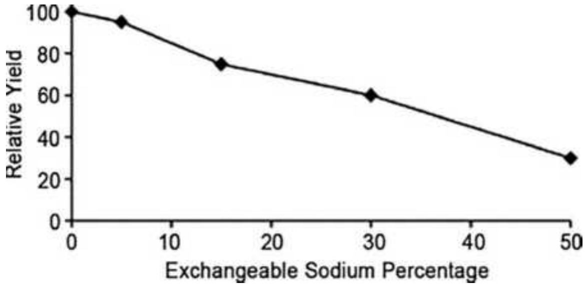


FIGURE 1 Relative yield of cereals grown in Australian sodic soils in relation to average root zone ESP.

Salt Accumulation in Root Zones of Sodic Soils

Soils with sodic subsoils are characterized by moderate to high exchangeable sodium and, in many cases, with high pH (>8.5) where carbonate and bicarbonate minerals are present. Subsoil sodicity restricts drainage beyond the root zone and as a result salts accumulate in this zone. The concentration of accumulated salts fluctuates with rainfall pattern, input of salt from agronomic practices, and soil weathering, as schematically explained in Figure 2. Dryland salinity or “seepage salinity” in many countries is associated with rising saline groundwater tables. However, the extent of subsoil salinity, also called “transient salinity,” not associated with saline groundwater is large in many landscapes dominated by subsoil sodicity. A relationship between rainfall, subsoil ESP, and ECe for northeastern Australian soils has been reported.[7] In dryland regions with annual rainfall between 250 and 600 mm, sodic subsoils have an EC_e between 2 and 20 that can dramatically affect crop production through osmotic effects during dry periods. Laboratory measured EC_e will increase several folds under field conditions as the soil layers dry in between rainy days. The combination of poor water storage and osmotic stress enhance water stress to crops under dryland cropping.

Root Zone Constraints in Dryland Sodic Soils

Multiple problems occur in soils with subsoil sodicity. Soil compaction, crusting, and induration of subsoil layers re quire “physical” reclamation. Sodicity, salt accumulation, and alkaline pH require “chemical” reclamation. All of these conditions cause, in addition to water stress, macro- and micronutrient deficiency, and toxicity owing to Na^+, Cl^-, HCO_3^-, CO_3^{2-}, B, $Al(OH)_4^-$, and others. Low organic matter and biological activity compound these problems encountered in sodic subsoils.

Management of Dryland Sodic Soils

Major criteria in increasing productivity in dryland sodic soil are improved water storage and transport in the root zone and crop water use efficiency. More information is available on agricultural management in sodic soils that is more relevant to irrigated lands.[6] Reclamation procedures involving high costs are prohibitive in dryland regions because of low benefit/cost ratio.

Diagnosis of multiple problems with large variations, vertically and horizontally across the paddock, is primarily important. Gypsum is the most commonly used compound to reclaim sodic soils. Subsoil reclamation may involve higher rates of gypsum application or deep placement of gypsum by deep ripping or deep plowing. Salt-tolerant plant species may alleviate subsoil salinity. Plants that can tolerate ion toxicity such as boron, carbonate, sodium, and chloride have also been identified. Strategies to improve subsoil fertility may include 1) mechanical means of placing nutrients deeper in the profile; 2) using nutrient sources of lower or higher mobility; 3) using deep-rooted legumes to fix nitrogen at depths; and 4) selection of plant species and genotypes better suited to acquiring nutrients from subsoils. Future research is needed on developing plants that modify the rhizosphere and adapt to edaphic conditions. Farming systems should be developed to prevent accumulation of salts and toxic elements in the root zone of sodic soils.

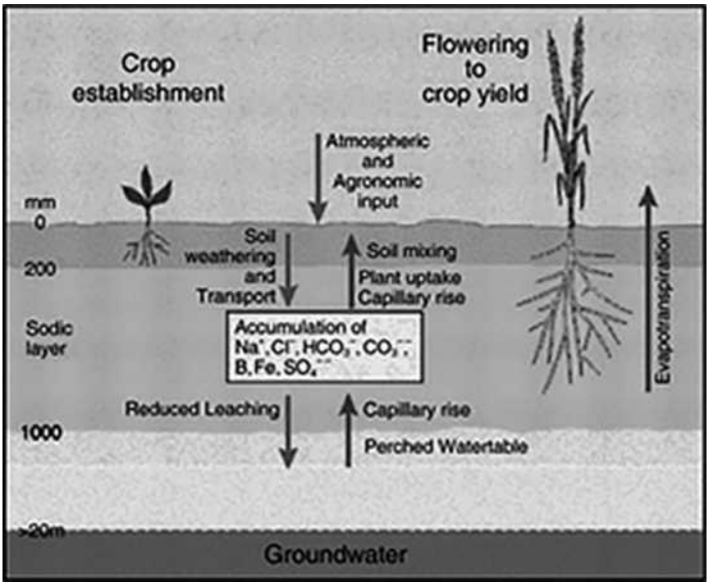


FIGURE 2 Salt accumulation in sodic subsoils.

Conclusions

Soils with a high proportion of exchangeable sodium are considered as sodic soils. On wetting these soils, clay particles swell and disperse degrading soil structure, and on further drying soils become dense. Poor water storage and restricted movement of air and water in soil profile lead to yield decline of many crops. Sodic subsoils cause accumulation of salts in the rootzone. In dryland sodic soils multiple problems such as salinity and toxic elements in addition to sodicity occur making management decisions difficult. Soil amelioration with gypsum and choice of tolerant crops are useful in farming these soils.

References

1. Bui, E.N.; Krogh, L.; Lavado, R.S.; Nachtergaele, F.O.; Toth, T.; Fitzpatrick, R.W. Distribution of sodic soils: the world scene. In *Sodic Soils: Distribution, Properties, Management and Environmental Consequences*; Sumner, M.E., Naidu, R., Eds.; Oxford University Press: New York, 1998; 19–33.
2. Rengasamy, P.; Olsson, K.A. Sodicity and soil structure. Aust. J. Soil Res. **1991**, *31,* 821–837.
3. French, R.J.; Schultz, J.E. Water use efficiency of wheat in a mediterranean-type environment. 1. The relation between yield, water use and climate. Aust. J. Agric. Res. **1984**, *35*, 765–775.
4. Rengasamy, P. Sodic soils. In *Methods for Assessment of Soil Degradation*; Lal, R., Blum, W.H., Valentine, C., Stewart, B.A., Eds.; CRC Press: New York, 1997; 265–277.
5. Letey, J. The study of soil structure: science or art. Aust. J. Soil Res. 1991, *29,* 699–707.
6. Oster, J.D.; Jayawardane, N.S. Agricultural management of sodic soils. In *Sodic Soils: Distribution, Properties, Management and Environmental Consequences*; Sumner, M.E., Naidu, R., Eds.; Oxford University Press: New York, 1998; 125–147.
7. Shaw, R.J.; Coughlan, K.J.; Bell, L.C. Root zone sodicity. In *Sodic Soils: Distribution, Properties, Management and Environmental Consequences*; Sumner, M.E., Naidu, R., Eds.; Oxford University Press: New York, 1998; 95–106.

26

Soil Degradation: Global Assessment

Ahmet Çilek,
Suha Berberoğlu,
Erhan Akça, Cenk Dönmez, Mehmet Akif Erdoğan,
Burçak Kapur, and
Selim Kapur

Introduction

One of the major challenges facing humanity in the coming decades is the need to increase food production in limited soil sources to cope with the ever-increasing population, which is indeed the greatest threat for food/soil security.[1] Although estimates on the prevalence on land degradation vary from 1 to 6 billion ha,[2] Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES)[3] reported that the worldwide degraded land is reached to 75% which now threatens 3.2 billion of people well-being in spite of the strict laws enacted for environmental protection. This is an unpalatable but important indicator of high external input dependent on global food, fiber and biofuel production, and a corresponding increase in soil/land damage, reflecting the expensive agricultural production.[4] Global efforts are in an increasing trend for achieving sustainable soil use in the world and the Global Soil Partnership established in 2012 as a mechanism to develop a strong interactive partnership and enhanced collaboration and synergy of efforts between all stakeholders for the protection of the soils against main soil degradation threats.[5] It is based upon the concept of soil multifunctionality and recognizes the various soil functions relevant for global economy and the environment. These functions are under threats of losses in soil organic carbon and soil biodiversity, soil sealing, nutrient imbalance, salinity-alkalinization, and pollution that are severely limiting soil functionality and therefore socioeconomic growth. Soils in the world are recognized as being the result of millennia of human interaction with the landscape and have never been as fragile as today.[6]

Background

Soil can even degrade without actually eroding. It can lose its nutrients and soil biota, and can become damaged by waterlogging and compaction. Erosion is only the most visible part of degradation, where the forces of gravity, water flow, or wind actively remove soil particles. Rather than taking the classical view that soil degradation was, is, and will remain an ongoing process mainly found in countries of the developing world, this phenomenon should be seen as a worldwide process that occurs at different scales and different time frames in different regions.[7] The causes of biophysical and chemical soil degradations are enhanced by socioeconomic interventions, which are the main anthropogenic components of this problem, together with agricultural mismanagement, overgrazing, deforestation, overexploitation, loss of soil organic carbon soil sealing, and pollution as the main reasons for erosion and chemical soil degradation.[5]

Soil degradation, the threat to "soil security," is ubiquitous across the globe in its various forms and at varying magnitudes, depending on the specific demands of people and the inexorably increasing pressures on land. Europe provides many telling examples of the fragile nature of soil security and the destructive consequences of a wide range of soil degradation processes. Asia, Africa, and South and North America are not only partly affected by the non-resilient impacts of soil degradation but also experiencing subtler destruction of soils via political developments, which seek to provide temporary relief and welfare in response to the demands of local populations. Several diverse data sets were used in this research (Table 1). The modeling of land degradation adopted here is based on soil types, water and wind erosion, land cover, terrestrial net primary productivity (NPP), percent tree cover, aridity, and chemical deterioration.

The global *soil map* derived by the United States Department of Agriculture/Natural Resources Conservation Service (USDA-NRCS, Soil Survey Division) was used as one of the inputs in this study. This map was produced in September 2005 and was based on a reclassification of the Food and Agriculture Organization of the United Nations Educational, Scientific and Cultural Organization (FAO-UNESCO) Soil Map of the World combined with a soil climate map. The soil map shows the global distribution of major soil types. The scale of the global soil map is 1:5,000,000.

The global *water erosion vulnerability map* was based on a reclassification of the global soil climate map and global soil map of the USDA-NRCS. The global soil climate map comprises the major biome types at the global scale, which was derived using global land cover maps of the European Space Agency. The global water erosion vulnerability map consisted of four vulnerability classes of the water. These classes were also based on soil climate and soil classification of the USDA-NRCS, Soil Survey Division. This map was produced in 1998 with a scale of 1:5,000,000.

The global *wind erosion vulnerability map* was derived using soil and soil climate maps of the USDA-NRCS, Soil Survey Division. The map was 1:5,000,000 scale and comprised four vulnerability classes.

The global *land cover map* was derived by ESA Climate Change Initiative—Land Cover led by UCLouvain. The land cover map was derived using AVHRR HRPT (1992–1999), SPOT-Vegetation (1999–2012), and PROBA-V (2013–2015) at 1 km spatial resolution, and comprises 22 land cover classes

TABLE 1 The Data Sets Used in the Modeling of Land Degradation

Data Sets	Source of Data
Global soil map	United States Department of Agriculture (USDA)
Global water erosion vulnerability	USDA
Global wind erosion vulnerability	USDA
Global saline/alkaline domains	USDA
Global land cover map	ESA Climate Change Initiative—Land Cover led by UCLouvain (2017). The Land Cover CCI Climate Research Data Package (CRDP)
Net primary production	NASA Earth Observations
Percent tree cover	Geospatial Information Authority of Japan, Chiba University, and collaborating organizations
Aridity map	CGIAR-CSI (International Research Center and CGIAR – Consortium for Spatial Information)
Chemical deterioration map	LADA

defined according to the UN Land Cover Classification System. This set of cover types included 11 categories of natural vegetation covers.[8]

The global NPP map indicated the net amount of carbon from the atmosphere into green plants. NPP is an important ecological variable that measures the energy input to the biosphere and terrestrial carbon dioxide assimilation and shows ecosystem performance. The Global NPP map was gathered from The Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC) for biogeochemical dynamics, which is one of the National Aeronautics and Space Administration-Earth Observing System Data and Information System (NASA-EOSDIS) data centers managed by the Earth Science Data and Information System (ESDIS) Project. This map was derived by using the NASA-CASA (Carnegie Ames Stanford Approach) NPP Model Approach.

The *percent tree cover map* was gathered from the Global Land Cover Facility/University of Maryland. It illustrates the global cover of woody vegetation, on a continuous scale from 0% to 100%. Advanced very high-resolution radiometer (AVHRR) images were used to derive the percent tree cover map at 1 km spatial resolution. The percent tree cover was estimated using the linear regression method. This map comprises three classes, namely, the (1) 10%–80% tree cover, (2) non-vegetated land, and (3) tree cover less than 10%.[9]

The global *aridity index map* was gathered from the -Consultative Group for International Agricultural Research – Consortium for Spatial Information (CGIAR-CSI). Aridity is expressed as a function of precipitation, potential evapo-transpiration (PET), and temperature, and is classified according to the climatic zones proposed by the UNEP.[10] The Aridity Index is used to quantify the precipitation deficit over atmospheric water demand. The Aridity Index map was derived using MODIS images with 1 km spatial resolution. The climate data sets were obtained from the WorldClim data set,[11] and the map was estimated with the following equation:

$$\text{Aridity Index (AI)} = \text{MAP} / \text{MAE},$$

where MAP is the mean annual precipitation and MAE is the mean annual evapotranspiration.

The *chemical deterioration map* was obtained from the Land Degradation Assessment in Drylands (LADA) organization. The chemical deterioration map used in this study comprised four main classes: (1) fertility decline and reduced organic matter content (not by erosion)—leaching, nutrient mining, oxidation, and volatization (N); (2) acidification—decreased soil pH; (3) soil pollution—soil contamination with heavy metals and toxic components; and (4) salinization/alkalinization—a net increase in salt and sodium contents of the (top) soil leading to a productivity decline.

Standardization and weighting of these inputs were the two critical steps while modeling the land degradation within a Geographical Information System (GIS) environment. The inputs were scaled between 0 and 1 in standardization using sigmoidal functions. The fuzzy approach was used in standardization to evaluate the possibility that each pixel belongs to a fuzzy set by evaluating any of a series of fuzzy set membership functions. The sigmoidal, J-shaped, and linear functions were controlled by four points ordered from low to high on the measurement scale. Factor-by-factor weighting was done (i.e., weak NPP will have the higher score and risk than dense NPP areas) following standardization. All factors were evaluated through overlapping each other.

Europe

The major problems concerning the soils of Europe are the quality and quantity loss of soils due to erosion, sealing/urbanization, flooding, loss of fertility due to deep ploughing, removing crop residues, mono-culture, decline in organic carbon and biodiversity along with local and diffuse soil contamination especially in industrial and urban areas, and soil acidification (Figure 1).[12,13] Salinity, although not widely distributed in Europe, is seen in some parts of Western and Eastern Europe,[14] whereas urbanization and construction of infrastructure at the expense of fertile land are widespread in Europe,

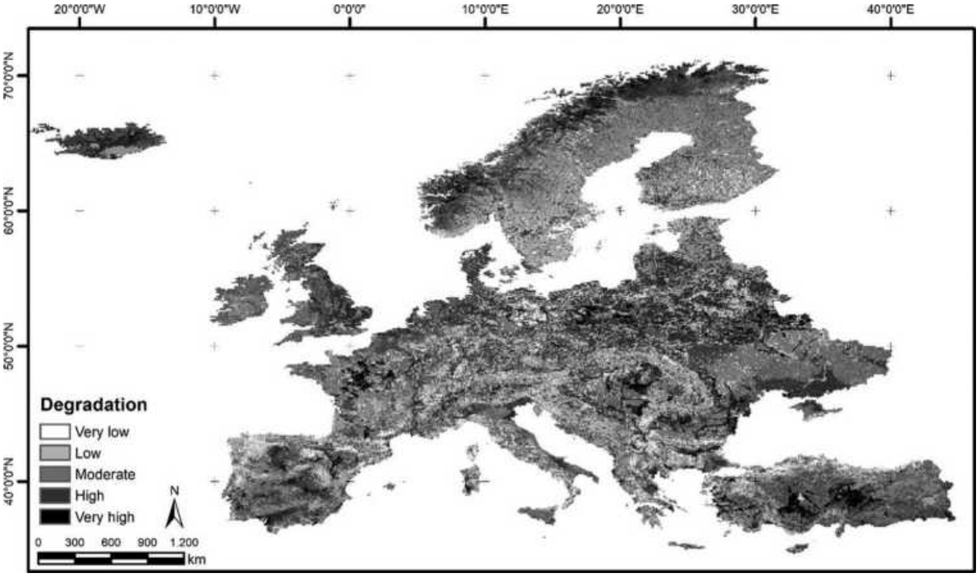


FIGURE 1 Soil degradation map of Europe.

particularly in the Benelux countries, France, Germany, and Switzerland, and such effects are most conspicuously destructive along the misused coasts of Spain, France, Italy, Greece, Turkey, Croatia, and Albania. The EU average soil sealing from 2000 to 2006 was 3%, but this figure raises to approximately 14% in Ireland, Cyprus, and Spain.[15] The drastic increase in the rate of urbanization since the 1980s is now expected to follow the Plan Bleu,[16] which seeks to create beneficial relationships between populations, natural resources, the main elements of the environment, and the major sectors of development in the Mediterranean Basin and to work for sustainable development in the Mediterranean region. The very appropriate term "industrial desertification" remains valid for the once-degraded soils of Eastern Europe, under the pressure of mining and heavy industry, as in the Ukraine where such land occupies 3% of the total land area of the country.[17]

There are three broad zones of "natural" erosion across Europe, including Iceland: (1) the southern zone (the Mediterranean countries); (2) a northern loess zone comprising the Baltic States and part of Russia; and (3) the eastern zone of Slovenia, Croatia, Bosnia-Herzegovina, Romania, Bulgaria, Poland, Hungary, Slovakia, the Czech Republic, and Ukraine (Figure 1). Seasonal rainfalls are responsible for severe erosion due to overgrazing and the shift from traditional crops. Erosion in Southern Europe is an ancient problem and still continues in many places, with marked on-site impacts and with significant decreases in soil productivity as a result of soil thinning. The northern zone of high-quality loess soils displays moderate effects of erosion with less intense precipitation on saturated soils. Local wind erosion on light textured soils is also responsible for the transportation of agricultural chemicals used in the intensive farming systems of the northern zone to adjacent water bodies, along with eroded sediments. The high erodibility of the soils of the eastern zone is exacerbated by the presence of large state-controlled farms that have introduced intensive agriculture at the expense of a decrease in the natural vegetation. Contaminated sediments are also present in this zone, particularly in the vicinity of former industrial operations/deserts, with high rates of erosion in Ukraine (41% of the total land area) and Russia (57% of the total land area) whose agricultural land has been subjected to strong water and wind erosion ever since the beginning of industrialization.[17]

Localized zones of likely soil contamination through the activities of heavy industry are common in northwestern and central Europe as well as Northern Italy, together with more scattered areas of known and likely soil contamination due to the intensive use of agricultural chemicals.[21] Sources of

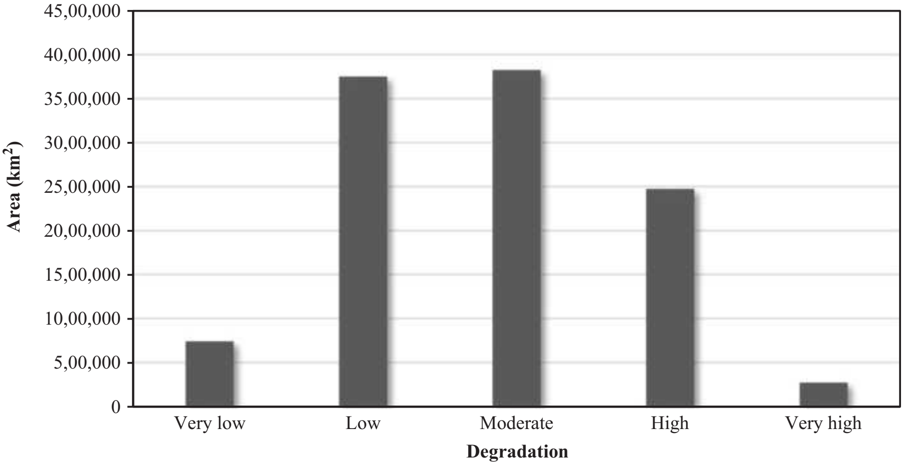


FIGURE 2 Distribution of soil degradation rates in Europe.

contamination are especially abundant in the "hot spots" associated with urban areas and industrial enclaves in the northwestern, southern, and central parts of the continent (Figure 1). Acidification through deposition of windborne industrial effluents and aerosols has been a long-standing problem for the whole of Europe; however, this is not expected to increase much further, especially in Western Europe, as a result of the successful implementation of the emission-control policies over 40 years.[18]

The desertification of parts of Europe has been evident for some decades, and the parameters of the problem are now becoming clear, with current emphasis on monitoring of the environmentally sensitive areas[19] on selected sites, seeking quality indicators for (1) soil, (2) vegetation, (3) climate, and (4) human management throughout the Mediterranean Basin. Apart from the human factor, these indicators are inherent. Nevertheless, the majority of the lands/soils in Europe are under the threat of low and moderate degradation levels as compared with other continents that have suffered extensive and very high degradation, especially through erosion, salinity, and loss of nutrients (Figure 2).

Asia

The most severe aspect of soil degradation on Asian lands has been desertification due to the historical, climatic, and topographic character of this region as well as the political and population pressures created by the conflicts of the past 500 years or so. At least US $10 billion is lost annually as a result of losses resulting from land degradation in India, Pakistan, Bangladesh, Iran, Afghanistan, Nepal, Sri Lanka, and Bhutan. FAO, UNDP, and UNEP estimated that this figure is equal to 2% of the region's Gross Domestic Product, which is about 7% of the value of its agricultural output.[20] Salinization caused by the rapid drop in the level of the Aral Sea and the waterlogging of rangelands in Central Asia due to the destruction of the vegetation cover by overgrazing and cultivation provide the most striking examples of an extreme version of degradation—desertification caused by the misuse of the land. Soil salinity has had severe effects at the northern and western parts of the Caspian Sea, north of India and Pakistan, Laos, Myanmar, and Thailand[21] mainly due to the shift to irrigated agriculture and destruction of the natural vegetation (Figure 3). A major threat to the Asian environment has occurred through the accumulation of soluble salts, mainly deposited from saline irrigation water or through mismanagement of available water resources, as in the drying Aral Sea and the Turan lowlands as well as the deterioration of the oases in Turkmenistan, with excessive abstraction of water in Central Asia (Figure 3).[21] Soil salinity costs Uzbekistan about US $1 billion annually.[21] Moreover, as soil salinity decreases plant productivity, it has negative effect on soil organic carbon inputs which lead to low soil organic carbon contents.

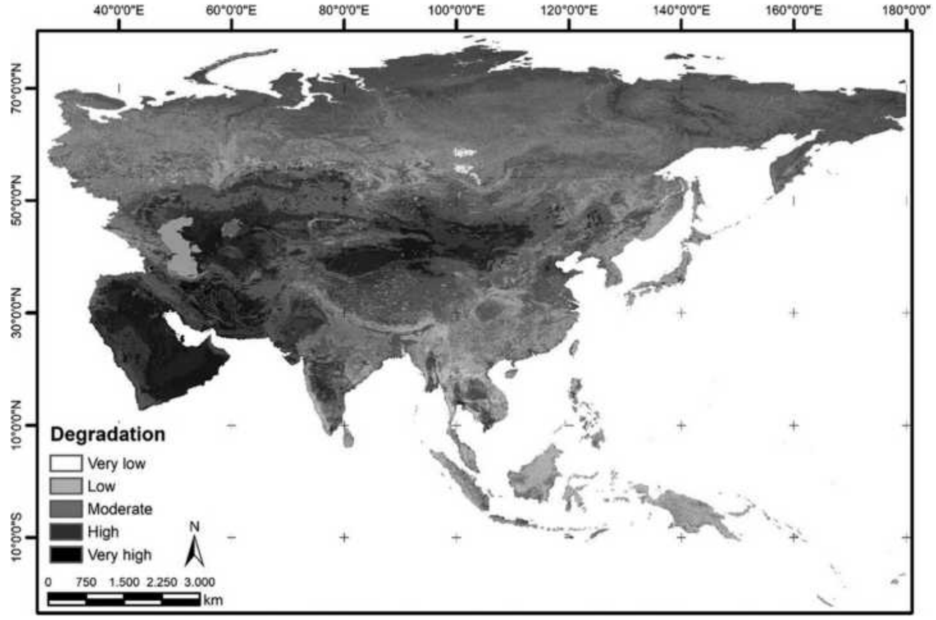


FIGURE 3 Soil degradation map of Asia.

The drylands of the Middle East have been degrading since the Sumerian epoch, with excessive irrigation causing severe salinity and erosion/siltation problems,[22] especially in Iraq, Syria, and Saudi Arabia. Iraq has been unique in the magnitude of the historically recorded buildup of salinity levels, with 8.5 million ha saline land, which is 64% of the total arable land surface (i.e., 90% of the land in the southern part of that country). The historical lands of Iran, Pakistan, Afghanistan, India, and China are also subject to ancient and ongoing soil/land degradation processes, which are subtle in some areas but evident and drastic in others (Figure 3). All ongoing contemporary loss of soil and land in the Asian continent is at an increasing trend even to the presence of the extensive untouched natural forest areas of the north of the continent (Figure 4).

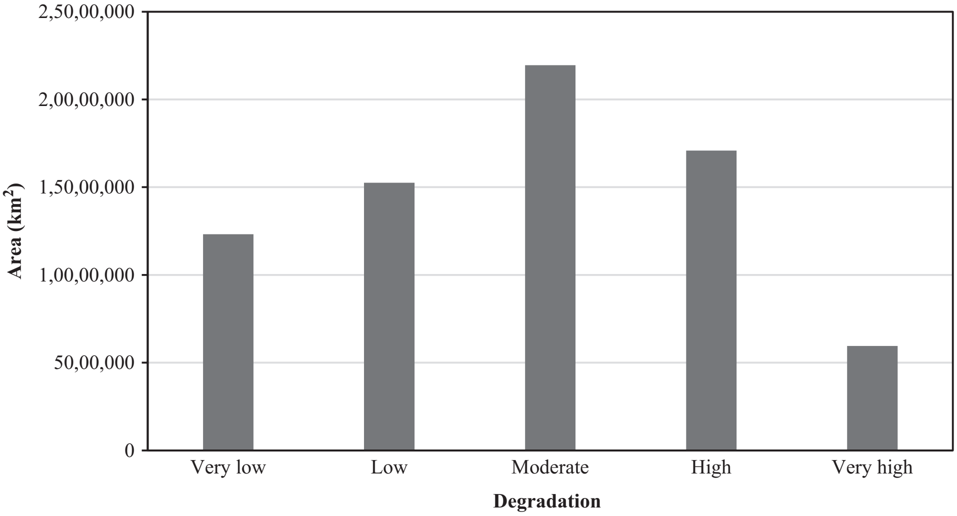


FIGURE 4 Distribution of soil degradation rates in Asia.

Africa

Africa's primary past and present concern has been the loss of soil productivity by nutrient depletion, i.e., the decreasing NPK levels (in kilograms per hectare) in cultivated soils following the exponential growth in population and the resulting starvation and migrations at high levels in the northern, central, and southern parts of the continent (Figure 5). Intensification of land use to meet the increased food demands combined with the mismanagement of the land leads to the degradation of the continental soils. Central African forests are converted to low productive agricultural lands and are used for charcoal and fuelwood production. Moreover, urbanization and uncontrolled mining are other major threats to African forests. This poses the ultimate question of how the appropriate sustainable technologies that will permit the increased productivity of the soils can be identified. This problem is illustrated by the example of the Sudan, where nutrient depletion has steadily increased through

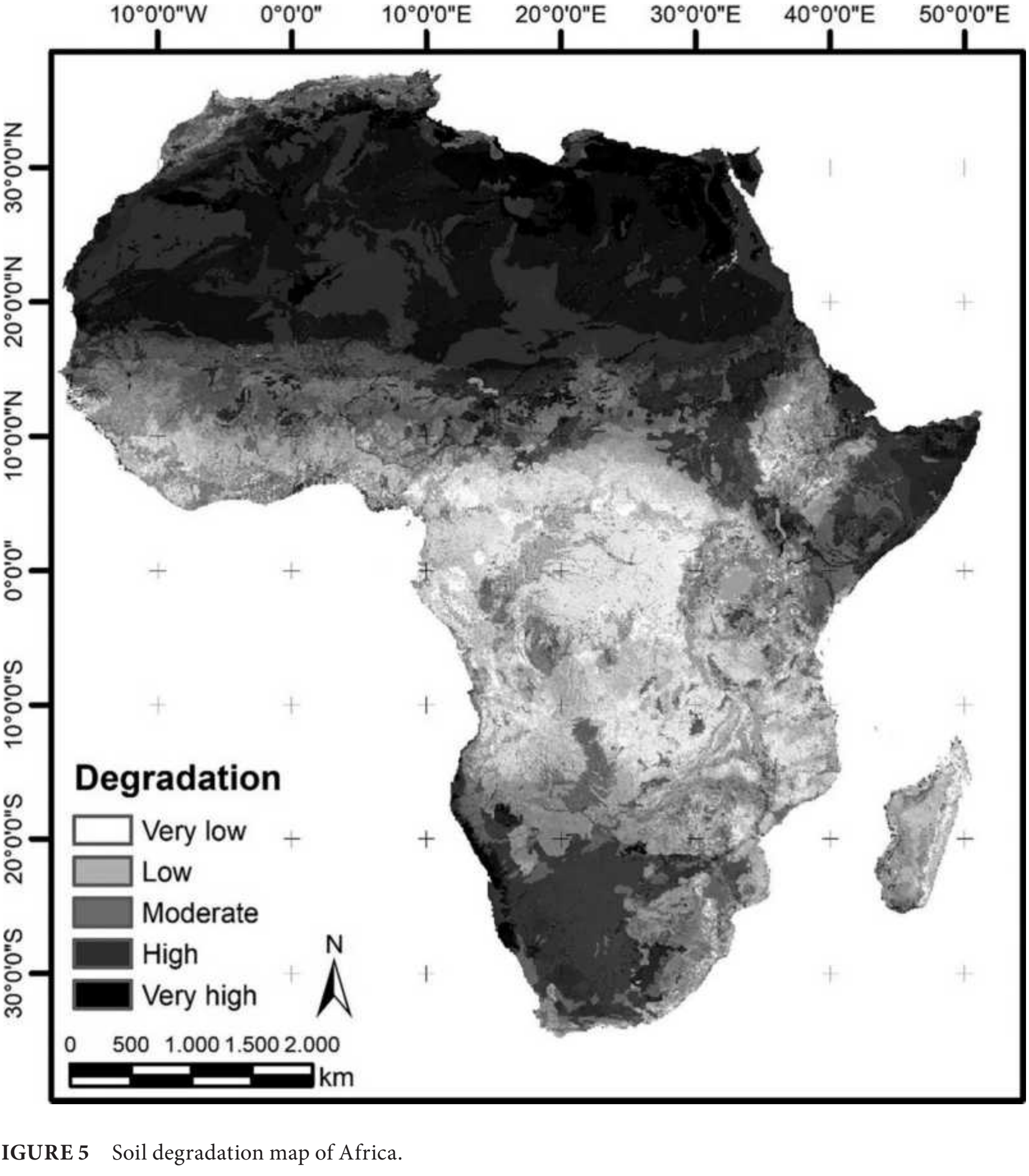


FIGURE 5 Soil degradation map of Africa.

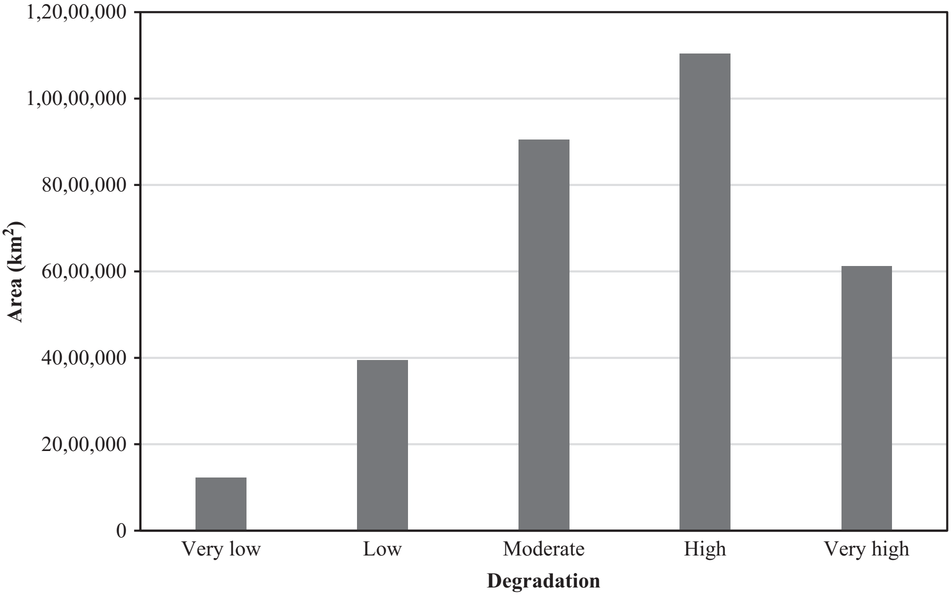


FIGURE 6 Distribution of soil degradation rates in Africa.

more mechanized land preparation, planting and threshing without the use of inorganic fertilizers, and legume rotations. Thus, aggregate yields have been falling as it became more difficult to expand the cultivated area without substantial public investments in infrastructure. In Burkina Faso, and in the other lands of Africa located around the southern and northern fringes of the Saharan desert and the desert margins, the decreased infiltration and increased runoff causing erosion is a further consequence of repeated cultivation. Thus, the technological measures to be identified for Burkina Faso and similarly to other African countries must be smart agriculture. This includes development of water retention technologies, while polyculture/rotations with proper manuring and fertilization for cost-efficient provisions of N and P by preferably green manuring are major actions to be taken. These actions in the Sudan and likewise other parts of Africa will provide balanced management of soil moisture, nutrients, and organic matter by enhancing C-sequestration in soil, which is the main goal for sustainability based on the sustainable management of the land, to ensure the security of both the soil and the global climate.[23,24] Salinity is progressively menacing the lands of the African continent by the increased salt accumulation in the soils of the Sahel (especially in Libya) and parts of the Sahara, which is irrigated and/or excess irrigated by fossil water.

The land area of the African continent that is highly and very highly affected by land/soil degradation is estimated to be above 17 million km^2, 1.7 times larger than 10 million km^2 Europe continent, which is mainly caused by depletion of the nutrients and increase of soil salinity (Figure 6). The slight erosion determined on the Saharan lands reflects a mere reality, i.e., the ultimate fate of the land depends on strict conservation measures, as well as detailed research on any fossil water considered for use. Only by such means can stable ecosystems be conserved in such water–soil poor regions.

South America

Water and wind erosion on the eastern and western coastal/ inland areas of the South American continent are the dominant soil degradation processes in these regions and have caused the loss of the topsoil at alarming rates due to the prevailing climatic and topographic conditions. Almost as important is the loss of nutrients from the Amazon Basin[13] (Figure 7). Graesse et al.[25] calculated 17% of new cropland and 57% of new pastureland replaced forests in Latin America from 2001 to 2013. These deforestations and overgrazing effects caused the degradation of 576 million ha of potential agricultural land, an area almost equal to the very high, high, and moderately high degraded land/soil areas (Figure 8). Another important factor has been the ever-increasing introduction of inappropriate agricultural practices derived from the so-called imported technology, which has not been properly adapted to indigenous land-use procedures. The traditional methods of permitting the land to recover naturally have been almost totally abandoned and has been replaced by unsuitable technological measures designed to maintain production levels (temporarily) and to overcome the loss of soil resilience, thus increasing chemical inputs.

The rapid industrialization/urbanization of the limited land resources in the Caribbean has been expelling agricultural communities to remote and marginal regions that are at present rich in

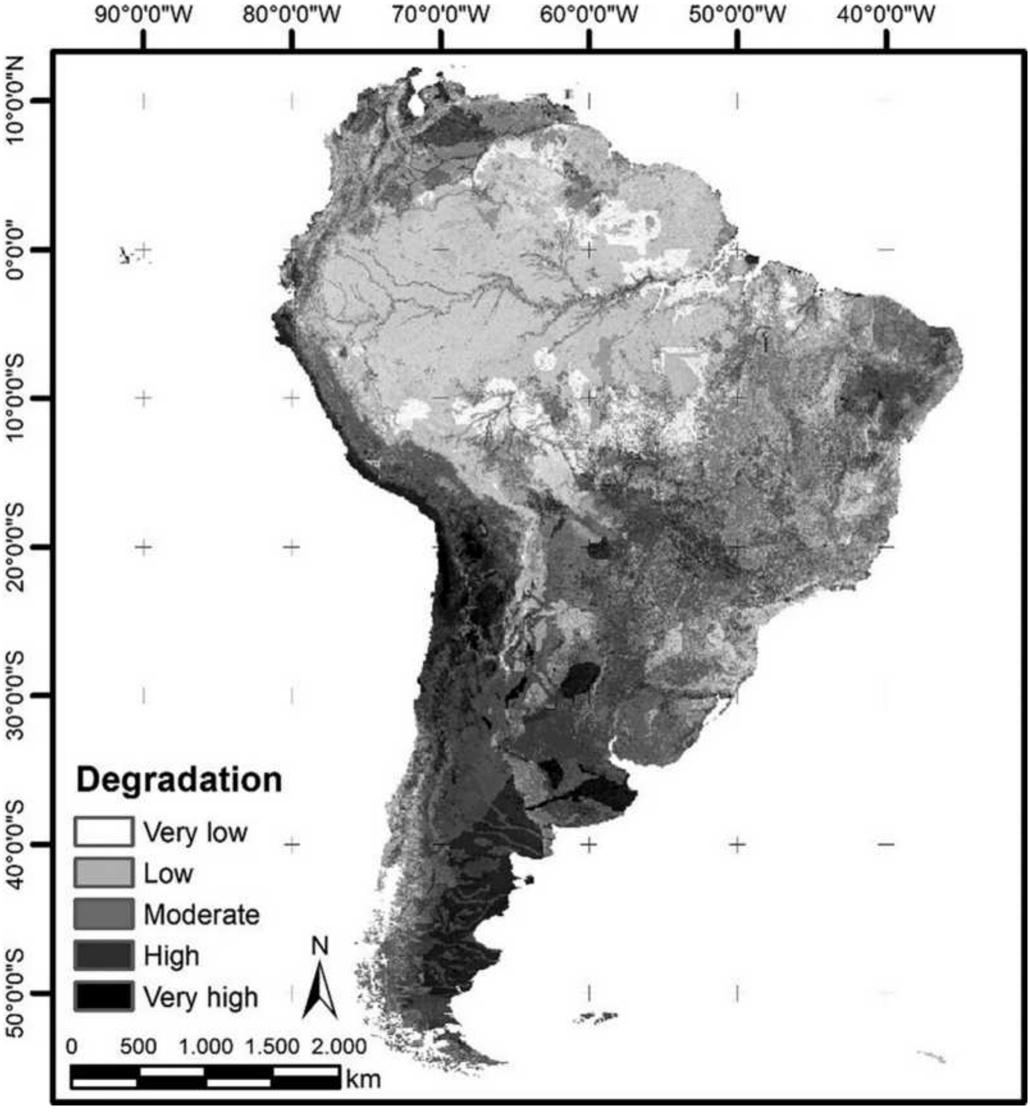


FIGURE 7 Soil degradation map of South America.

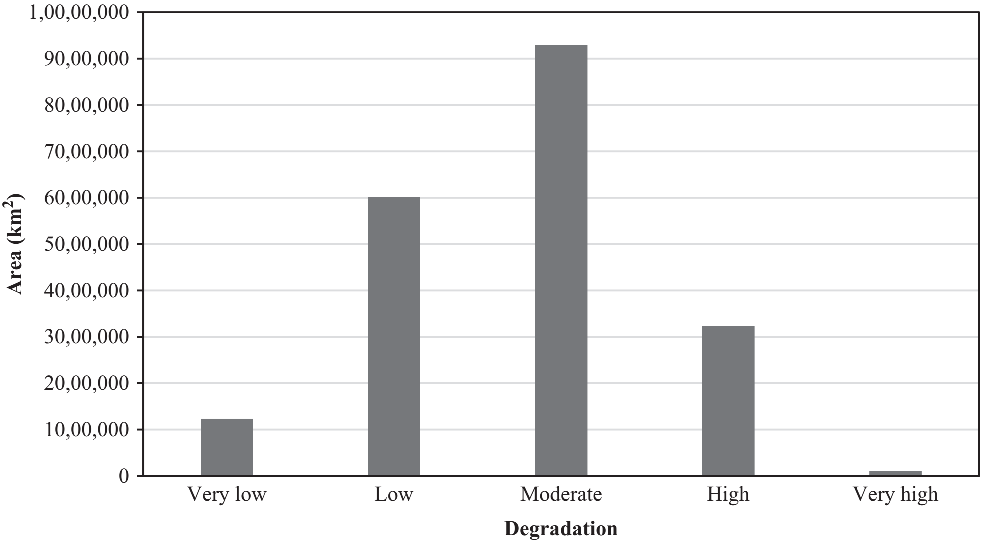


FIGURE 8 Distribution of soil degradation rates in South America.

biodiversity and biomass—a major global C sink. Moreover, large-scale livestock herding in Central and South America is also a major threat to soil security and has been responsible for degrading 35% of the pasture lands of Argentina.

North America, Central America, and the Caribbean Islands

The most prominent outcome of soil/land degradation (or more correctly desertification) in the United States is exemplified by the accelerated dust storm episodes of the 1930s—the Dust Bowl years, marked by the "Black Blizzards," which were caused by persistent strong winds, droughts, and overuse of the soils. These resulted in the destruction of large tracts of farmland in the south and central United States. Recently, salinization has become an equally severe problem in the western part of the country (Figure 9) through the artificial elevation of water tables by extensive irrigation, with associated acute drainage problems. An area of about 10 million ha in the west of the United States has been suffering from salinity-related reductions in yields, coupled with very high costs in both the Colorado River basin and the San Joaquin Valley.[26] Unfortunately, new irrigation technologies such as the Center-Pivot irrigation system (developed as an alternative to the conventional irrigation systems that caused the salinity problems) has caused a decline of the water table levels in areas north of Lubbock, Texas, by around 30–50 m, leading to a dramatic decrease (by 50%) in the thickness of the well-known Ogallala aquifer. In some areas, this has been followed by ground subsidence, which is an extreme form of soil structure degradation, i.e., loss of the physical integrity of the soil.

Loss of topsoil, as the result of more than 200 years of intensive farming in the United States, is estimated to vary from 25% to 75% and exceeds the upper limit in some parts of the country.[27,28] The United States provides good examples of the difficulties involved in erosion control, with its large-scale intensive agriculture damaging soil structure and also increasing the erosion of susceptible soils. This problem could be greatly mitigated by strict enforcement of the no-till system. Such no-till areas have increased tremendously in the United States and are forecast to increase in linear fashion.[29] Conservation farming is practiced in only about half of all US agricultural land and on less than half of the country's most erodible cropland. Conservation farmers are encouraged to use only the basic types of organic fertilizers, such as animal and green manure together with compost, mulch farming, improved pasture management, and crop rotation to conserve soil nutrients. In Alaska, scientists have identified a

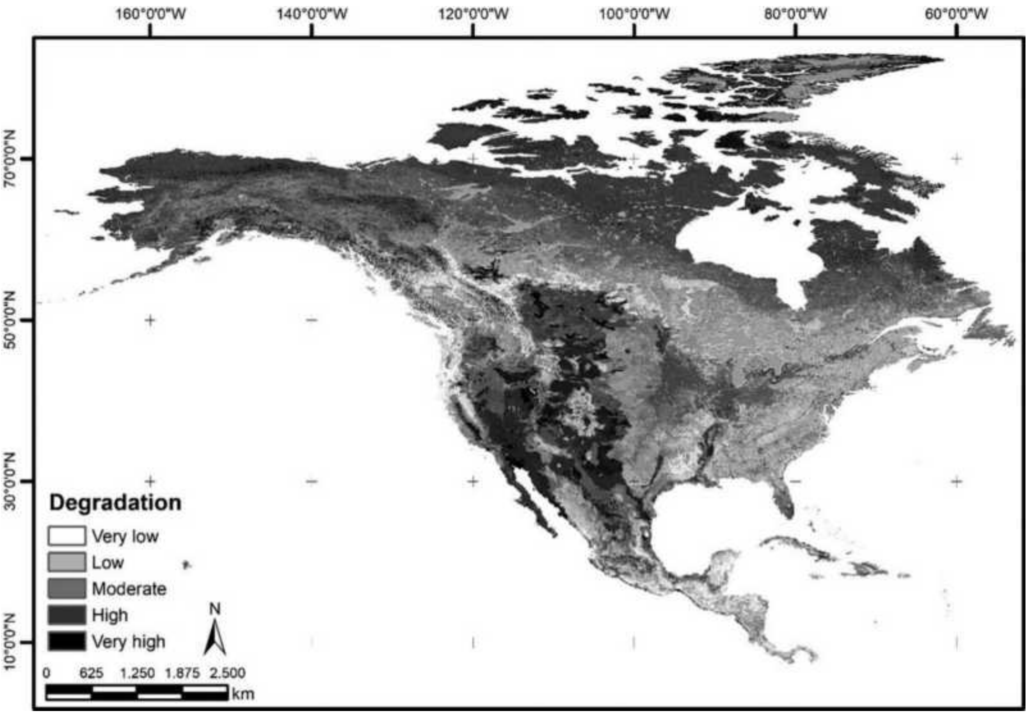


FIGURE 9 Soil degradation map of North and Central America and the Caribbean islands.

number of events that may have been the cause of a severe and rapid soil erosion, where the northern ice sheet is diminishing and consequently causing the increase of water temperatures, which means the increased development of storm events, more precipitations, and higher waves with a potential of eroding some 14 m of coastline each year, which is a much higher rate than the 6.1 m erosion recorded in the mid-1950s.[30] Jones et al. state that this might be a new era in ocean–land interactions and the reshaping of the Arctic coastline by erosion. Furthermore, they claim that the warming sea-surface temperatures and the rising sea level will most probably act to reduce the permafrost-dominated coastline and thaw the ice-rich bluffs and consequently increase erosion and degrade Arctic landscapes together with the long-standing cultural sites.

Canada is a large country where half of the 68 million ha of available land is cultivated, with an average farm size of 450 ha. It is reported that Canada has experienced annual soil losses on the prairies (through wind and water erosion) that are similar to the Asian steppes, amounting, respectively, to 60 and 117 million tons. These annual rates are much higher than the rate of soil formation, resulting in an annual potential grain production loss of 4.6 million tons of wheat. With regard to primary soil salinity, during historic times, the prairies have experienced steady increases, related partially to increasing groundwater levels. Major problems of secondary salinity are estimated to affect 2.2 million ha of land in Alberta, Saskatchewan, and parts of Manitoba, with an immense economic impact each year.

Central America is drastically suffering from soil erosion, where almost 26% of the territory is strongly affected via this irreparable menace caused mainly by overgrazing, which almost adds up to a total of 58% of the overall degradation in the southern tip of North America (Figure 10). Concrete precautions are being taken in the whole region by implementing the “Quesungual” agro-forestry method, concerning the planting of trees between the dispersed natural forest tree communities, converting pastures to forests (World Bank Project), and implementing ambitious projects such as the Lake Managua Watershed drainage to avoid landslides—a major component of degradation in Nicaragua.[31]

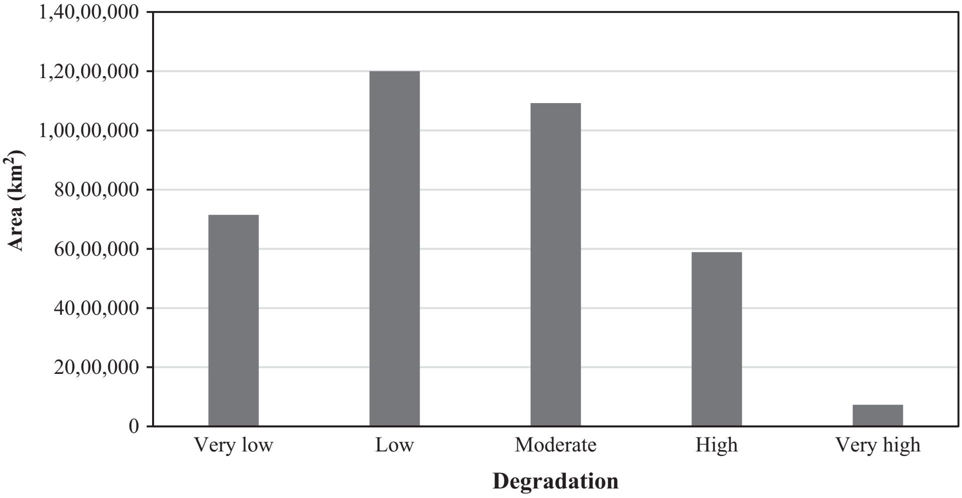


FIGURE 10 Distribution of soil degradation rates in North and Central America.

The Caribbean islands and the lands of Central America have long faced a common fate of soil degradation via intense conversion of forests to cultivated soil, i.e., deforestation, causing flooding, landslides, and development of land surfaces prone to erosion and degrading the cycles of carbon and water—the wonders that sustain our living globe. Costa Rica loses about 860 million tons of valuable topsoil every year.[32] Earthquakes of the Caribbean and Central America were/are followed primarily by volcanic eruptions with immediate and dramatic changes in the environment, where lava buries soils that would be covered for at least a generation-long period before developing into fertile lands rich in nutrients. Drastic examples reveal the inappropriate soil use systems imposed by the colonial powers of Europe back in the 16th century based on the creation of extensive deforestation areas replaced by sugar cane fields, which has been the beginning of the soil degradation era, particularly in Haiti in the Caribbean islands. Precautions to combat the vast ongoing soil erosion in the Caribbean islands, especially in Haiti, seek the development of tree plantations and in particular bamboo crops to overcome the soil-degradation- and political-unrest-oriented contemporary poverty in the country. The high rates of erosion, as sediment depositions in the sea, are also threatening one-third of the coral reefs around the Caribbean islands. High turbidity of inshore water and elevated algal cover on reefs are linked to the impacts of coastal development inducing erosion, with soil/sediment deposition being a major influence on the conditions of reefs.[33,34] The Caribbean Environment Program is active in the region, seeking to mitigate soil and land degradation by developing projects primarily on the "Best Management Practices in Agriculture" and "Integrating Watersheds and Coastal Areas Management—IWCAM" via the Global Environment Facility (GEF) included in National Programs of Action.[35]

Australia

The Australian agricultural/soil resource base has been long endangered by the "business as usual" concept on the continent aiming to achieve temporary economic betterment,[36] *which, in contrast, induced soil degradation.* Identification of the different types of soil degradation in Australia reveals that erosion has been the main component, primarily via dust storms, which still are a serious problem, especially where cropping practices do not include retention of cover and minimum tillage methods. Water erosion effects are also particularly severe in areas of summer rainfall and topographic extremities (Figure 11). Remedial actions for this include the well-known measures of maintaining adequate

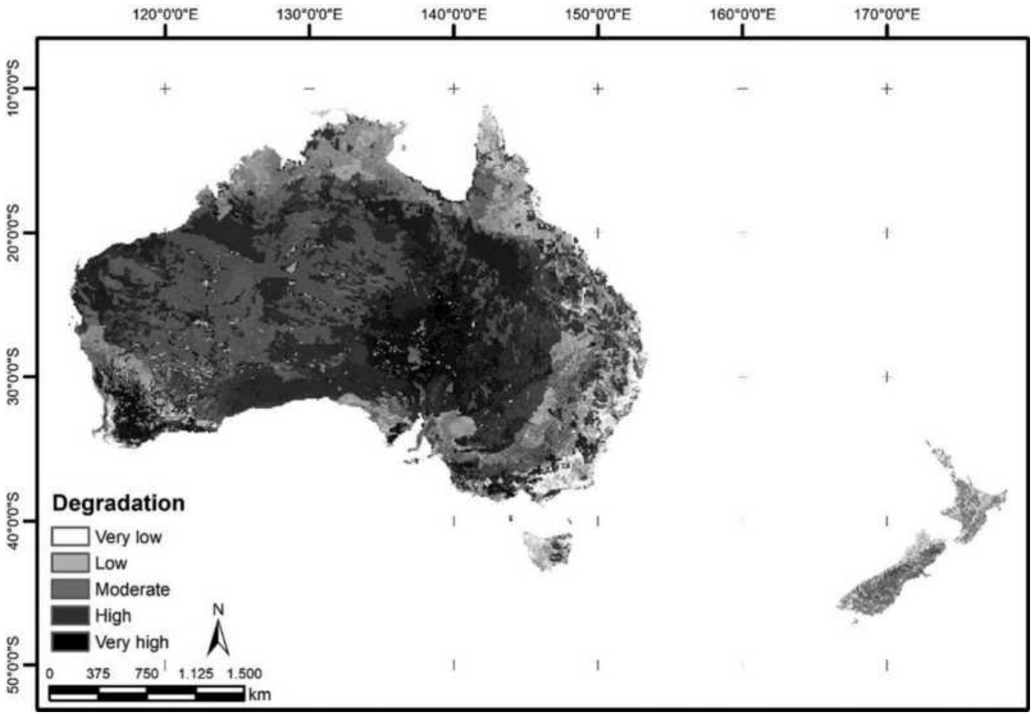


FIGURE 11 Soil degradation map of Australia and New Zealand.

cover and changing prevailing altitudes towards stock management, storage feed, redesign of watering sites, and management of riparian areas.

Part of the excess salinity in Australia is of primary origin and was retained in the subsoil by trees, which have now been cleared to create soil surfaces for cropping and pastures, allowing penetration of water to the saline subsoil, then followed by abstraction from the water table, thus leading to serious problems in the southwest, east, and central-eastern parts of the country (Figure 11). About 30% of Australia's agricultural land is sodic, creating poor physical soil conditions and impeded productivity. This problem can only be alleviated by massive revegetation programs and by taking extra care of the water table and plant cover. Despite the introduction of costly conventional measures for reclamation, salinity levels continue to increase across Australia in the dry and irrigated soils. The dryland salinity in the continent affects about 2.5 million ha of farmland and is expanding at a rate of 3%–5% yr^{-1} which may sum up to 8.8 million ha (33%) by 2050.[37]

The retardation of the organic matter levels also requires remediation measures, with economically justified fertilizer use strategies to be utilized throughout the continent. Moreover, overgrazing has resulted in the impoverishment of plant communities and loss of habitats as well as the decline in the chemical fertility of the soil by progressive depletion of organic matter in the topsoil, followed by deterioration in soil structure.

Acidification caused by legume-based (clover, etc.) mixed farming plus use of ammonia-based fertilizers threatens 55 million ha of Australian land, which is about one-half of the degraded land areas (high, moderate, and very high, respectively) of the continent (Figure 12). Liming seems to be the most effective present remedy, but this is costly, does not lead to rapid recovery, and is impractical for subsoil acidity. Thus, the precise remedies have yet to be developed for the conditions on this continent, utilizing careful, long-term monitoring and the experience of farmers to devise specific treatment and conservation procedures. Despite the common view that reflects the absence of soil loss in

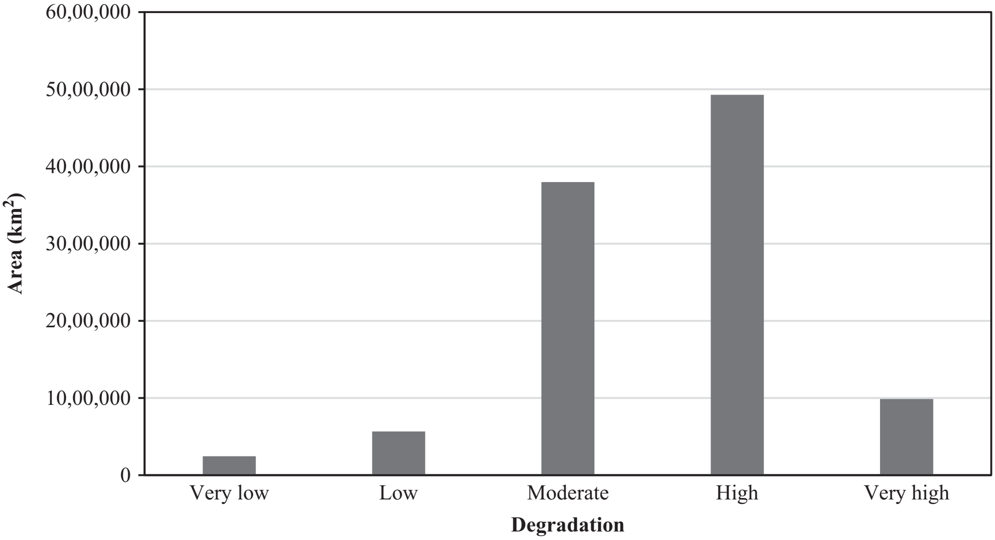


FIGURE 12 Distribution of soil degradation rates in Australia and New Zealand.

New Zealand due to the presence of the widespread forests, soil erosion has been a major threat to the natural resources as being accelerated by early colonial and contemporary deforestation followed by the unwise land-use practices such as overgrazing. Accelerated erosion is the most serious and the least reversible of soil degradation problems encountered in the country. The main forms of erosion in New Zealand are, namely, the mass, fluvial, surface, and stream-bank transportation of the soil in the hilly country, in the Marlborough and Manawatu regions, in the lands outside the hill country, and at the unstable areas cleared of tree cover. The long-standing experience and research provided the relevant bodies dealing with erosion in New Zealand, with the appropriate techniques to reduce the impacts of erosion in the pastoral lands. These techniques include the maintaining of the adequate vegetative cover (e.g., avoiding overgrazing and maintaining a dense pasture sward through regular applications of fertilizer and grass seed), the spaced or close tree planting, retiring the land from pasture, fencing off and planting riverbanks, and building debris dams to slow water flows in gullies. A similar system to the Australian "National Landcare Organisation" is being implemented in New Zealand in 1996 as the "New Zealand Landcare Trust" which is a nongovernmental organization facilitating sustainable land management (SLM) and biodiversity initiatives at a community-driven basis, in order to effective catchment management to improve the land and water quality of the country. The Landcare Trust seeks the assurance of implementing land management more sustainably.[38]

Conclusions

The state of soil/land degradation and its remediation through a multifunction–multi-impact approach has been defined through a driving force–pressure–state–impact–response (DPSIR) matrix developed by the European Environmental Agency[39] (Figure 13) leading to SLM[40–42] measures to be taken for the future. SLM is concerned with more soil/land-friendly farming practices that minimize the erosion potential of soils, together with the adoption by landholders of property management planning procedures that involve community actions, such as the Landcare Program of Australia.[38] Moreover, as Smyth and Dumanski[40] have stated, this approach combines socioeconomic principles with environmental concerns to enable production to be enhanced alongside a reduction in the level of risk to the sustenance of natural resources. Thus, degradation of soil and water quality, which have negative

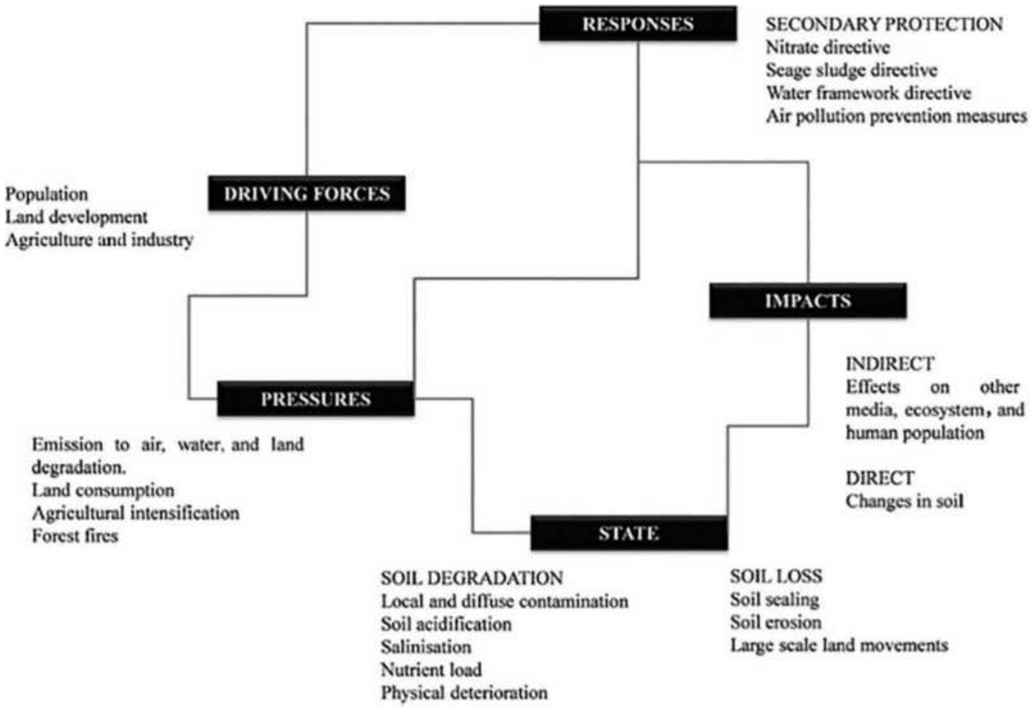


FIGURE 13 DPSIR framework applied to soil.
Source: UNEP/EEA.[39]

effects on climate[43] and biodiversity, can be prevented or minimized in a manner that is acceptable to the farming community. The methods to be adapted for SLM via community actions include contour farming, terracing, vegetative barriers, and other land-use practices amalgamated with indigenous (traditional) technical knowledge (ITC)[44,45] as applied to farming and landscape preservation. The impetus for the use of ITK by scientists and local communities in creating new strategies for sustainable resource management was provided in the United Nations Conference on Environment and Development (UNCED) held in Rio de Janeiro (Brazil) in 2012. Thus, at the 13th meeting of the Parties to the United Nations Convention to Combat Desertification, 113 countries had agreed to specify solid targets with clear indicators, to rehabilitate degraded lands and reverse degradation, which currently affects more than third of the world's prime land resources.[46] This approach was widely accepted and endorsed as the Land Degradation Neutrality approach for a primary solution to the degradation of global soil/land resources in the new UNCCD 2018–2030 Strategic Framework[47] in order to meet UN Sustainable Development Goals aimed to be achieved by 2030 for the global welfare and peace.

References

1. Alexandratos, N. World food and agriculture: the outlook for the medium and longer term. NAS Colloquium. Plants and Population: Is There Time? Beckman Center of the National Academy of Sciences: UC Irvine, 5–6 December 1998.
2. Gibbs, H.K.; Salmon, J. M. Mapping the world's degraded lands. *Applied Geography*, **2015**, *57*, 12–21.
3. IPBES. *Thematic Assessment of Land Degradation and Restoration*. Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services: Medellin, **2018**.

4. Dowdall, C.M.; Klotz, R.J. *Pesticides and Global Health: Understanding Agrochemical Dependence and Investing in Sustainable Solutions*. Routledge: New York, **2016**.
5. FAO and ITPS. *Status of the World's Soil Resources (SWSR)—Main Report*. Food and Agriculture Organization of the United Nations and Intergovernmental Technical Panel on Soils: Rome, **2015**.
6. Cherlet, M.; Hutchinson, C.; Reynolds, J.; Hill, J.; Sommer, S.; Von Maltitz, G. (Eds.). *World Atlas of Desertification*. Publications Office of the European Union: Luxembourg, **2018**.
7. FAO and ITPS. *Global Soil Organic Carbon Map (GSOCmap)—Technical Report*. Food and Agriculture Organization of the United Nations and Intergovernmental Technical Panel on Soils: Rome, **2018**.
8. Strahler, A.; Muchoney, D.; Borak, J.; Friedl, M.; Gopal, S.; Lambin, E.; Moody, A. *MODIS land cover product algorithm theoretical basis document (ATBD)*. Boston University Official Web Page, **2019**, available at http://modis.gsfc.nasa.gov/data/atbd/atbd_mod12.pdf (accessed August 2019).
9. DGS. *Land cover and land use change*. University of Maryland, Department of Global Geographical Sciences, **2019**, available at https://geog.umd.edu/research/landingtopic/land-cover-land-use-change (accessed August 2019).
10. Cherlet, M., Hutchinson, C., Reynolds, J., Hill, J., Sommer, S., von Maltitz, G. (Eds.), *World Atlas of Desertification*, Publication Office of the European Union, Luxembourg, **2018**.
11. Hijmans, R.J.; Cameron, S.E.; Parra, J.L.; Jones, P.G.; Jarvis, A. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **2005**, *25*, 1965–1978.
12. Panagos, P.; Imeson, A.; Meusburger, K.; Borrelli, P.; Poesen, J.; Alewell, C. Soil conservation in Europe: wish or reality? *Land Degradation & Development*, **2016**, *27(6)*, 1547–1551.
13. Lal, R. Soil erosion by wind and water: problems and prospects. In: *Soil Erosion Research Methods*. Lal, R., Ed. Routledge: New York, 2017; 1–10.
14. Daliakopoulos, I.N.; Tsanis, I.K.; Koutroulis, A.; Kourgialas, N.N.; Varouchakis, A.E.; Karatzas, G.P.;, Ritsema, C.J. The threat of soil salinity: a European scale review. *Science of the Total Environment*, **2016**, *573*, 727–739.
15. Gardi, C.; Panagos, P.; Van Liedekerke, M.; Bosco, C.; De Brogniez, D. Land take and food security: assessment of land take on the agricultural production in Europe. *Journal of Environmental Planning and Management*, **2015**, *58(5)*, 898–912.
16. UNEP/MAP. *Mediterranean Strategy for Sustainable Development 2016–2025*. Plan Bleu, Regional Activity Centre: Valbonne, **2016**.
17. Stankevich, S.A.; Kharytonov, N.N.; Dudar, T.V.; Kozlova, A.A. Risk assessment of land degradation using satellite imagery and geospatial modelling in Ukraine. In: *Land Degradation and Desertification—A Global Crisis*; Kaswamila, A., Ed. IntechOpen, **2016**, 53–77.
18. Crippa, M.; Janssens-Maenhout, G.; Dentener, F.; Guizzardi, D.; Sindelarova, K.; Muntean, M.; Van Dingenen, R.; Granier, C. Forty years of improvements in European air quality: regional policy-industry interactions with global impacts. *Atmospheric Chemistry and Physics*, **2016**, *16*, 3825–3841.
19. Kosmas, C.; Ferrara, A.; Briassouli, H.; Imeson, A. Methodology for mapping environmentally sensitive areas (ESAs) to desertification. In: *The MEDALUS Project. Mediterranean Desertification and Land Use Report*; Kosmas, C., Kirkby, M., Gecson, N., Eds. European Commission: Belgium, **1999**; 31–47.
20. Khor, M. 2019. *Land degradation causes $10 billion loss to South Asia annually*. https://www.globalpolicy.org/global-taxes/49705-land-degradation-causes-10-billion-loss-to-southasia-%-20annually-.html (accessed August 2019).
21. Ivushkin, K.; Bartholomeus, H.; Bregt, A.K.; Pulatov, A.; Kempen, B.; De Sousa, L. Global mapping of soil salinity change. *Remote Sensing of Environment*, **2019**, *231*, 111260.
22. Lowdermilk, W.C. *Conquest of the Land through Seven Thousands Years*. USDA, Soil Conservation Service, U.S. Government Printing Office: Washington, DC, **1997**; *99*, 1–30.

23. Sanders, J.H.; Southgate, D.D.; Lee, J.G. The Economics of Soil Degradation: Technological Change and Policy Alternatives. *SMSS Technical Monograph*; Department of Agricultural Economics, Purdue University, **1995**; *22*, 8–11.
24. Masso, C.; Baijukya, F.; Ebanyat, P.; Bouaziz, S.; Wendt, J.; Bekunda, M.; Vanlauwe, B. Dilemma of nitrogen management for future food security in sub-Saharan Africa—A review. *Soil Research*, **2017**, *55(6)*, 425–434.
25. Graesser, J.; Aide, T.M.; Grau, H.R.; Ramankutty, N. Cropland/pastureland dynamics and the slowdown of deforestation in Latin America. *Environmental Research Letters*, **2015**, *10*(3), 034017.
26. Barrow, C.J. *Land Degradation*. Cambridge University Press: Cambridge, **1994**.
27. USDA-NRCS. *Risk of Human Induced Desertification*. NRSC, Soil Survey Division, World Soil Resources: Washington, DC, **1999**.
28. USDA-NRCS. *Global Desertification Vulnerability*. NRSC, Soil Survey Division, World Soil Resources: Washington, DC, **1999**.
29. Kern, J.S.; Johnson, M.G. *Impact of Conservation Tillage Use on Soil and Atmospheric Carbon in the Contiguous United States. EPA/600/3-91/056.* Environmental Research Laboratory: Corvallis, OR and Orlando, FL, **1991**.
30. Jones, B.M.; Arp, C.D.; Jorgenson, M.T.; Hinkel, K.M.; Schmutz, J.A.; Flin, P.L. Increase in the rate and uniformity of coastline erosion in Arctic Alaska. *Geophysical Research Letters*, **2009**, *36*(3), L03503. Doi: 10.1029/2008GL036205.
31. Faber, D. A revolution in environmental justice and sustainable development: the political ecology of Nicaragua. In: *Environmental Justice*; Thompson, P., Ed. Routledge, **2017**, 39-70.
32. Butler, R. *Soil erosion and its effects*. https://rainforests.mongabay.com/0903.htm, **2012** (accessed August 2019).
33. UNEP/CEP. *Guidelines for Sediment Control Practices in the Insular Caribbean*, CEP Technical Report 32; UNEP Caribbean Environment Program: Kingston, **1994**.
34. DPNR/DEP and USDA/NRCS. *Unified Watershed Assessment Report—United States Virgin Islands, Virgin Islands Department of Planning and Natural Resources in cooperation with USDA Natural Resources Conservation Service, Caribbean Area, St. Croix, USVI*, **1998**.
35. UNEP/GEF/Kalmar Högskola/Cimab. *Global International Water Assessment (GIWA), Caribbean Islands Bahamas, Cuba, Dominican Republic, Haiti, Jamaica, Puerto Rico Regional Assessment 4, Kalmar Sweden*. **2004**.
36. Russell-Smith, J.; Sangha, K.K.; Costanza, R.; Kubiszewski, I.; Edwards, A. Towards a sustainable, diversified land sector economy for North Australia. In: *Sustainable Land Sector Development in Northern Australia*; Russell-Smith, J. James, G., Pedersen, H., Sangha, K.K. Eds. Taylor and Francis: Boca Raton, FL. **2018,** 85-132.
37. Lambers, H. Dryland salinity: a key environmental issue in southern Australia. *Plant and Soil*, **2003**, *257*, v–vii.
38. Landcare Australia. *Annual Report and Yearbook*. Landcare Australia: Melbourne, **1998**; 70.
39. UNEP/EEA. Soil degradation and sustainable development in Europe: a challenge for the 21st century. *4th Conference to the Parties*, Bonn, December **2000**.
40. Smyth, A.J.; Dumanski, J. FESLM: An International Framework for Evaluating Sustainable Land Management, *World Soil Resources Report. No. 73.* FAO: Rome, **1993**.
41. Kapur, S.; Akça, E.; Zucca, C.; Berberoğlu, S.; Miavaghi, S.R. Anthroscapes: a robust basis for mapping land quality and sustainable land use patterns. In: *Eastern Mediterranean Port Cities*; Yenişehirlioğlu, F., Özveren, E., Selvi Ünlü, T., Eds. Springer: Cham, **2019**, 63–77.
42. Kapur, S.; Eswaran, H.; Blum, W.E. (Eds.). *Sustainable Land Management: Learning from the Past for the Future*. Springer Science & Business Media: Heidelberg. **2010**.
43. Watanabe, T.; Kapur, S.; Aydın, M.; Kanber, R.; Akça, E. (Eds.). *Climate Change Impacts on Basin Agro-Ecosystems*. Springer International Publishing: Cham, **2019**.

44. Eswaran, H.; Beinroth, F.; Reich, P. Biophysical considerations in developing resource management domains. *Proceedings of the IBSRAM International Workshop on Resource Management Domains*, Kuala Lumpur, August 26–29, 1996; Leslie, R.N., Ed. The International Board for Soil Research and Management (IBSRAM): Bangkok, **1996**, 61–78.
45. Kapur, S.; Kapur, B.; Akça, E.; Eswaran, H.; Aydin, M. A. Research strategy to secure energy, water, and food via developing sustainable land and water management in turkey. *Hexagon Series on Human and Environmental Security and Peace*, **2009**, 4 (Part V), 509–518.
46. UNCCD. *Countries agree on a landmark 2030 strategy to save fertile lands* https://www.unccd.int/news-events/countries-agree-landmark-2030-strategy-save-fertile-lands. **2017** (accessed August 2019).
47. UNCCD. 2030 agenda for sustainable development: implications for the United Nations convention to combat desertification: the future strategic framework of the convention. *Conference of the Parties. Thirteenth Session*. Ordos, China, 6–16 September **2017**.

27

Soil Erosion and Carbon Dioxide

Pierre A. Jacinthe
and Rattan Lal

Introduction

The projected warming of global climate and the degradation of land resources by erosion are among the most pressing environmental challenges of the modern era. The recent warming trend in the world's climate is linked to the accumulation of radiatively active gases (CO_2, N_2O, CH_4) in the atmosphere.[1,2,3] Atmospheric CO_2 concentration has been increasing at a rate of 2.2 ppm C y^{-1},[4] which, if maintained, could profoundly alter the chemical composition and the energy balance of the Earth's atmosphere. Although fossil fuel burning is the primary contributor to the current atmospheric CO_2 buildup, it is estimated that agriculture and land-use change contributed 24% of total anthropogenic emissions of CO_2.[5] In soils, evolution of CO_2 occurs through soil respiration which includes root respiration and the microbial decomposition of crop residues and soil humus.

The soil organic carbon (SOC) content is positively related to structural stability and negatively related to the susceptibility of soils to erosion.[6,7] Cultivation reduces the stability of aggregates,[6] thereby increasing the erodibility of soils.[8] The effect of SOC on susceptibility to erosion is best exemplified by the use of SOC content in determining the soil erodibility factor (K) in the universal soil loss equation (USLE).[7] Worldwide, an estimated 1.11×10^9 ha of land are affected by water erosion,[9] resulting in the annual transport of $20–25\times10^9$ Mg of sediment to rivers and oceans.[10,11] While much is known about the protective role of SOC against soil erosion, information is limited regarding the fate of SOC mobilized by erosion. It is estimated that as much as 1.1 PgC [1 Pg$=10^{15}$ g$=$billion metric ton] may be emitted into atmosphere by erosion by water.[12] In this entry, we deliberate that water erosion, as a mechanism for SOC removal, has a profound effect on the global CO_2 budget, and we present an approach to assess erosion-induced CO_2 emission into the atmosphere.

Eroded Soil Organic Carbon: State of the Knowledge

Water erosion involves the detachment of soil aggregates, the release of aggregate-protected SOC, and the translocation and deposition of detached soil particles. Globally, estimates of SOC displaced in terrestrial ecosystems by water erosion are in the order 57 Pg of Cy^{-1}.[13,14] There is considerable uncertainty regarding the fate of the eroded SOC, however. The Intergovernmental Panel on Climate Change (IPCC) echoed this uncertainty in its handbook for conducting national greenhouse gas inventory by stating that the net effect of erosion on CO_2 evolution is unclear at the present time.[15]

The fate of SOC mobilized by erosion can include redistribution, long-term storage in aquatic and terrestrial systems, and mineralization during transport and deposition leading to emission of CO_2 (production of CH_4 possible depending on oxidation status).[14] During transport, detached soil particles are hydrodynamically sorted and redistributed over the eroding landscape as determined by flow condition, surface roughness, particle size, and geomorphology. Between 80% and 90% of the total soil mass mobilized by water erosion remains near the site of erosion.[15] A fraction of the eroded SOC is exported to streams, rivers, and lakes. Several publications have documented the magnitude of these exports which are in the order of 0.6–0.8 Pg C y^{-1} [16,17] or more.[12,14]

Eroded SOC in runoff is also retained in the low-lying portions of the eroding landscape, but the fate of SOC entrapped in these terrestrial deposits is also not fully understood. On the premise that the decomposition process may be O_2-constrained, it has been suggested that entrapment of eroded SOC in terrestrial deposits can be a mechanism for C sequestration.[18,19] It is estimated that a sequestration of 0.6–1.5 Pg C y^{-1} may be possible via this mechanism.[19]

Eroded soil materials typically contain twice as much (or more) C than the average topsoil.[18] In addition, these materials are enriched in labile C.[20] Given the propensity of these SOC fractions to be associated with the lighter and most erosion-prone soil particles, mineralization of eroded C during transport and deposition is a reasonable expectation. A study comparing CPMAS^{13}C-nuclear magnetic resonance spectrometry data of eroded and deposited soil samples indicated that 70% of SOC in eroded soil could be decomposed during transport and deposition.[21] A recent review suggests a lower (20%) mineralization rate.[13]

The information presented above clearly underscores the need for better quantitative assessments of the various fates of eroded SOC. Of particular interest is an assessment of water erosion's contribution to atmospheric CO_2. At the present time, there are few if any experimental data to support such an evaluation. However, an indirect determination of CO_2 production during transport and deposition of eroded SOC is possible using literature data pertaining to the other pathways identified above.

Assessment of Erosion-Induced Carbon Dioxide Fluxes from Existing Data

Conceptual Framework and Description of the Mass Balance Approach

In order to estimate CO_2 production during transport and deposition of eroded SOC, a mass balance approach, described in Figure 1, is proposed. This approach assumes a steady-state equilibrium with respect to SOC content. Augmentation of the initial pool of SOC via humification of crop residues compensates for SOC losses, which include soil respiration, leaching, and SOC translocation with runoff waters. It follows that

$$C_0 + h \cdot C_A = C_t + S_R + E \tag{1}$$

$$S_R = k \cdot C_t \tag{2}$$

$$E = L_C + S_w + E_W + S_L + E_L \tag{3}$$

$$\frac{E_L = (C_0 - C_t)}{t + h \cdot C_A - k \cdot C_t - (L_C + S_w + E_w + S_L)} \tag{4}$$

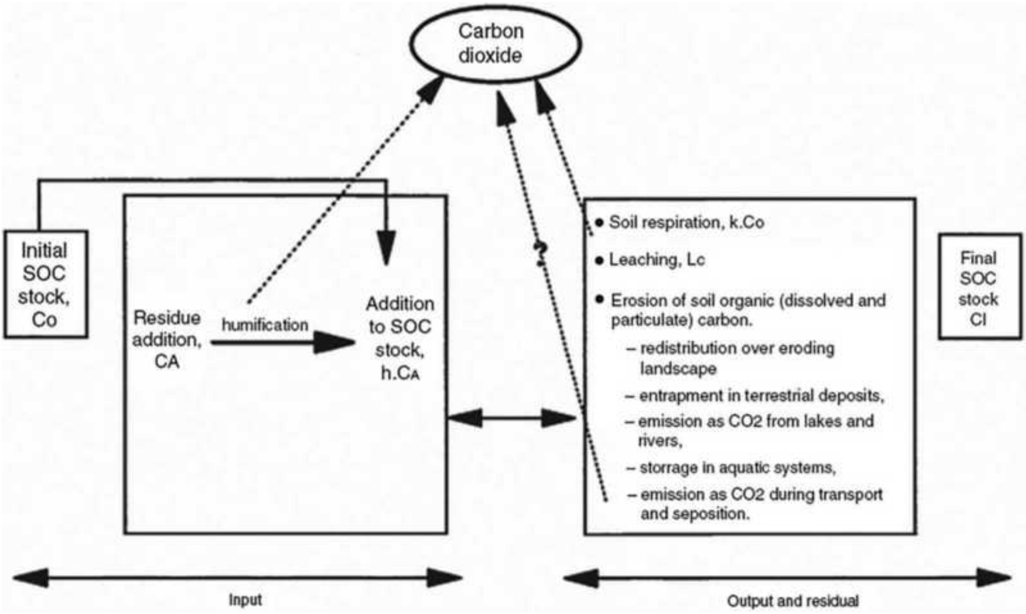


FIGURE 1 A conceptual mass balance model of SOC evolution in soils.

where C_0=initial SOC stocks, h=humification rate, C_A=annual residue addition, C_t=the SOC stocks after t years, S_R=annual SOC mineralization, k=coefficient of SOC mineralization, L_C=leaching loss of SOC, S_w=terrestrial C delivered to lakes and oceans, E_w=CO_2 evasion from aquatic systems (lakes, rivers, streams), S_L=C storage in terrestrial deposits, and E_L=mineralization and CO_2 evolution during transport and deposition of eroded SOC. The coefficients h and k are in units of y^{-1}, C_0 and C_t are in g C m^{-2}, and the other parameters are in units of g C m^{-2} y^{-1}.

Using Eq. (4), CO_2 evolution during transport and deposition of eroded SOC (E_L) can be computed. Values for the parameters in Eq. (3), assembled from various regional and global scale studies, are summarized in Table 1. We used the mass balance model and data from several long-term studies of SOC evolution to estimate EL.

TABLE 1 Estimates of the Major Pathways of Soil Organic Carbon Exported from Agricultural and Forested Watersheds

Description of SOC Loss Pathway	Symbol	Range (g C m^{-2} y^{-1})
Leaching of organic carbon in forest and cropland	L_C	0.5–5[a]
Terrestrial carbon delivered to lakes, rivers, and oceans	S_W	10–25[b]
Evasion CO_2 from aquatic systems	E_W	0.2–3.6[c]
Organic carbon storage in terrestrial deposits	S_l	6–15[d]

[a] Data from Guggenberger and Kaiser.[29]

[b] Terrestrial C delivery to the ocean [computed using global C input to the ocean (750×10^{12} g C/yr from Ludwig et al.[17]) and the total area of river draining to the ocean (89×10^{12} m^2 from Stallard[19])] was added to accumulation rate of terrestrial C in freshwater systems and man-made reservoirs (2–10 g C m^{-2} y^{-1} from Mulholland and Elwood[30]).

[c] Adapted from Spitzy and Leenheer.[31]

[d] Estimate computed using global C storage in terrestrial sediments ($0.6–15\times10^{15}$ g C y^{-1} from Stallard[19]) and total area of river basins (10^{14} m^2 from Meybeck[32]).

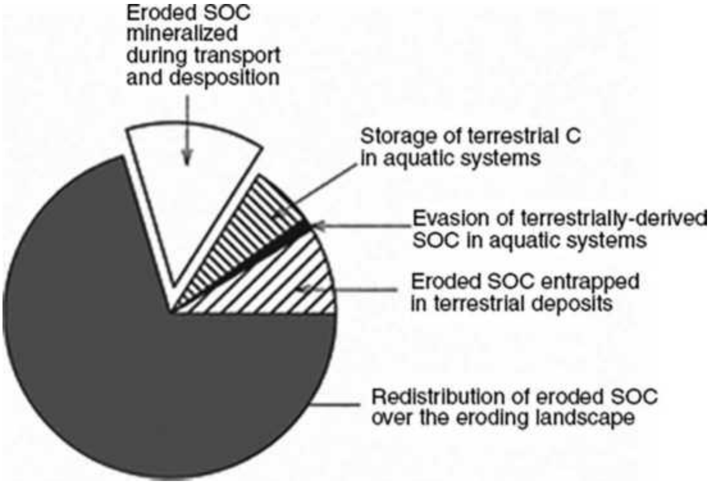


FIGURE 2 A hypothetical distribution of erosion-mobilized SOC among various possible fates.

Estimates of Erosion-Caused CO_2 Emissions

Analyses of the data available in the literature show that estimates of CO_2 production (E_L) during transport and deposition of eroded SOC range between 6 and 52 g C m^{-2} y^{-1} (Table 2). These estimates are necessarily crude given that the values assigned to the model parameters were obtained from different studies. It is important to note, however, that our parameter estimates (e.g., S_W, E_W) were similar to direct measurements made in studies conducted in various geographical regions including the United States,[22] Europe, and New Zealand.[16] Assuming a sediment delivery ratio of 15%[11] and a C-enrichment ratio of 2 for eroded soil material,[18] the total C exported from a typical eroded field averages 30% of the total C mobilized during erosional events. Using the C export data in Table 1, estimates of total C mobilized by water erosion range between 56 and 168 g C m^{-2} y^{-1}. Similar estimates (40 and 190 g C m^{-2} y^{-1}) of SOC mobilization are also obtained by assuming that computed E_L values represent 20% of the total SOC displaced by water erosion.[12] Using these ranges of erosion-induced SOC mobilization and the data in Table 1, a hypothetical distribution of eroded SOC among several pathways is presented in Figure 2. For example, for a soil containing $4–5\times10^3$ g C m^{-2}, the loss and redistribution of 1%–5% of the total SOC might go unnoticed for some time. Removal of SOC via sheet erosion could be particularly difficult to detect in the short term, hence a possible dismissal of water erosion as an important mechanism for SOC loss.[23] However, field study data support the view that erosion is a significant contributor to SOC decline and sometimes contributes more to SOC depletion than soil respiration.[24,25]

If the data in Table 2 are extrapolated to the global cropland (1.5×10^9 ha),[26] erosion-induced global CO_2 fluxes to the atmosphere ranging between 0.12 and 0.55 Pg C y^{-1} can be computed. These fluxes could represent up to one-third of the annual CO_2 emission (1.6 Pg C y^{-1}) due to deforestation and land-use change in the tropics.[13]

Conclusions

The information presented in this entry supports our contention that sizable amounts of CO_2 can be produced as a result of SOC mobilization by water erosion. Moreover, water erosion could elicit profound disturbances in the plant–soil–atmosphere segment of the global C cycle through a cascading process involving nutrient depletion, decreased biomass production, and alteration of the radiative and thermal properties of soils. The strong control of temperature on CO_2 emission[21] and the higher temperature

TABLE 2 Estimates of CO_2 Production during Transport and Deposition of Eroded Soil Organic Carbon Using a Mass Balance Approach

References	Cropping System and Location	Rate of Residue C Input	Rate of SOC Decline[c] (g C m^{-2} y^{-1})	Erosion-Induced CO_2 Flux, E[d]L[33]
Buyanovsky et al.[33]	Continuous wheat, Sanborn Field plots, Missouri, USA	698	42	17
Paul et al.[34]	Corn–soybean rotation, long-term ecological research (LTER) plots, Michigan, USA	425	42	21
Rasmussen et al.[35]	Wheat–fallow system, Pacific Northwest, USA	520–700[b]	0.6–2.8	6–52
Dumanski et al.[36][a]	Grain- and wheat-fallow systems in the Canadian prairies (Saskatchewan, Alberta) and Ontario, Canada	245–326	9.6–38.3	6–14
Lal[37]	Continuous corn, with and without N fertilization, International Institute of Tropical Agriculture (IITA), Nigeria	225[d]	25	29–32

[a] Compilation of several studies.

[b] This includes annual carbon input from wheat residue (252 g C m^{-2}) and manure (rates: 448–2680 g C m^{-2}).

[c] $(C_0 - C_t)/t$.

[d] To compute EL, values for h and k provided in the reference were used. If not available, h (0.16–0.2) and k (0.01–0.02) were used.

usually recorded in eroded soils[22] also indicate a shortened residence time of photosynthetically fixed C in the soil system. Although many of the effects remain to be evaluated experimentally, available information suggests a positive feedback of water erosion on atmospheric CO_2 and climate warming.

Acknowledgments

The authors gratefully acknowledge the technical support provided by Dr. J.M. Kimble of the National Soil Survey Center (NSSC) of Natural Resources Conservation Service (NRCS) at Lincoln, Nebraska.

References

1. IPCC. *The Physical Science Basis: Summary for Policy Makers*. Cambridge University Press: Cambridge, 2013.
2. IPCC. *Global Warming Special Report*. Cambridge University Press. Cambridge, 2018.
3. IPCC. *Land Use, Land Use Change and Forestry; Special Report; Intergovermental Panel on Climate Change*. Cambridge University Press: Cambridge, 2000.
4. WMO. *Greenhouse Gas Bulletin*. World Meterological Organization: Geneva, 2018.
5. Tubillo, F.N.; Salvatore, M.A.; Ferraro, A.F.; House, J. et al. The contribution of agriculture, forestry and other land use activities to global warming. *Global Change Biology* **2015**, *21*, 2655–2660.
6. Cook, G.D.; So, H.B.; Dalal, R.C. Structural degradation of 2 vertisols under continuous cultivation. *Soil and Tillage Research* **1992**, *24* (1), 47–64.
7. Wischmeier, W.H.; Smith, D.D. *Predicting Rainfall Erosion Losses: A Guide to Conservation Planning*; *Agriculture Handbook No. 537.* United States Department of Agriculture: Washington, DC, 1978; 58.
8. Lal, R. Soil erosion on Alfisols in western Nigeria. I. Effects of slope, crop rotation and residue management. *Geoderma* **1976**, *16*, 363–375.

9. Oldeman, L.R. The global extent of soil degradation. In *Soil Resilience and Sustainable Land Use*; Greenland, D.J., Szabolcs, I., Eds. CAB International: Wallingford, 1994; 99–118.
10. Milliman, J.D.; Syvitski, J.P.M. Geomorphic/tectonic control of sediment discharge to the ocean: the importance of small mountainous rivers. *Journal of Geology* **1992**, *100* (5), 325–344.
11. Walling, D.E. The changing sediment loads of the world rivers. *IAHS Publication* **2008**, *325*, 323–338.
12. Lal, R. Soil erosion and the global carbon budget. *Environment Intrenational* **2003**, *29*, 437–450.
13. Lal, R. Global soil erosion by water and carbon dynamics. In *Soil and Global Change*; Lal, R., Kimble, J.M., Levine, E., Stewart, B.A., Eds. CRC/Lewis: Boca Raton, FL, 1995; 131–142.
14. Lal, R.; Griffin, M.; Apt, J.; Lave, L.; Morgan, M.J. Managing soil carbon: a response. *Science* **2004**, *305*, 1567–1567.
15. IPCC. *Revised IPCC Guidelines for National Greenhouse Gas Inventories, Reference Manual*, vol. *3*. Intergovernmental Panel on Climate Change: Geneva, 1996.
16. Hope, D.; Billet, M.F.; Cresser, M.S. A review of the export of carbon in river water: fluxes and processes. *Environmental Pollution* **1994**, *84* (3), 301–324.
17. Ludwig, W.; Amiotte-Suchet, P.; Probst, J.L. River discharges of carbon to the world's oceans: determining local inputs of alkalinity and of dissolved and particulate organic carbon. *Comptes Rendus de l'Acade'mie des Sciences, Se'rie II., Sciences de la Terre et des Plane'tes* **1996**, *323* (12), 1007–1014.
18. van Noordwijk, M.; Cerri, C.; Woomer, P.L.; Nugroho, K.; Bernoux, M. Soil carbon dynamics in the humid tropical forest zone. *Geoderma* **1997**, *79* (1–4), 187–225.
19. Stallard, R.F. Terrestrial sedimentation and the carbon cycle: coupling weathering and erosion to carbon burial. *Global Biogeochemical Cycles* **1998**, *12* (2), 231–257.
20. Tiessen, H.; Stewart, J.W.B. Particle-size fractions and their uses in studies of soil organic matter: II. Cultivation effects on organic matter composition in size fractions. *Soil Science Society America Journal* **1983**, *47*, 509–514.
21. Beyer, L.; Frund, R.; Schleuss, U.; Wachendorf, C. Colluvisols under cultivation in Schleswig-Holstein. 2. Carbon distribution and soil organic matter composition. *Journal of Plant Nutrition and Soil Science* **1993**, *156* (3), 213–217.
22. Moeller, J.R.; Minshall, G.W.; Cummins, K.W.; Petersen, R.C.; Cushing, C.E.; Sedell, J.R.; Larson, R.A.; Vannote, R. Transport of dissolved organic carbon in streams of differing physiographic characteristics. *Organic Geochemistry* **1979**, *1*, 139–150.
23. Schlesinger, W.H. Changes in soil carbon storage and associated properties with disturbance and recovery. In *The Changing Carbon Cycle: A Global Analysis*; Trabalka, J.R., Reichle, D.E., Eds. Springer-Verlag: New York, 1986; 194–220.
24. Slater, C.S.; Carleton, E.A. The effect of erosion on losses of soil organic matter. *Soil Science Society America Proceedings* **1938**, *3*, 123–128.
25. Gregorich, E.G.; Anderson, D.W. Effects of cultivation and erosion on soils on four toposequences in the Canadian prairies. *Geoderma* **1985**, *36* (3–4), 343–354.
26. FAO. *Production Yearbook*, Vol. 52. Food and Agricultural Organization of the United Nations: Rome, 1998.
27. Fortin, M.C.; Rochette, P.; Pattey, E. Soil carbon dioxide fluxes from conventional and notillage smallgrain cropping systems. *Soil Science Society of America Journal* **1996**, *60* (5), 1541–1547.
28. Sutherland, R.A.; Menard, T.; Perry, J.L.; Penn, D.C. The influence of rolled erosion control systems on soil temperature and surface Albedo: Part I. A greenhouse experiment. *Land Degradation and Development* **1998**, *9* (2), 159–178.
29. Guggenberger, G.; Kaiser, K. Significance of DOM in the translocation of cations and acidity in acid forest soils. *Journal of Plant Nutrition and Soil Science* **1998**, *161* (2), 95–99.
30. Mulholland, P.J.; Elwood, J.W. The role of lake and reservoir sediments as sinks in the perturbed global carbon cycle. *Tellus* **1982**, *34* (5), 490–499.

31. Spitzy, A.; Leenheer, J. Dissolved organic carbon in rivers. In *Biogeochemistry of Major World Rivers*; Degens, E.T., Kempe, S., Richey, J.E., Eds. SCOPE 42. John Wiley and Sons: Chichester, NY, 1991; 214–232.
32. Meybeck, M. Concentrations des Eaux Fluviales en Eléménts Majeurs et Apports en Solution aux Océans. *Revue de Géologie Dynamique et de Géographie Physique* **1979**, *21*, 215–246.
33. Buyanovsky, G.A.; Kucera, C.L.; Wagner, G.H. Comparative analyses of carbon dynamics in native and cultivated ecosystems. *Ecology* **1987**, *68* (6), 2023–2031.
34. Paul, E.A.; Harris, D.; Collins, H.P.; Schulthess, U.; Robertson, G.P. Evolution of CO_2 and soil carbon dynamics in biologically managed, row crop agroecosystems. *Applied Soil Ecology* **1999**, *11* (1), 53–65.
35. Rasmussen, P.E.; Albrecht, S.L.; Smiley, R.W. Soil C and N changes under tillage and cropping systems in semiarid pacific northwest agriculture. *Soil and Tillage Research* **1998**, *47* (3–4), 197–205.
36. Dumanski, J.; Desjardins, R.L.; Tarnocai, C.; Monreal, C.; Gregorich, E.G.; Kirkwood, V.; Campbell, C.A. Possibilities for future carbon sequestration in Canadian agriculture in relation to land use changes. *Climatic Change* **1998**, *40* (1), 81–103.
37. Lal, R. Soil quality changes under continuous cropping for seventeen seasons of an alfisol in Western Nigeria. *Land Degradation and Development* **1998**, *9* (3), 259–274.

28

Soil Quality: Carbon and Nitrogen Gases

Philippe Rochette,
Sean McGinn, and
Reynald Lemke

Carbon Gases

Carbon is a major constituent of biomass and soil organic matter. However, more than 99% of global carbon is locked into sediments and fossil forms and is not available for biological processes. The small remaining active fraction of global carbon transits between atmospheric CO_2, biomass and soil organic matter, and detritus in the so-called carbon cycle (Figure 1). The carbon cycle is driven by photosynthetic fixation of atmospheric CO_2 by plants. In global terrestrial ecosystems, it is estimated that plant photosynthesis fixes more than 200 Gt of CO_2 every year.[1] Eventually, similar amounts are returned to the atmosphere by the respiration of animals and by the aerobic heterotrophic decomposition of soil organic matter and plant litter. In the absence of oxygen and other electron acceptors, CH_4 is the final product of soil organic matter decomposition. Human activities, including changes in land use and soil management, are contributing to unprecedented rapid increases in atmospheric CO_2 and

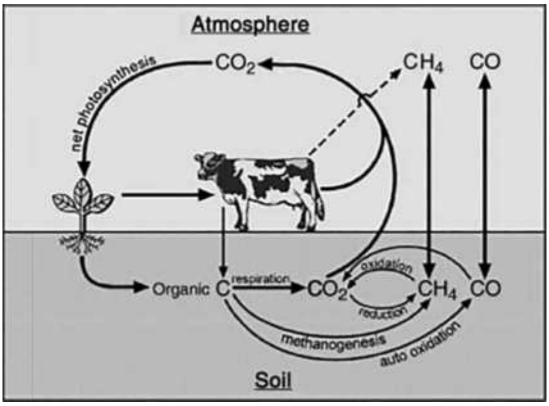


FIGURE 1 The soil carbon cycle. In well-aerated soils, decomposition of soil organic matter produces carbon dioxide. Significant amounts of methane are produced only under highly anaerobic (waterlogged) soilas.

CH_4 concentrations, which may result in important modifications of the Earth's climate. In dry soils, auto-oxidation of organic compounds can produce carbon monoxide (CO). Carbon monoxide can also be biologically oxidized to CO_2 in moist but well-aerated soils.

Carbon Dioxide

Soil-surface emitted CO_2, or soil respiration, is the sum of the CO_2 produced by root respiration and heterotrophic decomposition of root exudates, soil organic matter and plant litter. Decomposition processes, while dominated by soil microbes, are the result of complex interactions between soil fauna, fungi, actinomycetes, and bacteria. During decomposition, complex molecules like cellulose, hemicellulose, proteins, and lignin are broken down into low molecular weight substances and oxidized to CO_2 to produce energy and C for the growth of organisms.[2,3] The rate of decomposition is regulated by the quality and quantity of organic substrates and by physical/environmental properties of soil, such as temperature, moisture, and aeration.[4]

When ecosystems are in equilibrium, soil CO_2 emissions are the result of the natural recycling of nutrients and equal the amount of atmospheric CO_2 fixed by plant photosynthesis. However, in the past 150 years, human activity has broken this equilibrium by burning fossil carbon reserves and by decreasing soil organic matter through land- use changes. During this period, the additional CO_2 that entered the carbon cycle could not be completely absorbed by increases in autotrophic activity and atmospheric CO_2 concentrations increased from 280 to 365 ppm.

Several measures have been identified to reduce net CO_2 emissions from soils. Among them, it is believed that modifications in agricultural and forest management practices could result in increased C storage in soils as organic matter. Converting cultivated land into natural ecosystems usually increases soil carbon content because of increased return of plant litter and decreased decomposition. Also, tillage breaks down aggregates and increases soil carbon losses by exposing organic matter to microbes that decompose it. Therefore, reducing tillage intensity, fallow frequency, and the amount of cultivated land (by increasing productivity), and increasing perennial crops in rotation and permanent grasslands are seen as potential management practices for mitigating global warming.[5]

Methane

Methane is produced in soils by the decomposition of organic matter and by the reduction of CO_2 under highly anaerobic environments. Such conditions are found in wetlands and in rice paddies that, together with landfills, contribute about half of the total emissions of anthropogenic CH_4. Methane is stable in waterlogged soils and can be emitted to the atmosphere via diffusion, ebullition, and transport through plants. In the presence of oxygen, certain bacteria can oxidize CH_4 to CO_2 (Figure 1).[6]

Human intervention can greatly influence CH_4 production or consumption in soils. Flooding of soils in natural or agricultural ecosystems usually results in increased emissions[7] while drainage of wetlands can turn a source of methane into a sink.[8] Flooded rice fields are a major source of anthropogenic CH_4. Management practices have been proposed to reduce CH_4 emissions from rice paddies, including draining the fields during the growing season, replacing urea by other types of nitrogen fertilizers, and reducing the input of crop residues by using new cultivars and alternative cultural practices.[5] Reduced rates of CH_4 oxidation in well-aerated soils have been observed following cultivation and addition of nitrogen fertilizers.[9]

Nitrogen Gases

The passage of nitrogen (N) through ecosystems can be represented as a "loop-within-a-loop" (Figure 2). Nitrogen enters ecosystems primarily through biotic or abiotic processes that "fix" (convert molecular nitrogen [N_2] to biologically available forms) N_2 from the atmosphere. Within the soil-plant system, N

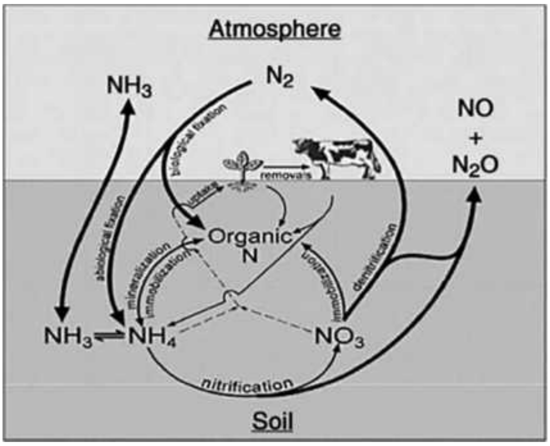


FIGURE 2 Schematic representation of the nitrogen cycle depicted as a “loop-within-a-loop.” The outer loop traces the passage of N from the atmosphere through ecosystems and back to the atmosphere, while the inner loop traces the interconversion of N between organic and inorganic forms within the soil–plant system.

undergoes a complex series of transformations, resulting in a continual transfer of N between inorganic and organic forms—the inner loop (Figure 2). Nitrous oxide and NO_x are both produced and consumed during these transformations. With time, most of the N entering ecosystems returns to the atmosphere as N_2, but an important fraction is emitted as gaseous NH_3, NO_x, and N_2O. More than 70% of the estimated 18 Tg N_2O–N entering the atmosphere each year is emitted by soils.[1,10] Above plant-canopy emissions of NO_x from soils are probably in the range of 3.3–21 Tg N yr^{-1}.[11,12] This exchange of N between the atmosphere and the soil-plant system— the outer loop—strongly mediates atmospheric concentrations of NH_3, NO_x, and N_2O.

Nitric and Nitrous Oxides

The majority of soil-emitted NO_x is nitric oxide (NO);[13] therefore, the rest of this discussion will focus on nitrogen gases other than nitrogen dioxide (NO_2). Nitric oxide is derived primarily via autotrophic nitrification,[14,15] whereas N_2O emissions can be a product of autotrophic nitrification,[15] denitrification,[16] or a combination of both.[14] Nitrous oxide, and particularly NO, may also be emitted from various chemical reactions collectively known as chemodenitrification.[17]

The magnitude of N_2O and NO emissions is strongly influenced by site-specific variables—most particularly soil attributes—and human intervention. Given that both gases are produced primarily during microbial transformations of inorganic N, the potential of a soil to produce and emit N_2O or NO increases with increasing N availability.[18] Anthropogenic activities that increase the flow of N into the system also increase emissions of N_2O and NO. Such activities include, for example, the cultivation of legume grain or forage crops,[19] and intensive use of nitrogen fertilizers.[20]

Total N_2O emissions tend to increase as soil organic carbon (SOC) content increases. This is understandable, since N turnover is closely tied to SOC turnover, and the amount of carbon (C) available to drive microbial processes is directly related to the quantity and quality of SOC. The relative availability of C and N is of particular importance when livestock manure or sewage sludge is added to soils. The addition of both C and N tend to increase N_2O and NO emissions more than N alone, particularly in C limited systems.[21,22]

Agricultural management practices that minimize soil disturbance tend to increase soil water status and soil bulk density, favoring the development of anaerobic conditions and higher N_2O[4] but possibly lower NO emissions.[23] Conversely, equal or lower emissions of N_2O have also been reported in no-tillage systems compared to conventional tillage systems.[24] In this instance, reduced soil disturbance appeared to alter N cycling, resulting in lower NO_3^- availability—hence lower N_2O emissions—during the spring thaw period.

In general, soil conditions or management that lead to accumulations of NH_4^+, NO_3^-, and particularly NO_2^-, favor gaseous N losses. This may happen if N is released during soil C mineralization at a time when plant uptake is absent or minimal, an example being bare soil fallow periods commonly employed in the semiarid regions of North America, or if fertilizer or organic N is added at a time when crop growth is not vigorous. Low soil pH levels (<5) inhibit NO_2^- oxidizers more than NH_4^+ oxidizers, resulting in accumulations of NO_2^- and abiotic production of NO.[25] The latter may occur even in near-neutral soils as a result of very high concentrations of NH_4^+ localized in urine patches, or concentrated in bands or nests of ammonium based fertilizers.[26]

Ammonia

The production of ammonia (NH_3) in soils is closely linked to the accumulation of ammonium (NH_4^+) resulting from the mineralization of organic matter (ammonification) by soil microbes. In agricultural soils, the NH_4^+ content is enhanced where inorganic nitrogen and livestock manure fertilizers are used. In soil solution, NH_4^+ dissociates and forms an equilibrium with NH_3, which can volatilize to the soil air and eventually reach the atmosphere [Eq. (1)]. Ammonium is also bound to negative exchange sites on soil organic matter and clay minerals (exchangeable NH_4^+) and can be "fixed" within clays. The latter is released slower than exchangeable NH_4^+.[27] Ammonium is converted through biological oxidation (nitrification) by soil bacteria (nitrifiers) to nitrate (NO_3^-) which is taken up by plants. Atmospheric NH_3 and NH_4^+ can be recycled back to the soil through dry and wet deposition (Figure 2).

$$\begin{array}{l} NH^+_{4(\text{solution})} \Leftrightarrow NH^+_{4(\text{exchangeable})} \Leftrightarrow NH^+_{4(\text{non-exchangeable})} \\ \Updownarrow \\ NH_{3(\text{solution})} \Leftrightarrow NH_{3(\text{soil air})} \Leftrightarrow NH_{3(\text{atmosphere})} \end{array}$$

The volatilized NH_3 from soil is not only a loss of valuable nitrogen for crop growth [28] but is a significant source of atmospheric NH_3[29,30] which can damage plants[31,32] and contribute to odor nuisances that are associated with the application of manure to soils. In the atmosphere, NH_3 neutralizes atmospheric acids and the resulting aerosols contribute to atmospheric haze. Atmospheric NH_3 can also cause acidification of ecosystems by augmenting the capture of sulfur dioxide in clouds[33] and through nitrification of deposited NH_4^+ and NH_3.[34]

Most management techniques that relate to minimizing NH_3 emissions pertain to managing inorganic and livestock manure fertilizers, which are the most significant sources of this gas.[30,35] The key to these techniques is to reduce exposure of the amendment to the atmosphere either by limiting the amount applied or by injection/incorporation of the amendment. Manipulating the chemistry of the soil can also be used to control NH_3 emissions especially in amended soils.[36,37] For example, the addition of acids to fertilized soils can potentially reduce NH_3 emissions. Adding heavy metals or organic compounds with urea fertilizer can inhibit the conversion of urea to NH_4^+ by retarding urease enzyme activity. There is also the potential of adding Ca or Mg salts with urea fertilizers to reduce NH_3 emissions.

Meteorological and soil conditions also modify the emission of NH_3. Application of manure when soil moisture is initially high but subject to rapid drying can result in higher than normal emissions of NH_3.[36,38] High soil moisture content would put more NH_4^+ into solution near the soil surface and increase volatilization. However, irrigation or rainfall following application can leach nearsurface NH_4^+ and decrease volatilization.

Conclusions

Soil quality and carbon and nitrogen gas dynamics are closely linked. In general, high-quality soils have high nutrient availability and aeration, good infiltration and retention of water, and a stable structure, all of which are factors that control the production and consumption of gases in soils. Improving soil quality through higher soil organic C content or better soil structure often results in a reduction of net emissions of greenhouse gases into the atmosphere. Also, soil degradation increases emissions of greenhouse gases as a result of greater use of nitrogen fertilizers, erosion, denitrification, deforestation, and grassland cultivation.[39] Therefore, adoption of most management practices aimed at improving soil quality also help to improve atmospheric quality.

References

1. Duxbury, J.M.; Harper, L.A.; Mosier, A.R. Contributions of agroecosystems to global climate change. In *Agricultural Ecosystem Effects on Trace Gases and Global Climate Change*; Harper, L.A., Mosier, A.R., Duxbury, J.M., Eds.; ASA Special Publication No. 55; American Society of Agronomy: Madison, WI, 1993; 1–18.
2. Kilham, K. The ecology of soil nutrient cycling. In *Soil Ecology;* Cambridge University Press: New York, 1994; 89–108.
3. Coleman, D.C.; Crossley, D.A. Decomposition and nutrient cycling. In *Fundamentals of Soil Ecology;* Associated Press: New York, 1996; 109–140.
4. Linn, D.M.; Doran, J.W. Effect of water-filled pore space on carbon dioxide and nitrous oxide production in tilled and nontilled soils. Soil Sci. Soc. Am. J. **1984**, *48,* 1267–1272.
5. Intergovernmental Panel on Climate Change. In *Impacts, Adaptations and Mitigation of Climate Change: Scientific- Technical Analyses;* Watson, R.T., Zinyowera, M.C., Moss, R.H., Eds.; Cambridge University Press: New York, 1995.
6. Topp, E.; Pattey, E. Soils as sources and sinks for atmospheric methane. Can. J. Soil Sci. **1997**, *77,* 167–178.
7. Duchemin, E.; Lucotte, M.; Canuel, R.; Chamberland, A. Production of greenhouse gases CH_4 and CO_2 by hydroelectric reservoirs of the boreal region. Global Biogeochem. Cycles **1995**, *9*, 529–540.
8. Roulet, N.T.; Ash, R.; Quinton, W.; Moore, T.R. Methane flux from drained northern peatlands: effect of a persistent water table lowering on flux. Global Biogeochem. Cycles **1993**, *7*, 749–769.
9. Mosier, A.R.; Scimel, D.; Valentine, D.; Bronson, K.; Par- ton, W. Methane and nitrous oxide fluxes in native fertilized and cultivated grasslands. Nature **1991**, *350,* 330–332.
10. Kroeze, C.; Mosier, A.; Bouwman, L. Closing the global N_2O budget: a retrospective analysis 1500–1994. Global Biogeochem. Cycles **1999**, *13* (1), 1–8.
11. Delmas, R.; Serca, D.; Jambert, C. Global inventory of NO_x sources. Nutr. Cycl. Agroecosyst. **1997**, *48,* 51–60.
12. Davidson, E.A.; Kingerlee, W. A global inventory of nitric oxide emissions from soils. Nutr. Cycl. Agroecosyst. **1997**, *48*, 37–50.
13. Hutchinson, G.L.; Davidson, E.A. Processes for production and consumption of gaseous nitrogen oxides in soil. In *Agricultural Ecosystem Effects on Trace Gases and Global Climate Change;* American Society of Agronomy: Madison, WI, 1993; 79–93.
14. Skiba, U.; Smith, K.A.; Fowler, D. Nitrification and denitrification as sources of nitric oxide and nitrous oxide in a sandy loam soil. Soil Biol. Biochem. **1993**, *25* (11), 1527-1536.

15. Hutchinson, G.L.; Guenzi, W.D.; Livingston, G.P. Soil water controls on aerobic soil emissions of gaseous nitrogen oxides. Soil Biol. Biochem. **1993**, *25* (1), 1–9.
16. Williams, P.H.; Jarvis, S.C.; Dixon, E. Emission of nitric oxide and nitrous oxide from soil under field and laboratory conditions. Soil Biol. Biochem. **1998**, *30* (14), 1885–1893.
17. Haynes, R.J.; Sherlock, R.R. Gaseous losses of nitrogen. In *Mineral Nitrogen in the Plant-Soil System;* Haynes, R.J., Ed.; Academic Press: New York, 1986; 242–286.
18. Matson, P.A.; Vitousek, P.M. Cross-system comparisons of soil nitrogen transformations and nitrous oxide flux in tropical forest ecosystems. Global Biogeochem. Cycles **1987**, *1*, 163–170.
19. Galbally, I.E. Biosphere-atmosphere exchange of trace gases over Australia. In *Australia's Renewable Resources: Sustainability and Global Change;* Gifford, R.M., Barson, M.M., Eds.; Bureau of Rural Resources: Canberra, Australia, 1992; 117–149.
20. Smith, K.A.; McTaggart, I.P.; Tsuruta, H. Emissions of N_2O and NO associated with nitrogen fertilization in intensive agriculture, and the potential for mitigation. Soil Use Management **1997**, *13*, 296–304.
21. Rochette, P.; Angers, D.A.; van Bochove, E.; Prevost, D.; Cote, D.; Bertrand, N. Soil C and N dynamics following application of pig slurry for the 19th consecutive year: II— mineral N and N_2O fluxes. Soil Sci. Soc. Am. J. **2000**, *64,* 1396–1403.
22. Thornton, F.C.; Shurpali, H.J.; Bock, B.R.; Reddy, K.C. N_2O and NO emissions from poultry litter and urea applications to Bermuda grass. Atmos. Environ. **1998**, *32* (9), 1623–1630.
23. Civerolo, K.L.; Dickerson, R.R. Nitric oxide soil emissions from tilled and untilled cornfields. Agri. Forest Meteorol. **1998**, *90* (4), 307–311.
24. Lemke, R.L.; Izaurralde, R.C.; Nyborg, M.; Solberg, E.D. Tillage and N-source influence soil-emitted nitrous oxide in the Alberta parkland region. Can. J. Soil Sci. **1999**, *79*, 15–24.
25. Yamulki, S.; Harrison, R.M.; Goulding, K.W.T.; Webster, C.P. N_2O, NO and NO_2 fluxes from a 1398 Quality and Production of Carbon and Nitrogen Gases grassland: effect of soil pH. Soil Biol. Biochem. **1997**, *29* (8), 1199–1208.
26. Burns, L.C.; Stevens, R.J.; Smith, R.V.; Cooper, J.E. The occurrence and possible sources of nitrite in a grazed, fertilized, grassland soil. Soil Biol. Biochem. **1995**, *27* (1), 47–59.
27. Mengel, K. Dynamics and availability of major nutrients in soils. Adv. Soil Sci. **1985**, *2*, 65–131.
28. McGinn, S.M.; Janzen, H.H. Ammonia sources in agriculture and their measurement. Can. J. Soil Sci. **1998**, *78*, 139–148.
29. Bouwman, A.F.; Lee, D.S.; Asman, W.A.H.; Dentener, F.J.; Van Der Hoek, K.W.; Olivier, J.G.J. A global high-resolution emission inventory for ammonia. Global Biogeochem. Cycles **1997**, *11* (4), 561–587.
30. Duxbury, J. The significance of agricultural sources of greenhouse gases. Fertilizer Res. **1994**, *38,* 151–163.
31. Van der Eerden, L.J.M. Toxicity of ammonia to plants. Ag- ric. Environ. **1982**, *7*, 223–235.
32. Van der Eerden, L.J.M.; de Visser, P.H.B.; van Dijk, C.J. Risk of damage to crops in the direct neighbourhood of ammonia sources. Environ. Pollut. **1998**, *102,* 49–53.
33. ApSimon, H.M.; Kruse, M.; Bell, J.N.B. Ammonia emission and their role in acid deposition. Atmos. Environ. **1987**, *21* (9), 1939–1946.
34. Fahey, T.J.; Wiliams, C.J.; Rooney-Varga, J.N.; Cleveland, C.C.; Postek, K.M.; Smith, S.D.; Bouldin, D.R. Nitrogen deposition in and around an intensive agricultural district in central New York. J. Environ. Qual. **1999**, *28,* 1585–1600.
35. Buijsman, E.; Maas, H.F.M.; Asman, W.A.H. Anthropogenic NH_3 emissions in Europe. Atmos. Environ. **1987**, *21* (5), 1009–1022.
36. Hargrove, W.L. Soil, environmental, and management factors influencing ammonia volatilization under field conditions. *In Ammonia Volatilization from Urea Fertilizers*; Brock, B.R., Kissel, D.E., Eds.; National Fertilizer Development Center, Tennessee Valley Authority: Muscle Shoals, AL, 1988; 17–36.

37. Fenn, L.B.; Hossner, L.R. Ammonia volatilization from ammonium or ammonium-forming nitrogen fertilizers. Adv. Soil Sci. **1985**, *1,* 124–169.
38. Brunke, R.; Alvo, P.; Schuepp, P.; Gordon, R. Effect of meteorological parameters on ammonia loss from manure in the field. J. Environ. Qual. **1988**, *17* (3), 431–436.
39. Lal, R.; Kimble, J.; Stewart, B.A. World soils as a source or sink of radiatively-active gases. In *Soil Management and Greenhouse Effect;* Lal, R., Kimble, J., Stewart, B.A., Eds.; Advances in Soil Science; CRC Press: Boca Raton, 1995; 1–7.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

29

Sustainable Agriculture: Soil Quality

John W. Doran and
Ed G. Gregorich

Introduction

Growing human populations, diminishing nonrenewable resources, social and political instability, and environmental degradation threaten the natural processes that sustain the global ecosphere and life on Earth.[1,2] With losses of farmland to degradation and urbanization and little new agricultural land to develop, meeting the needs of future populations for food, fiber, and other agricultural products depends on a significant increase in crop yields. Under current food production practices, this yield increase spells greater use of inputs (e.g., energy, fertilizer, pesticides), and raises the question of the environmental costs of increased production.

Agriculture and Environmental Sustainability

To a large extent, the economic and social viability of agriculture depends on the sustainability of the resources that agriculture both depends on and affects. Farmers have always intuitively known that good soil quality underpins the success of their operations, and soil conservation programs in many countries throughout the last century supported this recognition. Since the late 1980s, interest has grown in a fuller accounting of the environmental costs of agricultural production. These costs include declining water quality and competition for finite water resources, declining air quality, emission of greenhouse gases, loss of wildlife habitat, and decline in species and genetic diversity.[3] To a large extent, conserving and enhancing soil quality can mitigate these costs.

Soil

Agricultural land management in the past has taxed the soil system, often withdrawing more than it has returned. In particular, mechanical cultivation and the continuous production of row crops has resulted in the displacement and loss of soil by erosion, large decreases in soil organic matter, and a concomitant release of carbon as carbon dioxide into the atmosphere.[4] Within the last decade, inventories of soil productive capacity indicate human-induced degradation on nearly 40% of the Earth's arable land as a result of land clearing, extensive soil cultivation, soil erosion, atmospheric deposition of pollutants, over-grazing, salinization, and desertification.[5] The projected doubling of the human population in the next century threatens even greater degradation of soils and other natural resources.[6]

Water

Intensive farming practices have, in many parts of the world, jeopardized the quality of surface water and groundwater through the addition of agriculturally derived nitrate nitrogen, phosphorus, pesticides, sediment, and pathogens (e.g., bacteria from manure). Agriculture is considered the most widespread contributor to nonpoint-source water pollution in the U.S.[7] The major agricultural water contaminant in North America and Europe is nitrate nitrogen derived from atmospheric deposition, livestock manures, and commercial fertilizers. Human alterations of the nitrogen cycle have almost doubled the rate of nitrogen input to terrestrial ecosystems over the past 30 years, resulting in large increases in the transfer of nitrogen from land to the atmosphere and to rivers, estuaries, and coastal waters.[8–10]

Agriculture, competing as it does with many other uses of fresh water, is also at the heart of current discussions of water quantity and is a key consideration in the trend toward demand management, rather than supply management, of water resources. Global warming scenarios predict water shortages in some existing agricultural areas in the world, and water surpluses in others. Changing conditions will call for new ways of farming and of resolving conflict among water users.

Controlling soil erosion reduces the amount of sediments entering waterways, along with the chemicals (e.g., pesticides) and pathogens they carry. Nutrient management plans that account for all sources of nitrogen and phosphorus and match nutrient application more closely to plant needs reduce the risk of nitrogen and phosphorus entering groundwater and surface water. Maintaining soil quality improves the ability of soil to receive and partition water, making the most efficient use of this resource.

Air and Atmosphere

Atmospheric concentrations of greenhouse gases—particularly nitrous oxide, methane, and carbon dioxide—have been increasing dramatically over the past 20 years, enhancing the greenhouse effect by which the Earth's atmosphere is warmed. Agriculture is a significant contributor of these gases to the atmosphere; in Canada, for example, agriculture accounted for about 10% of total 1996 emissions of these gases, an increase of about 4% since 1981.[11] Soil management practices such as tillage, cropping patterns, and the use of manure and commercial fertilizers influence atmospheric quality through changes in the soil's capacity to produce or consume greenhouse gases.[12,13] Under the United Nations Framework Convention on Climate Change and its Protocols, signatory countries are considering ways to reduce emissions, including those from agriculture, and also to capitalize on agricultural sinks of greenhouse gases, such as by carbon sequestration in agricultural soils. The present threat of ozone depletion and global climate change necessitates a better understanding of the influence of land management on soil processes.

Agricultural emissions of ammonia (from fertilizer and manure) and particulate matter (dust released during soil tillage and subsequent erosion) have been linked to various environmental effects, such as acidification, eutrophication, and smog. Reducing the emission of these substances involves improving the quality of soils so that they are more resistant to erosion, reducing tillage and increasing the amount of soil cover, and practicing better nitrogen management (both manure and commercial fertilizer).

Agroecosystem Biodiversity

Agriculture benefits from biodiversity in many ways, but it has also reduced biodiversity over the years, mainly through the conversion of natural habitats, but also through effects on soil and water quality.

Improving soil quality not only maintains a healthy biological community in the soil itself, along with the functions these organisms carry out, but also supports the other flora and fauna dependent on this community. Soil quality improvements that benefit water quality can also improve the viability of aquatic ecosystems.

Assessment of Soil Quality

Assessment of soil quality or health is invaluable in determining the sustainability of land management systems.[14] Soil quality is the major link between the strategies of conservation management practices and the achievement of the major goals of agriculture.[15,16]

Assessment of soil quality can be used to identify problem production areas, make realistic estimates of food production, monitor changes in sustainability and environmental quality as related to agricultural management, and assist government agencies in formulating and evaluating sustainable agricultural and land use policies (Figure 1). Using simple indicators of soil quality and soil health that have meaning to farmers and other land managers is likely the most fruitful means of linking science with practice in assessing the sustainability of land management practices.[17]

Doran et al.[18] stressed the importance of holistic management approaches that optimize the multiple functions of soil, conserve soil resources, and support strategies for promoting soil quality. They proposed a basic set of indicators to assess soil quality and health in various agricultural management systems. Many of these key indicators are very useful to specialists (e.g., researchers, consultants, extension staff, and conservationists) but are beyond the expertise or time constraints of the producer to measure. Success in developing and implementing standards for assessing soil quality and sustainability hinges on partnership with agricultural producers, who are the primary stewards of the land and the chief decision makers regarding land use and management. In light of this consideration, Doran et al.[18] also presented strategies for sustainable management that include a more generic set of indicators to assess soil quality and health that are practical for producers (Table 1). A study conducted in the northern U.S. corn and dairy belt used a similar approach to determine how farmers assess soil quality and health and found that they ranked soil organic matter, crop appearance, and risk to erosion as the three most important properties for describing soil health and sustainable management.[17] Such strategies maximize the benefits of natural cycles, reduce dependence on nonrenewable resources, and help producers

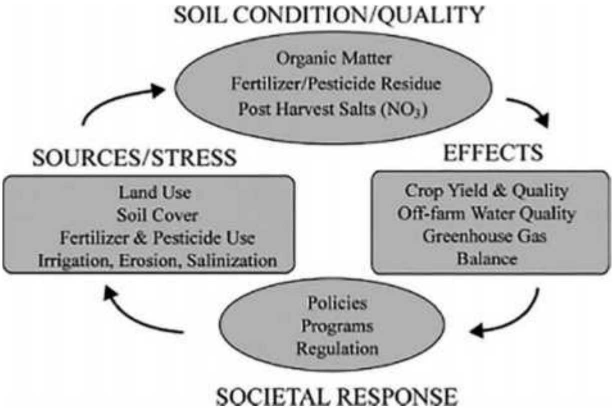


FIGURE 1 The politics of soil health. The assessment of soil condition/quality is needed to monitor changes in sustainability and environmental quality as related to agriculture management and to assist governmental agencies in formulating realistic agricultural and land use policies.

TABLE 1 Strategies for Sustainable Agricultural Management and Proposed Indicators of Crop Performance and Soil and Environmental Health

Sustainability Strategy	Indicators for Producers
Conserve soil organic matter through maintaining soil C and N levels by reducing tillage, recycling plant and animal manures, and/or increasing plant diversity where C inputs C outputs	Direction/change in organic matter levels with time (visual or remote sensing by color or chemical analysis), specific organic matter potential for climate, soil, and vegetation, soil water storage
Minimize soil erosion through conservation tillage and increased protective cover (residue, stable aggregates, covercrops, green fallow)	Visual (gullies, rills, dust, etc.) Surface soil properties (topsoil depth, organic matter content/texture, water infiltration, runoff, ponding, % cover)
Balance production and environment through conservation and integrated management systems (optimizing tillage, residue, water, and chemical use) and by synchronizing available N and P levels with crop needs during the year	Crop characteristics (visual or remote sensing of yield, color, nutrient status, plant vigor, and rooting characteristics) Soil physical condition/compaction Soil and water nitrate levels Amount and toxicity of pesticides used
Better use of renewable resources through relying less on fossil fuels and petrochemicals and more on renewable resources and biodiversity (e.g., crop rotations, legumes, manures, integrated pest management)	Input and output ratios of costs and energy Leaching losses/soil acidification Crop characteristics (as listed above) Soil and water nitrate levels

Source: Postel et al.[2]

identify long-term goals for sustainability that also meet short-term needs for production. Soil quality assessment must be directed at the financial survival of the farm as well as environmental preservation.

Although much remains to be done, useful models exist for translating soil science into practice. For example, Gomez et al.[19] provide a practical framework for determining the sustainability of hill country agriculture in the Philippines. It uses indicators that satisfy both the needs of the farmer (e.g., productivity, profitability, stability, and viability) and the conservation of soil and water resources. Threshold values for sustainability are identified relative to the average local conditions for crop yield, profit, risk of crop failure, soil depth, percent soil cover, and soil organic matter content. This conceptual framework for assessing sustainability could be expanded to include other needs of society and environmental conservation. In particular, adding a category for balancing energy input and output, as well as monetary costs, would better assess the short- and long-term sustainability of management and the value of greater reliance on renewable resources and less dependence on fossil fuels and petrochemicals in enhancing economic, ecological, and environmental resources. Expanding the list of resource conservation variables to include leachable salts (especially nitrate), measured as soil electrical conductivity at the time of fertilization and after harvest, would permit land managers to better quantify the impact of agricultural practices on air and water quality.

Confirmation of the effectiveness of systems for residue management, organic matter formation, nitrogen and carbon cycling, soil structure maintenance, and biological control of pests and diseases will assist in discovering approaches that are both profitable and environmentally sound. The challenge in the future will be to better use the diversity and resiliency of the soil biological community to maintain a quality ecosystem, thus fostering sustainability. Sustainability strategies must be fine-tuned using such practices as crop rotation for greater crop diversity and tighter cycling of nutrients, reduction of soil disturbance to maintain soil organic matter and reduce erosion, and development of systems that make greater use of renewable biological resources. Ultimately, the indicators of soil quality and strategies for sustainable management must be linked to the development of management systems that foster reduction in the inputs of nonrenewable resources, maintain acceptable levels of productivity, and minimize impact on the environment.

References

1. Bouma, J. Soil environmental quality: a European perspective. J. Environ. Qual. **1997**, *26*, 26–31.
2. Postel, S.; et al. Carrying capacity: earth's bottom line. In *State of the World; Brown,* L.R., Ed.; W.W. Norton & Co.: New York, 1994; 3–21.
3. Lal, R. Basic concepts and global issues: soil quality and agricultural sustainability. In *Soil Quality and Agricultural Sustainability*; Lal, R., Ed.; Ann Arbor Press: Chelsea, MI, 1998; 3–12.
4. Houghton, R.A.; Hobbie, J.E.; Melillo, J.M.; Moore, B.; Peterson, B.J.; Shaver, G.R.; Woodwell, G.M. Changes in the carbon content of terrestrial biota and soils between 1860 and 1980: a net release of CO_2 to the atmosphere. Ecol. Monogr. **1983**, *53*, 235–262.
5. Doran, J.W.; Sarrantonio, M.; Liebig, M.A. Soil health and sustainability. In *Advances in Agronomy;* Sparks, D.L., Ed.; Academic Press: San Diego, CA, 1996; 1–54.
6. Ruttan, V.W. The transition to agricultural sustainability. Proc. Natl. Acad. Sci. **1999**, *96*, 5960–5967.
7. National research council. *In Soil and Water Quality: An Agenda for Agriculture, Committee on Long-Range Soil and Water Conservation, Board on Agriculture, National Research Council;* National Academy Press: Washington, DC, 1993; 516 pp.
8. Matson, P.A.; Parton, W.J.; Power, A.G.; Swift, M.J. Agricultural intensification and ecosystem processes. Science **1997**, 277, 504–509.
9. Socolow, R.H. Nitrogen management and the future of food: lessons from the management of energy and carbon. Proc. Natl. Acad. Sci. **1999**, *96*, 6001–6008.
10. Vitousek, P.M.; Aber, J.D.; Howarth, R.W.; Likens, G.E.; Matson, P.A.; Schindler, D.W.; Schlesinger, W.H.; Tilman, D.G. Human alteration of the global nitrogen cycle: sources and consequences. Ecol. Appl. **1997**, *7*, 737–750.
11. McRae, T.; Smith, C.A.S.; Gregorich, L.J., Eds. *Environmental Sustainability of Canadian Agriculture*; Report of the Agri-Environmental Indicator Project; Agriculture and Agri-Food Canada: Ottawa, Ont, 2000; 224 pp.
12. Mosier, A.R. Soil processes and global change. Biol. Fert. Soils **1998**, *27*, 221–229.
13. Rolston, D.E.; Harper, L.A.; Mosier, A.R.; Duxbury, J.M. *Agricultural Ecosystem Effects on Trace Gases and Global Climate Change;* American Society of Agronomy Spec. Publ. 55; American Society of Agronomy: Madison, WI, 1993.
14. Karlen, D.L.; Mausbach, M.J.; Doran, J.W.; Cline, R.G.; Harris, R.F.; Schuman, G.E. Soil quality: a concept, definition, and framework for evaluation. Soil Sci. Soc. Am. J. **1997**, *61*, 4–10.
15. Acton, D.F.; Gregorich, L.J. *The Health of Our Soils: Toward Sustainable Agriculture in Canada*; Agriculture and Agri-Food Canada: Ottawa, Ont, 1995.
16. Parr, J.F.; Papendick, R.I.; Hornick, S.B.; Meyer, R.E. Soil quality: attributes and relationship to alternative and sustainable agriculture. Am. J. Altern. Agric. **1992**, *7*, 5–11.
17. Romig, D.E.; Garlynd, M.J.; Harris, R.F.; McSweeney, K. How farmers assess soil health and quality. J. Soil Water Conserv. **1995**, *50*, 229–236.
18. Doran, J.W.; Jones, A.J.; Arshad, M.A.; Gilley, J.E. Determinants of soil quality and health. In *Soil Quality and Soil Erosion;* Lal, R., Ed.; CRC Press: Boca Raton, FL, 1999; 17–36.
19. Gomez, A.A.; Swete Kelly, D.E.; Seyers, J.K.; Coughlan, K.J. Measuring sustainability of agricultural systems at the farm level. In *Methods for Assessing Soil Quality*; Doran, J.W., Jones, A.J., Eds.; Soil Sci. Soc. Am. Spec. Publ. 49; SSSA: Madison, WI, 1996; 401–410.

III

CSS: Case Studies of Environmental Management

30

Drought and Agricultural Production in the Central Andes

Claudia Canedo-Rosso, Ronny Berndtsson, and Cintia B. Uvo

Introduction

Agricultural production is highly sensitive to climate hazards, including droughts. Losses due to drought events represent a significant challenge to farmers, agricultural stakeholders, and policymakers worldwide (UNISDR, 2015). The scientific community predicts an intensification of negative impacts due to future climate change (IPCC, 2013). The UN Sustainable Development Goals (SDGs) state that the priority areas for adaptation to climate change are water and agriculture. These, in turn, are related to the largest climate hazards that are drought, floods, and higher temperature (UN, 2016). Droughts originate from the lack of precipitation during a specific period of time, resulting in water shortages in a specific region (Wilhite and Glantz, 1985). The rainy season in the Central Andes generally occurs during the austral summer (DJFM). In this period, the position of the Atlantic intertropical convergence zone (ITCZ) and the formation of the Bolivian High result in a prevalence of easterly winds over the Central Andes (Garreaud et al., 2003). Easterly winds transport humid air from the lowlands east of the Central Andes to the region, while the prevalence of westerly winds inhibits moisture transport (Lenters and Cook, 1999; Thibeault et al., 2012; Vuille et al., 2000). The summer rainfall represents more than 70% of the total annual precipitation (Canedo-Rosso et al., 2019b; Garreaud et al., 2003). Spatially, precipitation in the Central Andes presents a northeast-southwest gradient with annual precipitation above 800 mm in the northeastern decreasing to below 200 mm in the south-western Central Andes (Vuille et al., 2000).

Agricultural production is directly affected by climate hazards in the Central Andes, and drought conditions may lead to shortage of food for humans and animals (Garcia, 2003; Garcia et al., 2007). This is the case for the most important crops in the Central Andes: quinoa and potato, which are affected by precipitation variability, and losses in crop production are connected to drought (Garcia et al., 2003). Precipitation in the Central Andes is strongly associated with the El Niño Southern Oscillation (ENSO), which warm phase is connected to droughts (Garreaud and Aceituno, 2001; Thompson et al., 1984; Vicente-Serrano et al., 2015). The ENSO is a periodical variation in sea surface temperature (SST) over the tropical Pacific Ocean with intervals of 2–7 years and may be represented by three phases: neutral, warm (El Niño), and cold (La Niña) phases. In Bolivia, large socio-economic losses have been experienced in the past due to ENSO. For example, El Niño 1982–1983 affected 1.6 million people in rural areas (mainly Potosí and Oruro), 85% of potato and 50% of quinoa production were lost due to decreasing water supply, and a total loss of USD 400 million was estimated (Mariscal et al., 2011). In 1997–1998, 135,000 people were affected and USD 515 million loss was estimated (UNDP, 2011). More recently, during the 2015–2016 El Niño, 665,000 people were affected and over USD 450 million of losses were experienced (Guha-Sapir et al., 2016). To lessen the long-term impacts of similar events, the Bolivian national government has allocated large budget amounts for emergency operations to compensate part of the losses, which are usually evaluated in an ex-post fashion. Important in this context is that Bolivia has implemented a national program for risk disaster management that includes an early warning system informing about disaster occurrence a few days ahead of the event.

Precipitation Variability

Principal component analysis (PCA) was applied to the summer precipitation (DJFM) and Nino3.4 for seeking their relationship. The monthly precipitation at El Alto, Patacamaya, Oruro, and Potosi (Los Pinos) was obtained from the National Service of Meteorology and Hydrology (SENAMHI) of Bolivia from 1948 to 2016 (Figure 1). And, ENSO, Nino3.4 data were obtained from Rayner et al. (2003) for the region 3.4 for the same time period (see https://www.esrl.noaa.gov/psd/gcos_wgsp/Timeseries/Nino34). El Alto and Oruro datasets had no missing data. Patacamaya and Potosi time series had less than 10% missing data. The data gaps of the time series were replaced with the monthly mean of each time series. The summer precipitation (DJFM) totals for El Alto, Patacamaya, Oruro, and Potosi and the summer average of Nino3.4 from 1948 to 2016 were used in this study. Prior to the PCAs, all seasonal data were standardized to zero mean and unit standard deviation. PCA was developed by Pearson (1901) and Hotelling (1933). It reduces data dimensionality and concentrates variance into a small number of variables. All numerical computations were performed using the MATLAB software.

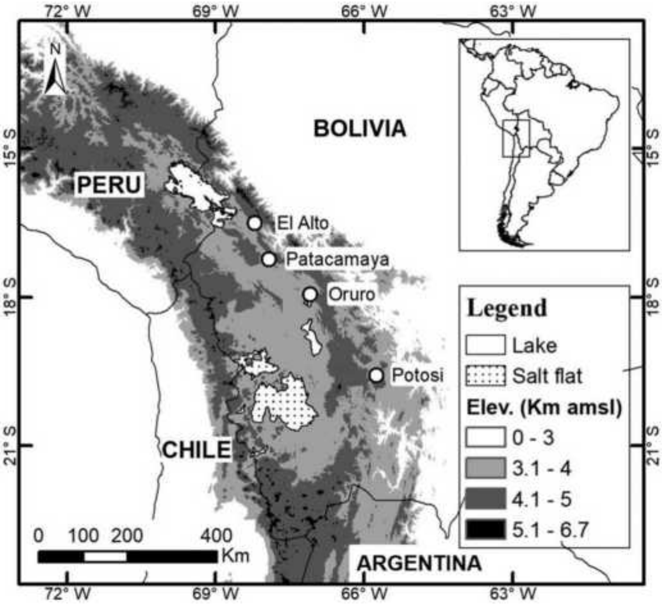


FIGURE 1 The Central Andes of South America.

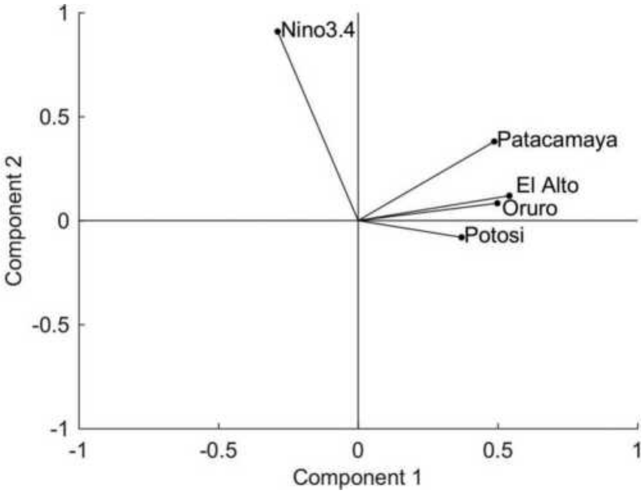


FIGURE 2 Biplot of the first and second modes for the PCA for the summer precipitation and Nino3.4.

The results of the PCA show that the first two modes explain about 70% of the total variance (Figure 2). The first mode explains about 50% of the variance and shows a negative relationship between summer precipitation and ENSO for all studied locations. The correlation coefficients of the first mode are 0.5 for El Alto, Patacamaya, and Oruro; 0.4 for Potosi; and −0.3 for Nino3.4. These results confirm previous findings that there is a negative relationship between precipitation and ENSO (see Canedo-Rosso et al., 2019b; Garreaud et al., 2003; Garreaud and Aceituno, 2001; Thibeault et al., 2012). This means that less rainfall occurs during warm ENSO phases, and this condition could exacerbate the occurrence of drought. The second mode explains about 20% of the total variance and shows a difference for summer precipitation pattern in the studied region. Patacamaya shows a larger variation of precipitation compared to the other locations, with a correlation coefficient of 0.4. The heterogenic topography could be a factor that increases variance of precipitation in the studied region. The Central Andes is divided into two mountain ranges, and between these is the Altiplano high plateau. El Alto, Patacamaya, and Oruro are located on the plateau, and Potosi is located on the eastern range of the Central Andes. This could explain the slight negative relationship between the precipitation for Potosi and the other locations.

Impacts of El Niño – Southern Oscillation on Crop Yield

Precipitation dependency by rainfed agriculture and frost risk limit farming production to the austral summer (Condori et al., 2014; Garcia et al., 2007). Due to the fact that farming is mostly rainfed, weather-related phenomena are very important for crop production. The ENSO impact on crop yield was analyzed using two-sample t-test and Wilcoxon's rank-sum tests for La Paz (including El Alto and Patacamaya), Oruro, and Potosi from 1981 to 2017. The crop yield for potato and quinoa was standardized and filtered by PCA, and the time series were reconstructed keeping about 90% of their original variability to reduce noise. To test the relationship, crop yield during neutral/moderate years was compared with crop yield during strong and very strong El Niño years (warm ENSO phase) (see Table 1). The results showed that quinoa yield during warm ENSO phase and neutral/moderate years presents a significant difference at 95% confidence level except for Oruro. The average yield during neutral/moderate years is higher compared to El Niño years in La Paz and Potosi. The crop yield during El Niño years is lower than the mean in Oruro, except for 1982–1983 and 2015–2016. This contradicts previous findings that reported large agricultural losses during 1982–1983, 1997–1998 (Mariscal et al., 2011; Santos, 2006),

TABLE 1 Classification of ENSO

Strong and [a]Very Strong El Niño	Neutral and Moderate	Strong La Niña
1982–1983[a]	1981–1982	1988–1989
1987–1988	1983–1987	1998–1999
1991–1992	1989–1991	1999–2000
1997–1998[a]	1992–1997	2007–2008
2015–2016[a]	2000–2007	2010–2011
	2008–2010	
	2011–2015	
	2016–2017	

Strong and very strong El Niño (≥ 1.5°C), strong La Niña (≤ –1.5°C), and neutral/moderate (–1.4 to 1.4°C) years for the time period from 1981 to 2017.
[a] Very strong El Niño years. From Null (2018).

and 2015–2016 (Valdez, 2017). On the other hand, the quinoa yield has constantly increased during the last years, mainly in Oruro. This could be explained by employment of advanced crop management strategies (e.g., selected crop varieties and application of agricultural innovations), as Oruro is one of the largest producers of quinoa in Bolivia and the world (Ormachea and Ramirez, 2013).

Despite the quinoa's high tolerance to environmental stress including droughts (Jacobsen et al., 2005; Jacobsen et al., 2003), it generally showed large losses during El Niño events (Figure 3). This may be due to the crop´s sensitivity for water stress during specific stages of the growing season, and the most sensitive stages are the emergence, flowering, and grain development (Geerts et al., 2008, 2009). The risk for crop yield reduction could be reduced here with irrigation during the sensitive phases of the quinoa crop development. A strategy like deficit irrigation could be employed (Geerts et al., 2008; Talebnejad and Sepaskhah, 2015). Another strategy to mitigate the crop yield reduction is implementation of crop varieties more resistant to water stress (Sun et al., 2014).

The t-test and rank-sum test results showed that potato yield during neutral/moderate and El Niño years is significantly different at 95% confidence level except for La Paz. The results showed that production during neutral/moderate years is higher in Oruro and Potosi. All regions showed the lowest potato yield during strong El Niño 1982–1983, with a yield reduction of four standard deviations below the mean. The yield reduction during El Niño events seems to have a larger effect in Oruro. In La Paz, potato yield showed lower vulnerability to El Niño, except during the strong events of 1987–1988 and 2015–2016 (Figure 3). This could be explained by closeness to the Lake Titicaca and other water bodies that might be used as a water source during precipitation deficit. Similar strategies for drought mitigation (e.g., irrigation and resistant crop varieties) could be implemented in order to avoid large crop losses. However, knowing that a very strong El Niño could lead to large agricultural losses, insurance policy could be assigned to farmers in order to manage the risk before the occurrence of a drought event (Canedo-Rosso et al., 2019a). For the implementation of any drought mitigation strategy, identification, evaluation, and monitoring of drought risk are crucial. Especially important are our findings that ENSO must be taken explicitly into account in such considerations.

As mentioned before, ENSO is the most important climate phenomenon that affects the variability of precipitation in the Central Andes (Canedo-Rosso et al., 2019b; Garreaud et al., 2003; Garreaud and Battisti, 1999; Vuille, 1999; Vuille et al., 2000), and it is known that its warm phase is connected to dryer and warmer conditions in the Central Andes. This information could be used for the development of seasonal forecast of precipitation and temperature over Central Andes as ENSO forecasts are developed at major global research and operational centers (e.g., International Research Institute for Climate and Society and Climate Prediction Center). When an El Niño is predicted, it is possible to assume that less

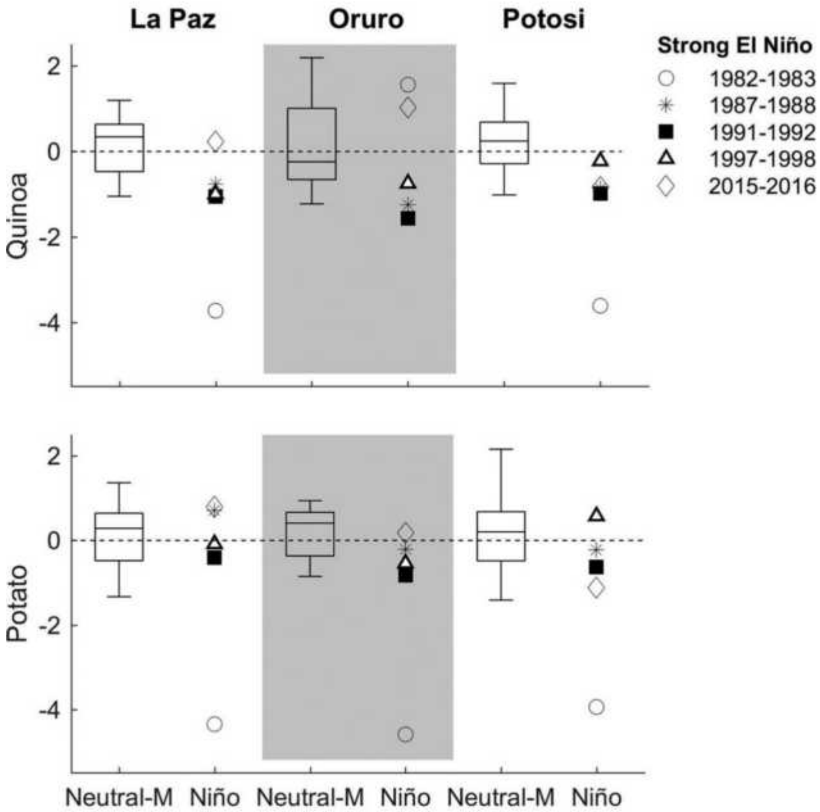


FIGURE 3 Crop yield during neutral/moderate and ENSO years. (Boxplots of quinoa and potato yield anomalies for neutral and moderate (Neutral-M) years for La Paz, Oruro, and Potosi. Crop yield anomalies during strong El Niño years (markers). Based on Canedo-Rosso et al. (2019a).)

precipitation and higher temperature could eventually lead to drought occurrence in the Central Andes during the growing season. Moreover, higher temperatures increase evapotranspiration rates and loss of water content in the soil. For a scenario of high probability of drought occurrence, the application of ex-ante strategies could mitigate its negative effects. Thus, in agriculture, these strategies could reduce crop production losses (Canedo-Rosso et al., 2019a). There are several ex-ante strategies for drought management in agriculture. Examples of these are as follows: use crop or varieties resistant to drought, install irrigation systems, and allocate financial compensation or insurance policy to farmers. For the implementation of any drought mitigation strategy, identification, evaluation, and monitoring of drought risk are needed.

Conclusions

Droughts in the Central Andes are generally driven by ENSO warm phase, and they affect the agricultural production by lowering quinoa and potato yields by decreasing summer precipitation. Regions that presented accessibility to other water supply sources than precipitation or have a better crop management showed lower influence of drought events. It is important to consider the implementation of ex-ante strategies (such as irrigation, agricultural technology, and insurance), in order to decrease the agricultural loss due to climate hazards in the region. Moreover, the evaluation of causes and consequences of droughts could support seasonal forecasting, drought risk reduction, and sustainable water management.

References

Canedo-Rosso, C., Hochrainer-Stigler, S., Pflug, G., Condori, B., and Berndtsson, R. (2019a). Drought risk in the Bolivian Altiplano associated with El Niño southern oscillation using satellite imagery data. *Natural Hazards and Earth System Sciences Discussion, 2019*, 1–21. doi:10.5194/nhess-2018-403

Canedo-Rosso, C., Uvo, C. B., and Berndtsson, R. (2019b). Precipitation variability and its relation to climate anomalies in the Bolivian Altiplano. *International Journal of Climatology, 39*(4), 2096–2107. doi:10.1002/joc.5937

Condori, B., Hijmans, R., Ledent, J., and Quiroz, R. (2014). Managing potato biodiversity to cope with frost risk in the High Andes: a modeling perspective. *PLoS One, 9*(1), e81510. doi:10.1371/journal.pone.0081510

Garcia, M. (2003). *Agroclimatic Study and Drought Resistance Analysis of Quinoa for an Irrigation Strategy in the Bolivian Altiplano*. Leuven: Katholieke Universiteit Leuven.

Garcia, M., Raes, D., and Jacobsen, S.-E. (2003). Evapotranspiration analysis and irrigation requirements of quinoa (Chenopodium quinoa) in the Bolivian highlands. *Agricultural Water Management, 60*(2), 119–134. doi:10.1016/S0378-3774(02)00162-2

Garcia, M., Raes, D., Jacobsen, S.-E., and Michel, T. (2007). Agroclimatic constraints for rainfed agriculture in the Bolivian Altiplano. *Journal of Arid Environments, 71*(1), 109–121. doi:10.1016/j.jaridenv.2007.02.005

Garreaud, R. D., and Aceituno, P. (2001). Interannual rainfall variability over the South American Altiplano. *Journal of Climate, 14*(12), 2779–2789. doi:10.1175/1520-0442(2001)014<2779:Irvots>2.0.Co;2

Garreaud, R. D., and Battisti, D. S. (1999). Interannual (ENSO) and interdecadal (ENSO-like) variability in the Southern Hemisphere tropospheric circulation. *Climate, 12*(7), 2113–2123. doi:10.1175/1520-0442(1999)012<2113:Ieaiel>2.0.Co;2

Garreaud, R. D., Vuille, M., and Clement, A. C. (2003). The climate of the Altiplano: observed current conditions and mechanisms of past changes. *Palaeogeography, Palaeoclimatology, Palaeoecology, 194*(1–3), 5–22. doi:10.1016/S00310-0182(03)00269-4

Geerts, S., Raes, D., Garcia, M., Miranda, R., Cusicanqui, J. A., Taboada, C., Mendoza, J., Huanca, R., Mamani, A., Condori, O., Mamani, J., Morales, B., Osco, V., and Steduto, P. (2009). Simulating yield response of quinoa to water availability with aquacrop. *Agronomy Journal, 101*(3), 499–508. doi:10.2134/agronj2008.0137s

Geerts, S., Raes, D., Garcia, M., Vacher, J., Mamani, R., Mendoza, J., Huanca, R., Morales, B., Miranda, R., Cusicanqui, J., and Taboada, C. (2008). Introducing deficit irrigation to stabilize yields of quinoa (Chenopodium quinoa Willd.). *European Journal of Agronomy, 28*(3), 427–436. doi:10.1016/j.eja.2007.11.008

Guha-Sapir, D., Hoyois, P., Wallemacq, P., and Below, R. (2016). *Annual Disaster Statistical Review 2016: The numbers and trends*. Retrieved from Brussels: emdat.be/sites/default/files/adsr_2016.pdf

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology, 24*(6), 417–441. doi:10.1037/h0071325

IPCC. (2013). *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and G. F. Midgley Eds.). Cambridge and New York: Cambridge University Press.

Jacobsen, S. E., Monteros, C., Christiansen, J. L., Bravo, L. A., Corcuera, L. J., and Mujica, A. (2005). Plant responses of quinoa (Chenopodium quinoa Willd.) to frost at various phenological stages. *European Journal of Agronomy, 22*(2), 131–139. doi:10.1016/j.eja.2004.01.003

Jacobsen, S. E., Mujica, A., and Jensen, C. R. (2003). The resistance of quinoa (Chenopodium quinoa Willd.) to adverse abiotic factors. *Food Reviews International, 19*(1–2), 99–109. doi:10.1081/FRI-120018872

Lenters, J. D., and Cook, K. H. (1999). Summertime precipitation variability over South America: role of the large-scale circulation. *Monthly Weather Review, 127*(3), 409–431. doi:10.1175/1520- 0493(1999)127<0409:Spvosa>2.0.Co;2

Mariscal, C. B., Tassi, N., Miranda, A. R., Canedo, L. A., and Cazorla, I. (2011). Rural migration in Bolivia: the impact of climate change, economic crisis and state policy. In IIED (Ed.), *Human Settlements Working Paper Series: Rural-Urban Interactions and Livelihood Strategies – 31* (pp. 1–49). London: International Institute for Environment and Development (IIED).

Null, J. (2018). *El Niño and La Niña Years and Intensities*. Retrieved from https://ggweather.com/enso/oni.htm

Ormachea, E., and Ramirez, N. (2013). *Propiedad colectiva de la tierra y producción agrícola capitalista: El caso de la quinua en el Altiplano Sur de Bolivia* (first ed.). Bolivia: Centro de Estudios para el Desarrollo Laboral y Agrario (CEDLA).

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine, 2*(11), 559–572. doi:10.1080/14786440109462720

Rayner, N. A., Parker, D. E., Horton, E. B., Folland, C. K., Alexander, L. V., Rowell, D. P., Kent, E. C., and Kaplan, A. (2003). Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *Journal of Geophysical Research: Atmospheres, 108*(D14). doi:10.1029/2002JD002670

Santos, J. L. (2006). The impact of El Niño – Southern oscillation events on South America. *Advances in Geosciences, 6*, 221–225. doi:10.5194/adgeo-6-221-2006

Sun, Y., Liu, F., Bendevis, M., Shabala, S., and Jacobsen, S.-E. (2014). Sensitivity of two quinoa (Chenopodium quinoa Willd.) varieties to progressive drought stress. *Journal of Agronomy and Crop Science, 200*(1), 12–23. doi:10.1111/jac.12042

Talebnejad, R., and Sepaskhah, A. R. (2015). Effect of deficit irrigation and different saline groundwater depths on yield and water productivity of quinoa. *Agricultural Water Management, 159*, 225–238. doi:10.1016/j.agwat.2015.06.005

Thibeault, J., Seth, A., and Wang, G. L. (2012). Mechanisms of summertime precipitation variability in the Bolivian Altiplano: present and future. *International Journal of Climatology, 32*(13), 2033–2041. doi:10.1002/joc.2424

Thompson, L. G., Mosley-Thompson, E., and Arnao, B. M. (1984). El Niño-southern oscillation events recorded in the stratigraphy of the tropical Quelccaya ice cap, Peru. *Science, 226*(4670), 50–53. doi:10.1126/science.226.4670.50

UN. (2016). *The Sustainable Development Goals Report 2016*. New York: United Nations.

UNDP. (2011). *Tras las huellas del cambio climático en Bolivia. Estado del arte del conocimiento sobre adaptación al cambio climático: agua y seguridad alimentaria*. La Paz: UNDP-Bolivia.

UNISDR. (2015). *Making Development Sustainable: The Future of Disaster Risk Management. Global Assessment Report on Disaster Risk Reduction*. Retrieved from Geneva, Switzerland.

Valdez, C. (2017). Severe Bolivian drought hurts crops, threatens capital. Phys.org.

Vicente-Serrano, S. M., Chura, O., López-Moreno, J. I., Azorin-Molina, C., Sanchez-Lorenzo, A., Aguilar, E., Moran-Tejeda, E., Trujillo, F., Martínez, R., and Nieto, J. J. (2015). Spatio-temporal variability of droughts in Bolivia: 1955–2012. *International Journal of Climatology, 35*(10), 3024–3040. doi:10.1002/joc.4190

Vuille, M. (1999). Atmospheric circulation over the Bolivian Altiplano during dry and wet periods and extreme phases of the southern oscillation. *International Journal of Climatology, 19*(14), 1579–1600. doi:10.1002/(SICI)1097-0088(19991130)19:14<1579::AID-JOC441>3.0.CO;2-N

Vuille, M., Bradley, R. S., and Keimig, F. (2000). Interannual climate variability in the Central Andes and its relation to tropical Pacific and Atlantic forcing. *Journal of Geophysical Research: Atmospheres, 105*(D10), 12447–12460. doi:10.1029/2000JD900134

Wilhite, D. A., and Glantz, M. H. (1985). Understanding: the drought phenomenon: the role of definitions. In D. A. Wilhite, W. E. Easterling, and D. A. Wood (Eds.), *Planning for Drought: Toward a Reduction of Social Vulnerability* (pp. 10–30). Boulder, CO: Westview Press.

31

Mines: Rehabilitation of Open Cut

Douglas J. Dollhopf

Introduction

Mining directly disturbs approximately 240,000km^2 of the Earth's surface.[1] Surface mining methods may be classified as: 1) open pit mining; 2) strip mining; 3) dredging; and 4) hydraulic mining.[2] Open pit mining includes quarries used to produce limestone, sandstone, marble, and granite; pits used to produce sand, gravel, and bentonite; and large excavations used to produce talc, copper, gold, iron, silver, and other metals. This mining method is distinguished by having one large pit or numerous small pits across the landscape. State or federal governments establish regulations that require the operator to restore the land to an approved land use(s) that is equal to or better than the premine land use. The approved postmine land use(s) is instrumental in determining the final graded topography and vegetation community.

Landscape Regrading

Reclamation of an open pit mine must address the open pit itself, waste rock removed from the pit to gain access to the ore, tailing impoundments, and access roads (Figure 1). For open pit mines, complete backfilling of the pit is not usually economically feasible. Because the pit remaining after mineral extraction may be hundreds of feet deep, the cost of moving waste rock back into the depression is prohibitive. Surface and groundwater may flow into the depression after mining is terminated and an impoundment develops. Pit slopes should be reduced to the limits required for safe access for humans, livestock, and wildlife. Pit slopes should not be permitted at gradients that jeopardize the success of postmine reclamation.

FIGURE 1 Large open pit mine in the U.S. with water impounded in bottom.

Coversoil Resources

Successful plant community establishment on an open pit mine site is a direct function of the coversoil quality applied. Coversoil suitability criteria vary in the US from state to state, but are similar to those presented in Table 1. The coversoil resource emanates from two procedures: 1) salvaging the natural soil resource from the area to be disturbed and 2) mining unconsolidated geologic stratum when the soil resource is absent. Under present US federal and state mine regulatory programs, all portions of the soil resource shall be salvaged in a project area if it meets physicochemical suitability criteria. Soil materials that do not meet these suitability criteria are generally not used since they may impair plant establishment and growth. Once the soil resource is salvaged, it should be directly hauled to an area that has been backfilled and is ready to receive coversoil. Direct hauled soil contains a viable seed bank, mycorrhizal associations, organic matter, and nutrients that aid in plant establishment. Conversely, if the soil resource is stockpiled, these resources will deteriorate with time.

Prior to implementation of mine land reclamation regulations during the 1970s and 1980s, open pit US mine operations often did not institute land reclamation. The soil resource was not salvaged, the disturbed landscape was not graded to the approximate original contour, and the site was not seeded. Consequently, the soil resource was lost. Today, these lands are being coversoiled using unconsolidated geologic stratum. Valley fill areas adjacent to the disturbed landscape often contain alluvial materials beneath the soil resource that meet all soil suitability criteria (Table 1), except organic matter is absent. This alluvial stratigraphy may be 10–30 m thick or more. Earth moving equipment is used to stockpile

TABLE 1 Coversoil Suitability Criteria

Soil Parameter	Suitability Criteria
pH	>6.5 and <8.0 standard units
Electrical conductivity	<4.0 dS/cm
Exchangeable sodium percentage	<12%
USDA textural classes (12 types)	All suitable except clay, loamy sand, sand
Rock content (particles >2mm diameter)	
Slope gradients <25%	<35%, weight basis
Slope gradients >25%	>35 and <60%, weight basis
As, Cd, Cu, Pb, Zn, other metals	Near back ground levels, i.e., no enrichment
Organic matter	>0.5%, weight basis

the soil resource, separating true topsoil (A-horizon) from subsoil (B- and C-horizons), to enable excavation of deeper alluvial material. Following placement of this alluvial coversoil on the disturbed landscape, the pit created is contoured and the stockpiled soil resource is replaced. Lands receiving coversoil emanating from a geologic stratum should have organic matter applied to expedite establishment of nutrient cycles in the plant root zone. Cattle manure and municipal compost are two sources of organic matter used in land reclamation projects.

Coversoil Thickness Requirements

The thickness of coversoil required for maximum vegetation performance is dependent on several factors including the 1) vegetation species present; 2) local climate; 3) coversoil quality; and 4) physicochemical quality of the substrate beneath the coversoil. The maximum rooting depth of many rangeland ecosystem plant species is less than 45 cm and generally does not exceed 100 cm.[3] Investigators found that a minimum of 40 cm to a maximum of 150 cm of coversoil is needed for optimum plant growth.[4–8] This wide variation of findings is a function of site specific conditions identified before. The Barth[4] investigation is representative of most research. This investigator found that generic spoil, defined as nonalkaline and nonsaline loams with sodium adsorption ratio (SAR) below seven, showed an increase in plant production up to a coversoil depth of approximately 50 cm. Sodic spoil, or material with SAR of 26 or higher, required a minimum of 70 cm of coversoil to reach maximum cool season grass production. Coversoiled acid spoils, defined as spoils exhibiting pH values between 3.6 and 4.3, revealed increased grass production up to the maximum coversoil depth used in this study (152 cm).

In Situ Soil Reclamation

In situ soil reclamation at open cut mine sites means no coversoil resource is used. Chemical and/or organic amendments are applied to the graded spoil, tailings, or waste rock landscape to enable plant establishment and growth.

In Situ Sodic Minesoil Remediation

Thousands of hectares of land in Wyoming and Montana were open pit mined for the mineral bentonite prior to passage of state regulations requiring land reclamation. The absence of a soil resource to reclaim these lands means the plant ecosystem must be established on graded spoils. Spoils are overburden materials cast aside to access the underlying mineral, bentonite. These overburden materials have a sodic condition (ESP 20–40), clay texture (60% clay) dominated by swelling clay minerals such as smectite, and preclude plant establishment and growth. In situ soil remediation requires a two-fold approach: 1) permanently reduce exchangeable sodium percentage (ESP) to less than 10 within the 0–30 cm soil profile and 2) provide an organic amendment to immediately prevent soil crust development and increase water infiltration. Crust development in these clayey-sodic soil systems is precluded with applications of wood chips from saw mill waste or manure.[9] Applications of gypsum, calcium chloride, magnesium chloride, and sulfuric acid to these sodic soils have been shown to be effective at reducing the ESP to acceptable levels in minesoils and provide a root zone suitable for plant growth.[10]

In Situ Acid Minesoil Remediation

Open pit mines associated with gold, silver, and metal extraction from sulfide ore bodies frequently produce tailings impoundments and waste rock areas that are acidic (pH 2.5–5.5) and fail to support plant growth. Soil acidity is formed when sulfide minerals, such as pyrite, are exposed to oxidizing conditions in the presence of water. In the absence of a coversoil resource, tailings and waste rock material are treated with alkaline amendments to permanently neutralize the soil acidity. Soil analysis of the acid base account and active acidity enable determination of the calcium carbonate requirement.[11] Amendments used include calcium carbonate, calcium hydroxide, and calcium oxide. In situ treatment with amendments should be at least to the 45 cm soil depth and special engineered plowing equipment is required to attain this depth of incorporation.[12]

Acid Mine Drainage

Upon exposure of open pit mine wastes to water and oxygen, sulfide minerals oxidize to form acidic, sulfate rich drainage.[13] Iron-oxidizing bacteria, e.g., *Thiobacillus ferrooxidans,* expedite the acid producing reaction rate up to a million times. Metal composition (Al, Cu, Fe, Mn, Pb, Zn, and others) and concentrations in acid mine drainage (AMD) depend on the type and quantity of sulfide minerals present. Approaches to prevent or treat AMD include: 1) removal of air and water from the sulfide body that is disturbed; 2) wetland construction to precipitate contaminants with oxidation ponds and anaerobic-microbial reactions in plant root zones; 3) underground anoxic-and above ground open-limestone drains to raise water pH; and 4) treatment of drainage with neutralizing chemicals such as hydrated lime, quicklime, soda ash, caustic soda, and ammonia.[14]

Steep Slope Reclamation

Open pit mines are frequently located in mountainous terrain and reconstructed slopes may have gradients as steep as 50%. Sediment yields tend to increase linearly with slope gradient. There is an inverse relationship between soil loss on slopes and rock cover.[15] High rock content in soils increases infiltration rate and surface roughness decreasing runoff and soil loss. As erosion occurs on rocky soil, rock cover increases as the coarse fragments below the surface are exposed. This armoring of the soil surface can help reduce soil erosion. It has been reported that plant growth may be impaired when soil rock content exceeds 35%.[16] High rock contents of 35%–60% applied on steep slopes (Table 1) may impair plant growth, but are considered a best management practice to facilitate plant establishment and slope stability.

Construction of shallow pits across steep slopes is commonly practiced to increase water storage on the slope and minimize runoff. Pitting techniques using equipment referred to as Dammer–Diker, gouger, and dozer basin blade have been shown to be effective in controlling sediment loss.[17]

Conclusions

During these past 30 years, land rehabilitation sciences at open pit mines developed at an exponential rate. Grading the mined landscape to the approximate original contour, use of adequate amounts of—and quality of—coversoil, and development of in situ soil treatments at sites where coversoil resources were absent have all provided means to establish a diverse plant community. Techniques have been developed to revegetate and stabilize slopes approaching a 50% gradient, which is an asset for mines located in mountainous terrain. Methods to control and treat AMD from open pit mines continue to improve, but this impact to surface water resources remains unresolved. Future research is required to develop better AMD treatment methods.

References

1. Solomons, W. *Mining impacts worldwide.* Proceedings of the International Hydrology Program; UNESCO, Asian Institute of Technology: Bangkok, 1988; 7–9.
2. Paone, J.; Struthers, P.; Johnson, W. Extent of disturbed lands and major reclamation problems in the United States. In *Reclamation of Drastically Disturbed Lands;* Schaller, F.W., Sutton, P., Eds.; American Society of Agronomy: Madison, WI, 1978; 11–22.

3. Wyatt, J.W.; Dollhopf, D.J.; Schafer, W.M. Root distribution in 1-to 48-year old strip-mine spoils in southeastern Montana. J. Range Mgmt **1980**, *33*, 101–104.
4. Barth, R.C. *Soil-depth Requirements to Reestablish Perennial Grasses on Surface-Mined Areas in the Northern Great Plains;* No. 1, Publication 0192–6179/84/2701-0001; Mineral and Energy Resources Research Institute, Colorado School of Mines: Golden, 1984; Vol. 27.
5. Doll, E.C.; Merrill, S.D.; Halvorson, G.A. *Soil Replacement for Reclamation of Strip Mined Lands in North Dakota;* Agricultural Experiment Statement Bulletin 514; North Dakota State University: Fargo, 1984; 23 pp.
6. Halvorson, G.A.; Melsted, S.W.; Schroeder, S.A.; Smith, C.M.; Pole, M.W. Topsoil and subsoil thickness requirements for reclamation of nonsodic mined-land. Soil Sci. Soc. Am. J. **1986**, *50,* 419–422.
7. Power, J.F.; Sandoval, F.M.; Ries, R.E. Topsoil–subsoil requirements to restore North Dakota mined lands to original productivity. Mining Engng 1979, December, 23–27.
8. Schuman, G.E.; Taylor, E.M.; Pinchak, B.A. Revegetation of mined land: influence of topsoil depth and mulching method. J. Soil Water Conserv. **1985**, *40,* 249–252.
9. Schuman, G.E.; King, L.A.; Smith, J.A. Reclamation of bentonite mined lands. In *Reclamation of Drastically Disturbed Lands;* Barnhisel, R.I., Darmody, R.G., Daniels, W.L., Eds.; American Society of Agronomy: Madison, WI, 2000; Vol. 41, 687–707.
10. Dollhopf, D.J.; Rennick, R.B.; Smith, S.C. *Long-Term Effects of Physicochemical Amendments on Plant Performance at a Bentonite Mine Site in the Northern Great Plains*; Reclamation Research Unit Publication 88-02, Montana State University: Bozeman, 1988; 126 pp.
11. Sobek, A.A.; Schuller, J.R.; Freeman, J.R.; Smith, R.M. *Field and Laboratory Methods Applicable to Overburdens and Minesoils*; Publication 600/2–78-054; U.S. Environmental Protection Agency, Office of Research and Development: Cincinnati, Ohio, 1978; 47–67.
12. Dollhopf, D.J. *Deep Lime Incorporation Methods for Neutralization of Acidic Minesoils;* Reclamation Research Unit Publication 9201; Montana State University: Bozeman, 1992; 94.
13. Stumm, W.; Morgan, J.J. *Aquatic Chemistry;* Wiley: New York, 1970; 178 pp.
14. Skousen, J.G.; Sexstone, A.; Ziemkiewicz, P.F. Acid mine drainage control and treatment. In *Reclamation of Drastically Disturbed Lands;* Barnhisel, R.I., Darmody, R.G., Daniels, W.L., Eds.; American Society of Agronomy: Madison, WI, 2000; Vol. 41, 131–168.
15. Kapolka, N.M.; Dollhopf, D.J. Effect of slope gradient and plant growth on soil loss on reconstructed steep slopes. Int. J. Surf. Mining, Reclam. Environ. **2001**, *15* (2), 86–89.
16. Munn, L.; Harrington, N.; McGirr, D.R. Rock fragments. In *Reclaiming Mine Soils and Overburden in the Western United States; Analytical Parameters and Procedures;* Williams, R.D., Schuman, G.E., Eds.; Soil Conservation Society of America: Ankeny Iowa, 1987; 259–282.
17. Larson, J.E. *Revegetation Equipment Catalog;* Stock No. 001–00518-5 Forest Service; U.S. Department of Agriculture, U.S. Government Printing Office: Washington, DC, 1980.

IV

DIA: Diagnostic Tools: Monitoring, Ecological Modeling, Ecological Indicators, and Ecological Services

32

Bioenergy Crops: Carbon Balance Assessment

Rocky Lemus and
Rattan Lal

Overview

Global production and use of bioethanol, derived from corn and sugarcane, doubled between 2007 and 2016.[1] Annual bioethanol production was 103 billion liter in 2016 and increased to 106 billion liter in 2017.[2] Several herbaceous species have been assessed as potential energy sources and for carbon (C) sequestration since the 1990s. Switchgrass (*Panicum virgatum* L.), a herbaceous species, and two short-rotation woody crops (SRWCs), hybrid poplar (*Populus* spp.) and willow (*Salix* spp.), are promising for C sequestration and biofuel production. Using these species as energy crops can partly offset CO_2 emitted by fossil fuel combustion. However, biomass yield of these promising species depends on environmental differences across years and across sites along with differences due to application of nitrogenous fertilizers.[3]

There are approximately 60 million ha (Mha) of degraded soil (severely eroded and mined land) in the United States that can be sown to bioenergy crops to decrease soil erosion and increase C sequestration.[4] The Midwest and the Southeast regions of the United States are the potential areas where bioenergy crops are most likely to compete with other traditional crops for land resources.[5] The use of bioenergy crops in these areas offers an opportunity to replenish the soil organic C (SOC) pool depleted by tillage and soil degradation.

Biomass fuels used in a sustainable manner can result in no net increase in atmospheric CO_2 and thus are C neutral. Indeed, sustainable use of biomass can result in a net decrease in the rate of enrichment of atmospheric CO_2. This is based on the assumption that all the CO_2 emitted by the use of biomass fuels is absorbed from the atmosphere by photosynthesis. However, the water and C footprint of bioethanol may vary widely. For example, Brazilian bioethanol from sugarcane has a better energy balance (17.7 MJ/L) and lower C footprint (38.5 gCO_{2e}/MJ) than corn-based ethanol in the United States (11.2 MJ/L and 44.9 gCO_{2e}/MJ, respectively).[1] Externalities associated with life-cycle emissions from corn-based ethanol must be addressed.[6,7] Increased substitution of fossil fuels with biomass-based fuels reduces the risk of global warming by enrichment of atmospheric CO_2. This entry discusses the importance of energy crops in offsetting fossil fuel combustion through C sequestration in biomass and soil.

Influence of Bioenergy Crops at the Temporal Scale

Long-term tillage and continuous cropping can reduce SOC pool[8,9] and increase atmospheric concentration of CO_2 (Figure 1).[10] Conversion of natural to agricultural ecosystems results in the net release of C into the atmosphere.[11] Conservation practices, such as the introduction of perennial crops on degraded agricultural soils and growing bioenergy crops, constitute a direct link between sink (SOC) and the source of CO_2 (fossil fuel). Bioenergy crops are the sink/source transition because the C incorporated into their biomass and root system has a high potential of being incorporated into the SOM pool. Most of the C is incorporated into the biomass and soil by the deep and extensive root systems. The SOC pool attains a new equilibrium after grassland restoration, but the time required for SOC to reach equilibrium is variable, especially under diverse climatic conditions.[12,13,14] Most models used to estimate C sequestration in bioenergy crops suggest that a period longer than 10 years may be necessary for pronounced soil-quality improvements.

Benefits of Bioenergy Crops

The SOC pool and its dynamics are the major indicators of soil quality and crop productivity.[13,15] Bioenergy crops can improve soil quality through increase in SOM, nutrient dynamics, erosion control, and improvement in soil structure and porosity. Most perennial crops and SRWCs have extensive root systems that reduce soil erosion and nonpoint-source pollution.

A shift from traditional agricultural crops into perennial bioenergy crops stabilizes agricultural soils, reduces erosion, and improves water quality.[14,16,17] Most of these benefits are due to the elimination of tillage leading to a significant decrease in both erosion and chemical runoff especially nitrate (NO_3–N). Soil erosion and nitrate runoff are reduced in 2-year-old stands of switchgrass when compared to no-till corn (Table 1). Conservation effectiveness of switchgrass is attributed to high root biomass, which stabilizes the soil,[18] depletes excess N and P,[19] and increases microbial activity.[20]

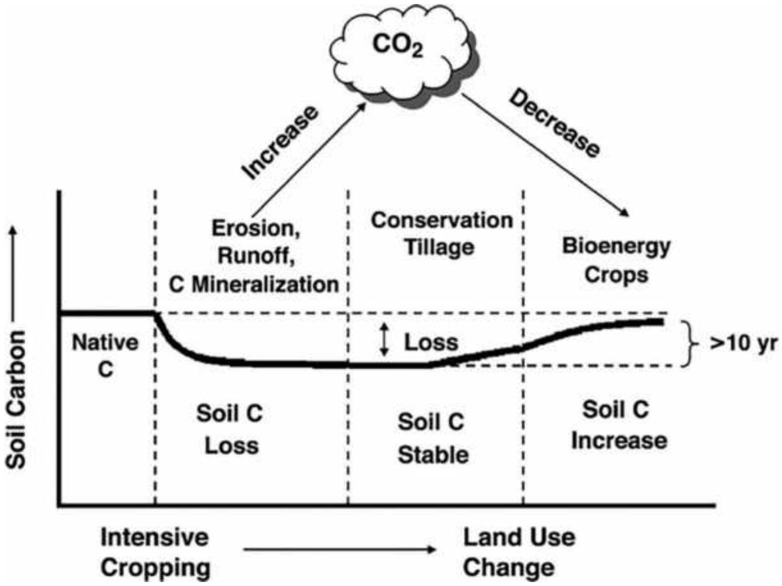


FIGURE 1 Influence of bioenergy crops on C sequestration and CO_2 mitigation with changes in agricultural practices.
(**Source:** Adapted from Janzen et al.[10])

TABLE 1 Environmental Benefits of Switchgrass (SWG) Compared to No-Till Corn (NTC) over Time

	Soil Erosion			N Loss in Water Runoff			Nitrate Concentration		
Year	SWG ($Mg\ ha^{-1}$)	NTC ($Mg\ ha^{-1}$)	Ratio SWG/NTC[a]	SWG ($Kg\ ha^{-1}$)	NTC ($Mg\ ha^{-1}$)	Ratio SWG/NTC	SWG (ppm)	NTC (ppm)	Ratio SWG/NTC
1	2.80	0.70	4.00	10.70	2.60	4.11	3.41	0.57	5.98
2	0.14	0.19	0.74	0.70	1.40	0.50	0.72	2.18	0.33
3	0.06	0.08	0.75	0.30	0.90	0.33	0.77	0.90	0.86

Source: Adapted from McLaughlin et al.[17]

[a] The ratio SWG/NTC of <1 indicates conservation effectiveness of a SWG system during the second and third years of its establishment.

Potential of C Sequestration

Bioenergy crops have the potential to store large quantities of C. Exploring their biomass potential through N fertilization and soil management are some of the proposed strategies to offset CO_2 emissions by fossil fuel combustion. However, a bioenergy crop is not a closed system and only some portion of the C sequestered might be conserved through the production-utilization cycle. Most of the C sequestered in the biomass is utilized for energy production allowing some C release back into the atmosphere. This system does not cause negative impacts as most of the C returned to the atmosphere is recaptured by the plants during the subsequent season.

Assessing the net C sequestration of bioenergy compared to agricultural crops requires analyses of C sequestration across the soil profile. Carbon sequestration is influenced by biological processes such as root biomass and crop species, and by soil physical and chemical aspects such as texture, bulk density, and pH. The amount of C in these types of management options is affected by soil perturbation which could increase SOM oxidation, exacerbating losses, especially at the soil surface. Estimations of C sequestration on land under the Conservation Reserve Program (CRP) have ranged from 0.6 to 1.0 Mg C ha^{-1} yr^{-1} in SOM,[21] compared to 2.0 Mg C ha^{-1} yr^{-1} in SRWC.[22] Root biomass and its relationship to SOC are always positive but may be influenced by soil depth and plant species.[23] Switchgrass improves soil quality via reduced nutrient loss (especially NO_3–N) and increases C sequestration through its deep rooting system, high biomass production, and perenniality.[24] Switchgrass root biomass can be as high as 16.8 Mg ha^{-1} up to 3.3 m deep with some root mass variability among soil type.[20]

The sustainability of soil and crop systems is typically affected by changes in the SOC pool. The largest gain in SOC pool occurs in the upper 30-cm layer. There are differences in the amount of C being sequestered by herbaceous crops and SRWCs, but they have the potential to increase C both in the soil and aboveground biomass. The data in Figure 2 show that after 5 years, SOC pool at 60-cm depth was 97.5 Mg ha^{-1} under willow, 80.5 Mg ha^{-1} under switchgrass, and 78.5 Mg ha^{-1} under corn. The mean rate of SOC sequestration, vis-à-vis corn, over 5 years was 3.8 Mg C ha^{-1} yr^{-1} under willow and 0.4 Mg C ha^{-1} yr^{-1} under switchgrass (Figure 2). The SOC fluctuations are more drastic in the top soil layers, probably because of the greater effect of precipitation, soil temperature, larger root biomass, and greater microbial activity. Long-term experiments on switchgrass in the United States indicate that it is productive and sustainable on rainfed and marginal lands east of the 100th meridian and stands can last up to 10 years with N input of 10–12 kg/ha.[26]

Conclusions

Bioenergy crops can influence the global C cycle. An important challenge in using bioenergy crops is assessing how much C is being sequestered and how much is being put back into the atmosphere by the cofiring process to produce electricity or other sources of energy. Available data indicate that perennial crops are moving in the right direction in improving soil quality while providing an alternative energy

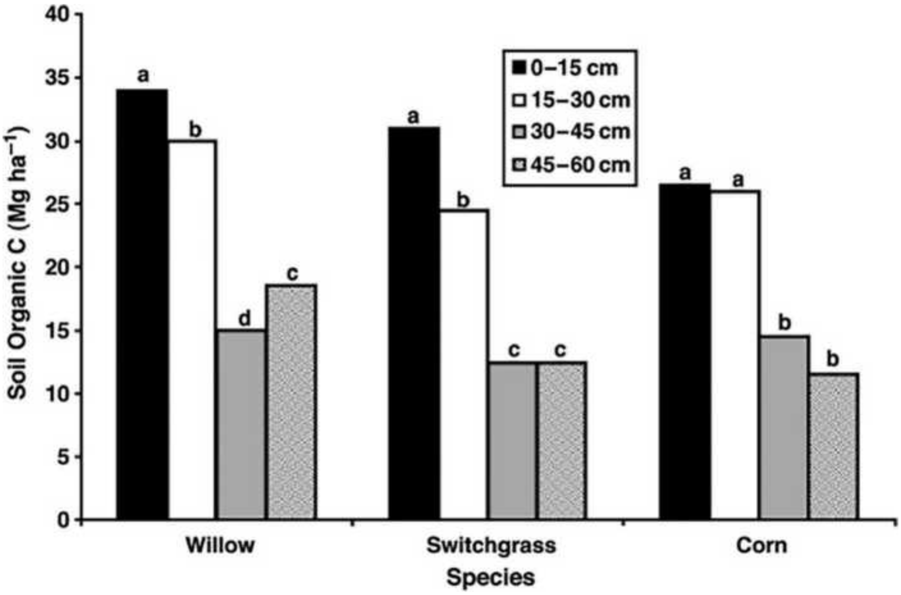


FIGURE 2 Comparison of C sequestration between bioenergy crops and corn at different soil depths 5 years after establishment.
(**Source:** Mehdi et al.[25])

source. A combination of soil C sequestration and bioenergy crops with high biomass capability and deep perennial root systems is a useful strategy of reducing the rate of enrichment of atmospheric CO_2. The use of bioenergy crops increases soil C and improves soil quality by eliminating C losses associated with annual cultivation. However, these improvements depend on the rate of soil C additions, the long-term capacity of the soil for C storage, and the stability or permanence of the sequestered C overtime.

References

1. Mekonnen, M.M.; Romanelli, T.L.; Ray, C.; Koekstra, A.Y.; Liska, A.J.; Neale, C.M.U. Water energy and carbon footprints of bioethanol from the U.S. and Brazil. *Environ. Sci. Technol.* **2018**, *52*, 14508–14518.
2. Renewables. Global Status Report. RNE21, Renewable Energy Policy Network for the 21st Century. Paris, **2018**. (ISBN 978-3-9818911-3-3), 324 pp.
3. Lee, D.K.; Aberle, E.; Anderson, E.K.; Anderson, W. Biomass Production of herbaceous energy crops in the United States: field trial results and yield potential maps from the multiyear regional feedstock partnership. *GCB Biol.* **2018**. doi:10.1111/gcbb.12493.
4. Kort, J.; Collins, M.; Ditsch, D. A review of soil erosion potential associated with biomass crops. *Biomass Bioenergy* **1998**, *14* (4), 351–359.
5. Walsh, M.E.; Ince, P.J.; De la Torre Ugarte, D.G.; Alig, R.; Mills, J.; Spelter, H.; Skog, K.; Slinsky, S.P.; Ray, D.E.; Graham, R.L. Potential of short-rotation wood crops as a fiber and energy source in the U.S. In *Proceedings of the Fourth Biomass Conference of the Americas*; Elsevier Science Ltd.: Oxford, 1999; 195–206.
6. Hoekman, S.K.; Broch, A.; Liu, X. Environmental implications of higher ethanol production and use in the U.S.: A literature review. Part I: Impacts on water soil and air quality. *Renew. Sust. Energy Rev.* **2018**, *81*, 3140–3158.

7. Hoekman, S.K.; Broch, A.; Liu, X. Environmental implications of higher ethanol production and use in the U.S.: A literature review. Part II: Biodiversity, land use change, GHG emissions and sustainability. *Renew. Sust. Energy Rev.* **2018**, *81*, 3159–3177.
8. Kern, J.S.; Johnson, M.G. Conservation tillage impacts on national soil and atmospheric carbon levels. *Soil Sci. Soc. Am. J.* **1993**, *57*, 200–210.
9. Schlamadinger, B.; Marland, G. The role of forest and bioenergy strategies in the global carbon cycle. *Biomass Bioenergy* **1996**, *10,* 275–300.
10. Janzen, H.H.; Campell, C.A.; Izaurralde, R.C.; Ellert, B.H.; Juma, N.; McGill, W.B.; Zenter, R.P. Management effects on soil C storage on the Canadian prairies. *Soil Tillage Res.* **1998**, *47*, 181–195.
11. Schulze, E.D.; Wirth, C.; Heimann, M. Managing forests after Kyoto. *Science* **2000**, *289,* 2058–2059.
12. Potter, K.N.; Tobert, H.A.; Johnson, H.B.; Tischler, C.R. Carbon storage after long-term grass establishment on degraded soils. *Soil Sci.* **1999**, *164,* 718–725.
13. Lal, R. Soil carbon sequestration impacts on global climate change and food security. *Science* **2004**, *304*, 1623–1627.
14. Lal, R. Digging Deeper: A wholistic perspective of factors affecting SOC sequestration. *Global Change Biol.* **2018**, *24* (8): doi:10.1111/gcb.14054.
15. Mann, L.; Tolbert, V. Soil sustainability in renewable biomass plantings. *Ambio* **2000**, *29* (8), 492–498.
16. McLaughlin, S.B.; Walsh, M.E. Evaluating environmental consequences of producing herbaceous crops for bioenergy. *Biomass Bioenergy* **1998**, *14* (4), 317–324.
17. McLaughlin, S.B.; De La Torre Ugarte, D.G.; Garten, C.T.; Lynd, L.R.; Sanderson, M.A.; Tolbert, V.R.; Wolf, D.D. High-value renewable energy from prairie grasses. *Environ. Sci. Technol.* **2002**, *36*, 2122–2129.
18. Gilley, J.E.; Eghball, B.; Kramer, L.A.; Moorman, T.B. Narrow grass hedge effects on runoff and soil loss. *J. Soil Water Conserv.* **2000**, *55*, 190–196.
19. Eghball, B.; Gilley, J.E.; Kramer, L.A.; Moorman, T.B. Narrow grass hedge effects on phosphorus and nitrogen in runoff following manure and fertilizer application. *J. Soil Water Conserv.* **2000**, *55*, 172–176.
20. Ma, Z.; Wood, C.W.; Bransby, D.I. Soil management on soil C sequestration by switchgrass. *Biomass Bioenergy* **2000**, *18,* 469–477.
21. Follet, R.F. Soil management concepts and carbon sequestration in zin cropland soils. *Soil Tillage Res.* **2001**, *61,* 77–92.
22. Tuskan, G.A. Short-rotation woody crop supply systems in the United States: What do we know and what do we need to know?. *Biomass Bioenergy* **1998**, *14* (4), 307–315.
23. Slobodian, N.; Van Rees, K.; Pennock, D. Cultivation induced effects on below ground biomass and organic carbon. *Soil Sci. Soc. Am. J.* **2002**, *66*, 924–930.
24. Sladden, S.E.; Bransby, D.I.; Aiken, G.E. Biomass yield, composition, and production costs for eight switchgrass varieties in Alabama. *Biomass Bioenergy* **1991**, *1* (2), 119–122.
25. Mehdi, B.; Zan, C.; Girouard, P.; Samson, R. Soil Carbon Sequestration Under Two Dedicated Perennial Bioenergy Crops. Research Reports. Resource Efficient Agricultural Production (REAP)-Canada, 1998. Available at http://www.reap-canada.com/Reports/reportsindex.htm (accessed April 6, 2004).
26. Mitchell, R.; Vogel, K.P.; Schmer, M. Switchgrass for biofuel production. Extension **2016**, eXtension.org.
27. Mitchell, R.; Vogel, K.P.; Sarath, G. Managing and enhancing switchgrass as a bioenergy feedstock. *Biofuels, Bioprod. Biorefin.* **2008**, *2*, 530–539.

33

Erosion by Water: Amendment Techniques

X.-C. (John) Zhang

Introduction

Soil erosion by water and wind is a worldwide problem. For example, approximately 36 billion tons of soils were removed from global lands in 2012.[1] Soil erosion is perhaps the major environmental threat to profitability, sustainability, and resilience of world agriculture. This entry focuses on soil erosion by water and its control with various amendment techniques. Soil erosion by water is defined as the amount of soil lost or removed in a specified time over an area of land by the action of raindrops and surface flow of water. Fundamental erosion processes involve detachment, transport, and deposition of soil particles by the two erosive agents. Detachment is the dislodging of soil particles from the soil mass, which is bound together by physical and chemical bonding forces. Transport is the movement of detached soil particles over the soil surfaces. Deposition is the downward motion of sediment particles settling out of the surface flow of water. Deposition occurs when sediment load in the surface flow is greater than the maximum carrying capacity of the flow due to a decrease in surface flow power or velocity.

Based on the predominant erosion forms and mechanisms, erosion by water may be categorized as splash, interrill, rill, ephemeral gully, gully, and stream-channel erosion. Raindrop impact is the primary force causing soil detachment in areas between rills (i.e., interrill area), while net downslope transport by raindrop splash due to the slope effect is very small. Erosive power of the shallow sheet flow in the interrill areas is mostly too low to detach the cohesive soil but high enough to carry the detached soil particles to rills with the assistance of raindrop agitation. Rills are small channels (usually 50–300 mm wide by 50–150 mm deep)[2] normally formed by the incision of concentrated flow and can be obliterated by tillage. Detachment and transport in rills are primarily by flowing water without aiding of raindrop impact. Flow in rills often converges in a few larger channels before leaving the field boundaries. These channels are ephemeral gullies, which are generally shallow and wide (a few meters) but can still be tilled across.[2] When concentrated flow erodes channels deeper than can be crossed with farm equipment, the eroded channels are called gullies.[2] Another important sediment source is erosion in streams, including downward cutting and bank erosion. However, gully and stream erosion are not discussed in this entry because they cannot be directly controlled by any amendment technique. In contrast, upland erosion, which includes splash, interrill, rill, and ephemeral gully erosion, is primarily a soil surface phenomenon and therefore can be effectively controlled by surface treatments with various types of soil amendments. Upland erosion has both on-site and off-site consequences. The main on-site impacts are loss of soil productivity and reduction of soil quality, while the main off-site effects are discharge of sediment and agricultural pollutants (e.g., nitrogen, phosphorous, pesticides, and herbicides) into streams and other surface water bodies. Thus, controlling upland erosion is of great importance in both preserving land productivity and protecting surface water quality.

Material added to soil to benefit soil or plant is called a soil amendment.[3] Soil amendments generally fall into two groups: inorganic and organic. Inorganic amendments include lime, gypsum, coal combustion by-products, stone, gravel, sand, and many others. Organic amendments comprise manure, compost, municipal and industrial wastes such as sewage sludge and paper mill sludge, synthetic polymer, various plant-derived mulches (e.g., crop residue, stubble, sawdust, woodchips, straw, leaves, and barks), etc. Soil amendments, based on their origins, may be arbitrarily grouped into five different categories: mined or mineral, synthetic polymer, manure and compost, municipal and industrial wastes, and mulches. Technically, chemical fertilizers are soil amendments, but conventionally they are not referred to as such. Many materials provide multiple benefits or functions to the soil and crops. Thus, soil amendments cannot be strictly defined. For example, manure has traditionally been used as a fertilizer for centuries, but it also conditions soil by improving its physical, chemical, and biological properties. This is true for almost all organic amendments, either raw or composted, which eventually decompose and release nutrients into soil through nutrient cycling. For comprehensiveness, manure, compost, and organic urban and industrial wastes are considered as soil amendments in this entry.

A closely related term to soil amendment is soil conditioner, which is a subset of soil amendments, aimed at creating favorable physical environments for plant growth such as improving aeration, water infiltration, water holding capacity, soil bulk density, aggregate stability, soil structure, resistance to erosion, and tilth. Soil amendments are either incorporated into the soil or applied on the surfaces. Incorporated materials intend to alter soil properties in the tillage layer, while surface-applied materials are used to primarily protect properties of the soil surface. For centuries, soil amendments such as animal waste, farmyard manure, green manure, crop residues, and composts derived from these materials are often incorporated into soils to recycle plant nutrients and to improve soil structure and tilth. During the past half-century, surface-applied mulches, notill mulches, stubble mulches, and living mulches have been widely used to protect the soil surface from wind and water erosion.[4] In the past decades, agricultural land applications of municipal and industrial wastes, such as sewage sludge, paper mill sludge, and coal combustion by-products have increased dramatically due to increasing landfill costs.[5] More recently, new and inexpensive polymer products have revived the interests in using synthetic polymers to reduce water runoff and soil erosion.

Overall, soil amendments impart beneficial effects on soil physical, chemical, and biological properties and promote plant growth.[3–5] Most amendments are also important agronomic measures for soil erosion control. Soil amendments are often chosen to increase soil aggregation and soil structural stability,[6] increase soil's resistance to erosion, protect soil from raindrop splash erosion, reduce surface seal formation (a thin, dense surface layer with very low water permeability, resulting from raindrop impact), increase water infiltration, decrease surface runoff volume, obstruct water flow, and therefore reduce soil erosion. The specific benefits for erosion control are derived from the decreased erosive power of both raindrop impact and flowing water as well as the increased soil's resistance to erosion. Thus, most amendments are effective in reducing soil erosion by water, though only a few (e.g., standing residue cover, living mulches) are effective in controlling erosion by wind.

Types of Amendment and Erosion Control

Lime

Liming materials such as quicklime (CaO) and limestone minerals are the most widely used amendments that are mined. Liming soil has a long history in the United States. Early settlers started to lime soil from the beginning of the 19th century. By the early 20th century, liming had become a routine agricultural practice to raise soil pH and increase crop productivity. Liming soil generally improves soil physical and chemical properties[3,7] in two ways. First, it neutralizes soil acidity, which reduces phytotoxicity of Al, Mn, and H ions and increases plant availability of Ca and many nutrients and microbial nutrient cycling processes. Second, it promotes soil aggregation and soil structural stability by increasing Ca^{2+} in soil solution and on soil exchange sites. However, overliming can result in soil clay dispersion and aggregate disintegration in highly weathered, high-Al^{3+} low-pH soils. Liming reduces soil erosion by improving soil aggregation and structural stability and by increasing surface cover due to enhanced plant growth. Lime has also been successfully applied as a soil stabilizer and cementing agent to control side sloughing of canal embankments and to prevent subsurface (piping) erosion on earthen dams, which are composed of highly dispersive and expansive clays.[3,8]

Gypsum

Mined and by-product gypsum are widely used as soil amendments to improve soil physical and chemical properties, reduce soil erosion, and increase nutrient availability of calcium and sulfur. Gypsum has been successfully used for decades to correct structurally unstable, high-Na^+ soils (e.g., saline and sodic soils). Because soil clay and organic matter are primary structural components (cementing agents) for aggregate formation, the displacement of clay by dispersion often leads to breakdown of soil aggregates and deterioration of soil structure, while clay flocculation promotes soil aggregation and structural stability. The beneficial effects of gypsum addition on controlling soil erosion result from increased stability of soil aggregates and structure due to elevated electrolyte concentration in soil solution and the predomination of calcium ions on the exchange sites of soil clays.

Cation charge and electrolyte concentration are the two major factors influencing clay flocculation or dispersion.[9] When two clay particles approach each other, their electric double layers (a negatively charged clay surface surrounded by a compensating cation cloud) begin to repel each other when the two cation clouds overlap. The cation cloud is pushed toward the particle surface when electrolyte concentration is increased or when a divalent cation replaces a monovalent cation. The reduction in the cloud thickness results in a decrease in this repulsive force. When a critical electrolyte concentration is reached, van der Waals' attractive force overcomes the repulsive force resulting in a rapid flocculation of clay particles. This concentration is termed critical flocculation concentration (CFC). Because cation charge has a more dominant effect on clay flocculation than electrolyte concentration, Al salts such as alum have been widely used to flocculate suspended particles in many water treatment plants. Among the common soil cations (e.g., Al^{3+}, Ca^{2+}, Mg^{2+}, K^+, Na^+, and H^+), Ca^{2+} has the greatest flocculation power apart from Al^{3+} (Al^{3+} is only significant in weathered, acid soils), while Na^+ has the least. This is why sodic soils have poor aggregate and soil structure stability, and gypsum addition to these soils helps to improve soil structure. Based on these principles, clay particles flocculate when electrolyte concentration is greater than CFC; otherwise, dispersion takes place. A visual demonstration of clay flocculation is illustrated in Figure 1. After 6-hour settling, more dispersed clay particles have formed into microaggregates and settled out to the bottom as solution concentrations of gypsum increase.

It is well known that soils having high concentration of Na^+ are not structurally stable. However, low-Na^+ soils can also be destabilized and dispersed by external forces such as raindrop impact due to low electrolyte concentration in rainwater,[10–12] resulting in seal formation on the soil surface. Clay dispersion and soil aggregate breakdown by raindrop impact are the two primary processes causing seal formation on soil surfaces. Surface seals often have water hydraulic conductivities several orders less than the same non-sealed soils. Thus, surface application of gypsum can prevent clay dispersion, reduce seal formation, increase infiltration, and reduce erosion of easily dispersed soils.

Overall, gypsum addition is effective in preventing clay dispersion, stabilizing soil aggregates and structure, reducing surface seal formation, increasing water infiltration, and reducing soil erosion by water.[3,10,11] It is well documented that gypsum can reduce soil erosion by (1) promoting water infiltration and reducing erosive power of runoff, (2) increasing soil resistance to erosion by increasing aggregate

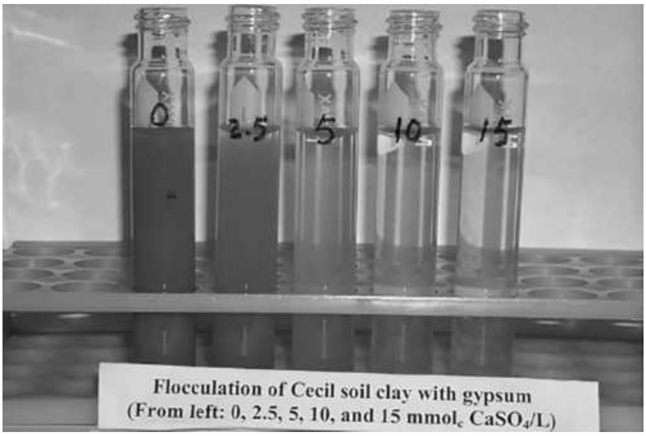


FIGURE 1 Gypsum flocculation test series after 6-hour settling of the Cecil soil clay in gypsum solution.

size and stability, and (3) enhancing flocculation and deposition of clay particles suspended in runoff.[13] Based on field experiment conducted in many countries, surface application of 3–10 Mg/ha of gypsum would considerably reduce surface runoff and soil erosion in most soils for several months after application or until mixed into soils by tillage.[14] If it is applied as a source of fertilizer calcium and sulfur, 0.1–0.5 Mg/ha is often adequate.

Synthetic Polymer

Polymers are repeated monomers chained together to form long extended molecules and are characterized by their molecular weight, configuration, and charge properties. Polyacrylamide (PAM) and polysaccharide are widely studied and used as soil conditioners. They can be cationic, nonionic, and anionic, among which anionic PAM is most widely used. The effectiveness of polymers in altering soil properties is related to their chemical properties, water quality (ionic type and strength), and soil characteristics. The irreversible interaction between polymers and fine soil particles, especially clays, lays the foundation for its ability to stabilize soil aggregates, promote water infiltration, and therefore reduce surface runoff and soil erosion.

The use of synthetic polymers to improve soil physical properties began in the 1950s. In early agricultural studies, polymers were usually incorporated into the entire plow layer to promote structural stability and water permeability. During the 1950s, reports that polymers were effective in reducing erosion, improving water infiltration, and increasing crop yields stimulated a great deal of interest among agriculturists. However, the widespread use of synthetic polymers available at the time was restrained by high cost and large quantities needed for soil incorporation.

Since the 1980s, new polymer products and new application strategies have rekindled interest in these soil conditioners. Only a small amount of polymer, when sprayed on the very soil surface or dissolved in irrigation water, can reduce surface seal and crust formation by preventing clay dispersion and stabilizing soil aggregates, increase water infiltration, and reduce soil erosion.[3] When surface applied, these new polymer products are especially effective with unstable, sealing-prone soils. The thread-like large polymer molecules or interwoven polymer nets are adsorbed strongly (often irreversibly) to soil aggregates and clay particles. The adsorbed polymer nets that surround soil aggregates physically protect them from disintegration by raindrop impact or shear force imparted by flowing water. Polymers can also chain clay particles together to promote flocculation. Polymers have been used to remove fine particles from water body in water treatment plants and in sedimentation ponds.

Anionic PAM, being similar to soil humus in charge property, is most extensively used as soil conditioners. The effectiveness of anionic PAM is usually improved by adding a source of cation such as gypsum because cations are essential for building cation bridges between negatively charged clay particles and PAM molecules. Anionic PAM has also been successfully used to reduce soil erosion in furrow and sprinkler irrigation, on construction sites, and on steep slopes such as road banks.[15] Generally, application rates vary with PAM properties, soil characteristics, and specific goals of a particular treatment. To enhance infiltration and reduce soil erosion in flood and sprinkler irrigation, normally less than 5 kg/ha when dissolved in irrigation water would be sufficient.[3] When sprayed on soil surfaces to control surface seal formation, rates of 10–30 kg/ha are adequate for most soils. To control erosion on steep slopes such as road embankments or on construction sites, greater amounts up to 100 kg/ha may be needed. The effectiveness of PAM in reducing runoff and soil erosion can last several weeks or even months following application.

Manure, Compost, and Organic Sludge

Animal, farmyard, and green manures have been used as organic soil amendments to improve soil productivity for many centuries. Since the 1990s, application of municipal biosolids (sewage sludge) and industrial waste compost such as paper sludge to agricultural lands has increased dramatically. These materials have the ability to increase soil aggregate stability, porosity, and water holding capacity; reduce water runoff and soil erosion; improve soil tilth; increase plant nutrient availability and soil cation exchange capacity; and elevate soil microbial activities.[3] Decomposed organic substances such as humus along with soil clays are primary binding agents for soil aggregation. An increase in soil organic matter content often leads to increases in soil aggregation and structural stability. Thus, addition of organic materials such as manure and compost reduces soil erosion by water by improving soil aggregation, reducing runoff, and increasing soil resistance to erosion. In addition, improved soil conditions and fertility enhance root and plant growth, resulting in greater plant cover and root biomass, which further protect soil and reduce soil erosion. Large amounts of these organic amendments must be applied over decades to substantially increase soil humus levels.

Organic and Inorganic Mulches

Most mulch materials used for runoff and erosion control are naturally occurring. Both inorganic and organic mulches have been used. Inorganic mulches include stone, gravel, sand, and many others, while organic mulches are normally undecomposed materials derived from living matter such as crop residue, hay, leaves, woodchips, and sawdust.[3] These materials dissipate raindrop impact energy, protect surface soil aggregates from disintegration by raindrop impact, prevent surface seal formation, increase water infiltration, reduce the sediment-carrying capacity of overland flow by obstructing flow, and reduce soil erosion.

One of the key benefits of mulches is erosion control. The great benefit of erosion control results from the ability of surface mulches to reduce not only soil detachment by raindrop impact and flowing water but also the sediment-carrying capacity of overland flow. Surface-applied mulches of crop residue, no-tillage mulches, and stubble mulches are found to be the most effective agronomic measures in controlling soil erosion.[16,17] Moreover, living mulches, such as cover crops that are later turned into soil as an organic amendment, permanent grass buffer/filter strips, and grass waterways, are even more effective than residue mulches in controlling soil erosion. These living mulches can increase soil organic matter and nutrient cycling, reduce soil splash by canopy cover, reduce downward scouring of concentrated flow by interlocked root networks, enhance sediment deposition by slowing down water flow, and reduce pesticide and chemical movement to surface waters by adsorption and filtration.

Overall, surface mulches are very effective in reducing surface runoff and soil erosion by water. Many studies have shown that both runoff volumes and soil loss rates decrease rapidly (often exponentially) with increases of surface mulches.[4,16,17] Buried or incorporated residues are also effective in reducing soil erosion, especially rill erosion. However, the reduction is generally less as compared to surface residues. In addition, all mulches that cover the soil surface (gravel) or anchor the soil (stubble) are also effective in reducing soil erosion by wind. Furthermore, organic mulches eventually decompose into humus and impart to the soil all the beneficial qualities derived from traditional organic matter additions as discussed earlier.

Conclusion

Soil amendments are generally used to improve the chemical, physical, and biological properties of the soil and therefore to promote plant growth. Lime neutralizes soil acidity, remedies Al and Mn toxicity, enhances microbial activity, increases plant availability of many nutrients, and promotes soil aggregation and structural stability. Gypsum and synthetic polymers prevent clay dispersion, reduce surface seal formation, protect and stabilize soil aggregates and structure, reduce soil erodibility, increase water infiltration, and reduce soil erosion by water. Organic amendments such as animal manure, compost, and municipal and industrial wastes add organic matter to soils, increase microbial activity, enhance nutrient cycling and availability, improve soil tilth, increase aggregate stability and water holding capacity, enhance water infiltration, and reduce surface runoff and soil erosion. Inorganic mulches like stone and sand physically prevent dislodgement and breakdown of soil surface aggregates by raindrop impact, reduce surface seal formation, suppress soil evaporation, increase water infiltration, obstruct overland flow, and reduce sediment transport capacity and soil erosion. Organic mulches such as crop stubble and grass filter strips impart to the soil the combined benefits of the inorganic mulches and organic amendments as discussed above. Being the most important and most effective agronomic measures for erosion control, organic mulches not only protect soil from detachment by raindrop impact and flowing water shear but also reduce the sediment-carrying capacity of the flow by impeding it.

References

1. Borrelli, P.; Robinson, D.A.; Fleischer, L.R.; Lugato, E.; Ballabio, C.; Alewell, C.; Meusburger, K.; Modugno, S.; Schutt, B.; Ferro, V.; Bagarello, V.; Van Oost, K.; Montanarella, L.; Panagos, P. An assessment of the global impact of 21st century land use change on soil erosion. *Nature Commun.* 2017, *8*. doi:10.1038/s41467-017-02142-7.
2. Foster, G.R.; Young, R.A.; Romkens, M.J.M.; Onstad, C.A. Processes of soil erosion by water. In Soil Erosion and Crop Productivity; Follett, R.F., Stewart, B.A., Eds.; ASA, CSSA, and SSSA: Madison, WI, 1985.
3. Wallace, A.; Terry, R.E., Eds. *Handbook of Soil Conditioners: Substances that Enhance the Physical Properties of Soil*; Marcel Dekker: New York, 1998.
4. Oschwald, W.R., Ed. *Crop Residue Management System;* ASA Spec. Publ. No. 31; ASA, CSSA, and SSSA: Madison, WI, 1978.
5. Rechcigl, J.E.; MacKinnon, H.C., Eds. *Agricultural Uses of By-products and Wastes*; ACS Symp. Series 668; ACS: Washington, DC, 1997.
6. Emersion, W.W.; Bond, R.D.; Dexter, A.R., Eds. *Modification of Soil Structure;* John Wiley & Sons: New York, 1978.
7. Adams, F., Ed. *Soil Acidity and Liming;* Agron. Monogr. 12; ASA, CSSA, and SSSA: Madison, WI, 1984.
8. Gutschick, K.A., Ed. *Lime for Environmental Uses;* ASTM: Philadelphia, PA, 1985.
9. Van Olphen, H. *An Introduction to Clay Colloid Chemistry;* John Wiley & Sons: New York, 1977.
10. Sumner, M.E.; Stewart, B.A., Eds. *Soil Crusting: Chemical and Physical Processes*; Lewis: Boca Raton, FL, 1992.

11. Shainberg, I.; Sumner, M.E.; Miller, W.P.; Farina, M.P.W.; Pavan, M.A.; Fey, M.V. Use of gypsum on soils: A review. *Adv. Soil Sci.* **1989**, *9,* 1–111.
12. Miller, W.P. Infiltration and soil loss of three gypsumamended Ultisols under simulation rainfall. *Soil Sci. Soc. Am. J.* **1987**, *51* (5), 1314–1320.
13. Levy, G.J.; Sumner, M.E. Mined and by-product gypsum as soil amendments and conditioners. In Handbook of Soil Conditioners: Substances that Enhance the Physical Properties of Soil; Wallace, A., Terry, R.E., Eds.; Marcel Dekker: New York, 1998; 187–216.
14. Miller, W.P. Environmental considerations in land application of by-product gypsum. In Agricultural Utilization and Urban and Industrial By-Products; Karlen, D.L., Wright, R.J., Kemper, W.O., Eds.; ASA, CSSA, and SSSA: Madison, WI, 1995.
15. Levy, G.J.; Ben-Hur, M. Some uses of water-soluble polymers in soil. In Handbook of Soil Conditioners: Substances that Enhance the Physical Properties of Soil; Wallace, A., Terry, R.E., Eds.; Marcel Dekker: New York, 1998; 399–428.
16. Unger, P.W., Ed. *Managing Agricultural Residues*; Lewis: Boca Raton, FL, 1994.
17. Wischmeier, W.H.; Smith, D.D. *Predicting Rainfall Erosion Losses: A Guide to Conservation Planning*; USDA Agric. Handb. 537; U.S. Gov. Print. Office: Washington, DC, 1978.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

34

Erosion by Water: Assessment and Control

José Miguel Reichert, Nadia Bernardi Bonumá, Gustavo Enrique Merten, and Jean Paolo Gomes Minella

Introduction

Soil water erosion is the process of detachment and transportation of soil particles from the point of origination by the action of rainfall and runoff. Although erosion is a natural process, human activities have increased erosion either due to inappropriate agricultural and deforestation practices or due to constructions and mining activities without considering the soil characteristics. Land degradation and water pollution owing to erosion have become severe problems worldwide. In order to assess and prevent the environmental impacts due to soil erosion, several control measures can be used, but first, it is necessary to understand the sources and factors affecting the erosion process. The specific objectives of this entry are to discuss 1) the causes of water erosion and the link between on-site and off-site environmental impacts as a result of soil losses; 2) erosion assessment by monitoring, use of mathematical models, and indicators; 3) the erosion control practices; and 4) an integrated environmental management and the role of farmers, constructors, scientists, and managers in order to reach a sustainable strategy to control soil erosion.

Erosion Problems

Soil erosion by water is the most significant land degradation problem worldwide.[1] Even though erosion is a natural geomorphic process occurring continually over the earth's surface,[2] human activities can magnify, minimize, or prevent the operation of this natural process.[3] Man has been the most active agent linked to accelerate erosion either due to inappropriate agricultural and deforestation practices or due to constructions and mining activities without considering the soil characteristics.

Each year, 75 billion tons of soil are removed due to erosion, with most coming from agricultural land, and around 20 million ha of land are lost.[4] In the United States, according to the U.S. Environmental Protection Agency and the federally mandated National Pollution Discharge Elimination System, soil erosion is considered the biggest contributor to nonpoint source pollution.[5] In the European Union, the main problems for soils are contamination and losses due to increasing soil erosion.[6] In the developing world, soil erosion is the single most important environmental degradation problem, which has far-reaching economic, political, social, and environmental implications due to both on-site and off-site damages.[4]

In agricultural areas, according to Renschler and Harbor,[3] farming operations directly affect soil movement through activities such as tillage,[7] root crop harvesting, and the trampling of soil and removal of vegetation by livestock.[8]

Soil losses from agriculture are of concern because of both agricultural on-site and environmental off-site impacts.[9] The agricultural on-site impacts are related to soil degradation and include increased bulk density, reduced aggregate stability, and depletion of nutrient and carbon content,[10] and consequently reduced soil productivity.[1] The problem may become so severe that the land can no longer be cultivated and must be abandoned, which have practical and economic implications for land managers.[6]

The environmental off-site impacts are the increase in sediment and other pollution loads in receiving waters.[1] Excessive sediment inputs to rivers can cause physical problems of turbidity in the water column and sedimentation in channels, reservoirs, estuaries, harbors, and the near-coastal zone. In turn, this sedimentation may affect channel morphology and behavior, and navigation.[11] The cost of dredging the several billion tons of sediment from rivers and harbors each year is about 15 times more than the cost of holding the soil on the land from which it eroded.[12] Sediment is also a pollutant in its own right and, through the chemicals adsorbed to it, can increase the levels of nitrogen and phosphorus in water bodies and result in eutrophication.[9]

Sources of Erosion and Sedimentation

Soil water erosion is the process of detachment and transportation of soil particles from the point of origination by the action of rainfall and runoff. Water erosion of soils starts when raindrops strike bare soil peds and clods, separating particles and causing the finer particles to move with the flowing water as suspended sediments.[12] Sources include agricultural lands; disturbed forest lands; construction sites; roadway embankments, cuts, and ditches; surface mines; and natural geological eroding "badlands."[13]

Erosion can happen within rills, interrill (the area between rills), tunnels, gullies, and stream channels. Interrill erosion process is rainfall dominated, whereas rill erosion is mostly defined by runoff. Rills are such small concentrations of running water that they can be completely removed by normal cultivation methods, whereas gullies cannot be.[14] Gully erosion is defined as the erosion process whereby runoff water accumulates and often recurs in narrow channels and, over short periods, removes the soil from this narrow area to considerable depths.[15] Tunnel erosion may occur in soils with sublayers that have a greater tendency to transport flowing water than does the surface layer.[14] Stream channels can be sources (stream channel erosion) or sinks of sediments (sedimentation).

Erosion Assessment

To address soil erosion processes and risk, scientists and engineers from various fields have developed physical parameters, equations, and models with the intention of implementing assessment tools for educational, planning, and legislative purposes.[3]

Monitoring

The purpose of the assessment and the size of the area (scale) will strongly influence the approach. Time scale and funding will influence the methods that can be used.[16] Field measurements may be classified into two groups: those designed to determine soil loss from small areas or erosion plots and those designed to assess erosion over a large area such as a drainage basin.[10]

Field Plots

In order to describe the factors that cause the erosion process and the effectiveness of erosion control techniques, several studies on experimental plots have been done.

The experimental plots may be soil pans or field plots. Single or several pans may be used in a static or mobile position.[17] Field erosion plots may be subject to natural or simulated rainfall. On small plots, a rotating simulator (Figure 1) can be set up between two plots, which can be rained on at the same time.[18] The standard plot is 22 m long and 1.8 m wide, though other plot sizes are sometimes used.[10]

One of the best uses for field plots is demonstration; examples are to demonstrate to farmers that serious erosion is taking place or to show that erosion is much less from a plot that has a good vegetative cover than from a bare plot. Another valid use is in comparative studies, for example, to test or get an approximate indication of the effect on runoff or erosion of a simple comparison such as with or without a surface mulch, or the amount of runoff at the top and at the bottom of a slope. A third possible use is to obtain data that are to be used to construct or to validate a model or equation to predict runoff or soil loss.[18] Data from experimental plots have revealed some precise mathematical relationships between erosion factors and erosion losses and have led to setting up of equations that enable calculations of precipitation interrill erosion to be made, and which maintain their validity.[19] The most thorough analysis of these relationships is the Universal Soil Loss Equation (USLE)[20] based on a database of approximately 10,000 plot-years of field experiments under natural rainfall in the United States, supplemented by data collected using rainfall simulators.

However, it is clear that using standard experimental plots will not give estimates of what is happening in the landscape as a whole as regards erosion rates.[6]

FIGURE 1 Rotating rainfall simulator.

Sediment Transport in Watersheds

River sediments present a variable grain size ranging from small particles, such as clays <0.002 mm to larger particles such as very coarse gravel (32–64 mm diameter) and boulder (2048–4096 mm). Depending on the size of these particles and the hydrodynamic conditions of flow (especially flow velocity), sediment is transported in suspension or by bedload. Fine particles, such as clay, silt, and medium sand fractions (<0.250 mm), are transported in suspension by the process of advection and diffusion, while larger particles, such as coarse sands and gravel (>0.250 mm), are transported preferentially by bedload over the bed.

Sediment transport in rivers (suspension and bedload) presents great spatial–temporal variability. In a river cross section, the suspended sediment concentration (SSC) varies both in the longitudinal direction (section width) and in the vertical one (section depth). In general (but not necessarily in this way), the SSC in a cross section varies from the center to the banks so that the highest SSCs are found in the region close to the center of flow and the lowest are found close to the banks. This condition is a function of the current velocity that is usually greater in the center and lower close to the banks (due to the effect of hydraulic roughness of the banks and the bed that caused the reduction of flow velocity). Also, depending on velocity, concentration along a vertical tends to increase with depth because coarser and/or denser sediment particles may not be homogeneously distributed.

The sediment flux represents how much sediment is transported by a river. Sediment flux is mass in time and the unit is represented in grams per second or metric tons per day. In order to measure bedload flux, samplers are positioned on the river bed for a few minutes so that a mass of sediments is collected in a given time (Figure 2). On the other hand, the suspended load flux is measured indirectly using the following relation:

$$\text{Supended sediment flux}\left(\text{tons day}^{-1}\right)=\text{SSC}\left(\text{mg L}^{-1}\right)\times Q\left(\text{m}^{3}\ \text{sec}^{-1}\right)\times 0.0864 \tag{1}$$

where SSC is the suspended sediment concentration (g m^{-3}) and Q is the flow rate ($\text{m}^3\ \text{sec}^{-1}$). In this way, to measure suspended sediment flux, it is necessary to obtain two variables, which are the SSC and Q, taking into account the spatial–temporal variability of SSC.

There are two principles involved in the collection of representative SSC samples. The first is isokinetic sampling that uses equipment (Figure 3) where the velocity of the river is used to fill the sampler.

FIGURE 2 Equipment used to measure bedload rates.

FIGURE 3 Equipment used to collect the water and suspended sediment to determine SSC.

Isokinetic sampling means that during the flow sampling to determine SSC, the sampler should be able to collect the mixture of water and sediments through a nozzle without changing flow velocity. If there is a change in this velocity, the SSC collected will not be representative of the SSC of flow. The second principle is depth-and-width integrated sampling, which collects water/sediment mixtures from the entire cross section. Details about methodology to sampler bedload and SSC can be found in the work by Edwards and Glysson.[21]

Automatic sampling equipment (Figure 4) does not collect isokinetically because the sample is pumped into the container, and it does not collect depth-and-width-integrated samples because it collects the sample from a single point in the cross section. However, it is possible to adjust the data generated from automatic samples using a so-called box equation. This equation is site specific and is developed by collecting concurrent samples (simultaneously collecting a manual isokinetic depth-and-width-integrated sample while pumping through the auto-sampler) over the typical range of flows at the site.[21]

FIGURE 4 Auto-sampler used to collect the water and suspended sediment to determine SSC.

The total amount of sediment that exits a drainage basin over a given period of time, usually a year, is called *sediment yield*.

Sediment yield can be measured using techniques that have been explained or estimated using the sediment delivery ratio (SDR) concept. The proportion of eroded sediment that exits the drainage basin is called SDR,[14] which, in other words, is the ratio of the primary erosion rate on hillslopes to the sediment yield at the basin outlet. Primary erosion hillslope rate is normally estimated using mathematical models and SDR is estimated using empirical relation based on drainage basin area.

Fingerprinting techniques[22] such as mineral magnetics and 137C can offer powerful but time-consuming approaches to establishing the source of the sediment.[6] In essence, the fingerprinting approach involves comparison of the geochemical properties of suspended sediment transported by the river with those of potential sources, in order to establish their relative importance.[23] At a relatively small scale of a few square kilometers, assessment of erosion can incorporate sediment fingerprinting to determine the source of deposits, e.g., in dams.[6]

Models

Mathematical models have been proven to be a cost-effective tool for improving our understanding of erosion processes and evaluating possible effects of land-use changes on soil erosion and water quality.1] Several water erosion and sediment transport models at hillslope and watershed scales can be found in the literature.[2,14,24–26] The mathematical models can be classified in various ways, but the most useful distinction is between empirical, conceptual, and physically based models.

The USLE[20] is an empirical model developed to estimate soil losses and to be used as a decision-making tool in soil conservation planning in cropping lands on a site basis.[27]

The USLE is often given as

$$A = R \cdot K \cdot L \cdot S \cdot C \cdot P \tag{2}$$

where A is the average annual soil loss, R is the rainfall erosivity factor, K is the soil erodibilty factor, L is the slope length factor, S is the slope gradient factor, C is the crop factor, and P is the conservation practice factor.

The USLE was originally applied to the prediction of soil losses from agriculture in the United States in order to preserve soil resources but has been extended for use in numerous countries.[28] In the way in which the USLE has generally been used (or misused), it provides an average rate of erosion over a long period of time—how that average is arrived at, or what events contribute to it, is not the issue. This simple point is often ignored or glossed over: erosion occurs when there is inadequate vegetation cover to offer protection, and sufficient runoff; both conditions must be satisfied. The USLE does not address this issue—it deals with average conditions.[6] Its modified version (MUSLE[29]) has been an attempt to compute soil loss for a single storm event. The USLE was also revised (RUSLE) and revisited[30] for improvement.

The Areal Nonpoint Source Watershed Response Simulation (ANSWERS[31]) includes a conceptual hydrological process and a physically based erosion process. The Agricultural NonPoint Source model (AGNPS[32]) simulates runoff, sediment, and nutrient transport from agricultural watersheds. The Water Erosion Prediction Project (WEPP[33]) is a model to predict soil erosion and sediment delivery from fields, farms, forests, rangelands, construction sites, and urban areas. The Griffith University Erosion System Template (GUEST[34]) was developed to interpret temporal fluctuations in sediment concentration from bare soil in single erosion events. The Limburg Soil Erosion Model (LISEM[35]) is a physically based runoff and erosion model that simulates the spatial effects of rainfall events on small watersheds. The European Soil Erosion Model (EU-ROSEM[36]) is a model for predicting soil erosion by water from fields and small catchments. The Soil and Water Assessment Tool (SWAT[37]) is a watershed scale model developed to predict the impact of land management practices on water, sediment, and agricultural chemical yields in complex watersheds with varying soils, land use, and management conditions over long periods of time.

Indicators

Visual Indicators

Monitoring soil erosion is costly, and the use of soil erosion models requires massive data input. Almost all soil erosion models, including USLE, are empirical and require careful calibration between actual soil erosion measurements and predictors prior to each application.[38] To improve the models, addition of an intermediate step, i.e., use of observed, site-specific soil erosion indicators, could bypass the massive data requirements of soil erosion models.

Examples of soil erosion indicators are defined as follows:[38]

- Eroding clods: located above flow surfaces, having predominantly convex forms, shaped by splash and disintegration
- Flow surfaces: almost flat areas with a smoothed micro-relief with (often parallel) linear flow marks or sediments, partly eroded by shallow non-concentrated flows
- Pre-rills: shallow channels (3–5 cm deep) cut by concentrated flows, slightly concave and not part of the micro-drainage system of the area
- Rills: deeper channels (5–30 cm) cut by concentrated surface flows, with clear lateral micro-scarps on the sides, usually part of the micro-drainage system of the area
- Soil accumulation: areas with smoothed micro-relief, caused by deposition of eroded materials

Quantitative Indicators

How much erosion can be tolerated? In order to answer this question, quantitative indicators that relate the rate of erosion and its effect should be determined.

For agricultural areas,[20] define the soil loss tolerance or allowable soil loss as the maximum level of soil erosion that will permit a high level of crop productivity to be maintained economically and indefinitely. Based on long-term average annual soil losses estimated by the USLE, values of soil loss tolerance were determined and a maximum value of 11 t/ha was chosen with lower values based primarily on the depth of soil favorable for plant root growth.[39] Nevertheless, these values are not applicable to construction or other nonfarm land use.[40] Although a soil tolerance value of 11 t/ha may maintain productivity, a value as low as 2 t/ha may be inadequate to control the offsite effects of erosion.[13] A modified definition of tolerable soil erosion is proposed by Verheijen et al.[41] as “any actual soil erosion rate at which a deterioration or loss of one or more soil functions does not occur,” actual soil erosion being “the total amount of soil lost by all recognized erosion types.”

Water Erosion Control

Soil erosion can be reduced by controlling either soil detachment or soil sediment transport or both.[12] The approach must be dynamic and must include the use of environmental technology, ecotechnology, and cleaner technology, as well as environmental policies.

$$\text{Environmental Technology} \times \text{Cleaner Technology}$$

The necessity of cleaner technology is an issue. Although no agricultural and engineering activity is totally clean, some practices of crop management and soil conservation can contribute to control soil water erosion. The use of environmental technology and ecotechnology includes mechanical methods such as contouring, terraces, contour ridges and stone terrace, waterways, stabilization structure, and geotextiles. Despite the decrease in soil erosion, these methods can cause some environmental impacts because of their mechanic nature, for instance, due to land movement and fuel consumption. On the other hand, some agricultural practices, such as no-tillage, not only can reduce soil erosion but also benefit the environment through C sequestration, energy conservation, and decreased N loss.

Soil Conservation Practices

Leaving vegetation and crop residue on the surface rather than burying it, applying mulches, and using soil-conservation cropping systems increase infiltration and reduce runoff.[42] A hypothetical example of how soil management affects soil erosion and water quality is shown in Figure 5.[43] The adoption of a farming system involving intense tillage for annual crops created favorable conditions for the occurrence of erosion, which generates a large amount of sediment. Moreover, the intense soil mobilization affects the water balance since the absence of soil cover decreases the infiltration of water. When the soil infiltration is reduced, the surface runoff volume increases and thereby reduces the amount of water for groundwater recharge. On the other hand, when the fields are cultivated with soil conservation practices, the coverage allowed above the soil surface reduces erosion and promotes water infiltration into the soil.

Soil-conservation practices often provide long-term benefits in exchange for immediate costs. The value of each ton of eroded soil varies according to its clay, organic-matter, and nutrient contents. A ton of eroded soil is usually more fertile, and therefore more valuable, than a ton of average soil because of textural sorting combined with the effects of organic-matter accumulation, fertilization, and liming that are concentrated in the upper part of the soil. Sedimentation damage occurs both in fields near the source of the sediment and in downstream areas. Some damage is unavoidable, but hundreds of millions of dollars per year could be saved through flood-prevention efforts.[43]

Contouring

Contouring consists of performing all field practices such as plowing, disking, planting, and cultivation on the contour (Figure 6). It is effective for low slopes and short slope length and for rainstorms of low intensity. The effectiveness of this practice is about 50%[44] depending upon slope grade and length, and this effectiveness can be increased by establishing a series of depressions or pits (Figure 7) that fill with water and sediment during rain.

Terraces

Terraces are mechanical structures constructed across a slope and consist of a channel and an earth or stone bank to intercept surface runoff, allow it to infiltrate and evaporate, or divert the excess to a stable outlet at a nonerosive velocity. Depending upon their function or type of construction, terraces are

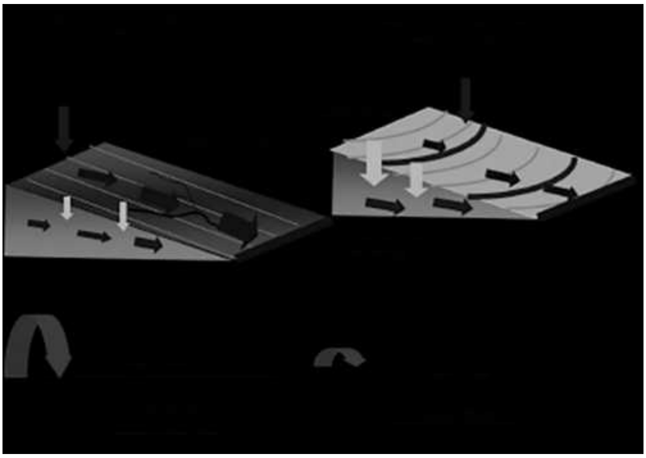


FIGURE 5 Comparative of hydrological effects and water quality impacts that occur in the watershed when subjected to a cultivated field with and without soil conservation practices.

FIGURE 6 Contouring in an agricultural field in southern Brazil.
Source: Reichert and Cassol.[43]

FIGURE 7 Sedimentation pits in a horticultural field in Costa Rica.
Source: Reichert and Cassol.[43]

classified as retention or absorption, diversion or graded, and bench terraces. Retention or absorption terraces are used when water is to be stored on the hillside of slopes less than 4.5°, having level channels able to accumulate runoff volume generated by a 10 years return period rainfall (Figure 8). Diversion or graded terraces aim at intercepting runoff and conducting it to a proper outlet through a channel of a slight grade, usually 1:250 (Figure 9) on slopes less than 7°. Bench terraces are used on steep slopes, up to 30°, alternating a series of shelves (cuts) and raises (fills) (Figure 10). Research data from several

FIGURE 8 Retention or absorption terraces in southern Brazil, accumulating runoff in a level channel.
Source: Reichert and Cassol.[43]

FIGURE 9 Diversion or graded terraces in southern Brazil.
Source: Reichert and Cassol.[43]

FIGURE 10 Bench terraces on high-grade hillslope in Guatemala.
Source: Reichert and Cassol.[43]

authors[10] indicate effectiveness in soil loss control of bench terraces up to 93%. Terraces must be properly designed to avoid even greater soil erosion damage[44] and should be planned considering the whole farm, a group of neighboring farms, or even a small hydrographic catchment,[45] to ensure optimization of land use.

Contour Ridges and Stone Terraces

Contour ridges are earth banks, similar to narrow-based terraces, constructed across the slope using hand tools, animal traction implements, and sometimes motorized farm machinery. These structures are frequently used on small-scale farms and hilly regions. Their effectiveness in reducing soil loss varies, but research data from several authors indicate values up to 80% and even 100% control.[10] Contour stone bunds (Figure 11) are an alternative technique on stony lands, where stone walls are set up in shallow trenches. These allow slow retention of sediment and are permeable to water. With time, upslope sedimentation will generate a gradual development into bench terraces. The construction of stone barriers requires considerable labor and time, but it is more permanent and easier to maintain than earth barriers. Spacing and construction details are presented in Manual on Integrated Soil Management and Conservation Practices, IITA and FAO.[46]

Waterways

A system of graded terraces requires waterways to receive the excess drained runoff and safely conduct at a nonerosive velocity to lower parts of the landscape. Waterways must be carefully established (considering location, construction, and stabilization) and maintained (managing grass species and controlling incipient erosion) to avoid serious erosion problems, even causing gully formation. On grass waterways, tile drains, or steep slopes, concrete structures (Figure 12) are located in natural depressions or, if needed, reshaped channels. The grass waterways reduce water-flow velocity because of the retardance effects of vegetation and increased soil resistance to erosion due to soil coverage and aggressive rooting system holding soil particles together. To work properly, waterway design follows defined principles from channel hydraulics[10,46] and the flow ways must be maintained and cared for.[45]

FIGURE 11 Contour stone barriers in an agricultural field in Cape Verde.
Source: Photo by L.D. Norton.

FIGURE 12 Grass waterway draining to a concrete drop inlet structure in southern Brazil.
Source: Reichert and Cassol.[43]

Conservation Structures

In certain situations, the only feasible way to convey safely large amounts of runoff water from a higher elevation to a lower one is through the use of conservation structures that are made of some type of nonerodible material, such as concrete or large rocks. Examples of these structures include chute drop inlets (Figure 12), flumes and drop spillways (Figure 13), and pipe spillways, which must follow standard functional requirements to maintain channel stabilization.[45] Selection and design of these structures are based upon existing or anticipated erosion problem, slope steepness, estimated maximum water flow discharge, and cost of alternative systems. Conservation structures are usually used in conjunction with other erosion-control techniques such as grass waterways (Figure 12).

FIGURE 13 Conservation structures showing drop spillway and flumes to convey water downslope, along with stone barriers, in Spain.
Source: Reichert and Cassol.[43]

FIGURE 14 Gully development on sandy soil in southern Brazil.
Source: Reichert and Cassol.[43]

Gully Stabilization Structures

Gullies represent an advanced stage of rill erosion (Figure 14) and are most important as a source of sediment in streams. If the gully is small, recovery might be suitable by filling it with soil, but in many cases, recovery is technically and economically not viable. In such cases, gully stabilization is recommended by constructing small dams using diverse materials depending on availability. These materials may include stones, wood planks or branches, earth, etc., to trap sediment, thus reducing gully depth and slope. Along with these stabilization structures, plants must be established to reclaim the affected areas (Figure 15). Sometimes, to control the overfall of water on the headwall on large gullies, more permanent structures are required.[9,44]

Geotextiles

Erosion on road banks or urban areas is best controlled by vegetation, but establishment on these highly erodible and usually infertile soil surfaces is difficult. Thus, other materials, including straw and hay, are used to help establish vegetation, but these do not hold tight to the soil and are washed away easily.[44] Alternative materials include geotextiles that are permeable textiles for erosion control interacting with soil and vegetation. The geotechnical material is in the form of a mat, sheet, grid, or web of natural or synthetic fiber.[10] The first type is biodegradable while the second type gives a more permanent protection to a slope, interacting with roots to improve soil cohesion and slope stability.

Environmental Legislation

Since the 1960s, Green Revolution technologies have been introduced in many countries actively supported by irrigation development, subsidized credit, and fertilizer programs.[4] Erosion is currently not regarded as simply due to water; there is a growing political awareness of the value of soil.[16] Both environmental legislation and economic instruments have the potential to improve the performance of soil erosion control.

FIGURE 15 Gully stabilization with used tires, stones, and vegetation in Costa Rica.
Source: Reichert and Cassol.[43]

Environmental legislation such as banning cultivation in hilly land and riparian zones is very important not only for erosion control but also for stream channel maintenance, flood control, and water quality. For instance, in Brazil, two major advances in legislation were made: the concepts of mandatory reservation areas, a portion of the properties left without any intervention, and the special preservation areas, such as those along river banks and around reservoirs and lakes.

In the European Community, different policies contribute to soil protection, particularly environment (e.g., air and water) and agricultural (agri-environmental and cross-compliance) policy. Nine member states have specific legislation on soil protection. However, these laws often cover only one specific threat, such as soil contamination, and do not always provide a coherent framework.[47]

Many developing country governments provided direct assistance to farmers to adopt soil conservation measures.[4] Direct payment programs to encourage sustainability in agriculture due its benefits to society are still incipient in developing countries. One example is the Program for Water Quality and Quantity Improvement in Rural Catchments proposed by The Brazilian Agency for Water Resources.[48] It provides financial compensation for farmers, known as Water Productor Program, which “buys” the benefits (products) generated by the participant (concept of “provider-recipient”) where the payments are proportional to erosion abetment. The parameter of erosion abatement performance consists of the ratio of soil loss in the present condition to the worst-case condition (bare soil).

Policies to facilitate the development of an efficient marketing system may encourage switch from low-value crops to high-value crops and improve returns on investment to soil conservation measures. There is a need to have more market-based policies so that resources will be allocated more efficiently. On the other hand, policies such as import controls and high prices can also enhance soil erosion; for instance, global food policies such as food subsidies in the United States and the European Union and food aid lower international food prices. The extent to which these are transmitted into domestic food prices and international policies can have an impact on soil erosion.[4]

Integrated Environmental Management

How can we reach a sustainable strategy to control soil erosion? What are the roles of scientists, farmers, constructors, and managers? These questions stray from the strictly scientific arena (certainly from the physical sciences) into areas of social science and management. The more difficult issues are those of the proper relationship between science and managers (e.g., farmers and policy makers).[6]

Scientists played an important role in the development of problem perception and control. It was mainly the results from scientific studies that revealed the magnitude and severity of soil loss and sediment delivery to people working in the administration, and later to policy makers as well. One of the most important roles that the soil erosion research community has played is to contribute to monitoring the problem. Whereas the administration recently recognized the problem and even started with some control measures, they almost completely neglected the monitoring part. In the framework of the erosion control program, scientists also provided more information on the various soil conservation measures and sediment delivery control techniques that are available and the locations in the landscape where these measures need to be taken to be most effective.[49]

In the past, farmers were little concerned about soil erosion on their fields. Farmers were also not concerned about downstream consequences of soil erosion, like muddy floods in the villages.[49] The expansion and concentration of cultivated acreage led to extensive land-use change and increased concern about soil erosion impacts. In response to increasing soil erosion hazards, especially from agricultural land, interest groups and policy-driving organizations worldwide fostered the development and use of techniques for soil and water conservation.[43]

Conservation prioritization is an important consideration for planning natural resources management, allowing decision makers to implement management strategies that are more sustainable in the long term.[50] Soil and water protection have to be considered together; the connection of land at risk of soil erosion to rivers has to be taken into account as a potential source of water pollution.[47]

Conclusions

The analysis of soil water erosion reveals that a holistic view of the problem is necessary to ensure that the causes and factors affecting the problem are understood for any control measures to be undertaken. In order to assess the problem, it is important to have a strong database especially from field monitoring of soil losses and sediment transport at hillslope and watershed scales. These data can be used in mathematical models to simulate the impacts of land uses and climate changes. As monitoring is costly and most of the models required massive input database, visual indicators can be used to improve the models. In order to relate the rate of erosion and its effect, quantitative indicators such as soil loss tolerance have been used; however, it is important to consider not only productivity but also the off-site environmental impacts. Environmental legislation and public policies must induct farmers and constructors to use the soil in their properties bearing in mind the need to have environmental and social practices in their economical enterprises. Erosion control practices need to be part of an integrated environmental management that includes farmers, constructors, scientists, and managers to ensure soil and water preservation to the present and future generations.

References

1. Deng, Z.Q.; Lima, J.; Jung, H.S. Sediment transport rate-based model for rainfall induced soil erosion. Catena **2009**, *76* (1), 54–62.
2. Zhang, L.; O’Neill, A.L.; Lacey, S. Modelling approaches to the prediction of soil erosion in catchments. Environ. Software **1996**, *11* (1–3), 123–133.

3. Renschler, C.S.; Harbor, J. Soil erosion assessment tools from point to regional scales—The role of geomorphologists in land management research and implementation. Geomorphology **2002**, *47,* 189–209
4. Ananda, J.; Herath, G. Soil erosion in developing countries: A socioeconomic appraisal. J. Environ. Manage. **2003**, *68* (4), 343–353.
5. Faucette, L.B.; Jordan, C.F.; Risse, L.M.; Cabrera, M.; Coleman, D.C.; West, L.T. Evaluation of stormwater from compost and conventional erosion control practices in construction activities. J. Soil Water Conserv. **2005**, *60* (6), 288–297.
6. Boardman, J.; Poesen, J.; Evans, R. Socio-economic factors in soil erosion and conservation. Environ. Sci. Policy **2003**, *6*, 1–6.
7. Quine, T.A.; Walling, D.E.; Chakela, Q.K.; Mandiringana, O.T.; Zhang, X. Rates and patterns of tillage and water erosion on terraces and contour strips: Evidence from caesium-137 measurements. Catena **1999**, *36,* 115–142.
8. Poesen, J.W.; Boardman, J.; Wilcox, B.; Valentin, C. Water erosion monitoring and experimentation for global change studies. J. Soil Water Conserv. **1996**, *51* (5), 386–390.
9. Bilotta, G.S.; Brazier, R.E.; Haygarth, P.M; Macleod, C.J.A; Butler, P.; Granger, S.; Krueger, T; Freer, J; Quinton, J.N. Rethinking the contribution of drained and undrained grasslands to sediment-related water quality problems. J. Environ. Qual. **2008**. *37,* 906–914.
10. Morgan, R.P.C. *Soil Erosion and Conservation;* Blackwell: Oxford, 2005.
11. Owens, P.N.; Batalla, R.J.; Collins, A.J.; Gomez, B.; Hicks, D.M.; Horowitz, A.J.; Kondolf, G.M.; Marden, M.; Page, M.J.; Peacock, D.H.; Petticrew, E.L.; Salomons, W; Trustrum, N.A. Fine-grained sediment in river systems: Environmental significance and management issues. River Res. Appl. **2005**, *21* (7), 693–717.
12. Miller, R.W; Gardiner, D.T. *Soils in Our Environment,* 8th Ed.; Prentice-Hall: Upper Saddle River, NJ, 1998.
13. Foster, G.R. Modeling the erosion process. In *Hydrologic Modeling of Small Watersheds;* Haan, C.T., Johnson, H.P., Brakensiek, D.L., Eds.; American Society of Agricultural Engineers: St. Joseph, 1982; 296–380.
14. Aksoy, H.; Kavvas, M.L. A review of hillslope and watershed scale erosion and sediment transport models. Catena **2005**, *64,* 247–271.
15. Poesen, J.; Nachtergaele, J.; Verstraeten. G.; Valentin, C. Gully erosion and environmental change: Importance and research needs. Catena **2003**, *50*, 91–133.
16. Boardman, J. Soil erosion science: Reflections on the limitations of current approaches. Catena **2006**, *68,* 73–86.
17. Agassi, M.; Bradford, J.M. Methodologies for interrill soil erosion studies. Soil Tillage Res. **1999**, *49,* 277–287.
18. FAO. *Field measurement of soil erosion and runoff.* FAO Soils Bulletin 68, by Hudson, N.W. 1993.
19. Zachar, D. *Soil erosion: Developments in soil science.* Elsevier Scientific: New York, 1982; 547.
20. Wischmeier, W.H.; Smith, D.D. *Predicting rainfall erosion losses; a guide to conservation planning.* U.S. Department of Agriculture: Washington, DC, 1978; 58.
21. Edwards, T.K.; Glysson, G.D. *Field methods for measurement of fluvial sediment.* U.S. Geological Survey: Reston, 1999; 89.
22. Minella, J.P.G; Walling, D.; Merten, G. Combining sediment source tracing techniques with traditional monitoring to assess the impact of improved land management on catchment sediment yields. J. Hydrol. **2008**, *348*, 546–563.
23. Walling, D.E. Linking land use, erosion and sediment yields in river basins. Hydrobiologia **1999**, *410,* 223–240.
24. Bryan, R.B. Soil erodibility and processes of water erosion on hillslope. Geomorphology **2000**, *32,* 385–415.

25. Merritt, W.S.; Latcher, R.A.; Jakeman, A.J. A review of erosion and sediment transport models. Environ. Modell. Software **2003**, *18,* 761–799.
26. Jetten, V.; Roo, A.; Favis-Mortlock, D. Evaluation of field-scale and catchment-scale soil erosion models. Catena **1997**, *37,* 521–541.
27. Wischmeier, W.H. Use and misuse of the Universal Soil Loss Equation. J. Soil Water Conserv. **1976**, *31,* 5–9.
28. Kinnell, P.I.A. Event soil loss, runoff and the Universal Soil Loss Equation family of models: A review. J. Hydrol. **2010**, *385,* 384–397.
29. Williams, J.R.; Berndt, H.D. Sediment yield prediction based on watershed hydrology. Trans. ASAE **1977**, *20* (6), 1100–1104.
30. Renard, K.G.; Foster, G.R.; Weesies, G.A.; McCool, D.K.; Yoder, D.C. Predicting soil erosion by water: A guide to conservation planning with the Revised Universal Soil Loss Equation (RUSLE). U.S. Department of Agriculture Agricultural Handbook. No. 703. US Department of Agriculture: Washington, DC, 1997.
31. Beasley, D.B.; Huggins, L.F.; Monke, E.J. ANSWERS: A model for watershed planning. Trans. ASAE **1980**, *23,* 938–944.
32. Young, R.A.; Onstad, C.A.; Bosch, D.D.; Anderson, W.P. AGNPS: A nonpoint-source pollution model for evaluating agricultural watersheds. J. Soil Water Conserv. **1989**, *44*, 168–173.
33. Nearing, M.A.; Foster, G.R.; Lane, L.J.; Finkner, S.C. A process-based soil erosion model for USDA-Water Erosion Prediction Project technology. Trans. ASAE **1989**, *32* (5), 1587–1593.
34. Misra, R.K.; Rose, C.W. Application and sensitivity analysis of process-based erosion model GUEST. Eur. J. Soil Sci. **1996**, *47,* 593–604.
35. De Roo, A.P.J., Wesseling, C.G., Ritsema, C.J. LISEM: A single-event physically based hydrological and soil erosion model for drainage basins: I. Theory, input and output. Hydrol. Processes **1996**, *10,* 1107–1117.
36. Morgan, R.P.C.; Quinton, J.N.; Smith, R.E.; Govers, G.; Poesen, J.W.A.; Auerswald, K.; Chisci, G.; Torri, D.; Styc-zen, M.E. The European Soil Erosion Model (EUROSEM): A dynamic approach for predicting sediment transport from fields and small catchments. Earth Surf. Processes Landforms **1998**, *23*, 527–544.
37. Arnold, J.G.; Fohrer, N. SWAT2000: Current capabilities and research opportunities in applied watershed modeling. Hydrol. Processes **2005**, *19,* 563–572.
38. Bie, C.A.J.M. Assessment of soil erosion indicators for maize-based agro-ecosystems in Kenya. Catena **2005**, *59,* 231–251.
39. USDA-SCS. Advisory notice, Soils-6, Washington, DC, 1973.
40. Hall, G.F.; Logan, T.J.; Young, K.K. Criteria for determining tolerable erosion rates. In *Soil Erosion and Crop Productivity;* Stewart, B.A; Follet, R.F., Eds. American Society of Agronomy: Madison, 1985; 213–234.
41. Verheijen, F.G.A.; Jones; R.J.A.; Rickson, R.J.; Smith, C.J. Tolerable versus actual soil erosion rates in Europe. Earth-Sci. Rev. **2009**, *94*, 23–38.
42. Troeh, F.R.; Hobbs, J.A.; Donahue, R.L. *Soil and Water Conservation: Productivity and Environmental Protection*, 3rd Ed.; Prentice-Hall: Upper Saddle River, NJ, 1999.
43. Reichert, J.M.; Cassol, E.A. Erosion and sedimentation control: Engineering techniques. In *Encyclopedia of Soil Science,* 2nd Ed.; Lal, R., Ed.; Marcel Dekker: New York, 2006; online version, 548–552.
44. Hudson, N. *Soil Conservation,* 2nd Ed.; Cornell University Press: Ithaca, 1995.
45. Schwab, G.O.; Fangmeier, D.D.; Elliot, W.J. *Soil and Water Conservation Engineering,* 4th Ed.; Wiley: New York, 1992.
46. Manual on Integrated Soil Management and Conservation Practices, IITA and FAO: Rome, 2000.
47. Volk, M.; Möller, M. Wurbs, D. A pragmatic approach for soil erosion risk assessment within policy hierarchies. Land Use Policy **2010**, *27,* 997–1009.

48. ANA. Agência Nacional de Águas. O Programa do Produtor de Água, available at http://www.ana.gov.br/gestaoRecHidricos/ConservacaoAguaSolo/default2.asp (accessed October 2, 2008).
49. Verstraeten, G.; Poesen, J.; Govers, G.; Gillijns, K.; Rom-paey, A.V.; Oost, K.V. Integrating science, policy and farmers to reduce soil loss and sediment delivery in Flanders, Belgium. Environ. Sci. Policy **2003**, *6*, 95–103.
50. Zhang, X.; Wu, B.; Ling, F.; Zeng, Y.; Yan, N.; Yuan, C. Identification of priority areas for controlling soil erosion. Catena **2010**, *83,* 76–86.

35

Erosion by Water: Empirical Methods

John M. Laflen

Introduction

Assessment of soil erosion requires some estimate of mass, volume, depth, and/or location of the source and/or deposition of eroded soil material and the frequency with which it occurs. Assessment may also require an estimate of the physical, chemical, and/or biological characteristics of eroded materials. Moreover, assessment usually includes some estimate of the impact of management and treatment on soil erosion. Usually, these estimates are made using a model. In this entry, the focus will be on empirical models to estimate the mass of soil erosion by water.

Why use models? Erosion is highly variable in time and space. Figure 1 shows annual sediment yields from a small western Iowa watershed cropped continuously to corn over a 24 years period. The coefficient of variation [CV (standard deviation/mean)] of annual values is 135%, while the CV for annual runoff is 85% and the CV for annual precipitation is only 20%. While annual values are shown in Figure 1, at Columbia, Missouri, USA, over an 11 yr period for continuous corn or continuous soybeans, 50% of the erosion occurred in a single storm, and 80% occurred in only seven storms.[1] Others have also shown that a major portion of sediment yield occurs in only a few storms.[2] This extreme variability in soil erosion makes it virtually impossible to make meaningful measurements of erosion for assessment purposes in a feasible period of time. Hence, the use of models for such purposes is required.

An empirical model is a representation of data.[3] The empirical models described here include variables generally based on our understandings of erosion processes, with the coefficients and relationships based on data. Generally, because they are based on data, empirical models give reliable answers for situations encountered in the data used for development. However, answers for other situations or outside the range of the experimental data are less reliable and, in fact, could be contrary to that expected.

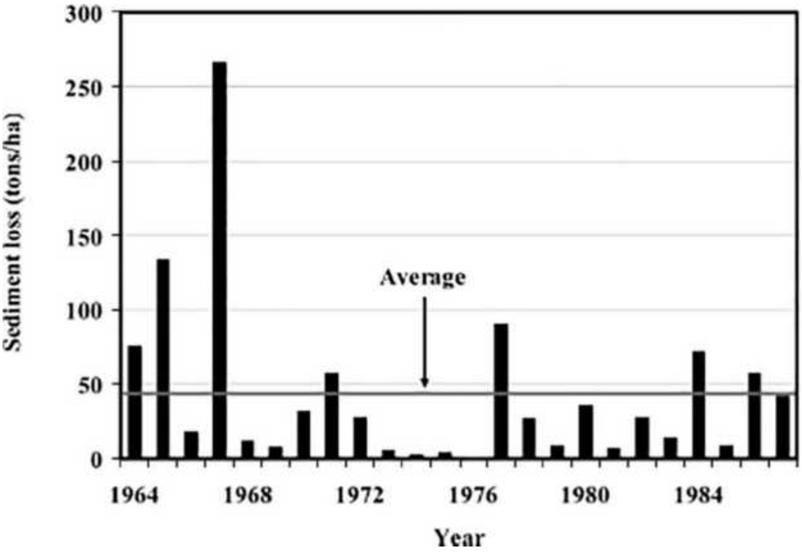


FIGURE 1 Annual sediment loss from a small watershed near Treynor, Iowa, USA, for 1964–1987 (unpublished data). The CV for annual sediment loss values was 135%, while the CV for the annual runoff was 85% and the CV for annual precipitation was 20%.

Empirical models can be easy to use and parameterize, an important factor in model selection. Equally important is that users may find them easier to understand than more sophisticated models. However, sophisticated models that are process based may find a wider range of use, a particularly important factor as needs for erosion assessment and other associated environmental assessments become more complex.

Models described here are "field scale" models for the most part. Models for large watersheds are not included, although some empirical models are components of some watershed models.

Empirical Erosion Model Development

Empirical model development began with experiments to understand the effect of various factors—climate, topography, soil, management, and conservation practices—on soil erosion. Early models considered a limited set of factors; these were expanded as the number of workers in the field increased the amount of information available on which to expand the models. As the science progressed, the complexity increased, with relationships of factor values to factor characteristics being developed based on experimental data.

The major factor in the development of empirical models was establishment of erosion plots at numerous locations in the United States.[4] These plots provided an immense data set that is still used today for evaluating soil erosion and hydrologic models. More than 10,000 plot years of data were collected on these plots.[5] Additionally, many plots were located on experimental farms that contained small watersheds on which conservation practices were installed, monitored, and evaluated. These experimental areas provided the necessary data for developing the relationships in water erosion empirical models.

Erosion plots were also established in other regions of the world.[6] Erosion plots can be found on every continent except Antarctica. They are frequently used to evaluate the applicability of empirical and non-empirical erosion models to specific regions, and to develop appropriate factor values for practices common to those regions.

TABLE 1 Evolution of U.S. Empirical Water Erosion Models

	Soil Loss	Coefficient or Rainfall Factor	Soil Erodibility Factor	Length–Slope Factor	Cropping and Management Factor	Conservation Practice Factor
Zingg 1940[31]	A	C'		$L^{0.6}\ S^{1.4}$		
Smith 1941[32]	A	C''		$L^{0.6}\ S^{1.4}$		P
Browning 1947[33]	A	*Cò*	K'	$L^{0.6}\ S^{1.4}$		P
Musgrave 1947[34]	A'	$P_{30}{}^{1.75}$	K''	$L^{0.35}\ S^{1.35}$	C^*	P
USLE 1961[9]	A	EI_{30}	K	$(L/72.6)^{0.5}(0.065+0.045\ S+0.0065\ S^2)$	C	P
USLE 1978[10]	A	EI_{30}	K	$(L/72.6)^{0.5}(65.4\ \sin^2\Theta+4.56\ \sin\ \Theta+0.065)$	C	P
MUSLE 1975[14]	A^*	$95(Qq_p)^{0.56}$	K	$(\mathrm{L}/72.6)^{0.5}(65.4\ \sin^2\Theta+4.56\ \sin\Theta+0.065)$	C	P
RUSLE 1997[25]	A	EI_{30}	K	$(\mathrm{L}/72.6)^{\mathrm{m}}(a\ \sin\Theta+b)$	C	P

A—soil loss (tons/acre). A'—soil loss (inches/year). A^*—sediment delivery (tons/year).
P30—maximum precipitation amount (inches) falling in 30 min in a storm.
Q—storm runoff volume (acre-feet). qp—peak runoff rate (cubic feet per second).
Qo—storm runoff volume (inches). qo—peak runoff rate (inches per hour).
E—storm rainfall energy (hundreds of foot-tons per acre).
I30—maximum rainfall intensity in a 30 min period within a storm (inches per hour).
K', K'', K—soil erodibility factors.
L—slope length (feet). S—slope (%). Θ—slope angle (degrees).
m—exponent on length term (values depend on slope or slope and rill/inter-rill ratio).
a, b—coefficients in function making up slope term (values depend on slope).
C', C'', *Cò*—coefficients. C^*—vegetal cover factor. C—cropping and management factor (before 1978) and cover and management factor (after 1978).
P—conservation practice factor.

The development of empirical models has been a classic example of the operation of science and its application to solving real-world problems. The steady progression of science is illustrated in Table 1, where the steps in the development of U.S. empirical erosion modeling are given chronologically. The work initiated in 1940 in the United States continues today, each step built on top of earlier work.[4,7]

Empirical Models

Universal Soil Loss Equation

The Universal Soil Loss Equation (USLE) is written as

$$\mathrm{A} = \mathrm{RKLSCP},$$

where A is soil loss per unit area (units depend on units used for R and K); R is the rainfall and runoff factor, the number of rainfall erosion index units, plus factors for runoff from snowmelt or applied water; K is the soil erodibility factor, the soil loss rate per rainfall erosion index unit on a plot 72.6 ft long of uniform 9% slope maintained in a clean tilled fallow; L is the slope-length factor, the ratio of soil loss for a given slope length to a 72.6 ft long slope; S is the slope-steepness factor, the ratio of soil loss for a given slope to soil loss from a 9% slope; C is the cover and management factor, the ratio of soil loss for a given land use and management to soil loss from a continuous clean tilled fallow; and P is the support practice factor, the ratio of soil loss with a support practice, such as contouring, terracing, or stripcropping, to soil loss for up and down the slope. Much more detail on factor values and assistance in computing these values is available elsewhere.[5]

The USLE was the culmination of more than two decades of work in developing a national model of soil erosion by water.[4] As shown in Table 1, the form and factors were long identified before the USLE was developed. The defining scientific finding in the development of the USLE was that a single rainfall variable was selected that could be used over the United States to model soil erosion from rainstorms.[8] This allowed the application of factor values in one region to be used in another region. Climatic databases were developed for the entire United States, and tables and charts were published to allow use of the USLE for most U.S. conditions.[5,9,10] It is being replaced by the Revised Universal Soil Loss Equation (RUSLE)[11] in the United States.

The USLE technology was the water erosion model of choice over much of the world through the last century, with major efforts in developing *R* values for the areas to which it would be applied. It served as the platform from which a number of other empirical technologies evolved.

Although the USLE is a good predictor of average annual erosion, it is a poor predictor of individual storm soil erosion. It does not contain variables associated with the hydrologic condition of the soil at the time the storm occurs. Additionally, the USLE does not consider deposition on a field and does not predict sediment yield, a needed estimate for many applications. The USLE's area of application is from the top of a hillslope (or divide) to where deposition begins, or runoff enters a channel.

Modified Universal Soil Loss Equation

The Modified Universal Soil Loss Equation (MUSLE) is written as

$$A^{*} = 95\left(Qq_{p}\right)^{0.56} KLSCP,$$

where A* is sediment delivery; *Q* is the total volume of runoff; q_p is the peak rate of surface runoff; and *K, L, S, C*, and *P* are the same as for the USLE.

MUSLE was developed to predict individual storm sediment yield, something that the USLE did not do. For individual storms, sediment concentration for five small watersheds near Riesel, Texas was highly correlated with the ratio of peak runoff rate per square mile to total volume of runoff.[12] The USLE was used to predict average annual sediment yield for the same five small watersheds, using a delivery ratio based on watershed characteristics.[13] The various factors in the USLE (except for the rainfall factor) were weighted based on drainage area of the various factor values, channel lengths, channel slopes, and cultivated area.

Data from 18 watersheds were used to examine three forms of a runoff factor.[14] The runoff factor $95(Qq_p)^{0.56}$ (Q—volume of runoff in acre-feet, q_p—peak flow rate in cubic feet per second) was the best form for estimating individual storm sediment yield from small watersheds. The 18 watersheds ranged in size from 2.7 to 4380 acres. Two watersheds were located at Hastings, Nebraska, the remainder were located at Riesel, Texas, and included the five small watersheds used in earlier analyses.

MUSLE requires inputs, storm volume of runoff, and peak rate of runoff, not used in the USLE. Simulation models can generate this information or it can be estimated for design storms.

MUSLE continues to be heavily used in various modeling activities. It is an option in the EPIC (Erosion Productivity Impact Calculator) model[15] for computing sediment yield. While MUSLE is widely used, there are few publications that detail the accuracy of MUSLE, outside of the original work in its development.

Soil Loss Estimation Model for Southern Africa

The Soil Loss Estimation Model for Southern Africa (SLEMSA) was introduced as a reliable management tool developed from limited resources.[16] The development team evaluated the USLE for adoption in Southern Africa but was concerned about adopting factor values developed in America.[17] Additionally, there was considerable concern about differences in farming practices. They felt that they did not have the resources to implement the USLE, and a simpler approach was needed.

SLEMSA is written as

$$Z = KCX,$$

where Z is the mean annual soil loss, K is the mean annual soil loss from a standard field plot 30 m × 10 m at 4.5% slope for a soil of known erodibility under a weedfree bare fallow, C is the ratio of soil lost from a cropped plot to that lost from bare fallow land, and X is the ratio of soil lost from a plot length L and slope percent S to that lost from the standard plot.

K values are computed as

$$\mathrm{Ln}\, K = b \ln E + a,$$

where E is seasonal rainfall energy, and a and b are given by

$$a = 2.884 - 8.1209F$$

$$b = 0.4681 - 0.7663F,$$

where F is soil erodibility. Based on soil texture, F is assigned a value between 4 (light textured, or sandy soil) and 6 (heavy textured, or clay soil). Additional adjustments are made based on hydrologic characteristics of the soil, tillage, soil structure, and crusting, giving a range of F values from 1 to 9.

C is computed as $C = e^{(-0.06i}$) when $\mathrm{i} < 50\%$ and as $\mathrm{C} = (2.3 - 0.01\mathrm{i})/30$ when $\mathrm{i} \geq 50\%$, where i is the percentage rainfall energy intercepted by the crop (used crop cover %).

X is computed as

$$X = (l)^{1/2}\left(0.76 + 0.53_S + 0.076s^2\right)/25.65$$

This is derived from the USLE LS factor in use in the 1970s. The equation above converted the USLE LS factor to the metric system and adjusted it to a 4.5%, 30 m long slope. The variables l and s are length (m) and slope (%), respectively.

The USLE and SLEMSA are very similar in many respects. The major differences are in the climate variables and the cropping-management approaches. The USLE climate factor contains both rainfall energy and rainfall intensity where SLEMSA contains only rainfall energy. The USLE C value incorporates both tillage and cropping while the C value for SLEMSA is a function of the interception of rainfall energy by crops. However, SLEMSA does capture some of the tillage component in the F value.

ABAG (German USLE)

The USLE was applied almost intact for German conditions.[18] Experimentally, most of the USLE factors were validated, and few modifications were needed to apply it under German conditions. One modification was that soils highly enriched in potassium were found to be more erodible than expected, based on the soil erodibility nomograph of Wischmeier et al.[19]

While no modification of the USLE was found necessary, ABAG factors were necessary for German conditions, specifically the R and C factors. A book detailing the use of the USLE for extension workers in Germany has been published.[20]

Soiloss

A computer program (SOILOSS), which used the USLE approach to estimate average annual soil losses and gave recommendations on ways to reduce soil loss through changes in land and crop management practices, was developed.[21] SOILOSS was developed for use in New South Wales, Australia. It used a slightly different equation for computing rainfall energy than that used in the USLE. It also used both the USLE nomograph[19] for soil erodibility, plus experimentally derived values for New South Wales. In a later revision,[22] equations to compute *LS* values were adopted from RUSLE.[11] From the earliest version of SOILOSS, the subfactor approach of Laflen et al.[23] that was implemented in RUSLE[11] was used to estimate the *C* value in SOILOSS. It contained *P* value estimates similar to those in the USLE,[5] but some *P* value estimates were from experimental work in New South Wales. SOILOSS also included a benefit from sediment deposited in bank (terrace) channels.

A SOILOSS user guide [24] was developed for high school students. It included the program, an introduction to the USLE, complete parameterization for most Australian conditions, and worksheets and examples.

In a number of ways, SOILOSS could be considered an early RUSLE, using much of the science and approaches that were eventually used in RUSLE. It was also developed in the time frame where RUSLE was developed.

Revised Universal Soil Loss Equation

RUSLE[25] maintains the form of the USLE in terms of the factors, but since it is operated on a computer, many improvements in computations, particularly those including interactions, have been incorporated. Major improvements in the RUSLE include an improved *R* value map for the United States, time variation in soil erodibility, an improved *LS* factor that incorporates recent science, a subfactor approach for computing the cover and management factor, and *P* values based on modeling analyses that include storm severity, ridge height, drainage, and off-grade contouring.

Development of RUSLE has continued. The latest version is named RUSLE2 and it is used by the Natural Resource Conservation Service of the United States Department of Agriculture (USDA) for estimating soil erosion on U.S. land.[26] RUSLE improvements include those in science, ease of use, supporting databases, and applications related to sediment deposition and delivery.

Accuracy of Erosion Predictions

A knowledge of the accuracy expected in an assessment of soil erosion is critical in many applications. For soil erosion runoff plots, the CV for replicated plots was found to be a function of the value of the measured soil erosion (M),[27–29] and the relationship could be expressed as

$$\mathrm{CV} = 0.73M^{-0.306} \tag{1}$$

where M is expressed in tons per hectare. The relationship held for erosion for all time periods.

Eq. 1 was used to derive a relationship between measured erosion and confidence intervals (CIs) for measured erosion. This relationship is

$$\mathrm{CI}(tons\,/\,ha) = t_{(\alpha+1)/2} M\mathrm{CV}, \tag{2}$$

where t is the cumulative t distribution value for $(\alpha+1)/2$ for an infinite number of points, and α is the probability level selected. For a 95% CI, and substituting Eq. 1 for CV in Eq. 2, CI can be written as

$$\mathrm{CI}_{95} = 1.43\, M^{0.694}, \tag{3}$$

and lower and upper bounds (LB and UB) for the 95% CI (Figure 2) are given by

$$\mathrm{LB}(\mathrm{tons/ha}) = M - \mathrm{CI}_{95} \tag{4}$$

$$\mathrm{UB}(\mathrm{tons/ha}) = M + \mathrm{CI}_{95} \tag{5}$$

Table 2 shows measured and USLE-predicted soil loss from erosion plots at 20 locations in the United States.[30] These data had been used in the development of the USLE. There were multiple plots and treatments at most locations. The same data are shown in Figure 2, plus the 95% CIs computed using Eqs. 3, 4, and 5. On the average, soil erosion was overpredicted by 22%, and 6 of the 20 soil erosion values predicted fell outside the 95% CI.

The UB and LB on the CI shown in Figure 2 are based on variability in measured data. It would be unrealistic to expect better estimates from any model of soil erosion—empirical or otherwise.[35]

TABLE 2 Comparison of Average Annual Measured and Predicted Soil Loss from Erosion Plots at 20 U.S. Locations

Site	Record Length Plot Years	Measured Soil Loss (tons/ha)	Predicted Soil Loss (tons/ha)*
Hays, Kansas	88	3.1	6.7
Tifton, Georgia	64	3.6	7.6
Ithaca, New York	79	6.5	9.1
Raleigh, North Carolina	10	7.1	25
Madison, South Dakota	72	17.1	12
Morris, Minnesota	40	18	18.8
Presque Isle, Maine	45	19.9	15
Dixon Springs, Illinois	96	20.9	20.5
Guthrie, Oklahoma	153	22.6	28.5
Geneva, New York	57	22.9	20.8
Marcellus, New York	79	24	32.3
Temple, Texas	105	28.8	26.2
Watkinsville, Georgia	119	32.1	28.8
Statesville, North Carolina	72	54.1	119.9
Clarinda, Iowa	117	55	47.2
Bethany, Missouri	90	57.7	23.8
Clemson, South Carolina	6	57.9	81.8
La Crosse, Wisconsin	234	66	54.4
Castana, Iowa	44	76.5	145.8
HollySprings, Mississippi	24	88.8	109.7
Average	79.7	34.1	41.7

Predictions were made by the USLE.

Source: Tiwari et al.[30]

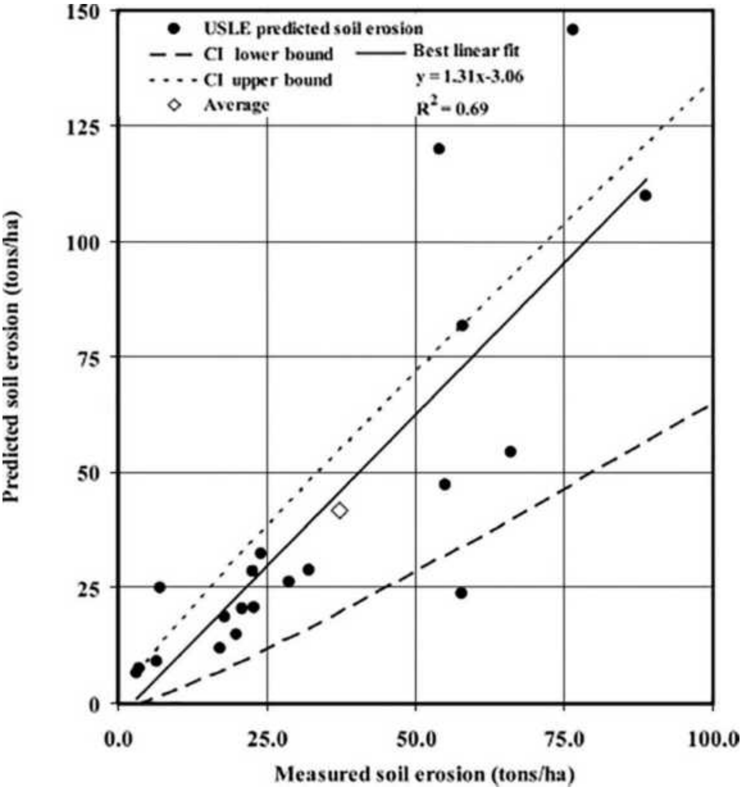


FIGURE 2 Measured versus predicted soil erosion from runoff plots. Also shown are CIs and the mean measured and predicted soil erosion. CIs (95%) for measured erosion, based on the work of Nearing and others[27–29] and of measured and predicted data from Tiwari et al.[30]

Conclusion

Soil erosion is highly variable with a large proportion of soil erosion occurring in only a few storms.

The USLE technology has been adapted to apply to most of the earth's land surface. Nearly all empirical technology for erosion prediction is based in large part on the USLE technology, and the development of the USLE was clearly an evolutionary process.

Soil erosion is highly variable. Measured erosion on replicated plots has high CV; it should be expected that erosion assessments based upon models (or measurements) will have similar levels of predictability.

References

1. Ghidey, F.; Alberts, E.E. Comparison of measured and WEPP predicted runoff and soil loss for Midwest claypan soil. Trans. Am. Soc. Agric. Eng. **1996**, *39* (4), 1395–1402.
2. Lane, L.J.; Wigmosta, M.S.; Hernandez, M.; Coleman, A. Extreme events and stochastic processes of sediment yield. Proceedings of the World Environmental and Water Resources Congress, American Society of Civil Engineers, 2008.
3. Haan, C.T. *Statistical Methods in Hydrology;* Iowa State Press: Ames, Iowa, 1977; 394 pp.

4. Meyer, L.D. Evolution of the Universal Soil Loss Equation. J. Soil Water Conserv. **1984**, *39* (2), 99–104.
5. Wischmeier, W.H.; Smith, D.D. *Predicting rainfallerosion losses—A guide to conservation farming*, Agricultural Handbook No. 537; U.S. Department of Agriculture: Washington, DC, 1978; 58 pp.
6. Hudson, N.W. Erosion control research. Rhod. Agric. J. **1957**, *54* (4), 297–323.
7. Laflen, J.M.; Moldenhauer, W.C. *Pioneering Soil Erosion Prediction—The USLE Story*; World Association of Soil and Water Conservation, Special Publication 1, 2003.
8. Wischmeier, W.H.; Smith, D.D. Rainfall energy and its relationship to soil loss. Trans. Am. Geophys. Union **1958**, *39* (2), 285–291.
9. Wischmeier, W.H.; Smith, D.D. *A universal equation for predicting rainfall-erosion losses—An aid to conservation farming in humid regions*, ARS Special Report 22–66; Agricultural Research Service, U.S. Department of Agriculture: Washington, DC, 1961; 11 pp.
10. Wischmeier, W.H.; Smith, D.D. Predicting rainfall-erosion losses from cropland east of the Rocky Mountains—Guide for selection of practices for soil and water conservation, Agricultural Handbook No. 282; U.S. Department of Agriculture: Washington, DC, 1965; 47 pp.
11. Renard, K.G.; Foster, G.R.; Weesies, G.A.; Porter, J.P. RUSLE: Revised Universal Soil Loss Equation. J. Soil Water Conserv. **1991**, *46* (1), 30–33.
12. Williams, J.R.; Hiler, E.A.; Baird, R.W. Prediction of sediment yields from small watersheds. Trans. Am. Soc. Agric. Eng. **1971**, *14* (6), 1157–1162.
13. Williams, J.R.; Berndt, H.D. Sediment yield computed with universal equation. J. Hydraul. Div., Am. Soc. Civ. Eng. **1972**, *98* (HY12), 2087–2098.
14. Williams, J.R. 1975. Sediment yield prediction with universal equation using runoff energy factor. In *Present and Prospective Technology for Predicting Sediment Yields and Sources*, Sediment-Yield Workshop, USDA Sedimentation Laboratory: Oxford, Mississippi, November 28–30, 1975; ARS-S-40, Agricultural Research Service, U.S. Department of Agriculture: Washington, DC, 1975; 244–252.
15. Sharpley, A.N.; Williams, J.R. *EPIC—Erosion/Productivity Impact Calculator: 1. Model Documentation*, Technical Bulletin No. 1768; U.S. Department of Agriculture: Washington, DC, 1990; 235 pp.
16. Elwell, H.A. Modeling soil losses in Zimbabwe Rhodesia. American Society of Agricultural Engineers Paper Number 79–2051; American Society of Agricultural Engineers: St. Joseph, MI, 1979.
17. Elwell, H.A. Modeling soil losses in Southern Africa. J. Agric. Eng. Res. **1978**, *23*, 117–127.
18. Auerswald, K. 2000. Empirical soil erosion models. Personal letter, 4–12–2000.
19. Wischmeier, W.H.; Johnson, C.B.; Cross, B.V. A soil erodibility nomograph for farmland and construction sites. J. Soil Water Conserv. **1971**, *26* (5), 189–193.
20. Schwertmann, U.; Vogl, W.; Kainz, M. *Bodenerosion durch Wasser-Vorhersage des Abtrags und Bewertung von Gegenmaznahmen*; Ulmer Verlag: Stuttgart, Germany, 1987; 64 pp.
21. Rosewell, C.J.; Edwards, K. *SOILOSS—A program to assist in the selection of management practices to reduce erosion*, Technical Handbook No. 11; Soil Conservation Service of New South Wales: Sydney, NSW, 1988; 71 pp.
22. Rosewell, C.J. *SOILOSS—A program to assist in the selection of management practices to reduce erosion*, 2nd Ed.; Technical Handbook No. 11; Soil Conservation Service of New South Wales: Sydney, NSW, 1993; 108 pp.
23. Laflen, J.M.; Foster, G.R.; Onstad, C.A. Simulation of individual storm soil loss for modeling impact of soil erosion on crop productivity. In *Soil Erosion and Conservation*; El-Swaify, S.A., Moldenhauer, W.C., Lo, A., Eds.; Soil Conservation Society of America: Ankeny, IA, 1985; 285–295.
24. Keats, J.; Rosewell, C. *SOILOSS—High School User Guide;* Department of Conservation and Land Management, NSW: Australia, 87 pp.

25. Renard, K.G.; Foster, G.R.; Weesies, G.A.; McCool, D.K.; Yoder, D.C. *Predicting soil erosion by water: A guide to conservation planning with the Revised Universal Soil Loss Equation (RUSLE)*, Agricultural Handbook No. 703; U.S. Department of Agriculture: Washington, DC, 1997; 384 pp.
26. U.S. Department of Agriculture, National Soil Erosion Research Laboratory. Revised universal soil loss equation, version 2. Official NRCS RUSLE2 Program, available at http://www.ars.usda.gov/Research/docs.htm?docid=18095.
27. Nearing, M.A. Why soil erosion models over-predict small soil losses and under-predict large soil losses. Catena **1998**, *32,* 15–22.
28. Nearing, M.A. Evaluating soil erosion models using measured plot data: Accounting for variability in the data. Earth Surf. Processes Landforms **2000**, *25,* 1035–1043.
29. Nearing, M.A.; Govers, G.; Norton, L.D. Variability in soil erosion data from replicated plots. Soil Sci. Soc. Am. J. **1999**, *63,* 1829–1835.
30. Tiwari, A.K.; Risse, L.M.; Nearing, M.A. Evaluation of WEPP and its comparison with USLE and RUSLE. Trans. Am. Soc. Agric. Eng. **2000**, *43* (5), 1129–1135.
31. Zingg, A.W. Degree and length of land slope as it affects soil loss in runoff. Agric. Eng. **1940**, *21* (2), 59–64.
32. Smith, D.D. Interpretation of soil conservation data for field use. Agric. Eng. **1941**, 22 (5), 173–175.
33. Browning, G.M.; Parish, C.L.; Glass, J. A method for determining the use and limitations of rotation and conservation practices in the control of soil erosion in Iowa. J. Am. Soc. Agron. **1947**, *39* (1), 65–73.
34. Musgrave, G.W. The quantitative evaluation of factors in water erosion—A first approximation. J. Soil Water Conserv. **1947**, 2 (3), 133–138, 170.

36

Erosion by Water: Process-Based Modeling

Mark A. Nearing

Introduction

Historically models of soil erosion by water have been primarily either empirically based or process-based (sometime referred to as physically based). The first models of soil erosion were empirically based. The prime example of the empirically based model is the Universal Soil Loss Equation (USLE).[1–2] More recent models have been based on equations which describe the physical, biological, and/or chemical processes which cause or affect soil erosion. It is important to understand that both the process-based and the empirically based models possess a major empirical component, in the sense that the constitutive equations use parameters based on experimental data.

Process-Based Models

Process-based erosion models address soil erosion on a relatively fundamental level using mass balance differential equations for describing sediment continuity on a land surface. The fundamental equation for mass balance of sediment in a single direction on a hillslope profile is given as:

$$\partial(cq)/\partial x + \partial(ch)/\partial t + S = 0 \tag{1}$$

where c (kg/m^3) is sediment concentration, q (m^{2}s) is unit discharge of runoff, h (m) is depth of flow, x (m) is distance in the direction of flow, t (s) is time, and S [kg/(m^{2}s)] is the source/sink term for sediment generation. Eq. 1 is an exact one-dimensional equation. It is the starting point for development of all existing process-based models. The differences in various erosion models are primarily: a) whether the partial differential with respect to time is included; and b) differing representations of the source/sink term, S. If the partial differential term with respect to time is dropped then the equation is solved for the steady state, whereas the representation of the full partial equation represents a fully dynamic model. The source/sink term for sediment, S, is generally the greatest source of differences in soil erosion models. It is this term that may contain elements for soil detachment, transport capacity terms, and sediment deposition functions. It is through the source/sink term of the equation that empirical relationships and parameters are introduced.

One of the earliest process-based erosion models was presented by Meyer and Wischmeier.[3] This model considered soil detachment by rain and by runoff, and then compared the sediment load generated from this detachment to the sediment transport capacity of the flow. If sediment load exceeded sediment transport capacity, then deposition was calculated. The ideas from this early work were expanded and revised by Foster and Meyer,[4] and ultimately were worked into the Chemicals, Runoff, and Erosion from Agricultural Systems (CREAMS) model,[5] which is a hybrid model in the sense that it uses process-based equations with parameters from the USLE. Experience from developing CREAMS then led to the development of the fully process-based Water Erosion Prediction Project (WEPP) model.[6]

Continuous Simulation Models

Process-based models for soil erosion by water may be either continuous simulation or event models. Continuous simulation models predict soil erosion for series of individual storms, and have auxiliary components for updating system parameters between storms. Event models simply predict erosion for individual storm events, but require the user to provide system information for each event. As such, the continuous simulation model is useful for analyzing the effects of crop management systems on erosion, whereas the event model is relatively limited in that regard.

The Water Erosion Prediction Project (WEPP) model is an example of a continuous simulation model. WEPP includes nine major components: climate, infiltration, water balance, winter process and snowmelt erosion, plant growth and residue decomposition, irrigation, surface runoff, rainfall erosion, and channel routing for watersheds.[6] System parameters that are updated on a daily basis within the continuous simulation include soil moisture, soil density, soil resistance to erosion, plant canopy, surface residue, soil surface roughness, soil surface sealing and crusting, buried residue amounts, frozen soil layers, snow cover, and roots. All of these system parameters define the antecedent conditions for each erosion event.

The WEPP climate model has been tested for erosion and well parameterized for the United States.[7] The infiltration component of the hillslope model is based on the Green and Ampt equation as modified by Mein and Larson,[8] with the ponding time calculation for an unsteady rainfall.[9] The water balance and percolation component of the profile model is based on the water balance component of SWRRB (Simulator for Water Resources in Rural Basins),[10,11] with some modifications for improving estimation of percolation and soil evaporation parameters. The plant growth component of the model simulates plant growth and residue decomposition for cropland and rangeland conditions. The residue and root decomposition model simulates decomposition of surface residue (both standing and flat), buried residue, and roots for the annual crops specified in the WEPP User Requirements[12] plus perennial crops of alfalfa and grasses. Surface runoff is calculated using a kinematic wave equation. Flow is partitioned into broad sheet flow for interrill erosion calculations and concentrated flow for rill erosion calculations. The erosion component of the model uses a steady-state sediment continuity equation that calculates net values of detachment or deposition rates along the hillslope profile.[13] The erosion process is divided into rill and interrill components where the interrill areas act as sediment feeds to the rills, or small channel flows. The model is applicable to hillslopes and small watersheds.

Because the model is based on all of the processes described above, and more, it is possible with a continuous simulation, process-based model to have an enormous array of possible system interactions represented in the simulations. Just to name a very few examples, slope length and steepness effects are functions of soil consolidation, surface sealing, ground residue cover, canopy cover, soil water content, crop type and many other factors. Ground residue cover is a function of biomass production rates, tillage implement types, residue type, soil moisture, temperature and solar radiation, previous rainfall, and many other factors. Rill erosion rates are a function of soil surface roughness, ground cover, consolidation of the soil, soil physical and chemical properties, organic matter, roots, interrill erosion rates, slope, and runoff rates, among other factors. The lists continue ad infinitum. These are interactions which are simply not possible to represent with an empirical model. The continuous simulation, process-based model is quite complex in this sense.

The disadvantage of the process-based model is also the complexity of the model. Data requirements are huge, and with every new datum element comes the opportunity to introduce uncertainty, as a first order error analysis would clearly indicate. Model structure interactions are also enormous in number, and with every structural interaction comes the opportunity for error, as well. In a sense, the goal in using the process-based model is to capture the advantages of the complexity of model interactions, while gaining the accuracy and dependability associated with the simpler empirically based model. This can be done, and was done with the WEPP model, using a combination of detailed sensitivity analyses and calibration of the model to the large database of natural runoff plot information used to develop the USLE. Without the tie between model and database, and without knowledge of the sensitive input variables so as to know where to focus efforts, turning a complex model such as WEPP into a useful conservation tool would not be possible. Thus in a sense, even though WEPP routines are process-based descriptors of various components of the erosional system, ultimately the model must be empirically based on the same type of data as was used to develop the USLE, along with additional experimental data collected specifically for WEPP.

Examples of process-based event models of soil erosion by water include EUROSEM,[14] the Hairsine and Rose model,[15,16] and GUEST.[17] EUROSEM is a fully dynamic model that simulates the erosion, transport and deposition of sediment over the land surface by interrill and rill processes. The model requires break-point rainfall data, ideally of one-minute resolution, as input along with a detailed description of the soils, slopes and land cover of the watershed. The model computes the interception of rainfall by the plant cover, the generation of runoff as infiltration-excess, soil detachment by raindrop impact and runoff, the transport capacity of runoff and the deposition of sediment. EUROSEM uses the runoff generator and the water and sediment routing routines of KINEROS.[18] The model of Hairsine and Rose[15,16] develops the ideas presented earlier by Rose and colleagues,[19,20] where erosion is calculated as a net balance of instantaneous processes of sediment entrainment, deposition, and re-entrainment. The GUEST model[17] is a practical conservation tool developed from these same concepts.

References

1. Wischmeier, W.H.; Smith, D.D. A Universal soil-loss equation to guide conservation farm planning. Trans. Int. Congr. Soil Sci. **1960**, 418–425.
2. Wischmeier, W.H.; Smith, D.D. *Predicting Rainfall Erosion Losses. A Guide to Conservation Planning*; Agriculture Handbook No. 537; USDA-SEA, U.S. Govt. Printing Office: Washington, DC, 1978.
3. Meyer, L.D.; Wischmeier, W.H. Mathematical simulation of the process of soil erosion by water. Trans. Am Soc. Agric. Eng. **1969**, *12* (6), 754–758, 762.
4. Foster, G.R.; Meyer, L.D. A closed-form soil erosion equation for upland Areas. In *Sedimentation* (Einstein); Shen, H.W., Ed.; Colorado State Univ.: Ft. Collins, CO, 1972.
5. Knisel, W.G., Ed.; *CREAMS: A Field Scale Model for Chemicals, Runoff, and Erosion from Agricultural Management Systems*; USDA-ARS: Washington, DC, 1980; 643.
6. Flanagan, D.C.; Nearing, M.A. *USDA-Water Erosion Prediction Project: Hillslope Profile and Watershed Model Documentation*; NSERL Report No. 10; USDA-ARS National Soil Erosion Research Laboratory: West Lafayette, IN, 1995; 47097–1196.
7. Baffaut, C.; Nearing, M.A.; Nicks, A.D. Impact of climate parameters on soil erosion using CLIGEN and WEPP. Trans. Am. Soc. of Agricultural Eng. **1996**, *39*, 447–457.
8. Mein, R.G.; Larson, C.L. Modeling infiltration during a steady rain. Water Resources Research **1973**, *9* (2), 384–394.
9. Chu, S.T. Infiltration during an unsteady rain. Water resources research **1978**, *14* (3), 461–466.

10. Williams, J.R.; Nicks, A.D. SWRRB, a simulator for water resources in rural basins: an overview. Proceedings of the Natural Resources Modeling Symposium, Pingree Park, Oct 16–21, 1983; DeCoursey, D.G., Ed.; USDA-ARS: Washington, DC, 1985; 17–22.
11. Arnold, J.G.; Williams, J.R.; Nicks, A.D.; Sammons, N.B. *SWRRB: A Basin Scale Simulation Model for Soil and Water Resource Management*; Texas A&M University Press: College Station, TX, 1990; 142 pp.
12. Flanagan, D.C.; Livingston, S.J. *USDA-Water Erosion Prediction Project: WEPP User Summary*; NSERL Report No. 11; USDA-ARS National Soil Erosion Research Laboratory: West Lafayette, IN, 1995; 47097–1196.
13. Nearing, M.A.; Foster, G.R.; Lane, L.J.; Finkner, S.C. A Process-based soil erosion model for USDA-water erosion prediction Technology. Trans. Am Soc. Agric. eng. **1989**, *32* (5), 1587–1593.
14. Morgan, R.P.C.; Quinton, J.N.; Smith, R.E.; Govers, G.; Poesen, J.W.A.; Auerswald, K.; Chisci, G.; Torri, D.; Styczen, M.E. The European soil erosion model (EUROSEM): a dynamic approach for predicting sediment transport from fields and small catchments. Earth Surface Processes and Landforms **1998**, *23*, 527–544.
15. Hairsine, P.B.; Rose, C.W. Modeling water erosion due to overland flow using physical principles: 1. sheet flow. Water Resource Res. **1992**, *28* (1), 237–243.
16. Hairsine, P.B.; Rose, C.W. Modeling water erosion due to overland flow using physical principles: 2. rill flow. Water Resource Res. **1992**, *28* (1), 245–250.
17. Misra, R.K.; Rose, C.W. Application and sensitivity analysis of process-based erosion model GUEST. Eur. J. Soil Sci. **1996**, *47*, 593–604.
18. Woolhiser, D.A.; Smith, R.E.; Goodrich, D.C. *KINEROS:* A *Kinematic Runoff and Erosion Model: Documentation and User Manual*; USDA Agricultural Research Service ARS-77, 1990.
19. Rose, C.W.; Williams, J.R.; Sander, G.C.; Barry, D.A. A mathematical model of soil erosion and deposition processes: I. Theory for a plane land element. Soil Sci. Soc. Am. J. **1983**, *47*, 991–995.
20. Rose, C.W. Developments in soil erosion and deposition models. In *Advances in Soil Science*; Springer-Verlag: New York, 1985; Vol 2.

37

Erosion by Wind: Source, Measurement, Prediction, and Control

Brenton S. Sharratt and R. Scott Van Pelt

Introduction

Wind erosion is the loss of soil caused by wind. Wind velocity in excess of that required to initiate movement of sand or soil particles results in particles rolling or creeping along the surface, bouncing or saltating on the soil surface, or being ejected from the surface directly into the atmosphere. Soil in motion is thus displaced from its original position; although displacement can be on the order of millimeters or kilometers, a net loss of soil may occur from the source area such as a field or dune. Wind erosion not only affects the productivity of the soil resource that supports plant growth but also can impact other natural resources. For example, soil eroded from a landscape and transported in the atmosphere can be deposited downwind into lakes or streams affecting water quality or linger over communities affecting air quality. On a larger scale, suspended sediment in the atmosphere can affect the radiation balance and therefore the earth's climate. Likewise, wind erosion of dunes can cause encroachment of sand onto roads or into communities.

This entry identifies causes of wind erosion as well as the implications to humans and ecosystems around the world. Discussion then focuses on a wide array of techniques and sensors used to monitor wind erosion. Indeed, measurements of wind erosion are integral to developing forecasting tools and conservation practices to protect our natural resources. This entry concludes with the authors' perspective on preventing wind erosion, recognizing that wind erosion continues today despite the availability of best management practices.

Causes

Wind erosion occurs as a result of displacement or movement of particles by wind at the earth's surface. Movement can occur as a result of rolling or creeping along the surface, bouncing or saltating on the surface, or suspension. The latter described a process whereby particles, once ejected from the surface, remain suspended and are transported off-site in the atmosphere. Creep is generally limited to particles 500 to 2000 μm in diameter, saltation to particles 100 to 500 μm in diameter, and suspension to particles smaller than 100 μm in diameter. Movement is governed by the shape, size, and density of the particle and occurs when the aerodynamic forces of wind overcome the gravitational and attraction forces imposed upon the particle at the surface. Aerodynamic forces result from wind shear at the surface; drag exerted by the surface on wind flow causes wind shear or a change in wind velocity with height above the surface. The wind velocity profile within the surface boundary or constant flux layer can be expressed as:

$$u_z = \left(\frac{u^*}{k}\right)\ln\left(\frac{z}{z_o}\right), \tag{1}$$

where u_z is wind velocity (m s^{-1}) at height z (m), u^* is friction velocity (m s^{-1}), z_o is aerodynamic roughness (m), and k is von Karman's constant (0.4). Friction velocity is proportional to the shear stress exerted by wind on the surface and is typically obtained from the wind velocity profile under conditions of neutral atmospheric stability. Particle movement occurs when u^* exceeds the threshold friction velocity of the particle (u^*_{tp}). The threshold velocity of a dry sand particle was originally described by Bagnold[1] as:

$$u^*_{\mathrm{tp}} = A\sqrt{\frac{\sigma - \rho}{\rho} gd} \tag{2}$$

where σ and ρ are the respective particle and air density (Mg m^{-3}), g is acceleration due to gravity (m s^{-2}), d is mean particle diameter (m), and A is an empirical coefficient of turbulence. As mentioned previously, the shape and size of a particle are important characteristics that affect movement. For example, flat particles are more easily entrained by wind than spherical particles due to greater lift forces at the surface of flat particles.[2] Thus, efforts have been made by scientists[3] to account for differences in particle shape and size using an equivalent particle size in estimating u^*_{tp}.

The cause of wind erosion is associated with factors that influence u^* and threshold friction velocity of the soil surface (u^*ts) since erosion can only occur when u^* exceeds u^*_{ts}. In fact, the streamwise sand transport rate and vertical dust flux from an eroding field are typically expressed as a function of the difference in u^* and u^*_{ts}.[4,5] Dust is used in this entry to describe the component of eroded soil, usually particles <100 μm in diameter, transported by suspension processes. Note that u^*_{ts} is not the same as u^*_{tp} since soil physical and chemical properties and protection afforded particles by nonerodible elements on the soil surface affect the threshold velocity. The principle factor affecting u^* is wind, and according to Eq. 1, u^* increases with wind velocity. Surface roughness also affects u^*. Roughness elements such as vegetation, stubble, ridges, and aggregates on the soil surface cause drag on the wind flow and an increase in u^*. Factors that influence u^*_{ts} include soil wetness, soil crusting, soil surface roughness, coverage of the soil surface with nonerodible aggregates and prostrate biomass, and frontal area of standing stubble or vegetation protruding above the soil surface. Shao[6] formulated an expression for u^*_{ts} based upon these soil and surface characteristics as follows:

$$u^*_{\mathrm{ts}} = u^*_{\mathrm{tp}} f_\theta(\theta) f_c(c) f_r(r) f_n(n) f_b(b) \tag{3}$$

where θ is soil moisture content, c is crust cover, r is roughness, n is nonerodible element surface cover, b is biomass frontal area, and f_θ, f_c, f_r, f_n, and f_b are respective corrective functions for soil moisture, crust cover, surface roughness, nonerodible element cover, and biomasss frontal cover. The corrective functions account for shear stress partitioning at the soil surface caused by soil wetness, crusting, roughness, nonerodible elements, and biomass. These functions are described as the ratio of $u*_t$ for a wet, crusted, rough, cloddy, or vegetative surface to $u*_t$ for a dry, loose, smooth, fine, or denuded surface. Therefore, the value of each function is ≥1. This mathematical portrayal of factors affecting $u*_{ts}$ indicates that roughening or wetting the surface or increasing coverage of crust or nonerodible elements on the surface will increase $u*_{ts}$ and therefore decrease erosion. Similarly, smoothing or drying the surface or decreasing coverage of crust or nonerodible elements on the surface will decrease $u*_{ts}$ and therefore increase erosion.

Wind erosion generally occurs in dry and windy environments. This would include the arid and semiarid regions of the world (see the section on "Source Regions") where precipitation deficits restrict or limit vegetative growth, thus allowing greater exposure of the soil surface to the forces of wind. Desertification and wildfires in these and other regions escalate wind erosion since removal of vegetation only enhances exposure of the soil surface. Thus, land use and management have a strong impact on the susceptibility of the soil to erode. Management practices that bury vegetation or biomass, destroy soil crusts or aggregates, smooth the soil surface, and aid soil drying will enhance the erodibility of the soil surface. Cultivation of land typically buries biomass on the soil surface, and although the soil surface may be rough after cultivation, crop residue generally provides greater protection than roughness to a smooth surface.[7] The type of tillage equipment and frequency of tillage operations can also affect the erodibility of the soil. Crop rotations are integral to processes that determine soil quality (e.g., biological activity, density, aggregation, water-holding capacity, crusting) and thus can influence the susceptibility of the soil to wind erosion. Intensifying the use of crops through annual cropping in drier regions or double cropping in wetter regions may enhance vegetative or crop residue cover on the soil surface throughout the year. However, rotations proven to be successful in drier regions have employed fallow in the rotation to conserve soil water for the subsequent crop. While fallow is beneficial to the subsequent crop, controlling weeds in fallow through tillage often reduces residue cover and destroys aggregates.

Climate change has a profound influence on wind erosion. Over a period of years to decades, drought will reduce vegetative growth and cover and thus enhance wind erosion. On a geologic time scale, wind erosion was thought to be enhanced during the Pleistocene as a result of modifications in wind pattern, hydrologic cycle, and the size of source areas. These factors contributed to the vast loessial deposits evident in Asia, Europe, North America, and South America.[8]

Implications

Wind erosion defines a process of moving soil particles across the landscape by wind. Wind erosion is a selective process whereby the uppermost layer of soil (topsoil) is removed from the soil profile. In regions where soils support crop production, the removal of topsoil represents an invaluable resource. In some regions of the world, loss of topsoil is an irreplaceable resource since loss of topsoil may exceed the rate of soil formation. Topsoil represents the most fertile and biologically active part of the soil profile. Eroded soil is generally enriched in carbon and nutrients[9] as well as soil microorganisms that are an important source of carbon, nitrogen, and macronutrients in desert regions.[10] Erosion of topsoil can also negatively affect the inherent physical properties of the soil such as waterholding capacity and density.[11] Thus, loss of topsoil will diminish the soil resource and the ability of the soil to support plant growth over time. Our ability to feed the world therefore deteriorates with the loss of the soil resource unless other means are used to offset the loss of nutrients and organic matter that are found in topsoil.

Wind erosion can cause immediate economic hardship to the agricultural and transportation industry. Damage to plants caused by saltation (sand blasting leaves and stems) can result in a reduction in yield or complete crop failure. Deposition of windblown soil can bury emerging seedlings and thus jeopardize plant establishment. Removal of topsoil by wind can expose newly sown seed or seedlings (Figure 1) and thus necessitate reseeding, hilling, or some other mechanical operation to reestablish the crop. Windblown sediment can also carry pesticides, weeds, and diseases harmful not only to sensitive crops and lands but also to humans. Windblown soil deposited in ditches or roads necessitates the removal of soil from these repositories. Transportation can also be impacted by failure of engines when exposed to high atmospheric dust concentrations, a particular concern in the aviation industry.

Wind erosion also contributes to poor air quality. Air quality refers to chemical constituents of the atmosphere that affects not only our health but also our visual assessment. Chemical constituents such as pesticides carried by windblown sediment can impair health or cause off-site damage to sensitive crops. In addition, inhalable particulate matter can be removed from soils by wind and travel for great distances in the atmosphere affecting communities downwind (Figure 2). Elevated PM10 [particulate

FIGURE 1 Etiolated stem and roots of potato plants exposed by wind erosion in southcentral Washington, USA. About 50 mm of topsoil was eroded during a high wind event in May 2010.

FIGURE 2 View looking to the west over Pullman, Washington, USA, as obstructed by windblown dust (top image) on October 28, 2003. Windblown dust emanated from sources 50 to 200 km west of the community and resulted in a PM10 concentration of 500 μg m^{-3} at the time of the photo. The bottom image was taken on a clear day after the dust storm.

matter ≤10 µm in diameter regulated by government entities such as the Chinese State Environmental Protection Administration, the European Commission, and the United States Environmental Protection Agency (USEPA)] concentrations downwind of eroding agricultural fields have been a concern in the Columbia Plateau region of the United States for the past two decades.[12] Wind erosion also contributes to poor visibility as a result of enhanced dust concentrations near the earth's surface. Regional haze caused by elevated dust concentrations can impair the value of natural and historic objects such as wilderness areas, national parks, and national monuments. Elevated dust concentration can also impair visibility along roadways, causing vehicular damage and loss of life.

Fine dust particulates suspended in the atmosphere as a result of wind erosion can impact regional and global climates. Atmospheric dust contributes to the scattering and absorbing of solar radiation and is a source of cloud condensation nuclei.[13] The efficiency with which dust particles affect these processes is complex and dependent on the chemical and physical properties of particles.

Deposition of windblown soil into water bodies can also impair the quality of water. Deterioration of water quality results from nutrients and pesticides being carried to water sources by airborne soil particles. In particular, phosphates bound to windblown soil trigger algae blooms that can impair both human and marine life. Recent evidence of windblown soil affecting water quality was reported in 2009 when a dust storm originating in South Australia on September 22 moved eastward, depositing suspended sediment into Sydney Harbor. The nitrogen and phosphorus bound to this sediment resulted in an explosion of plant life in the harbor, an event previously observed in Australian waters.[14]

Source Regions

Wind erosion occurs primarily in arid and semiarid regions of the world. This does not negate, however, the occurrence of wind erosion in humid regions. Nanney et al.,[15] for example, reported an annual erosion rate of 193 mt ha^{-1} in Indiana where the annual precipitation is 900 mm. Robinson[16] also reported 50 mm of topsoil eroded during a 5-day windstorm in Lincolnshire, England, where annual precipitation is nearly 600 mm. Wind erosion can also occur from roads and construction sites, but these sources are localized. Salt-affected soils are a major source of dust and are usually associated with ephemeral or dry lakes. Sparse vegetative growth and cover, characteristic of saline and sodic soils, and dispersion of aggregates in sodic soils enhance the vulnerability of salt-affected soils to wind erosion.

Wind erosion occurs on every continent of earth, even in Antarctica where rates of sand movement in McMurdo Dry Valleys rival those in many other desert regions of the world.[17] Wind erosion in Australia escalated over the past two centuries as a result of clearing lands for agricultural production. Dust storms occur with the greatest frequency in Western and South Australia. The Lake Eyre Basin located in South Australia is one of the largest sources of dust in the world.[18] Wind erosion in Africa occurs with the greatest frequency in the Saharan Desert. Chad, which lies at the southern boundary of the Sahara Desert, is the largest source of dust in the world. Warren et al.[19] suggested that the Bodele Depression in northern Chad is the dustiest place on earth. Dust storms in the Bodele Depression occur about 100 days per year. Wind erosion in Asia is prevalent in the Middle East and central Asia. Dust storms in the Middle East can occur on more than 35 days per year in Iran and Kuwait near the Persian Gulf,[20] whereas dust storms in central Asia can occur on more than 60 days per year in Turkmenistan[21] and in northwestern China.[22] The annual frequency of dust storms in Asia may seem of little consequence, but dust originating from sources such as the Taklamakan Desert in China and Gobi Desert in northern China and southern Mongolia was integral to the formation of the Loess Plateau. Wind erosion in Europe is of particular concern in the southeast region of Romania, Ukraine, and Russia.[23] Soils in these regions are generally sands. Wind erosion is not widespread in northern Europe, but degradation of agricultural land by wind is of concern from sands and silts in northern Germany, western Denmark, eastern England, southern Sweden, and the Netherlands.[24] Wind erosion in North America occurs most extensively within the Great Plains that extend from Texas, USA, into the provinces of Alberta, Manitoba, and Saskatchewan in Canada. The southern Great Plains was

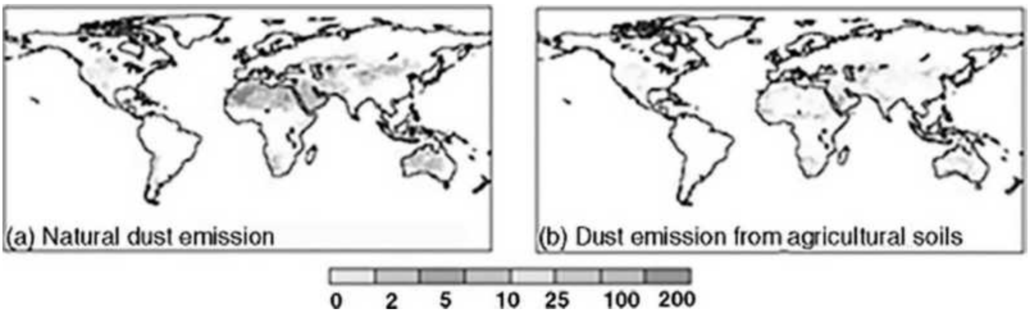


FIGURE 3 Simulated annual dust emission (g m^{-2}) from natural and agricultural or cultivated soils around the world.
Source: Reproduced from Tegen et al.[25] by permission of the American Geophysical Union.

known as the "Dust Bowl" in the 1930s when wind erosion created one of the most severe environmental catastrophes in U.S. history. Wind erosion continues to be problematic in other regions such as the Great Basin and Columbia Plateau of the United States. Wind erosion in South America is of particular concern in Argentina due to the extent of arid and semiarid lands and desertification. One of the major sources of dust in the world is the Salar de Uyuni in Bolivia.[18]

Source regions of wind erosion that contribute to the atmospheric load of dust around the world were simulated by Tegen et al.[25] The major sources of dust emanating from natural (desert, shrubland, and forest) and cultivated soils (rangeland and cropland) are from Africa, Asia, and Australia. Europe, North America, and South America also contribute as sources of dust emanating from natural and cultivated soils (Figure 3).

Tools for Monitoring

Although the redistribution of soil by wind is often quite visible, erosion and deposition can be difficult to quantify. Spatial and temporal variability of factors controlling erosion make simple measurements unreliable if not impossible.[26,27] For example, differences in topography and soil texture often create conditions where one location in the field or landscape erodes more easily than another location. Soil surface conditions such as soil roughness and crusting vary as a result of mechanical disturbance and weather. Drought reduces vegetative growth and cover and limits the ability of the land manager to protect the surface with tillage or longer lasting effects of cover crops. Wind speed and direction fluctuate on scales of seconds, and the number of erosive high wind events often varies between months and years. Stroos-nijder[28] points out that sediment cannot be collected at a single location in a field because of variations in wind direction, discharge with height above the soil surface, and distance across the field. Despite these obstacles, research over the past eight decades has resulted in numerous methodologies to assess and predict the redistribution of soil by wind.

Direct Measurement

The direct measurement of soil redistribution by wind is rarely found in the literature. Early in the history of wind erosion research, soil erosion pins or stakes were the only available devices to measure soil redistribution. The concept behind the use of soil pins is simple—anchor the pin in the soil to avoid heave or change in elevation of the pin from normal soil processes such as freezing and thawing, initially measure the distance between the top of the pin and soil surface, and return at some future date for a repeated measurement. If the distance between the top of the pin and soil surface has increased, net erosion has occurred and if the distance has decreased, net deposition has occurred at the location of

FIGURE 4 Pin frame with pins conforming to the soil surface. Dark pins on both outside edges of the frame are resting on erosion pins.

the pin. In a study in southern New Mexico, erosion pins installed in 1933 have been measured multiple times since, revealing interdecadal changes in patterns and rates of soil redistribution.[29] Selkirk and Saffigna also used erosion pins to document deflation of 43 mm yr^{-1} at the north end and accretion of 28 mm yr^{-1} at the south end of a mixed sand and peat field in Australia.[30]

Since an erosion pin measures localized erosion or deposition, a large number of spatially distributed pins are needed to obtain a spatial estimate of soil redistribution. Using the pin as a reference point and taking multiple measurements of the soil surface reduces the number of pins needed and increases the accuracy of the estimate. This technique necessitates using a pin frame such as shown in Figure 4. These measurements can be time-consuming and the frame pins may sink into unconsolidated soil, reducing measurement accuracy. Advances in technologies using lasers and photoreceptors have allowed very precise and spatially intensive measurement of surface elevations.[31] Recently, photoelectronic pins with built-in photoreceptors have been used to investigate the short-term fluctuations in erosion.[32] Direct measurement of soil erosion has also been made from observations of exposed crop roots[33] and native vegetation such as shown in Figure 5. Direct measurement of deposition is often possible along field margins where fences and vegetation trap the sediment, resulting in buried fences.

Horizontal Mass Flux

The concept of mass balance is often used to measure wind erosion. In practice, soil erosion or deposition is determined by the difference in sediment discharge between a windward and leeward location in the field. Sediment discharge is typically assessed by measuring the horizontal mass flux (HMF), or the mass of sediment moving across a small vertical plane, at several heights above the surface.[34] An increase in HMF between the windward and leeward locations represents a net loss of sediment between the two locations. There is a limit to the amount of sediment that may be entrained by the wind as the entrained particles exert a drag on the wind and reduce its erosivity.[35] The upper limit of sediment that wind may carry is termed *transport capacity*; this concept is used to describe the variation in HMF as a function of distance across an eroding field using the relationship:

$$f_x = f_{mx}\left(1 - e^{-x/b}\right), \tag{4}$$

FIGURE 5 Root of a shrub used to measure depth of soil excavated by wind erosion in southern Kansas. the distance to a protected surface (m), and *b* is the characteristic horizontal length scale or the distance at which $fx = 0.632\, fmx$.[36]

where f_x is the HMF in the windward direction (g $m^{-2}day^{-1}$), f_{mx} is the maximum value of HMF at a given height (g $m^{-2}day^{-1}$), x is the horizontal coordinate usually taken to be The HMF typically diminishes with height above the soil surface. A theoretical expression describing the vertical distribution of sediment in saltation is:[37]

$$f = f_o\left(1+\frac{y}{\sigma}\right)^{\beta}, \tag{5}$$

where f is the suspended soil mass flux (g m^{-2}), f_o is the soil mass flux moving in the interval of 0–3 mm above the surface (g m^{-2}), y is the height above the soil surface (m), $\boldsymbol{\sigma}$ is the height below which 50% of total HMF occurs for $\boldsymbol{\beta} = -2$ (m), and $\boldsymbol{\beta}$ is a dimensionless power term describing the slope of the height as a log transformed function of the measured HMF. Other investigators[38] reported that suspended sediment, which is above the layer of creep and saltation, conforms to the power function:

$$w = ay^{-b}, \tag{6}$$

where w is the mass of sediment per unit area of sampler opening (g m^{-2}) collected at height y (m) and a and b are regression coefficients. Sediment transported by creep and saltation were described by the relationship:

$$w = ce^{dy}, \tag{7}$$

where w and y are the same as Eq. 6 and c and d are regression coefficients. Eqs. 6 and 7 may be combined to describe the vertical distribution of creep, saltation, and suspension using the expression:

$$w = ay^{-b} + ce^{dy}. \tag{8}$$

Integrating either Eq. 5 or Eq. 8 throughout the height of vertical flux and entering these integrated vertical sediment fluxes into Eq. 4 results in the field-wise estimate of soil loss.[37]

HMF Samplers

Reliable measurements of HMF at several heights above the surface must be made to estimate sediment discharge at various locations within a field. Specialized samplers have been designed and utilized for measuring HMF. Among the first attempts to measure HMF was the vertical slot sampler of Bagnold.[1] Chepil improved on this design by adding a wind vane and mounting the sampler so that it could pivot into the wind.[39] Many designs have been tested over the years, but few have been truly accepted by the aeolian research community. Use of active samplers, those relying on a motive force to move air through the sampling orifice and sedimentation chamber or filter, is rare in the scientific literature. Passive samplers include the Modified Wilson and Cooke (MWAC),[40] the Big Spring Number Eight (BSNE),[41] and the Wedge Dust Flux Gauge (WDFG).[42] These samplers require frequent inspection to reinitialize and manually collect sediment prior to or after a high wind event. The SUSpended Sediment TRAp (SUSTRA) is a pivoting passive sampler that weighs the collected sediment, allowing for automated real-time HMF measurements.[43]

Among passive HMF samplers most widely used include the MWAC (Figure 6) and the BSNE (Figure 7) samplers. HMF samplers vary in sampling efficiency. The MWAC sampler has an 80%–102% efficiency while the BSNE sampler has a 35%–45% efficiency in trapping particles with a diameter ≤63 μm at wind speeds of 2 to 5 m s^{-1}.[44] In a similar investigation, the MWAC sampler was found to be 90%–120% efficient and the BSNE sampler was found to be 70%–130% efficient in trapping particles with diameters between 132 and 287 μm at wind speeds ranging from 6 to 14 m s^{-1}.[45] While the efficiency of the MWAC sampler was relatively independent of wind speed, the efficiency of the BSNE sampler tended to decrease with increased wind speed.

FIGURE 6 A profile of MWAC samplers used to measure the vertical distribution of airborne sediment. Inlet tubes (to the left of the bottles) are oriented into the wind by the sail (to the right of the image) and trap airborne sediment as the mast rotates in the soil mount (at the bottom of the image).

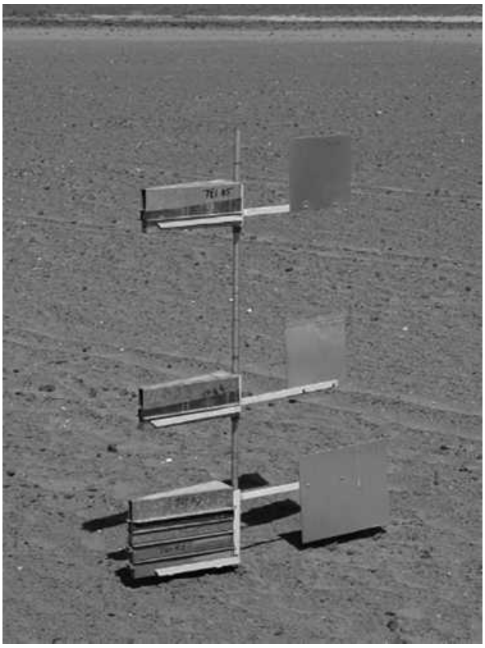


FIGURE 7 Field installation of BSNE samplers. The two uppermost samplers are single samplers while the bottom sampler is specially designed to sample HMF at three heights in the saltation layer (below a height of 0.25 m).

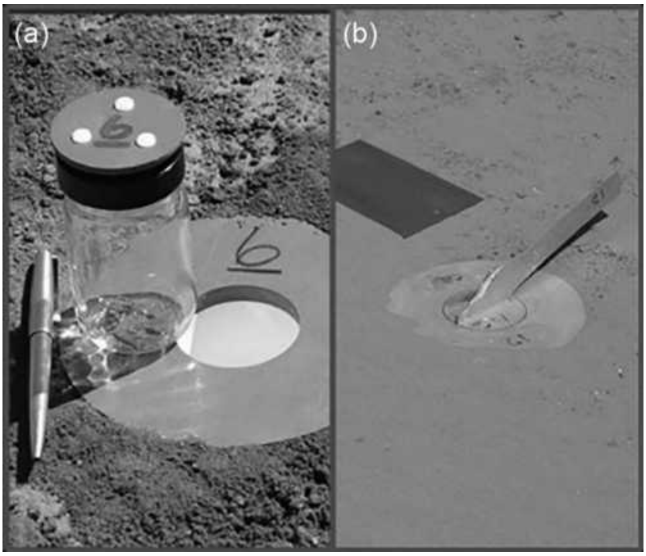


FIGURE 8 Two types of creep samplers include an omnidirectional sampler consisting of a jar with holes in the raised metallic lid (a) and a pivoting sampler that traps creep and saltation at a height of 0 to 0.02 m above the soil surface (b).

Movement of the creep component of HMF is measured by specialized samplers. The simplest design utilizes a jar with a small hole in the lid that collects particles passing over the hole. A more sophisticated design is a wind vane oriented sampler that is installed flush with the soil surface and samples particles moving along and very close to the soil surface.[46] An example of both these designs is shown in Figure 8.

Temporal resolution of saltation in the field is often measured using piezoelectric or acoustical impact sensors such as the Sensit[47] shown in Figure 9 or the saltiphone[48] shown in Figure 10. These sensors develop an electrical pulse when struck by a saltating particle. A temporal resolution of less than 1 sec allows these impact sensors to be used to estimate the threshold wind velocity of erodible surfaces.[49] A recent study found that the sensitivity varies among individual sensors of the same type and with wind speed and particle diameter. The authors concluded that impact sensors should only be used to

FIGURE 9 A Sensit omnidirectional saltation impact sensor with blue piezo sensing element. The Sensit is mounted vertically by burying the massive base or by inverting the Sensit and supporting the base above the soil surface as shown in the image.

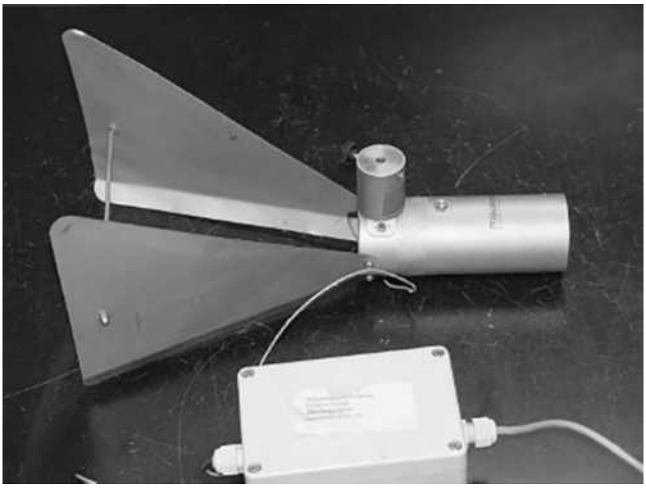


FIGURE 10 A saltiphone pivoting saltation impact sensor. The saltiphone is an acoustical sensor with a microphone diaphragm in the tubular body. The fins orient the saltiphone tube and microphone into the wind.

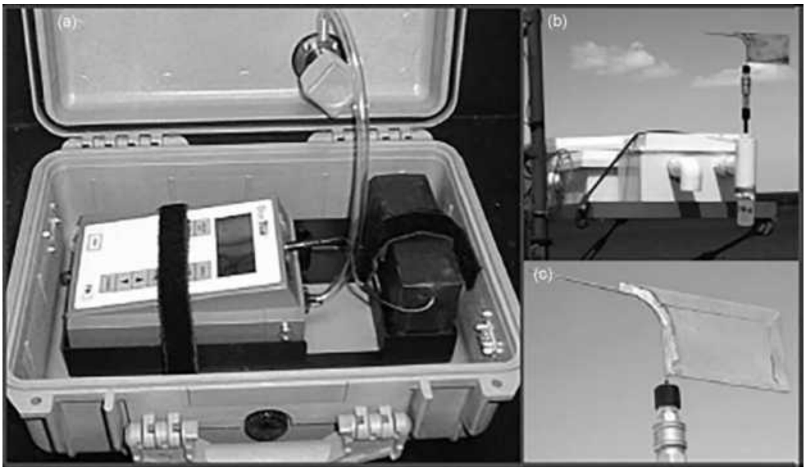


FIGURE 11 A DustTrak suspended sediment sampler. A case houses the sampler, battery, and inlet tube (a) and is mounted on a tower at an erosion field site (b) with the inlet tube oriented into the wind by the fin (c).

investigate temporal and spatial patterns of wind erosion unless carefully calibrated to the specific conditions expected in the field.[50] Recent advances in optical sensors promise more quantitative time-resolved measurements of saltation.[51]

Fugitive dust in suspension is often the most visible evidence of active wind erosion. Passive HMF samplers can be used to trap particles in suspension, but these samplers are inefficient at trapping dust due to resuspension of small particles within the sampling chamber. The suspension component of HMF is most often measured with aspirated sampling devices. Optical sensors such as the DustTrak (Figure 11), E-sampler, or Grimm particle sampler aspirate the sediment-laden air past a light source that measures the optical occlusion caused by the suspended sediment. Optical sensors have been used with saltation impact sensors to examine the linkage between saltation and fine dust emissions.[52] The Tapered Element Oscillating Microbalance (TEOM), as shown in Figure 12, aspirates the sedimentladen air, filters the particulates, and periodically weighs the collected dust. The TEOM is a USEPA-approved device for measuring real-time PM10. Filter-based samplers, such as the low-volume or high-volume sampler, are also used to measure dust in suspension. These samplers aspirate air at a known rate through a preweighed hydrophobic filter, thus providing a time-integrated measure of dust in suspension.

Radioisotopic Techniques

Radioisotopes, especially the anthropogenic isotopes such as ^{137}Cs, provide ideal soil tracers for studying long-term (50 years) soil redistribution rates. These radioisotopes were formed as a result of nuclear bomb detonations and distributed globally in the stratosphere. From 1954 to 1970, these isotopes were deposited on the earth's surface by wet and dry deposition processes, collectively termed fallout. The peak of radioisotopic fallout was in 1963 and 1964 as nations rushed to use test warheads prior to the nuclear test ban treaty of 1963.[53] Certain anthropogenic radioisotopes, especially ^{137}Cs and $^{239+240}$Pu are bound very tightly to the exchange surface of soil particles and only move in the environment with movement of the soil particle to which they are bound.[53]

Radioisotopes have been widely used to study the rates and patterns of soil redistribution by water,[54] but studies of aeolian soil redistribution using radioisotopes are limited. One of the first studies to look at redistribution of ^{137}Cs on the landscape as an indicator of aeolian soil movement assumed that the erosion on a field knoll might be due to wind.[55] Subsequent to this study, the ^{137}Cs technique has been used to study rates and patterns of soil redistribution by wind in North America,[56,57] Europe,[58]

FIGURE 12 A TEOM dust sampler. Air is aspirated through the inlet located above the white enclosure in the field (a) and into a tube before passing through a filter (in the tan base of b). The weight of the filter is periodically recorded by the controller unit (gray panel in upper left of b).

Africa,[59] Australia,[60] and Asia.[61] Models developed for estimating aeolian redistribution rates based upon loss of ^{137}Cs from the soil agree well with observations taken at field sites with documented erosion and deposition histories.[62] Future work will hopefully develop the use of the longer-lived $^{239+240}$Pu for estimating wind erosion and allow for calibration of wind erosion models based on 50 years average erosion rates.

Wind Erosion Models

The measurement of wind erosion during high wind events has allowed the development of a wide variety of models to predict sediment entrainment and transport at field to global scales.[34] The models range from early empirical models, such as the Wind Erosion Equation (WEQ) released in 1965, to more process-based models, such as the Wind Erosion Prediction System (WEPS). References and

TABLE 1 Commonly Used Wind Erosion Models along with Citations and Climatic and Soil Surface Input Variables[34]

Model	Citation	Climate Variables	Roughness	Aggregation	Crust	Soil
WEQ	[65–67]	Precipitation; potential evapotranspiration; monthly wind direction, preponderance, and % erosive energy; irrigation dates	Ridge height, ridge spacing, standing, and flat biomass	Fraction <0.84 µm	Presence or absence	Percent sand, silt clay, organic matter, and calcium carbonate
RWEQ	[75,76]	Wind speed, precipitation, air temperature, solar radiation, irrigation dates	Ridge height, ridge spacing, random roughness, ridge orientation, standing, and flat biomass	Function of sand, silt, clay, organic matter, and calcium carbonate content	Clay and organic matter content	Sand, silt, clay, organic matter, and calcium carbonate content
WESS	[76,77]	Wind speed, wind speed perturbation factor	Ridge interval, roughness factor, standing, and flat biomass	Soil erodibility as a function of soil texture	Not considered	Bulk density, erodible particle diameter, soil water content, and drying rate
WEPS	[68–70, 72]	Air density, wind direction, daily wind speed (in specified intervals per day or Weibull equation parameters and fraction calm)	Ridge height, ridge spacing, ridge width, ridge orientation, random roughness, dike spacing, standing, and flat biomass	Aggregate density, aggregate stability, aggregate diameter (mean and standard deviation), maximum and minimum aggregate size	Crust fraction, thickness, density, and stability; fraction and mass of LEM on crust	Number of layers and layer thickness; bulk density; sand silt and clay fraction; rock volume, water content, wilting point water content
WEAM	[78]	Wind speed, precipitation, evaporation	Aerodynamic roughness length, plant and soil frontal area index	Not considered	Not considered	Coarse, medium, and fine sand content

WEQ, Wind Erosion Equation; RWEQ, Revised Wind Erosion Equation; WESS, Wind Erosion Stochastic Simulator; WEPS, Wind Erosion Prediction System; WEAM, Wind Erosion Assessment Model; LEM, loose erodible material.

input parameters for selected wind erosion models are presented in Table 1. Although other wind erosion models such as the Wind Erosion on European Light Soils model[63] and Dust Production Model[64] can be found in the literature, discussion hereafter is limited to the WEQ and WEPS. These two models represent the oldest and one of the newest technologies available to predict wind erosion. The WEPS also dynamically simulates changes in soil properties and surface characteristics that govern wind erosion.

Wind Erosion Equation

In the United States, the WEQ has been used extensively over the past several decades to assess the effect of crop and management on wind erosion.[65,66] The WEQ was designed to predict long-term average annual wind erosion based on a specific set of climatic and field conditions; the model can be used to predict wind erosion for specific time periods in which the soil and climate factors do not change. Annual soil loss is then obtained by summing soil loss for all individual time periods in a year. Details for using WEQ are found in the U.S. Department of Agriculture-Natural Resources Conservation Service (USDA-NRCS) National Agronomy Manual.[66] The WEQ is defined by the following function:

$$E = f\left(IKCLV\right), \tag{9}$$

where E is the estimated annual soil loss (Mg ha^{-1}), f indicates a non-linear functional relationship among the variables, I is the soil erodibility index (Mg ha^{-1}), K is the soil surface roughness factor, C is the climatic factor, L is the unsheltered distance, and V is the vegetative cover factor.

In practice, I is determined by the surface soil texture and assigned a value determined by a predetermined effect of soil texture on the aggregate formation and stability of the soil. The absolute value of I is defined as the potential wind erosion for an unsheltered, bare, level, smooth, and loose soil at a location where C is 100. The other factors are then multiplied with I using nomographs in the USDANRCS National Agronomy Manual[66] to determine appropriate values.

Being an empirical model, WEQ tends to perform well within the range of original data used to develop the model. In a Global Change and Terrestrial Ecosystem/Soil Erosion Network (GCTE/SEN) exercise comparing the performance of wind erosion models during 46 high wind events across six states in the United States, WEQ was found to accurately estimate erosion in the central Great Plains but overestimate erosion in the northern Great Plains and underestimate erosion in the southern Great Plains.[67] The WEQ is very sensitive to soil texture, surface roughness, and residue. The high sensitivity of WEQ to soil texture necessitates adjusting the I factor to local conditions.[67]

The Revised Wind Erosion Equation (RWEQ) was later developed[75] to estimate erosion during a high wind event based upon assessments of soil moisture, aggregate size, crust cover, surface roughness, and crop residue characteristics. Although empirical, RWEQ was developed using field data and intended to predict real-time soil loss from a field of a known length while WEQ was developed using wind tunnel data and intended to estimate annual soil loss.

Wind Erosion Prediction System

Recent advances in computational power and in the understanding of the physical processes governing wind erosion have resulted in the advent of process-based wind erosion models. The WEPS was developed by USDA scientists to provide more accurate results across climatic zones and land management systems. The WEPS simulates, on a daily time scale, the fundamental processes involved in soil redistribution by wind.[68] The WEPS represents the computer implementation and integrations of several science-based submodels. The weather submodel uses historical statistical information of several weather parameters and stochastic techniques to develop likelihood-based weather scenarios for individual growing seasons. These weather scenarios drive processes in other submodels. The hydrology submodel uses inputs from the weather and other submodels to compute water content of various soil layers and at the soil surface. The management submodel simulates the primary anthropogenic actions that affect a field's susceptibility to wind erosion. These actions include cultural practices such as tillage, planting, harvesting, and irrigation. The soil submodel simulates soil temporal properties that affect the erodibility of the soil. This submodel uses information from the weather, hydrology, and management submodels. The crop submodel is a crop growth model that produces roots, leaves, stems, and reproductive tissues on a daily basis. This submodel is integral to the development of the canopy and stems that modify wind speeds near the surface and flat biomass that protect the surface. The decomposition submodel simulates the microbial breakdown of the biomass created by the crop submodel. Finally, the erosion submodel uses parameters provided by other submodels to simulate the entrainment and transport of soil particles.

The WEPS user interface allows the operator to define the size and orientation of the field, the location of the field, the cropping system including tillage and crop rotation, and the soil series. The location of the field defines the weather station database that is accessed by the weather submodel. The soil series specifies parameters for the soil and hydrology submodels; these paremters are obtained from the USDA-NRCS Soil Survey Geographic (SSURGO) database.

Initial tests in using the WEPS indicated good performance. Reasonable agreement between measured and simulated wind erosion ($R^2=0.71$) was found in the GCTE/SEN exercise.[69] Even better agreement between measured and simulated wind erosion ($R^2>0.9$) was found in Germany.[70] Sensitivity analyses have found WEPS to be sensitive to soil surface conditions including surface wetness, dry aggregate stability, oriented roughness, and residue management.[71] Some of the WEPS simulation inaccuracies noted in the GCTE/SEN exercise have been attributed to errors in soil parameter values or time-dependent changes in soil erodibility.[69] In a very detailed sensitivity analysis,[72] investigators found that the WEPS was most sensitive to changes in biomass flat cover, near-surface soil water content, ridge height, and other parameters including crust cover and random roughness.

Control Strategies

Land management plays a key role in controlling wind erosion. Establishing native perennial species in deserts, rangelands, and shrublands will provide some stability against disease and drought. Likewise, eliminating annual weeds from these ecosystems will offer some protection against the spread of wildfire and thus denuding the land. Limiting grazing on rangelands will ensure adequate vegetative cover to protect the soil surface while limiting recreational pursuits along coast and deserts will protect the sparse vegetation and surface crusts in these landscapes. Regulating water levels of interior seas or lakes may reduce the exposure of beaches to wind, but more stringent efforts such as applying chemical stabilizers may be needed to reduce dust emissions from beaches where desiccation is a natural process.[73]

Land management practices are available in most agricultural regions that will protect the soil resource within a cropping system. No tillage or conservation tillage provides greater protection to the soil surface against the forces of wind as compared with conventional tillage because conservation tillage enhances crop residue cover and minimizes degradation of aggregates or crusts (Figure 13). Ridge tillage, with ridges oriented perpendicular to prevailing winds, provides some protection to the soil surface by enhancing roughness. Intensifying crop rotations or the use of cover crops can also minimize the time when soils are exposed to wind, but may be impractical without irrigation in arid regions. Other practices such as amending the soil with straw, manure, or synthetic binders can be effective but costly. Wind strips or barriers or strip cropping is effective in reducing wind velocities at the soil surface but is only practical in regions with unidirectional winds. Perennial plants are frequently used in wind strips across agricultural fields; their effectiveness in reducing erosion will depend on plant height, spacing, leafiness, and rigidity.

Adoption of practices by land managers to control wind erosion is influenced by an awareness of the practice, availability of resources to apply the practice, effectiveness of the practice, and economics of adopting the practice. In arid and semiarid regions, failure to adopt practices is generally governed by

FIGURE 13 Conservation tillage used to control wind erosion in irrigated corn in southern Washington. Corn was sown into a cover crop that was established the previous autumn (conservation tillage) or using conventional tillage practices. The transition between conservation and conventional tillage is noted by change in color of soil surface along the vertical in the center of the image.

economics (low profit margin). For example, the cost of establishing a wind erosion control practice in an agricultural field cannot be compensated by an increase in crop yield. As such, assistance is typically needed by federal or state governments to offset costs of adopting control practices. Government programs have been effective in promoting the use of conservation practices to reduce soil erosion to tolerable levels in some regions of the world, but these programs are costly. Nonprofit organizations such as the Northern Agricultural Catchments Council of Australia have provided incentives to land managers to adopt conservation practices to control wind erosion. These organizations, however, typically rely on both public and private funding to support adoption of practices. Private industry has also provided incentives to land managers to adopt erosion control practices. For example, Shepherd's Grain contracts crops grown by farmers using sustainable practices for manufacturing food products in the Pacific Northwest, USA. These industries are generally localized and fill a small niche in the market. Therefore, public support is needed to ensure the sustainability of our air, soil, and water across broad regions such that today's land management practices do not jeopardize the air, soil, and water resources to meet the needs of future generations. This support could be in the form of national policies that encourage research and development of simple and cheap wind erosion control technologies.

Conclusion

Wind erosion continues to threaten air, soil, and water resources across the world. Loss of fertile topsoil by wind has contributed to the demise of civilizations.[74] Fine particles emitted from soils by high winds are transported great distances in the atmosphere and have contributed to pollution of our air and water resources. To ensure the sustainability of the soil resource to meet the needs of future generations, cheap and readily available management practices that will protect the soil surface from the forces of the wind are needed. This is a particular challenge in the arid and semiarid regions of the world, but one that can be met through continued education, development, and research. Tools for monitoring and predicting wind erosion have rapidly expanded over the past three decades; these tools will be invaluable in monitoring the effectiveness of new control practices across vast landscapes. Development of technologies to control wind erosion are costly, but must be shared by today's generation to ensure the well-being of future generations.

References

1. Bagnold, R.A. *The Physics of Blown Sand and Desert Dune*; Dover Publications: Mineola, New York, 1941.
2. Rice, M.A. Grain shape effects on aeolian sediment transport. Acta Mech. **1991**, *1,* 159–166.
3. Allen, T. *Particle Size Measurement*; Chapman and Hall: London, England, 1981.
4. Lettau, H.H.; Lettau, K. Experimental and micrometeorological studies of dune migration. In *Exploring the World's Driest Climate;* Lettau, H.H., Lettau, K., Eds.; Institute of Environmental Studies Report 101; University of Wisconsin: Madison, Wisconsin, 1978; 110–147.
5. Gillette, D.A.; Passi, R. Modeling dust emission caused by wind erosion. J. Geophys. Res. **1998**, *93* (D11), 14233–14243.
6. Shao, Y. *Physics and Modeling of Wind Erosion*; Kluwer Academic Publishers: Dordrecht, The Netherlands, 2000.
7. Horning, L.; Stetler, L.; Saxton, K.E. Surface residue and soil roughness for wind erosion protection. Trans. Am. Soc. Agric. Eng. **1998**, *41,* 1061–1065.
8. Busacca, A.; Chandler, D. Climate change and erosion by wind. In *Encyclopedia of Soil Science;* Chesworth, W., Ed.; Marcel Dekker, Inc.: New York, 2002; 484–488.
9. Li, J.; Okin, G.S.; Epstein, H.E. Effects of enhanced wind erosion on surface soil texture and characteristics of windblown sediments. J. Geophys. Res. **2009**, *114,* doi:10.1029/2008JG000903.
10. Belnap, J. Surface disturbances: Their role in accelerating desertification. Environ. Monit. Assess. **1995**, *37,* 39–57.
11. Mahli, S.S.; Izaurralde, R.C.; Nyborg, M.; Solberg, E.D. Influence of topsoil removal on soil fertility and barley growth. J. Soil Water Conserv. **1994**, *49,* 96–101.
12. Sharratt, B.S.; Lauer, D. Particulate matter concentration and air quality affected by windblown dust in the Columbia Plateau. J. Environ. Qual. **2006**, *35*, 2011–2016.
13. Tegen, I. Modeling the mineral dust aerosol cycle in the climate system. Quat. Sci. Rev. **2003**, *22*, 1821–1834.
14. Shaw, E.C.; Gabric, A.J.; McTainsh, G.H. Impacts of Aeolian dust deposition on phytoplankton dynamics in Queensland coastal waters. Mar. Freshwater Res. **2008**, *59*, 951–962.
15. Nanney, R.D.; Fryrear, D.W.; Zobeck, T.M. Wind erosion prediction and control. Water Sci. Technol. **1993**, *28,* 519–527.
16. Robinson, D.N. Soil erosion by wind in Lincolnshire, England. East Midl. Geogr. **1968**, *4*, 351–362.
17. Lancaster, N. Flux of Aeolian sediment in the McMurdo Dry Valleys, Antarctica: A preliminary assessment. Arct. Alp. Res. **2002**, *34,* 318–323.
18. Goudie, A. The history and nature of wind erosion in deserts. Annu. Rev. Earth Plant. Sci. **2008**, *36,* 97–119.
19. Warren, A.; Chappell, A.; Todd, M.C.; Bristow, C.; Drake, N.; Engelstaedter, S.; Martins, V.; M'bainayel, S.; Washington, R. Dust-raising in the dustiest place on earth. Geomorphology **2007**, *92,* 25–37.
20. Safar, M.I. *The Frequency of Dust in Day-time Summer in Kuwait;* Meteorological Department: State of Kuwait, 1980.
21. Goudie, A.S. Dust storms in space and time. Prog. Phys. Geogr. **1983**, *7,* 502–530.
22. He, Q.; Zhao, J.; Nagashima, H. The distribution of sand storms in Taklimakan Desert. J. Arid Land Stud. **1996**, *5,* 185–193.
23. European Environment Agency. *Europe's Environment: The Second Assessment;* European Environment Agency; Elsevier Science Ltd: Copenhagen, Denmark, 1998.
24. Riksen, M.; Brouwer, F.; de Graaff, J. Soil conservation policy measures to control wind erosion in northwestern Europe. Catena **2003**, *52,* 309–326.
25. Tegen, I.; Werner, M.; Harrison, S.P.; Kohfeld, K.E. Relative importance of climate and land use in determining present and future global soil dust emission. Geophys. Res. Lett. **2004**, *31,* doi:10.1029/2003GL019216.

26. Chepil, W.S.; Woodruff, N.P. The physics of wind erosion and its control. Adv. Agron. **1963**, *13,* 211–302.
27. Visser, S.M.; Sterk, G. Techniques for simultaneous quantification of wind and water erosion in semi-arid zones. In *Soil Erosion Research for the 21st Century,* International Symposium for Soil Erosion Research for the 21st Century, Honolulu, HI, USA, Jan. 3–5, 2001; Ascough, J.C., Flanagan, D.C., Eds.; American Society of Agricultural Engineers: St. Joseph, MI, USA, 2001; 544–547.
28. Stroosnijder, L. Measurement of erosion: Is it possible? Catena **2005**, *64,* 162–173.
29. Gibbens, R.P.; Tromble, J.M.; Hennesey, J.T.; Cardenas, M. Soil movement in mesquite dunelands and former grasslands of southern New Mexico from 1933 to 1980. J. Range Manage. **1983**, *36,* 145–148.
30. Selkirk, J.M.; Saffigna, L.J. Wind and water erosion of a peat and sand area on subantarctic Macquarie Island. Arc., Antarct. Alp. Res. **1999**, *31,* 412–420.
31. Huang, C.; Bradford, J.M. Portable laser scanner for measuring soil surface roughness. Soil Sci. Soc. Am. J. **1990**, *54,* 1402–1406.
32. McDermott, J.P.; Sherman, D.J. Using photo-electronic erosion pins for measuring bed elevation changes in the swash zone. J. Coastal Res. **2009**, *25*, 788–792.
33. Chepil, W.S. Conversion of relative field erodibility to annual soil loss by wind. Soil Sci. Soc. Am. J. **1960**, *24,* 143–145.
34. Zobeck, T.M.; Sterk, G.; Funk, R.; Rajot, J.L.; Stout, J.E.; Van Pelt, R.S. Measurement and data analysis methods for field-scale wind erosion studies and model validation. Earth Surf. Processes Landforms **2003**, *28,* 1163–1188.
35. Owen, P.R. Saltation of uniform sand grains in air. J. Fluid Mech. **1964**, *20,* 225–242.
36. Stout, J.E. Wind erosion within a simple field. Trans. Am. Soc. Agric. Eng. **1990**, *33*, 1597–1600.
37. Fryrear, D.W.; Stout, J.E.; Hagen, L.J.; Vories, E.D. Wind erosion: Field measurement and analysis. Trans. Am. Soc. Agric. Eng. **1991**, *34,* 155–160.
38. Vories, E.D.; Fryrear, D.W. Vertical distribution of winderoded soil over a smooth, bare field. Trans. Am. Soc. Agric. Eng. **1991**, *34,* 1763–1768.
39. Chepil, W.S. Width of field strips to control wind erosion; Kan. Agric. Exp. Sta. Tech. Bull. 92; Kansas State University: Manhattan, Kansas, 1957.
40. Wilson, S.J.; Cooke, R.U. Wind erosion. In *Soil Erosion;* Kirkby, M.J.; Morgan, M.P.C., Eds.; John Wiley & Sons: Chichester, England,1980; 217–251.
41. Fryrear, D.W. A field dust sampler. J. Soil Water Conserv. **1986**, *41*, 117–120.
42. Hall, D.J.; Upton, S.L.; Marsland, G.W. Designs for a deposition gauge and a flux gauge for monitoring ambient dust. Atmos. Environ. **1994**, *28,* 2963–2979.
43. Janssen, W.; Tetzlaff, G. Entwicklimg und eichung einer registrierenden suspension falle. Z. Kulturtech. Landensentwickl. **1991**, *32,* 167–180.
44. Goossens, D.; Offer, Z.Y. Wind tunnel and field calibration of six aeolian dust samplers. Atmos. Environ. **2000**, *34,* 1043–1057.
45. Goossens, D.; Offer, Z.; London, G. Wind tunnel and field calibration of five aeolian sand traps. Geomorphology **2000**, *35*, 233–252.
46. Stout, J.E.; Fryrear, D.W. Performance of a windblown particle sampler. Trans. Am. Soc. Agric. Eng. **1989**, *32,* 2041–2045.
47. Stockton, P.H.; Gillette, D.A. Field measurement of the sheltering effect of vegetationon erodible surfaces. Land Degrad. Rehabil. **1990**, *2*, 77–85.
48. Spaan, W.P.; van den Abeele, G.D. Wind borne particle measurements with acoustic sensors. Soil Technol. **1991**, *4*, 51–63.
49. Stout, J.E. A method for establishing the critical threshold for aeolian transport in the field. Earth Surf. Processes Landforms **2004**, *29,* 1195–1207.
50. Van Pelt, R.S.; Peters, P.; Visser, S. Laboratory wind tunnel testing of three commonly used saltation impact sensors. Aeolian Res. **2009**, *1*, 55–62.

51. Mikami, M.; Yamada, Y.; Ishizuka, M.; Ishimaru, T.; Gao, W.; Zeng, F. Measurement of saltation process over gobi and sand dunes in the Taklimakan Desert, China with a newly developed sand particle counter. J. Geophys. Res. **2005**, *110,* doi: 10.1029/2004JD004688.
52. Zobeck, T.M.; Van Pelt, R.S. Wind-induced dust generation and transport mechanics on a bare agricultural field. J. Hazard. Mater. **2006**, *132,* 26–38.
53. Ritchie, J.C.; McHenry, J.R. Application of radioactive fallout cesium-137 for measuring soil erosion and sediment accumulation rates and patterns: A review. J. Environ. Qual. **1990**, *19,* 215–233.
54. Ritchie, J.C. ^{137}Cs use in estimating soil erosion: 30 years of research. In *Use of ^{137}Cs in the Study of Soil Erosion and Sedimentation;* Zapata, F., Pennock, D., Eaglesham, A.R.J., Eds.; International Atomic Energy Agency TECDOC-1028; International Atomic Energy Agency: Vienna, Austria, 1998; 5–12.
55. de Jong, E.; Villar, H.; Bettany, J.R. Preliminary investigations on the use of ^{137}Cs to estimate erosion rates in Saskatchewan. Can. J. Soil Sci. **1982**, *62,* 673–683.
56. Sutherland, R.A.; Kowalchuk, T.; de Jong, E. Cesium-137 estimates of sediment redistribution by wind. Soil Sci. **1991**, *151*, 387–396.
57. Ritchie, J.C.; Herrick, J.E.; Ritchie, C.A. Variability in soil redistribution in the northern Chihuahuan Desert based on ^{137}Cs measurements. J. Arid Environ. **2003**, *55*, 737–746.
58. Chappell, A.; Warren, A. Spatial scales of ^{137}Cs-derived soil flux by wind in a 25 km^2 arable area of eastern England. Catena **2003**, *52*, 209–234.
59. Chappell, A.; Warren, A.; Oliver, M.A.; Charlton, M. The utility of ^{137}Cs for measuring soil redistribution rates in southwest Niger. Geoderma **1998**, *81,* 313–337.
60. Harper, R.J.; Gilkes, R.J. Evaluation of the ^{137}Cs technique for estimation wind erosion losses for some sandy western Australian soils. Aust. J. Soil Res. **1994**, *32,* 1369–1387.
61. Ping, Y.; Dong, Z.; Dong, G.; Zhang, X.; Zhang, Y. Preliminary results of using 137Cs to study wind erosion in Qinghai-Tibet Plateau. J. Arid Environ. **2001**, *47,* 443–452.
62. Van Pelt, R.S.; Zobeck, T.M.; Ritchie, J.C.; Gill, T.E. Validating the use of 137Cs measurements to estimate rates of soil redistribution by wind. Catena **2007**, *70,* 455–464.
63. Bohner, J.; Schafer, W.; Conrad, O.; Gross, J.; Ringeler, A. The WEELS model: Methods, results and limitations. Catena **2003**, *52*, 289–308.
64. Marticorena, B.; Bergametti, G. Modeling the atmospheric dust cycle: 1. Design of a soil-derived dust emission scheme. J. Geophys. Res. **1995**, *100,* 16415–16430.
65. Woodruff, N.P.; Siddoway, F.H. A wind erosion equation. Soil Sci. Soc. Am. Proc. **1965**, *29,* 602–608.
66. USDA-NRCS Soil Survey Staff. *National Agronomy Manual,* 2nd Ed.; 190-V; USDA-Natural Resources Conservation Service: Washington, DC, 1988.
67. Van Pelt, R.S.; Zobeck, T.M. Validation of the Wind Erosion Equation (WEQ) for discrete periods. Environ. Modell. Software **2004**, *19,* 199–203.
68. Hagen, L.J. A wind erosion prediction system to meet user needs. J. Soil Water Conserv. **1991**, *46,* 106–111.
69. Hagen, L.J. Evaluation of the Wind Erosion Prediction System (WEPS) erosion submodel on cropland fields. Environ. Modell. Software **2004**, *19,* 171–176.
70. Funk, R.; Skidmore, E.L.; Hagen, L.J. Comparison of wind erosion measurements in Germany with simulated soil losses by WEPS. Environ. Modell. Software **2004**, *19,* 177–183.
71. Hagen, L.J.; Wagner, L.E.; Skidmore, E.L. Analytical solutions and sensitivity analyses for sediment transport in WEPS. Trans. Am. Soc. Agric. Eng. **1999**, 42, 1715–1721.
72. Feng, G.; Sharratt, B. Sensitivity analysis of soil and PM_{10} loss in WEPS using the LHS-OAT method. Trans. Am. Soc. Agric. Eng. **2005**, *48,* 1409–1420.
73. Argaman, E.; Singer, A.; Tsoar, H. Erodibility of some crust forming soils/sediments from the Souuthern Aral Sea Basin as determined in a wind tunnel. Earth Surf. Processes Landforms **2006**, *31,* 47–63.

74. Montgomery, D.R. *Dirt: The Erosion of Civilizations;* University of California Press: Berkeley, California, 2007.
75. Fryrear, D.W.; Saleh, A.; Bilbro, J.D.; Schomberg, H.M.; Stout, J.E.; Zobeck, T.M. Revised Wind Erosion Equation (RWEQ), Tech. Bull. 1; Wind Erosion and Water Conservation Research Unit, Southern Plains Area Cropping Systems Research Laboratory: Lubbock, Texas, 1998.
76. Van Pelt, R.S.; Zobeck, T.M.; Potter, K.N.; Stout, J.E.; Popham, T.W. Validation of the wind erosion stochastic simulator (WESS) and the revised wind erosion equation (RWEQ) for single events. Environ. Modell. Software **2004**, *19,* 191–198.
77. Potter, K.N.; Williams, J.R.; Larney, F.J.; Bullock, M.S. Evaluation of EPIC's wind erosion submodel using data from southern Alberta. Can. J. Soil Sci. **1998**, *78*, 485–492.
78. Shao, Y.; Raupach, M.R.; Leys, J.F. A model for prediction of aeolian sand drift and dust entrainment on scales from paddock to region. Aust. J. Soil Res.**1996**, *34*, 309–342.

38

Erosion Control: Tillage and Residue Methods

Richard Cruse, Jerry Neppel, John Kost, and Krisztina Eleki

Tillage and Soil Erosion Processes

The processes of soil detachment, transport, and deposition occur during erosion. Soil surface conditions created by tillage greatly influence these processes and therefore soil erosion losses. Tillage practices that increase (or at least do not decrease) soil structural stability, leave plant residues on the soil surface, slow surface-water flow velocity, and/ or promote high infiltration rates favor soil conservation.

Surface residue cover greatly influences soil erosion. When 30% of the surface is covered, soil erosion losses are reduced by approximately 50% compared with a bare, tilled soil (Figure 1). Conservation tillage is considered to be any tillage system that has at least 30% of the soil surface covered by plant residues after planting[1] (Figure 1).

Tillage Effect on Soil Properties Critical to Erosion

Residues intercept raindrops and minimize soil detachment. This reduces soil available for transport and also limits surface seal development. This improves infiltration, which in turn reduces the amount of water runoff and transport potential. Residues also slow surface flow velocity, by acting like little dams on the surface. This causes soil deposition to occur on the upslope side of the residue pieces where water flow slows.[2]

Surface roughness and structural stability play dual roles. A rough surface stores water in the surface depressions between the clods or aggregates during heavy rainfall. This slows runoff and limits transport. The large pores of a rough surface also require more soil detachment to create a surface seal than a smooth surface, minimizing transport. Similarly, contour tillage, tillage occurring across the slope, reduces runoff by in-creasing surface water storage and slowing water runoff velocity.

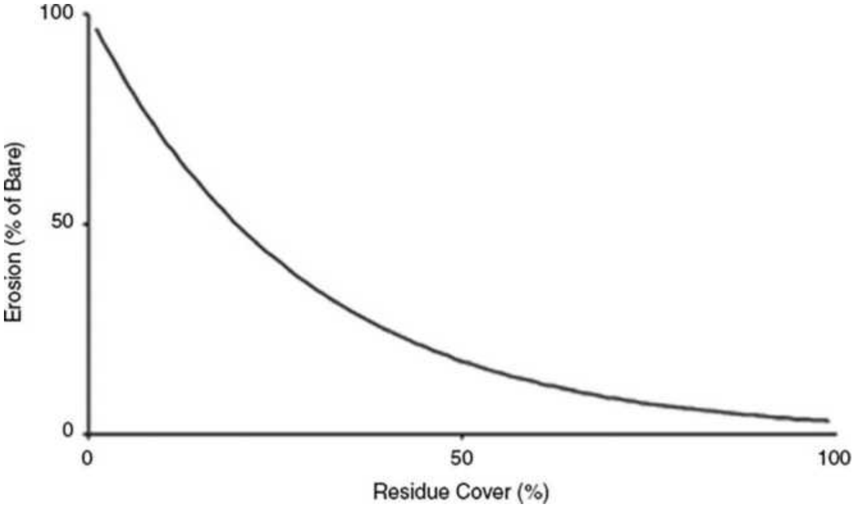


FIGURE 1 Effect of surface residue cover on soil erosion by water, expressed as the percent of erosion observed for bare soil. **Source:** Adapted from Renard et al.[2]

Open pores from the subsurface to the soil surface are critical for high infiltration rates and therefore low transport potential. Tillage practices that promote stable structure and result in surface residues to intercept raindrop impact promote stable open pores. Most tillage practices, however, weaken structure and therefore promote soil detachment from raindrop impact. Tillage also disrupts earthworm activity. Earthworms can play a major role in producing large open pores on the soil surface and very high infiltration rates with selected management systems. Four basic tillage/management systems will be discussed. Many variations of each system exist. Also, other systems using the principles described in this entry have been developed and can be located in other literatures.[3]

No-Till

No-till results in minor soil disturbance only during planting, leaving the greatest possible amount of surface residue after planting (Figure 2).

Compared to cleanly tilled systems, no-till can reduce erosion by as much as 95%.[4] The accumulation of residue from season to season reduces erosion by protecting the soil surface from impacting raindrops, as well as improving structural stability, pore size, and pore stability. Earthworm activity is promoted by no-till. Where large earthworm populations exist and the population is active on or near the surface, runoff can be very low or non-existent, even for large rainfall events.

No-till is best adapted for semiarid regions, sloping soils, and/or soils with good internal drainage. Low soil temperatures can be a problem in cooler climates, especially in early spring if soil is wet. One variation of no-till creates a residue- free band over the planted row with the planter. This helps the row zone warm and dry faster than if the soil is residue covered. In hotter climates, the reduced soil temperature caused by residue can be advantageous. In the absence of tillage, weed control is typically done through herbicide application. However, a combination of herbicide and cultivation for weed control can also be practiced.

Residue-covered surfaces also reduce soil water evaporation. No-till practices may conserve sufficient water under semiarid conditions (or on droughty soils) to significantly increase crop yield relative to that for other tillage methods (Figure 3).

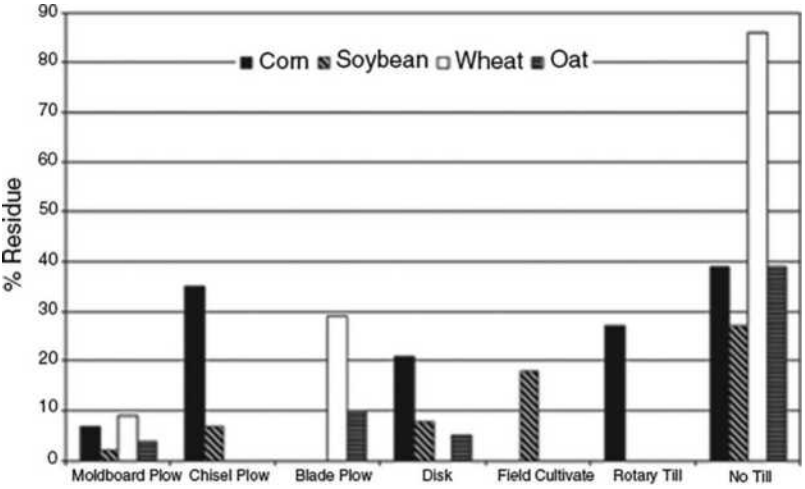


FIGURE 2 Percent residue by tillage method.
Source: Adapted from *Conservation Tillage Systems and Management,* Midwest Plan Service, Iowa State University, 1992.

FIGURE 3 No-till corn following wheat.

Ridge Tillage

With ridge tillage the soil surface "is left undisturbed from harvest to planting" except for strips up to one-third of the row-width. Planting is completed on the ridge and usually involves removal of the ridge top. Planting is completed with sweeps, disk openers, coulters, or row cleaners. Residue is left on the surface between ridges. Weed control is accomplished with crop protection products (frequently banded) and/or cultivation. Ridges are rebuilt during row cultivation.[1]

Ridge tillage historically is practiced on wetter, poorly drained soils in northern climatic row cropping regions.[5] However, it is also a viable option in semiarid, rain-fed row cropping regions where soil moisture conservation is a necessity.

FIGURE 4 Harvested corn on ridge-till soil.[6]

Selected aspects of ridge tillage systems protect the soil from soil erosion. Crop residues remain on the soil surface from harvest until planting of the succeeding crop. Only a portion of the soil surface is disturbed at planting. This leaves a high percentage of crop residues on the soil surface and maintains the large pores for rapid water infiltration. Crop residues from the row (ridge) are placed in the interrow (valley) area. This "extra" residue decreases detachment and slows water runoff, enhancing infiltration in the interrow area. Ridge tillage promotes controlled traffic, the practice of maintaining a fixed traffic pattern in the field, such that only certain interrow areas experience wheel-caused compaction. Controlled wheel traffic limits soil compaction to preselected interrows. The non-traffic areas maintain large pores and stable structure, enhancing infiltration. Ridge tillage is much more effective at conserving soil if ridging and planting are done on the contour than if up and down hill management is used.

Ridge till also offers opportunities to reduce weed control costs through banding of herbicides and row cultivation for weed management (Figure 4). Nutrient losses may be reduced by injection and/or subsurface application of fertilizers.[7] Also, ridged soil warms quicker in the spring than no-till soil, permitting earlier planting in many situations.[8]

Mulch Tillage

Mulch tillage is a full width conservation tillage system involving one or more soil loosening operations prior to planting. Mulch tillage maintains a substantial amount of plant residue cover before and after crop establishment. Tillage tools such as chisel plows, field cultivators, disks, or blades are typically used for primary tillage. Secondary tillage is minimized to conserve surface residue.[9]

Mulch can be from any crop material. It is normally retained on the surface during harvest of the previous crop. The amount of mulch left on the surface depends on the sequence of the tillage, tool(s) used, and the mulch material of the previous crop. In general, the higher the crop yield, the more surface residue will exist.[10]

Surface mulch reduces the evaporation of soil water, increasing soil water content, relative to that occurring with a bare surface. Consequently soil warming can be slower in the spring, which can slow plant emergence and early development. The higher soil water content can also favorably affect crop yield under dry conditions. Mulch tillage increases soil organic matter content compared to more intensive tillage systems (Figure 5). Weed control can be done mechanically, with herbicides, or with a combination of the two.

FIGURE 5 Mulch tillage procedure.

Strip Tillage

Strip tillage involves tilling only in the crop row zone. The interrow area is untilled with surface residue left undisturbed. Tillage may be done in fall or spring. However, fall tillage is more commonly done. Typically a row cleaner, coulter, shank, and covering disks till each row area. This combination of components is equally spaced to match planter rowspacing so that the succeeding crop is planted in the tilled zones.[11] This system offers the combined advantages of no-till in the interrow zone and conventional tillage in the planted zone. The tilled zone is normally warmer and drier than if no tillage were performed.

Weed control can be through herbicide application, cultivation, or a combination of these methods. Fertilizer can be applied during the tillage operation at the base of the tilled depth and/or during the cropping season. Strip tillage is used only for row crop production. Strip tillage on the contour is much more effective at conserving soil than planting up and down hill.

Summary

Tillage systems that leave residues on the surface, promote stable soil structure, and/or result in open pores to the soil surface favor water infiltration and soil conservation. Surface residue management is closely related to stable soil structure development and open surface pores. No-till, ridge tillage, mulch tillage, and strip tillage exemplify management systems that use these principles for favorable crop production and reduced soil erosion rates.

References

1. http://www.ctic.purdue.edu/Core4/CT/Definitions.html (accessed August 29, 2001).
2. Renard, K.G.; Foster, G.R.; Weesies, G.A.; McCool, D.K.; Yoder, D.C. Predicting Soil Erosion by Water: A Guide to Conservation Planning with the Revised Universal Soil Loss Equation (RUSLE), Agricultural Handbook No. 703; U.S. Department of Agriculture: Washington, DC, 1997; 384 pp.3. http://www.inform.umd.edu/EdRes/Topic/AgrEnv/ndd/agronomy/DEFINITIONS_OF_TILLAGE_SYSTEMS_FOR_CORN.html (accessed August 27, 2001).
4. Midwest plan service, water erosion. *Conservation Tillage Systems and Management,* 1st Ed.; Iowa State University: Ames, 1992; 9 pp.

5. Uri, N.D. Introduction. *Conservation Tillage in U.S. Agriculture: Environmental, Economic, and Policy Issues*; Food Products Press: Binghamton, NY, 1999; 1–4.
6. http://www.ctic.purdue.edu/Core4/CT/PhotosGraphics.html (accessed August 29, 2001).
7. Kanwar, R.S.; Colvin, T.S.; Karlen, D.L. Ridge, moldboard, chisel, and no-till effects on tile water quality beneath two cropping systems. J. Prod. Agric. **1997**, *10* (2), 227–234.
8. Radke, J.K. Managing early season soil temperatures in the northern corn belt using configured soil surfaces and mulches. Soil Sci. Soc. Am. J. **1982**, *46,* 1067–1071.
9. http://www.agric.gov.ab.ca/crops/wheat/wtmgt06.html (accessed August 27, 2001).
10. http://muextension.missouri.edu/xplor/agguides/agengin/ g01650.htm (accessed August 27, 2001).
11. http://www.ag.ohio-state.edu/~ohioline/aex-fact/0507.html (accessed August 29, 2001)

39

Pest Management: Modeling

Andrew Paul
Gutierrez

Introduction

Pest management is a key component of sustainable agriculture and may be defined as applied population ecology focusing on human managed populations of plants and/or domesticated animals and their pests and natural enemies in environments modified by weather and agro-technical inputs. The complexity of managing pests in an agroecosystem requires the development of models that enable separation of losses due to pests from yield variation due to weather and agronomic practice.[1] Such models provide the bases for evaluating the dual objectives in modern agriculture of *minimizing* inputs that cause adverse environmental, human, and animal health effects and *maximizing* net profits. In modern societies, these goals may result in conflicts between public and private interests, but this is usually not the case in subsistence ones[2] where the goal is often yield stability.

Some components of crop pest management research are illustrated in Figure 1a, but the same approach applies to medical and veterinary pest management. The methods fall under the ambit of agro-ecosystem analysis.[3] The analysis must be tri-trophic in scope because natural processes such as biological control, when correctly managed, may replace disruptive pesticide inputs that may induce resurgence of target pests, outbreaks of secondary pests, pesticide resistance, and pollution. Misuse of other agronomic inputs (e.g., fertilizers, water) may also exacerbate pest levels and must be considered in the pest management system. Several modeling approaches have been used in pest management, and they may be broadly classed as empirical, statistical, operations research, analytical, and simulation.[1–6] To be truly sustainable, the system must be ecologically, economically, and socially sustainable (Figure 1b).

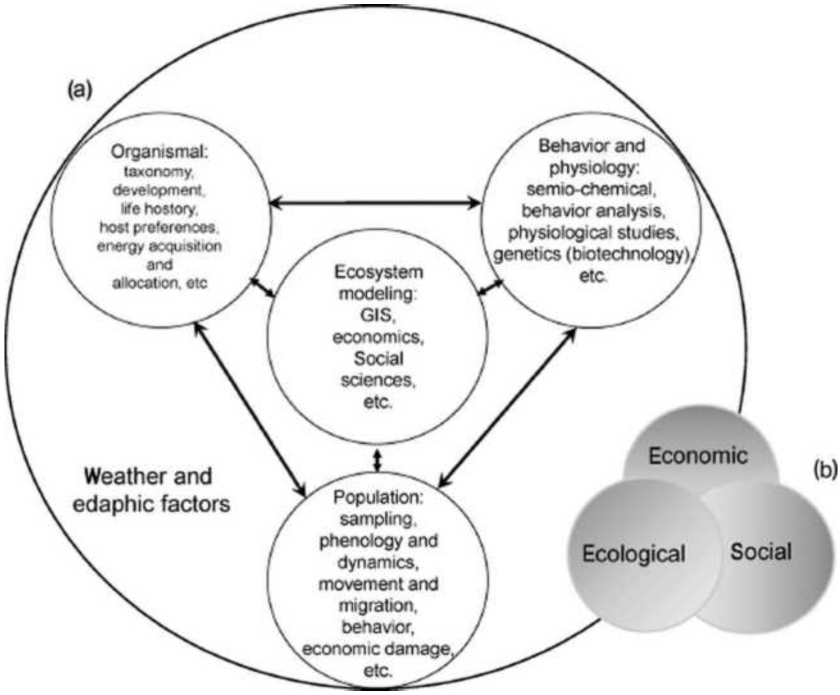


FIGURE 1 Components of pest management research using (a) systems analysis as a unifying tool and (b) components of sustainability.
Source: Courtesy of Center for the Analysis of Sustainable Agricultural Systems (CASAS).

Modeling Approaches

Empirical models based on trial and error have a long history and accumulated common wisdom. Traditional societies worldwide developed pest management strategies based on common wisdom that reduced pest damage and led to sustainable crop production. These systems may fail when modern agro-technical inputs are introduced and may require the accumulation of new experience to resolve management problem.

Population sampling and agronomic trials have been used to develop statistical models to assess the costs and benefits of pest levels and timing agronomic inputs (e.g., economic threshold studies). These models tend to be static and hence are time and place specific. They may not be able to be used to evaluate factors beyond the range of the data used to develop the model. Such models have, however, yielded useful pest management decision rules.

Attempts have been made to use operations research methods from engineering and economics to determine optimal pest management strategies. Such methods require sufficient knowledge of the system to formulate robust optimization models. Often this information is lacking, and/or the biology may be too complicated, requiring large reductions in the dimensions of the model. These simplifications often obviate their general utility except for general strategic analysis.

Pest management is *applied population ecology* and hence lends itself to demographic modeling methods. Mathematical models of the population dynamics may be simple descriptors of the system having analytical solutions or they may include considerable biological detail requiring numerical simulation. Analytical models may give strategic insights about the system, while richer simulation models may yield tactical recommendations for specific abiotic and biotic conditions. Modeling for tactical decision making requires a detailed understanding of the biology of the interacting species as modified by weather, edaphic, and agronomic factors. Biologically rich mathematical models are usually implemented as computer simulations.

There are many approaches to simulation modeling, but increasingly, mechanistic, physiologically based models of the energy dynamics of cropping systems are being developed because the important role of the plant as the integrator of all factors is being increasingly recognized. Energy is the currency of biology that when multiplied by price leads directly to economic models.

Simulation Models of Crop Systems

Crops are age-structured populations of plants each with *time-varying mass-age-structured* subunit populations of fruits, leaves, stems, and roots. The number in each age-mass class may vary overtime in response to factors that affect birth–death rates of whole plants and/or their subunits as well as those of higher trophic level populations. The crop may be viewed as a canopy of average plants with populations of pests and natural enemies within, or as individual plants each with its own set of pest and natural enemy subpopulations with migration between plants or fields (i.e., a meta-population). Modeling of individuals in pest and natural enemy populations (individual-based models) is not recommended in pest management because the decision rules for individual behavior are rarely known and the computations may be extensive. Simpler approaches that capture the *meta* details of physiologically based approaches may also be used.[7]

A basic premise of the physiological approach is that all organisms, including the economic one, face the same processes of resource (energy) acquisition and allocation.[1,8] Physiological models often assume energy allocation priority first to respiration (maintenance costs in economics), then reproduction (profit) and if assimilate (revenues) remains, then allocation to growth and reserves (infrastructure and savings). The shapes of the acquisition functions are concave (photosynthesis,

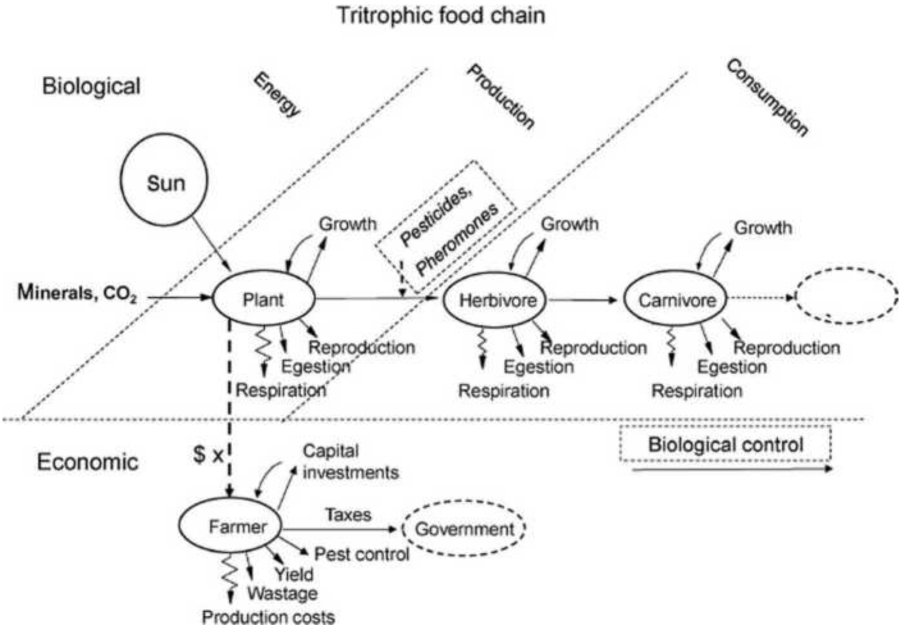


FIGURE 2 Analogies between trophic levels including the economic one.
Source: Courtesy of John Wiley and Sons, New York.

predations, etc.), and those for the maintenance costs are positive exponential (the Q_{10} rule) with the concave net of the two functions being the amount of resources available for allocation. These analogies[3] allow the use of the same model to describe the dynamics of all species (including the economic one) at the per-capita, population and regional levels (Figure 2). Each organism is assumed to try to satisfy a genetic (economic) demand for resources, but the process involves imperfect search, and the supply of a resource obtained is always less than (or equal to) the demand, and shortfall cause reductions in growth, reproduction, and survival rates from the maximum. In the model, all vital rates are controlled by one or more limiting supply/demand ratio with only the units and interpretation of the flow rates differing among species.

How biotic and abiotic factors affect plant growth and development is central to developing plant system models. These factors may affect either the supply (production) or the demand (sinks, e.g., fruits) side of the supply/demand ratio. Occasionally, both sides may be affected, and in other cases, the pests may attack the standing crop. The supply/demand paradigm simplifies model development, allowing assessment of yield loss and in some cases yield compensation in the face of pest damage. Such models facilitate evaluation of the costs and benefits of pest management options basic to the development of dynamic economic thresholds. Incorporating the biology of these factors in the model enables the development of highly realistic models.

Supply-Side Pests

Supply-side pests reduce the photosynthetic rate in various ways. Important supply-side pests are defoliators, sapsuckers, spider mites, nematodes, and diseases. Defoliation depletes leaves and may cause wound healing losses, but the effects on yield depend on the age of leaves attacked, the loss rate, and compensation due to increased light penetration to still intact leaves. In contrast, spider mites kill leaf cells, reducing photosynthesis in damaged leaves that are not shed, and reducing light penetration to lower leaves. Stem borers and vascular plant diseases may slow the photosynthetic rate by reducing the translocation of water and nutrients, and some may kill whole plants. Pests such as thrips and armyworms may damage the apical meristem, inducing developmental delays and reducing yield.

Plant Demand-Side Pests

Pests may attack fruit (e.g., net sinks), causing premature abscission, thereby altering present and future demands for photosynthate. High abscission rates may cause rank growth as the photosynthate is allocated to vegetative growth. Many plant species have reproductive capacity that allow varying degrees of compensation. Some plants may be determinant as all fruit are produced at roughly the same time (e.g., apple) and others are indeterminate (e.g., cotton), producing fruit until the maximum load is reached or the plant is limited by other factors. Of crucial importance in compensation is the time and energy lost in abscised fruits. Little time and energy may be lost when new buds and small fruit are abscised, and in indeterminate species such as cotton, replacement buds may be produced at rates sufficient for complete compensation. In determinate species such as apple, the time for compensation may be very short and compensation may be impossible if the attack rate is too high. In both growth types, attacks on older fruits may involve considerable losses in time and energy that often precludes compensation.

The ratio of the cumulative buds initiated to the cumulative numbers abscised may provide the basis for determining whether and how much compensation is possible. Such data yield a concave function that estimates the *compensation point* and may provide a rule of thumb for estimating the economic threshold. For example, losses of 30% of fruit bud in many cotton varieties do not affect yield.

Other Kinds of Plant Pests

Some pests affect both sides of the supply/demand ratio. For example, pests attacking the apical meristem of a plant or branch kill primordial tissues that introduce time delays and alter future fruit and vegetative dynamics. Other pests may attack the standing crop of fruit without causing fruit to shed and without appreciable reductions in sink demands, making plant compensation in such cases unlikely. Furthermore, damage in such cases accrues over time.

Modeling the Impact of Biotechnology

Increasingly, crop plants are being genetically modified to protect them from insect pests and to make them tolerant to herbicides designed to kill weed competitors, and these developments have added an important genetic level to modeling pest management. One of the most common genetic modifications is the insertions of genes from the bacterium *Bacillus thuringiensis* into transgenic crops for the production of Bt toxin. This innovation is thought to be environmentally friendly, and numerous recent economic studies concluded high benefits and good prospects for the technology (mainly in cotton and maize) in the United States, Australia, Argentina, China, South Africa, and other parts of Africa and Asia.[9] In many of these benefit studies, econometric methods were applied to cross-sectional data from farm surveys or experimental data, and while these methods can provide a good assessment of the static productivity of pest control agents (and other inputs), they are less suitable for capturing the interaction between control decisions and dynamic ecosystem reactions. The complexity introduced by biotechnology is exemplified by the cotton system: There are more than a dozen herbivore pests in cotton [e.g., pink bollworm, tobacco budworm, cotton bollworm, rough bollworm, fall armyworm, beet army-worm, cabbage looper, soybean looper, plant bugs (*Lygus* spp.), whitefly, cotton bollweevil, spider mites, and others], and each of them is affected differently by Bt toxin(s). In addition, their natural enemies may also be affected but at levels considerably less than by pesticides.[10] This complexity was examined using modeling that suggested the potential for secondary outbreaks of pests (i.e., plant bugs) induced by the use of Bt cotton.[10] Increases in secondary pests in Bt cotton have occurred in the Southeastern United States, India, and China, [11] with concomitant increased use of insecticides.[12] The pyramiding of new technologies (more Bt and other toxins) adds more complexity to the analysis. Modeling studies in transgenic crops are also made difficult by intellectual property rights that limit field scientific investigation.

The use of herbicide-resistant crop has increased herbicide use, led to the development of herbicide-resistant weeds, and increased environmental pollution and nontarget effects (e.g., amphibious species).[13] Rather than simplifying the pest management system, biotechnology has greatly increased the complexity in the long run.

Regional Pest Management

Modeling, geographic information system (GIS), and marginal analysis may be used to assess the regional distribution and abundance of pest and their biological control agents due to weather and over the long run of climate. Figure 3a illustrates how the same model at the individual, tri-trophic, field, area, or regional level applies. Weather data [daily maximum and minimum temperatures (°C), solar radiation (kcal $cm^{-2}day^{-1}$), rainfall (mm), daily runs of wind (km day^{-1}), and relative humidity] for each location on a grid (or numerous locations) are used to run the model over given time periods (days, months, years). Appropriate summary variables from the model may be mapped using GIS technology (Figure 3b) and used to show the aggregate regional effects of weather on biological system dynamics.[14] The GIS simulation data may also be analyzed using various statistical methods

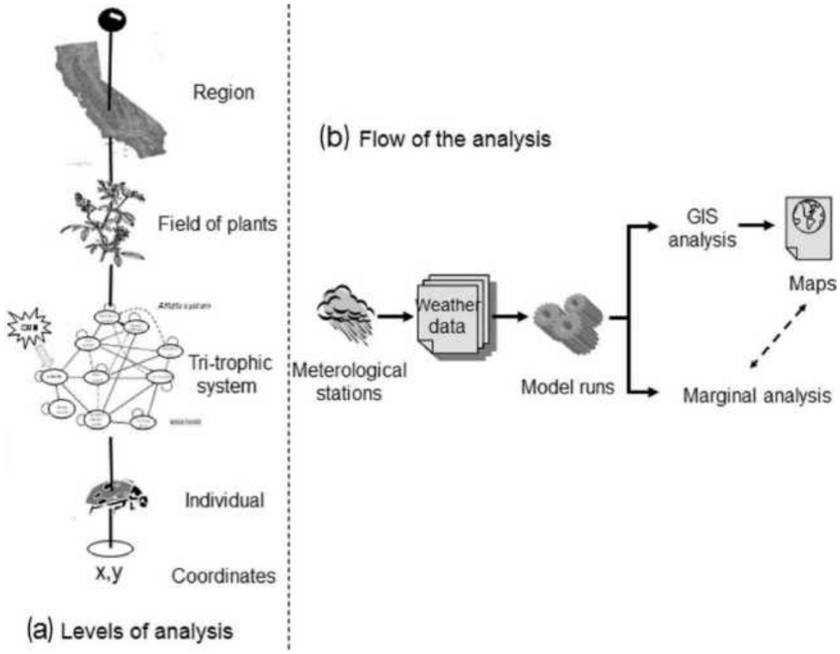


FIGURE 3 (a) Application of the same dynamics model (see Figure 2) at the individual, population, field, or larger geographic area, and (b) the use of the simulation model in regional GIS and marginal analysis.

(e.g., linear multivariate regression), keeping only independent variables and interactions with slopes significantly different from zero (t values, $p < 0.05$) in the statistical model. As in econometrics, marginal analysis may be used to estimate the average magnitude and direction of large effects ($\partial y / \partial x_i$).

Conclusions

Interactions among pests and higher trophic levels (top–down effects) are complicated, time varying, and best integrated via the plant's physiology and dynamics as modified by abiotic factors (bottom–up effects). Tri-trophic models based on the same supply/demand paradigm enable assembling and understanding the top-down biological relationships.[15,16] These models provide the bases for developing ecologically sound pest management strategies that may be implemented as part of a total farm expert system or regionally using GIS technology to develop region-wide pest management strategies. All of these methods fall under the ambit of agro-ecosystems analysis.

References

1. Gutierrez, A.P., Ed. *Applied Population Ecology: A Supply-Demand Approach*; John Wiley and Sons, Inc.: New York, 1996.
2. Altieri, M.A. Sustainable agriculture. Encycl. Agric. Sci. **1994**, *4,* 239–247.
3. Gutierrez, A.P.; Curry, G.L. Conceptual framework for studying crop-pest systems. In *Integrated Pest Management Systems and Cotton Production;* Frisbie, R.E., El-Zik, K.M., Wilson, L.T., Eds.; John Wiley and Sons: New York, 1989; 37–64.
4. Curry, G.L.; Feldman, R.M., Eds. *Mathematical Foundations of Population Dynamics,* Series No. 3; TEES Monograph, 1987, Texas A&M Press, College Station, Tx.

5. deWit, C.T.; Goudriaan, J., Eds. *Simulation of Ecological Processes,* 2nd Ed.; PUDOC Publishers: the Netherlands, 1978.
6. DiCola, G.; Gilioli, G.; Baumgärtner, J. Mathematical models for age-structured population dynamics. In *Ecological Entomology,* 2nd Ed.; Huffaker, C.B., Gutierrez, A.P., Eds.; John Wiley and Sons: New York, 1998.
7. Gutierrez, A.P.; Daane, K.M.; Ponti, L.; Walton, V.L.; Ellis, C.K. Prospective evaluation of the biological control of the vine mealybug: Refuge effects. J. Appl. Ecol. **2007**, *44*, 1–13.
8. Regev, U.; Gutierrez, A.P.; Schreiber, A.P.; Zilberman, D. Biological and economic foundations of renewable resource exploitation. Ecol. Econ. **1998**, *26,* 227–242.
9. Qaim, M.; Zilberman, D. Yield effects of genetically modified crops in developing countries. Science **2003**, *299,* 900–902.
10. Gutierrez, A.P.; Adamcyzk, J.J. Jr.; Ponsard, S. A physiologically based model of *Bt* cotton-pest interactions: II. Bollworm–defoliator–natural enemy interactions. Ecol. Modelling **2006**, *191,* 360–382.
11. Lu, Y.; Wu, K.; Jiang, Y.; Xia, B.; Li, P.; Feng, H.; Wyckhuys, K.A.G.; Guo, Y. Mirid bug outbreaks in multiple crops correlated with wide-scale adoption of Bt cotton in China. Science **2010**, *328,* 1151–1154.
12. Pemsl, D.; Gutierrez, A.P.; Waibel, H. The economics of biotechnology under ecosystems disruption. Ecol. Econ. **2007**, *66*, 177–183.
13. Relyea, R.A. The lethal impact of roundup on aquatic and terrestrial amphibians. Ecol. Appl. **2005**, *15* (4), 1118–1124.
14. Gutierrez, A.P.; Ponti, L.; Ellis, C.K.; d'Oultremont, T. Analysis of climate effects on agricultural systems: A report to the Governor of California sponsored by the California Climate Change Center, **2006**, available at http://www.climatechange.ca.gov/climate_action_team/reports/index.html.
15. Gutierrez, A.P.; Neuenschwander, P.; van Alphen, J.J.M. Factors affecting the establishment of natural enemies: Biological control of the cassava mealybug in West Africa by introduced parasitoids: A ratio-dependent supply–demand driven model. J. Appl. Ecol. **1993**, *30,* 706–721.
16. Graf, B.; Gutierrez, A.P.; Rakotobe, O.; Zahner, P.; Delucchi, V. A simulation model for the dynamics of rice growth and development. II. Competition with weeds for nitrogen and light. Agric. Syst. **1990**, *32,* 367–392.

40

Soil Quality: Indicators

Graham P. Sparling

Criteria for Selecting Soil Quality Indicators

Many soil properties and characteristics have been proposed as indicators of soil quality.[1,2] In order to be considered as a valuable indicator, a soil property must meet certain criteria: 1) it needs to get our attention, and 2) it needs to tell us something useful about the condition of the soil. Desirable attributes include being quantitative and responsive within the specified time scale; interpretable; cost effective; scientifically justifiable; socially acceptable; internationally recognized; and preferably a part of historical monitoring procedures.[1–4]

There is some overlap between the categories of soil quality and land quality, of which soil quality is a subset.[2,5] Indicators for land quality generally describe the inherent and intrinsic characteristics of the soil, usually on a landscape scale. Inherent characteristics, such as soil mineralogy and texture, soil depth, and stoniness, are used to assess land capability or suitability for use, and generally are not greatly affected by land management. Other inherent characteristics, such as slope and aspect, overlap with soil quality indicators because they influence the more rapidly changing and dynamic soil properties that are responsive to land management. The variety of proposed indicators reflects the many ways in which soil quality has been defined[5] and the many components of a soil's chemical, physical, and biological attributes that contribute to its overall character. Of its many definitions, quality can be most briefly described as "fitness for use."[1,5] Within a soil context, a quality rating depends on how well the soil characteristics match the suitability for a particular use. Soil characteristics are not fixed—what is good soil quality for one land use may be deemed poor soil quality for another use. Thus, quality ratings are not absolute, but are instead values based on interpretation depending on the intended use of the soil.[6] The soil itself does not change, only our perception of it in relation to human needs.

A Minimum Data Set

Because it is not feasible to measure all soil quality characteristics, indicators need to be selected to address the issues of greatest concern. There are three major categories of soil quality concerns: soil erosion and redistribution; chemical and biological contamination; and soil degradation and depletion. The examples of these categories given in Table 1 are of concern because they adversely influence the role of the soil in ecosystem functioning and services.[5] Their degree of importance varies in different parts of the world.

TABLE 1 Major Aspects of Soil Quality Requiring Indicators

Soil Quality Aspect	Examples
Erosion and deposition	Soil redistribution by wind and water, mass movement, slips and slumps, and tillage displacement
Contamination/pollution	Presence of potentially toxic chemicals Pathogenic organisms
Degradation/depletion	Depletion of soil organic matter Loss of fertility Soil acidification Salinization Structural degradation

Time and Spatial Aspects

Selected indicators must be appropriate to the scale being considered, within both spatial and time dimensions. We tend to classify these items based on human perceptions, and soil processes that geologists consider rapid may be considered slow or relatively static to soil biologists. Examples of time and spatial scales and some relevant soil quality characteristics are shown in Table 2. Human-induced changes to soil, including soil redistribution through erosion and deposition, organic matter depletion, and chemical and radioactive contamination, can take hundreds of years to remediate.

TABLE 2 Spatial and Time Scales of Soil Properties Important for Soil Quality

Spatial Scale	Time Scale	Soil Properties Showing Change	Soil Quality Aspects
		Rapidly changing, highly dynamic characteristics:	
1–100 mm	Minutes and hours	Moisture content and temperature	Moisture and warmth for plant growth and biological activity
100 mm-1 m	Hours and days	Infiltration and drainage Biological activity	Compaction due to treading or wheeled traffic, risk of run-off during rainfall, waterlogging in poorly draining soils, poor root environment Decomposition of fresh organic matter
		Intermediate characteristics:	
1 m–100m	Days and months	Available nutrients and mineralization Soil structure	Nutrient depletion and supply, nutrient imbalances for plant growth Soil damage from compaction; influence on root environment, aeration, moisture retention, and drainage
100 m–1 km	1–5 years	Organic matter content Salinity Acidification Erosion and redistribution	Turnover of moderately decomposable organic matter, nutrient supply Saline water in root zone, toxic stress Soil chemical process leading to acidification in root zone, nutrient imbalances, toxicity Loss of topsoil, soil deposition
		Slowly changing characteristics:	
1km–10km	5–25 years	Soil loss and formation Humus formation or loss Salinity Acidification Weathering	Loss of topsoil and decrease in soil depth, soil deposition Changes in soil structure and nutrient storage due to loss of organic matter Salinity in the root zone, toxic stress, rising water tables Increasing acidity due to leaching and cation/anion imbalance, toxic stress Nutrient release and soil formation from rapidly weathering minerals
10 km–100 km	25–100 years	Stable organic matter	Humus provides exchange sites to retain ions in soil and contributes to soil aggregate formation
>100 km	>100 years	Soil formation Mineralogy	Nutrient release and soil formation from slowly weathering minerals Proportions of sand, silt, and clay, exchange characteristics of the clays

Interpretation

An interpretative framework is needed to make sense of the numeric values obtained for the various soil characteristics. We need to know what constitutes a high or low value, and what target value is desirable for each particular soil and land use. An interpretive framework should also indicate whether there might be off-site consequences for the wider ecosystem. For example, soluble fertilizer contents in soil may benefit crop production (good quality), but not matching the fertilizer to plant needs may result in an excess, increasing the risk of contaminating receiving waters (poor quality). In general, interpretative frameworks are better defined for contaminants than for other soil quality measures, because the presence of a contaminant can be related to an increased risk to plant and animal health. A critical value can be defined based on toxicity. Such dose–response curves are not available for most indicators used for soil loss, deposition, or soil depletion, and in many cases the response curve is an ill-defined continuum rather than a critical point.[2] Much work still needs to be done to define justifiable targets for soil properties, and even the broad categories of "more is better," "less is better," or "optimum range" have been the subject of ongoing discussion.[7] The units of expression can also influence interpretation—expression on a volume or area basis is generally preferred, particularly for comparisons between soils and land uses with differing bulk densities.[8]

Trends

In the absence of defined critical values for soil quality indicators, the importance of the trend over time has been emphasized.[9] Trends away from a target range are interpreted as a decline in soil quality; trends toward the target are interpreted as improved soil quality. However, it can also be argued that short-term trends away from a target value could still be regarded as sustainable, provided the trend can be reversed within an acceptable time frame. Within the present context, recovery or restoration within a human generation of 20–25 years could be regarded as preserving intergenerational equity.[10] Such exploitation-recovery cycles were much used during the shifting agriculture phase of human development. Soils differ in their resistance to change, and the rates at which they re-cover[11,12] affects their soil quality rating. Generally, soils that are resistant to change, or do change but show a rapid rate of recovery, are regarded as having better quality.[11]

Combining indicators

Because multiple data sets are difficult to comprehend, it has been proposed that indicator sets be combined into a single index.[1,2] These single indices are usually additive or multiplicative combinations of the individual indicators, and a weighting factor is used to adjust for the relative importance of each factor.[13] While a single number is easier to remember and communicate, its use has the disadvantage of possibly masking information. When too many indicators are combined, poor-quality soil aspects may be masked by those of modest or good quality. Reporting on any soil environmental quality should reflect the limiting quality rather than the average value. For example, excellent chemical fertility is of little use if the physical structure of the soil is badly degraded. This poor physical structure would therefore dominate the soil characteristics.

Conclusions

Indicators must attract attention and inform. They need to be sufficiently sensitive for their intended task and to respond in a timely manner. Critical values can be set so that a particular indicator value will trigger a management response. The critical values will differ, depending on the soil and land use, and management options. The ability of soils to recover, and their susceptibility to change under use, will define which indicators are useful and where their critical limits may be set. Multiple indicators can be combined into a single index, using weighting factors to adjust for the relative importance of each indicator. When combining indicators, care is needed not to mask poor indicator scores, and a safer option is to identify which soil characteristics are limiting factors.

References

1. Doran, J.W.; Parkin, T.B. Defining and assessing soil quality. In *Defining Soil Quality for a Sustainable Environment,* SSSA Special Publication No. 35; Doran, J.W., Coleman, D.C., Bezdicek, D.F., Stewart, B.A., Eds.; Soil Science Society of America: Madison, Wisconsin, 1994; 3–21.
2. Karlen, D.L.; Mausbach, M.J.; Doran, J.W.; Cline, R.G.; Harris, R.F.; Schuman, G.E. Soil quality: a concept, definition, and framework for evaluation (Guest Editorial). Soil Sci. Soc. Am. J. **1997**, *61,* 4–10.
3. Doran, J.W.; Coleman, D.C.; Bezdicek, D.F.; Stewart, B.A., Eds. *Defining Soil Quality for a Sustainable Environment,* SSSA Special Publication Vol. 35; Soil Science Society of America: Madison, Wisconsin, 1994; 244.
4. Cornforth, I.S. Selecting indicators for assessing sustainable land management. J. Environ. Manage. **1999**, *56,* 173–179.
5. Carter, M.R.; Gregorich, E.G.; Anderson, D.W.; Doran, J.W.; Janzen, H.H.; Pierce, F.J. Concepts of soil quality and their significance. In *Soil Quality for Crop Production and Ecosystem Health*; Gregorich, E.G., Carter, M.R., Eds.; Elsevier: Amsterdam, 1997; 1–19.
6. Singer, M.J.; Ewing, S. Soil quality. Interdisciplinary aspects of soil science. In *Handbook of Soil Science*; Sumner, M.E., Ed.; CRC Press: New York, 1999; 272–298.
7. Sojka, R.E.; Upchurch, D.R. Reservations regarding the soil quality concept. Soil Sci. Soc. Am. J. **1999**, *63* (5), 1039–1054.
8. Reganold, J.P.; Palmer, A.S. Significance of gravimetric versus volumetric measurements of soil quality under biodynamic, conventional, and continuous grass management. J. Soil Water Conserv. **1995**, *50* (3), 298–305.
9. Larson, W.E.; Pierce, F.J. The dynamics of soil quality as a measure of sustainable management. In *Defining Soil Quality for a Sustainable Environment,* SSSA Special Publication No. 35; Doran, J.W., Coleman, D.C., Bezdicek, D.F., Stewart, B.A., Eds.; Soil Science Society of America: Madison, Wisconsin, 1994; 37–51.
10. Lunney, D.; Pressey, B.; Archer, M.; Hand, S.; Godhelp, H.; Curtin, A. Integrating ecology and economics: illustrating the need to resolve the conflicts of space and time. Ecol. Econ., **1997**, 23, 135–143.
11. Greenland, D.J.; Szabolcs, I. *Soil Resilience and Sustainable Land Use*; CAB International: Wallingford, England, 1994.
12. Seybold, C.A.; Herrick, J.E.; Brejda, J.J. Soil resilience: a fundamental concept of soil quality. Soil Sci. **1999**, *164,* 224–234.
13. Andrews, S.S.; Karlen, D.L.; Mitchell, J.P. 2002 A comparison of soil quality indexing methods for vegetable production systems in Northern California. Agric. Ecosyst. Environ. **2002**, *90,* 25–45.

V

ELE: Focuses on the Use of Legislation or Policy to Address Environmental Problems

41

Acid Sulfate Soils: Management

Michael D. Melville
and Ian White

Introduction

Acid sulfate soils (ASSs) is the name given to all those soils or unconsolidated sediments that contain reduced sulfur minerals and compounds that have been, or have the potential to be, oxidized and thereby produced acidity.[1–5] They occur globally and are an important problem requiring careful management. In our earlier publication on these soils,[6] we stressed the general coastal location of these soils, which contain a range of materials due to the sulfate in seawater that provides a large reservoir of dissolved sulfur that can be reduced biochemically in the presence of organics to form various monosulfides, and eventually the disulfide, pyrite (FeS_2). However, it is now clear that we should also recognize materials and problems from sulfidic hard rocks, generally freshwater bodies, and landscapes in inland areas. This need has been admirably illustrated by Fitzpatrick and Shand,[7,8] working on a wide range of sites and issues in Australia, but the principles and examples that they illustrate have global application.

The biochemical processes in ASSs are analogous to those of acid mine drainage (AMD)/acid rock drainage (ARD), and the rain-induced drainage from some mine sites with these materials also have analogues in the hydrology of acid sulfate soil (ASSs). Nevertheless, there are important differences between problems and management with AMD/ARD and other ASS landscapes, mostly because of their respective scales, material, and mineral characteristics, but particularly in their respective economics that affect the viable management options.

There are three fundamental questions underlying the use of ASS landscapes. Firstly, how best to use the existing environment productively while minimizing any sulfide mineral oxidation. Secondly, how to effectively neutralize any existing or newly created acidity formed during a particular land use. Thirdly, how to eliminate or minimize downstream environmental impacts from ASS drainage waters, which may include deoxygenation of soil or surface waters and the release of metals and nutrients. Therefore, best management of ASS must address these questions.

Existing and Emerging Problems

Much of the early research on ASS, as illustrated in the contents of the early series (from 1973 until 1992) of international symposia organized by ILRI (International Institute for Land Reclamation and Improvement) from Wageningen, the Netherlands, was directed towards improving their agronomic usefulness. However, the 2002 international ASS conference in Australia emphasized the range of environmental problems with ASS and their best management, and the most recent (2008) conference in Guangzhou, China, continued that theme and was combined with research and management for ARD.

While ASSs were recognized in Europe more than 250 years ago,[1] appreciation of their existence and importance in terms of environmental impacts in Australia was limited.[9] The most important ecological impact that changed this appreciation was the 1987 acidity discharge event on the Tweed River in northeastern New South Wales. Twenty-three kilometers of the Tweed River were completely clarified and sterilized of fish, crustaceans, and most benthic organisms after the flooding caused acidic drainage discharges from ASSs of the river floodplain. It was the astute observations of a local entomologist, published in an amateur fisherman's magazine,[10] which identified the true cause of this devastating event that took up to 18 months to heal.[6]

While Australia was slow to become involved in ASS research and management, the Tweed River fish kill initiated broad interest and involvement of stakeholders, researchers, industry groups, regulators, and politicians. The pathway of change with these groups on the issue of ASS in Australia has not always been straightforward without conflict; neither has it been simple.[9] However, a generally very positive outcome has been achieved and Australia now probably provides much of the impetus and methods for global adoption of best management practices with respect to ASS. Each State and the National Commonwealth of Australia have adopted policies and practices for ASS best management. Much of the materials necessary for the dissemination of information on ASS are freely available on the web.[7,23–25] The Queensland Government publications are updated from time to time. A quarterly newsletter, *ASSAY,* which has now been in publication for about 13 years, is available for free electronic subscription.[11]

As we pointed out in White,[9] many coastal floodplains around the world are composed of Holocene-age (<10,000 years BP) sediments that contain iron sulfide minerals,[2] mostly as pyrite, but in some regions, monosulfides are important.[12] It is estimated that the global extent of such ASS is between 10^7 and 10^8 ha, with large deposits in Southeast Asia, the Far East, Africa, and North and South America.[13] The total extent of ASS now mapped in Australia[14] is about 2.2×10^6 ha, of which about 0.6×10^6 ha is coastal and 1.6×10^6 ha is inland.[12] When exposed to air by natural or human-induced drainage, sulfides in these materials oxidize to sulfuric acid that may leach iron, aluminum, and other metals into sediment pore waters. Drainage of these pore waters has significant detrimental ecological impacts,[2,10,15,16] corrosion of engineering infrastructure,[17,18] blooms of cyanobacteria,[19–21] and emission of toxic and greenhouse gases.[22,23] The existence and problems of non-coastal ASS, and the need for considering hard rock sulfidic materials, increase the scope of management strategies.[8,25]

There are many global issues that relate to ASS and that are likely to emerge in the future, but there are two that seem to us are particularly problematic. Firstly, with global warming is the possible oxidation of sulfidic material in organic-rich landscapes presently preserved or protected by permafrost in high latitudes of the Northern Hemisphere (e.g., around James and Hudson Bays in Canada). Drainage of oxidation products and emission of greenhouse and sulfurous gases are likely. Secondly, perhaps also to some extent due to global warming but from population growth and political decisions for development, the increasing upstream abstraction of water resources from major rivers so that delivery of freshwater to coastal deltas or estuaries decreases markedly and wetlands underlain by sulfidic materials or sediments drain and oxidize and/or experience subsidence and increased saline intrusion (e.g., the Mekong Delta of Vietnam, and with the Brahmaputra and Ganges Rivers into Bangladesh). Any sea level rise associated with global warming increases this tidal inundation risk.

In southeastern Australia, the severe drought from 2005 to about 2009 markedly reduced the delivery of discharge to the outflow lakes area of the Murray-Darling Basin (MDB). This caused oxidation of the sulfidic materials and sediments that had accumulated during the various water management regimes that were initiated in the lakes and supply rivers following European settlement (Fitzpatrick and Shand,[8] pp. 38–43). Many different management regimes have been proposed and begun to be implemented to overcome these problems. However, it seems clear that licensing of water abstraction, predominantly for agricultural irrigation in the midcatchment during the multidecadal "flood-dominated rainfall regime" after 1946,[27] exceeded the capacity of the river system to supply water to the outflow lakes and maintain their health when a "drought-dominated rainfall regime"[27] possibly began.

In western Finland, the rapid isostatic uplift (<12 mm/yr) from retreat of the last northern European glacial sheet is exposing sulfidic sediments to natural oxidation, but this is being exacerbated by artificial drainage systems to enable agricultural land uses.[28] Drainage from these ASS landscapes is markedly polluting streams and waters of the Gulf of Bothnia, and it is estimated that the heavy metal pollution from this source is equal to that from all Finnish secondary industry.

It is clear that management of landscapes in which ASSs occur must make due cognizance of the landscape's ability to accommodate human impacts and probable variability in climatic or geomorphic conditions.

In recent years, the distinction between the characteristics, processes, problems, and management options of ASS and ARD have almost disappeared. Soil scientists such as Fitzpatrick and his colleagues are researching on what might previously have been considered ARD issues [e.g., see Fitzpatrick and Shand[8] (pp. 31–37) and Skwarnecki and Fitzpatrick[29]], and management techniques used commonly in AMD sites, such as capping to prevent sulfide oxidation and constructed wetlands for acidity dis-charge neutralization,[30] have been used successfully for ASS management.

Management Approaches

We have previously described[9] how early major problems from ASS on East Coast Australia created significant conflict between fishermen, sugarcane farmers, and environmentalists, and how this was resolved through a cooperative learning and coastal stewardship approach. This began the process that has underlain Australia's becoming a global leader in the understanding and management of ASS.

In complex environmental situations and problems, such as are likely to occur with ASS, significant knowledge gaps are likely to exist as to what is the problem and how best to manage it. Nevertheless, action is thought necessary. Adaptive management is sometimes proposed[9] on the basis that any mistakes due to incorrect information can be corrected provided there is rigorous monitoring. However, adaptive management assumes linear processes where any mistakes can easily be reversed. Unfortunately, with many situations of environmental management, including with ASS, processes are not linear (and are probably generally hysteretic) and environments are susceptible to dramatic collapse. Therefore, environmental degradation is not easily reversed.[31,32]

Many of the problems arising with ASS are the outcome of environmental change from natural processes, anthropogenic activities, or some combination of these. Within the population affected by such environmental changes, management of problems arising is best achieved if the reasons for the change are understood. Appropriate, reliable information, communicated in a relevant way, can provide an important catalyst for changed attitudes. The challenge is to collect and communicate such complex information in a manner that is trusted and accepted by all. Therefore, the first step is to describe the nature and cause of the problem and indicate some likely consequences, but without causing further and new concerns. The next step needs to identify and bring together all stakeholders in a participatory way so that consensus and compromise agreement can be reached.[33] This may require the input of some external, trusted facilitators. Strangely, because many see these personnel as rather "ivory towerish," universities can fulfill this role. Unfortunately, government employees are often unable to take this role because they have a regulatory role in environmental management.[25] In Australia, the problems of managing coastal resources have come from the many top-down, conflicting visions and disparate goals of successive governments and individuals. These goals, past policies, and legislation have concerned issues such as environmental protection and/or rehabilitation, economic development, and regional employment and growth. Politics is an ever-present factor.

Environmental stewardship has been proposed as one way by which conflicts might be reduced so that ownership and pride in the environment and its heritage are encouraged. In Australia, the indigenous original owners of the land have a culture of custodial stewardship and knowledge that can be used for best management practices, but the early European settlers were unable or unwilling to accept this. Now, theirs is a voice that must be heard in environmental management, including that with ASS. Stewardship involves voluntary compliance, strong commitment, and a willing participation in the sustainable use of resources through wise practices. There are probably well-founded concerns that without strong underpinning regulations, voluntary compliance agreements do not have effective mechanisms to address persistent breaches of agreements.

The best management of ASS involves a range of activities that address the issues of minimizing the export of acidity, metals, and nutrients into downstream environments. This involves two issues: 1) the creation of new acidity by sulfide mineral oxidation and 2) the management of existing acidity in the landscape. The latter has been frequently ignored. Across the range of land uses of northern NSW, the existing acidity in the sulfuric layer averages approximately 50 tons of sulfuric acid per hectare.[34] Potential acidity represented in the sulfidic minerals of the deeper subsoil is many times this amount. Nevertheless, the annual discharge of acidity is <0.5 tons/ha.[35] The degree to which soil acidification has been caused by artificial drainage is uncertain but natural processes are also involved.[36] It is clear, however, that artificial drainage networks, which decrease by orders of magnitude the time of inundation of floodplain backswamps from their natural conditions, provide the conduit by which acidity is transferred rapidly to the estuary and potentially causes downstream impacts.[37] Decreasing the density of drains has a major impact on the export of acidity. Laser grading provides a technique for removing surface waters and decreasing drainage density and allows more land to be planted to crops.[37]

The actual impact caused by rain-induced acidic discharges depends upon the magnitude of the discharge relative to the dilution by upland flows and neutralization capacity of receiving waters. Land users on coastal ASS floodplains must appreciate that other parts of the estuarine ecosystem depend on the dissolved alkalinity of the receiving water, consumed during acidity neutralization. Floodplain land uses should be undertaken so as to avoid creating any new acidification and minimizing the export of any acidity.

Overall, best management of ASS involves four possible approaches: 1) education and assessment; 2) avoidance; 3) oxidation prevention; and 4) acidity containment and neutralization. These approaches for ASS management have been included to a varying extent by cases described in Jiggins[31] and Harris[32] and detailed in Queensland's ASS Management Guidelines.[24] The latter work also provides insight to the risks associated with the various management approaches. ASS management is not necessarily straightforward and amenable to a prescriptive set of firm rules. Each project must align with the particular existing and proposed environmental situations.

Education and Assessment

Best management of ASSs requires knowledge of their distribution; the depth of the sulfidic material from the soil surface; the acidity stored in sulfuric material; the hydrological behavior of the soil profiles, landscape, and drainage system; the climatic regime; and the magnitude, tidal characteristics, and water quality of the receiving waters.[37] All of these are infrequently considered.

Australian soil scientists were generally slow in appreciating both the presence and the environmental problems from ASSs, and in raising the awareness of the public, land managers, and policy makers. The first national symposium on ASSs was only in 1993.[8] Nevertheless, in the past 20 years, a major shift has occurred, and now all States and Territories, and the Commonwealth of Australia, have included

ASS management in their environmental policies. Australia is one of the few countries with a national strategy on ASSs. The shift in New South Wales (NSW) has been greatly helped by publication of "Acid Sulphate Soil Risk Maps"[38] for the entire NSW coastline (scale 1:25,000). These maps are being used as the basis for landuse planning instruments ("Local Environment Plans," or LEPs) of local government authorities. These LEPs require submission of a Development Application for any activity that disturbs more than 1 ton of ASS.

The sugar industry, a major user of ASSs in NSW, has been granted State-wide exemption from these requirements for normal farming and drain cleaning activities because each of the 700 cane growers has signed a contract to comply with an approved code of best management practice for ASSs. Less than 20 years ago, the sugar industry was in a state of denial about the existence and problems of ASSs in their cane lands. The industry's new sense of land stewardship with ASSs is exemplary and indications are that economic rewards from up to 30% increase in yields are occurring. Education and capacity building of the cane growers has been helped through an ASS survey demonstration and assessment on each grower's land. Compliance to the agreement is independently audited each year, and to date, there have been very few inadvertent noncompliances found and these were quickly rectified.

The existence and characteristics of ASS across Australia are now provided in the Atlas of Australian Acid Sulfate Soils[14] that is accessible online through the Australian Soil Resource Information System (ASRIS: http://www.aris.gov.au). The management of complex data sets, such as arise in studies of ASS landscapes, can be managed using a system such as that described in Baker.[39] The methods for sampling, laboratory analysis, and assessment of ASS are provided by the ever-evolving Queensland Acid Sulfate Soils Manual.[24] Many organizations across Australia now provide training courses specifically for various aspects about ASS, and these are used for training groups such as stakeholders and regulators (e.g., see various issues of *ASSAT*).[11]

Avoidance

A primary preventative consideration with ASSs should be an avoidance strategy. Such strategies include the decision not to drain ASS wetlands with sulfidic material and to divert or relocate a proposed land use to an alternative site. If use of the site is unavoidable, then treatment so as to prevent sulfide mineral oxidation and export of any existing acidity is necessary.

This seems a relatively easy and sensible first step and one that we advocated as the first principle in an acid sulfate best management consultancy report to the NSW Roads Traffic Authority (RTA), prior to their undertaking major construction of the 1000 km divided highway between Sydney and Brisbane.[40] However, the RTA ASS Guidelines document did not include "avoidance" but most other parts of our report were included verbatim. Nevertheless, we believe that RTA has considered avoidance in assessing the final route of this construction but otherwise has ensured that appropriate treatment has been included in their ASS management plans for the various highway sections that encounter ASS. In assessing the need for avoidance or otherwise, land-use planners will consider relative land values and political advice as well as any engineering issues. This is a complex assessment process.

Oxidation Prevention

This might involve separation of the sulfidic material and its capping with material that is impermeable to oxygen and rainfall infiltration. The use of other sulfidic clays as capping materials must be avoided.[4] Decisions to not drain coastal wetlands are intended not only to avoid the creation of new acidity from sulfide mineral oxidation but also to address the issue of retaining existing acidity.[36] Management techniques such as raising water table elevations are seen as a means of protecting sulfide minerals from oxidation, but this should also consider the hydrological effects with possible increase of existing acidity and the social and economic impact of any subsequently necessary changed land use.

Acidity Containment and Neutralization

Neutralization of ASS acidity by the alkalinity of seawater is the ultimate natural process in the geochemical sulfur cycle. The use of this approach with human-induced acidification and increased acidity discharge needs to be approached with great caution. In eastern Australia, where existing acidity concentrations can exceed 50 tons/ha,[34] application of sufficient crushed limestone or other neutralizing material is impracticable for agriculture. Far better options are land-use practices that minimize acidity export and the strategic application of crushed limestone to drains through which the acidity is exported. These practices can be incorporated into normal activities at the individual paddock scale of intensive agriculture such as with sugarcane. In the NSW sugarcane industry, the adoption of laser grading on estuary floodplains has allowed the marked reduction in the number of field drains that are the main source of acidity export.[36] This has also improved the productivity of the cane land. Low-value land uses, such as with grazing, will be precluded from these new developments, and a more communal approach with some external funding may be necessary. The use of emerging technologies from AMD experience may be useful.[30,41]

Conclusion

The understanding of the need and techniques for best management of ASSs has shifted markedly in the past 20 years. This has been particularly noticeable in Australia and the Baltic coasts and now in inland problems arising in Australia as well. The boundaries between what were previously considered "acid sulfate soil" and "acid rock drainage" or "acid mine drainage" issues have almost disappeared as researchers from both discipline areas have worked together in understanding the chemical and hydrological processes that are associated with their common sulfidic mineral oxidation, and acidity and dissolved metal drainage processes. The issue of ASSs and their problems has involved not only scientific and engineering researchers but also the stakeholders managing land uses and the general public. Across Australia, at least, ASSs and their best management have become incorporated into government policy, regulation, and practice with all three levels of government. It is hoped that adoption by all stakeholders and land managers of a culture of environmental stewardship will greatly reduce the problems arising from ASSs.

References

1. Pons, L.J. Outline of the genesis, characteristics, classification and improvement of acid sulphate soils. In Proceedings of the 1972 International Acid Sulphate Soils Symposium, Wageningen, the Netherlands, Aug. 13–29, 1972; Dost, H., Ed.; International Institute for Land Reclamation and Improvement, ILRI Pub. No. 18, 1973; Vol. 1, 3–27.
2. Dent, D.L. *Acid Sulphate Soils: A Baseline for Research and Development;* International Institute for Land Reclamation and Improvement: Wageningen, the Netherlands, 1986; ILRI Pub. No. 39, 1–204.
3. Dent, D.L.; Pons, L.J. A world perspective on acid sulfate soils. Geoderma **1995**, *67,* 263–276.
4. Fanning, D.S.; Burch, S.N. Coastal acid sulfate soils. In *Reclamation of Drastically Disturbed Lands;* Barnhisel, R.I.; Darmody, R.G.; Daniels, W.L., Eds.; Agronomy 41, American Society of Agronomy: Madison, WI, 2000; 921–937.
5. Fanning, D.S. Acid sulfate soils, definitions, and classification. In *The Encyclopedia of Soil Science*; Lal, R., Ed.; Marcel Dekker: New York, 2002; 11–13.
6. Melville, M.D.; White, I. Acid sulfate soils: Problems and management. In *The Encyclopedia of Soil Science*; Lal, R., Ed.; Marcel Dekker: New York, 2002; 19–22.

7. Fitzpatrick, R.; Shand, P., Eds. *Inland acid sulfate soil systems across Australia,* CRC LEME Open File Report 249, CRC LEME: Perth, Australia, 2008; 1–303. Available online at http://crcleme.org.au/Pubs/OFRSSindex.html (accessed June 2010).
8. Fitzpatrick, R.; Shand, P. Inland acid sulfate soils in Australia; overview and conceptual models (Ch. 1). In *Inland Acid Sulfate Soil Systems across Australia*; Fitzpatrick, R., Shand, P., Eds.; CRC LEME Open File Report 249, CRC LEME: Perth, Australia, 2008; 6–74.
9. White, I.; Melville, M.; Macdonald, B.; Quirk, R.; Hawken, R.; Tunks, M.; Buckley, D.; Beattie, R.; Williams, J.; Heath, L. From conflicts to wise practice agreement and national strategy: Cooperative learning and coastal stewardship in estuarine floodplain management, Tweed River, eastern Australia. J. Cleaner Prod. **2007**, *15,* 1545–1558.
10. Easton, C. The trouble with the Tweed. Fishing World, **1989**, *3,* 58–59.
11. Walsh, Ed. ASSAY; Quarterly publication. Issues 1–52 available at http://www.dpi.nsw.gov.au/aboutus/resources/periodicals/newsletters/assay/. Subscription free from simon.walsh@industry.nsw.gov.au (accessed June 2010) with subject: "subscribe ASSAY".
12. Bush, R.T.; McGrath, R.; Sullivan, L.A. Occurrence of marcasite in an organic-rich Holocene estuarine mud. Aust. J. Soil Res. **2005**, *42,* 617–621.
13. Brinkman, R. Social and economic aspects of reclamation of acid sulphate soil areas. In Proceedings of the Bangkok Symposium on Acid Sulphate Soils, Bangkok, Thailand, Jan. 8–24, 1981; Dost, H., van Breemen, N., Eds.; Institute for Land Reclamation and Improvement: Wageningen, the Netherlands, 1982; ILRI Pub. No. 31, 21–36.
14. Fitzpatrick, R.; Marvanek, S.; Powell, B. Atlas of Australian acid sulfate soils (Ch. 2). In *Inland Acid Sulfate Soil Systems across Australia;* Fitzpatrick, R., Shand, P., Eds.; CRC LEME Open File Report 249, CRC LEME: Perth, Australia, 2008; 75–89.
15. Sammut, J.; Melville, M.D.; Callinan, R.D.; Fraser, G.C. Estuarine acidification: Impacts on aquatic biota of draining acid sulphate soils. Aust. Geogr. Stud. **1995**, 33, 89–100.
16. Degens, B.; Shand, P.; Fitzpatrick, R.; George, R.; Rogers, S.; Smith, M. Avon Basin, WA wheatbelt: Acidification and formation of inland ASS materials in salt lakes by acid drainage and regional groundwater discharge (Ch. 11). In *Inland Acid Sulfate Soil Systems across Australia*; Fitzpatrick, R., Shand, P., Eds.; CRC LEME Open File Report 249, CRC LEME: Perth, Australia, 2008; 176–188.
17. White, I.; Melville, M.D.; Sammut, J.; Wilson, B.P.; Bowman, G.M. Downstream impacts from acid sulfate soils. In *Downstream Impacts of Landuse;* Hunter, H., Eyles, A., Raymont, G., Eds.; Department of Natural Resources: Brisbane, 1996, 165–172.
18. Shand, P.; James-Smith, J.; Hodgkin, T.; Fitzpatrick, R.; McLure, S.; Raven, M.; Stadter, M.; Hill, T. Ancient acid sulfate soils in Murray Basin sediments: Impacts on borehole clogging by $Al(OH)_3$ and salt interception scheme efficiency (Ch. 5). In *Inland Acid Sulfate Soil Systems across Australia;* Fitzpatrick, R., Shand, P., Eds.; CRC LEME Open File Report 249, CRC LEME: Perth, Australia, 2008; 129–136.
19. Dennison, W.C.; O'Neil, J.M.; Duffy, E.; Oliver, P.; Shaw, G. Blooms of the alga cyanobacterium *Lyngbya majuscula* in coastal waters of Queensland. In Proceedings of the International Symposium on Marine Cyanobacteria, Paris, Nov. 24–28, 1997; Charpy, L., Larkum, A.W.D., Eds.; Bulletin de L'Institut Oceanographique, Monaco, 1999; 632.
20. Albert, S.; O'Neil, J.M.; Udy, J.W.; Ahern, K.S.; O'Sullivan, C.M.; Dennison, W.C. Blooms of the cyanobacterium *Lyngbya majuscula* in coastal Queensland, Australia: Disparate site, common factors. Mar. Pollut. Bull. **2005**, *51* (1–4), 428–437.
21. Abram, N.J.; Gagan, M.K.; McCulloch, M.T.; Chappell, J.; Hantoro, W.S. Coral reef death during the 1997 Indian Ocean dipole linked to Indonesian wildfires. Science **2003**, *301,* 952–955.
22. Macdonald, B.C.T.; Denmead, O.T.; White, I.; Melville, M.D. Natural sulfur dioxide emissions from sulfuric soils. Atmos. Environ. **2004**, *38,* 1473–1480.

23. Hicks, W.; Fitzpatrick, R. Greenhouse gas emissions and toxic gas missions from soil organic matter and carbonates associated with acid sulfate soils (Ch. 6). In *Inland Acid Sulfate Soil Systems across Australia*; Fitzpatrick, R., Shand, P., Eds.; CRC LEME Open File Report 249, CRC LEME: Perth, Australia, 2008; 137–148.
24. Ahern, C.R.; McElnea, A.E.; Sullivan, L.A. Acid sulfate soils laboratory guidelines (Version 2.1). In *Queensland Acid Sulfate Soil Technical Manual;* Queensland Department of Natural Resources, Mines and Energy: Indooroopilly, Queensland, Australia, 2004; i-I2–4. (June 2010, current Queensland publication on ASS available at http://www.derm.qld.gov.au/land/ass/).
25. Dear, S.E.; Moore, N.G.; Dobos, S.K.; Wattling, K.M.; Ahern, C.R. Soil management guideline (Version 3.8). In *Queensland Acid Sulfate Soil Technical Manual*; Queensland Department of Natural Resources and Mines: Indooroopilly, Queensland, Australia, 2002; 1–63.
26. Dear, S.E.; Moore, N.G.; Wattling, K.M.; Fahl, D.; Dobos, S.K. Legislation and policy guidelines (Version 2.2). In *Queensland Acid Sulfate Soil Technical Manual*; Queensland Department of Natural Resources and Mines: Indooroopilly, Queensland, Australia, 2004; 1–64.
27. Erskine, W.D.; Warner, R.F. Geomorphic effects of alternating flood and drought dominated regimes on NSW coastal rivers. In *Fluvial Geomorphology of Australia;* Warner, R.F., Ed.; Academic Press: Sydney, 1988; 223–244.
28. Astrom, M.E.; Nystrand, M.; Gustafsson, J.P.; Osterholm, P.; Nordmyr, L.; Reynolds, J.K.; Peltola, P. Lanthanoid behavior in an acidic landscape. Geochim. Cosmochim. Acta **2010**, *74* (3), 829–845.
29. Skwarnecki, M.; Fitzpatrick, R. Geochemical dispersion in acid sulfate soils: Implications for mineral exploration in the Mount Torrens-Strathalbyn area, South Australia (Ch. 4). In *Inland Acid Sulfate Soil Systems across Australia;* Fitzpatrick, R., Shand, P., Eds.; CRC LEME Open File Report 249, CRC LEME: Perth, Australia, 2008; 98–128.
30. Quirk, R.; Melville, M.; Kinsela, A.; Zwemer, T.; Hancock, M.; Macdonald, B.; White, I. Treatment of drainage from acidic canelands using a constructed wetland. Sugar Technol. **2009**, *11* (1), 73–76.
31. Jiggins, J. Interagency learning process. In The OECD Workshop on: An Interdisciplinary Dialogue: Agricultural Production and Integrated Ecosystem Management of Soil and Water, Ballina, NSW, Nov. 10–16, 2002. OECD, University of Western Sydney, NSW Agriculture, 2002, 86–94.
32. Harris, G.P. Comparison of the biogeochemistry of lakes and estuaries: Ecosystem processes, functional groups, hysteresis effects and interactions between macro- and microbiology. Mar. Freshwater Res. **1999**, *50,* 791–811.
33. Thom, B.G.; Harvey, N. Triggers for late 20th century reform of Australian coastal management. Aust. Geogr. Stud. **2000**, *38*, 275–290.
34. Smith, J.; Marston, H.; Melville, M.D.; Macdonald, B.C.T. Spatial distribution and management of actual acidity in an acid sulfate soil environment, McLeods Creek, northeastern NSW, Australia. Catena **2003**, *51,* 61–79.
35. Wilson, B.P.; White, I.; Melville, M.D. Floodplain hydrology, acid discharge and change in water quality associated with a drained acid sulfate soil. Mar. Freshwater Res. **1999**, *50,* 149–157.
36. Kinsela, A.; Melville, M.D. Mechanisms of acid sulfate soil oxidation and leaching under sugarcane cropping. Aust. J. Soil Res. **2004**, *42,* 569–578.
37. White, I.; Melville, M.D.; Wilson, B.P.; Sammut, J. Reducing acidic discharges from coastal wetlands in eastern Australia. Wetlands Ecol. Manage. **1997**, *5*, 55–72.
38. Naylor, S.D.; Chapman, G.A.; Atkinson, G.; Murphy, C.L.; Tulau, M.J.; Flewin, T.C.; Milford, H.B.; Morand, D.T., Eds. *Guidelines for the Use of Acid Sulphate Soil Risk Maps.* NSW Soil Conservation Service, Department of Land and Water Conservation: Sydney, 1995.
39. Baker, A.; Fitzpatrick, R. A new web-based approach for the acquisition, collation and communication of complex inland acid sulfate soil data (Ch. 9). In *Inland Acid Sulfate Soil Systems across Australia*; Fitzpatrick, R., Shand, P., Eds.; CRC LEME Open File Report 249, CRC LEME: Perth, Australia, 2008; 162–168.

40. White, I.; Melville, M.D. Treatment and containment of potential acid sulphate soils: Formation, properties and management of potential acid sulphate soils. In *Report for Roads and Traffic Authority, NSW;* Technical Report T53; CSIRO Centre for Environmental Mechanics: Canberra, 1993.
41. Waite, T.D.; Desmier, R.; Melville, M.D.; Macdonald, B.C.T. Preliminary investigations into the suitability of permeable reactive barriers for the treatment of acid sulfate soil discharge. In *Handbook of Groundwater Remediation of Trace Metals, Radionuclides and Nutrients with Permeable Reactive Barriers*; Academic Press imprint of Elsevier Science: San Diego, 2002; 68–106.

42

Agricultural Water Quantity Management

X. S. Qin and Y. Xu

Introduction

During the past decades, integrated water resource management has been widely recognized as a major challenge to many countries around the world.[1,2] With the rapid economic development and population growth, the water consumption rates have kept rising, leading to serious water shortage problems; meanwhile, the problems are further compounded by a drastic decrease in water quality, resulting from contamination by industrial wastes, domestic sewage water, and agricultural fertilizers and pesticides.[3] Hence, the degradation of water quality is closely linked to the reduction of available water resources. Among various economic activities, the agricultural system has received intensive concerns from government and public, due to the fact that it not only provides necessities for human living but also poses a threat to ambient environmental quality.[4] In a typical agricultural system, the crops' growth requires the nutrients provided by fertilizer and manure. As a consequence, a significant amount of nitrogen and phosphorus may be released from the applied fertilizer and manure, leading to the increase of nutrient levels and deterioration of water quality in the receiving water bodies. In addition, the irrigation activities would consume a huge amount of water, which could exacerbate water shortage problems under disadvantageous conditions.[5] Therefore, an integrated water management strategy becomes important for agricultural development.

Previously, many measures have been proposed and implemented to save water resources or protect water quality for agricultural activities, including promulgation of environmental regulations, reduced application of fertilizers or manures, adoption of soil/water conservation practices, cultivation of low-water-consumption crops, and improvement of farming practices. However, the water shortage and contamination problems have not been significantly ameliorated.[6,7,8,4] One of the major reasons is that management of agricultural water systems is a rather complex issue, involving many interacting components (e.g., type of crops, tillable land areas, and livestock numbers) or influencing factors (e.g., available water resources, environmental loading capacities, and cost implications); single control measures can hardly solve the problem as they are trying to tackle the problem from an unilateral viewpoint instead of a holistic one. On the one hand, the interactive relationships among various system components would lead to difficulties in decision making. For example, the application of the manures and fertilizers provide necessary nutrients to the crops, which are the main source of system benefits; meanwhile, it is related to the applied cost and release of pollutants. On the other hand, an agricultural system involves many social, economic, and environmental factors. For example, a variety of agricultural activities (e.g., crop cultivation and livestock breeding) require enormous infrastructural investment and resources (i.e., land and water); meanwhile, they also pose influences on the socioeconomic activities, which are the main receivers of agricultural products; the socioeconomic activities could also affect agricultural activities through adopting various policies and strategies, such as prescribed pollution-control standards and fertilizer-application practices. Another challenging issue is the existence of uncertainties, which are associated with many system components in an agricultural water management system, such as the economic benefit of crop cultivation, the applied cost of fertilizer and manure, allowable pollutant discharge amounts, and the environmental capacities of receiving water bodies. These uncertainties could be derived from the random nature of the system, errors in measurement, or subjective human judgment, and would directly affect the generation of rational decision schemes.

Over the past decades, many optimization approaches were advanced for handling the above complexities and helping managers for better decision making. Traditionally, to quantify available water resources and identify the relationships among crop yield, crop quality, and irrigation water requirement, the local water authorities should conduct a thorough survey to obtain full meteorological and hydrological information, such as rainfall, surface water inflow and outflow, land areas, crop type and yield, and irrigation frequencies and quotas. Afterward, simulation models could be used to describe the interrelationships among various components within the agricultural system.[9–13] To generate cost-efficient water management strategies, optimization models could be applied.[14–17] Generally, the main objective is to maximize the economic benefit through crop cultivation and livestock breeding; meanwhile, the constraints, such as available water resources and allowable pollutant discharges, should be satisfied.

However, many of the earlier versions of optimization models did not take uncertainty into consideration, leading to oversimplified or inaccurate solutions. In recent years, a large number of inexact optimization techniques were proposed for dealing with uncertainties in agricultural water management fields.[8,4,18,5] A majority of these methods focused on stochastic mathematical programming (SMP), fuzzy mathematical programming (FMP), and interval linear programming (ILP). The selection of uncertainty-analysis techniques depends on the data availability and uncertainty characteristics associated with the system components. Stochastic mathematical programming is used mainly to tackle uncertainties expressed as random variables with probabilistic distribution functions (PDFs). Fuzzy mathematical programming uses fuzzy sets to handle uncertain data showing features of vagueness. The fuzzy membership functions are estimated empirically and can be obtained through human subjective judgments. Interval linear programming can express uncertainties as interval numbers and is effective in situations when little parameter information is available. Conclusively, a variety of optimization models were proposed in the agricultural water management field. In this entry, we aim to systematically introduce a number of the related methodologies, with a focus on demonstration of their applicability through a simplified agricultural water management case. The obtained results will effectively reflect the advantages of various uncertain optimization algorithms. Some implications of the findings and the potential research directions in this field will also be discussed.

General Inexact Optimization Models

In order to realize an integrated agricultural water quantity and quality management, the simulation models should be used to identify the interactive relationships among various system components. The model should be calibrated and verified based on the collection and analysis of long-term historical data. Then, depending on data availability and uncertain features of system components, specific optimization models can be formulated for generating cost- effective management schemes for decision makers. The alternatives of different simulation models or equations can be seen in the work of Zhang et al.,[19] Doorenbos and Kassam[9], Wang[20] Sang,[21] and Lin and Liang.[22] In this entry, the general forms of a number of inexact optimization models will be introduced.

Stochastic Mathematical Programming

Stochastic mathematical programming is used mainly to tackle uncertainties expressed as random variables with PDFs. A series of solutions under different probability levels can be obtained. However, the applicability of SMP is restricted by its rigorous data requirement to specify parameter probability distributions and intensive computational burden.[23] A general SMP model can be written as follows:

$$\text{Minimize} f = \sum_{j=1}^{J} c_j(s) x_j \tag{1a}$$

Subjected to

$$\sum_{j=1}^{J} a_{ij}(s) x_j \le b_i(s)\ \forall i, s \tag{1b}$$

$$\sum_{j=1}^{J} d_{rj} x_j \le e_{r,}\ \ \forall r \tag{1c}$$

$$x_j \ge 0, \forall j \tag{1d}$$

$$c_j(s), a_{ij}(s), d_{rj} \ne 0, \forall i,\ r,\ j,\ s \tag{1e}$$

where x_j are decision variables; $c_j(s)$, $a_{ij}(s)$, and $b_i(s)$ are sets with random elements defined on a probability space s, s ϵ S;[24] and d_{rj} and e_r are fixed coefficients for deterministic constraints. The main step of solving Eq. 1 is to transform the random variables into deterministic ones based on different algorithms. Currently, SMP handles random variables in different ways, such as stochastic chance-constrained programming (SCCP),[8,24–26] two- stage stochastic programming (TSP),[27,28] multistage stochastic programming,[29,30] and stochastic robust opti- mization.[31,26,28,32] The detailed solution procedures can be found in the related references.

Fuzzy Mathematical Programming

Fuzzy mathematical programming is derived through incorporating the fuzzy set theory into the conventional mathematical programming framework. Fuzzy mathematical programming uses fuzzy sets to handle uncertain data, which are characterized by vagueness or imprecision. The fuzzy membership functions are estimated empirically and can be obtained based on subjective judgments. A general FMP can be written as follows:

$$\text{Minimize}\,\tilde{f} = \sum_{j=1}^{J} \tilde{c}_j x_j \tag{2a}$$

Subject to

$$\sum_{j=1}^{J} \tilde{a}_{ij} x_j \le \tilde{b}_i \ \ \forall i \tag{2b}$$

$$\sum_{j=1}^{J} d_{rj} x_j \le e_{r,} \ \ \forall r \tag{2c}$$

$$x_j \ge 0, \forall j \tag{2d}$$

$$\tilde{c}_j, \tilde{a}_{ij}, d_{rj} \ne 0, \forall i,\ r,\ j \tag{2e}$$

where $\tilde{\boldsymbol{f}}$ is the objective function presented in a fuzzy form; *xj* are decision variables; $\tilde{c}_{j,}$ $\tilde{a}_{ij,}$ and $\tilde{\boldsymbol{b}}_i$ are assumed to be different types of fuzzy sets (e.g., triangular, trapezoidal, and exponential forms) with fuzzy membership functions being denoted as $\mu(\tilde{c}_j)$, $\mu(\tilde{a}_{ij})$, and $\mu(\tilde{\boldsymbol{b}}_i)$, respectively; and d_{rj} and e_i are fixed coefficients. To solve Eq. 2, the fuzzy sets are also required to be transformed into deterministic variables. Currently, FMP could be sorted into the following categories based on different modeling structures and solution algorithms: fuzzy flexible programming,[33,34] fuzzy possibilistic programming,[34] fuzzy robust programming (FRP),[23,4] and fuzzy chance-constrained programming (FCCP).[35,3] The detailed solution procedures can be referred to in the related references.

Interval Linear Programming

Interval linear programming is capable of handling the interval-type uncertainties in optimization problems, with all or part of the model parameters being expressed as interval numbers. The model could be solved by the interactive two-step algorithm proposed by Huang et al.[36] A general ILP model can be formulated as follows:

$$\text{Minimize}\,f^{\pm} = \sum_{j=1}^{n} c_j^{\pm} x_j^{\pm} \tag{3a}$$

Subject to

$$\sum_{j=1}^{n} a_{ij}^{\pm} x_j^{\pm} \le b_i^{\pm}\ \forall i \tag{3b}$$

$$x_j^{\pm} \ge 0, \forall j \tag{3c}$$

$$c_j^{\pm}, a_{ij}^{\pm} \ne 0, \forall i,\ j \tag{3d}$$

where $x_j^{\pm}$ are the decision variables presented in interval forms; ± is a boundary symbol for an interval number, representing a spectrum of possible values between the lower and upper bounds; and $a_{ij}^{\pm}$, $b_i^{\pm}$, and $c_j^{\pm}$ are coefficients represented as interval numbers. The term "interval number" is expressed as

$a_{ij}^{\pm} = \left[a_{ij}^{-}, a_{ij}^{\pm}\right]$ where the items a_{ij}^{-}, and a_{ij}^{+} are the lower and upper bounds of a_{ij}^{+} espectively. The objective value and decision variables are obtained as $f_{\text{opt}}^{\pm} = \left[f_{\text{opt}}^{-}, f_{\text{opt}}^{+}\right]$ and $x_{j,\,\text{opt}}^{\pm} = \left[x_{j,\,\text{opt}}^{-}, x_{j,\,\text{opt}}^{+}\right]$, respectively. Previously, many applications of ILP in dealing with environmental management problems were reported.[33,36–39,26,28,32]

Case Study

Overview of the Study Case

To demonstrate the applicability of the introduced inexact optimization methods, a simplified agricultural water quantity and quality management case is adopted.[7,8] The study system is conceptualized as a typical rural area (as shown in Figure 1), consisting of two subareas, where crop cultivation (including wheat, vegetable, and potato) and livestock breeding (including swine and poultry) are the two major economic activities. The water consumption is mainly required by crop irrigation; the pollutants generated from manure/fertilizer applications are discharged into the receiving water bodies. Each subarea has a canal to supply water. The decision makers are responsible for designing optimal farming patterns within a multiperiod horizon, in light of the economic return from the agricultural products, the local land/water resource conditions, pollutant discharge standards, and environmental loading capacities of the receiving water bodies. Figure 2 shows the management framework of the system. For uncertain or deterministic parameters that will be used in different methodologies, they are drawn from literature.[4,5,7,8] In order to better demonstrate methodologies, some parameters are directly assumed based on literature, and the scale of the problem (e.g., the type of agricultural activities and the number of subareas) is considerably simplified.

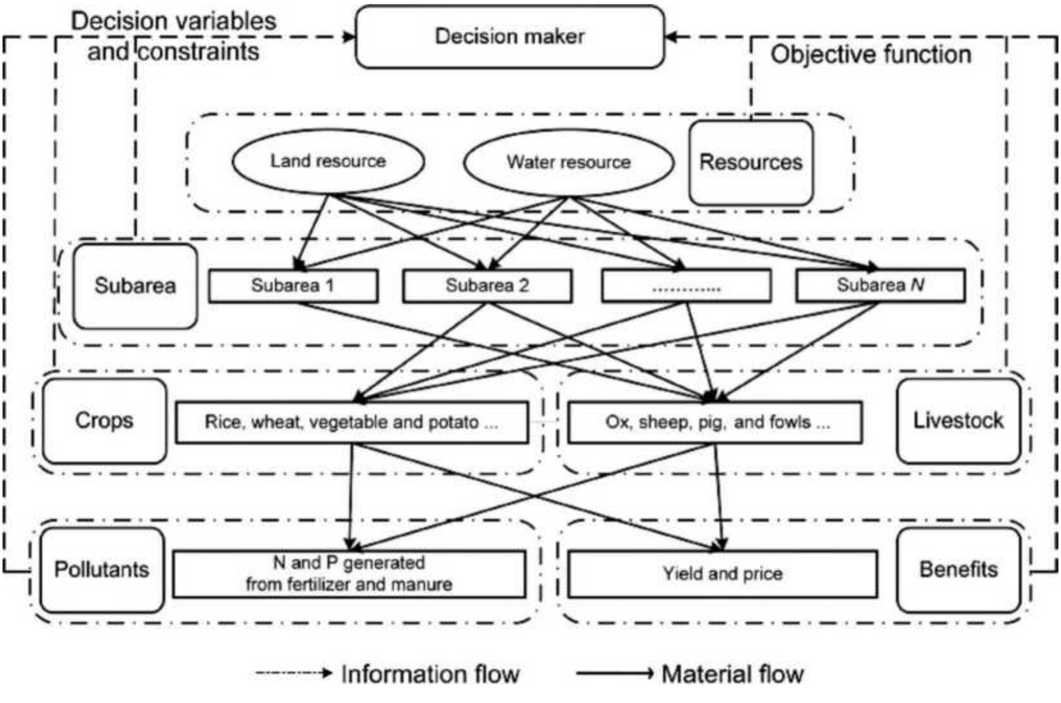


FIGURE 1 Overall structure of the studied system.

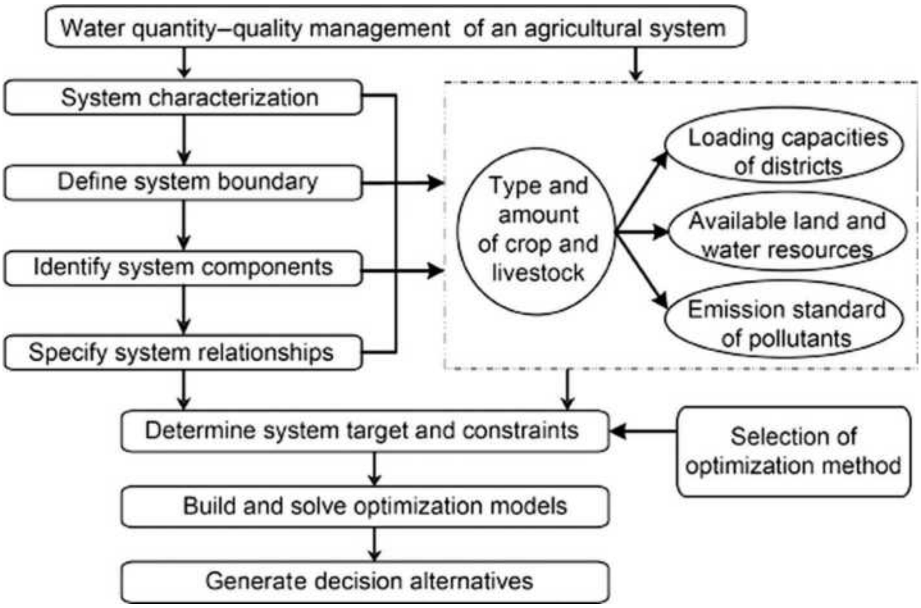


FIGURE 2 Procedures of agricultural water quantity and quality management.

General Model Formulation

A general model for agricultural water quantity–quality management can be formulated as follows:[4,5,7,8]

$$\text{Maximize} f = \sum_{t=1}^{T}\sum_{i=1}^{I}\delta_i Y_i S_{it} + \sum_{j=1}^{J}\eta_j Z_j - \sum_{t=1}^{T}\sum_{i=1}^{I} G_i S_{it}$$

$$-G_M \sum_{i=1}^{I} M_i - G_F \sum_{i=1}^{I} F_i - \sum_{t=1}^{T}\sum_{i=1}^{I} v_{it} W_{it} S_{it} \tag{4a}$$

Subject to

1. Manure mass balance:

$$\sum_{i=1}^{I}\boldsymbol{M}_i - \sum_{j=1}^{J}\boldsymbol{B}_j\boldsymbol{Z}_j = 0 \tag{4b}$$

2. Crop nutrient balances:

$$(1-p_1)\, g_i \boldsymbol{M}_{i+}(1-\boldsymbol{p}_2)\boldsymbol{F}_i - \sum_{t=1}^{\boldsymbol{T}}\boldsymbol{q}_i\boldsymbol{S}_{it} \ge 0, \quad \forall \boldsymbol{i} \tag{4c}$$

3. Energy and digestive protein requirements:

$$\sum_{\boldsymbol{t}=1}^{T}\sum_{\boldsymbol{i}=1}^{I}\boldsymbol{Y}_i\boldsymbol{\beta}_i\boldsymbol{S}_{it} - \sum_{j=1}^{J}\boldsymbol{E}_j\boldsymbol{Z}_j \ge 0 \tag{4d}$$

$$\sum_{t=1}^{T}\sum_{i=1}^{I} Y_i \gamma_i S_{it} - \sum_{j=1}^{J} D_j Z_j \ge 0 \tag{4e}$$

4. Pollutants losses:

$$\sum_{i=1}^{I}\left(M_i g + F_i - q_i \sum_{t=1}^{T} S_{it} \right) \le a \sum_{t=1}^{T} K_t \tag{4f}$$

$$\sum_{i=1}^{I} S_{it} \le K_t \quad \forall t \tag{4g}$$

$$\sum_{t=1}^{T}\sum_{i=1}^{I} L_i S_{it} \le b \sum_{t=1}^{T} K_t \tag{4h}$$

$$\sum_{t=1}^{T}\sum_{i=1}^{I} h_1 L_i S_{it} \le c_1 \sum_{t=1}^{T} K_t \tag{4i}$$

$$\sum_{t=1}^{T}\sum_{i=1}^{I} h_2 L_i S_{it} \le c_2 \sum_{t=1}^{T} K_t \tag{4j}$$

$$\sum_{t=1}^{T}\sum_{i=1}^{I} \left(R_{1i} N_{1i} + R_{2i} N_{2i} \right) S_{it} \le u_1 \sum_{t=1}^{T} K_t \tag{4k}$$

$$\sum_{t=1}^{T}\sum_{i=1}^{I} \left(R_{1i} P_{1i} + R_{2i} P_{2i} \right) S_{it} \le u_2 \sum_{t=1}^{T} K_t \tag{4l}$$

5. Water quantity constraints:

$$\sum_{i=1}^{I} W_{it} S_{it} \le Q_t, \quad \forall t \tag{4m}$$

6. Nonnegativity constraints:

$$S_{it}, M_i, F_i, T_j \ge 0, \ \forall i, j, t \tag{4n}$$

where f=net system income (\$); t, i, and j(t=1, 2,..., T; i=1, 2,..., I;j=1, 2,..., J) are indexes of subarea crops, and livestock, respectively; T, I and J are numbers of subarea, crops, and, livestock, respectively; Y_i=unit yield of crop i (kg/ha); S_{it}=area of crop i in subarea t (ha); Z_j=number of livestock j in the study area; M_i=amount of manure applied to crop i (t); F_i=amount of fertilizer nitrogen applied to crop i (kg);

W_{it} = flow rate of irrigation water required by crop *i* in subarea *t* [(m^3/s)/ha]; *Bj* = unit amount of manure generated by livestock *j* that needs to be disposed (t/unit); p_1 = nitrogen volatilization/denitrification rate of manure (%); *g* = nitrogen concentration of manure (kg/t); p_2 = nitrogen volatilization/denitrification rate of fertilizer (%); q_i = unit nitrogen requirement of crop *i* (kg/ha); β_i = net energy potential of crop *i* (Mcal/kg); *Ej* = unit net energy requirement of livestock *j* (Mcal/unit); γ_i = digestible protein content of crop *i* (%); D_j = unit digestible protein requirement of livestock *j* (kg/unit); K_t = tillable area in subarea *t* (ha); *a* = maximum allowable total nitrogen loss rate (kg/ha); h_1 = nitrogen content of soil (%); h_2 = phosphorus content of soil (%); N_{1i} = dissolved nitrogen concentration in wet season runoff from land planted to crop *i* (mg/l); N_{2i} = dissolved nitrogen concentration in dry season runoff from land planted to crop *i* (mg/l); R_{1i} = wet season runoff from land planted to crop *i* (mm); *R2i* = dry season runoff from land planted to crop *i* (mm); u_1 = maximum allowable loss rate of dissolved nitrogen by runoff (kg/ha); u_2 = maximum allowable loss rate of dissolved phosphorus by runoff (kg/ha); p_{1i} = dissolved phosphorus concentration in wet season runoff from land planted to crop *i* (mg/l); p_2i = dissolved phosphorus concentration in dry season runoff from land planted to crop *i* (mg/l); Q_t = maximum canal flow within subarea *t* (m^3/s); δ_i = unit price of crop *i* ($/kg); *ηj* = unit average return from livestock *j* ($/unit); G_i = unit farming cost for crop *i* ($/ha); G_M = unit cost of manure collection and disposal (dollars/t); G_H = unit cost of fertilizer application ($/kg); v_{it} = unit cost to deliver water to S_{it} {$/(m^3/s)}; L_i = soil loss rate from land planted to crop *i* (kg/ha); *b* = maximum allowable soil loss rate (kg/ha); c_1 = maximum allowable solid-phase nitrogen loss rate (kg/ha); and c_2 = maximum allowable solid-phase phosphorus loss rate (kg/ha).

As discussed in the "Introduction" section, a variety of uncertainties are associated with the agricultural water management system. The optimization model can effectively reflect the relationships between the agricultural and social–economic system and among various system components. The uncertainties can be expressed as random variables, fuzzy sets, or discrete intervals based on the uncertain features of parameters and data availability. In the next section, three inexact optimization models will be used to tackle problem.

Interval Parameter Water Quality Management Model

Huang[7] presented an interval parameter water quality management model (IPWM) for agricultural water quantity and quality planning. The model allowed uncertain information of parameters, expressed as interval numbers, to be effectively communicated into the optimization process and the obtained solutions. Compared with SMP and FMP, the generation of interval numbers has the least data requirement. In a typical agricultural system, many hard-to- get data could be assumed as interval numbers. Table 1 shows part of the interval parameters for the study case. The interval constraints of an IPWM model for the pollutant losses part could be written as follows:

$$\text{Maximize}\, f^{\pm} = \sum_{t=1}^{r}\sum_{i=1}^{m} \delta_i^{\pm} Y_i^{\pm} S_{it}^{\pm} + \sum_{j=1}^{n} \eta_j^{\pm} Z_j^{\pm} - \sum_{t=1}^{r}\sum_{i=1}^{m} G_i^{\pm} S_{it}^{\pm}$$

$$-G_M^{\pm}\sum_{i=1}^{m} M_i^{\pm} - G_F^{\pm}\sum_{i=1}^{m} F_i^{\pm} - G_F^{\pm}\sum_{t=1}^{r}\sum_{i=1}^{m} v_{it}^{\pm} W_{it}^{\pm} S_{it}^{\pm} \tag{5a}$$

Subject to

$$\sum_{i=1}^{m} M_i^{\pm} - \sum_{j=1}^{n} B_j^{\pm} Z_j^{\pm} = 0 \tag{5b}$$

TABLE 1 Part of the Model Parameters Presented as Discrete Intervals

Model Parameters	Values
Nitrogen content of manure (kg/t)	[10, 13]
Unit cost of manure collection/disposal ($/t)	[4, 5]
Unit cost of fertilizer application ($/kg)	[0.7, 0.9]
Nitrogen content of soil (%)	[0.0020, 0.0025]
Phosphorus content of soil (%)	[0.0009, 0.0011]
Nitrogen volatilization/denitrification rate of manure (%)	[0.3, 0.35]
Phosphorus volatilization/denitrification rate of manure (%)	[0.10, 0.12]
Maximum allowable dissolved nitrogen loss rate by runoff (kg/ha)	[2.2, 2.3]
Maximum allowable dissolved phosphorus loss rate by runoff (kg/ha)	[0.20, 0.22]
Maximum allowable nitrogen losses (kg/ha)	[38, 40]
Maximum allowable soil loss (kg/ha)	[3650, 5600]
Maximum allowable soil-phase nitrogen loss rate (kg/ha)	[10, 11]

Source: Adapted from Huang.[7]

$$\left(1-P_1^{\pm}\right)g_i^{\pm}M_i^{\pm}+\left(1-P_2^{\pm}\right)F_1^{\pm}-\sum_{t=1}^{T}q_i^{\pm}S_{it}^{\pm}\geq 0,\ \forall i \tag{5c}$$

$$\sum_{t=1}^{T}\sum_{i=1}^{I}L_i^{\pm}S_{it}^{\pm}\leq b^{\pm}\sum_{t=1}^{T}K_t \tag{5d}$$

$$\sum_{t=1}^{T}\sum_{i=1}^{I}h_1^{\pm}L_i^{\pm}S_{it}^{\pm}\leq c_1^{\pm}\sum_{t=1}^{T}K_t \tag{5e}$$

$$\sum_{t=1}^{T}\sum_{i=1}^{I}h_2^{\pm}L_i^{\pm}S_{it}^{\pm}\leq c_2^{\pm}\sum_{t=1}^{T}K_t \tag{5f}$$

$$\sum_{t=1}^{T}\sum_{i=1}^{I}\left(R_{1i}^{\pm}N_{1i}^{\pm}+R_{2i}^{\pm}N_{2i}^{\pm}\right)S_{it}^{\pm}\leq u_1^{\pm}\sum_{t=1}^{T}K_t \tag{5g}$$

$$\sum_{t=1}^{T}\sum_{i=1}^{I}\left(R_{1i}^{\pm}P_{1i}^{\pm}+R_{2i}^{\pm}P_{2i}^{\pm}\right)S_{it}^{\pm}\leq u_2^{\pm}\sum_{t=1}^{T}K_t \tag{5h}$$

$$\sum_{i=1}^{I}W_{it}^{\pm}S_{it}^{\pm}\leq Q_t^{\pm},\ \forall t \tag{5i}$$

As proposed by Huang et al.[36,38], the solution for Eq. 5 can be obtained through an interactive two-step method. A submodel corresponding to f^+ (when the objective function is to be maximized) is first formulated and solved, and then the relevant submodel corresponding to f can be formulated based on the solution of the first submodel. Finally, the objective value and decision variables as discrete intervals can be obtained as $f_{\text{opt}}^{\pm} = \left[f_{\text{opt}}^{-}, f_{\text{opt}}^{+} \right]$ and $x_{j,\,\text{opt}}^{\pm} = \left[x_{j,\,\text{opt}}^{-}, x_{j,\,\text{opt}}^{+} \right]$ The interactive method does not lead to complicated intermediate models and thus has low computational requirements.

Table 2 shows the solutions from IPWM for the study case. Some results are represented as interval numbers. For example, the objective function value (i.e., net income) would range from \$121.38 × 10^3 to \$310.99 × 10^3. The upper bound of the objective function represents an optimal decision alternative with the highest net income. Correspondingly, the obtained decision variables of cropping area and livestock amount would reach their upper bounds, and the manure and fertilizer would achieve their lower- bound values. For example, the cropping area of the vegetables in subarea 1 is 39.60 km^2; the number of swine is 324 units; the applied manure and fertilizer amounts of the vegetables are 606.12 × 10^6 t and 4004.15 kg, respectively. These upper-bound solutions are obtained under an advantageous condition. For example, the discharge standards of pollutants are relatively loose, and the loading capacities of the receiving water bodies and the available water amounts are relatively larger. This also means that higher system benefits correspond with higher levels of pollutant discharge and/or more serious water shortage problems. Conversely, the lower-bound solution of net income is under conservative consideration, where the conditions are stricter (e.g., the allowable pollutant amounts and provided water amounts are lower). In such a case, the water quality could be maintained at a higher level, and the water shortage crisis could be alleviated somewhat. Based on the obtained interval solutions, a variety of decision alternatives can be generated through adjusting continuously within their solution intervals. In fact, decision variables presented as discrete intervals are rather flexible for decision makers, since they are capable of assisting decision makers in making the decision based on their preferences (e.g., economic development or environmental protection).

Although the ILP method has many advantages, it also has limits. Firstly, the representation of uncertainties by discrete intervals is relatively simple, where much useful information is neglected. Secondly, although ILP has been proven as an effective approach in dealing with uncertainties, it may lead to the difficulties when the model's right- hand-side coefficients are highly uncertain. Therefore, further investigations are needed to improve the applicability of ILP. In fact, many other types of uncertainty methods, such as SCCP, TSP, and FRP, have potential to be further integrated into an ILP framework for mitigating the drawbacks of ILP and reflecting more complex conditions.

TABLE 2 Interval Solutions of IPWM Model

Model Solutions	Values
Cropping area of wheat in subarea 1 (km2)	0
Cropping area of wheat in subarea 2 (km2)	0
Cropping area of vegetable in subarea 1 (km2)	[22.38, 39.60]
Cropping area of vegetable in subarea 2 (km2)	[10.96, 19.23]
Cropping area of potato in subarea 1 (km2)	0
Cropping area of potato in subarea 2 (km2)	0
Applied manure amounts of wheat (106 t)	[0, 8.19]
Applied manure amounts of vegetable (106 t)	606.12
Applied manure amounts of potato (106 t)	0
Applied fertilizer amounts of wheat (kg)	0
Applied fertilizer amounts of vegetable (kg)	4004.15
Applied fertilizer amounts of potato (kg)	0
Swine amounts (unit)	[311, 324]
Poultry amounts (unit)	[379,557]
Net income (103 \$)	[121.38, 310.99]

Inexact Stochastic Water Management Model

Stochastic chance-constrained programming is one of the SMP methods that can tackle uncertain parameters presented as PDFs. Its main advantage is that it does not require that all of the constraints be strictly satisfied. Instead, they can be satisfied in a proportion of cases with given probabilities.[40] However, SCCP is restricted by data availability to generate stochastic distributions.[37,38] Interval linear programming is widely applicable for addressing all uncertain parameters; however, it does not allow any violation of system constraints and might be infeasible when the right-hand-side parameters in constraints are highly uncertain. The two methods have varied strengths and weaknesses, with a potential for compensating each other when they are combined.

Huang[8] proposed an interval-stochastic chance- constrained programming (ISCCP) model for agricultural water quantity and quality management. The ISCCP model is an integration of ILP and SCCP but improves upon both. It allows the right-hand-side parameters in the constraints to be expressed as PDFs and the coefficients in the objective function and the left-hand-side parameters in the constraints to be described as interval numbers. For the introduced study case, the available water resource amount, the maximum allowable soil loss, the maximum allowable soil-phase nitrogen loss, and the maximum allowable soil-phase phosphorus loss are assumed as random variables (expressed as PDFs). Table 3 lists the random parameters used in the ISCCP model. Other uncertain parameters are tackled as discrete intervals, similar to those in the IPWM model.

Thus, the interval-stochastic part of an ISCCP model can be written as follows:

$$\text{Maximize} f^{\pm} = \sum_{t=1}^{r}\sum_{i=1}^{m}\delta_i^{\pm}Y_i^{\pm}S_{it}^{\pm} + \sum_{j=1}^{n}\eta_j^{\pm}Z_j^{\pm} - \sum_{t=1}^{r}\sum_{i=1}^{m}G_i^{\pm}S_{it}^{\pm}$$

$$G_M^{\pm}\sum_{i=1}^{m}M_i^{\pm} - G_F^{\pm}\sum_{i=1}^{m}F_i^{\pm} - \sum_{m}^{r}\sum_{i=1}^{m}V_{it}^{\pm}W_{it}^{\pm}S_{it}^{\pm} \tag{6a}$$

Subject to

$$\sum_{i=1}^{m}M_i^{\pm} - \sum_{j=1}^{n}B_j^{\pm}Z_j^{\pm} = 0 \tag{6b}$$

$$\sum_{t=1}^{T}\sum_{i=1}^{I}L_i^{\pm}S_{it}^{\pm} \le b(s)\sum_{t=1}^{T}K_t \tag{6c}$$

TABLE 3 Random Parameters in ISCCP Model

Model Parameters	Values
Supplied water amounts of canal 1 (m3/ha-s)	(3.6, 0.12)
Supplied water amounts of canal 2 (m3/ha-s)	(0.9, 0.04)
Maximum allowable soil loss (kg/ha)	(4600, 150)
Maximum allowable soil-phase nitrogen loss (kg/ha)	(10.5, 0.5)
Maximum allowable soil-phase phosphorus (kg/ha)	(4.5, 0.2)

Note: (*a1, a2*) represents a random variable where a1and a2 are mean and standard deviation, respectively.

$$\sum_{t=1}^{T}\sum_{i=1}^{I} h_1^{\pm} L_i^{\pm} S_{it}^{\pm} \le c_1(s)\sum_{t=1}^{T} K_t \tag{6d}$$

$$\sum_{t=1}^{T}\sum_{i=1}^{I} h_2^{\pm} L_i^{\pm} S_{it}^{\pm} \le c_2(s)\sum_{t=1}^{T} K_t \tag{6e}$$

$$\sum_{i=1}^{I} W_{it}^{\pm} S_{it}^{\pm} \le Q_t(S),\ \forall t \tag{6f}$$

In the solution process of Eq. 6, the constraints in Eq. 6c–f are required to be satisfied with at least a probability of *1–ph* According to Charnes et al.,[24] these constraints can be written as[24]

$$\boldsymbol{P}_r\left[\sum_{t=1}^{\boldsymbol{T}}\sum_{i=1}^{\boldsymbol{I}} \boldsymbol{L}_i^{\pm} \boldsymbol{S}_{it}^{\pm} \le \boldsymbol{b}(\boldsymbol{s})\sum_{t=1}^{\boldsymbol{T}} \boldsymbol{K}_t\right] \ge 1-\boldsymbol{p}_{h},\ \forall \boldsymbol{t}, \boldsymbol{h} \tag{7a}$$

$$\boldsymbol{P}_r\left[\sum_{t=1}^{\boldsymbol{T}}\sum_{i=1}^{\boldsymbol{I}} \boldsymbol{h}_1^{\pm} \boldsymbol{L}_i^{\pm} \boldsymbol{S}_{it}^{\pm} \le \boldsymbol{c}_1(\boldsymbol{s})\sum_{t=1}^{\boldsymbol{T}} \boldsymbol{K}_t\right] \ge 1-\boldsymbol{p}_{h},\ \forall \boldsymbol{t}, \boldsymbol{h} \tag{7b}$$

$$\boldsymbol{P}_r\left[\sum_{t=1}^{\boldsymbol{T}}\sum_{i=1}^{\boldsymbol{I}} \boldsymbol{h}_2^{\pm} \boldsymbol{L}_i^{\pm} \boldsymbol{S}_{it}^{\pm} \le \boldsymbol{c}_2(\boldsymbol{s})\sum_{t=1}^{\boldsymbol{T}} \boldsymbol{K}_t\right] \ge 1-\boldsymbol{p}_{h},\ \forall \boldsymbol{t}, \boldsymbol{h} \tag{7c}$$

$$\boldsymbol{p}_r\left[\sum_{i=1}^{I} W_{it}^{\pm} \boldsymbol{S}_{it}^{\pm} \le \boldsymbol{Q}_t(\boldsymbol{S})\right] \ge 1-\boldsymbol{p}_{h},\ \forall \boldsymbol{t}, \boldsymbol{h} \tag{7d}$$

where $\boldsymbol{p}_h$ means the allowable probability of violation given by the decision makers. In this entry, the values of $\boldsymbol{p}_h$ are considered to be 0.01 and 0.05, respectively. Referring to Charnes et al.,[24] the constraints in Eq. 7a–7d can be transformed to their equivalent forms as follows:

$$\sum_{t=1}^{T}\sum_{i=1}^{I} L_i^{\pm} S_{it}^{\pm} \le b(s)^{Ph}\sum_{t=1}^{T} K_t \tag{8a}$$

$$\sum_{t=1}^{T}\sum_{i=1}^{I} h_1^{\pm} L_i^{\pm} S_{it}^{\pm} \le c_1(s)^{Ph}\sum_{t=1}^{T} K_t \tag{8b}$$

$$\sum_{t=1}^{T}\sum_{i=1}^{I} h_2^{\pm} L_i^{\pm} S_{it}^{\pm} \le c_2(s)^{Ph}\sum_{t=1}^{T} K_t \tag{8c}$$

$$\sum_{i=1}^{I} W_{it}^{\pm} S_{it}^{\pm} \le Q_t(S)^{Ph},\ \forall t, h \tag{8d}$$

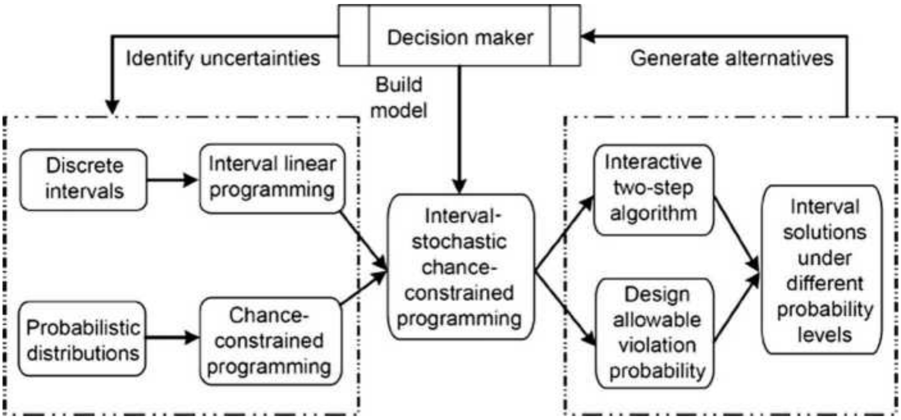


FIGURE 3 The framework for applying the ISCCP model.

where $a(s)^{ph} = Fh^{-1}[a(s)]$, given the cumulative distribution function of $a(s)$ and the probability of violating constraint h (i.e., p_h). Finally, Eq. 8 can be transformed into pure ILP models under various constraint-violation levels and then solved by an interactive two-step method.[36] Figure 3 shows the solution procedures of an ISCCP model.

Table 4 lists the solutions from the ISCCP model under probability levels of 0.01 and 0.05. Since ISCCP is an integration of ILP and SCCP, the obtained solutions reflect characteristics of both methods. Firstly, due to the existence of ILP, a majority of solutions present as interval numbers, which are feasible and stable in the given decision space; thus, the decision schemes can be generated in the ranges of their solution intervals. For example, at a significance level of $p_h = 0.01$, the total system benefits are [124.38, 256.03] $\times 10^3$. The upper bounds of objective function are preferable under advantageous conditions; conversely, the lower bounds of objective function value are more desirable under conservative conditions. The local managers can make decisions based on their preferences.

The solutions of ISCCP also possess features of SCCP where the trade-off between cost and risk can be evaluated (reflected by adjusting *ph*). For example, when *ph* increases from 0.01 to 0.05, the total system benefit would increase from [124.38, 256.03] to [127.40, 262.25] $\times 10^3$. This is due to the fact that, as *ph* increases, the allowable degree of violating constraints would increase (e.g., looser pollutant discharge standards and environmental loading capacities). This indicates that a higher system benefit could lead to a higher system risk. Conversely, if the planner aims toward a lower system benefit, a more reliable management alternative would be generated. The p_h values can be used to help analyze the trade-off between system economy and failure risk and offer local water managers a spectrum of alternatives for decision making.

ISCCP also shows some limitations. Firstly, the conversion from a stochastic constraint into an "equivalent" deterministic one applies only in conditions when the right-hand-side items are stochastic. The interval treatment for the left-hand-side uncertainties is a compromised approach, which may lead to the loss of valuable information if additional distribution information of the interval parameters is available.

Inexact Double-Sided Fuzzy Chance-Constrained Programming

The fuzzy approach is another alternative to deal with uncertainties. Referring to Xu and Qin,[5] an inexact double-sided fuzzy chance-constrained programming (IDFCCP) model was used to solve the agricultural water quantity and quality management problem. The IDFCCP model is formulated through

TABLE 4 Solutions of ISCCP Model at Probability Levels of 0.01 and 0.05

Model Solutions	Probability (pi)	Values
Cropping area of wheat in subarea 1 (km2)	0.01	0
	0.05	0
Cropping area of wheat in subarea 2 (km2)	0.01	0
	0.05	0
Cropping area of vegetable in subarea 1 (km2)	0.01	[23.22, 32.87
	0.05	[23.78, 33.66]
Cropping area of vegetable in subarea 2 (km2)	0.01	[11.10, 15.58]
	0.05	[11.37, 15.96]
Cropping area of potato in subarea 1 (km2)	0.01	0
	0.05	0
Cropping area of potato in subarea 2 (km2)	0.01	0
	0.05	0
Applied manure amounts of wheat (106 t)	0.01	0
	0.05	0
Applied manure amounts of vegetable (106 t)	0.01	499.11
	0.05	511.24
Applied manure amounts of potato (106 t)	0.01	0
	0.05	0
Applied fertilizer amounts of wheat (kg)	0.01	0
	0.05	0
Applied fertilizer amounts of vegetable (kg)	0.01	3297.27
	0.05	3377.35
Applied fertilizer amounts of potato (kg)	0.01	0
	0.05	0
Swine amounts (unit)	0.01	267
	0.05	273
Poultry amounts (unit)	0.01	458
	0.05	470
Net income (103 $)	0.01	[124.38, 256.03
	0.05	[127.40, 262.25]

Notes: [a,b] represents an interval number, where a and b are the lower and upper bounds, respectively.

incorporating ILP into a double-sided fuzzy chance-constrained programming (DFCCP) framework and could be used to deal with uncertainties expressed as fuzzy sets and interval parameters. Recently, FCCP has been presented as a novel FMP method through incorporating predefined confidence levels of fuzzy constraint satisfaction into optimization models. Similar to SCCP models, FCCP requires that the fuzzy chance constraints be transformed to deterministic equivalents at predetermined confidence levels. Fuzzy chance-constrained programming is sorted into single-sided FCCP and DFCCP. In many real-world problems, it is more common that both sides of model constraints be associated with uncertainties. Interval linear programming is capable of handling the uncertainties associated with other parameters in the optimization model if they do not have either fuzzy or stochastic distribution information.

For the introduced study case, the environmental loading capacities (such as allowable soil loss and pollutant discharge amounts) are subject to human judgment and described by fuzzy sets. Table 5 lists

TABLE 5 Fuzzy Parameters in IDFCCP Model

Model Parameters	Values
Supplied water amounts of canal 1 (m3/ha-s)	(0.8, 0.88, 1)
Supplied water amounts of canal 2 (m3/ha-s)	(3.2, 3.5, 4.0)
Maximum allowable soil loss (kg/ha)	(3650, 4600, 5600)
Maximum allowable soil-phase nitrogen loss (kg/ha)	(10, 10.4, 11)
Maximum allowable soil-phase phosphorus (kg/ha)	(4.3, 4.5, 4.7)

Notes: (*a, b, c*) represents a triangular fuzzy set, where *a* and *c* are the minimum and maximum possible values, and b is the most likely value

the related fuzzy parameters. Other uncertain parameters are treated as discrete intervals. The specific constraints related to the IDFCCP model can be formulated as follows:[5]

$$Pos\left\{\tilde{L}_i,\ \tilde{b}\,\middle|\,\sum_{t=1}^{r}\sum_{i=1}^{m}\tilde{L}_i S_{it}^{\pm} \le \tilde{b}\sum_{t=1}^{r}K_t\right\} \ge \alpha \tag{9a}$$

Subject to

$$Pos\left\{\tilde{L}_i,\ \tilde{b}\,\middle|\,\sum_{t=1}^{r}\sum_{i=1}^{m}\tilde{L}_i S_{it}^{\pm} \le \tilde{b}\sum_{t=1}^{r}K_t\right\} \ge \alpha \tag{9b}$$

$$Pos\left\{\tilde{L}_i,\ \tilde{c}_1\,\middle|\,\sum_{t=1}^{r}\sum_{i=1}^{m}h_1^{\pm}\tilde{L}_i S_{it}^{\pm} \le \tilde{c}_1\sum_{t=1}^{r}K_t\right\} \ge \alpha \tag{9c}$$

$$Pos\left\{\tilde{L}_i,\ \tilde{c}_2\,\middle|\,\sum_{t=1}^{r}\sum_{i=1}^{m}h_2^{\pm}\tilde{L}_i S_{it}^{\pm} \le \tilde{c}_2\sum_{t=1}^{r}K_t\right\} \ge \alpha \tag{9d}$$

$$Pos\left\{\tilde{Q}_t\,\middle|\,\sum_{i=1}^{m}W_{it}^{\pm} S_{it}^{\pm} \le \tilde{Q}_t\right\} \ge \alpha,\ \forall t \tag{9e}$$

where *Pos*{·}denotespossibility of events in {·}; and α is a predetermined confidence level, where the values of α are 0.3, 0.6, and 0.9, respectively. According to predetermined confidence levels, a fuzzy chance constraint can be converted to two crisp equivalents as follows:[41]

1. Confidence levels under the minimum reliability.

$$\sum_{t=1}^{r}\sum_{i=1}^{m}L_i^{L}(\alpha)\,S_{it} \le b^{R}(\alpha)\sum_{t=1}^{r}K_t \tag{10a}$$

$$\sum_{t=1}^{r}\sum_{i=1}^{m}h_1^{\pm}L_i^{L}(\alpha)\,s_{it}^{\pm} \le c_1^{R}(\alpha)\sum_{t=1}^{r}K_t \tag{10b}$$

$$\sum_{t=1}^{r}\sum_{i=1}^{m} h_2^{\pm} L_i^L(\alpha) s_{it}^{\pm} \le c_2^R(\alpha) \sum_{t=1}^{r} K_t \tag{10c}$$

$$\sum_{i=1}^{I} W_{it}^{\pm} S_{it}^{\pm} \le Q_t^R(\alpha), \ \forall t \tag{10d}$$

2. Confidence levels under the maximum reliability.

$$\sum_{t=1}^{r}\sum_{i=1}^{m} L_i^R(1-\alpha) S_{it}^{\pm} \le b^L(1-\alpha) \sum_{t=1}^{r} K_t \tag{11a}$$

$$\sum_{t=1}^{r}\sum_{i=1}^{m} h_1^{\pm} L_i^R(1-\alpha) S_{it}^{\pm} \le c_1^L(1-\alpha) \sum_{t=1}^{r} K_t \tag{11b}$$

$$\sum_{t=1}^{r}\sum_{i=1}^{m} h_2^{\pm} L_i^R(1-\alpha) S_{it}^{\pm} \le c_2^L(1-\alpha) \sum_{t=1}^{r} K_t \tag{11c}$$

$$\sum_{i=1}^{I} W_{it}^{\pm} S_{it}^{\pm} \le Q_t^L(1-\alpha), \ \forall t \tag{11d}$$

The solutions at different confidence levels for IDFCCP can be then obtained through the two-step method.[36] Finally, two groups of objective value and decision variables (presented as interval numbers) corresponding to two scenarios of reliabilities can be obtained. Tables 6–8 list the solutions from IDFCCP at confidence levels of 0.3, 0.6, and 0.9 with two reliabilities. It can be found that the objective function value and part of the decision variables would present as discrete intervals. The solutions also have considerable variations under different combinations of α-cut levels and reliability scenarios. Generally, the system income would decrease as α-cut level increases; meanwhile, the system income under the minimum reliability would be lower than that under the maximum one. For example, at confidence levels of 0.3, 0.6, and 0.9 under the minimum reliability, the system incomes are [145.53, 299.48], [139.94, 287.98], and [134.36, 276.47] ($\times 10^3$), respectively. Under the maximum reliability, the system incomes are [129.01, 265.48], [125.53, 258.33], and [122.06, 251.17] ($\times 10^3$), respectively. This is due to the fact that, when α-cut level increases, the confidence levels of constraint satisfaction would increase, and the constraints of pollutant discharge standards, loading capacities, and available water resources would become stricter. This indicates that a high economic benefit could lead to a high system failure risk. Conversely, a conservative decision alternative is more desirable for mitigating water quality and quantity crisis. A trade-off between system benefit and reliability of satisfying model constraints needs to be analyzed in order to gain an in-depth insight into the characteristics of the management system.

One of the limitations associated with IDFCCP is that the determination of the confidence levels is rather arbitrary. There are currently no effective tools or guidelines to help choose proper confidence levels for satisfying constraints. Secondly, IDFCCP generates many sets of solutions under different confidence levels and reliability scenarios, which may complicate the decision-making process. Multicriteria decision analysis tools may be a potential tool to mitigate such a problem.

TABLE 6 Solutions of ISCCP Model under α = 0.3

Model Solutions	Reliability	Values
Cropping Area of Wheat in Subarea 1 (km2)	Minimum	0
	Maximum	0
Cropping area of wheat in subarea 2 (km2)	Minimum	0
	Maximum	0
Cropping area of vegetable in subarea 1 (km2)	Minimum	[26.92, 38.12]
	Maximum	[23.85, 33.76]
Cropping area of vegetable in subarea 2 (km2)	Minimum	[13.21, 18.54]
	Maximum	[11.73, 16.46]
Cropping area of potato in subarea 1 (km2)	Minimum	0
	Maximum	0
Cropping area of potato in subarea 2 (km2)	Minimum	0
	Maximum	0
Applied manure amounts of wheat (106 t)	Minimum	0
	Maximum	0
Applied manure amounts of vegetable (106 t)	Minimum	583.68
	Maximum	517.41
Applied manure amounts of potato (106 t)	Minimum	0
	Maximum	0
Applied fertilizer amounts of wheat (kg)	Minimum	0
	Maximum	0
Applied fertilizer amounts of vegetable (kg)	Minimum	3855.96
	Maximum	3418.12
Applied fertilizer amounts of potato (kg)	Minimum	0
	Maximum	0
Swine amounts (unit)	Minimum	312
	Maximum	276
Poultry amounts (unit)	Minimum	536
	Maximum	475
Net income (103 $)	Minimum	[145.53, 299.48]
	Maximum	[129.01, 265.48]

Further Discussions

The methodologies introduced in this study were based on integration of SMP, FMP, and ILP techniques. The major reason is that different types of uncertainty-analysis tools have varied strengths and weaknesses, in terms of the sensitivity to data quality, accuracy of information interpretation, and requirement of computational efforts. Integration of different algorithms could better deal with dual or multiple uncertainties. In agricultural water management, many types of uncertainties may exist. For example, the available water amount is affected by hydrological and meteorological conditions and may show random features; it can normally be expressed as a stochastic variable; the parameters related to allowable pollutant discharge amounts and environmental loading capacities are subject to human judgments and could be expressed by fuzzy membership functions; parameters related to the benefit and the crop yield may fluctuate in a range where their distribution information may be difficult to obtain due to a lack of sufficient historical data, and therefore, they are more suitable to be described by discrete intervals. Depending on the quality of data, how to choose proper uncertainty-analysis methods becomes a critical challenge for decision makers.

TABLE 7 Solutions of ISCCP Model under $\alpha = 0.6$

Model Solutions	Reliability	Values
Cropping area of wheat in subarea 1 (km2)	Minimum	0
	Maximum	0
Cropping area of wheat in subarea 2 (km2)	Minimum	0
	Maximum	0
Cropping area of vegetable in subarea 1 (km2)	Minimum	[25.87, 36.63]
	Maximum	[23.22, 32.87]
Cropping area of vegetable in subarea 2 (km2)	Minimum	[12.71, 17.85]
	Maximum	[11.40, 16.00]
Cropping area of potato in subarea 1 (km2)	Minimum	0
	Maximum	0
Cropping area of potato in subarea 2 (km2)	Minimum	0
	Maximum	0
Applied manure amounts of wheat (106 t)	Minimum	0
	Maximum	0
Applied manure amounts of vegetable (106 t)	Minimum	561.25
	Maximum	503.47
Applied manure amounts of potato (106 t)	Minimum	0
	Maximum	0
Applied fertilizer amounts of wheat (kg)	Minimum	0
	Maximum	0
Applied fertilizer amounts of vegetable (kg)	Minimum	3707.76
	Maximum	3326.06
Applied fertilizer amounts of potato (kg)	Minimum	0
	Maximum	0
Swine amounts (unit)	Minimum	300
	Maximum	269
Poultry amounts (unit)	Minimum	516
	Maximum	462
Net income (103 $)	Minimum	[139.94, 287.98]
	Maximum	[125.53, 258.33]

The study case used in this entry is only for demonstration purpose, and it is therefore highly simplified. The water quantity and quality levels are conceptualized into simple mass-balance expressions (i.e., the left-hand sides of model constraints). In real-world cases, the related processes are not necessity linear, and rigid simulation models are required to quantify the related relationships. For example, the prediction of nonpoint sources of pollution on a watershed scale requires a sophisticated hydrological model. Thus, a coupling of simulation and optimization models is desired to tackle more complicated problems. However, how to effectively realize such integration is a challenging topic and deserves further investigations.[42] Another issue is about the acquisition of data. Many of the parameters used in this study were obtained through subjective assumption or referred to in the literature. In actual applications, the data quality significantly determines the results of models and thus influences the decision-making processes. Although uncertainty-analysis tools are available, they cannot be considered as alternatives to good-quality data. Efforts are still needed to acquire as much accurate information as possible in order to ensure more reliable modeling results and robust management strategies.

TABLE 8 Solutions of ISCCP Model under α = 0.9

Model Solutions	Reliability	Values
Cropping area of wheat in subarea 1 (km2)	Minimum	0
	Maximum	0
Cropping area of wheat in subarea 2 (km2)	Minimum	0
	Maximum	0
Cropping area of vegetable in subarea 1 (km2)	Minimum	[24.83, 35.15]
	Maximum	[22.59, 31.98]
Cropping area of vegetable in subarea 2 (km2)	Minimum	[12.22, 17.15]
	Maximum	[11.07, 15.54]
Cropping area of potato in subarea 1 (km2)	Minimum	0
	Maximum	0
Cropping area of potato in subarea 2 (km2)	Minimum	0
	Maximum	0
Applied manure amounts of wheat (106 t)	Minimum	0
	Maximum	0
Applied manure amounts of vegetable (106 t)	Minimum	538.82
	Maximum	489.54
Applied manure amounts of potato (106 t)	Minimum	0
	Maximum	0
Applied fertilizer amounts of wheat (kg)	Minimum	0
	Maximum	0
Applied fertilizer amounts of vegetable (kg)	Minimum	3559.57
	Maximum	3234
Applied fertilizer amounts of potato (kg)	Minimum	0
	Maximum	0
Swine amounts (unit)	Minimum	288
	Maximum	261
Poultry amounts (unit)	Minimum	495
	Maximum	450
Net income (103 $)	Minimum	[134.36, 276.47]
	Maximum	[122.06, 251.17]

Conclusion

This entry gave a general introduction of how to conceptualize and formulate inexact optimization models for supporting integrated agricultural water quantity and quality management under uncertainty. A number of inexact optimization approaches based on stochastic, fuzzy, and interval programming models were introduced. Each of them is capable of handling a certain type of uncertain information, which is dependent on the uncertain features of the system components. A typical agricultural water management problem was used to demonstrate the applicability of different methods. The study results effectively showed the characteristics of various optimization approaches. Limitations associated with different modeling algorithms were also discussed.

References

1. Johnson, C.; Handmer, J. Water supply in England and Wales: Whose responsibility is it when things go wrong? Water Policy **2002**, *4* (4), 345–366.
2. Al-Salihi, A.H.; Himmo, S.K. Control and management study of Jordan’s water resource. Water Int. **2003**, *28*(1), 1–10.

3. Xu, Z.X.; Li, J.Y. Sustainable water resource management in China: A great challenge. Alliance for Global Sustainability Bookseries **2010**, *18* (1), 3–10.
4. Nie, X.H.; Huang, G.H.; Wang, D.; Li, H.L. Robust optimisation for inexact water quality management under uncertainty. Civ. Eng. Environ. Syst. **2008**, *25(2),* 167–184.
5. Xu, Y.; Qin, X.S. Rural effluent control under uncertainty: An inexact double-sided fuzzy chance-constrained model. Adv. Water Resour. **2010**, *33* (9), 997–1014.
6. Haith, D.A. *Environmental Systems Optimization;* John Wiley and Sons, Inc.: New York, 1982.
7. Huang, G.H. IPWM: An interval parameter water quality management model. Eng. Optim. **1996**, *26,* 79–103.
8. Huang, G.H. A hybrid inexact-stochastic water management model. Eur. J. Oper. Res. **1998**, *107,* 137–158.
9. Doorenbos, J.; Kassam, A.H. *Yield Response to Water,* FAO Irrigation and Drainage Paper 33; FAO: Rome, 1979.
10. Rao, N.H. Field test of a simple soil-water balance model for irrigated areas. J. Hydrol. **1987**, *91,* 179–186.
11. Howitt, R.E. Positive mathematical-programming. Am. J. Agric. Econ. **1995**, 77 (2), 329–342.
12. Doll, P.; Siebert, S. Global modeling of irrigation water requirements. Water Resour. Res. **2002**, *38* (4), 1037.
13. Yang, Y.M.; Yang, Y.H.; Moiwo, J.P.; Hu, Y.K. Estimation of irrigation requirement for sustainable water resources reallocation in North China. Agric. Water Manage. **2010**, *97,* 1711–1721.
14. Amir, I.; Fisher, F.M. Analyzing agricultural demand for water with an optimizing model. Agric. Syst. **1999**, *61,* 45–56.
15. Salman, A.Z.; AI-Kaeablieh, E.K.; Fisher, F.M. An interseasonal agricultural water allocation system (SAWAS). Agric. Syst. **2001**, *68,* 233–252.
16. Montesinos, P.; Camacho, E.; Alvarez, S. Seasonal furrow irrigation model with genetic algorithms (optimec). Agric. Water Manage. **2001**, 52(1), 1–16.
17. Pais, M.S.; Ferreira, J.C.; Teixeira, M.B.; Yamanaka, K.; Carrijo, G.A. Cost optimization of a localized irrigation system using genetic algorithms. Intell. Data Eng. Autom. Learn. **2010**, *6283,* 29–36.
18. Zhang, X.D.; Huang, G.H.; Nie, X.H. Optimal decision schemes for rural water quality management planning with imprecise objective. Agric. Water Manage. **2009**, *96,* 1723–1731.
19. Zhang, W.S.; Wang, Y.; Peng, H.; Li, Y.T.; Tang, J.S.; Wu, K.B. A coupled water quantity–quality model for water allocation analysis. Water Resour. Manage. **2010**, *24,* 485–511.
20. Wang, J. The estimation of evapotranspiration with Penman–Monteith and evaporator methods. Agric. Res. Arid Areas **2002**, 20(4), 67–71.
21. Sang, S. Simulation–optimization method for crop irrigation scheduling with limited water supply. J. Tsinghua Univ. **2005**, *45* (9), 1179–1183.
22. Lin, R.; Liang, Y. Characteristics of dry matter accumulation and partitioning in the process of yield formation in different rice cultivate. China Agric. Sci. Bull. **2006**, *22* (2), 185–190.
23. Liu, L.; Huang, G.H.; Liu, Y.; Fuller, G.A. A fuzzy- stochastic robust optimization model for regional air quality management under uncertainty. Eng. Optim. **2003**, 35 (2), 177–199.
24. Charnes, A.; Cooper, W.W.; Kirby, P. Chance constrained programming: An extension of statistical method. In *Optimizing Methods in Statistics;* Academic Press: New York, 1972; 391–402.
25. Qin, X.S.; Huang, G.H. An inexact chance-constrained quadratic programming model for stream water quality management. Water Resour. Manage. **2009**, *23,* 661–695.
26. Xu, Y.; Huang, G.H.; Qin, X.S.; Cao, M.F. SRCCP: A stochastic robust chance-constrained programming model for municipal solid waste management under uncertainty. Resour., Conserv. Recycl. **2009**, 53, 352–363.
27. Huang, G.H.; Loucks, D.P. An inexact two-stage stochastic programming model for water resources management under uncertainty. Civ. Eng. Environ. Syst. **2000**, *17,* 95–118.

28. Xu, Y.; Huang, G.H.; Qin, X.S. An inexact two-stage stochastic robust optimization model for water resources management under uncertainty. Environ. Eng. Sci. **2009,** *26,* 1765–1776.
29. Li, Y.P.; Huang, G.H.; Nie, S.L. An interval-parameter multi-stage stochastic programming model for water resources management under uncertainty. Adv. Water Resour. **2006**, *29*, 776–789.
30. Li, Y.P.; Huang, G.H.; Yang, Z.F.; Nie, S.L. IFMP: Interval-fuzzy multistage programming for water resources management under uncertainty. Resour., Conserv. Recycl. **2008**, 52, 800–812.
31. Watkins, D.W., Jr.; McKinney, D.C. Finding robust solutions to water resource problems. J. Water Resour. Plann. Manage. **1997**, *123,* 49–58.
32. Xu, Y.; Huang, G.H.; Qin, X.S.; Cao, M.F.; Sun, Y. An interval-parameter stochastic robust optimization model for supporting municipal solid waste management under uncertainty. Waste Manage. **2010**, *30* (2), 316–327.
33. Huang, G.H.; Baetz, B.W.; Patry, G.G. A grey fuzzy linear programming approach for municipal solid waste management planning under uncertainty. Civ. Eng. Syst. **1993**, *10,* 123–146.
34. Zimmermann, H.J. *Fuzzy Set Theory and Its Applications,* 3rd Ed; Kluwer Academic Publishers, 1995.
35. Liu, B.D.; Iwamura, K. Chance-constrained programming with fuzzy parameters. Fuzzy Sets Syst. **1998**, *94,* 227–237.
36. Huang, G.H.; Baetz, B.W.; Patry, G.G. A grey linear programming approach for municipal solid waste management planning under uncertainty. Civ. Eng. Syst. **1992**, 9, 319–335.
37. Huang, G.H.; Baetz, B.W.; Patry, G.G. Grey integer programming: An application to waste management planning under uncertainty. Eur. J. Oper. Res. **1995**, *83,* 594–620.
38. Huang, G.H.; Baetz, B.W.; Patry, G.G. Grey fuzzy integer programming: An application to regional waste management planning under uncertainty. Socio-Economic Plann. Sci. **1995**, 29, 17–38.
39. Qin, X.S.; Huang, G.H.; Zeng, G.M.; Chakma, A.; Huang, Y.F. An interval-parameter fuzzy nonlinear optimization model for stream water quality management under uncertainty. Eur. J. Oper. Res. **2007**, *180* (3), 1331–1357.
40. Loucks, D.P.; Stedinger, J.R.; Haith, D.A. *Water Resource Systems Planning and Analysis;* Prentice-Hall: Englewood Cliffs, NJ, 1981.
41. Fiedler, M.; Nedoma, J.; Ramík, J.; Rohn, J.; Zimmermann, K. Linear Optimization Problems with Inexact Data; Springer Science Business Media, Inc., New York, 2006.
42. Wardlaw, R.; Bhaktikul, K. Application of a genetic algorithm for water allocation in an irrigation system. Irrig. Drain. **2001**, *50,* 159–170.

43

Erosion by Water: Vegetative Control

Seth M. Dabney and
Silvio J. Gumiere

Introduction

Vegetation controls erosion by dissipating the erosive forces of rainfall and runoff (erosivity) and by reducing the susceptibility of soil to erosion (erodibility). Vegetation alters the partitioning of rainfall between infiltration, surface storage, and surface runoff. Erosivity is reduced because rainfall kinetic energy is absorbed, runoff volume is reduced due to increased infiltration, and runoff velocity is slowed through increased surface detention and reduced development of areas of concentrated flow. Vegetation also reduces soil erodibility by increasing soil aggregation, binding aggregates together with roots, and lowering soil matric potential. Vegetation may cover the entire soil surface, as with crops, cover crops, or forests, or it may be limited to specific critical areas, as with various types of conservation buffers. This entry reviews the mechanisms and processes by which vegetation reduces soil erosion by water and decreases yield of sediment and associated pollutants from agricultural fields. Emphasis is placed on vegetative buffers. Crop residue effects are considered in another entry.

General Mechanisms

Slower Runoff

Theoretically, if runoff occurs uniformly over a plane, its depth increases in a predictable manner as slope length increases. In practice, the development of concentrated-flow areas of high velocity limits the depth of sheet flows. By slowing runoff, vegetation can reduce or delay the development of rills and associated concentrated-flow erosion. Vegetation may increase runoff depth 10-fold compared with an equivalent discharge over a smooth surface or fivefold deeper than rainfall-impacted flow over a natural bare soil surface.[1] By increasing water depth fivefold, average velocity, V, is reduced fivefold. Since erosivity of runoff is proportional to V^2 and its sediment transport capacity is proportional to V^5,[2] vegetation reduces concentrated-flow erosion and increases deposition of previously eroded sediment. Sediment trapping is most efficient where the flow is shallow and uniform and the filter is not submerged.[3–5]

The retardation of surface runoff is a critical aspect of the functioning of conservation buffers. Figure 1 illustrates the situation where sediment-laden runoff encounters a vegetated buffer. Because of the additional hydraulic resistance of stems and leaves, flow depth within the buffer, D_2, is greater than in the influence area upslope of the buffer, D_0. The depth at the upslope edge of the buffer, D_1, however, is greater even than that within the buffer (D_2) because of the following: 1) enhanced vegetation growth at the buffer margin; 2) compression of stems into a denser barrier; and 3) loading of the buffer edge with trapped residues and thatch. In many studies, more than half of the sediment trapped by vegetated buffers is deposited in the ponded area upslope of the buffer. Where the ponded area is deep and slow flowing, transport capacity is negligible, and the water surface approaches horizontal. In these circumstances, [6] the fraction of particles with fall velocity V_{si} that will be trapped (T_{i}) is given by

$$T_i = 1 - \exp\left[-V_{\mathrm{si}}\ L/q\right] \tag{1}$$

where q is the specific discharge and L is the length of the pond (Figure 1). When the ponded area retains significant transport capacity, trapping efficiency is reduced, and a transport capacity or sediment re-entrainment term must be added.[7] To accurately capture the sediment trapping capability of narrow strips of vegetation, models must consider backwater effects and realistic flow rates.[8] We will consider achievable sediment trapping efficiencies in a later section.

Increased Infiltration of Water into Soil

Vegetation increases infiltration by reducing the development of surface seals that limit infiltration rates, increasing soil water storage capacity through evapotranspiration, and developing soil macroporosity through root growth and enhanced activities of mesofauna such as earthworms and ants. By covering the soil and absorbing the kinetic energy of raindrops, vegetation can prevent the detachment and rearrangement of soil particles that result in the creation of soil surface seals.[9] Although water use varies with species and climate, vegetation transpires approximately 0.3 m^3 of water for each kilogram of aboveground dry matter produced.[10] This transpiration leaves more capacity in the soil for infiltration of subsequent rains and thus reduces runoff and erosion.[11] Vegetation increases soil macroporosity directly through root growth[12] and indirectly by improving the habitat and activity of mesofauna[13] so that the benefits from vegetation on infiltration increase over time.[14] By slowing runoff due to hydraulic roughness, vegetation increases the depth of ponded water and the area of soil that is submerged, thus increasing opportunities for macropore flow infiltration.

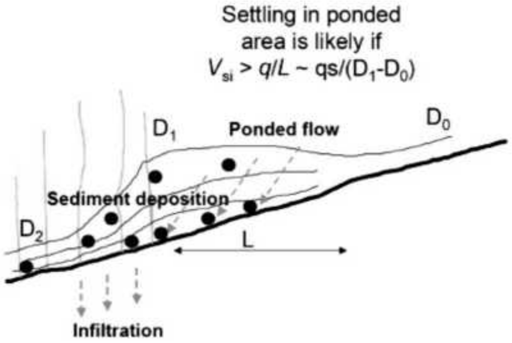


FIGURE 1 Schematic illustration of how vegetative buffers slow runoff, increasing flow depth and trapping sediment.

Reduced Soil Erodibility

Soil erodibility refers to the ease with which soil particles (primary or aggregates) can be detached and transported by the shear forces associated with raindrop splash or flowing water. Soil with increased organic matter content has greater aggregate stability[15] and, hence, greater resistance to detachment and transport. Vegetation may influence soil erodibility due to the additions of soil organic matter from vegetation from litter and roots.[16,17] The effects of vegetation on reducing erodibility include consolidation of soil with time after tillage and binding together of soil particles by roots and by microorganisms, particularly fungi, that use plant biomass and exudates as a food source.[18–20] Although seedling corn roots were not found to significantly reduce interrill erosion,[21] mechanical reinforcement increases soil shear strength and reduces rill, gully, and stream-bank erosion.[22,23]

Vegetative Buffers

Buffer Types

Conservation buffers designed to reduce soil erosion and/ or sediment delivery are usually areas of perennial vegetation placed at critical points in a landscape. These buffers may be located along stream banks or along the edges of fields or may be placed within fields. To distinguish among these buffer types, the nomenclature of the United States Department of Agriculture–Natural Resources Conservation Service (NRCS) is adopted.

The 10 types of conservation buffers that help to prevent soil erosion are summarized in Table 1.[24] The first seven of these buffers are intended to reduce water erosion and sediment delivery in runoff, and the last three help control wind erosion. The first three water erosion control practices are located at the edges of fields, while the other four are usually placed within agricultural fields. Six of the water erosion control buffer types are illustrated in Figure 2.[25] In addition to controlling erosion and/or reducing sediment delivery, many of these buffers can also serve additional purposes, such as improving water quality, sequestering carbon, and providing wildlife habitat (Table 1). Current national standards for these practices are given in the NRCS *National Handbook of Conservation Practices*.[26] Descriptive information about each practice can be found in the CORE4 training materials.[27] Local specification criteria can be found in the local NRCS *Field Office Technical Guide*.

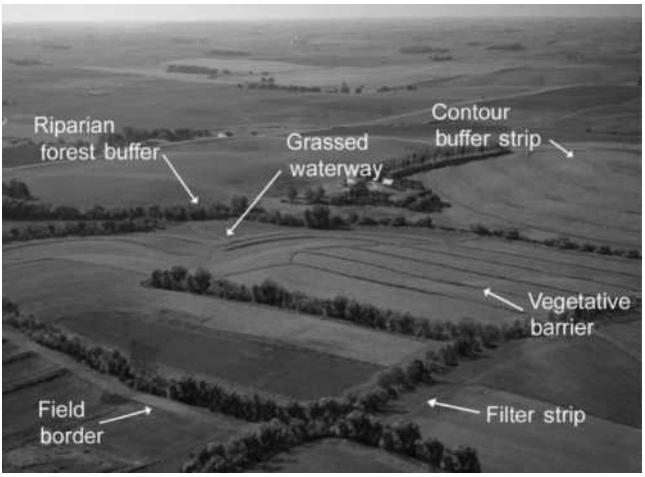


FIGURE 2 Schematic illustration of six in-field and edge-of-field buffer types.
Source: Dabney et al.[25]

TABLE 1 Comparison of Selected Purposes and Criteria from the USDA-NRCS *National Handbook of Conservation Practices* for the 10 CORE4 Buffer Types and Some Related practices

		Erosion Control Purposes					Other Purposes					Criteria (Minimum or Maximum)				
Practice	NRCS code	Reduce sheet and rill erosion	Reduceconc. flow erosion	Reduce wind erosion	Reduce sediment delivery	Increase wildlife habitat	Reduce contam. transport	Increase carbon storage	Produce harvest	Protect crops	Field slope	A long-strip gradient	Strip width (SW)	Strip spacing	Field length	Stem density
CORE4 buffer types																
Riparian forest buffer	391^a		+		+	+	+	+	+			Along stream	5–30 m			
Field border	386		+	+	+	+	+	*				Field edge	>6 m			
Filter strip	393^a				+	+	+	*			1% to 10%	<0.5 %	>6 m		<50*SW	1500/m^2
Grassed waterway	412^a		+		+	*	+	*				Along gradient		Concentrated-flow areas		n-VR curve, permissible velocity
Alley cropping	311	+		*	+	+	+	+	+	+		Contour	>6 m	Species light requriements		
Contour buffer strip	332^a	+			+	*	+	*			2% to 8%	<2%	>5 m (grass), >9 m (legume)	1/2 RUSLE crit. slope length (CSL)	RUSLE CSL	540/m^2 (grass), 320/m^2 (legume)
Vegetative barrier	601	+	+		+	*		*				<1%	>1 m	1.3 to 2.0 m vertical int.		Depends on stem diameter
Shelter belt	38^a			+	+	+	*	+	+	Across prevailing			12 H, +			

(*Continued*)

TABLE 1 (***Continued***) Comparison of Selected Purposes and Criteria from the USDA-NRCS *National Handbook of Conservation Practices* for the 10 CORE4 Buffer Types and Some Related Practices

		Erosion Control Purposes					Other Purposes					Criteria (Minimum or Maximum)				
Practice	NRCS code	Reduce sheet and rill erosion	Reduceconc. flow erosion	Reduce wind erosion	Reduce sediment delivery	Increase wildlife habitat	Reduce contam. transport	Increase carbon storage	Produce harvest	Protect crops	Field slope	A long-strip gradient	Strip width (SW)	Strip spacing	Field length	Stem density
Crosswind trap strip	589C			+	+	+	+	*	+	Wind direction		5 to 8 m	Upwind of protected area			
Herbaceous wind b arrier	422A[a]			+	+	+	*	*	+			1 m	<10 H, +	At least 2 rows, 0.5 m tall (H)		
Related practices with buffering attributes																
Constructed wetland	656				+	*	+	*								
Channel vegetation	322		+		+		+	*		Channel banks						
Terrace	600	+	+		+		+		*	Across slope			>27 m			
Water and sediment cont. basin	638		+		+		+		*	Across flow areas						
Grade stabilization structure	410	*	+		+		+		*	Field edge, side inlet						
Farm ponds/ in-stream wetlands	378		*		+	*	+	*	+							

Source: Lowrance et al.[24]

+ indicates a purpose for which practice may be designed; * reflects an area of potential impact but not a design purpose. [26]

[a] Eligible for cost sharing under national CRP continuous sign-up.

FIGURE 3 Vegetative barriers of vetiver grass *(Vetiveria zizanioides)* planted in rows on contour lines to hold the soil in St. Vincent, British West Indies, during the 1950s.
Source: Vélez.[29]

The three edge-of-field water erosion control buffers are riparian forest buffer (RFB), filter strip (FS), and field border (FB). An RFB is a forested area adjacent to a water body and is frequently combined with grass buffers. An FB is a grassed field margin. Because it may be used for parking and turning equipment, an FB is also usually wider than the minimum indicated in Table 1. In contrast to an FB, traffic is usually excluded from an FS, and vegetation and slope requirements are far more stringent (Table 1). Generally, edge-of-field buffers are designed primarily to trap sediment and infiltrate water, not to control erosion. The RFB is an exception in that it can control concentrated-flow erosion, but this function refers to reducing the velocity of out-of-bank flood flows, thereby minimizing floodplain scour and promoting deposition, and not to reducing erosion from locally generated concentrated runoff. To properly function, these edge-of-field buffers require that runoff pass through them as diffuse sheet flow.

The next four buffer types in Table 1 function within fields and are designed to control in-field erosion. Three of these buffers, alley cropping, contour buffer strip, and vegetative barrier (VB), control sheet-and-rill erosion by interrupting hillslopes with strips of permanent vegetation aligned close to the contour. The widths of these buffers are often varied so that the edges of each cropped zone stay parallel and within strip gradient specifications (Table 1, Figure 2). Alley cropping involves growing crops and forages between strips of trees. Vegetative barriers are usually narrow strips of large stiff-stemmed grasses (Figure 3). Contour buffer strips are somewhat wider strips with less stringent vegetation and contour alignment requirements (Table 1).

Only two buffer practices, grassed waterway and VB, may be specifically designed to control in-field concentrated-flow erosion. Grassed waterways are oriented up and down the slope and are planted with vegetation that is intended to be submerged while functioning. In contrast, VBs designed to control concentrated-flow erosion are planted perpendicular to the flow direction and are intended to remain unsubmerged while retarding runoff.

Buffer Hydraulic Resistance

For fluid flowing over vegetated areas, the total drag force can be divided into two components: 1) the drag force due to the stems and leaves; and 2) the drag force generated by the bed boundary roughness. The drag force due to vegetation may be described by Chow:[28]

$$F_D = \frac{1}{2} C_D \rho A V^2 \tag{2}$$

where C_D is the drag coefficient of vegetation, V is the free stream velocity moving past the stems, A is the cross-sectional momentum-absorbing area of vegetation projected in a plane normal to the flow, and p is the water density.

The hydraulic resistance of vegetation frequently is parameterized with Manning's equation:

$$V = \frac{1}{n} R^{2/3} S^{1/2} \tag{3}$$

where R is the hydraulic radius (flow area divided by wetted perimeter), S is the land slope gradient, and n is a hydraulic resistance parameter (Manning's roughness coefficient). For steady uniform flow where total shear stress is equal to $\rho g R S$, the Manning coefficient for vegetation can be expressed as:

$$n = \left(\frac{1}{2g} R^{1/3} C_D \left(\frac{A}{bl} \right) \right)^{1/2} \tag{4}$$

where g is the gravity acceleration, and b and l are the width and length of the area containing projected vegetal area A, so that the dimensionless term (A/bl) is the vegetal projected area density.

Fig. 4 illustrates how Manning's n varies with the product V and R for three kinds of buffer vegetation. For broad uniform flows, the hydraulic radius reduces to the flow depth, H, and VR is approximately equal to the specific discharge. Also, since the viscosity of water is generally constant, the VR product is also directly related to the flow Reynolds number. At very low flow rates, surface drag caused by soil surface roughness or drag on surface residues may dominate total flow resistance and buffers may have little effect on erosion beyond their ability to reduce detachment. At somewhat higher flow rates, where depths have become several multiples of the diameter of the vegetal elements, the dominant component of hydraulic resistance is drag on emergent stems. When the stems are uniform with height, such as with the simulated filter strips made of brush bristles (Fig. 4) in a flume with a smooth floor,[30] average velocity remains constant with increasing flow and n increases in proportion to the 2/3 power of discharge, as predicted from equation (3).[31,32]

Once vegetation becomes submerged, Manning's n decreases with increasing discharge. At very high flows, all of the vegetation is deeply submerged, the main factor determining hydraulic resistance is the

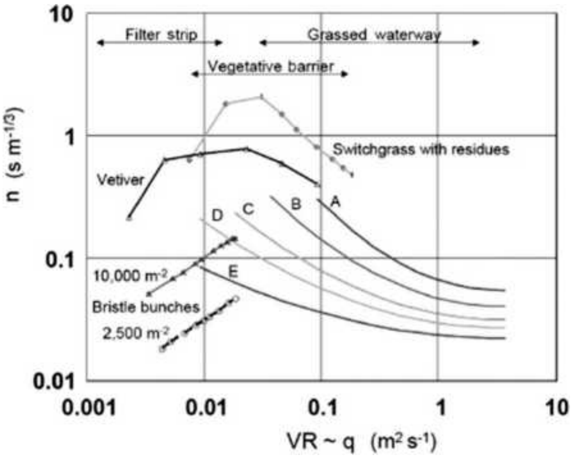


FIGURE 4 Hydraulic roughness of vegetated areas first increases with increasing flow as more vegetation interacts with the flow, then decreases with increasing flow as flow approaches the height of the vegetation and submerges it. Data for A through E from brush bristle data from Jin et al.[30] and Temple et al.[33]; switchgrass (*Panicum virgatum*) data from Dabney[37]; and data on vetiver from Dabney (unpublished).

length of the stems that are dragging in the flow, [33] and momentum transfer is dominated by shear-layer vortices.[34] The vegetal retardance curve labeled “A” in Figure 4 represents 0.9 to 1.0 m tall vegetation while “E” reflects vegetation that had been burned or mowed at about 4 cm height. In designing a GW, the erodibility of the underlying soil and the growth characteristics of the vegetal cover determine a maximum permissible velocity or the allowable hydraulic stress on the soil, and the channel is designed with dimensions great enough that, with expected vegetation, the permissible velocity or stress will not be exceeded at the design discharge.

Vegetative barriers have application at specific discharges that span the range between those of FS and GW (Figure 4) and can thus be used to complement other buffer types by spreading out concentrated runoff. At low flows, the hydraulic resistance of VB increases more rapidly than the 2/3 power of discharge because stems and leaves become less clumped together, increasing projected area with increasing height in the lower canopy.[35] At greater discharges, flow depth increases to the point where stems begin to thin out or bend. Then average velocity increases, the flow resistance, expressed as Manning’s n, reaches a peak value. As with grain roughness,[36] this peak occurs slightly before complete inundation. Beyond this point, hydraulic resistance begins to decline, even while flow depth may continue to increase with increasing discharge.[37] The stiff grasses used to form VB remain erect and emergent at greater flows than other vegetation types in Figure 4 because the large-diameter stems are stiffer and on the order of 2 m tall. The enhanced growth and residue loading noted to occur at the edge of all buffers are also important factors that give VB greater hydraulic resistance than retardance class “A” vegetation. Riparian forest buffer vegetation, of course, remains erect at even greater flows than does VB vegetation, but usually offers less hydraulic resistance at low flows.

Riparian buffers affect flow resistance in streams as well as on floodplains, and they can greatly increase the stability of stream banks through a combination of root reinforcement and hydrologic effects.[38] Riparian buffers generally reduce flow conveyance through drag and by supplying large wood debris to the channel.[39] Reach-scale roughness coefficients of narrow channels may be increased as much as three-or fourfold when grassy banks are replaced by dense woody growth, but effects tend to be much less pronounced for wider channels.[28] The flow resistance induced by a riparian buffer depends on vegetation density, whether the vegetation is rigid or flexible, and whether it is submerged or emergent. Models have been developed to simulate riparian vegetation effects on flow, sediment transport,[40–42] and channel morphology.[43]

Sediment Trapping Efficiency

Native or planted vegetation can remove sediments and other pollutants from surface runoff by filtration, deposition, infiltration, adsorption, absorption, decomposition, and volatilization.[44] However, vegetal effects on flow retardation and sediment transport greatly depend on specific flow rates (Figure 4). Research conducted using rainfall simulators and small plots usually collects buffer performance data over only a narrow range of very low flow rates (below the scale of Figure 4) that cannot represent probable performance under real-world conditions. Unfortunately, such studies are common. For example, a recent extensive review[44] included only two specific flow rates equal to or greater than $0.001\,m^2s^{-1}$. Under those conditions, buffers comprised of either tilled cropland or grass usually trap more than 50% of incoming sediment.[45,46]

At realistic flow rates, the benefits of very narrow, continuous buffers will be underestimated if backwater effects are ignored. Physically based sediment transport models that employ a kinematic wave approximation and ignore[47] or oversimplify[48] sediment deposition in backwater areas can greatly underestimate sediment trapping by narrow buffers[49] where the backwater length may be as large or larger than the buffer width. More physically realistic representations or approximations would recognize the influences of deposited sediment on increasing the water elevation at the buffer and reducing the flow’s sediment transport capacity. Deposited sediment initially flattens local slope, causing a progressive extension of the upslope sediment depositional area[6,8] until the flow over a steepened deposition area has sufficient capacity to transport the sediment load into the buffer.[50]

Flow concentration due to topographic nonuniformity can lead to higher specific flow rates through some portions of buffers than others.[51,52] Sediment trapping efficiency will go down with increasing discharge unless depth and backwater length increase more than specific discharge (Eq. 1). The VB behavior at specific flow rates between 0.001 and 0.01 $m^2 s^{-1}$ (Figure 4) suggests that this may occur under a limited set of conditions. The ability of emergent vegetation to maintain an even distribution of flow despite microscale variation in topography or vegetation density may contribute to the finding that the addition of a VB to an FS increased its sediment retention effectiveness, even when flow was concentrated.[53]

Sediment retention effectiveness of grass[54] and woody[55] vegetative buffers improves over a period of years after establishment due to increased density of grass stem density and/or increased infiltration rate in vegetated areas. Significant gaps or channels through buffers can greatly reduce performance of a buffer of any size. Tillage-induced ridges and channels at the edges of vegetated areas may alter the hydraulic flow patterns of surface runoff, which has the potential for altering the effectiveness of buffers in real-world applications.[56] Field studies demonstrate a wide variety of sediment trapping efficiencies that reflect the local impacts of vegetation buffers on hydrological and sedimentological connectivities. Sediment trapping efficiency was about 50% for grass buffers on 45 m long plots with a 12% to 17% slope in Norway,[57] and a 6 m buffer reduced sediment yield from 35 m long, 1.8% slopes by over 95% in northeast Italy.[58] A riparian buffer along the last 300 m of an ephemeral stream draining a 249 ha watershed trapped 18% of the total sediment load in China.[59] Riparian buffers receiving runoff from hillslopes removed 80% to 90% of incoming sediment in North Carolina, United States.[60,61]

The spatial organization of natural and planned vegetative buffers can significantly alter the hydrological and sedimentological connectivity and, hence, the global water and sediment yields.[62–64] In some cases, the location of the vegetation has more effect than a variation in the size of the buffer.

Conclusions

The effectiveness of vegetation to control erosion and reduce sediment delivery is due to its ability to reduce both detachment of soil particles and the transport capacity of sediment-laden flows. The first increment of buffer has a larger impact on sediment transport than any subsequent increment.[25] Erect emergent vegetation helps to spread runoff and maintain the sediment retention functionality of buffers over a wider range of variation in topography and flow rate. Buffers alter sedimentological connectivity both at the local scale between two adjacent fields and in a distributed way at the watershed scale. Optimal buffer placement in a landscape will rarely resemble a uniformly wide green strip along a stream. Rather, optimal arrangement of buffers for multiple resource concerns calls for a variety of buffer types, sizes, and shapes at different landscape locations starting in watershed headwater areas.[24]

References

1. Parsons, D.A. *Depths of Overland Flow,* SCS-TP-82; U.S. Dept. of Agriculture-Soil Conservation Service: Washington, DC, 1949; 1–33.
2. Meyer, L.D.; Wischmeier, W.H. Mathematical simulation of the process of soil erosion by water. Trans. ASAE **1969**, *12,* 754–758, 762.
3. Barfield, B.J.; Tollner, E.W.; Hayes, J.C. Filtration of sediment by simulated vegetation I. Steady-state flow with homogenous sediment. Trans. ASAE **1979**, *22*, 540–548.
4. Dillaha, T.A.; Reneau, R.B.; Mostaghimi, S.; Lee, D. Vegetative Filter strips for agricultural non-point source pollution-control. Trans. ASAE **1989**, *32,* 513–519.
5. Deletic, A.; Fletcher, T.D. Performance of grass filters used for stormwater treatment—A field and modelling study. J. Hydrol. **2006**, *317,* 261–275.

6. Dabney, S.M.; Meyer, L.D.; Harmon, W.C.; Alonso, C.V.; Foster, G.R. Depositional patterns of sediment trapped by grass hedges. Trans. ASAE **1995**, *38* (6), 1719–1729.
7. Beuselinck, L.; Govers, G.; Steegen, A.; Harisine, P.B.; Poesen, J. Evaluation of the simple settling theory for predicting sediment deposition by overland flow. Earth Surf. Processes Landforms **1999**, *24,* 993–1007.
8. Hussein, J.; Yu, B.; Ghadiri, H.; Rose, C. Prediction of surface flow hydrology and sediment retention upslope of veti-ver buffer strip. J. Hydrol. **2007**, *338*, 261–272.
9. Römkens, M.J.M.; Prasad, S.N.; Whisler, F.D. Surface sealing and infiltration. In *Process Studies in Hillslope Hydrology*; Anderson M.G., Burt, T.P., Eds.; John Wiley and Sons Ltd.: Chichester, U.K., 1990; 127–172.
10. Tanner, C.B.; Sinclair, T.R. Efficient water use in crop production: Research or re-search. In *Limitations to Efficient Water Use in Crop Production;* Taylor, H.M. Ed.; American Society of Agronomy: Madison, WI, 1983; 1–27.
11. Zhu, J.C.; Gantzer, C.J.; Anderson, S.H.; Alberts, E.E.; Beuselinck, P.R. Runoff, soil and dissolved nutrient losses from no-till soybean with winter cover crops. Soil Sci. Soc. Am. J. **1989**, *53,* 1210–1214.
12. Elkins, C.B.; Harland, R.L.; Homeland, C.S. Grass roots as a tool for penetrating soil hardpans and increasing crop yields. Proc. Southern Pasture Forage Crop Imp. Conf. **1977**, *34,* 21–26.
13. Tomlin, A.D.; Shipitalo, M.J.; Edwards, W.M.; Protz, R. Earthworms and their influence on soil structure and infiltration. In *Earthworm Ecology and Biogeography in North America;* Hendrix, P.F., Ed.; Lewis Pub.: Boca Raton, FL, 1995; 159–183.
14. Le Bissonnais, Y.; Lecomte, V.; Cerdan, O. Grass strip effects on runoff and soil loss. Agronomie **2004**, *24,* 129–136.
15. Angers, D.A. Changes in soil aggregation and organic carbon under corn and alfalfa. Soil Sci. Soc. Am. J. **1992**, *56,* 1244–1249.
16. Le Bissonnais, Y. Aggregate stability and assessment of soil crustability and erodibility: I. Theory and methodology. Eur. J. Soil Sci. **1996**, *47,* 425–437.
17. Cerdà, A. Soil aggregate stability under different Mediterranean vegetation types. Catena **1998**, *32,* 73–86
18. Renard, K.G.; Foster, G.R.; Weesies, G.A.; McCool, D.K.; Yoder, D.C. (coordinators). *Predicting Soil Erosion by Water: A Guide to Conservation Planning with the Revised Universal Soil Loss Equation (RUSLE)*, Agriculture Handbook 703; U.S. Department of Agriculture-Agricultural Research Service: Washington, DC, 1997; 1–384.
19. Cosentino, D.; Chenu, C.; Le Bissonnais, Y. Aggregate stability and microbial community dynamics under drying-wetting cycles in a silt loam soil. Soil Biol. Biochem. **2006**, *38*, 2053–2062.
20. Chenu, C.; Le Bissonnais, Y., Arrouays, D. Organic matter influence on clay wettability and soil aggregate stability. Soil Sci. Soc. Am. J. **2000**, *64,* 1479–1486.
21. Bui, E.; Box, J.J.E. Growing corn root effects on interrili soil erosion. Soil Sci. Soc. Am. J. **1993**, *57,* 1066–1070.
22. De Baets, S.; Poesen, J.; Gyssels, G.; Knapen, A. Effects of grass roots on the erodibility of topsoils during concentrated flow. Geomorphology **2006**, *76,* 54–67.
23. Pollen, N. Temporal and spatial variability in root reinforcement of streambanks: Accounting for soil shear strength and moisture. Catena **2007**, *69,* 197–205.
24. Lowrance, R.; Dabney, S.; Schultz, R. Improving water and soil quality with conservation buffers. J. Soil Water Con-serv. **2002**, *57* (2), 36A–43A.
25. Dabney, S.M.; Moore, M.T.; Locke, M.A. Integrated management of in-field, edge-of-field, and after-field buffers. J. Am. Water Resour. Assoc. **2005**, *42* (1), 15–24.
26. National Handbook of Conservation Practices, available at http://www.nrcs.usda.gov/technical/standards/nhcp.html (accessed May 2012).

27. CORE4 training materials, available at http://www.nrcs. usda.gov/Internet/FSE_DOCUMENTS/nrcs143_025540.pdf (accessed May 2012).
28. Chow, V.T. *Open-Channel Hydraulics;* McGraw-Hill: New York, 1959.
29. Vélez, I. Soil conservation practices in the Caribbean archipelago. The Scientific Monthly **1952**, *74* (3), 183–185.
30. Jin, C.X.; Römkens, M.J.M.; Griffioen, F. Estimating Manning's roughness coefficient for shallow overland flow in non-submerged vegetative filter strips. Trans. ASAE **2000**, *43* (6), 1459–1466.
31. Petryk, S.; Bosmajian, G. Analysis of flow through vegetation. J. Hydraul. Div. ASCE **1975**, *101* (HY7), 871–884.
32. Nepf, H.M. Drag, turbulence, and diffusion in flow through emergent vegetation. Water Resour. Res. **1999**, *34* (2), 479–489.
33. Temple, D.M.; Robinson, K.M.; Ahring, R.M.; Davis, A.G. *Stability Design of Grass-lined Open Channels*, Agriculture Handbook 667; U.S. Department of Agriculture–Agricultural Research Service: Washington, DC, 1987.
34. Nepf, H.; Ghisalberti, M. Flow and transport in channels with submerged vegetation. Acta Geophys. **2008**, *56* (3), 753–777.
35. Järvelä, J. Flow resistance of flexible and stiff vegetation: A flume study with natural plants. J. Hydrol. **2002**, *269*, 44–54.
36. Lawrence, D.S.L. Hydraulic resistance in overland flow during partial and marginal surface inundation: Experimental observations and modeling. Water Resour. Res. **2000**, *36,* 2381–2393.
37. Temple, D.; Dabney, S. Hydraulic performance testing of stiff grass hedges. In the Proceedings of the Seventh Federal Interagency Sedimentation Conference, Reno, Nevada, March 25–29, 2001; Vol. 2(XI), 118–124.
38. Pollen-Bankhead, N., Simon, A. Hydrologic and hydraulic effects of riparian root networks on streambank stability: Is mechanical root-reinforcement the whole story? Geomorphology **2010**, *116,* 353–362.
39. Shields, F.D., Jr.; Gippel, C.J. Prediction of effects of woody debris removal on flow resistance. J. Hydraul. Eng. **1995**, *121,* 341–354.
40. Wu, W.; Shields, F.D., Jr.; Bennett, S.J.; Wang, S.S.Y. A depth-averaged two-dimensional model for flow, sediment transport, and bed topography in curved channels with riparian vegetation. Water Resour. Res. **2005**, *41,* W03015.
41. Wilson, C.A.M.E.; Stoesser, T.; Bates, P.D. Modelling of open channel flow through vegetation. In *Computational Fluid Dynamics: Applications in Environmental Hydraulics;* Bates, P.D., Lane, S.N., Ferguson, R.I., Eds., John Wiley and Sons Ltd.: Chichester, U.K., 2005; 395–428.
42. Wu, W.; He, Z. Effects of vegetation on flow conveyance and sediment transport capacity. Int. J. Sediment Res. **2009**, *24,* 247–259.
43. Li, S.S.; Millar, R.G. A two-dimensional morphodynamic model of gravel-bed river with floodplain vegetation. Earth Surf. Processes Landforms **2011**, *36,* 190–202.
44. Gumiere, S.J.; Bissonnais, Y.L.; Raclot, D.; Cheviron, B. Vegetated filter effects on sedimentological connectivity of agricultural catchments in erosion modeling: A review. Earth Surf. Processes Landforms **2011**, *36,* 3–19.
45. Young, R.A.; Huntrods, T., Anderson, W. Effectiveness of vegetated buffer strips in controlling pollution from feedlot runoff. J. Environ. Qual. **1980**, *9*, 483–487.
46. Magette, W.L.; Brinsfeld, R.B.; Palmer, R.E.; Wood, J.D. Nutrient and sediment removal by vegetated filter strips. Trans. ASAE **1989**, *32,* 663–667.
47. Flanagan, D.C.; Foster, G.R.; Neibling, W.H.; Burt, J.P. Simplified equations for filter strip design. Trans. ASAE **1989**, *32* (6), 2001–2007.
48. Muñoz-Carpena, R.; Parsons, J.E.; Gilliam, J.W. Modeling hydrology and sediment transport in vegetative filter strips. J. Hydrol. **1999**, *214,* 11–129.

49. Muñoz-Carpena, R.; Parsons, J.E. A design procedure for vegetative filter strips using VSMOD-W. Trans. ASAE **2004**,*74* (4), 1933–1941.
50. Foster, G.R. Modeling the erosion process. In *Hydrologic Modeling of Small Watersheds;* Hann, C.T., Ed.; American Society of Agricultural Engineers: St. Joseph, MI, 1982; 296–380.
51. Dosskey, M.G.; Helmers, M.J.; Eisenhauer, D.E.; Franti, T.G.; Hoagland, K.D. Assessment of concentrated flow through riparian buffers. J. Soil Water Conserv. **2002**, *57* (6), 336–343.
52. Helmers, M.J.; Eisenhauer, D.E.; Dosskey, M.G.; Franti, T.G.; Brothers, J.M.; McCollough, M.C. Flow pathways and sediment trapping in a field-scale vegetative filter. Trans. ASAE **2005**, *48* (3), 955–968.
53. Blanco-Canqui, H.; Gantzer, C.J.; Anderson, S.H. Performance of grass barriers and filter strips under interrill and concentrated flow. J. Environ. Qual. **2006**, *35*, 1969–1974.
54. Dabney, S.M.; McGregor, K.C.; Wilson, G.V.; Cullum, R.F. How management of grass hedges affects their erosion reduction potential. Soil Sci. Soc. Am. J. **2009**, *73* (1), 241–254.
55. Dosskey, M.G.; Hoagland, K.D; Brandle, J.R. Change in filter strip performance over ten years. J. Soil Water Conserv. **2007**, *62* (1), 21–32.
56. Dabney, S.M.; Vieira, D.A.N. Terrace formation by tillage erosion. In *Encyclopedia of Environmental Management*, this volume.
57. Syversen, N.; Borch, H. Retention of soil particle fractions and phosphorus in cold-climate buffer zones. Ecol. Eng. **2005**,*25*, 382–394.
58. Borin, M.; Vianello, M.; Morari, F.; Zanin, G. Effectiveness of buffer strips in removing pollutants in runoff from a cultivated field in North-East Italy. Agric. Ecosyst. Environ. **2005**, *105,* 101–114.
59. Wang, X.H.; Yin, C.Q.; Shan, B.Q. The role of diversified landscape buffer structures for water quality improvement in an agricultural watershed. North China. Agric. Ecosyst. Environ. **2005**, *107,* 381–396.
60. Daniels, R.B.; Gilliam, J.W. Sediment and chemical load reduction by grass and riparian filters. Soil Sci. Soc. Am. J. **1996**, *60,* 246–251.
61. Cooper, J.R.; Gillian, J.W.; Daniels, R.B.; Robarge, W.P. Riparian areas as filters for agricultural sediment. Soil Sci. Soc. Am. J. **1987**, *51,* 416–420.
62. Lecomte, V. Transferts de produits phytosanitaires par le ruisellement et l'érosion de la parcelle au bassin versant. Modélisation spatiale. PhD thesis. ENGREF et INRA-Orléans, 1999.
63. Cammeraat, L.H.; Imeson, A.C. The evolution and significance of soil-vegetation patterns following land abandonment and fire in Spain. Catena **1999**, *37,* 107–127.
64. Fitzjohn, C.; Ternan, J.L.; Williams, A.G. Soil moisture variability in a semi-arid gully catchment: Implications for runoff and erosion control. Catena **1998**, *32,* 55–70.

44

Erosion Control: Soil Conservation

Eric T. Craswell

Introduction

Soil formation is a biologically driven process that is so slow, soil may be considered to be a nonrenewable resource. In most areas of the Earth's surface, the loss of soil due to wind or water erosion by natural processes is balanced by the soil formation creating a dynamic equilibrium so that soil depth depends largely on the climax vegetation. The exceptions are landscapes where volcanic activity exposes new parent material or changes the slope, or lack of precipitation, which leaves soil vulnerable to natural water and wind erosion. By far, the most widespread cause of erosion is human-induced land transformation for infrastructure, forestry, grazing, and crop production, which now covers 98 million km^2 of the Earth's land surface. Soil erosion by water can have serious local off-site impacts especially where the construction of roads and buildings occurs on unstable slopes in high rainfall areas. However, the most widespread water erosion of soils occurs in the vast areas of land developed for crop production and exposed by plowing to rainfall and runoff. The expansion of cropped land to more than 15 million km^2 has been driven by population growth, which is projected to continue to reach 9 billion in 2050. On the other hand, suitably stocked rangelands used for grazing, which retain a relatively complete vegetative cover, are generally protected from water erosion. On the 47 million km^2 of grazing lands, and especially in very dry areas, overgrazing causes de-vegetation, making erosion by wind, rather than by water, the major threat. Either cause of erosion results in loss of the surface layer of soil, which generally contains the highest level of nutrients and organic matter, and hence, the residual soil fertility is significantly reduced. Off-site impacts of the water- and wind-borne soil particles on water bodies and coastal zones are exacerbated by the nutrients they contain, which perturbate natural ecosystem functioning and nutrient cycling processes.

This entry reviews current knowledge of the problem of water and wind erosion of soils, focusing on the environmental impacts, then describes some of the technologies developed to conserve soil and minimize off-site impacts and community-based approaches to promote the technologies.

The Scope and Impact of Soil Erosion

Available global estimates[1] indicate that 7.48 million km^2 of land are affected by moderate to serious soil erosion by water and 2.8 million km^2 are affected by wind erosion. The widespread extent of erosion has raised concerns that the future of agriculture and civilization itself may be threatened.[2,3] The estimates[1] indicate that 6.73 million km^2 of the total soil erosion is occurring in Asia and Africa, which have large areas of tropical soils with high rainfall and steep slopes. These lands, which traditionally were cultivated infrequently by shifting cultivators, are being cropped more intensively or destabilized by inappropriate clearing methods used to establish plantation crops. The recent trend to expand biofuel crop production in these areas may accelerate soil erosion and increase the rates of sediment flow.

The movement of eroded soil in watersheds involves both transport and deposition of sediments in water bodies. Recently published global estimates suggest that human activity has increased the flow of sediments in the world's rivers by 2.3 ± 0.6 billion metric tons per annum.[4] However, the construction of dams has reduced the flow of sediment to the coasts by 1.4 ± 0.3 billion metric tons per annum. The estimates indicate that globally in the last 50 years, 100 billion tons of sediment accumulated in reservoirs, resulting in massive economic costs due to reduced storage capacity, damage to power-generating turbines, and reduced water quality for domestic and industrial uses. Such large off-site economic impacts of soil erosion make a strong case for investments to promote soil conservation. The most significant sediment outflows to the coast occur in areas with short, fast-flowing rivers in archipelago countries such as Indonesia and the Philippines. From an environmental protection perspective, an important ultimate destination of sediments whether natural or human induced is the coastal zone, where assets such as pristine coastal reefs can be severely affected by changed water quality and nutrient inputs. This kind of damage is widespread in some tropical coastal areas where negative economic impacts on the tourist and fishing industries can be serious.

Other off-site impacts of soil erosion by water include both positive and negative effects. Historically, sediment deposition in floodplain areas has contributed to the fertility of the soil of deltas that have become major population centers. However, sediment deposition in water bodies such as streams and lakes affects freshwater biodiversity and recreational fisheries and can cause streams to change course, leading to damage to infrastructure such as roads and bridges. On the other hand, poorly implemented construction of roads and bridges, and urban and industrial development in general, can be a major cause of localized soil erosion, the engineering aspects of which are addressed by organizations such as the International Erosion Control Association.[5] Drainage structures and the stabilization of steep banks with vegetation and screens can prevent surface erosion, but construction must also address the risk of landslips caused by subsurface flows.

Viewed at the landscape level, soil erosion on agricultural land may channel runoff, creating gullies that worldwide are responsible for 44% of soil erosion.[6] Thus, measurements of soil erosion on small plots, which dominated much of the soil erosion research in the past, do not capture the full extent and complexity of the problem. Furthermore, as with infrastructure developments, subsurface flow of water on hill slopes and stream banks also contributes significantly to the generation of sediment that impairs stream flow.[7] These considerations, which have only recently become better understood, strengthen the case for a landscape approach not only to research on soil erosion but also to the development and implementation of soil and water conservation practices.

The extent of wind erosion is less well understood than water erosion but is understood to be episodic and largely restricted to dryland areas. In China alone, 1.91 million km^2, equal to 20% of the land area of the country, is estimated to be affected.[8] The extent of the problem is increasing due to intensification of soil cultivation and increased grazing pressure on native pastures. This pattern is repeated in some other dryland and desert areas such as those in West Africa, Australia, and North America (where the infamous "dust bowl" of the 1930s occurred). Some scientists have labeled wind erosion as "dust storm erosion," indicating that the off-site impact on air quality should be emphasized. Although episodic and sometimes uncommon, these air quality impacts affect the health and lives of millions of urban dwellers far from the sites where wind erosion is occurring. The impacts of wind erosion and dryland degradation are considered so important by the international community that the United Nations Convention to Combat Desertification (UNCCD) was created. The UNCCD estimates that 70% of the Earth's drylands are degraded.[9] The extent of this problem is projected to expand in many regions due to climate change.

Productivity Impacts of Soil Erosion

The extent of soil erosion by water depends on the steepness and length of the slope of the land, the rainfall and runoff intensity/erosivity, the erodibility of the soil, and the management of the soil surface and other support practices such as strip cropping, vegetative strips, etc. These factors from the Revised Universal Soil Loss Equation (RUSLE) can be combined in an equation to calculate the average soil loss per annum.[10] By comparing this rate with an estimate of "tolerable" soil loss, soil conservationists can assess needs and plan soil and water conservation measures. Soil loss tolerance (T) is the maximum amount of soil loss in tons per acre per year that can be tolerated and still permit a high level of crop productivity to be sustained economically and indefinitely. The RUSLE has been extensively studied and used in North America where an extensive database provides a sound basis for estimating the empirical parameters for the case of mechanized agriculture. Other simulation models based on an understanding of the physical process of soil erosion provide an alternative to the RUSLE and can be adapted more easily to areas such as tropical steeplands. One example of such a model is the Griffith University Erosion System Template (GUEST),[11] which has been successfully applied to steep- land areas with high rainfall intensities in several Asian countries and Australia. The validated soil erosion model provides a basis for developing new tools for research including decision support tools for the evaluation of sustainable land management.[12]

The on-site impacts of soil erosion on productivity vary considerably and appear to be site specific. A review[13] of extensive experimental data sets showed that relative impacts of past erosion on crop yields measured in North America and Europe were two to three times smaller than those in Africa, Asia, Australia, and Latin America. Apparently, the lower crop yields in the latter regions caused yields to decline proportionally more significantly with the same amount of soil loss. Contributing to this difference may have been factors such as the shallower depth of the soil and the absence of the effects of higher fertilizer applications in North America and Europe that might mask the nutrient loss due to erosion. Further analysis of published data[14] showed that, on an annual global average basis, soil erosion caused a 0.3% loss of yield of six crops (maize, millet, potatoes, sorghum, soybeans, and wheat), resulting in an annual economic loss of $523 million. Such global averages do not take account of significant variations, especially in humid tropical areas where shallow soils, steep slopes, and high rainfall intensities make soil erosion by water a highly significant threat to the sustainability of smallholder farming.

Soil Erosion Control

The problem of soil erosion in steeplands spawned conservation measures that date back centuries and are still practiced in some developing countries. These technologies based on indigenous knowledge utilize the same principles as modern soil conservation measures and include a wide range of methods that comprise terraces, handmade ridges, natural vegetation strips, stone bunds, logs, etc. Some examples are listed in Table 1.

Table 2 shows the wide array of vegetative and mechanical approaches to soil conservation in agricultural landscapes. Some of these technologies are finding especially widespread applicability, depending on the farming system and site characteristics. Protecting the soil surface from raindrop or wind impacts is the first principle of onsite erosion prevention. In modern mechanized agriculture, this increasingly is being achieved through the adoption of no-till practices based on the development of machines that can plant through mulched crop residues and on the availability of herbicides to control weeds. The protection of the soil surface afforded by the mulch reduces the erodibility of the soil while arresting the decline in soil organic matter caused by conventional plowing and reducing the use of

TABLE 1 Some Indigenous Technologies for Soil Conservation and Erosion Control[15,16]

Country	Technology	Description
China	Ditches along hillsides	Grassed contour ditches between rows of fruit trees
Ethiopia	Stone bunds	Stones collected and placed in lines on contours
India	Kana bandi	Windbreaks of dead wood
Kenya	Fanya-juu terraces	Soil thrown uphill to form contour bunds
Laos	Lao Theung systems	Sustainable shifting cultivation on mid-altitude steeplands
Nepal	Terraces	Manually built on steep slopes; terrace risers sliced regularly

TABLE 2 Technologies for Soil Erosion Control in Agricultural Landscapes

Vegetative	Mechanical
Cultivated crop land	Cultivated crop, forest, and grazing land
• Mulching	• Conservation tillage
• Crop management	Contour tillage
Dense planting	Ridging and ridge tying
Multiple cropping/agroforestry (contour strips or rotations)	Minimum till and no till
Cover cropping	• Terracing
Forest land	• Waterways
• Tree planting	• Structures
• Agri-silviculture	
Grazing land	
• Plant grasses	
• Plant shrubs	
Wastelan	
• Plant trees, shrubs, grasses	
• Biological engineering Wattling and staking Brush matting	
• Hydroseeding	
Riparian zones	
• Plant trees and shrubs	

Source: Adapted from El-Swaify et al.[17]

fossil fuels for tillage. The adoption of no-till practices in the major grain-producing and -exporting countries is gathering pace[18]; in major South American food-producing countries, average adoption ranges from 38% to 67%, whereas adoption in Canada, the United States, and Australia is 26%, 14%, and 18%, respectively. Future expansion in the adoption of no-till practices has enormous potential to reduce on- and off-site impacts of soil erosion in large-scale agricultural production areas.

As shown in Table 2, traditional soil conservation measures in areas with mechanized agriculture were based on a wide array of structures such as contour banks, but recently, this principle has been extended to contour plowing, vegetative strips, and even controlled traffic systems and geospatial technologies such as global positioning systems, remote sensing, and geographic information systems. Utilizing the last-mentioned technologies, precision conservation[19] provides a means to plan conservation on a whole-farm or watershed basis rather than on a singlefield basis. Tools such as erosion hazard maps help land managers to pinpoint hotspots and prioritize conservation measures accordingly. Thus, cultivated fields, vegetation strips, and riparian zones can be managed on a precise site- specific basis to reduce soil erosion and sediment outflows.

As mentioned above, soil conservation presents special and difficult challenges in many developing countries. The steepland areas in the humid tropics are a particular concern because of the high rainfall intensities and steep slopes, which make soil erosion a potent risk. Additionally, the poverty of subsistence farmers and their lack of tenure over the land add to the problems faced by soil conservation agencies. In these areas, agroforestry, based on contour plantings of shrubs, is a promising way to create biological terraces that stabilize the soil and provide level areas in between on which food crops can be planted.[20] The inclusion of fruit trees, or other perennial cash trees and shrubs, in the hedgerows provides a source of extra income and, at the same time, helps to stabilize the hillside.

From a broader whole-farm or watershed perspective, modern approaches to soil and water conservation can be seen to encompass concepts such as the multifunctionality of agricultural land. This concept promoted in Organisation for Economic Co-operation and Development (OECD)[21] countries entails food security, economic and social functions of land, and environmental functions. Although the multifunctionality concept may be (mis)used in some developed countries to justify farm subsidies and hence play a role in trade negotiations,[22] the application of this holistic concept has been used to analyze and illustrate the benefits of integrated farming practices in countries such as Thailand.[23] In developing countries, the various farm enterprises such as cropping, aquaculture, livestock production, forestry, and fruit production have considerable potential for synergies that increase productivity and profitability while protecting soil and enhancing ecosystem services.

Attaining Widespread Soil Erosion Control

In the 20th century, concern about the on- and off-site impacts of soil erosion prompted governments, mainly in developed countries, to establish soil conservation services. The traditional approach to these services in areas of mechanized agriculture was for a government agency to provide trained soil conservationists to advise and help farmers plan contour banks, strip cropping, minimum and zero tillage, and other measures to reduce erosion in their fields. While these efforts have not been successful in many countries, in the United States, despite massive investments in soil conservation since the 1930s, some experts believe that up until recently, the rates of soil erosion have not declined.[24] To replace the old top–down approach to soil conservation extension, many practitioners now advocate community-based approaches.

The usual starting point for the introduction of soil conservation measures is the farm because individual land managers are directly responsible for managing their soil and water resources in the context of the needs of the watershed and community as well as their own needs. The following examples from China and Uganda show that farmer commitment and soil conservation policy implementation can yield success in defined regions within a country. On the Loess Plateau, which had a history of massive erosion during times of war, and even the cultural revolution, the introduction of the farmer responsibility system and the adoption of watershed-based comprehensive soil conservation policies by the Chinese government have significantly reduced erosion and the flows of sediment into the Yellow River.[25] The support of farmers in the Kigezi district, or at least an absence of resistance from them, has proven to be the key to the successful implementation of soil conservation policies introduced in colonial Uganda.[26] Farmers understood that their livelihoods depended on their conservation of the soil.

In developed countries, successful movements such as the landcare program in Australia[27] and conservation districts in the United States[28] are based on a bottom–up approach in which members of rural communities work together to conserve not only the soil and water resources but also, on the basis of a more holistic watershed approach, the biodiversity and landscape integrity. These movements encourage principles of stewardship of natural resources while sustaining the productivity of the farms. The success of the landcare approach—Australia has 6000 landcare and coastcare groups—is built on a bottom–up organization based on networking through boards and committees and leaders who are passionate about conservation. Governments at the local, state, province, or national level provide support through coordination at the broader catchment and regional levels, as well as through developing long-term strategies, identifying knowledge gaps, and supporting research and development. The ultimate scale is the river basin in which the cumulative effects of soil and water management on runoff, nutrient, and sediment flows combine with riparian management impacts to determine the nature and extent of human influences on estuarine and coastal zones. In the case of large river basins, national and international levels of governance may be involved.

While it is difficult to precisely pinpoint the drivers of the shift in thinking towards a broad-scale, community- based approach to soil conservation, one of its important dimensions has been the widespread recognition of the importance of the ecosystem services provided by farmers. The associated increase in emphasis on the need to reduce the off-site negative impacts of soil erosion has culminated in the proposition that farmers be paid Green Water Credits (Green Water Credits is a mechanism for transfer of cash to rural people in return for water management activities that determine the supply of all freshwater at source—these activities are presently unrecognised and unrewarded, see Green Water Credits).[29] This approach, and the promise of benefits from carbon credits, may help accelerate the adoption of soil conservation technologies in developing countries, because while the landcare model has been adopted in parts of some developing countries, e.g., the Philippines, a comprehensive and expensive national-scale approach such as that in Australia and the United States has not been feasible in many parts of the developing world. Ultimately, a broad view of sustainable land management is needed, and this can be achieved at a range of scales provided it entails a community-based approach with farmer participation; these are the keys to success in both developed and developing countries, although the most appropriate frameworks for community, institution, and government involvement vary from place to place. Nevertheless, progress is being made, as indicated by a comprehensive review of 286 "bright spots" in 57 countries.[30] The bright spots were defined as examples of intervention—where poor farmers have increased their productivity and income—having a positive impact on land and water resources. Factors identified as keys to success were investment in appropriate land and water management technologies and human factors such as the aspirations of the leaders and community to improve their circumstances and entrepreneurship. Applying such lessons to facilitate the widespread adoption of soil conservation in every country with an agricultural base presents a major future challenge to land managers, technologists, and policy makers.

References

1. Oldeman, L.R.; Hakkeling, R.T.A.; Sombroek, W.G. *World Map of the Status of Human-Induced Soil Degradation: An Explanatory Note*; International Soil Reference and Information Centre: Wageningen, 1991; 1–34.
2. Montgomery, D.R. Is agriculture eroding civilization's foundation? Geol. Soc. Am. Today **2007**, *17* (10), 4–9.
3. Pimental, D.; Pimental, M. Global environmental resources versus world population growth. Ecol. Econ. **2006**, *59,* 195–198.
4. Syvitski, J.P.M.; Vorosmarty, C.J.; Kettner, A.J.; Green, P. Impact of humans on the flux of terrestrial sediment to the global coastal ocean. Science **2005**, *308,* 376–380.
5. International Erosion Control Association. Knowledge Base, available at http://www.ieca.org/ (accessed January 2011).
6. Poesen, J.; Nachtergaele, G.; Verstraeten, G.; Valentin, C. Gully erosion and environmental change: Importance and research needs. Catena **2003**, *50*, 91–133.

7. Fox, G.A.; Wilson, G.V. The role of subsurface flow in hillslope and stream bank erosion: A review. Soil Sci. Soc. Am. J. **2009**, *74,* 717–733.
8. Wang, X.; Oenema, O.; Hoogmoed, W.B.; Perdok, U.D.; Cai, D. Dust storm erosion and its impact on soil carbon and nitrogen losses in northern China. Catena **2006**, *66,* 221–227.
9. UNCCD. The causes of desertification. Available at http://www.unccd.int/publicinfo/factsheets/showFS.php?number=2 (accessed September 2010).
10. RUSLE. RUSLE factors. Available at http://www.iwr.msu.edu/rusle/factors.htm (accessed January 2011).
11. Yu, B.; Sajjapongse, A.; Yin, D.; Eusof, Z.; Anecksamphan, C.; Rose, C.W.; Cakurs, U. Application of a physically based soil erosion model, GUEST, in the absence of data on runoff rates II. Four case studies from China, Malaysia, and Thailand. Aust. J. Soil Res. **1999**, *37,* 13–31.
12. Rais, M.; Craswell, E.T.; Gameda, S.; Sajjapongse, A.; Bechstedt, H.-D. Decision support system for sustainable land management: A Southeast Asia case. Decis. Support Syst. Sustainable Dev. **2002**, *II*, 183–196.
13. den Biggelaar, C.; Lal, R.; Wiebe, K.; Breneman, V. The global impact of soil erosion on productivity I: Absolute and relative erosion-induced yield losses. Adv. Agron. **2004**, *81,* 1–48.
14. den Biggelaar, C.; Lal, R.; Wiebe, K.; Eswaran, H.; Breneman, V.; Reich, P. The global impact of soil erosion on productivity II: Effects on crop yields and production over time. Adv. Agron. **2004**, *81,* 49–95.
15. Pongsapich, A. *Indigenous Technical Knowledge for Land Management in Asia;* International Board for Soil Research and Management: Bangkok, 1998, Issues in Sustainable Land Management No. 3.
16. WOCAT (World Overview of Conservation Approaches and Technologies). Database on SLM Technologies. Available at http://www.wocat.org/ (accessed January 2011).
17. El-Swaify, S.A.; Dangler, E.W.; Armstrong, C.L. *Soil Erosion by Water in the Tropics*; College of Tropical Agriculture and Human Resources, University of Hawaii: Honolulu, 1982; 1–173.
18. Fischer, R.A.; Byerlee, D.; Edmeades, G.O. Can technology deliver on the yield challenge to 2050? Paper presented at the Expert Meeting on How to Feed the World in 2050, Rome, Italy, 24–26, June 2009. Food and Agriculture Organization of the United Nations. Rome, Italy. Available at ftp://ftp.fao.org/docrep/fao/012/ak977e/ak977e00.pdf (accessed January 2011).
19. Delgado, J.A.; Joseph, K.; Berry, J.K. Advances in precision conservation. Adv. Agron. **2008**, *98,* 1–48.
20. Craswell, E.T.; Sajjapongse, A.; Howlett, D.J.B.; Dowling, A.J. Agroforestry in the management of sloping lands in Asia and the Pacific. Agroforestry Syst., **1997**, *38,* 121–137.
21. OECD. *Multifunctionality: Towards an Analytical Framework*; Organization for Economic Cooperation and Development: Paris Cedex, France, 2001.
22. Potter, C.; Burney, J. Agricultural multifunctionality in the WTO—Legitimate non-trade concern or disguised protectionism? J. Rural Stud. **2002**, *18* (1), 35–47.
23. Tipraqsa, P.; Craswell, E.T.; Noble, A.D.; SchmidtVogt, D. Resource integration for multiple benefits: Multifunctionality of integrated farming systems in Northeast Thailand. Agric. Syst. **2007**, *94,* 694–703.
24. Trimble, S.W.; Crosson, P. U.S. soil erosion rates—Myth and REALITY. Science **2000**, *289* (5477), 248–250.
25. Wei, J.; Zhou, J.; Tian, J.L.; He, X.B.; Tang, K.L. Decoupling soil erosion and human activities on the Chinese Loess Plateau in the 20th century. Catena, **2006**, *68,* 10–15.
26. Carswell, G. Soil conservation policies in colonial Kigezi, Uganda: Successful implementation and an absence of resistance. In *Social History and African Environments;* Beinart, W., McGregoreds, J., Eds.; Heineman and James Curry: Oxford, 2003; 131–154.
27. Landcare Australia. Available at http://www.landcareon-line.com.au/ (accessed January 2011).
28. National Association of Conservation Districts. Available at http://www.nacdnet.org/ (accessed January 2011).

29. Green water credits. Available at http://www.isric.org/UK/About+ISRIC/Projects/Current+Projects/Green+Water+Credits.htm (accessed January 2011).
30. Noble, A.D.; Bossio, D.A.; Penning de Vries, F.W.T.; Pretty, J.; Thiyagarajan, T.M. *Intensifying Agricultural Sustainability: An Analysis of Impacts and Drivers in the Development of 'Bright Spots',* Comprehensive Assessment of Water Management in Agriculture Research Report 13; International Water Management Institute: Colombo, 2006.

45

Farming: Organic Pest Management

Ján Gallo

Introduction

Organic farming in its essence is defined as a system of efficient production that is expected to create an environmentally beneficial, integrated, human, and economically stable agricultural system. Organic farming changes the philosophy of human consideration of nature from anthropocentric and technocratic arrogance to the gradual return of a holistic view in which humans are a part in the sense of the old trueism: "*We have not received soil, water and landscape from our parents, but we have borrowed them from our children.*" According to this philosophy, nature is the uniform whole with its internal natural value and from this point of view ecological, economic, and social aspects of agricultural production are derived. The organic system of farming is based on a maximally closed flow of energy without any additional outside inputs; therefore, it is a sustainable system that uses local and renewable resources. Control of harmful biotic (diseases, pests, and weeds) and abiotic factors is an integral part of the system.[1,2]

Strategy of Control against Animal Pests

The strategy of controlling harmful animal organisms in the organic farming system is based first of all on natural factors of resistance. It includes mainly internal factors (genetic predisposition of organisms) as well as the external ones that act inside the ecosystem. External ecosystem factors are of an abiotic (pedo-hydrometeo factors) and biological character. The biotic resistance of the environment has been applied prevailingly against animal pests using mainly the factor of natural mortality.[2,3]

Basic strategic objectives for pest regulation can be summarized as follows:

A species that seems to be a potential pest can survive under certain acceptable abundance. This means that the population density of an individual species is regulated and those measures that are directed to species eradication at all costs are not realized. The term "pest" is relative, and for this reason, animals in agricultural crops within organic farming cannot be subdivided to harmful, indifferent, and useful.

The unit of regulation is the ecosystem; and for this reason, there is a need to contribute to maintaining its autoregulation capability in maximum range.

Great emphasis is placed upon knowledge of natural antagonists (parasitoids, predators, and pathogens). Miscellaneous species must be maintained within the landscape.

Measures against pests have to be in accordance with economic aspects.

Organic farming does not represent the simple substitution of synthetic chemical pesticides by biological ones, it is a system of complex measures that promotes the good health of plants and not the total annihilation of pathogens.[2–4]

Regulation of Animal Pests

Animal pests live in all field crops, but within each individual crop they require different conditions for their development. Under our conditions, 69% of field crop pests are staying on arable soil during the year. This linkage to arable soil is manifested in the population dynamics of individual species.

This strategy of control against animal pests requires that the regulation of animal pest occurrences is realized mainly through the utilization of

1. the phenomenon of biotic resistance of the environment and
2. anthropogenic measures

Phenomenon of Biotic (Natural) Resistance of the Environment

Natural mortality (which is caused first of all by the activity of predators, parasitoids, and parasites) is a significant natural regulator (bioregulator) of plant pest populations in agriculture. This regulation is manifested in the long-term maintenance of pest population density below the level of harmfulness. Bioregulators in the organic farming system belong mainly to the group of spiders (Araneida), predacious mites (Acarina), and parasitoidic insects (Insecta) predominantly within the order of Hymenoptera and Diptera. Another group of bioregulators can be found within the group of phytopathogenic microorganisms (for instance fungi, bacteria, and others). Spiders (Araneida) occur in all biotopes on soil surfaces and on the above-ground parts of plants and can effectively reduce the population of harmful insects, mainly aphids (Aphidoidea), in field crops. Predacious mites (Acarina) belong to edaphic fauna and within stands of field crops are effective natural regulators of pests that live in the soil. Predacious polyphagous bugs, (Heteroptera) for example, within the Anthocoridae family feed on the larvae of thrips (Thysanoptera), aphids (Aphidoidea), and mites (Acarina). The golden-eyed fly (Chrysopa perla) is able to destroy 200 to 500 aphids during its larval stage, but it can also feed on Diaspididae, Tetranychidae, and the larvae of flies. Parasitoidic Hymenoptera, predominantly the species that belongs to the families of Ichneumonidae, Braconidae, and Chalcidoidae, in the larval stage feed on the different stages of insect, starting with eggs and finishing with imagos. Imagos of these parasites as well as their larvae frequently injure the body of the host, feed on its body fluid, and finally cause the death of the host. For example, *Apanteles, Angitia,* and *Microgaster* are some of the main regulators of white butterflies (Pieridae). *Tersilochus heterocerus* is an important parasitoid of rape beetle (*Meligethes aeneus*). Beetles of the Carabidae family represent a significant group of predacious insects (for instance, *Calosoma sycophanta, C. inquisitor*) in field crops. They feed on aphids (Aphidoidea), larvae of beetles (Coleoptera), caterpillars of butterflies (Lepidoptera), and larvae of Diptera that they chase on the soil surface. Beetles *(Coleoptera)* within the Staphilinidae family stay on the surface of soil and feed on aphids, eggs, larvae, and cocoons of insects, partially also on fungi (*Aleochara bilineata* significantly regulates *Delia brassicae*). Another very important group of predacious insects in field crops is ladybugs (Coccinellidae), the larvae and imagos of which feed on aphids (Aphidoidea) and help to reduce their numbers (an imago consumes 40 to 60 aphids per day). Syrphidae live in all cultural crops and their larvae feed on aphids and belong to the effective factors through which an occurrence of aphids is regulated in cereals, sugar beet, potatoes, and legumes (a larva consumes about 100 aphids per day).[3,5]

It is necessary to protect the above-mentioned components of natural resistance in the environment and create optimum conditions for their existence. However, in many cases, the overpopulation of pests cannot be regulated only by means of their natural enemies (bioregulators).

Anthropogenic Measures

In the cases when biotic resistance of environment is less effective and there is a danger of substantial reduction in yields or deterioration of bioproduct quality, anthropogenic measures can be applied. Anthropogenic measures fulfill two essential roles:

1. To prevent pests from attacking healthy plants
2. To improve the health of attacked plants

These are mainly indirect (preventive or prophylactic) methods for pest regulation. Direct (repressive or therapeutic) methods can be used only in the most necessary cases, and only those that are in accordance with International Federation of Organic Agriculture Movements (IFOAM) rules. Indirect measures include crop rotation, selection of proper species and cultivars, appropriate soil cultivation, suitable fertilization, green fertilizing, and outside and inside quarantine. Crop rotation seems to be the dominant sustainable procedure on the farm that protects, renews, and supports biodiversity and also plays a key role in soil fertility preservation. By means of the above-mentioned indirect principles, overpopulation of pests can be restricted completely, or at least very substantially. *Zabrus gibus* damages winter wheat five to eight times less when winter wheat follows after sunflowers or maize within crop rotation. If cereal crops are grown several years in the same place, they are frequently damaged by *Scotia segetum*, nematodes (*Heterodera*), and frit fly (*Oscinella frit*). Within organic farming, the portion of cereals in crop rotation would not exceed 50%. Early plowing of stubble can eliminate 70% to 80% of thrips (*Haplothrips tritici*). On the other side, late sowings of peas are more likely to be attacked by pea aphids (*Acyrthosiphon onobrychys*) and weevils (*Apion* spp.). Harvesting maize to the shortest possible stubble (0.08 to 0.10 m) substantially decreases the number of overwintering carterpillars of European corn borer (*Ostrinia nubilalis*). If the harvest is done early enough it enables about 80% of pests to be removed together with plant material. Occurrence of bean weevils (*Acanthoscelides obsoletus*) on bean plants can be effectively reduced if rows of beans are alternated with rows of maize, which by its "odor" causes this pest to become disoriented as it is flying toward bean plants. Fertilization within organic farming is based predominantly on own farm organic fertilizer utilization, however, its application must follow the rules of IFOAM.[3,5]

For the future, solutions are needed for problems related to bioproducts from the raw materials grown within organic farming and that have received a certificate according to the rules of IFOAM. We have to face the fact that growing plants are exposed to biotic and abiotic stresses and in some cases they react by forming "natural pesticides" that can be dangerous for humans (e.g., carcinogenic psoralens in celery, parsley, and parsnip). While the above-mentioned and similar matters have to be solved, the medicinal advantages of bioproducts resulting from organic farming must not be doubted.[1,5]

In addition to the principles of sustainable development, appropriate laws should be created including legislative treatment of organic farming and biofoods production for each respective country (224/1998 Z.z in the Slovak Republic). By means of the law a new historical paradigm will be defined, within which nature will become a globally respected value.

References

1. Klinda, J. *Agenda 21 a ukazovatele trvalo udrzatel'ného rozvoja (Agenda 21 and Indicators of Sustainable Development)*; MZP SR: Bratislava, 1996; 1–517, ISBN 80-88833-03-5.
2. Gallo, J.; Sedivy, J. *Integrovanáochrana rastlín (Integrated Pest Management)*; VES VSP: Nitra, 1992; 1-163ISBN 80-73137-061-4.

3. Neuerburg, W.; Padel, S. *Organisch-biologischer Landbau in der Praxis;* BRD: Munchen, 1992; 1–460.
4. Lampkin, N. *Organic Farming;* Farming Press Books: Ipswich, United Kingdom, 1990; 214–271, ISBN 0-85236-191-2.
5. Petr, J.; Dlouhy, J. *Ekologické zemêdêlství (Ecological Agriculture);* Zemédélství nakladatelství Brázda: Praha; 1992; 1–305, ISBN 80-209-0233-3.

46

Integrated Nutrient Management

Bal Ram Singh

Introduction

The basic concept underlying integrated nutrient management (INM) is the maintenance or adjustment of soil fertility/productivity and of optimal plant nutrient supply for sustaining the desired level of crop productivity.[1] The objective is to accomplish this through optimization of the benefits from all possible sources of plant nutrients, including locally available ones, in an integrated manner while ensuring environmental quality. This provides a system of crop nutrition in which plant nutrient needs are met through a preplanned integrated use of mineral fertilizers, organic manures/fertilizers (e.g., green manures, recyclable wastes, crop residues, and FYM), and biofertilizers.[1] However, to safeguard the environment from degradation and to maintain clean and safe water, air, and food, the use of chemical fertilizers must be based on sound scientific and ecological lines.[2] The use of nutrient sources and their combination can vary greatly depending on the landuse system and on ecological, social, and economic conditions at the local level.[1]

Nutrient mining (implying more extraction of nutrient than returned back to the soil) in many developing countries and excessive fertilization (more fertilizer application than required for maximum yield) in developed countries are major causes for soil fertility decline and environmental degradation. These situations have created the need for the use of INM practices. In low- and medium-input agriculture such as in Africa and in intensive cropping systems with high cropping intensity of high yielding crops in many Asian countries, nutrient balance is increasingly becoming negative. For example, in Bangladesh, high-yielding crops remove considerable quantities of nutrient leading to annual depletion of nutrients (NPK) to as much as 250 kg ha^{-1} yr^{-1}. Besides the depletion of NPK, S and Zn deficiency areas have been identified in more than 4 and 2 million ha, respectively.[3]

Excessive fertilization is wasteful in terms of fertilizer, and it not only creates antagonistic effects that disturb the nutrient balance and thus leads to reduced crop yield and produce quality but also produces suboptimal economic returns. Although excessive fertilization, for example, with P could lead to significant residual benefits, it blocks the farmer's capital unnecessarily on one hand and can also lead to higher P losses to water resources in erosion-prone areas on the other hand. Excessive fertilization with N often results in poor N use efficiency, greater possibility of lodging and pest and disease attack, and greater N losses leading to negative impacts on the environment.

INM identifies the best combination of various types of plant nutrients in different fields for a balanced plant nutrient and plant yield, while sustaining soil fertility and controlling nutrient losses. Therefore, the cropping system (rather than a single crop) and the farming system (rather than an individual field) are the focus of INM approach for developing the practices for the main agro-ecological region of a country and for the various categories of farms.[1] It is envisaged that locally available materials of plant or animal origin as by-products of agricultural activities be used or, where such materials are not available, in situ production of organics, such as fast-growing leguminous shrubs or green manure crops, blue green algae (BGA), and phosphate-solubilizing bacteria (PSB), be practiced. Experiments conducted at several locations in north India have shown that green manure crops of sesbania (*Sesbania aculeata*) or sunnhemp (*Crotalaria juncea*) along with recommended fertilizers resulted in increased productivity of the system by 8%–20%. Furthermore, large biomass of narrow C:N ratio incorporated into the soil contributed to a sizable amount of plant nutrients, especially N, and improved soil physical properties.[2] In contrast to organic farming, INM involves a needs-based external input approach, taking into account a holistic view of soil fertility.

For a successful INM, we have to have judicious use of mineral fertilizers, maximum use of organic materials, and minimum negative environmental impacts. To achieve this, we need to include several steps, such as 1) availability of on-farm and off-farm resources; 2) yield target based on resource availability; 3) nutrient requirement to achieve desired yield; 4) integration of mineral and organic fertilizers; 5) time and method of fertilizer application; 6) efficient soil and water conservation measures; and 7) sustaining physical, chemical, and biological fertility of soils.[4]

This entry presents the integrated management of plant nutrients coming from mineral fertilizers and biofertilizers to increase crop productivity, nutrient use efficiency, and sequestration of soil carbon and other nutrients in soils and to avoid environmental degradation. Examples are cited from different soil and environmental conditions to illustrate the significance of INM for sustainable use of nutrient resources and soil carbon sequestration and to safeguard the environment for future generations.

Components of INM Systems

The main components of an INM system are shown in Figure 1. The figure shows a judicious integration of mineral and synthetic fertilizers, suitable minerals [e.g., phosphate rock (PR), elemental sulfur], crop residues, green manures, organic manures of plant or animal origin (e.g., compost, farm yard manure, animal slurry) and recyclable organic wastes from industry and households, and microbial inoculants (e.g., biofertilizers). At the farm level, INM aims to optimize the productivity of the nutrient flows through the soil/crop/livestock system during a crop rotation. Furthermore, INM improves the production capacity of a farm through the application of external plant nutrient sources and amendments and the efficient processing and recycling of crop residues and on-farm organic wastes. The nutrient cycling in soil–plant–air–water systems is shown in a schematic diagram (Figure 2).

The INM involves a more rational use of plant nutrients by balanced, efficient, yield-targeted, site- and soil- specific nutrient supply and a combination of mineral and organic fertilizers on a cropping-system or crop-rotation basis. It puts emphasis on monitoring and controlling the unwanted side effects of fertilization and the possible adverse consequences for soil health, crop diseases, and pollution of water and air. The main aim of INM is to obtain high yields and good product quality with

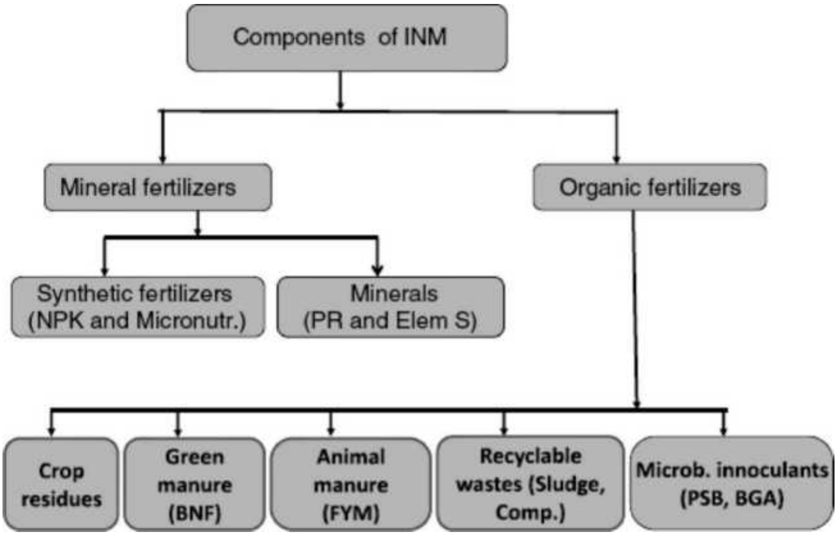


FIGURE 1 Components of integrated nutrient management system.

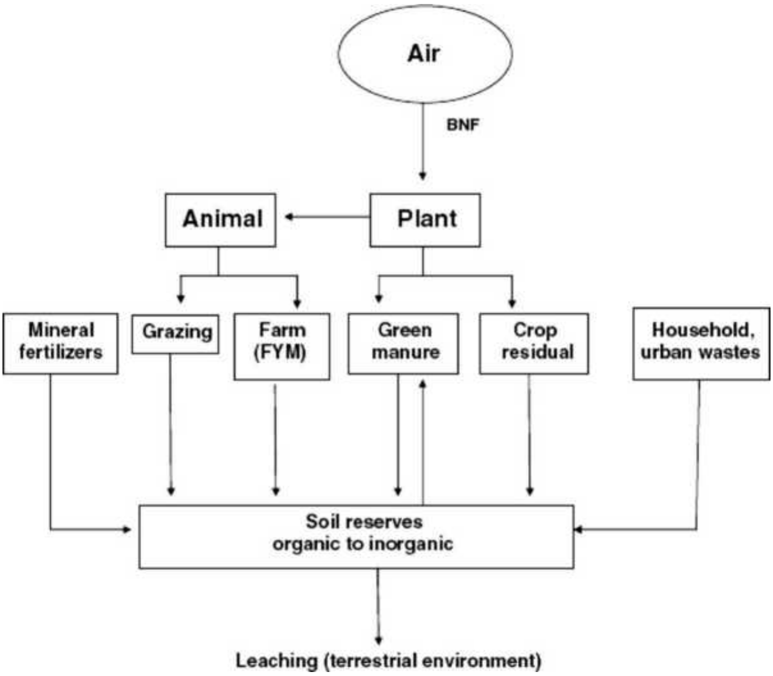


FIGURE 2 Nutrient cycling in soil–plant–air–water systems. BNF, biological N fixation; FYM, farmyard manure.

minimum damage to the environment. It also provides opportunities for saving resources, protecting the environment, and promoting more economical cropping for the farmers. It also promotes changes in land use, crop rotations, and interactions between forestry, livestock, and cropping systems as part of agricultural intensification and diversification.[1] Different components are described briefly below.

Mineral and Synthetic Fertilizers

Fertilizer use in the past has accounted for about 50% of the annual global food harvest and thus has played a very important role in increasing the food supply for the evergrowing population in the world. Improvements in agricultural production in many parts of the world, especially in Asia, have primarily come from intensive use of chemical fertilizers. A good example of the role played by fertilizers in increasing food production by many fold can be seen from India. During the last five decades, grain production in India has gone hand in hand with fertilizer consumption (Figure 3). The grain production increased from 52 million megagrams (M Mg) in 1950/1951 to 231 M Mg in 2007/2008 while the fertilizer (NPK) consumption during the same period increased from 0.07 to 24.7 M Mg.[5] It needs, however, to be emphasized here that along with fertilizer consumption, use of high-yielding varieties, increased irrigation facilities, and enhanced use of pesticides have also played a key role in this spectacular rise in grain production in India. At the same time, the quality of soil, air, and water has deteriorated and the area of natural ecosystems has decreased because of the intensification and extension of agricultural production.[6,7] On the contrary, the increase in food production in Africa and the Cerrito of Brazil is caused by the expansion of the cultivated area.

Worldwide mineral fertilizer nutrient use is expected to increase from 142 M Mg in 2002/2003 to 165 M Mg in 2009/2010, to 175 M Mg in 2015, to 199 M Mg in 2030.[8] However, the demand differs considerably among the regions (Table 1). The largest share of mineral fertilizers will be used by East Asia, followed by South Asia. These two regions together will account for about half of world mineral fertilizer use by 2030. The average application rates of mineral fertilizers in different development regions

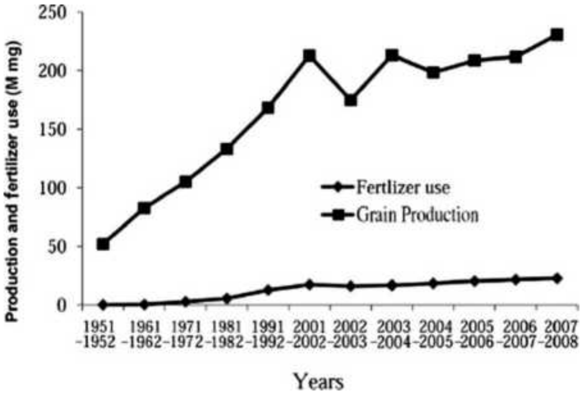


FIGURE 3 Fertilizer use and grain production scenario in India.

TABLE 1 Mineral Fertilizer Use in Developing Regions

Region	Total Nutrients ($N+P_2O_5+K_2O$) 2002/2003 (M Mg)	2009/2010 (M Mg)
Sub-Saharan Africa	2.3	
Near East and North Africa	7.9	
East Asia	50.6	59.5
South Asia	20.9	25.7
Latin America	13.2	18.3
World	141.6	165.0

Source: Food and Agriculture Organization.[8]

of the world have also increased greatly while the rate applied in Sub-Saharan African (SSA) countries have continued to be very low. This has resulted in continuous decrease in per capita food production in Africa in contrast to sustained increase in Asia and Latin America.

Mineral fertilizers normally supply three major nutrients (N, P, and K), but since the last few decades, incorporation of micronutrients in mineral fertilizers has also become very common in many regions of the world because of increasing cases of their deficiencies in food and feed crops and added focus on human and animal health. A brief description of these fertilizers is presented below.

Nitrogen

The nitrogen requirement of 1 Mg of cereal grains is 15–30 kg N,[9] and globally, 85 M Mg of N through fertilizers is used,[8] which makes nearly 60% of all fertilizers used. Even in the future, N will be the most needed nutrient because of increasing food demands. Losses of N can range from 30% to 50% of mineralized N. If not taken by plants, it can be leached to underground layers because it is not retained by soils. Leaching losses coupled with N loss to the atmosphere through denitrification and ammonium volatilization lead to low N use efficiency of fertilizers. The average recovery of N by crops depends on crops and fertilizer used and the soil and management conditions. Nutrient recovery from different fertilizers is shown in Table 2. The table shows that N recovery varied from 20% to 70% depending upon the N source. In a long-term experiment (22 years) in India, it was found that N recovery was 17% in maize and 32% in wheat with application of N alone, but the recovery was almost doubled (33% in maize and 65% in wheat) when P and K along with N were applied.[10] The later treatment also resulted in buildup of organic C in the soil. Singh and Swarup[11] also made similar observations based on the results from several long-term experiments from different parts of India. They also found that continuous use of N alone resulted in the greatest decline in yield at all centers, indicating that other major nutrients and micronutrients were becoming limiting factors, and adequate response to N could not be obtained unless the yield-limiting factors were remedied.

Nitrogen balance is very important in the INM as it is an excellent tool in choosing appropriate nitrogen sources. The N cycle, considering natural to anthropogenic fluxes, is currently under the influence of anthropogenic control in many parts of the world. Nitrogen pool and flux studies should adopt a comprehensive and consistent treatment of the reservoirs and cycling of terrestrial N, so as to improve the treatment of N cycling in models and mass balance studies. This will facilitate better scientific explanation of historical N-related environmental changes and more closely balance N budgets on a range of geographical and temporal scale as the N balance surplus is an important indicator for the evaluation of the sustainability of agriculture.[12]

Although biomonitoring techniques are usually simple and relatively inexpensive on an individual basis, the natural variability of the environment may require large numbers of samples to be taken to meet the required statistical precision. Consequently, biomonitoring should not necessarily be seen as a low-cost approach to compliance assessment. The requirement for a statistically sound method places stringent constraints on the sampling design. Total tissue nitrogen (N) has been measured for many

TABLE 2 Average Utilization Rate of Fertilizer Nutrients by the First Crop

Nutrient and Source	Utilization Rate (% recovery)
Nitrogen, mineral	50–70
Nitrogen slurry	30–50
Nitrogen manure	20–40
Phosphorus, mineral	10–20
Potassium, mineral	50–60
Micronutrient mineral	0.5–5

Source: Roy et al.[1]

years in all types of plant tissue to assess atmospheric nitrogen de- position.[13] This method requires both field sampling and laboratory analyses. On a limited scale, the method has also been used to assess the N need of plants during their growth to work out for overdressing of N fertilizers as well as to determine the N requirements of plants.

Phosphorus

The dynamics of phosphate in soil presents special problems because of the low solubility of most P compounds. Phosphorus added through soluble fertilizers first enters the soil solution, but much of it is converted into adsorbed P within a few hours. Furthermore, organic manure is not able to meet crop requirement and hence P fertilizer is necessary to overcome P depletion. As shown in Table 2, the recovery of P is much lower than that of N, being only 10%–20%, but P, unlike N, gets accumulated into its reserves in soil as it is not lost from the soil except through soil erosion. In the high P-sorbing soils of Cerrado, Brazil, large application of P can replenish P and the residual effects of P can last for 5 to 10 years.[14] These corrective applications, along with subsequent maintenance application and sound agronomic practices, have proved very successful in this region.[14]

Reactive PRs can be used in acid soils and for long- duration crops. Their suitability depends on the reactivity of the rock, its particle size, soil pH, and type of crop. In several African countries, deposits of high to medium reactive (>15 g citrate-soluble P kg^{-1}) PRs have been found, which can be used without any pretreatment. The effectiveness of PRs not only is related to the reactive “available” portion but also depends on the P-mobilization capacity of the soil, which is related to pH, moisture status, and biological activity. [15] On-farm trials in P-deficient soils in western Kenya demonstrate that Minjingu PR was found to be as effective as triple superphosphate (TSP, 20% P) at equal P rates. The benefits are most pronounced with the integration of agroforestry technologies that improve soil fertility. Besides Minjingu PR, Busumbu PR from Uganda (BPR) is potentially another source of P.[16] Sharma et al.[17] calculated the relative agronomic effectiveness (RAE), defined as follows:

$$\text{RAE}\ (\%)\ = \frac{\text{Yield with PR control}}{\text{Yield with DAP control}} \times 100 \tag{1}$$

They[17] found that the RAE of Mussoorie rock phosphate (MRP) + PSB in relation to diammonium phosphate (DAP) as judged by the total productivity was 53%–65% in the first cycle but reached 69%–106% in the third cycle of the cropping system. The P balance (application—crop removal) was generally more positive for MRP + PSB than for DAP, and the highest P balance was recorded with an application of 52.5 kg P ha^{-1} as MRP + PSB, and it resulted in the highest 0.5 M NaHCO3 extractable P content in soil.

Potassium

Potassium (K) dynamics are determined by the rate of K exchange from the clay and organic matter surfaces and the rate of release from soil minerals. Potassium is generally stored on the surface of negatively charged clay minerals. This easily replaceable supply provides the soil solution with additional K when the soil-solution K concentration decreases as a result of crop uptake.[1] Furthermore, most K is retained in straw, and hence, if crop residue is returned to the soil, the bulk of the depleted K can be replenished. However, soils such as oxisols and ultisols are too deficient in reserve K and profitable yields cannot be obtained without K application.

The integrated use of FYM and inorganic fertilizer maintained a high amount of plant-available ammonium acetate extractable K, as compared to GM and inorganic fertilizer alone. It was observed that, under intensive cropping, application of the recommended dose of NPK along with FYM reduced the mining of K from the soil. The regression analysis of the data showed that water-soluble, exchangeable, and non-exchangeable K together accounted for 36%, 74%, and 99% variability in grain yield, crop K uptake, and available K in soil, respectively.[18]

Micronutrients

Widespread deficiencies of micronutrients have been found in many parts of the world, and in the last decades, their importance is increased both for increased crop productivity and for their effects on animal and human health. About 3 billion people in the world are affected with micronutrient malnutrition. Similar to major nutrients, micronutrients are present in several forms and fractions, and their soil concentration is very low. However, most micronutrients are in a dynamic equilibrium with their exchangeable and adsorbed forms, which sustain the soil solution and replenish it continuously as the nutrients are taken up by plants roots. Gupta[19] reported that the majority of Indian soils are deficient in zinc, followed by iron, copper, and manganese. Application of micronutrient fertilizers can mitigate the deficiency of these nutrients and help exploit the potential of crops. INM is another option for meeting the deficiency of micronutrients in various cropping systems.

Biofertilizers

Continuous application of chemical fertilizers alone has been reported to deteriorate soil health, and the application of organic manures alone does not produce required yields due to their low nutrient status. Hence, a judicious combination of mineral fertilizers with organic and biological sources of nutrients is being promoted. Such integrated applications are not only complementary but also synergistic as organic inputs have beneficial effects beyond their nutrient content.[1,20] Since nutrients are slowly released from biofertilizers, their combined use with mineral fertilizers (fast nutrient release) in INM system is capable of synchronizing nutrient supply with nutrient needs throughout crop growth, thereby maximizing nutrient use and minimizing nutrient loss and environmental pollution. Currently, farmers make limited use of this concept, primarily because there are no routine techniques available to rapidly assess the nutrient availability from biofertilizers and then adjust the mineral fertilizer rates accordingly. Among biofertilizers, biologically fixed nitrogen, crop residue, organic manures and composts, and sewage sludge are commonly used, but the use of microbial inoculants, artificially multiplied cultures of certain soil organisms, is also spreading, though on a limited scale because of limited supply to farmers.

Biological Nitrogen Fixation, Leguminous Crop Rotations, and Agroforestry

The major sources of biological nitrogen fixation (BNF) are the *legume–Rhizobium* symbiosis and *Cyanophyta* (BGA) and non-symbiotic N fixers *(Azospirillum* and *Azotobacter)*. However, others such as *Azolla* and Actinorrhizae *(Frankia)* make a small but significant contribution to N cycling in natural ecosystems. Estimates of global terrestrial BNF range from 100 to 290 M Mg of N yr^{-1}. Of this total, 40–48 M Mg is estimated to be biologically fixed in agricultural crops and fields.[1] As compared to this, only 87 M Mg of fertilizer N was supplied through mineral fertilizers in 2000, thus emphasizing the contribution of BNF to global N cycle.[1] Green manures of leguminous crops can add substantial amounts of organic matter and N as well as other nutrients. The bulk of the N input through leguminous green manures comes from BNF. Using rice culture as an example, this can range from 50 to 200 kg N ha^{-1}.[1] Agroforestry that includes legume trees provides another venue for BNF benefits, besides recycling of mineral nutrients brought up from below the root zone of annual crops; organic amendments with PR also enhance P availability.[21]

Most of the biological N fixed is present in organic form, and therefore, it must be mineralized before the following crop can utilize it. It thus has the potential to complement the crop nutrient needs with readily available nutrients such as from mineral and synthetic fertilizers in an INM system. Legume green manuring and incorporation of catch crops could also be used towards that purpose although limitation of land and water for this practice and cheaper N fertilizers restrict its widespread use.

Biosolids, Manures, Composts, and By-products (e.g., Crop Residue, Sewage Sludge)

Manures and composts not only provide N, P, and K as well as other nutrients to crops but also affect both physical and biological properties of soil. Average, well-rotted FYM contains 0.5%–1.0% N, 0.15%–0.20% P_2O_5, and 0.5%–0.6% K_2O. The desired C:N ratio in FYM is 15–20:1. Average content of different organic manure is presented in Table 4. In an integrated crop–animal farming system, nutrient contribution from applied manures and composts becomes significant and synchronizes nutrient supply to crops when complemented by mineral and synthetic fertilizers. For example, Mollah et al.[22] found significantly higher average yield (3 years) of Boro and T Aman rice under INM as compared to chemical fertilizer alone, while the nutrient (NPK) supply was generally lower under INM (Table 3). Similarly, the higher wheat yield (though not statistically significant) was obtained in the treatment with 50% N from FYM compost and 50% from urea as to 100% N supplied with urea.[23] Since N release from manure occurs over a number of seasons (20%–40% in the first season and 10%–20% in subsequent seasons), the cumulative effect of manure application is much greater than that assessed in a single season. Furthermore, since N in compost is more stabilized than in manure, the rate of N release from the former would be slower.

Since P contained in rock phosphate is less available than that in superphosphate, its availability is enhanced when it is added in manure and organic wastes and composted together (phospho-compost) before its application to the field.

In low-input farming systems, return of crop residues provides significant quantities of N, P, K, and Ca; in the aboveground crop biomass, the latter two exceed more than 60% of the total nutrient uptake. The low N concentration of straw presents a special problem for its decomposition where the soil contains insufficient available N. In spite of the low concentrations, as much as 125%–250% kg K_2O can be added to the soil by 10 Mg of cereal straw or 5 Mg of oilseed rape straw. This becomes especially critical in some tropical ultisols, oxisols, and alfisols, with low exchangeable Ca and K in these soils. For example, Sanchez and Benites[24] reported that the retention of rice and cow- pea crop residues returned to an ultisol (94% Ca and 89% K) that was accumulated by the crops. Addition of legume residues of *Sesbania*, *Gliricidia*, and *Leucaena* and supplementation with fertilizers are an important component of INM system, especially in low-input farming systems.

Urbanization has necessitated the need for safe and economic disposal of sewage sludge. Application of sewage sludge to soil ensures the recycling of nutrients and organic matter and benefits plant production. The general composition of sewage sludge is 1.1%–2.3% N, 0.8%–2.1% P_2O_5, and 0.5%–1.7% K_2O. Significant ammonia volatilization could occur from surface application of sludge, especially liquid

TABLE 3 Nutrient Supply and Rice Yields under Different Nutrient Management Systems

	Boro Rice		T Aman Rice	
	NPK		NPK	
Treatment	(kg ha–1)	Mg ha–1	(kg ha–1)	Mg ha–1
T2 (chemical fertilizers)	254	5.42a	159	4.56a
T3 (chemical fertilizers + FYM)	219	5.48a	129	4.75b

Values with the same letter are not statistically significant at $P < 0.05$.

Source: Extracted and modified from Mollah.[22]

TABLE 4 Environmental Problems Associated with Fertilizer Use and Possible Solutions

Problem	Causes	Possible Solutions
Groundwater contamination	Leaching of nitrate (most important), chloride, sulfate, etc.	Balanced use of fertilizers; optimal loading rates of organic fertilizers, increased N efficiency; deep placement of N fertilizer and integrated N and water management
Eutrophication	Nutrient loss from soils with erosion, surface runoff, or groundwater discharge	Reduce runoff, grow cover crops, adopt water harvesting and controlled irrigation, control soil erosion
Methemoglobinemia	Consumption of high-nitrate drinking water and food	Reduce leaching losses of N, improve water quality
Acid rain	Nitric acid formation in the air, ammonia volatilization, and sulfur dioxide emissions	Reduce denitrification, adopt proper N application methods to reduce NH3 volatilization, correct high soil pH, increase CEC by organic additions
Ozone depletion and global warming	Nitrous oxide emission from soil through denitrification	Use of nitrification inhibitors and urease inhibitors, increase N-use efficiency, prevent denitrification
Itai-itai disease	Eating rice and drinking water contaminated with Cd	Soil management such as liming or water control in rice fields, reduction in Cd content of PR and chemical fertilizers
Fluorosis in animals	Ingestion of soil or fertilizer treated with high-fluoride PR	Monitor the F content of PR applied directly to acid soils

Source: Extracted and modified from Pathak et al.[25]

sludge and animal slurries. Since sewage sludge contains relatively more P than N required by crops, the application rates based on P requirements of the crops require supplementary N applications. Thus, it would require the INM approach to balance the N and P needs of the crops.

Nitrogen-Fixing Biofertilizers

1. *Rhizobium* (RHZ), known for its ability to fix atmospheric nitrogen in symbiotic association with plants forming nodules in roots. Inoculation of grain legumes such as pulses is associated with a N gain of 20–40 kg N ha^{-1}.
2. *Azotobacter* (AZT), a free-living and non-symbiotic nitrogen-fixing organism that also produces certain substances good for the growth of plants and antibodies that suppress many root pathogens.
3. *Azospirillum* (AZS), also a nitrogen-fixing microorganism beneficial for non-leguminous plants (e.g., onion, plantation crops). It can fix 20–50 kg N ha^{-1} in association with roots. It also produces hormones such as indole acetic acid (IAA), gibberellic acid (GA), cytokinins, and vitamins.
4. BGA and *Azolla,* photosynthetic nitrogen fixers and free-living organisms found in rice fields. These are used as biofertilizers for wetland rice (paddy) and can provide 25–30 kg N ha^{-1} in one crop season, or up to 50 kg N ha^{-1} yr^{-1}.

Phosphate-Solubilizing Biofertilizers

PSB help in the solubilization of P from PR and other sparingly soluble forms of soil P by secreting organic acids, and in the process decreasing their particle size, reducing it to nearly amorphous forms. In addition to bacteria, the fungus *Penicillium belaji* has been shown to increase P availability from native soil and PR sources in calcareous soils.[1] Although their use is rather limited, in some countries like India, PSB use is becoming popular, ranking next to BNF.

Nutrient-Mobilizing Biofertilizers

The common nutrient mobilizers in the soil are mycorrhi- zae, which form association with plant roots and improve the capability of roots to acquire nutrients. The vesicular-arbuscular mycorrhizae (VAM) have been associated with enhanced accumulation of P, Zn, Cu, and S, primarily through greater soil exploration by the mycorrhizal hyphae, thus leading to higher plant growth.

Adoption of INM at Farm

At the farm level, INM optimizes the productivity of the nutrient flows through the soil/crop/livestock system during a crop rotation by making a balance sheet of every nutrient and especially of the major nutrients N, P, and K. The efficiency of a production system depends on the importance of crop uptake versus the total supply of nutrients. Exploitation of plant nutrient stocks is permissible as long as it does not affect the supply of nutrients and the general status of soil fertility.

Moreover, INM improves the production capacity of a farm through the application of external plant nutrient sources and amendments and the efficient processing and recycling of crop residues and on-farm organic wastes. It empowers farmers by increasing their technical expertise and decision-making capacity.[1] During the adoption of INM, special attention should be given to sources of nutrients that may be mobilized by the farmers themselves (manures, crop residues, soil reserves, BNF, etc.). Minimization of losses and replenishment of nutrients from both internal and external sources are of major interest. Although the effects of organic inputs go beyond the nutritional aspects, by contributing to improving soil physical properties and to a better efficiency of fertilizer use, the recycling of organic materials does not suffice to fully replenish the nutrients that are removed by crop harvests. Therefore, an increased and more efficient use of mineral fertilizers in most developing countries is required in the medium term.[8]

A quantitative knowledge of the depletion of plant nutrients from soils may be helpful in devising nutrient management strategies. Nutrient balance exercises serve as instruments to provide indicators for the sustainability of agricultural systems. Nutrient budget and nutrient balance methodologies using various approaches for different situations have been applied widely in recent years at a variety of levels: plot, farm, regional, national, and continental.[1]

Environmental Concerns of INM

Nutrients added through fertilizers, manures, and composts can have both positive and negative effects on the environment depending on the management practices. In this section, the focus is on negative effects. Nutrients, if not taken by plants, can be immobilized by the soil or lost from the soil system. Depending on the nutrient and soil condition, losses can be by volatilization to atmosphere, through soil and water erosion, and by leaching in the soil profile. Leached N can also be lost to the atmosphere through denitrification. A summary of the environmental problems associated with fertilizer use and general strategies to minimize them is presented in Table 4.[25] In general, the negative effects can be summarized as follows:

1. The constant removal of nutrients by crop products without their sufficient replenishment causes a steady decline in soil fertility. The use of low levels of input places additional stress on soil nutrient supplies, resulting in excessive mining of soil nutrients and in depletion of soil fertility, leading to land degradation.
2. Low crop yields resulting from nutrient depletion force farmers to cultivate land under forests or marginal soils that are subject to erosion or desertification, thus leading to soil degradation.
3. Large areas of soils in the tropics are inherently poor in soil nutrients and in addition suffer from problems of acidity, salinity, alkalinity, and Al toxicity.

Nutrient Losses

Of all the inputs, N applications have had the single largest effect on crop yields but N application also creates a matter of greater environmental concern. Nitrogen, if not taken up by plants, can be lost through soil erosion and surface runoff, leached to groundwater, volatilized as ammonia, and denitrified as nitrous oxide (NO) to the atmosphere. The losses can range from 0 to >100 kg N h^{-1} yr^{-1} and they are more in soils with large rainfall intensities and low storage capacity. INM through organic and inorganic input of N can also minimize the leaching losses by improving soil chemical and physical properties.

Nitrate lost by leaching or transported in surface runoff can result in increased nitrate concentrations in drinking water, eutrophication of surface waters, and increased production of NO. It has been estimated that the groundwater under some 22% of the cultivated land in the European Union (EU) has NO_3^- concentrations exceeding the EU upper limit of 20 mg L^{-1}.[1] Nitrate leaching has another associated negative effect because it causes loss of basic cations (Ca^{2+}, Mg^{2+}, and K^+), leading to increased soil acidity. Ammonia volatilization from soil and vegetation contributes to about 21 M Mg yr^{-1} of N,[23] and 14% of mineral N fertilizer used globally is lost as ammonia.[27] The highest emissions of ammonia are in regions with intensive animal production activity (Europe), widespread use of urea (India), and application of ammonium carbonate fertilizer (China). The dominant source of ammonia emission is animal manure, as about 30% N in urine and dung is lost through this route.

Denitrification losses as gaseous dinitrogen (N2) amount to about 14 M Mg yr^{-1}, and N_2O and NO from nitrification/denitrification contribute about another 8 M Mg yr^{-1}N to the total loss.[26] It is estimated that N_2O contributes 5%–6% to the present greenhouse gas effect.

Most of P is lost through crop removal and surface erosion. The $N:P_2O_5$ ratio in most animal manures is about 1:1, whereas plants remove about 2.4–4.5 times more N than P_2O_5. Such residual organic forms of P can thus be leached. However, inorganic forms of P are bound strongly to clays and oxide surfaces in acid soils and precipitated as relatively insoluble calcium phosphates in alkaline soils. Therefore, leaching and surface runoff of phosphate in solution do not generally contribute to eutrophication.

Neither fixed nor exchangeable K is very mobile in soils containing silicate minerals. However, oxisols and ultisols have limited capacity to maintain adequate K. Leaching losses in tropical soils are reported to vary from 52 to 165 kg K ha^{-1} yr^{-1} and they are especially higher under slush and burn.[28] INM with organic matter input is important in the later soils to increase their cation exchange capacity. The loss of K through leaching and erosion is a waste of resources, but it is not known to constitute any environmental or health hazard.

Soil Carbon Sequestration

INM has been shown to have a beneficial impact on C sequestration in soils. For example, in India, the use of NPK and farmyard manure maintained the SOC at 15 g kg^{-1} for a 25 years period compared with a decline to 8.0 g kg^{-1} with NPK alone and 5.0 g kg^{-1} with no fertilizer use. The data in Table 5 from India show an increase in SOC content by application of NPK and manure in five out of six soils, and increase in two out of six soils by use of chemical fertilizers.[29] There was a drastic decrease in SOC content in all soils, which received neither fertilizer nor manure (Table 5). It was also reported that in both rice–wheat and maize–wheat cropping systems, application of FYM or balanced fertilization with NPK resulted in higher C sequestration.[30] The SOC concentration was higher with FYM than with NPK application in both rice–wheat and maize–wheat systems after a period of 32 years.[27] Smith and Powlson[31] reported that 820 M Mg $years^{-1}$ of manure is produced in Europe but only 54% is applied to arable land and the remainder is applied to non-arable agricultural land. They observed that applying manure to cropland can enhance its SOC pool more than it does on pasture land. Smith and Powlson[31] estimated that if all manure were incorporated into arable land in the EU, there would be a net sequestration of 6.8 Tg C yr^{-1}, which is equivalent to 0.8% of the 1990 CO_2–C emissions for the region.

TABLE 5 Average Content of Nutrients in Organic Manures

Manure	N (%)	P_2O_5 (%)	K_2O (%)
Cattle dung	0.3	0.1	0.15
Sheep/goat dung	0.65	0.5	0.03
Human excreta	1.2–1.5	0.8	0.5
Farmyard	0.5	0.15	0.5
Poultry manure	2.9	2.9	2.4
Urban compost	0.5	0.2	0.5

Source: Extracted from Roy et al.[1]

Toxic Accumulation

Along with atmospheric deposition, phosphate fertilizers (both chemically produced ones and PRs) can be a major source of metal addition, especially Cd, to agricultural soils.[32] PRs are the basic raw materials for the production of phosphate fertilizers and PR can contain hazardous elements including undesired heavy metals (e.g., Cd, Cr, Hg, and Pb) and radioactive elements (e.g., U) that are considered to be toxic to human and animal health.[15] Undesirable heavy metals can also originate from finished fertilizers and organic manures (Table 6). In addition to P fertilizers, sewage sludge and animal manure may also contain high amount of metals (Table 6).[33] Repeated application of such wastes is reported to increase crop uptake of metals.[34] Soils from fertilized fields and especially in Europe and Northern America have shown accumulation of Cd. Cadmium ingested by animals and humans accumulates in the kidneys and can lead to organ dysfunction. Therefore, in many countries, the regulatory limits for food, so-called “maximum permissible concentrations (MPCs),” have been worked out. Some countries have also introduced voluntary or mandated concentrations of Cd in fertilizers. Such limits are under constant review, and efforts are also being made to reduce the Cd content of the chemically produced fertilizers.

Soil Fertility Degradation/Improvement

Widespread fertility degradation occurs under many cropping systems even on soils with good initial soil fertility, because nutrient mining or depletion is a widespread problem in low- and medium-input agriculture. This is a major threat to productive sustainable farming. Negative nutrient balances are quite common in many African and Asian countries. For example, in India, the annual gap between nutrients removed by crops and those added through fertilizers has been in the range of 8–10 M Mg of $N+P_2O_5+K_2O$ for several years.[35] Furthermore, soil organic matter (SOM) losses from the surface soils under continuous cultivation are a common phenomenon in the tropics and subtropics. For example, the SOM content of an oxisol in Zambia decreased at a rate of 1.6 Mg h^{-1} yr^{-1} in a long-term experiment.[36] Similarly, long-term studies in subtropics[9] and temperate regions[37] have shown that the

TABLE 6 Soil C Sequestration through INM for 20 Years in Some Indian Soils

Location	Soil	Cropping System	Initial SOC (g kg–1)	SOC after 20 Years (g kg–1)		
				Control	NPK	NPK+FYM
Bhubaneswar	Inceptisol	Rice–rice	2.7	4.1	5.9	7.6
Pantnagar	Mollisol	Rice–wheat	14.8	5.0	9.5	15.1
Barrackpur	Inceptisol	Rice–wheat–jute	7.1	4.2	4.5	5.2
Palampur	Alfisol	Maize–wheat	7.9	6.2	8.3	12.0
Karnal	Alkali soil	Fallow–rice–wheat	2.3	3.0	3.2	3.5
Trivandrum	Ultisol	Cassava	7.0	2.6	6.0	9.8

Source: Extracted from Lal.[29]

TABLE 7 Total Concentration of Heavy Metals in P Fertilizers and Manures

Fertilizers and Manures	Cd mg kg–1	Cr	Pb
Triple superphosphate	9	92	3
Cow manure	1	56	16
Sewage sludge	5	350	90

Source: Extracted from Tandon.[35]

organic matter of soil can be maintained or raised modestly by proper fertilization and especially by organic manures and crop residue management. On the other hand, even under conditions of low-input cropping and with nutrient-depleted soils, fertility degradation can be reversed in acid soils, through better P supply with phosphate fertilizers with lime application and N input via N fixation by legumes. While soil improvements may result in 50% higher yields at a low-input level, more impressive results can be obtained at a high-input level (Table 7). A good example is that of the formerly degraded and low-yielding, but now highly productive soils of Western Europe with present wheat yields of 8–10 Mg ha^{-1}, while the original cereal yield ranged from 0.5 to 1.5 Mg ha^{-1}.[1]

Conclusions

Persistent decline in soil fertility in many parts of the developing world and overfertilization in many temperate developed countries have created nutrient and fertility imbalances in terms of nutrient depletion and hence low crop productivity on one hand, and the excessive nutrients have resulted in the pollution of soil, water, and air on the other. To overcome such problems, the use of the INM system is considered as the most useful strategy because it supplies adequate nutrient, in synchrony with crop demands within a framework of improved land and soil management practices. Implementation of INM systems not only could increase crop production per unit area of land to meet the demands of an increasing population by using chemical fertilizers in integration with nutrients in organic resources available at the community level, coupled with microbial inoculants (e.g., N_2 fixer, PSB, mycorrhizae), but also improves the environment by utilizing these resources more efficiently and in synchrony with crop demands to avoid their losses. Furthermore, the improved understanding of nutrient availability from these resources and the fate of nutrients in an integrated (soil–plant–water–air) system is required as they impinge on soil's chemical, biological, and physical functions for sustainable production and environmental protection.

References

1. Roy, R.N.; Finck, A.; Blair, G.J.; Tandon, H.L.S. *Plant Nutrition for Food Security—A Guide for Integrated Nutrient Management;* FAO Fertilizer and Plant Nutrition Bulletin 16; FAO: Rome, 2006.
2. Singh, R.; Kumar, A.; Shivkumar, B.G. Integrated nutrient management. Available at http:/opals.iitk.ac.in9000/word press/index.php/ipnm (accessed August 24, 2010).
3. Karim, Z. A threat to sustainable agriculture production systems in Bangladesh. Presented at the Meeting of Soil, April 4, 2010; Water Conservation Society, 2007. Available at http://www.allaacademic.com/meta/P173589.
4. Singh, G.B.; Biswas, P.P. Balanced and integrated nutrient management for sustainable crop production. Fert. News (FAI) **2000**, *45*, 55–60.
5. FAI. *Fertilizer Statistics*; Fertilizer Association of India: New Delhi, 2008.

6. Galloway, J.N.; Townsend, A.R.; Erisman, J.-W.; Bekunda, M.; CAI, Z.; Freney, J.R.; Martinelli, L.A.; Seitzinger, S.P.; Sutton, M.A. Transformation of the nitrogen cycle: Recent trends, questions, and potential solutions. Science **2008**, *320,* 889–892.
7. Steinfeld, H.; Gerber, P.; Wassenaar, T.; Castel,V.; Rosales, M.; de Haan, C., Eds. *Livestock's Long Shadow: Environmental Issues and Options*; FAO: Rome, 2006.
8. FAO. *Current World Fertilizer Trends and Outlook to 2009/2010;* Food and Agriculture Organization: Rome, 2005.
9. Doberman, A.; White, P.F. Strategies for nutrient management in irrigated and rain fed lowland rice cultivation. Nutr. Cycling Agroecosyst. **1999**, *48,* 1–18.
10. Bembi, D.K.; Biswas, C.R. Nitrogen balance and N recovery after 22 years of maize. Wheat-cowpea cropping in a long-term experiment. Nutr. Cycling Agroecosyst. **1997**, *47,* 107–114.
11. Singh, G.B.; Swarup, A. Lessons from long-term fertility experiments. Fert. News (FAI) **2000**, *45,* 13–24.
12. Velmurugan, A.; Dadhwal, V.K.; Abrol, Y.P. Reginal nitrogen cycle: An Indian perspective. Curr. Sci. **2008**, *94,* 1455–1468.
13. Pitcairn, C.E.R.; Leith, I.D.; Sheppard, L.J.; Sutton, M.A.; Fowler, D.; Munro, R.C.; Tang, S.; Wilson, D. The relationship between nitrogen deposition, species composition and foliar nitrogen concentrations in woodland flora in the vicinity of livestock farms. Environ. Pollut. **1998**, *102,* 41–48.
14. Lopes, A.S. Soils under Cerrado: A success story in soil management. In IPA-PPI Regional Conf. for Latin America and the Caribbean, Mexico City, June 25–28, 1996; IPA-PPI: Mexico City.
15. FAO. *Use of Phosphate Rocks for Sustainable Agriculture;* FAO Fertilizer and Plant Nutrition Bulletin No. 13; Rome, 2004; 148 pp.
16. Jama, B.; Van Straaten, P. Potential of East African phosphate rock deposits in integrated nutrient management strategies. An. Acad. Bras. Cienc. **2006**, *78,* 781–790.
17. Sharma, S.N.; Prasad, R.; Shivay, Y.S.; Dwivedi, M.K.; Kumar, S.; Davari, M.R.; Ram, M.; Kumar, D. Relative efficiency of diammonium phosphate and mussoorie rock phosphate on productivity and phosphorus balance in a rice-rapeseed-mungbean cropping system. Nutr. Cycling Agroecosyst. **2010**, *86,* 199–209.
18. Mavi, M.S.; Bembi, D.K. Potassium dynamics under integrated nutrient management in rice-wheat system. Agrochimica **2008**, *52*, 83-91.
19. Gupta, A.P. Micronutrient status and fertilizer use scenario in India. J. Trace Elem. Med. Biol. **2008**, *18,* 325–331.
20. Sujatha, M.G.; Lingaraju, B.S.; Palled, Y.B.; Ashalatha, K.V. Importance of integrated nutrient management practices in maize under rainfed condition. Karnataka J. Agric. Sci. **2008**, *21,* 334–338.
21. Buresh, R.J.; Smithson, R.C.; Hellums, D.T. Building soil phosphorus capital in Africa. In *Replenishing Soil Fertility in Africa;* Buresh, R.J., Sanchez, P.A., Calhoun, F., Eds.; SSSA Special Publ. 51; SSSA: Madison, Wisconsin, USA, 1997; 97–109.
22. Mollah, M.R.A.; Asaduzzaman, M.; Khalequzzamanl, K.M.; Siddquie, M.N.A.; Rahim, M.A. Integrated nutrient management for Boro-T.aman rice cropping pattern in the level barind tract area (aez-25). Int. J. Sustainable Crop Prod. **2007**, *2*, 23–27.
23. Khodaei Joghan, A.; Ghalavand, A.; Aghaalikhani, M.; Gholamhoseini, M.; Dolatabadian, A. Comparison among different integrated nutrition management for soil micro and macro elements after winter wheat harvesting and yield. Not. Sci. Biol. **2010**, *2*, 107–111.
24. Sanchez, P.A.; Benites, J.R. Low-input cropping for acid soils of the humid tropics: A transition technology between shifting and continuous cultivation. In *Land Development and Management of Acid Soils in Africa;* Latham, M., Ahn, P., Eds.; International Board for Soil Research Management: Bangkok, Thailand, 1987; 85–106.

25. Pathak, H.; Singh, U.K.; Patra, A.K.; Kalra, N. Fertiliser use efficiency to improve environmental quality. Fert. News, **2004**, *49,* 95–106.
26. FAO/International Fertilizer Industry Association (FAO/ IFA). *Global Estimates of Gaseous Emissions from Agricultural Land;* Rome, 2001; 85 pp.
27. Smil, V. Nitrogen in crop production: An account of global flows. Biogeochemical Cycle **1999**, *13,* 647–662.
28. Malavolta, E. Potassium status of tropical and sub-tropical region soils. In *Potassium in Agriculture*; Munsen, R.D., Ed.; ASA Madison, Wisconsin, 1985; 163–200.
29. Lal, R. The potential of carbon sequestration in soils of south Asia ISCO 2004. In *Conserving Soil and Water for Society: Sharing Solutions,* 13th International Soil Conservation Organisation Conference, Brisbane, July 2004.
30. Kukal, S.S.; Rasool, R.; Benbi, D.K. Soil organic carbon sequestration in relation to organic and inorganic fertilization in rice–wheat and maize–wheat systems. Soil Tillage Res. **2009**, *102,* 87–92.
31. Smith, P.; Powlson, D.S. Considering manure and carbon sequestration. Science **2000**, *287,* 428–429.
32. Singh, B.R. Trace element availability to plants in agricultural soils, with special emphasis on fertilizer inputs. Environ. Rev. **1994**, *2*, 133–146.
33. Webber, M.D.; Singh, S.S. Contamination of agricultural soils. In *The Health of Our Soils—Toward Sustainable Agriculture in Canada*; Acton, D.F., Gregorich, L.J., Eds.; Centre for Land and Biological Resources Research: Ottawa, ON, Canada, 1995; 87–96.
34. Krebs, R.; Gupta, S.K.; Furrer, G.; Schulin, R. Solubility and plant uptake of metals with and without liming of sludge-amended soils. J. Environ. Qual. **1998**, *27,* 18–23.
35. Tandon, H.L.S. *Fertilisers in Indian agriculture—From 20th to 21st Century*; Fertiliser Development and Consultation Organisation: New Delhi, 2004; 239 pp.
36. Singh, B.R.; Goma, H.C. Long-term soil fertility management in eastern Africa. In *Soil Management—Experimental Bias for Sustainability and Environmental Quality;* Lal, R., Stewart, B.A., Eds.; Lewis Publishers: Boca Raton, 1995; 347–382.
37. Campbell, C.A.; Myers, R.J.K.; Curti, D. Managing nitrogen for sustainable crop production. Fert. Res. **1995**, *42,* 277–296.

47

Integrated Pest Management

H.F. van Emden

Introduction

There have been many definitions of integrated pest management. The most practical operational definition is probably that by President Jimmy Carter of the United States in his 1979 environmental message: Integrated Pest Management "uses a system approach to reduce pest damage to tolerable levels through a variety of techniques … and, when necessary and appropriate, chemical pesticides." An excellent, more detailed definition is given by Flint and van den Bosch: "Integrated pest management (IPM) is an ecologically based pest control strategy that relies heavily on natural mortality factors such as natural enemies and weather and seeks out control tactics that disrupt these factors as little as possible. IPM uses pesticides, but only after systematic monitoring of pest populations and natural control factors indicates a need. Ideally, an integrated pest management program considers all available pest control actions, including no action, and evaluates the potential interaction among various control tactics, cultural practices, weather, other pests, and the crop to be protected."[1]

IPM as defined above represents, as a result of usage, a rather different concept from Apple and Smith's[2] original 1976 definition (see below) and is so well used that it is often written or spoken as IPM without further explanation. IPM is now the preferred option to replace both prophylactic (= insurance) reliance on insecticides and the concept of replacing insecticides with single-component biological control. Both these latter methods are now universally seen as components of IPM, though biological control alone still has its niche in low value subsistence agriculture and other situations where insecticide cost is prohibitive or where the pest has multiple tolerance to insecticides.

Derivation of Integrated Pest Management

Following on from problems created by the extensive use of pesticides in the 1940s and 1950s, particularly the selection of pest strains resistant to insecticides and damage to biological control, workers at the University of California at Berkeley evolved the concept of "integrated control." This was based on their work combining biological control with reduced and more selective pesticide doses against pesticide-resistant spotted alfalfa aphid (*Therioaphis maculata*), published in a landmark paper in the scientific journal *Hilgardia* in 1959.[3] The more selective pesticide doses improved the survival of natural enemies while still killing a proportion of the aphids. This principle was reflected in the definition of "integrated control" by the Californians as "pest control which combines and integrates biological and chemical control. Chemical control is used as necessary and in a manner which is least disruptive to biological control." This episode is the origin of the "I" in IPM.

Following on from an Australian concept of "managing" pest populations, the Canadian Beirne proposed "Pest Management" at a conference in Raleigh, North Carolina, in 1970.[4] It was an all-embracing expression for the selection of that combination of control measures or that single measure that was "in the best long term interests of mankind." Beirne envisaged that sometimes a single measure, for example, biological control alone, might be a better solution in environmental terms than the use of more than one measure implicit in "integrated control." Whether this represented any substantive change in thinking is debatable, especially as Beirne himself conceded that most often "integrated control" would be the preferred "pest management" solution. Be this as it may, the replacement of the word "control" with "management" was appealing and topical, and it caught on.

In 1976, there appeared a book[2] entitled *Integrated Pest Management* edited by Lawrence Apple of North Carolina State University and Ray Smith of The University of California at Berkeley. Their description of IPM ran as follows: "The concept of pest management has now been broadened to include all classes of pests (pathogens, insects, nematodes, weeds) and in this context is commonly referred to as IPM with the implication of both methodological and disciplinary integration." As pointed out earlier, the "disciplinary integration" implications have never been generally accepted. Today, the next step in the evolution of the terminology is to take IPM in Apple and Smith's sense and to see this as just one aspect of crop production. The term "Integrated Crop Management" now seems to be gathering momentum.

Need for Integrated Pest Management

The intensification of agriculture with the overuse of and overreliance on synthetic pesticides in the 1940s and 1950s caused serious side effects including the resurgence of the target pests, the appearance of new pests, and the development of tolerance to individual pesticides in many pest species. It was therefore clear that pests had been "mismanaged," and the following failures to manage them correctly could be identified:

1. Overdosing with pesticides, leading to the development of tolerance;
2. Loss of biological control as result of pesticide use and a reduction of biodiversity in agroecosytems;
3. The introduction of genetically uniform, high-yielding but pest susceptible new cultivars over large areas of monoculture;
4. Agronomic changes such as abandonment of cultural controls and of mixed cropping.

These translate respectively into the four main building blocks of IPM[5]:

a. *Chemical control:* Decisions to spray are guided by economic thresholds, and selective materials are chosen where available;
b. *Biological control:* Use of pesticides in a selective manner. Natural enemies can also be introduced or recolonized and/or promoted by habitat modification, including the planning of biodiversity into farm management;
c. *Host plant resistance:* The introduction of crop varieties with at least partial plant resistance, based as far as possible on more than a single mechanism;
d. *Cultural control:* Can break the life cycle of a pest or greatly improve conditions for natural enemies, but must be compatible with farm management systems.

Other components of IPM, such as insect growth regulators, the confusion pheromone technique, and sterile male release, are also available in relation to particular targets.

As IPM for a single target implies managing the pest population by more than one method, the following two "golden rules" of IPM are particularly important:

a. *If* a single method gives adequate control on its own, then there is danger of a tolerant pest strain increasing in gene frequency and no opportunity to use a second method in addition. The method therefore needs to be made less efficient (reduced dose of pesticide, partial host plant resistance rather than immunity) for there to be value in introducing another control method to supplement it;
b. Methods are increasingly worth combining to the extent that the control then achieved exceeds the additive effects of the two methods in isolation.

Development of Pest Management Packages

There are three distinct philosophies to the synthesis of pest management packages.[5]

Menu Systems

These provide considerable choice for a farmer to select control methods for combination in relation to the pest spectrum in the area. The results of many scientist-years of research into different methods enable menus to be provided but, as the choice of the farmer may vary, the methods will be highly target-specific and not dependent on interaction with each other. Often several methods may be available against one pest species. Cotton pest management is a good example of a menu system. Menu systems may be transferable to the same crop in another country or continent with little modification.

Computer-Derived Packages

If sufficient data are available, a computer should be able to select the components of a menu that are in the best interests of whatever selection criteria the computer is given. Some sophisticated models driven by crop growth via the weather are already available for orchard systems. Again, however, interactions between control methods are unlikely to be exploited, as they are rarely adequately understood.

Protocols

These are "dish-of-the-day" menus—that is, a complete package on a "take it or leave it" basis. These are characteristic of crop scenarios where there is a paucity of the detailed long-term research needed for a menu system. Protocols may be very simple. For example, quite effective IPM has been introduced in developing agriculture by little more than combining the introduction of economic thresholds with more selective pesticidal materials (e.g., bacterial toxins). Later stages may be the release of mass-produced beneficials (particularly the egg parasitoid *Trichogramma)* and the introduction of partial plant resistance against a key pest following cultivar testing and plant breeding at national research stations.

Protocols may also be derived directly by experimentation, without the need for the extensive background biological and ecological understanding needed for menu systems.

Conclusions

Apart from IPM being seen as one component of a holistic integration of crop production (integrated crop management), the concept is the widely accepted view of what pest control in modern crop production should be. In many countries around the world it has been declared as the national policy; it is also the policy of international agencies such as the Food and Agriculture Organization. It is the approach being used by many crop consultants and the method favored in developed countries by the large supermarkets, which have such an influence on farm management today.

Until relatively recently, the integrated approach to pest control was accepted temporarily by growers at times of a crisis of pest tolerance to insecticides, to be abandoned just as soon as a new and effective pesticide was marketed. However, IPM is now being introduced even where there is not a current crisis; at long last the message seems to have sunk in, that it is a way of preventing pesticide crises in the first place.

References

1. Flint, M.L.; van den Bosch, R. *Introduction to Pest Management;* Plenum: New York, 1981; 1–240.
2. Apple, J.L.; Smith, R.F. *Integrated Pest Management;* Plenum: New York, 1976; 1–200.
3. Stern, V.M.; Smith, R.F.; van den Bosch, R.; Hagen, K.S. The integrated control concept. Hilgardia **1959**, *29* (2), 81–101.
4. Beirne, B.P. The Practical Feasibility of Pest Management Systems. In *Concepts of Pest Management;* Rabb, R.L., Guthrie, F.E., Eds.; North Carolina State University: Raleigh, NC, 1970; 158–169.
5. van Emden, H.F.; Peakall, D.B. *Beyond Silent Spring: Integrated Pest Management and Chemical Safety;* Chapman and Hall: London, 1996; 1–322.

48

Integrated Weed Management

Heinz Müller-Schärer and Alexandra Robin Collins

Introduction

Agrochemical companies promise that transgenic crops will simplify pest management programs through the use of singular chemical tactics. This "silver-bullet" approach has consistently failed and almost certainly will again because of a failure to understand the ecological relationships governing population size and diversity.[1] Furthermore, in many countries, pesticide policies have called for significant use reductions together with the promotion of biodiversity in agro-ecosystems.[2] Initiatives to reduce reliance on herbicides will require a much fuller understanding of how management practices complement one another to maintain weed populations at low equilibrium densities. Biological control approaches require, but also provide, detailed insight into weed–crop interactions and how they are influenced by both the biotic and abiotic environments. They can, thus, be viewed as the basis for integrated production.[3] In most cases, only combinations with other weed management tools will result in acceptable levels of weed control. Various types of integration can be envisaged, of which preventative measures will be most important for developing sustainable agricultural production.

Weed Control, Weed Science, and Integrated Weed Management

Agricultural weed management in farming systems is diverging in two distinct directions. In one set of farming systems, farmers rely primarily on herbicides to suppress weeds. This approach is exemplified by the extensive maize (*Zea mays* L.)/soybean [*Glycine max* (L.) Merr.] system of the midwestern United States, where >110 million kg of herbicide active ingredients are applied annually to >95% of the area planted with those two crops.[4] In a second set of farming systems, herbicides are largely or entirely avoided, and weeds are mainly suppressed using physical and ecological tactics. The existence and risk of development of herbicide resistance make herbicide-dependent cropping systems increasingly vulnerable. Moreover, widespread concern about environmental side effects of herbicides, combined with fear for public health, has resulted in several herbicides being banned in some countries and increasing pressure on farmers to reduce their use.[4]

In contrast to the disciplines of plant pathology and entomology, the "how to control" approach was shaped early on in weed science and, until recently, has dominated the discipline. The fact that weeds have been regarded as a problem that can be controlled with herbicides, rather than managed through cropping system design,[5] has resulted in a time lag in developing integrated weed management systems, as compared to integrated pest and disease management systems.[1] The United Nations Conference on Environment and Development (UNCED), in its Agenda 21, recognized integrated pest management (IPM) as the preferred strategy to achieve sustainable agricultural production.[6] IPM typically involves a reduction in the reliance on chemical pesticides, including herbicides.[7] Furthermore, the Convention on Biological Diversity[8] and prominent researchers in the field make the case that biological control should be given priority as a component of future pest management.[9,10]

Methods Used to Control Crop Weeds Biologically

Three principal methods of biological weed control can be distinguished (Figure 1).[3,11] First, the "inoculative" or "classical" approach aims to control naturalized weeds by the introduction of exotic control organisms from the weed's native range. They are released over only a small area of the total weed infestation and control is achieved gradually. Successful control depends on favorable conditions promoting an increase in the control agent's population, establishment of epiphytotics, and, thus, reduction of the target weed population. Second, the "inundative" or "bioherbicide" method uses periodic releases of an abundant supply of the control agent over the entire weed population to be controlled. Such biological agents generally are manufactured, formulated, standardized, packaged, and registered like chemical herbicides. Compared to the other two approaches, this approach is characterized by higher application costs and a relatively short time period to achieve a potential control success. Though there have been a number of successful biological control programs against crop weeds with some products resulting in commercial registration (Table 1), bioherbicides have still not managed to occupy a sizable share of

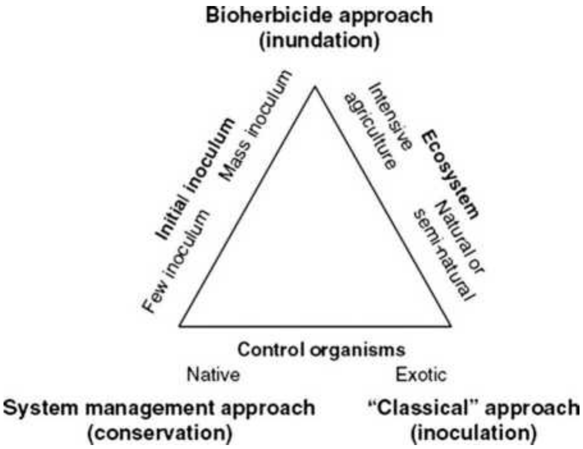


FIGURE 1 Schematic diagram showing the three methods of biological weed control in agro-ecosystems (from Müller- Schärer and Schaffner[11]) with respect to the initial size of the inoculum released, the origin of the control organisms used, and the ecosystem where the biological control program was implemented. See text for details.

TABLE 1 Selected Examples of Applied Integrated Weed Management Involving Biological Control, and Present Product Status

Target Weed	Common Pathogen(s)	Crop/Habitat	Type of Integration[a]	Type of Biological Control Strategy and Status of Product	Reference(s)
Northern jointvetch (.*Aeschynomene virginica*)	Collego™: *Colletotrichum gloeosporioides* f. sp. *aeschynomene*	Rice and soybean fields	Pathogen–pathogen–herbicide [H] (PSA)	Inundative method Reregistered as Lockdown.	[19]
Stranglervine (*Morrenia odorata*)	DeVine®: *Phytophthora palmivora*	Citrus groves	Pathogen–herbicide [H]	Inundative method Status unknown; EPA issued a Reregistration Eligibility Document (RED) in 2006; it appears that no one has come forward to reregister DeVine.	[20,21]
Velvetleaf (*Abutilon theophrasti*)	*Colletotrichum coccodes*	Corn, soybean	Pathogen–herbicide [H;V/P] (PSA)	Classical method Several organisms are being studied for use as biocontrol agents, though none are currently available for release.	[22,23]
Barnyard grass (*Echinochloa crus-galli*)	*Colletotrichum graminicola*	Various crops	Pathogen–herbicide [V/P]	Classical method Several organisms are being studied for use as biocontrol agents, though none are currently available for release.	[24]
Nutsedges (*Cyperus* spp.)	Dr BioSedge: *Puccinia canaliculata*	Various crops	Pathogen–herbicide [V/P] (PSA)	Inundative method Registered, but commercial development abandoned due to uneconomic production system and resistance in some weed biotypes	[25,26]
Spotted knapweed (*Centaurea maculosa*)	*Cyphocleonus achates, Agapeta zoegana, Larinus* spp., *Urophora* spp., etc. (see Julien and Griffiths 1998)	Rangeland	Insect–plant competition [V/E] Insect–herbicide [V/P or PS]	Classical method Several organisms are being studied for use as biocontrol agents, though none are currently available for release.	[27–30]

(*Continued*)

TABLE 1 (*Continued*) Selected Examples of Applied Integrated Weed Management Involving Biological Control, and Present Product Status

Target Weed	Common Pathogen(s)	Crop/Habitat	Type of Integrationa	Type of Biological Control Strategy and Status of Product	Reference(s)
Nodding thistle (*Carduus nutans*)	*Rhinocyllus conicus, Trichosirocalus horridus. Cassida rubiginosa*	Rangeland	Insect–herbicide [V/P] Insect–plant competition [V/E]	Classical method Several organisms are being studied for use as biocontrol agents, though none are currently available for release.	[31]
St. John's wort (*Hypericum perforatum*)	*Chiysolina hyperici, Chrysolina quadrigemina*	Rangeland	Insect–fire [V/E] Insect–plant competition [V/E]	Classical method Several organisms are being studied for use as biocontrol agents, though none are currently available for release.	[32–34]
Water hyacinth (*Eichhomia crassipes*)	*Neochetina eichhorniae, Neochetina bruchi, Sameodes albiguttalis*	Aquatic	Insect–herbicide [V/P] Insect–pathogen–herbicide [PS]	Classical method Currently, there are no registered bioherbicides, but several candidates have been identified.	[35]
Floating fern (*Salvinia molesta*)	*Cyrtobagous salviniae*	Aquatic	Insect–herbicide [V/PS or P] Insect–fertilizer [V/P]	Classical method Currently, there are no registered bioherbicides, but several candidates have been identified.	[36]
Dyer's woad (*Isatis tinctoria*)	Woad Warrior: *Puccinia thlaspeos*	Farms, rangeland, waste areas, and roadsides	Pathogen (rust)–herbicide [V/P]	Inundative method Registered product. Consists of rust spores on finely ground leaf and stem pieces of infected dyers woad.	[37]
Alders, aspen, and other hardwoods	Chontrol™: *Chondrostereum purpureum*	Rights of way and forests	Pathogen (fungus) – herbicide [V/P]	Inundative method Registered in Canada and the United States and commercially available.	[38–39]
Dandelion (*Taraxacum*)	Sanitor: *Sclerotinia minor*	Lawn and turf	Pathogen (fungus)–herbicide [H;V/P]	Inundative method Registered and commercially available in Canada. U.S. registration is pending.	[40–12]

the market.[12] This is mainly due to the fact that the reliability of field efficacy has not reached levels comparable with that of chemical herbicides.[13] According to Charudattan,[14] of the bioherbicide projects underway, only 8% of them were successful, leaving 91.5% of the projects uncertain, untried, or ineffective. Thus, it has become increasingly important to prioritize projects with high pathogen aggressiveness, high speed of disease increase, and high rates of population increase, what Charudattan[9] refers to as "killer traits." Third, and more recently, the "system management approach" of biological weed control had been described.[15,16] It is related to the conservation and augmentative approaches distinguished by some authors. Its aim is to shift the competitive weed–crop relationship in favor of the latter, mainly by stimulating the buildup of a disease epidemic or insect outbreak on the target weed population. The approach excludes the use of exotic organisms (classical approach) and the use of mass amounts of inoculum applied like a herbicide to the whole weed population (bioherbicide approach).

Integrating Biological Control with Other Methods of Weed Management

Weed problems in agro-ecosystems are rarely caused by single weed species. Clearly, biological control, with its inherently narrow species-specific approach, has to be considered as an integrated component of a well-designed pest management strategy, not as a cure by itself. In most cases, combinations of biological agents with other weed management tools will be needed to produce acceptable levels of overall weed control. Such integration can be viewed as a vertical integration of various control tactics against a single weed species, or as a horizontal integration across different weed species in one crop (Table 1).[17] Horizontal integration mainly involves the combination of microbial herbicides with chemical herbicides or mechanical methods to broaden the spectrum of weed species controlled. Furthermore, in situations where particularly high doses of herbicides are needed to control a single weed species while the rest of the weed flora could be controlled by lower amounts, biological control may allow considerable reduction of herbicide inputs and contribute to maintaining species diversity in crops. Three possible types of vertical integration of biological control with other methods of weed management can be distinguished, both in time and space: purpose-specific approaches, ecological integration, and physiological integration (Table 1).[18]

Purpose-Specific Approaches

The type and level of control are chosen according to the context-specific requirements. This often involves different management methods to be applied at different sites. For instance, for a weed that is still spreading, chemical herbicides may well be the method of choice to remove new infestations, while biological control may be relied on to give long-term control of large, established infestations.[43]

Ecological Integration

This term is given to situations where different approaches are used often at the same time on the same infestation. Integration with herbicides[44,45] and with plant (crop) competition[146–481] is most widely envisaged. This type of integration essentially summarizes holistic approaches that encompass all modifications to the environment, which may favor the effectiveness of biological control agents and facilitate the management of a weed population.[49]

Physiological Integration

This type of integration exploits synergistic interactions between changes in the biochemistry of weeds, often produced by sublethal effects of herbicides and the effectiveness of biological control agents.

Herbicides (or other "synergists") are known to increase incidence of infection and to enhance the growth of pathogens,[50–55] but infection by the pathogen may also facilitate the uptake of herbicides, mainly by injuring the cuticle and epidermis of the host. In addition, various studies have shown greatly increased disease severity and agent effects when combined with phytotoxic metabolites produced by the pathogen[56] or with specific formulation and delivery techniques of microbial herbicides.[53,57] Thus, physiological integration is directed toward combined effects with biological control agents on plant individuals.

Ultimately, optimal management, with minimal disruptive interventions, requires a good understanding of the weed's biology and, especially, population dynamics.[58] Biological weed control requires, and provides, a detailed ex ante analysis of the problem situation, especially of the crop environment, revealing interactions between the various components and their underlying interactions. It should, therefore, be the strategy that is basic to integrated production systems. Bridges between different disciplines need to be built to optimize the fit of biological control into existing management systems.[3,18,59]

Conclusions

When weeds are no longer regarded as a problem to be resolved by curative tactics, then prevention becomes the keyword and integrated cropping management becomes the new focus, of which integrated weed management is an important component. Much work remains to be done by scientists spanning a broad range of disciplines in order to be able to integrate soil, crop, and weed management effectively.[4] Further challenges for weed science are the elaboration of effective practices for new crops, new production systems for enlarged farms and fields, and the consequences of climate change. Furthermore, adequate answers need to be found for the increased concern about the conservation of biodiversity and the growing consumer demands on food safety.[60] In parallel, to transfer the scientific knowledge into farming practices, a considerable amount of time must be spent with farmers in order to understand the true practical dimensions of the increasingly complex study systems. In this cropping system design approach, numerous fitness-reducing and mortality events are integrated to manage weed populations, with herbicides being used as a last resort. Prevention involves any aspect of management that favors the crop relative to the weed. This includes the development of competitive crop cultivars, crop rotation, mixed cropping, and allelopathy.[61] Preventative control requires a detailed insight into weed biology and ecology and the ways in which they interact with the crop. Biological control provides a fundamental tool for successful management of weed populations, where weed control no longer considers crop production in a weed-free environment, but instead as a reduction of weed-induced yield losses. By that, it greatly contributes to promoting biodiversity in human-influenced landscapes, a central pillar of modern sustainable agriculture.

Acknowledgments

We greatly thank Karen L. Bailey, Graëme Bourdôt, Raghavan (Charu) Charudattan, Gary Peng, and Maurizio Vurro for sending us information and references for updating Table 1.

References

1. Mortensen, D.A.; Bastiaans, L.; Sattin, M. The role of ecology in the development of weed management systems: An outlook. Weed Res. **2000**, *40* (1), 49–62.
2. Scheepens, P.C.; Müller-Schärer, H.; Kempenaar, C. Opportunities for biological weed control in Europe. BioControl **2001**, *46* (2), 127–138.
3. Müller-Schärer, H.; Scheepens, P.C.; Greaves, M.P. Biological control of weeds in European crops: Recent achievements and future work. Weed Res. **2000**, *40* (1), 83–98.

4. Liebman, M.; Davis, A.S. Integration of soil, crop and weed management in low-external-input farming systems. Weed Res. **2000**, *40* (1), 27–47.
5. Buhler, D.D. Challenges and opportunities for integrated weed management. Weed Sci. **2002**, *50* (3), 273–280.
6. Agenda 21: Programme of Action for Sustainable Development. U.N. GAOR, 46th Sess., Agenda Item 21, UN Doc A/Conf.151/26, 1992.
7. Hatcher, P.E.; Melander, B. Combining physical, cultural and biological methods: Prospects for integrated nonchemical weed management strategies. Weed Res. **2003**, *43* (5), 303–322.
8. Convention on Biological Diversity. Decision VIII/28 Impact Assessment: Voluntary Guidelines on Biodiversity-Inclusive Impact Assessment, 2006. Available at http://www.cbd.int/decision/cop/?id=11042 (accessed February 2011).
9. Charudattan, R. A reflection on my research in weed biological control: Using what we have learned for future applications. Weed Technol. **2010**, *24* (2), 208–217.
10. Müller-Schärer, H.; Frantzen, J. An emerging system management approach for biological weed control in crops: *Senecio vulgaris* as a research model. Weed Res. **1996**, *36* (6), 483–491.
11. Müller-Schärer, H.; Schaffner, U. Classical biological control: Exploiting enemy escape to manage plant invasions. Biol. Invasions **2008**, *10* (6), 859–74.
12. Bastiaans, L.; Paolini, R.; Baumann, D.T. Focus on ecological weed management: What is hindering adoption? Weed Res. **2008**, *48* (*6*), 481–491.
13. Hallett, S.G. Where are the bioherbicides? Weed Sci. **2005**, *53* (3), 404–415.
14. Charudattan, R. Ecological, practical, and political inputs into selection of weed targets: What makes a good biological control target? Biol. Control **2005**, *35* (3), 183–196.
15. Müller-Schärer, H.; Frantzen, J. An emerging system management approach for biological weed control in crops: *Senecio vulgaris* as a research model. Weed Res. **1996**, *36* (6), 483–491.
16. Grace, B.S.; Müller-Schärer, H. Managing crop–weed interactions: Biological control of *Senecio vulgaris* in carrots (*Daucus carota*) with the rust fungus *Puccina lagenopho- rae*. Basic Appl. Ecol. **2003**, *4* (4), 375–384.
17. Watson, A.K.; Wymore, L.A. Biological control, a component of integrated weed management. In *VII International Symposium of Biological Control of Weeds;* Delfosse, E.S., Ed.; Ist. Sper. Patol. Veg. (MAF): Rome, Italy, 1989; 101–106.
18. Cullen, J.M. Integrated Control and Management. In *Proceedings of the IX International Symposium on Biological Control of Weeds*; Moran, V.C. and Hoffman, J.H. Eds.; University of Cape Town, South Africa: Stellenbosch, South Africa, 1996; 483–486.
19. Bailey, K.L.; Boyetchko, S.M.; Langle, T. Social and economic drivers shaping the future of biological control: A Canadian perspective on the factors affecting the development and use of microbial biopesticides. Biol. Control **2010**, *52*, 221–229.
20. Campbell, C.L.; McCaffrey, J.P. Populations trends, seasonal phenology and impact of *Chrysolina quadrigemina*, *C. hyperici* (Coleoptera, Chrysomelidae), and *Agrilus hyperici* (Coleoptera, Buprestidae) associated with *Hypericum perforatum* in Northern Idaho. Environ. Entomol. **1991**, *20,* 303–315.
21. Charudattan, R. The mycoherbicide approach with plant pathogens. In *Microbial Control of Weeds;* Tebeest, D.O., Ed.; Chapman and Hall: New York, 1991; 24–57.
22. Ditommaso, A.; Watson, A.K. Impact of a fungal pathogen, *Colletotrichum coccodes* on growth and competitive ability of *Abutilon theophrasti.* New Phytol. **1995**, *131,* 51–60.
23. Dumas, M.T.; Wood, J.E.; Mitchell, E.G.; Boyonoski, N.W. Control of stump sprouting of *Populus tremuloides* and *P. grandidentata* inoculation with *Chondrostereum purpureum.* Biol. Control **1997**, *10,* 37–41.
24. Goeden, R.D.; Andrés, L.A. Biological control of weeds in terrestrial and aquatic environments. In *Handbook of Biological Control: Principles and Applications*; Bellows, T.S., Fisher T.W., Eds.; Academic Press: San Diego, New York, 1999; 1046.

25. Harris, P. Effects of *Urophora affinis* Frfld. and *Urophora quadrifasciata* (Meig) (Diptera, Tephritidae) on *Centaurea diffusa* Lam. and *Centaurea maculosa* Lam. (Compositae). J. Appl. Entomol. **1980**, *90,* 190–201.
26. Harris, P.; Maw, M. *Hypericum perforatum* L., St. John's- wort (Hy-pericaceae). In *Biological Control Programmes againstInsects and Weeds in Canada;* Kelleher, J.S.; Hume, M. A, Eds.; Commonwealth Agriculture Bureau: London, 1984; 171–177.
27. Harris, P.; Peschken, D.; Milroy, J. Status of biological control of weed *Hypericum perforatum* in British Columbia. Can. Entomol. **1969**, *101,* 1–15.
28. Kok, L.T. Classical biological control of nodding and plumeless thistles. Biol. Control **2001**, *21,* 206–213.
29. Kropp, B.R.; Hansen, D.R.; Thomson, S.V. Establishment and dispersal of *Puccinia thlaspeos* in field populations of dyer's woad. Plant Dis. **2002**, *86*, 241–246.
30. Julien, M.H.; Griffiths M.W. *Biological Control of Weeds. A World Catalogue of Agents and Their Target Weeds*. CABI Publishing: Wallingford, 1998; 1–233.
31. Lang, R.F.; Richard, R.D.; Parker, P.E.; Wendel, L. Release and establishment of diffuse and spotted knapweed biocontrol agents by USDA, APHIS, PPQ, in the United States. Pan-Pac. Entomol. **2000**, *76,* 197–218.
32. Maddox, D.M. Biological control of diffuse knapweed (*Centaurea diffusa*) and spotted knapweed (*C. maculosa*). Weed Sci. **1982**, *30,* 76–82.
33. Mèller, H.; Schroeder, D.; Gassmann, A. *Agapeta zoegana* (L) (Lepidoptera, Cochylidae), a suitable prospect for biological control of spotted knapweed, *Centaurea maculosa* Delamarck, Monnet and *Centaurea diffusa* Delemark, Monnet (Compositae) in North America. Can. Entomol. **1988**, *120,* 109–124.
34. Müller-Schärer, H. The impact of root herbivory as a function of plant density and competition: survival, growth and fecundity of *Centaurea maculosa* (Compositae) in field plots. J. Appl. Ecol. **1991**, *28*, 759–776.
35. Müller-Schärer, H.; Schroeder, D. The biological control of *Centaurea* spp. in North America: Do insects solve the problem? Pestic. Sci. **1993**, 37, 343–353.
36. Phatak, S.C.; Callaway, M.B.; Vavrina, C.S. Biological control and its integration in weed management systems for purple and yellow nutsedge (*Cyperus rotundus* and *Cyperus esculentus).* Weed Technol. **1987**, 1, 84–91.
37. Phatak, S.C.; Sumner, D.R.; Wells, H.D.; Bell, D.K.; Glaze, N. C. Biological control of yellow nutsedge with the indigenous rust fungus *Puccinia canaliculata*. Science **1983**, *219,* 1446–1447.
38. Riddle, G.E.; Burpee, L.L.; Boland, G.J. Virulence of *Sclerotinia sclerotiorum* and *S. minor* on Dandelion (*Taraxacum officinale).* Weed Sci. **1991**, *39,* 109–118.
39. Room, P.M.; Harley, K.L.S.; Forno, I.W.; Sands, D.P.A. Successful biological control of the floating weed *Salvinia*. Nature **1981**, *294,* 78–80.
40. Schnick, P.J.; Stewart-Wade, S.M.; Boland, G.J. 2,4-D and *Sclerotinia minor* to control common dandelion. Weed Sci. **2002**, *50,* 173–178.
41. Tebeest, D.O.; Templeton, G.E. Mycoherbicides—Prog- ress in the biological control of weeds. Plant Dis. **1985**, *69*, 6–10.
42. Wymore, L.A.; Poirier, C.; Watson, A.K.; Gotlieb, A.R. *Colletotrichum coccodes*, a potential bioherbicide for control of velvetleaf (*Abutilon theophrasti*). Plant Dis. **1988**, 72, 534–538.
43. Yang, Y.K.; Kim, S.O.; Chung, H.S.; Lee, Y.H. Use of *Colletotrichum graminicola* KA001 to control barnyard grass. Plant Dis. **2000**, *84,* 55–59.
44. Müller-Schärer, H.; Schroeder, D. The biological control of *Centaurea* spp. in North America: Do insects solve the problem? Pestic. Sci. **1993**, *37* (4), 343–353.
45. Scheepens, P.C. Joint action of *Cochliobolus lunatus* and atrazine on *Echinochloa crus-galli* (L.) Beauv. Weed Sci. **1987**, *27* (1), 43–47.

46. Wymore, L.A.; Watson, A.K.; Gotlieb, A.R. Interaction between *Colletotrichum coccodes* and thidiazuron for control of velvetleaf (*Abutilon theophrasti*). Weed Sci. **1987**, *35* (3), 377–383.
47. Sheppard, A.W. The interaction between natural enemies and interspecific plant competition in the control of invasive pasture weeds. In *IX Int. Symposium on Biological Control of Weeds;* Moran, V.C., Hoffman, J.H., Eds.; Stellenbosch, South Africa, 1996; 19–26.
48. DiTommaso, A.; Watson, A.K.; Hallett, S.G. Infection by the fungal pathogen *Colletotrichum coccodes* affects vel- vetleaf (*Abutilon theophrasti*)-soybean competition in the field. Weed Sci. **1996**, *44* (4), 924–933.
49. Müller-Schärer, H.; Rieger, S. Epidemic spread of the rust fungus *Puccinia lagenophorae* and its impact on the competitive ability of *Senecio vulgaris* in celeriac during early development. Biocontrol Sci. Technol. **1998**, *8* (1), 59–72.
50. Newman, R.M.; Thompson, D.C.; Richman, D.B. Conservation strategies for the biological control of weeds. In *Conservation Biological Control;* Barbosa, P., Ed.; Academic Press: San Diego, USA, 1998; 371–396.
51. Hasan, S.; Ayres, P.G. The control of weeds through fungi: Principles and prospects. New Phytol. **1990**, *115* (2), 201–222.
52. Sharon, A.; Amsellem, Z.; Gressel, J. Glyphosate suppression of an elicited defence response. Increased susceptibility of *Cassia obtusifolia* to a mycoherbicide. Plant Physiol. **1992**, *98* (2), 654–659.
53. Gressel, J.; Amsellem, Z.; Warshawsky, A.; Kampel, V.; Michaeli, D. Biocontrol of weeds: Overcoming evolution for efficacy. J. Environ. Sci. Health, Part B **1996**, *31* (3), 399–404.
54. Duke, S.; Wedge, D.; Cerdeira, A.; Matallo, M. Interactions of synthetic herbicides with plant disease and microbial herbicides. In *Novel Biotechnologies for Biocontrol Agent Enhancement and Management;* Springer: Netherlands, 2007; 277–296.
55. Gressel, J. Herbicides as synergists for mycoherbicides, and vice versa. Weed Sci. **2011**, *58* (3), 324–328.
56. Peng, G.; Byer, K.N. Interactions of *Pyricularia setariae* with herbicides for control of green foxtail (*Setaria viridis*). Weed Technol. **2005**, *19* (3), 589–598.
57. Vurro, M.; Bottalico, A.; Capasso, R.; Evidente, A. Cy- tochalasins from phytopathogenic *Ascochyta* and *Phoma* species. In *Toxins in Plant Disease Development and Evolving Biotechnology;* Mukherji, K.G.; Upadhyay, R.K., Eds.; IBH Publishing Co. Pvt, Ltd.: Oxford, 1997; Vol. 7, 127–147.
58. Greaves, M.P. Microbial herbicides—factors in development. In *Crop Protection Agents from Nature: Natural Products and Analogues;* Copping, L.G., Ed.; Royal Society of Chemistry: Cambridge, U.K., 1996; 444–467.
59. Cousens, R.; Mortimer, M., Eds. *Dynamics of Weed Populations*; Cambridge University Press: London, 1995; 332.
60. Fernandez-Quintanilla, C.; Quadranti, M.; Kudsk, P.; Barberi, P. Which future for weed science? Weed Res. **2008**, *48* (4), 297–301.
61. Zimdahl, R.L., Ed. *Fundamentals of Weed Science;* Academic Press, Inc.: San Diego, CA, 1993; 450.

49

Manure Management: Compost and Biosolids

Bahman Eghball
and Kenneth
A. Barbarick

Manure

Manure (animal waste) is generated in beef cattle feedlots, swine operations, dairy barns, poultry houses, and other livestock operations. The number of animals and the number of large production facilities in U.S.A. have significantly increased in the past 10 years.[1] Manure, as well as composts and biosolids, is a renewable resource and an excellent source of macro- (N, P, K, Ca, Mg, and S) and micronutrients (Zn, Cu, Fe, Mn, etc.), that are essential for growing plants. For centuries manure was used throughout the world for improving soil fertility and enhancing crop productivity. However, with the advent of synthetic fertilizers after World War II, manure was considered more a liability than a nutrient resource for crop production.

When animals are grazing on pastures and rangelands, manure is dispersed across a large area and little management is needed because the material is not concentrated and decomposes rapidly. However, when animals are concentrated in small feeding areas, the quantity of manure requiring proper management increases greatly. Significant amounts of manure are generated each year in U.S.A. from the confined feedlots of major livestock species (Table 1). The amount of N, P, and K present in the manure from these species would replace 25%, 25%, and 45% of the purchased N, P, and K fertilizers, respectively, in U.S.A., if utilized at agronomic application rates (Table 1). However, because of the high hauling cost, replacement of fertilizer is presently limited to specific areas in the country where the animal feeding operations are located. Crop producers are also reluctant to use manure because of factors such as hauling and spreading costs, potential introduction of weed seeds, nonavailability of manure where needed, uncertainty about availability of manure nutrients to plants, and problems of odor and application uniformity. The global numbers of major livestock species are given in Table 2.

Even though manure is an excellent source of multiple nutrients and organic matter, it can also contribute to water, air, and land pollution because of the potential for environmental loading with excess phosphorus, nitrate, salts, undesirable microorganisms, pathogens (disease-causing organisms), and greenhouse gases. Manure application in excess of crop needs can cause a significant buildup of P, N, trace elements (As, Cd, Pb, Hg, Mo, Ni, Cu, Fe, Mn, Se, and Zn), and salts in soils. Trace-element limits in soil are given in Table 3. The elevated P and N levels in soil are of environmental concern when these nutrients are carried by runoff to streams and lakes and cause "eutrophication," which is the nutrient

TABLE 1 Annual Manure, N, P, and K Generated by Animals Confined in Beef Cattle Feedlots, Dairy Barns, and Poultry and Swine Operations, and Fertilizer Use in U.S.A

Animal Species	Animals on Feed[a] (million)	Manure (Dry Weight) (million Mg)	N[b] (Mg × 1000)	P[b] (Mg × 1000)	K[b] (Mg × 1000)
Beef cattle	13.22	31.67	602	206	633
Dairy cows	13.14	22.01	782	140	522
Chickens (broilers and layers)[c]	8263.00	13.26	544	186	278
Turkeysc	284.00	3.09	142	65	65
Swine	59.41	15.15	709	451	709
Total		85.18	2779	1048	2207
1999 Fertilizer use in U.S.A.[d]			12436	4345	5016
1996 Global fertilizer use[d]			78353	13543	17516
(Manure nutrient/U.S. fertilizer use) ×100 (%)			25	25	45

[a] From United States department of agriculture.[1]

[b] Manure weight and N, P, and K contents taken from United States Department of Agriculture.[2]

[c] Yearly production numbers.

[d] From Fertilizer Statistics.[3]

TABLE 2 Global Numbers of Major Livestock Species in 1997

Animal Species	Animals[a] (million)
Cattle	1333
Chickens	14156
Sheep and goats	1754
Swine	837

[a] From Food and Agricultural Organization.[4]

TABLE 3 United States Environmental Protection Agency (40 CFR 503.13, revised July 1, 1999) Trace Element Limits, and Concentrations in Littleton/Englewood, CO Biosolids, July 27, 1999

Trace Element	Agronomic Rate Concentration Limit (mg/kg)	Ceiling Concentration Limit (mg/kg)	Annual Soil Loading Limit (kg/ha)	Cumulative Soil Loading Limit (kg/ha)	Littleton/ Englewood Biosolids (mg/kg)
Arsenic (As)	41	75	2.0	41	2.7
Cadmium (Cd)	39	85	1.9	39	5.6
Copper (Cu)	1500	4300	75	1500	256
Lead (Pb)	300	840	15	300	46
Mercury (Hg)	17	57	0.85	17	1.2
Molybdenum (Mo)	—	75	—	—	8.0
Nickel (Ni)	420	420	21	420	15
Selenium (Se)	100	100	5.0	100	4.6
Zinc (Zn)	2800	7500	140	2800	198

Note: Concentration and quantities are on dry weight basis.

enrichment of water that can promote algal growth and depletion of dissolved oxygen in water. This oxygen is essential for aquatic animals. Pathogens (such as bacteria, viruses, and parasites) in runoff from fields treated with manure can be another source of water pollution. Pathogens and odorous materials can also be carried by wind from the feeding operations to neighboring areas. Excess manure application can contaminate the groundwater with nitrate-N. Nitrate is a water-soluble ion that moves with water into the soil and can reach the groundwater within a few days after application. The U.S. Environmental Protection Agency (USEPA) has set a 10 mg NO_3–N/L standard for drinking water.

Compost

Composting is the aerobic decomposition of organic materials in the thermophilic temperature range of 40–65°C. The composted material should be an odorless, fine textured, low-moisture content material that can be bagged and sold for use in gardens, potting, and nurseries or used as a source of nutrients and organic matter on cropland with little fly-breeding potential. Other advantages of composting include improving the handling characteristics of any organic residue by reducing its volume and weight. Composting also has the potential to kill pathogens and weed seeds. Disadvantages of composting organic residues include loss of N and other nutrients during composting, the time taken for processing, cost of handling equipment, need for available land for composting, odors during composting, marketing, diversion of manure or residue from cropland, and slow release of available nutrients. Similar to manure or biosolids, composts can cause water, air, and land pollution if not used properly.

Temperature, water content, C : N ratio, pH level, aeration rate, and the physical structure of organic materials are important factors influencing the rate and efficiency of the composting process. Ideal values for these factors include a temperature of 54–60°C, C : N ratio of 25 : 1–30 : 1, 50%–60% moisture content, oxygen concentration > 5%, pH of 6.5–8.0, and particle size of 3–13 mm. The requirement set by the U.S. Environmental Protection Agency regulations for composting municipal waste is that the temperature should be maintained at 55°C or above for at least three days so as to destroy the pathogens. A temperature of 63°C within the compost pile is needed to destroy the weed seeds.

Homogeneous manure solids can be composted alone without mixing with bulk materials. Bulking agents are required to provide structural support when manure solids, or other organic residues, are too wet to maintain air space within the composting pile, and to reduce water content and/ or to change the C : N ratio. Dry and fibrous materials, such as saw dust, leaves, and finely chopped straw or peat moss, are good bulking agents for composting wet manure or organic residues. Depending on the ambient temperature, a complete composting process may take two to six months.

There are a number of methods for composting organic materials. These include active windrow (with turning), passive composting piles, passively aerated windrow (supplying air through perforated pipes embedded in the windrow), active aerated windrow (forced air), bins, rectangular agitated beds, silos, rotating drums, containers, anaerobic digestion, and vermicompost (using earthworms). Carcass composting can be done by using all types of animals. Mortality composting can be accomplished in backyard-type bins, indicator composter bins, and in temporary open bins using layers of saw dust or chopped straw and dead animals. Water content is an important factor to be considered when composting dead animals, and should be maintained at about 40–50%.

Biosolids

Treatment of municipal wastewater results in a mostly organic by-product known as “biosolids.” Land application of biosolids for beneficial use has been practiced since the early 20th century in U.S.A. The USEPA announced requirements regarding beneficial use of biosolids with promulgation of the 40 CFR503 regulations in February 1993. The USEPA and the state agencies that control land application of biosolids encourage the judicious recycling of biosolids on crop- or rangeland, as they contain essential plant nutrients and organic matter.

A key aspect of USEPA and Colorado Department of Health (CDH) regulations requires the application of biosolids at an agronomic rate. The CDH[5] defines agronomic rate as "the rate at which biosolids are applied to land such that the amount of nitrogen required by the food crop, feed crop, fiber crop, cover crop or vegetation grown on the land is supplied over a defined growth period, and such that the amount of nitrogen in the biosolids which passes below the root zone of the crop or vegetation grown to groundwater is minimized." The USEPA trace-element limits for land application of biosolids are shown in Table 3. State agencies that control biosolids recycling on land are required to adopt these limits as minimum requirements to protect the environment and public health. Risk assessment of different biological pathways served as the foundation for establishing the trace-element restrictions. For example, the concentrations for Littleton/Englewood biosolids shown in Table 3 indicate that it meets the agronomic-rate limits and can, therefore, be applied at an agronomic rate with minimal restriction. New, aggressive pretreatment programs have significantly reduced trace-element concentration in biosolids since about 1970, and therefore, environmental and public health risks are even more minimal.

The USEPA requires municipal wastewater treatment facilities to reduce pathogens and to reduce the attraction of insects and animals before applying biosolids to land. Most municipal wastewater treatment plants use heat and attack by beneficent micro-organisms through anaerobic (without air) or aerobic (with air) digestion to kill potential pathogens and reduce odors that may reside in wastewater. Municipalities accomplish further reduction of pathogens and stabilization by composting, drying, or other techniques.

The major reason that the USEPA promotes land application of biosolids is that the plant nutrients and organic matter can benefit the soil–plant agroecosystem. For example, Littleton/Englewood biosolids used at two research locations contained up to 5.0% organic–N, 1.3% ammonium–N, 140 mg/kg nitrate–N, 3.7% P, and 0.30% K. Biosolids can also provide plant micronutrients such as Fe and Zn. The organic carbon in biosolids can help to develop and stabilize soil structure with a concomitant increase in precipitation capture and decrease in soil erosion. Efficacious land application of biosolids changes the perspective from disposal of a waste (i.e., a nuisance) to recycling a valuable resource (i.e., a beneficial process).

Conclusions

Organic residuals (wastes)can serve as excellent sources of plant nutrients such as N, P, and micronutrients such as Fe and Zn. Proper management is required, however, to match nutrient amounts supplied by the organic materials with crop needs so as to avoid potential pollution problems.

References

1. United States department of agriculture. In *Agricultural Statistics*;; Government Printing Office: Washington, DC, 2000.
2. United States department of agriculture. In *Agricultural Uses of Municipal, Animal, and Industrial Byproducts*; Conserv. Research Report No.44; Government Printing Office: Washington, DC, 1998.
3. *Fertilizer Statistics*; The Fertilizer Institute: Washington, DC, 2001; Available at http://www.tfi.org.
4. Food and Agricultural Organization. In *Statistical database*; United Nations: New York, 2001; Available at http://apps. fao.org.
5. Colorado Department of Health. In *Biosolids Regulation 4.9.0;* 1993, 46 pp.

50

Manure Management: Dairy

H.H. Van Horn
and D.R. Bray

Introduction

Water use is essential for all dairies. Drinking water is indispensable for cattlelife; some amount of water is necessary for cleaning and sanitation procedures; moderate amounts are important during periods of heat stress for evaporative cooling of cows to improve animal production and health; additional amounts can be used in labor-saving methods to move manure and clean barns by flushing in properly designed facilities; and the recovered wastewater can be recycled to supplement water requirements of forage crops grown to meet roughage requirements of the dairy herd. Extensive water use, however, increases the potential of surface runoff and its penetration into the ground with possible environmental impacts offsite. Heightened environmental concerns and the need for resource conservation, in many cases, have caused implementation of water-use permits. Thus, it is important to determine various essential uses of water, other uses that are important to management, and also consider whether reuse of some water is possible and if it is necessary to do so.

Some of the useful unit conversions are listed as follows:

1 gal of water = 8.346 lb.
1 ft^3 of water = 7.48 gal.
1 acre = 43,560 ft^2.
1 acre in. of water = 27,152 gal

Calibration methods to estimate use: Water flow meters should be installed on major water supply lines. If water meters are not in place to measure gallons pumped, it becomes necessary to estimate the usage. This can be achieved by capturing flow through various water lines for specified times and multiplying by the time the water flows through these lines every day.

Drinking

Table 1 provides estimates of drinking water requirements in gallons per cow per day. Consumption of 25–30 gal of water per day by lactating cows is common, which varies depending on milk yield, dry matter intake (DMI), temperature, and other environmental conditions.[1]

Cow Washing

Presently most dairies, in warm climates, bring cows to be milked into a holding area equipped with floor-level sprinklers, which spray water upward to wash cows. Each cow usually has a holding area of about 15 ft^2 and are typically washed for 3 min. Amount of water used per cow should be calculated for each dairy. An estimate for conservative use is that a holding area for 300 cows is 30 × 150ft^2 (15 ft^2 per cow) and is equipped with sprinklers with 5-ft spacing (say 7 across and 30 rows) having 210 sprinklers. If each sprinkler applies 5 gal min^{-1}, total usage is 1050gal min^{-1} or 3150 gal for 3 min, the average consumption per cow would be 3150/300 = 10.5 gal per cow per wash cycle. If cows are milked three times this would require 31.5 gal per cow per day.

The washing system previously described also helps in cooling of cows while they are crowded together waiting to be milked. However, the cooling effect could be achieved by sprinkling a little amount of water from above, alternatively with fans to give evaporative cooling, if cows were clean enough so that extensive washing was not required and water conservation was necessary.

Washing Milking Equipment and Milking Parlor

Use of water for these purposes is not as directly related to the number of cows as for other uses. For washing milking equipment, a common wash vat volume is 75 gal. If this is filled for rinse, wash, acid rinse, and sanitizing at each of three milkings, this amounts to 900 gal for the herd, e.g., with 300 cows, only 3 gal per cow per day. This is an extremely small component of the total water budget. The amount used to wash out the milking parlor varies largely. If only hoses are used, the amount may be as little as 2 gal per cow per milking or 6 gal per cow per day if cows are milked three times daily. If flush tanks are used, the amount may be more, i.e., nearly 3000gal per milking or 9000 gal day^{-1} for three times, equivalent to 30 gal per cow per day for a 300-cow system.

TABLE 1 Predicted Saily Water Intake of Dairy Cattle as Influenced by Milk Yield, DMI, and Season[a,b]

	Cool Season (e.g., February)		Warm Season (e.g., August)	
Milk Yield (lb)	DMI (lb)	Water Intake (gal)	DMI (lb)	Water Intake (gal)
0	25	11.5	25	16.3
60	45	22.2	44	26.8
100	55	28.6	48	31.9

[a] Drinking water intake predicted from equation of Murphy et al., J. Dairy Sci., 1983, 66, 35: Water intake (lb day^{-1}) = 35.2 × DMI (lb day^{-1}) + 0.90 × milk produced (lb day^{-1}) + 0.11 × *sodium* intake (g day^{-1}) + 2.64 × weekly mean minimum temperature [°C = (°F −32) ×5/9]. For examples above, diet dry matter was assumed to contain 0.35% Na. Predicted water intakes (lb) from formula calculations were divided by 8.346lb water per gal to convert to gallons.

[b] Average minimum monthly temperatures for February (43.5°F) and August (71°F) used with prediction equation were 70 years averages for specified months at Gainesville, FL (Whitty et al., Agronomy Dept, Univ. FL, 1991).

Sprinkling and Cooling

Sprinklers along with fans are used for evaporative cooling to relieve heat stress in dairy cows during hot periods of the year. Their use has shown increased cow comfort (lowered body temperature and respiration rates) and economic increases in milk production and reproductive performance.[2,3] Application rates used by dairymen vary. Florida experiments compared application rates of 51 gal per cow per day, 88 gal per cow per day, and 108 gal per cow per day at 10 psi in one experiment and 13 gal per cow per day, 25 gal per cow per day, and 40 gal per cow per day in another experiment. The application rate, 13 gal per cow per day, is close to the estimated evaporation rate from the cow and surrounding floors. This component should be considered in water use but not in runoff water that must be managed in the manure management system. We estimate 25 gal per cow per day as the minimum practical application rate in order to get adequate coverage of cows to cool them because often they are not in the sprinkled area. Total application days per year vary from 120 days to 240 days. A separate water well, or reserve tank and booster pump, may be needed to supply short-term high demand required by the sprinkler system.

Flushing Manure

Flushing manure can be made a clean and labor-saving process, if facilities include concrete floors with enough slope so that water flow propelled by gravity could be used to move manure. Amounts of water used per cow vary widely depending on size and design of facilities and frequency of flushing. However, usually a flush of about 3000 gal is required to clean an alley width of 10–16 ft. If 4 alleys are common for every 400 cows and alleys are flushed twice daily, this would amount to an average use of 60 gal per cow per day. Many dairies use more flushings per day.

Recycling Dairy Wastewater through Irrigation of Forage Crops

Most often nitrogen is the nutrient on which manure application rates are budgeted. To maximize nutrient uptake, crop growth should be as vigorous as possible. This requires irrigation during most of the year in many dairy regions for the disposal of flushed wastewater. In southern regions, multiple cropping systems are possible, which will recycle effectively nitrogen excretions from 100 cows on a sprayfield or manure application field of about 30 acres.[4]

Tentative estimates of total water needs of the growing crops in warm climates average about 1.75 in. of water per week (0.25 in. per day) from irrigation plus rainfall with a minimum of 0.5 in. per week tolerated even in rainy season on sandy soils.[5,6] Table 2 provides estimates of water requirements for two triple cropping forage systems that are common in southern climates. In sandy soils that hold only about 1.0 in. of water per foot of soil depth, some amount of rainfall cannot be stored. Therefore, even in heavy rainfall seasons, judicious irrigation is often needed during lower rainfall weeks. Limited data are available on the maximum amount of water that could be applied and not reduce yield or quality of forage and not result in pollution of groundwater with nitrates and other minerals. However, the maximum probably is at least 35–45 in. per year above the acre totals in Table 2.

Rainwater from Roofs and Concrete Areas

Rainwater entering wastewater holding areas can be significant. For example in the dairy representing typical minimum water usage with a flush system in southeast United States (Table 3), the net accumulation during the hot season was calculated as follows: assumed wastewater holding area is 1 acre surface area per 100 cows, net rainfall accumulation in holding area is 3 in. more than evaporation per month, concrete areas and/or undiverted roof areas that capture rainfall are 15,000 ft^2 per 100 cows that divert 15,000/43,560 ft^2 per acre of the 3 in. to the wastewater holding facility. Thus, 3 in. + $0.344 \times 3 = 4.03$ acre in. mo^{-1} or essentially 1.0 acre in. per week per 100 cows (approximately 27,000 gal per 100 cows).

TABLE 2 Crop Yield and Water Requirement Estimates for Two Triple Cropping Forage Systems

		Silage Yield				Water Required		
Crop No.	Name	Ton/A 35% DM	Ton/A DM	lb/A DM	lb/lb DM	lb/A Total	gal/A Total	A-in. Total
1	Wheat	10	3.5	7000	500	3,500,000	419,362	15.4
2	Corn	24	8.4	16,800	368	6,182,400	740,762	27.3
3	Corn	14	4.9	9800	368	3,606,400	432,111	15.9
	Total	48	16.8	33,600		13,288,800	1,592,235	58.6
1	Rye	10	3.5	7000	500	3,500,000	419,362	15.4
2	Corn	24	8.4	16,800	368	6,182,400	740,762	27.3
3	F. Sorghum	18	6.3	12,600	271	3,414,600	409,130	15.1
	Total	52	18.2	36,400		13,097,000	1,569,254	57.8

[a] A = acre; No. = number; DM = dry matter.

Developing a Water Budget

A wide range exists in water usage on dairy farms. For most dairy waste management systems designed to utilize flushed manure nutrients through cropping systems grown under irrigation, water amounts are small in relation to irrigation needs for crop production. Costs for construction of storage structures for holding wastewater until used for irrigation warrant consideration. For example, water-use budgets given in Table 3 show that water usage is small in comparison to irrigation needs when there are 30 acre of sprayfield crop production per 100 cows. Conversely, the amounts used in most dairy systems would be large and unmanageable if application through irrigation is not an option or if less acreage for irrigation is available than needed for application of all manure nutrients.

If a dairy does not have acreage available close by to utilize manure nutrients and water through an environmentally accountable sprayfield application system, it would be necessary to export nutrients off the farm, preferably as solid wastes to avoid excessive hauling or pumping costs. If the water and manure nutrients cannot be used through irrigation, a non-flush system should be utilized. However, usually some irrigation is possible, permitting dairymen to use cow washers and limited flushing if they scrape and haul manure from some areas.

Strategies to minimize water usage: Table 3 presents one column indicating a theoretical minimum amount of water use in a dairy. This system implies that cows are clean and cool enough so that sprinkler washers are not required to clean and cool cows while being held for milking. In addition, it is assumed that all of the manure is scraped and hauled to manure disposal fields or transported off the dairy in some other fashion. Intermediate steps that might be taken include the following:

1. Scraping and hauling manure from high use areas such as the feeding barn so that this manure can be managed off the dairy.
2. Using wastewater rather than fresh water to flush manure from feeding areas and freestall barns.
3. Using a housing system that will keep cows clean enough so that cow washers are not required to clean cows before milking. This system, however, may require use of alternating sprinklers and fans to keep crowded cows cool during hot weather conditions.

If flushing is desired in conjunction with scraping and hauling from heavy use areas, perhaps the feeding area could be flushed with recycled water after scraping to clean the area. These procedures would reduce total nutrient loads retained in wastewater and would significantly reduce the size of the sprayfield needed for water and manure nutrient recycling.

TABLE 3 Estimated Water Budgets for Three Example Dairies

	Flush Systems			
Water Use in the Dairy	Typical Need during Hot Season	Common Usage on Some Dairies	Non-Flush Theoretical Minimum	Worksheet for Your Dairy
Drinking (cows)	25	25	25	
Cleaning cows	32	150	0	
Cleaning milking equipment	3	5	3	
Cleaning milking parlor	30	30	6	
Sprinklers for cooling	25	130	12	
Flushing manure	60	80	0	
Total use per cow per day	175	400	46	
Total use per 100 cows per day	17,500	40,000	4600	
Use per 100 cows per week	122,500	280,000	32,200	
Water in milk per 100 cows per week	4500	4500	4500	
Estimated evaporation (at 20% of use)	24,500	56,000	6440	
Average rainfall and watershed drainage into storage facility per 100 cows per week	27,000	27,000	13,000	
Wastewater produced from 100 cows/week	120,500	246,500	38,760	
Acre in. per 100 cows per week	4.44	9.08	1.43	
in. per week if 30 acre in sprayfield	0.15	0.30	0.05	

[a] All values are in gal unless otherwise noted.

Example calculations (column 1): Total use per cow per day = 175 gal; total use per 100 cows per week = 122,500 gal less 4500 in milk and 24,500 gal evaporation = 93,500 gal week^{-1}; net rainfall and watershed drainage to storage per 100 per cows per week = 27,000; acre in. per 100 cows per week = (93,500 + 27,000)/27,152 gal per acre in. = 4.44.; if 30 acre were in sprayfield, 4.44/30 = 0.15 in. week^{-1}.; if crop needed 1.75 acre in. week^{-1} (a common average), a total of 1.75 in. × 30 acre × 27,152 gal per acre in. = 1,425,480 gal is needed of which only 120,500 gal (8.5%) would come from dairy wastewater. The remaining (91.5% of total) would have to come from rainfall or fresh irrigation water.

References

1. Beede, D.K. Water for dairy cattle. *Large Dairy Herd Management*; American Dairy Science Assoc.: Champaign, IL, 1992; 260–271.
2. Bray, D.R.; Beede, D.K.; Bucklin, R.A.; Hahn, G.L. Cooling, shade, and sprinkling. *Large Dairy Herd Management*; American Dairy Science Assoc.: Champaign, IL, 1992; 655–663.
3. Van Horn, H.H.; Bray, D.R.; Nordstedt, R.A.; Bucklin, R.A.; Bottcher, A.B.; Gallaher, R.N.; Chambliss, C.G.; Kidder, G. Water Budgets for Florida Dairy Farms; Circular 1091; Florida Coop. Ext., Univ. Florida: Gainesville, 1993.
4. Van Horn, H.H.; Nordstedt, R.A.; Bottcher, A.V.; Hanlon, E.A.; Graetz, D.A.; Chambliss, C.F. Dairy Manure Management: Strategies for Recycling Nutrients to Recover Fertilizer Value and Avoid Environmental Pollution; Circular 1016; Florida Coop. Ext., Univ. Florida: Gainesville, 1998; 1–24.
5. North Florida Research and Education Center. AREC Research Report 77–2; IFAS, University of Florida, Gainesville, 1977.
6. Wesley, W.K. Irrigated Corn Production and Moisture Management; Bul. 820; Coop. Ext. Serv., Univ. Georgia College of Agric. and USDA, 1979.

51

Manure Management: Phosphorus

Rory O. Maguire,
John T. Brake, and
Peter W. Plumstead

Recent Progress in Nutrient Management Practices

With increasing human population and awareness of environmental problems associated with pollution of rivers and streams with P, the end of the 20th century saw a refocusing of nutrient management strategies from increasing production toward environmental protection, in developed countries.[1] Initially, nutrient management strategies were aimed at avoiding overapplication of nutrients and voluntary best management practices were developed. However, manure still tended to be applied at a rate to supply crops with sufficient N, which typically leads to overapplication of P, as manures have N: P ratios of around 2: 1 to 4: 1 (depending on species) while crop requirements have an N: P ratio of 8: 1. When P is overapplied to soils, it builds up the soil test P (STP) concentration and over a period of many years, this leads to STP rising to levels in excess of crop requirement.[2] Soil test P in excess of crop requirement is undesirable, as studies have shown that when STP rises above agronomically optimum values P losses to rivers and streams increase to the detriment of water quality, [2] and it can also lead to a negative effect on some micronutrients, e.g., Zn.

Nutritional Strategies to Reduce the Manure Phosphorus Concentration

In the past, dietary P levels frequently included a safety margin that allows for any potential variation in both the requirement of the animals and the degree of dietary P utilization, with little attention paid to the resultant manure P concentration. However, recent research in broilers and broiler breeders demonstrates that levels of inorganic P, supplemented to both broiler and broiler breeder diets may be considerably reduced relative to minimum requirements published by the National Research Council without an adverse effect on performance.[3,4]

When considering strategies to increase P digestion and retention, species can be split into two categories according to their ability to digest phytate-P: ruminants (cattle, sheep) can digest phytate-P, while monogastric species (poultry, swine) can digest little. As approximately 66% of the P in corn and soybean is in the form of phytic acid, and these ingredients frequently constitute about 60%–90% of the diet, up to 80% of the total fecal P output may originate directly from corn and soybean meal in poultry diets.[5] The primary strategy to address the low digestibility of phytate-P in monogastric species has been the application of exogenous phytase enzymes to these diets and is done on the basis of replacing 0.1% of available P from supplemental calcium phosphate with a more or less standard amount of enzyme. Alternate strategies other than the use of phytase include the development of low phytate variants of corn and soybean cultivars, because these possess similar, if not greater, potential in reducing manure P excretion. However, the commercial application of these new cultivars has to date been limited as a result of the practical complications arising from the identity preservation of these ingredients from harvest to point of feeding.

Impact of Diet Modification on Nutrient Management

The total P and soluble P concentrations in manures have been shown to be affected by dietary P concentration, P contributing ingredients, and dietary phytase supplementation. Total P is important because applications of manure total P contribute to long-term changes in STP. Once optimum STP for crop production is attained, it is beneficial to balance manure total P applications with crop removal of P. This will maintain crop production without raising STP to levels of environmental concern; although in some soils, P can revert to less available forms so P availability needs to be monitored with soil testing. Applications of soluble P in manure have been correlated to soluble P losses in runoff from manure-amended soils, so minimizing the soluble P concentration in manure is important when P losses to rivers and streams are of concern. Reducing the P concentration in animal feeds (high vs. low nonphytate P diets) and using phytase (phytase vs. no phytase diets) can reduce the total and soluble P concentrations in poultry wastes (Figure 1).[6] It is estimated that using currently available technology, dietary amendment can potentially de-crease the total P concentration in poultry by 40%, swine by 50% and dairy by 30%, while future advances could raise these figures to 60% for both poultry and swine.[7] Research on the impact of dietary amendments on soluble P in manure and P losses in runoff from manure-amended soils is at an early stage; however, trends are starting to appear. Supplying P at an adequate level to satisfy animal requirements is generally achieved through a reduction in the added inorganic dietary calcium phosphate component, and results in a reduction in both total and soluble P in the manures generated. While most studies show that feed additives such as phytase reduce manure total P while having little effect on soluble P, this effect is very dependent on the level of dietary P relative to the requirements of the animals.[6,7] Phytase supplementation to diets containing P in excess of animal requirements, or without an adequate reduction in the P from supplemental calcium phosphate, will result in a potential increase in the water soluble P when expressed on a percentage basis.[5]

Traditionally, manure application rates to cropland have been decided by individual farmers. However, increasing concerns over the impact of the land application of nutrients in excess of crop requirement on water quality in rivers and streams, have led to increasing regulation of land application of manures in many developed countries. These regulations are primarily based on N and P application rates that are often linked to problems identified in localized areas and are based on the current levels of soil N and P, and the annual rate of removal of these nutrients through their incorporation into harvested crops. As forage crops typically have an N : P ratio of 8 : 1 and manure contains N : P ratios of approximately 2 : 1 to 4 : 1, nutrient management regulations based on P application may require up to four times more cropland than previous N-based application guidelines. The limitation of available land in areas proximate to intensive livestock production can cause problems associated with what to do with the manure if it cannot be land applied, as other uses of manure are currently fairly limited and transport costs limit the potential transportation of the wastes over any appreciable distance. If nutrient management regulations are based on P, then reducing the P concentration in manures could greatly alleviate these problems. For example, if one could reduce the total P concentration in a manure by 40%, then application of 40% more manure by weight to meet the same total P application limit would be possible. This may eliminate problems with excess manure P in some areas of intensive animal production, but may not be sufficient for other areas that already have high STP.

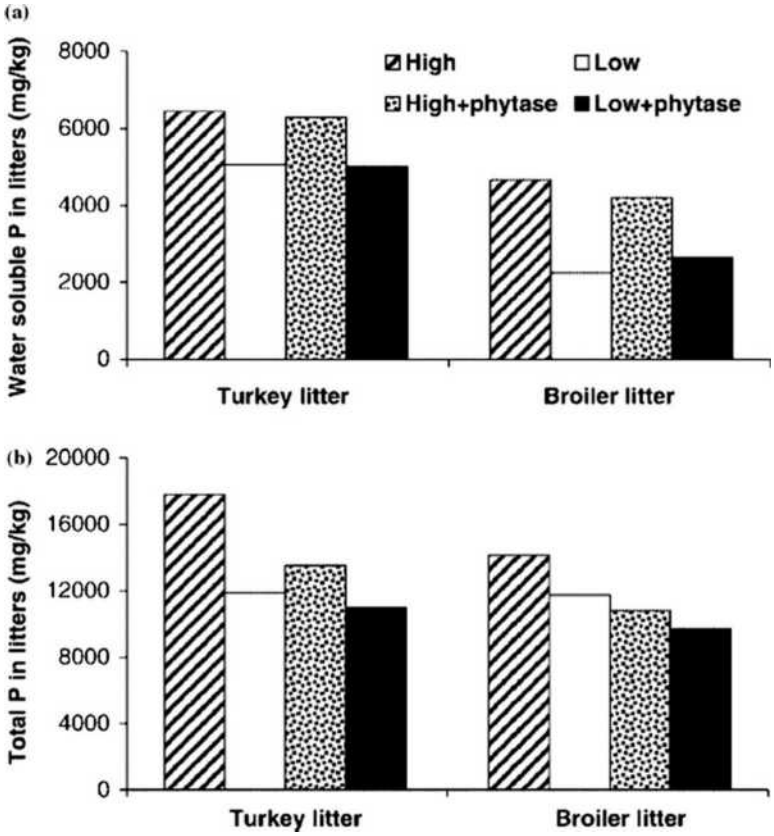


FIGURE 1 Effects of diets high and low in dietary nonphytate P, and the impact of dietary phytase on (a) water soluble P in litters, and (b) total P in litters produced.
Source: Maguire et al.[6]

Conclusions

Dietary amendment to reduce the P concentration in feeds and hence manures generated will help address concerns over the fate of manure nutrients in areas of intensive animal production. If manures have the same N concentrations but reduced P concentrations, then these manures will come closer to meeting crop N needs without oversupply of P. However, dietary amendment alone will not solve the problem of excessive manure P in all areas of animal production. Dietary amendment will, however, certainly help alleviate the current situation where applying animal manures to meet crop N requirements leads to overapplication of P, buildup of STP, and eventually increased losses of P to rivers and streams.

References

1. Beegle, D.B.; Carton, O.T.; Bailey, J.S. Nutrient management planning: justification, theory, practice. J. Environ. Qual. **2000**, *29*, 72–79.
2. Sims, J.T.; Edwards, A.C.; Schoumans, O.F.; Simard, R.R. Integrating soil phosphorus testing into environmentally based agricultural practices. J. Environ. Qual. **2000**, *29*, 60–71.
3. National research council. In *Nutrient Requirements of Poultry,* 9th revised ed.; National Academy Press: Washington, DC, 1994.

4. Brake, J.; Williams, C.V.; Lenfestey, B.A. Optimization of dietary phosphorus for broiler breeders and their progeny. In *Nutritional Biotechnology in the Feed and Food Industries,* Proceedings of Alltech 19th International Symposium; Lyons, T.P., Jacques, K.A., Eds.; Nottingham University Press: Nottingham, U.K., 2003, 77–84.
5. Brake, J.; Plumstead, P.; Gernat, A.; Maguire, R.; Kwanyuen, P.; Burton, J. Reducing phytate versus adding phytase to reduce fecal phosphorus. In *Session PCP 3: Phytate in Soybeans and Related Environmental Concerns*, Proceedings of the 95th AOCS Meeting, Cincinnati, OH, 2004, 118.
6. Maguire, R.O.; Sims, J.T.; Saylor, W.W.; Turner, B.L.; Angel, R.; Applegate, T.J. Influence of phytase addition to poultry diets on phosphorus forms and solubility in litters and amended soils. J. Environ. Quality **2004**, *33*, 2306–2316.
7. CAST (council for agricultural science and technology). In *Animal Diet Modification to Decrease the Potential for Nitrogen and Phosphorus Pollution.* Issue Paper No. 21; CAST: Washington, DC, 2002.

52

Manure Management: Poultry

Shafiqur Rahman
and Thomas R. Way

Introduction

The U.S. Environment Protection Agency (EPA) includes a number of species, including chicken, turkey, and duck, in their definition of poultry. In the United States, approximately 8.55 billion broilers, 337 million laying hens, and 247 million turkeys were raised in 2009.[1] The total revenue from the poultry industry in 2008 was $35.9 billion, which was up 11% as compared with 2007 ($32.2 billion). Of the combined total revenue, the broiler, layer, and turkey industries contributed 64%, 23%, and 12%, respectively. Modern poultry production occurs primarily in confined facilities, which allows larger intensive production in smaller areas. Owing to intensive poultry production systems, large amount of manure (i.e., feces, urine, undigested feed, spilled water, bedding used in poultry houses, etc.), is produced in a smaller area, which leads to air pollution (including odors and pollutant gas emissions) and water pollution (i.e., eutrophication) problems. It is the producer's responsibility to manage manure in an environmentally safe way. Therefore, poultry manure management is the focus of this entry.

Manure Production and Management System

Table 1 lists the total number of broilers, layers, and turkeys produced and the estimated manure production by bird type. A large quantity of manure or litter is produced daily, which needs to be managed properly to maximize agronomic benefits (i.e., nutrient values) and minimize environmental concerns (i.e., water and air pollution). Poultry manure management is composed of manure collection and handling, storing, and utilization. Manure collection and handling depend on manure properties, and they are influenced by bird type, bird age, diet, bird productivity, and management system.[2] Manure can be handled as a solid (>20 total solids), semisolid (10%–20% total solids), slurry (4%–10% solids), or liquid (<4% solids).[3] Solid manure can be stacked, while liquid manure can be stored in earthen anaerobic lagoons or storage structures (e.g., concrete or steel). For economical reasons, liquid manure is typically stored in anaerobic lagoons.

TABLE 1 Poultry Manure Production, as Excreted

Bird Type	Live Market Weight (kg)[a]	Manure per 1000 Birds/Day (kg)[a]	Total Number of Birds (1000s)[b]	Total Manure Production/Day (tons)
Broiler	2.0	80	8,550,500	684,040
Layer	1.8	118	337,376	39,810
Turkey (tom, light)	10.0	267	249,914	66,727
Total manure production				790,577

[a] From Poultry Waste Management Handbook, NRAES-132.[4]
[b] From National Agricultural Statistics Service.[1]

In broiler and turkey facilities, bedding or litter materials (e.g., sawdust, wood shavings, peanut hulls, rice hulls, sunflower hulls, etc.), are used, and litter is handled as a solid (Figure 1). Litter is sometimes completely removed after each flock or partially cleaned between flocks when litter is wet and packed around feeders and waterers. After partial cleaning, a thin layer of fresh bedding material is added.[2] It is also common that producers decake litter between flocks using a decaking implement to remove the

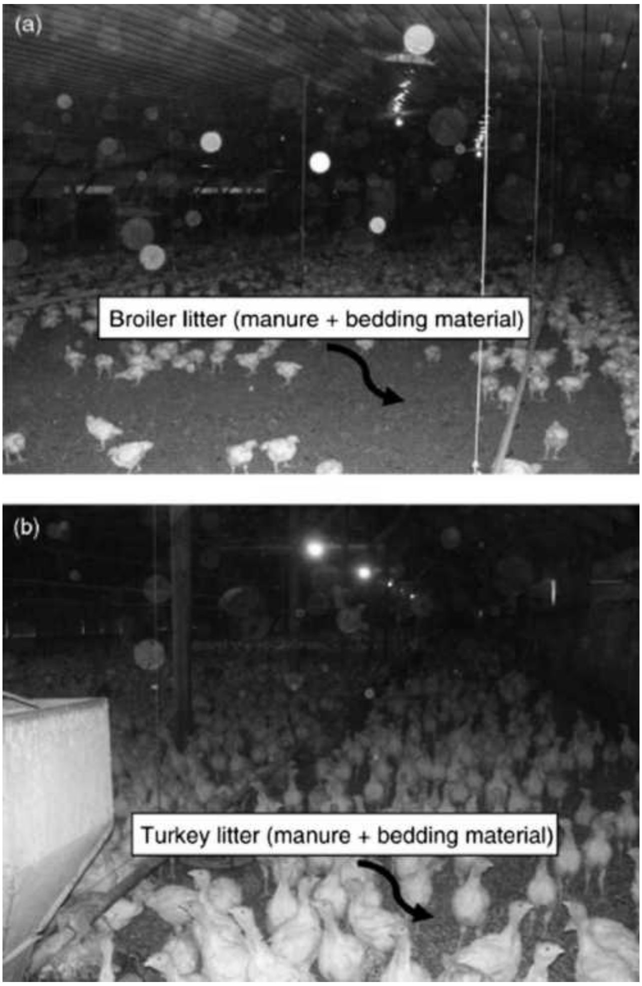


FIGURE 1 Poultry production facility and litter for (a) broiler and (b) turkey.

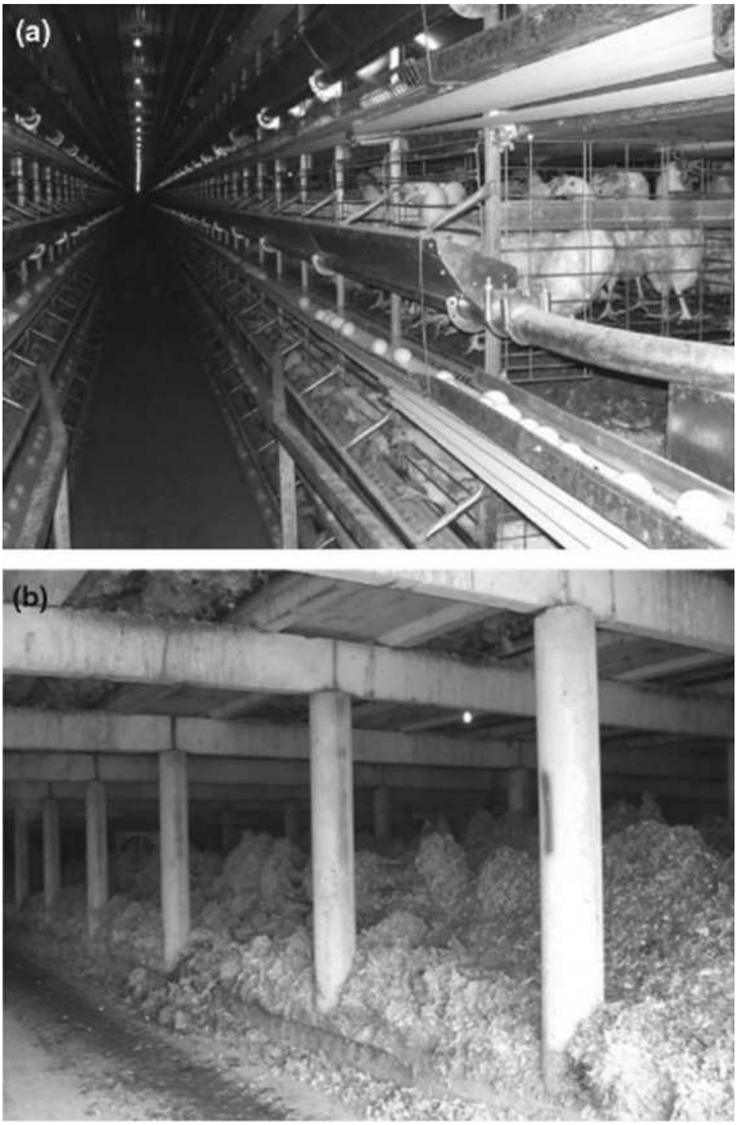


FIGURE 2 (a) High-rise layer house and (b) manure storage pit (stockpiled beneath cages).

larger clumps of litter while leaving the smaller clumps as bedding for the next flock. Depending on the management practices, litter cleanout varies significantly from farm to farm.

High-rise layer facilities (elevated cages) (Figure 2a) produce solid manure, and that manure is stored in the portion of the poultry house beneath the cages, as shown in Figure 2b. Manure is removed once or twice a year depending on individual farm management practices and regulatory requirements.

Some layer facilities with a shallow pit remove manure two or three times per week using a scraper (Figure 3a) and may produce semisolid or liquid manure and store it in an earthen anaerobic lagoon or outdoor storage (Figure 3b). However, manure cleaning or removing frequency may vary depending on the storage capacity, manure storage conditions (dry or wet), and environmental (indoor air quality) conditions.

Manure Storage

Manure or litter can be stored both indoors and outdoors depending on manure type and management practices. Both options have some pros and cons as discussed below.

Indoor Storage

Most of the high-rise layer houses (Figure 2a) stockpile manure directly beneath the birdcages, i.e., deep-pit storage, as shown in Figure 2b. This option minimizes runoff from precipitation, and manure can be stored for a longer time to allow scheduling of land application around crop nutrient requirements.[5] Although this option minimizes water pollution from runoff, it may cause air quality issues with the house, if not managed properly. Typically, high-rise layer houses are cleaned of manure once or twice a year; however, a manure storage pit should be inspected daily for any sign of water leakage. Ammonia (NH_3) and hydrogen sulfide (H_2S) emissions from indoor stockpiled manure may cause poor air quality and may affect bird health and productivity as well as worker health and safety. It is important to maintain recommended ventilation rates and to keep manure as dry as possible to minimize anaerobic conditions in manure stockpiles. Fans are commonly used in the pit to keep the manure pile relatively dry. If any portion of the manure stockpile is wet from water leakage, it is important to promptly remove manure from indoors and correct any water leakage problem. Typically, wet manure will require more frequent removal than dry manure.

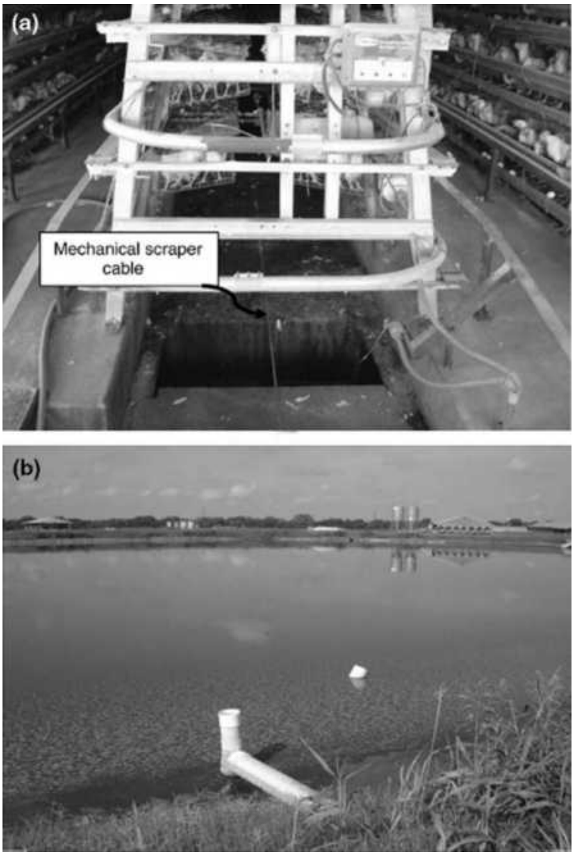


FIGURE 3 (a) Layer house with mechanical scraper and (b) anaerobic earthen lagoon system.

Outdoor Storage

Poultry litter should be removed from the house between flocks. The litter should be stored outside if no land is available for immediate manure application or if there is insufficient indoor storage room for a period of 6 to 9 mo. If solid manure is stored outside, proper management is necessary to minimize runoff and groundwater pollution. Outdoor stockpiles should be at least 30 m from the nearest surface water and should be separated from groundwater by an impervious soil layer having a minimum thickness of 1.2 m. To minimize runoff resulting from rainfall and to reduce nutrient losses, a cover on top of the litter pile is required to limit manure exposure to air. If regulatory guidelines and time permit, and if cropping conditions are appropriate, outside stockpiled manure should be applied to cropland as soon as possible.

Liquid manure or slurry is removed from under caged birds by mechanical scrapers (Figure 3) or mechanical belts or flushed with water and stored in an outdoor anaerobic lagoon or containment pond (Figure 3b). Manure is typically stored in an anaerobic earthen lagoon for a period of 6 to 9 mo and applied to cropland near or during the growing season to optimize use of manure nutrients. However, during manure pumping, it is very important to maintain the minimum design volume or treatment volume of an anaerobic lagoon to provide a dual function of storing and treating the manure.[6] Otherwise, anaerobic lagoons may be a significant source of nuisance odor.

Manure Utilization

Poultry manure contains essential nutrients (e.g., nitrogen, phosphorus, potassium, etc.), and can be applied to land as fertilizer to meet crop nutrient requirements. However, when land application of poultry manure as fertilizer is not an environmentally and economically sound option, manure (especially solid manure) can be composted (Figure 4), and the compost can be used as a fertilizer source for crops, gardens, and nurseries. In that case, nuisance odor can be minimized, and nutrients become bound in organic forms. However, for proper composting, several parameters (carbon-nitrogen ratio, pile moisture content, pile temperature, oxygen concentration, etc.), need to be maintained appropriately. Otherwise, anaerobic conditions may prevail, and the pile may become a source of nuisance odor.

FIGURE 4 Turkey litter composting pile and temperature sensors.

During storage, nutrient losses are a normal phenomenon. A portion of nutrient losses occurs during land application depending on the application method, such as surface vs. subsurface injection. Typically, most solid manure is applied on the surface, often followed by incorporation. Solid manure is commonly applied using truck-mounted spreaders (Figure 5). There are many options for solid manure spreaders, including side delivery and rear delivery (spinner spreader; single or double horizontal beater; vertical beaters). With a single beater, it is difficult to have uniform manure distribution to take full advantages of nutrients in manure, and it is difficult to break litter cake or large clumps. Researchers[6] concluded that distribution of manure across the application swath of a spreader needs to be relatively uniform to take full advantage of the fertilizer value of solid manure. To overcome this issue, double vertical beaters might be a better option (Figure 5b and c).

Traditional surface application of manure may cause nutrient losses, nuisance odor, and nutrient runoff, and reduces nutrient availability. Although incorporation will minimize some of these concerns, delayed incorporation (more than 24 hr) can result in increased nuisance odor, nitrogen nutrient losses through volatilization, and surface runoff.

To overcome problems associated with surface application of solid manure, manure can be subsurface banded. Until recently, subsurface application has been available for only liquid manure or slurry. Research has been conducted in both Canada and the United States to develop implements for subsurface band application of solid manure from poultry and feedlots.[7–9] Applying the manure in a controlled way decreases odor and greenhouse gas emissions while improving air quality and the social acceptance of the practice.[7] The U.S. Department of Agriculture–Agricultural Research Service (USDA-ARS) National Soil Dynamics Laboratory (Auburn, Alabama) has developed a four-row prototype implement for subsurface band application of poultry litter (Figure 6) and has used the implement to apply poultry litter in row crops and pastures.

The implement is equipped with conveyors, one for each trencher. Each conveyor drops the litter into a trencher, where the litter falls by gravity down into the trencher. The band spacing is adjustable from 0.25 to 1.0 m in increments of 25 mm. The implement is capable of applying poultry litter in a side-dressing manner to row crops, and the band spacing is then typically equal to the crop row spacing. Also, the implement is capable of applying litter to pastures, and typical band spacings that have been used for pastures are 250 to 380 mm. One disadvantage of the implement is that litter clumps measuring larger than 25 mm across have to be run through a hammer mill before being loaded into the hopper. Also, the litter moisture content should be <40% (wet basis). The implement is being modified to include on-the-go grinding as part of the machine. Subsurface banding of broiler litter to cotton with this implement has been shown to increase lint yield relative to surface broadcast application of litter.[10]

Subsurface band application of broiler litter to tall fescue and bermudagrass pastures using band spacings of 250 and 380 mm was found to produce forage yields equivalent to those for surface broadcast application of litter.[11] When broiler litter has been subsurface banded in a pasture with the implement, and rainfall has been applied using rainfall simulation, concentrations of phosphorus and nitrogen in runoff water have typically been reduced by 80%–95% relative to those for surface broadcast application of litter.

In contrast, liquid manure can be applied on both the surface and subsurface, and technologies are improving very fast. Injection of liquid manure limits the exposure of manure to the surface water and air, resulting in reduced nutrient losses and nuisance odor. For a liquid manure application system, a liquid storage tank with a flush spreader or injector assembly is used. However, a fully loaded tank may cause soil compaction, especially in wet clay soil. To overcome compaction issues, manure can be injected using a drag-hose system (Figure 7a). One new technology, namely, AerWay SSD® (Holland Equipment Inc., Norwich, Ontario, Canada), can place liquid manure in the active root zone and can substantially reduce any risk of groundwater contamination (Figure 7b).

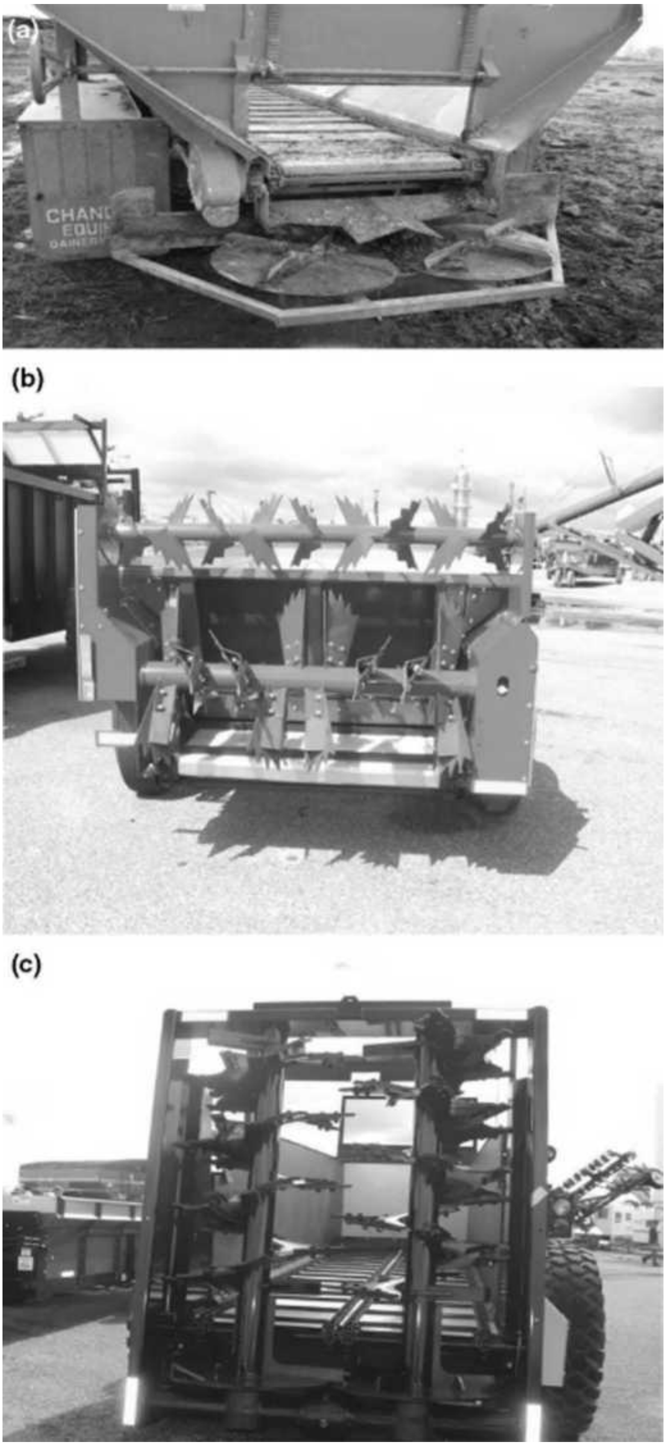


FIGURE 5 Rear delivery manure spreaders: (a) spinner spreader, (b) double horizontal beater, and (c) double vertical beater.

FIGURE 6 Side view of prototype subsurface band applicator implement for poultry litter developed by USDA-ARS at Auburn, Alabama.
Source: Dr. Thomas Way.

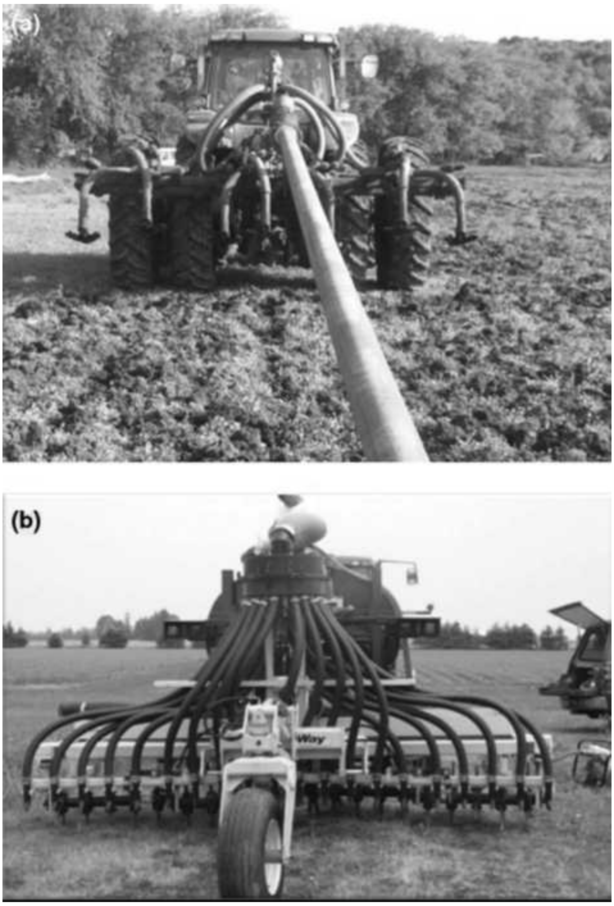


FIGURE 7 Liquid manure injection system: (a) drag-hose system and (b) AerWay SSD system.

Energy Uses

As of April 2010, there are 151 anaerobic digester systems operating at commercial livestock farms in the United States. Of these, only three anaerobic digester systems were using poultry manure. Poultry manure has higher biodegradable organic matter content than other livestock wastes,[12] and a substantial amount of water would be required to dilute poultry litter. They also found that efficiency of organic matter conversion to methane decreased with increasing organic loads to the digesters. As a result, limited anaerobic digester systems are using poultry manure.

Conclusions

Poultry manure management poses challenges at different steps of manure management, including manure collection, storage and handling, and utilization or disposal. Utilization or disposal of manure in an environmentally safe way is one of the biggest challenges. Poultry manure can be handled as either a solid or a liquid. Owing to enriched nutrient content in manure, it may be applied to crops to meet crop nutrient requirements. However, selection of manure application tools and methods is important in minimizing environmental concerns and maximizing agronomic benefits from manure.

References

1. National Agricultural Statistics Service (NASS). 2010, available at http://usda.mannlib.cornell.edu/MannUsda/viewDocumentInfo.do?documentID=1509 (accessed August 16, 2010).
2. Collins, E.R.; Barker, J.C.; Carr, L.E.; Brodie, H.L.; Martin, J. H., Jr. *Natural Resource;* Agriculture and Engineering Service (NRAES), NRAES-132, Ithaca, NY, 1999.
3. MWPS-18. *Livestock Waste Facilities Handbook,* 2nd Ed.; Midwest Plan Service, Iowa State University: Ames, Iowa, 1985.
4. *Poultry Waste Management Handbook, NRAES-132;* Natural Resource, Agriculture, and Engineering Service (NRAES), Cooperative Extension, Ithaca, NY, 1999.
5. Fulhage, C.; Hoehne, J.; Jones, D.; Koelsch, R. *Manure Storages,* MWPS-18, Section 2; Midwest Plan Service, Iowa State University: Ames, IO, 2001.
6. Norman-Ham, H.A.; Hanna, H.M.; Richard, T.L. Solid manure distribution by rear- and side-delivery spreaders. Trans. ASABE **2008**, *51* (3), 831–843.
7. Laguë, C.; Agnew, J.M.; Landry, H.; Roberge, M.; Iskra, C. Development of a precision applicator for solid and semi-solid manure. Appl. Eng. Agric. **2006**, *22* (3), 345–350.
8. Farm Show Publishing, Inc. Prototype applicator buries poultry litter. Farm Show **2009**, *33* (2), 25; Farm Show Publishing, Inc.: Lakeville, MN.
9. Way, T.R.; Siegford, M.R.; Rowe, D.E. Applicator system and method for the agricultural distribution of biodegradable and non-biodegradable materials. U.S. Patent Number 7,721,662, May 25, 2010.
10. Tewolde, H.; Armstrong, S.; Way, T.R.; Rowe, D.E.; Sistani, K.R. Cotton response to poultry litter applied by subsurface banding relative to surface broadcasting. Soil Sci. Soc. Am. J. **2009**, *73,* 384–389.
11. Warren, J.G.; Sistani, K.R.; Way, T.R.; Mays, D.A.; Pote, D.H. A new method of poultry litter application to perennial pasture: Subsurface banding. Soil Sci. Soc. Am. J. **2008**, *72,* 1831–1837.
12. Bujoczek, G.; Oleszkiewicz, J.; Sparling, R.; Cenkowski, S. High solid anaerobic digestion of chicken manure. J. Agric. Eng. Res. **2000**, *76,* 51–60.

53

Organic Matter: Management

R. Cesar Izaurralde
and Carlos C. Cerri

Introduction

Soil organic matter (SOM) consists of a complex array of living organisms such as bacteria and fungi, plant and animal debris in different stages of decomposition, and *humus*—a rather stable brown to black material showing no resemblance to the organisms from which it originates. Because SOM is or has been part of living tissues, its composition is dominated by carbon (C), hydrogen, oxygen and—in lesser abundance—by nitrogen, phosphorus, sulfur among other elements. Levels of SOM are expressed in terms of soil organic carbon (SOC) concentration (g kg^{-1}) or mass per unit area (g m^{-2}) to a given depth. The level of SOC in virgin soils reflects the action and interaction of the major factors of soil formation: climate, vegetation, topography, parent material, and age. These factors control SOC content by regulating the balance between C gains via photosynthesis and losses via autotrophic and heterotrophic respiration, as well as C losses in soluble and solid form. The SOC content usually ranges between 5 and 100 g kg^{-1} in mineral soils. These concentrations appear modest but at 1500 Pg, the amount of organic C stored globally in soils is second only to that contained in oceans and at least twice that found in either terrestrial vegetation or the atmosphere.

Cultivated soils usually contain less SOC than virgin soils[1] due to the magnification of two biophysical processes: 1) net nutrient mineralization accompanied by release of CO_2 due to microbial respiration and 2) soil erosion. SOC losses of up to 50% have been reported within 30–70 years of land use conversions under temperate conditions.[2–5] SOC losses reported in subtropical and tropical environments often match or even surpass those observed under temperate conditions.[6–8] In subtropical and tropical environments, shifting cultivation systems appear to conserve more SOC than forestlands permanently cleared for cultivation.[9]

Major Processes Leading to Carbon Losses from Soil

Mineralization Processes

Depending on its frequency and kind, tillage changes the soil biophysical environment in ways that affect the net mineralization of nutrients and the release of carbon. These changes can be described in terms of increases or decreases in soil porosity, disruption of soil aggregates, and redistribution in the proportion of soil aggregate size, as well as alteration of energy and water fluxes. All these changes enhance, at least temporarily, the conversion of organic C into CO_2[10] and the net release of nutrients from SOM. Much of the success of past agricultural practices relied heavily on the control of decomposition processes through tillage operations to satisfy plant nutrient demands. All this came at a price, however, for a heavy reliance on soil nutrients to feed crops without proper replenishment led to the worldwide declines of SOM.[11]

Soil Erosion Processes

Agricultural ecosystems normally experience soil losses at rates considerably greater than natural ecosystems because of an incomplete plant or residue cover of the soil during rainy or windy conditions. When surface and environmental conditions are right (i.e., bare soil, sloping land, intense rain, windy weather), the kinetic energy embedded in wind and water is transferred to soil aggregates causing them to be detached and transported away from their original position across fields or downhill. Besides the physical loss of soil particles and on-site impact on soil productivity, the detachment and transport processes also cause aggregate breakdown, thereby exposing labile C to microbial activity. This aggregate breakdown also facilitates the preferential removal of soil materials comprised mainly of humus and clay or silt fractions. Consequently, water- and windborne sediments become enriched in C with respect to the contributing soil. Carbon enrichment ratios ranging from 3 to 360 have been reported.[12,13] The fate of these C-enriched sediments is not well known, for while transport and burial of C in eroded sediments may lead to “sequestration,” [14] it may also result in part of it being emitted back to the atmosphere as CO_2.[15]

Restoring Soil Organic Matter: The Emerging Science of Soil Carbon Sequestration

Role of Long-Term Field Experiments

SOM is an essential attribute of soil quality[16] and has an essential role in soil conservation and sustainable agriculture. Many practices—some involving land use changes—have been shown to increase SOM and thus received considerable attention for their possible role in climate change mitigation.[17–19] Carbon sequestration in managed soils occurs when there is a net removal of atmospheric CO_2 because C inputs (nonharvestable net primary productivity) are greater than C outputs (soil respiration, C costs related to fossil fuels and fertilizers). Soil C sequestration has the additional appeal that all its practices conform to principles of sustainable agriculture (e.g., reduced tillage, erosion control, diversified cropping systems, improved soil fertility). Long-term field experiments have been instrumental to increase our understanding of SOM dynamics.[20,21] The first and longest standing experiment was started at Rothamsted, England, by J. B. Lawes and J. H. Gilbert who in 1843 began documenting the impact of nutrient manipulation on crop yields and soil properties.[22] Other experiments were initiated thereafter in America, Europe, and Oceania with the goal of discovering interactions among climate, soil, and management practices. The knowledge that emerged from these experiments has been instrumental for the development and testing of agroecosystem and SOM models.[23]

Global Importance of Soil C Sequestration

There appears to be a significant opportunity for managed ecosystems to act as C sinks. For example, results from inverse modeling experiments suggest that during 1988–1992, terrestrial ecosystems may have been sequestering atmospheric C at rates of 1–2.2 Pg y^{-1}.[24] Some of the likely causes include the growth of new forest in previously cultivated land[25] and the "CO_2-fertilization effect."[26] Globally, agricultural soils have been estimated to have the capacity to sequester C at rates of 0.6 Pg y^{-1}[11] during several decades. The realization of this potential C sequestration would not be trivial since it would offset roughly about one-tenth of the current emissions from fossil fuels. In the U.S., annual gains in soil C from improved agricultural practices have been estimated at 0.14 Pg yr^{-1}.[25] Whether or not soil C sequestration practices are widely adopted will depend on their value relative to other C capture and sequestration technologies.

Mechanisms of Soil C Sequestration

Recent reviews of experimental results have contributed to organize our understanding of the environmental and management controls of soil C sequestration in grassland[27,28] and agricultural[29,30] ecosystems. The use of C balance, soil fractionation, and isotope techniques have been instrumental to reveal how new C (from crop residues, roots, and organic amendments) enters soil, resides shortly (for a few years) in labile soil fractions, and finally becomes a longtime constituent (for hundreds of years) of recalcitrant organo-mineral complexes.[31] Figure 1 contrasts young (labile) organic matter fractions extracted

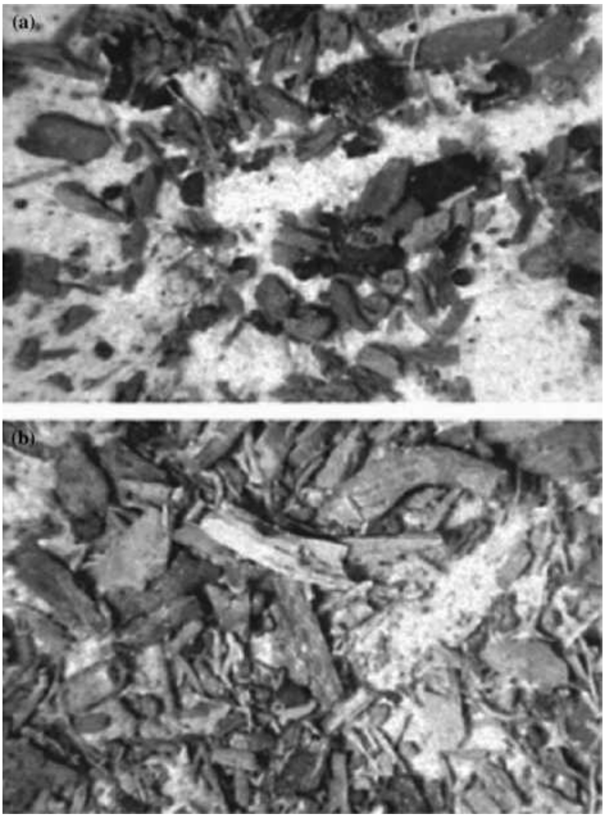


FIGURE 1 Young organic matters (fine roots and other organic debris) extracted from two Cryoboralfs under cereal cropping for 13 years receiving N at annual rates of 0 (a) and 50 kg ha^{-1} (b). The black material is charcoal. **Source:** Solberg et al.[32]

from two cultivated soils with and without N fertilizer.[32] The amounts of labile organic matter—fine roots and other organic debris— present in each soil reflect differences in crop productivity induced by addition of N at annual rates of 50 kg ha^{-1} for 13 years. "Terra Preta" soil—in tropical regions of South America and West Africa—represents a prime example of ancient wisdom applied to develop sustainable agriculture through the improvement of soil fertility and SOM.[33]

The quantity and quality of C entering soil as well as the interaction of this C with the soil biophysical environment are major factors determining the rate and duration of soil C sequestration. The quantity of C added to soil in the form of roots, crop residues, and organic amendments has been shown to play a dominant role in defining the trajectory of SOC over time.[34] Management practices geared toward optimizing nutrient supply and building nutrient reserves (e.g., fertilization, use of legumes in crop rotations) are almost guaranteed to increase soil C stocks. The quality of crop residues and the timing of their incorporation to soil also have an influential role on C decomposition and, thus, on soil C storage.[35] The degree of soil disturbance— through its impact on soil aggregation—constitutes another major factor regulating C decomposition and retention in soil.[36] In this context, no tillage agriculture has come to represent one of the most significant technological innovations of the last 30 years because it allows farmers the possibility of growing crops economically while reducing erosion and improving both quantity and quality of SOM. A few examples of the management impacts on soil C sequestration from around the world are presented in Table 1.

TABLE 1 Examples of Worldwide Land Use and Management Impacts on SOC

Region/ Country	Climate	Soil	Duration	Crop/Land Use	Treatment	Soil Organic Carbon (kg m–2) Initial	Final	Depth (cm)	Reference
Argentina	Temperate humid	Argiudoll	17	Corn–wheat–soybean	Moldboard plow		4.95	20	[37]
					No tillage		5.46		
Chaco, Argentina	Subtropical semiarid	Alfisol	20	Highly restored			7.05	20	[6]
			10	Moderately restored			3.10		
			60	Highly degraded			1.50		
Rondonia, Brazil	Tropical humid			Forest			4.33	50	[38]
			5	Pasture			5.85		
			9	Pasture			5.26		
			20	Pasture			5.28		
			41	Pasture			6.56		
			81	Pasture			6.12		
Georgia, U.S.	Temperate humid	Hapludult	5	Bermudagrass	Unharvested	1.39	1.74	6	[39]
					Lightly grazed		2.01		
					Heavily grazed		2.00		
					Hayed		1.59		

(*Continued*)

TABLE 1 (*Continued*) Examples of Worldwide Land Use and Management Impacts on SOC

						Soil Organic Carbon (kg m–2)			
Region/ Country	Climate	Soil	Duration	Crop/Land Use	Treatment	Initial	Final	Depth (cm)	Reference
Kentucky	Temperate humid	Paleudalf	20	Conventional tillage corn	0kg N/ha^{-1}		4.89	30	[40]
					84 kg N/ha^{-1}		5.63		
					168 kg N/ha^{-1}		5.64		
					336 kg N/ha^{-1}		6.14		
				No till corn	0 kg N/ha^{-1}		5.54		
					84 kg N/ha^{-1}		5.84		
					168 kg N/ha^{-1}		5.89		
					336 kg N/ha^{-1}		6.63		
Kuztown, Pennsylvania	Temperate humid	Fragiudalf	15	Corn	Conventional	4.20	4.30	15	[35]
					Organic	4.40	5.00		
					Manure	4.10	5.30		
Michigan	Cool temperate humid	Hapludalf	7	No crops—natural succession	Tillage		2.06	15	[41]
					Control		2.22		
Swift Current, Canada	Cold semiarid	Haploboroll	10	Cont. wheat	Minimum tillage	3.05	3.52	15	[42]
				Fallow–wheat–wheat		2.99	3.34		
				Green manure–wheat–wheat		2.89	3.22		
Breton, Canada	Cold subhumid	Cryoboralf	51	Wheat–fallow	Nil	2.64	1.81	15	[34]
					Fertilizer		2.13		
					Manure		3.11		
				Wheat–oat–barley–hay–hay	Nil		2.91		
					Fertilizer		3.37		
					Manure		4.32		
Russia	Cool temperate humid	Mollisol	300	Native grassland			2.07	50	[5]
			50	Hay			2.13		
			100	Continuous cropping			1.59		
			50	Continuous fallow			1.51		
Punjab, India	Subtropical subhumid	Alluvial	6	Corn–wheat	Minimum tillage, residue		0.48	15	[43]

(*Continued*)

TABLE 1 (*Continued*) Examples of Worldwide Land Use and Management Impacts on SOC

						Soil Organic Carbon (kg m–2)			
Region/ Country	Climate	Soil	Duration	Crop/Land Use	Treatment	Initial	Final	Depth (cm)	Reference
					retained				
					Minimum tillage, residue removed		0.48		
					Conventional tillage		0.50		
Morocco	Warm temperate semiarid	Calcixeroll	11	Continuous wheat and other rotations	Conventional tillage	3.20	3.73	20	[44]
					No tillage		3.39		
Western Nigeria			20	Bush fallow			2.77	15	[45]
			25	Bush fallow			2.96		
			10	Bermudagrass			3.90		
			10	Cultivation			1.29		

Soil Organic Matter, Energy, and Full C Accounting

Land is the natural habitat of humans. Humans dwell on it and use it as a resource for the production of food, fiber, and other goods. Simply put, land is managed when there is a manipulation of energy and matter flows in order to meet certain economic and social objectives. Farm mechanization and fertilizers are two of the many technical innovations that—though they rely on the utilization of fossil energy—have brought dramatic increases in food production during the last century. Changes in management practices that include soil C sequestration as an objective require careful evaluation of their impact not only on soil C gains but also on C costs from the use of fossil energy (e.g., manufacture of fertilizers)[46,47] and on the net greenhouse gas emissions.[48]

Role of Soil Organic Matter in the 21st Century

SOM has played and will continue to play a central role in sustainable land management. The restoration of SOM at global scales offers a unique opportunity to mitigate global warming. As population levels and affluence increase, demands on land to produce food, fiber, biomass, and other products will remain high. Because land is finite, important decisions will have to be made in order to balance such demands with functional objectives such as the preservation of natural ecosystems. As part of any climate policy, the impact of land use changes and management on SOM storage should be included as a criterion for making these decisions. Depending on their degree of expansion, several evolving agricultural technologies—such as genetically modified crops, conservation tillage, organic farming, and precision farming—may have important implications for soil C sequestration.[19] Their ultimate impact on C sequestration will depend not only on the economic benefit realized by individual producers but also on whether society recognizes the value of soil C storage to mitigate global warming.

References

1. Davidson, E.A.; Ackerman, I.L. Changes in soil carbon following cultivation of previously untilled soils. Biogeochemistry **1993**, *20,* 161–164.
2. Dalal, R.C.; Mayer, R.J. Long-term trends in fertility of soils under continuous cultivation and cereal cropping in southern Queensland. II. Total organic carbon and its rate of loss from the soil profile. Aust. J. Soil Res. **1986**, *24,* 281–292.
3. Mann, L.K. Changes in soil carbon after cultivation. Soil Sci. **1986**, *142,* 279–288.
4. Ellert, B.H.; Gregorich, E.G. Storage of carbon, nitrogen and phosphorus in cultivated and adjacent forested soils of Ontario. Soil Sci. **1996**, *161,* 587–603.
5. Mikhailova, E.A.; Bryant, R.B.; Vassenev, I.I.; Schwager, S.J.; Post, C.J. Cultivation effects on soil carbon and nitrogen contents at depth in the Russian chernozem. Soil Sci. Soc. Am. J. **2000**, *64,* 738–745.
6. Abril, A.; Bucher, E.H. Overgrazing and soil carbon dynamics in the western chaco of Argentina. Appl. Soil Ecol. **2001**, *16,* 243–249.
7. Lal, R. Deforestation and land use effects on soil degradation and rehabilitation in western Nigeria, II. Soil chemical properties. Land Degrad. Dev. **1996**, *7,* 87–98.
8. Lobe, I.; Amelung, W.; Du Preez, C.C. Losses of carbon and nitrogen with prolonged arable cropping from sandy soils of the South African highveld. Eur. J. Soil Sci. **2001**, *52*, 93–101.
9. Houghton, R.A. Changes in the storage of terrestrial carbon since 1850. In *Soils and Global Change*; Lal, R., Kimble, J., Levine, E., Stewart, B.A., Eds.; CRC/Lewis Publishers: Boca Raton, FL, 1995; 45–65.
10. Reicoski, D.C.; Lindstrom, M.J. Fall tillage method: effect from short-term carbon dioxide flux from soil. Agron. J. **1993**, *85,* 1237–1243.
11. Cole, V.; Cerri, C.; Minami, K.; Mosier, A.; Rosenberg, N.J.; Sauerbeck, D. Agricultural options for mitigation of greenhouse gas emissions. In *Climate Change 1995: Impacts, Adaptations and Mitigation of Climate Change*; Watson, R.T., Zinowera, M.C., Moss, R.H., Eds.; Report of IPCC Working Group II; Cambridge University Press: London, 1996; 745–771.
12. Sterk, G.; Herrmann, L.; Bationo, A. Wind-blown nutrient transport and soil productivity changes in southwest Niger. Land Degrad. Dev. **1996**, *7,* 325–336.
13. Zobeck, T.M.; Fryrear, D.W. Chemical and physical characteristics of wind-blown sediment. Trans. Am. Soc. Agric. Eng. **1986**, *29,* 1037–1041.
14. Stallard, R.F. Terrestrial sedimentation and the carbon cycle: coupling weathering and erosion to carbon burial. Global Biogeochem. Cycles **1998**, *12,* 231–257.
15. Lal, R. Global soil erosion by water and C dynamics. In *Soils and Global Change*; Lal, R., Kimble, J., Levine, E., Stewart, B.A., Eds.; CRC/Lewis Publishers: Boca Raton, FL, 1995; 131–141.
16. Doran, J.W.; Coleman, D.C.; Bezdicek, D.F.; Stewart, B.A., Eds.; *Defining Soil Quality for a Sustainable Environment;* Soil Science Society America Special Publication No. 35; SSSA: Madison, WI, 1994; 244 pp.
17. Batjes, N.H. Mitigation of atmospheric CO_2 concentrations by increased carbon sequestration in the soil. Biol. Fert. Soils **1998**, *27,* 230–235.
18. Post, W.M.; Kwon, K.C. Soil carbon sequestration and land-use change: processes and potential. Global Change Biol. **2000**, *6*, 317–327.
19. Izaurralde, R.C.; Rosenberg, N.J.; Lal, R. Mitigation of climatic change by soil carbon sequestration: issues of science, monitoring and degraded lands. Adv. Agron. **2001**, *70,* 1–75.
20. Powlson, D.S.; Smith, P.; Smith, J.U., Eds.; *Evaluation of Soil Organic Matter Models Using Existing Long-Term Datasets;* NATO ASI Series I; Springer: Heidelberg, 1996; Vol. 38, 429 pp.
21. Paul, E.A.; Paustian, K.; Elliott, E.T.; Cole, C.V., Eds.; *Soil Organic Matter in Temperate Agroecosystems: Long-Term Experiments in North America;* NATO ASI Series I; CRC/ Lewis Publishers: Boca Raton, FL, 1997; 414 pp.

22. Jenkinson, D.S. The rothamsted long-term experiments: are they still of use? Agron. J. **1991**, *83,* 2–10.
23. Smith, P.; Smith, J.U.; Powlson, D.S.; McGill, W.B.; Arah, J.R.M.; Chertov, O.; Coleman, K.W.; Franko, U.; Frolking, S.; Jenkinson, D.S.; Jensen, L.S.; Kelly, R.H.; Klein- Gunnewiek, H.; Komarov, A.S.; Li, C.; Molina, J.A.E.; Mueller, T.; Parton, W.J.; Thornley, J.H.M.; Whitmore, A.P. A comparison of the performance of nine soil organic matter models using datasets from seven long-term experiments. Geoderma **1997**, *81,* 153–225.
24. Fan, S.; Gloor, M.; Mahlman, J.; Pacala, S.; Sarmiento, J.; Takahashi, T.; Tans, P. A large terrestrial carbon sink in North America implied by atmospheric and oceanic carbon dioxide data and models. Science **1998**, *282,* 442–446.
25. Houghton, R.A.; Hackler, J.L.; Lawrence, K.T. The U.S. carbon budget: contributions from land use change. Science **1999**, *285,* 574–578.
26. Kimball, B.A. Carbon dioxide and agricultural yield: an assemblage and analysis of 430 prior observations. Agron. J. **1983**, *75*, 779–782.
27. Conant, R.T.; Paustian, K.; Elliott, E.T. Grassland management and conversion into grassland: effects on soil carbon. Ecol. Appl. **2001**, *11*, 343–355.
28. Scott, N.A.; Tate, K.R.; Ford-Robertson, J.; Giltrap, D.J.; Smith, C.T. Soil carbon storage in plantations and pastures: land-use implications. Tellus **1999**, *51B,* 326–335.
29. Janzen, H.H.; Campbell, C.A.; Izaurralde, R.C.; Ellert, B., H.; Juma, N.; McGill, W.B.; Zentner, R.P. Management effects on soil c storage on the Canadian prairies. Soil Till. Res. **1998**, *47,* 181–195.
30. Paustian, K.; Collins, H.P.; Paul, E.A. Management controls on soil carbon. In *Soil Organic Matter in Temperate Ecosystems: Long-Term Experiments in North America*; Paul, E.A., Paustian, K., Elliott, E.T., Cole, C.V., Eds.; CRC/Lewis Publishers: Boca Raton, FL, 1997; 15–49.
31. Jastrow, J.D. Soil aggregate formation and the accrual of particulate and mineral-associated organic matter. Soil Biol. Biochem. **1996**, *28,* 665–676.
32. Solberg, E.D.; Nyborg, M.; Izaurralde, R.C.; Malhi, S.S.; Janzen, H.H.; Molina-Ayala, M. Carbon storage in soils under continuous cereal grain cropping: N fertilizer and straw. In *Management of Carbon Sequestration in Soil;* Lal, R., Kimble, J., Follett, R., Stewart, B.A., Eds.; CRC/Lewis Publishers: Boca Raton, FL, 1998; 235–254.
33. Glaser, B.; Haumaier, L.; Guggenberger, G.; Zech, W. The "Terra Preta" phenomenon: a model for sustainable agriculture in the humid tropics. Naturwissenschaften **2001**, *88*, 37–41.
34. Izaurralde, R.C.; McGill, W.B.; Robertson, J.A.; Juma, N.G.; Thurston, J.T. Carbon balance of the Breton classical plots after half a century. Soil Sci. Soc. Am. J. **2001**, *65,* 431–441.
35. Drinkwater, L.E.; Wagoner, P.; Sarrantonio, M. Legume- based cropping systems have reduced carbon and nitrogen losses. Nature **1998**, *396,* 262–265.
36. Six, J.; Elliott, E.T.; Paustian, K. Soil macroaggregate turnover and microaggregate formation: a mechanism for C sequestration under no-tillage agriculture. Soil Biol. Biochem. **2000**, *32,* 2099–2103.
37. Alvarez, R.; Russo, M.E.; Prystupa, P.; Scheiner, J.D.; Blotta, L. Soil carbon pools under conventional and no-tillage systems in the argentine rolling pampa. Agron. J. **1998**, *90,* 138–143.
38. Neill, C.; Cerri, C.C.; Melillo, J.M.; Feigl, B.J.; Steudler, P.A.; Moraes, J.F.L.; Piccolo, M.C. Stocks and dynamics of soil carbon following deforestation for pasture in Rondoônia. In *Soil Processes and the Carbon Cycle;* Lal, R., Kimble, J., Follett, R., Stewart, B.A., Eds.; CRC/Lewis Publishers: Boca Raton, FL, 1998; 9–28.
39. Franzluebbers, A.J.; Stuedemann, J.A.; Wilkinson, S.R. Bermudagrass management in the southern piedmont USA: I. Soil and surface residue carbon and sulfur. Soil Sci. Soc. Am. J. **2001**, *65*, 834–841.
40. Ismail, I.; Blevins, R.L.; Frye, W.W. Long-term notillage effects on soil properties and continuous corn yields. Soil Sci. Soc. Am. J. **1994**, *58,* 193–198.

41. Richter, D.D.; Babbar, L.I.; Huston, M.A.; Jaeger, M. Effects of annual tillage on organic carbon in a finetextured udalf: the importance of root dynamics to soil carbon storage. Soil Sci. **1999**, *149,* 78–83.
42. Curtin, D.; Wang, H.; Selles, F.; McConkey, B.G.; Campbell, C.A. Tillage effects on carbon fluxes in continuous wheat and fallow-wheat rotations. Soil Sci. Soc. Am. J. **2000**,*64,* 2080–2086.
43. Ghuman, B.S.; Sur, H.S. Tillage and residue management effects on soil properties and yields of rainfed maize and wheat in a subhumid subtropical climate. Soil Till. Res. **2001**, *58,* 1–10.
44. Mrabet, R.; Saber, N.; El-Brahli, A.; Lahlou, S.; Bessam, F. Total, particular organic matter and structural stability of a calcixeroll soil under different wheat rotations and tillage systems in a semiarid area of Morocco. Soil Till. Res. **2001**, *57,* 225–235.
45. Lal, R. Land use and soil management effects on soil organic matter dynamics on alfisols in western Nigeria. In *Soil Processes and the Carbon Cycle;* Lal, R., Kimble, J., Follett, R., Stewart, B.A., Eds.; CRC/Lewis Publishers: Boca Raton, FL, 1998; 109–126.
46. Schlesinger, W.H. Carbon and agriculture: carbon sequestration in soils. Science **1999**, *284,* 2095.
47. Izaurralde, R.C.; McGill, W.B.; Rosenberg, N.J. Carbon cost of applying nitrogen fertilizer. Science **2000**, *288*, 811–812.
48. Robertson, G.P.; Paul, E.A.; Harwood, R.R. Greenhouse gases in intensive agriculture: contributions of individual gases to the radiative forcing of the atmosphere. Science **2000**, *289,* 1922–1925.

54

Pest Management

E.F. Legner

Pest management is a broad concept that involves considerations of genetics, climate, ecology, natural enemies, and cultural or chemical applications. Therefore, it is difficult to define this category exactly. A high level of sophistication is required to manage events in the environment for the efficient production of food and fiber and the abatement of public health and nuisance pests. A principal objective to the addition of sound environmental management is the reduction of pesticide usage albeit at the irritation of large commercial interests.[1,2]

Although scientific investigations in colleges and universities have led to a high level of production and pest abatement, deployment continues to face obstacles that are largely related to the absence of competent supervisory personnel. As expertise resides largely in the research community, this group is encumbered by an academic system that continues to stress research and teaching and to minimize the deployment aspect. The most successful programs in environmental management regularly require 5 or more years to develop. Investigator survival in the system demands frequent publication but not in the kind of journals that stress implementation. This distracts from the ultimate goal of deployment, which diminishes the amount of time an investigator has to be directly involved in an advisory capacity. Several examples of successful projects that have receded in the absence of this supervision but that could be reactivated with the proper advisory personnel present will explain some of the problems and difficulties involved.

Navel Orangeworm Management in Almond Orchards

The almond industry in California has suffered from the invasion of the navel orangeworm, *Amyelois transitella* (Walker), from Mexico and South America. Two external insect larval parasites, *Goniozus legneri* Gordh and *Goniozus emigratus* (Rohwer), and one internal egg-larval parasite, *Copidosomopsis plethorica* Caltagirone, which are dominant in south Texas, Mexico, Uruguay, and Argentina, were successfully established in irrigated and non-irrigated almond orchards in California.[3–4] Separate k-value analyses indicated significant regulation of their navel orangeworm host during the warm summer season. There is a diapause (hibernation) in the host triggered by several seasonally varying factors and a diapause in the parasites triggered by hormonal changes in the host. Possible latitudinal effects on diapause also are present. The ability of the imported parasites to diapause with their host enables their permanent establishment and ability to reduce host population densities to below economic levels.[5]

Although navel orangeworm infestations have decreased with the establishment of the three parasites,[6] the almond reject levels are not always below the economic threshold of 4%. Such rejects are sometimes due to other causes, such as ant damage and fungus infections. In certain years, the peach tree borer, *Synanthedon exitiosa* (Say), has been involved as its attacks stimulate oviposition by navel orangeworm moths and subsequent damage attributed to the latter.

In some orchards, the growers have sustained a reject level of 2.5% or less through 2008. Storing rejected almond mummies in ventilated sheds through winter allows for a buildup of natural enemies and their subsequent early entry into the fields to reduce orangeworm populations before the latter have an opportunity to increase. Commercial insectaries have harvested *G. legneri* from orchards for introductions elsewhere. *C. plethorica* and *G. legneri,* and to a lesser extent, *G. emigratus*, successfully overwinter in orchards year after year. However, only *Copidosomopsis* can consistently be recovered at all times of the year. The *Goniozus* species are not recovered in significant numbers until early summer. Therefore, pest management in almond orchards may require periodic releases of *G. legneri* to reestablish balances that were disrupted by insecticidal drift or by the absence of overwintering rejected almond refuges through aggressive sanitation practices. Although sanitation in this case may appeal to the grower, it is a costly procedure that also disrupts natural balances at low pest densities.

G. legneri has been reared from codling moth and oriental fruit moth in peaches in addition to navel orangeworm from almonds. A reservoir of residual almonds that remain in the trees after harvest is desirable to maintain a synchrony of these parasites with navel orangeworms in order to achieve the lowest pest densities. In fact, such reservoirs often exceed 1000 residual almonds per tree through the winter months and produce navel orangeworm densities at harvest that are below 1% on soft-shelled varieties. Superimposed upon the system is the diapausing mechanism in both the navel orangeworm and the parasites.[5] All of these forces must be considered for a sound, reliable integrated management. Almond producers have to make reasonable decisions on whether or not to remove residual almonds, a very costly procedure, or to use within-season insecticidal sprays. However, orchard managers rarely understand population stability through the interaction of natural enemies and their prey. Because the management of this pest with parasitic insects depends heavily on the perpetuation of parasites in orchards, it can be accomplished only by an understanding of the dynamics involved. Storing rejected almonds in protective shelters during winter months increases parasite abundance. This allows the parasites to reproduce in large numbers for subsequent spread throughout an orchard in the spring when outdoor temperatures rise. Complete sanitation of an orchard by removal of all rejected almonds is counterproductive to successful management as this also eliminates natural enemies.

Australian Bush Fly Management in Micronesia

Pestiferous flies in the Marshall Islands provide a classic example of the adaptation of invading noxious insects to an area with a salubrious climate. With nearly perfect temperature–humidity conditions for their development, an abundance of carbohydrate and protein-rich food in the form of organic wastes and excreta provided by humans and their animals, and a general absence of effective natural enemies, several species were able to reach maximum numbers.

There are principally four types of pestiferous flies in Kwajalein Atoll of the Marshall Islands, with the African–Australian bush fly, *Musca sorbens* Wiedemann, being by far the most pestiferous species. The common housefly, *Musca domestica* L., of lesser importance, frequents houses and is attracted to food in recreation areas. The remaining two types are the Calliphoridae (*Chrysomya megacephala* [Fab.] and [Wiedemann]) and the Sarcophagidae (*Parasarcophaga misera* [Walker] and *Phytosarcophaga gressitti* Hall and Bohart). These latter species are abundant around refuse disposal sites and wherever rotting meat and decaying fish are available. Most of the fly species differ from the common housefly and the bush fly in being more sluggish and noisy and by their general avoidance of humans. Because residents do not distinguish the different kinds of flies, non-pestiferous types are often blamed as nuisances when in fact they may be considered to fulfill a useful role in the biodegradation of refuse and rotting meat.

An initial assessment of the problem led to the expedient implementation of breeding source reduction to reduce the population of the housefly, *M. domestica* L., and both the Calliphoridae and Sarcophagidae to inconspicuous levels. These involved slight modifications of refuse disposal sites to disfavor fly breeding. These simple measures resulted in an estimated one-third reduction of total population of flies concentrating around beaches and residential areas. Because the housefly especially enters dwellings, the reduction in its numbers was desirable for the general health of the community, and fly annoyances indoors diminished. Thorough surveys of breeding sites and natural enemy complexes revealed that *M. sorbens* population reduction would not be quickly forthcoming, however. A schedule of importation of natural enemies was begun, and other integrated management approaches were investigated, e.g., baiting and breeding habitat reduction.

Bush Fly Origin and Habits

This species is known as the bazaar fly in North Africa, a housefly in India, and the bush fly in Australia.[7] It was first described from Sierra Leone in West Africa in 1830, where it is a notorious nuisance to humans and animals. The flies are attracted to wounds, sores, and skin lesions, searching for any possible food sources such as blood and other exudations. Although not a biting species, its habits of transmitting eye diseases, enteric infections, pathogenic bacteria, and helminth eggs make it a most important and dangerous public health insect.[8–11]

The bush fly has spread through a major portion of the Old World, Africa, and parts of Asia.[12] In Oceania, its distribution is in Australia,[13] New Guinea,[13] Samoa and Guam,[14] and the Marshall Islands.[15] In Hawaii, Joyce[16] first reported it in 1950. Later, Hardy[17] listed it in the *Catalog of Hawaiian Diptera,* and Wilton[18] reported its predilection for dog excrement. The importance of the bush fly increased in the 1960s, when it was incriminated as a potential vector of beta-hemolytic streptococci in an epidemic of acute glomerulonephritis.[8]

On the islands of Kwajalein Atoll, a substantial portion of the main density of *M. sorbens* emanated from dog, pig, and human feces. Inspections of pig droppings in the bush of 10 widely separated islets revealed high numbers of larvae (over 100 per dropping), making this dung, as in Guam,[15] a primary breeding source in the atoll. Pigs that are corralled on soil or concrete slabs concentrate and trample their droppings, making them less suitable breeding sites. In such situations, flies were able to complete their development only along the periphery of corrals. Coconut husks placed under pigs in corrals results in the production of greater numbers of flies by reducing the effectiveness of trampling. Kitchen and other organic wastes were not found to breed *M. sorbens,* although a very low percentage of the adult population could originate there judging from reports elsewhere. Nevertheless, this medium is certainly not responsible for producing a significant percentage of the adult densities observed in the Atoll.

Management Efforts Worldwide

Successful partial reduction of the bush fly population had been achieved only in Hawaii, through a combination of the elimination of breeding sites, principally dog droppings, and the activities of parasitic and predatory insects introduced earlier to combat other fly species, e.g., *M. domestica.*[19] The density of the bush fly population varies in different climatic zones in Hawaii, but the importance of this fly is minimal compared with Kwajalein. At times, hymenopterous parasites have been found to parasitize over 95% of flies sampled in the Waikiki area (H.S. Yu, unpublished data). Other parts of Oceania (e.g., Australia) either were not suitable for maximum effectiveness of known parasitic species or had principle breeding habitats that were not attractive to the natural enemies. Therefore, in Australia, a concerted effort has been made to secure scavenger and predatory insects from southern Africa that are effective in managing fly populations on range cattle and sheep dung, the principal fly-producing source.[20]

Kwajalein Atoll

Integrated fly management had reached a level of partial success by 1974. Initial surveys for natural enemies of *M. sorbens* revealed the presence of four scavenger and predatory insects: the histerid *Carcinops troglodytes* Erichson, the nitidulid *Carpophilus pilosellus* Motschulsky, the tenebrionid *Alphitobius diaperinus* (Panzer), and the dermapteran *Labidura riparia* (Pallas). Dog numbers were significantly reduced, and all privies were reconstructed or improved on one island, Ebeye. Dogs were reduced or tethered on Kwajalein Island and refuse fish, etc., disposed of thoroughly on Illeginni and other islands with American residents. Importations of natural enemies were made throughout the atoll, and the average population density of *M. sorbens* on Ebeye was subsequently reduced from an estimated 8.5 flies attracted to the face per minute to less than 0.5 fly per minute, which was readily appreciated by the inhabitants. The single most important cause appeared to be the partial elimination of breeding sources, with natural enemies playing a secondary role.

For the further reduction of bush fly numbers, the integration of a nondestructive insecticidal reduction measure was desirable. Sugar bait mixtures that have been used for houseflies in years prior to 1972 were wholly ineffective for killing adult *M. sorbens* due to their almost complete lack of attractiveness. However, a variety of decomposing foodstuffs including rotting eggs and rotting fish sauces were very highly attractive. Experiments using a 6-day-old mixture of one part fresh whole eggs to one part water[21] attracted over 50,000 bush flies that were then killed by a 0.5 ppm Dichlorvos additive. The poisoned mixture was poured in quantities of 100 mL each in flat plastic trays with damp sand at 20 sites in the shade and spaced every 10 m along a public beach on Kwajalein. Baits placed above the height of 1 m or against walls in open pavilions were only weakly attractive. After 48 hours, flies were reduced to inconspicuous levels all over Kwajalein Island. This condition endured for at least 3 days, after which newly emerging and immigrating flies managed to slowly increase to annoying levels as the baits ceased to be attractive. However, the former density of flies was never reached even one week after the baiting; these populations were subsequently reduced to even lower levels by applying additional fresh poisoned baits.

Baiting was extended to other islands in the atoll with the result of sustained reductions of bush fly numbers to below general annoyance levels (less than 0.01 attracted per minute on Kwajalein, Roi-Namur, Illeginni, and Meck Islands). A new attractant that augmented the rotting egg mixture consisted of beach sand soaked for one week in the decomposing body fluids of buried sharks. This new attractant was far superior to rotting eggs in both rate and time of attraction, the latter sometimes exceeding 5 days. The baiting method could be used effectively if applied initially twice a week, and only biweekly applications were necessary in the following months.

After January 2000, in the absence of specialist supervision, the baiting procedure in the atoll has not continued with the sophistication initially determined necessary. In the absence of supervision, the number of flies was not adequately reduced. Periodic personnel changes precluded the passing on of accurate information critical to managing the fly densities. Of vital importance is habitat reduction, the proper preparation of baits, and the latter's placement in shaded wind-calm areas of the islands. Because such sites are generally out of sight of the public, baiting has instead shifted to populated areas, where only very conspicuous but non-pestiferous species of flies are attracted to the baits in large numbers. Sometimes, even ammonia baits were substituted that attract harmless blow fly species but not the targeted bush fly.

Aquatic Weed Management by Fish in Irrigation Systems

Imported fish species have been used for clearing aquatic vegetation from waterways, which has also reduced mosquito and chironomid midge abundance. In the irrigation systems, storm drainage channels, and recreational lakes of southern California, the California Department of Fish and Game authorized the introduction of three species of African cichlids: *Tilapia zillii* (Gervais), *Oreochromis (Sarotherodon) mossambica* (Peters), and *Oreochromis (Sarotherodon) hornorum* (Trewazas). These became established over some 2000 ha of waterways.[22] Their establishment reduced the biomass of emergent aquatic vegetation that was slowing down the distribution of irrigation water but that also provided a habitat for such encephalitis vectors as the mosquito *Culex tarsalis* Coquillet. Previous aquatic weed reduction practices had required an expensive physical removal of vegetation and/or the frequent application of herbicides.

One species, *T. zillii* can reduce mosquito populations by a combination of direct predation and the consumption of aquatic plants by these omnivorous fishes.[23–25] As Legner and Sjogren[22] indicated, this is a unique example of persistent biological suppression and probably applicable only for relatively stable irrigation systems where a permanent water supply is assured and where water temperatures are warm enough in winter to sustain the fish.[26] A threefold advantage in the use of these fish is as follows: 1) clearing of vegetation to keep waterways open; 2) mosquito abatement; and 3) a fish large enough to be used for human consumption. However, optimum management of these cichlids for aquatic weed reduction often is not understood by irrigation district personnel,[27–28,19] with the result that competitive displacement by inferior cichlids minimizes or eliminates *T. zillii*, the most efficient weed-eating species.[29]

The three imported fish species varied in their influence in different parts of the irrigation system. Each fish species possessed certain attributes for combating the respective target pests.[30,31] *T. zillii* was best able to perform as both a habitat reducer and an insect predator. It also had a slightly greater tolerance to low water temperatures, which guaranteed the survival of large populations through the winter months; at the same time, it did not pose a threat to salmon and other game fisheries in the colder waters of central California. It was the superior game species and most desirable as human food. Nevertheless, the agencies supporting the research (mosquito abatement and county irrigation districts) acquired and distributed all three species simultaneously throughout hundreds of kilometers of the irrigation system, storm drainage channels, and recreational lakes. The outcome was the permanent and semipermanent establishment of the two less desirable species, *S. mossambica* and *S. hornorum,* over a broader portion of the distribution range. This was achieved by the competitively advantaged *Sarotherodon* species that mouth-brood their fry, while *T. zillii* did not have this attribute strongly developed. It serves as an example of competitive exclusion such as conjectured by Ehler.[32] In the clear waters of some lakes in coastal and southwestern California, the intense predatory behavior of *S. mossambica* males on the fry of *T. zillii* could be easily observed, even though adults of the latter species gave a strong effort to fend off these attacks.

This outcome was not too serious for chironomid reduction in storm drainage channels because the *Sarotherodon* species are quite capable of permanently suppressing chironomid densities to below annoyance levels.[23,26] However, for the management of aquatic weeds, namely, *Potamogeton pectinatus* L., *Myriophyllum spicatum* var. *exalbescens* (Fernald) Jepson, *Hydrilla verticillata* Royle, and *Typha* species, they showed little capability.[31] Thus, competition excluded *T. zillii* from expressing its maximum potential in the irrigation channels of the lower Sonoran Desert and in the recreational lakes of southwestern California. Furthermore, as the *Sarotherodon* species were of a more tropical nature, their populations were reduced in the colder waters of the irrigation canals and recreational lakes. Although *T. zillii* populations could have been restocked, attention was later focused on a potentially more environmentally destructive species, the white amur, *Ctenopharyngodon idella* (Valenciennes), and other carps. The competitively advantaged *Sarotherodon* species are permanently established over a broad geographic area, which encumbers the reestablishment of *T. zillii* in storm drainage channels of southwestern California.

Management of Filth Fly Abundance in Dairies and Poultry Houses

The most important of muscoid fly species are broadly defined as those most closely associated with human activities. Breeding habitats vary from the organic wastes of urban and rural settlements to those provided by various agricultural practices, particularly ones related to the management and care of domestic animals. Their degree of relationship to humans varies considerably with the ecology and behavior of the fly species involved. Some are more often found inside dwellings.

Research to reduce fly abundance has centered on the highly destructive parasitic and predatory species, such as the encyrtid *Tachinaephagus zealandicus* Ashmead, five species of the pteromalid genus *Muscidifurax,* and *Spalangia* species that destroy dipterous larvae and pupae in various breeding sources. The natural enemies are capable of successful fly suppression if the correct species and strains are applied in the right locality.[24,33–36] Other approaches have included the use of pathogens and predatory mites and inundative releases of parasites and predators.[37,38] Although partially successful, none of these strategies has become the sole method for fly abatement, and the choice of a ineffective parasite strain may have detrimental results.[19] Instead, the focus is on integrated management including habitat reduction, adult baiting, and aerosol treatments with short residual insecticides. Also, it is generally agreed that existing predatory complexes exert great influences on fly densities[39] and that many natural enemies of these flies have a potential to significantly reduce their abundance if managed properly.[29,40] Because climatic and locality differences dictate which abatement strategies are effective, simple instructions to the public are impossible, and the involvement of skilled personnel is required. Of primary importance for successful management is the provision of relatively stable breeding habitats and their natural enemy complexes. Periodic cleaning operations should stress the partial removal of breeding sites and the deposition of such waste into large stacks, which favors the generation of destructive heat while minimizing the area and attractiveness for fly oviposition. Nevertheless, this management procedure is difficult for abatement personnel to grasp in the absence of competent supervision.

References

1. Garcia, R.; Legner, E.F. Biological control of medical and veterinary pests. In *Handbook of Biological Control: Principles and Applications*; Fisher, T.W., Bellows, T.S., Jr. Eds.; Academic Press: San Diego, CA, 1999; 935–953.
2. Pimentel, D.; McLaughlin, L.; Zepp, A.; Lakitan, B.; Kraus, T.; Kleinman, P.; Vancini, F.; Roach, W.J.; Graap, E.; Keeton, W.S.; Selig, G. Environmental and economic impacts of reducing U.S. agricultural pesticide use. In *Handbook of Pest Management in Agriculture,* 2nd Ed.; Pimentel, D., Ed.; CRC Press: Boca Raton, Florida, 1991; Vol. I., 679–718.
3. Caltagirone, L.E. A new *Pentalitomastix* from Mexico. Pan-Pac. Entomol. **1966**, *42,* 145–151.
4. Legner, E.F.; Silveira-Guido, A. Establishment of *Goniozus emigratus* and *Goniozus legneri* [Hym: Bethylidae] on navel orangeworm, *Amyelois transitella* [Lep: Phycitidae], in California and biological control potential. Entomophaga **1983**, *28,* 97–106.
5. Legner, E.F. Patterns of field diapause in the navel orange-worm (Lepidoptera: Phycitidae) and three imported parasites. Ann. Entomol. Soc. Am. **1983**, *76,* 503–506.
6. Legner, E.F.; Gordh, G. Lower navel orangeworm (Lepi- doptera: Phycitidae) population densities following establishment of *Goniozus legneri* (Hymenoptera: Bethylidae) in California. J. Econ. Entomol. **1992**, *85* (6), 2153–2160.
7. Yu, H. The biology and public health significance of *Musca sorbens* Wied. in Hawaii, M.S. Thesis, University of Hawaii, 1971; 72 pp.
8. Bell, T.D. Epidemic glomerulonephritis in Hawaii [mimeograph]. Rep. Pediat. Serv., Dep. Med., Tripler Army Hospital: Honolulu, Hawaii, 1969; 25 pp.
9. Greenberg, B. *Flies and Disease*; Ecology; Classification and Biotic Associations; Princeton University Press: Princeton, NJ, 1971; Vol. I, 856 pp.
10. Hafez, M.; Attia, M.A. Studies on the ecology of *Musca sorbens* Wied. in Egypt. Bull. Soc. Entomol. Egypt **1958**, *42,* 83–121.
11. McGuire, C.D.; Durant, R.C. The role of flies in the transmission of eye disease in Egypt. Am. J. Trop. Med. Hyg. **1957**, *6*, 569–75.
12. Van Emden, F.I. The fauna of India and the adjacent countries. *Diptera;* Muscidae, Pt. I; Gov. Publ. India: Delhi, India, 1965; Vol. 7.

13. Patterson, H.E.; Norris, K.R. The *Musca sorbens* complex: The relative status of the Australian and two African populations. Aust. J. Zool. **1970**, *18,* 231–45.
14. Harris, A.H.; Down, H.A. Studies of the dissemination of cysts and ova of human intestinal parasites by flies in various localities on Guam. Am. J. Trop. Med. **1946**, *26,* 789–800.
15. Bohart, G.E.; Gressitt, J.L. *Filth inhabiting flies of Guam;* Bull. B.P. Bishop Museum, Honolulu 1951, *204;* 152 pp.
16. Joyce, C.R. Notes and exhibitions. Proc. Hawaii. Entomol. Soc. **1950**, *16* (3), 338.
17. Hardy, D.E. Additions and corrections to Bryan's check list of the Hawaiian Diptera. Proc. Hawaii. Entomol. Soc. **1952**, *14* (3), 443–84.
18. Wilton, D.P. Dog excrement as a factor in community fly problems. Proc. Hawaii. Entomol. Soc. **1963**, *28* (2), 311–17.
19. Legner, E.F. Diptera. Medical and veterinary pests. In *Introduced Parasites and Predators of Arthropod Pests and Weeds: A Review;* Clausen, C.P., Ed.; U.S. Department of Agriculture Technology Report, 1978; 1012–1019, 1043–1069.
20. Bornemissza, G.F. Insectary studies on the control of dung breeding flies by the activity of the dung beetle, *Onthophagus gazella* F. (Coleoptera: Scarbaeinae). J. Aust. Entomol. Soc. **1970**, *9,* 31–41.
21. Legner, E.F.; Sugerman, B.B.; Hyo-sok, Y.; Lum, H. Biological and integrated control of the bush fly, *Musca sorbens* Wiedemann, and other filth-breeding Diptera in Kwajalein Atoll, Marshall Islands. Bull. Soc. Vector Ecol. **1974**, (1), 1–14.
22. Legner, E.F.; Sjogren, R.D. Biological mosquito control furthered by advances in technology and research. J. Am. Mosq. Control Assoc. **1984**, *44* (4), 449–456.
23. Legner, E.F.; Fisher, T.W. Impact of *Tilapia zillii* (Gervais) on *Potamogeton pectinatus* L., *Myriophyllum spicatum* var. *exalbescens* Jepson, and mosquito reproduction in lower Colorado Desert irrigation canals. Acta Oecol., Oecol. Appl. **1980**, *1* (1), 3–14.
24. Legner, E.F.; Murray, C.A. Feeding rates and growth of the fish *Tilapia zillii* [Cichlidae] on *Hydrilla verticillata*, *Potamogeton pectinatus* and *Myriophyllum spicatum* var. *exalbescens* and interactions in irrigation canals in southeastern California. J. Am. Mosq. Control Assoc. **1981**, *41* (2), 241–250.
25. Legner, E.F.; Pelsue, F.W., Jr. Contemporary appraisal of the population dynamics of introduced cichlid fish in south California. Proc. Calif. Mosq. Vector Control Assoc., Inc. **1983**, *51,* 38–39.
26. Legner, E.F.; Medved, R.A.; Pelsue, F. Changes in chironomid breeding patterns in a paved river channel following adaptation of cichlids of the *Tilapia mossambica-hornorum* complex. Ann. Entomol. Soc. Am. **1980**, *75* (1), 293–299.
27. Hauser, W.J.; Legner, E.F.; Medved, R.A.; Platt, S. *Tilapia*—A management tool for biological control of aquatic weeds and insects. Bull. Am. Fish. Soc. **1976**, *1,* 15–16.
28. Hauser, W.J.; Legner, E.F.; Robinson, F.E. Biological control of aquatic weeds by fish in irrigation channels. In Proceedings of Water Management for Irrigation and Drainage, ASC/Reno, Nevada, Jul 20–22, 1977; 139–145.
29. Legner, E.F. Biological control of aquatic Diptera; *Contributions to a Manual of Palaearctic Diptera*; Science Herald: Budapest, 2000; Vol. 1, 847–870.
30. Legner, E.F.; Medved, R.A. Influence of *Tilapia mossam- bica* (Peters), T. *zillii* (Gervais) (Cichlidae) and *Molliene- sia latipinna* LeSueur (Poeciliidae) on pond populations of *Culex* mosquitoes and chironomid midges. J. Am. Mosq. Control Assoc. **1973**, *33*, 354–64.
31. Legner, E.F.; Medved, R.A. Predation of mosquitoes and chironomid midges in ponds by *Tilapia zillii* (Gervais) and T. *mossambica* (Peters) (Teleosteii: Cichlidae). Proc. Calif. Mosq. Control Assoc., Inc. **1973**, *41,* 119–121.
32. Ehler, L.E. Foreign exploration in California. Environ. Entomol. **1982**, *11*, 525–30.
33. Axtell, R.C.; Rutz, D.A. Role of parasites and predators as biological control agents in poultry production facilities. Misc. Publ. Entomol. Soc. Am. **1986**, *61,* 88–100.

34. Legner, E.F.; Greathead, D.J.; Moore, I. Equatorial East African predatory and scavenger arthropods in bovine excrement. Environ. Entomol. **1981**, *10,* 620–25.
35. Mandeville, J.D.; Mullens, B.A.; Meyer, J.A. Rearing and host age suitability of *Fannia canicularis* (L.) for parasitization by *Muscidifurax zaraptor* Kogan and Legner. Can. Entomol. **1988**, *120,* 153–59.
36. Pawson, B.M.; Petersen, J.J. Dispersal of *Muscidifurax zaraptor* (Hymenoptera: Pteromalidae), a filth fly parasitoid, at dairies in eastern Nebraska. Environ. Entomol. **1988**, *17,* 398–402.
37. Ripa, R. Survey and use of biological control agents on Easter Island and in Chile. In *Biological Control of Muscoid Flies*; Patterson, R.S., Rutz, D.A., Eds.; Misc. Publ. Entomol. Soc. Am. 61; 1986; 39–44.
38. Ripa, R. Biological control of muscoid flies in Easter Island. In *Biocontrol of Arthropods Affecting Livestock and Poultry*; Rutz, D.A., Patterson, R.S., Eds.; Westview Press: Boulder, CO, 1990; 111–119.
39. Geden, C.J.; Axtell, R.C. Predation by *Carcinops pumilio* (Coleoptera: Histeridae) and *Macrocheles muscaedomesticae* (Acarina: Macrochelidae) on the housefly (Diptera: Muscidae): Functional response, effects of temperature and availability of alternative prey. Environ. Entomol. **1988**, *17,* 739–44.
40. Mullens, B.A.; Meyer, J.A.; Mandeville, J.D. Seasonal and diel activity of filth fly parasites (Hymenoptera: Pteromalidae) in caged-layer poultry manure in southern California. Environ. Entomol. **1986**, *15,* 56–60.

55

Pest Management: Ecological Agriculture

Barbara Dinham

Introduction

Strategies to develop ecological farming methods stem from economic, health, environmental, and practical concerns with chemical pesticides. Ecological agriculture, also called sustainable agriculture,[1] agroecology, low-external input, regenerative agriculture,[2] and farmer-participatory integrated pest management (IPM), has no single definition.[3] The Food and Agriculture Organisation (FAO) of the United Nations suggests that such sustainable development (in the agriculture, forestry, and fisheries sectors) conserves land, water, plant, and animal genetic resources, is environmentally nondegrading, technically appropriate, economically viable, and socially acceptable.[4]

Ecological agriculture implies approaches to prevent or minimize application of chemical pesticides, which promote local inputs, and where increased farmer knowledge becomes the basis of managing pests and improving yields and sustainability.[5] Organic agriculture is distinctive in requiring an approved body to certify that no chemical fertilizers or pesticides have been used. This provides a guarantee for consumers, and can enable farmers to receive a premium for their crops.

Pest Management Strategies

Ecological pest management practices have been adopted in both industrialized and developing countries. Most farming systems still use certain aspects: crop rotation, field clearance to destroy pest refuges, resistant varieties, early or late planting regimes. More specific ecological strategies employed by farmers vary widely, according to the cropping system, whether the farmer is in an industrialized or a developing country, and the locally available inputs.

As interest in ecological pest management grows, it is increasing the demand for biological technologies. The biological control agent in most widespread use, *Bacillus thuringiensis* (Bt) can be easily produced in large quantities, and can be used like a chemical. As a result it is applied in both conventional and organic agriculture in many countries. Bt has proved extremely effective in controlling persistent pests such as the diamondback moth, which plagues cabbages and related crops, and has led to a pesticide treadmill in many areas.[6]

Biological pest controls require the development of breeding centers for insect predators and parasitoids. Cuba has the most advanced program globally with a countrywide network of over 300 centers for the reproduction of entomopathogens and entomophages supplying bacteria, fungi, viruses, and insect parasitoids. These include *Lixophaga diatraege* for cane stem borer, the parasitic wasp *Trichogramma*, and the fungal disease *Beauvaria bassiana* against a total of seven pests.[7]

Particularly in developing countries, many indigenous plants and locally adapted technologies are used against a variety of pests. Most widely known is the neem tree (*Azardirachta indica*), native to India and also found in parts of Africa, whose leaves are effective against many pests. Other common solutions are pyrethrum, chili peppers, wood ash, and castor oil seeds.[8,9]

Spreading Ecological Practices

Ecological strategies for pest management are characterized by a holistic approach and not only the substitution of biological controls for chemical pesticides. Management strategies include soil conservation, seed selection, and maintenance of agricultural biodiversity. Participation of the women and men farmers to ensure cultural and local appropriateness is of central importance in the development of new strategies. Farmer field schools (FFS) have proved a successful training approach, where training takes place at field level, focusing on recognition of pests and predators and their life cycles. Designing in-field experiments, examining economic losses from pest damage, and encouraging observation of local plants that may act as a trap crop for pests or have repellent properties, become part of the armory.[10]

Ecological Diversity for Diverse Cropping Systems

Fragile tropical soils may benefit most from ecological strategies, but these farming approaches are not restricted by crop, climate, or continent. Table 1 indicates a range of cropping systems and countries that have benefited from IPM programs. One of the most successful examples is in European glasshouses where 70% of commercial glasshouses have been managed through IPM based on biological controls for more than 15 years. In these intensive production systems, IPM enables growers to control all major pests and avoid most pesticide use.

Cotton is the crop that uses the most chemical insecticides worldwide. Nevertheless, cotton IPM and organic strategies have demonstrated that ecological alternatives are successful. An FFS trainer- and farmer-training program in Pakistan prevented insecticide applications in the first 8–10 weeks after planting, allowing natural enemy populations to build up, and giving higher yields in seven of the 10 demonstration plots.[11] Similar successes have been achieved in India, Zimbabwe, and elsewhere.

Adverse environmental impacts of pesticide use in conventional cotton systems and the problems of insect resistance to pesticides have encouraged farmers to invest in organic systems. While still a small proportion of overall cotton production, the growing consumer support for ecological fiber is likely to further encourage producers. The United States is the largest single producer with 32% of the total certified organic cotton fiber production in 1997, followed by Turkey with 22%. Organic cotton production is well suited to small-scale cropping systems and in the same year 15% of certified organic cotton fiber came from India, 19% from Africa, and 11% from Latin America (mainly Peru).[12]

In Indonesia IPM strategies based on improving farmers' knowledge of ecological pest management were highly successful against infestations of brown rice-hopper that decimated the rice crop in the mid-1980s. An IPM rice program supported by the FAO in South and South East Asia that began in 1980 has targeted farmers using high chemical inputs. More than 500,000 farmers have been trained and now save on average $10 per hectare per season, while maintaining or increasing yields.[13]

Farmers develop their own strategic improvements. In an area of low rainfall and high soil erosion in Burkina Faso, local groups and villagers worked with the government and local organizations to develop an ecological approach ranging from tree planting to increased use of manure. Covering more than 200 villages, farmers increased sorghum yields from 870 kg/ha to 1650–2000 kg/ha. In the semi-arid

TABLE 1 Impact of Selected IPM Programs on Pesticide Use, Crop Yields, and Annual Savings

Country and Crop	Average Change in Pesticide Use (as % of Conventional Treatments)	Changes in Yields (as % of Conventional Treatments)	Annual Savings of Program $1000
Togo, cotton	50	90–108	11–13
Burkina Faso, rice	50	103	No data
Thailand, rice	50	No data	5–10,000
Philippines, rice	62	110	5–10,000
Indonesia, rice	34–42	105	50–100,000
Nicaragua, maize	25	93[a]	No data
United States, nine commodities	No. of applications up, volumes applied down	110–130	578,000
Bangladesh, rice	0–25	113–124	No data
India, groundnuts	0	100	34
China, rice	46–80	110	400
Vietnam, rice	57	107	54
India, rice	33	108	790
Sri Lanka, rice	26	135	1,000

[a] Lower yields, but higher net returns.

Source: Pretty.[2]

Machakos region of Kenya farmers developed appropriate agricultural techniques, building terracing, selective animal grazing, and manure collection, significantly increasing soil fertility and yields.[14]

The major physical constraints to adoption of ecological pest management relate to agricultural production systems rather than crops, regions, and climate. Large-scale monoculture, for example, does not easily lend itself to ecological approaches because of dependence on single varieties, loss of natural soil fertility, and emergence of specialized pests that kill natural enemies. The most sustained improvements can be found when government policies support ecological practices.

Environmental and Economic Benefits

Because of the many starting points, farmers have different motivations for adopting ecological approaches. In industrialized countries, where farmers generally rely heavily on external inputs, ecological practices may lower yields, but improve the environment.

In developing countries, some 2.3–2.6 billion people are supported by agriculture using the higher inputs of green revolution technologies and in these areas farmers would generally stabilize or achieve slightly higher yields with ecological pest management, also gaining environmental benefits. The remaining 1.9–2.2 billion people are largely fed by traditional agriculture and farmers. Here, ecological management strategies can substantially improve yields and income.[15]

The environmental benefits of ecological agriculture stem from reductions in chemical inputs, which threaten biodiversity, pollute water sources, and kill fish and other nontarget beneficials, often including cattle or domestic animals. Some entomologists believe that a significant proportion of the most serious insect pest problems have been introduced or worsened as pesticide use eliminates local natural enemies.

Economically, dependence on pesticides and poor management strategies can have devastating impacts, and the health and environmental costs of pesticide use are rarely calculated. In India cotton production accounts for more than 50% of pesticides used and poor application practices have resulted in insect pest resistance to chemicals. Farmers lacked the know-how to develop alternatives, and became deeply indebted to money lenders and pesticide dealers. The FAO has estimated that IPM could reduce pesticide use in Asian rice crops by 50% or more, without compromising yields, and maintain or improve net returns to the farmer—savings could amount to $1 billion.[16]

In areas of fragile and problem soil, intensive agriculture can lead to desertification, while resource-conserving systems in arid and semi-arid regions can deliver sustainable and often increasing yields. Farmers in these regions can rarely afford the external inputs and small improvements in yield would have profound impacts on the lives of the largely poorer populations that these agricultural systems support.

Risks in Adopting Ecological Strategies

The risks involved in adopting ecological pest management strategies are lower for small-scale farmers who are not yet using chemical inputs, and for farmers on a pesticide treadmill caused by high dependence and pest resistance to chemical inputs. In these instances the main risks are from short-term projects imposed without involvement of women and men in the farming communities who may swap one kind of dependency for another—the goodwill of donors. Farmers need knowledge to transfer to more ecological pest management strategies, and even successful IPM training can be limited by lack of funds for follow up.

Farming systems in industrialized countries that are effectively managing pesticide usage are likely to lose yields when adopting ecological strategies, particularly during the transitional stage. These farmers are also likely to face higher labor costs, though offset to some extent by lower input costs. The benefits are longer term and less tangible: improvements to the environment and to health, and in maintaining yields over time. For farmers who adopt organic practices and seek certification, the initial lower yields can be offset by a premium for their crops.

References

1. In *Promoting Sustainable Agriculture and Rural Development, Agenda 21* Ch. 14, UNCED United Nations Conference on Environment and Development, Rio de Janiero, 1992. http://www.igc.apc.org/habitat/agenda21/ (accessed January 2001).
2. Pretty, J. Local Groups and Institutions for Sustainable Agriculture. In *Regenerating Agriculture: Policies and Practice for Sustainability and Self-Reliance*; Earthscan Publications: London, 1995; 131–162.
3. UNDP. In *Benefits of Diversity: An Incentive Toward Sustainable Agriculture;* UNDP: New York; 1992.
4. In *The den Bosch Declaration and Agenda for Action on Sustainable Agriculture and Rural Development*, Report of the Conference, FAO/Netherlands Conference on Agriculture and the Environment, FAO and Ministry of Agriculture, Nature Management and Fisheries of the Netherlands, the Netherlands, April 15–19, 1991.
5. Reijntjes, C.; Haverkort, B.; Waters-Bayer, A. Low-External- Input and Sustainable Agriculture (LEISA): An Emerging Option. In *Farming for the Future: An Introduction to Low- External-Input and Sustainable Agriculture*; Macmillan: London, 1992; 2–21.
6. CAB International. In *Global Crop Protection Compendium*; CAB International: Wallingford, U.K., 1999; (CD ROM).
7. Management of Insect Pests, Plant Diseases and Weeds and Soil Management: A Key to the New Model. In *The Greening of the Revolution: Cuba's Experiment with Organic Agriculture;* Benjamin, M., Rosset, P., Eds.; Ocean Press: Melbourne, 1994; 35–65.
8. Stoll, G. Methods of Crop and Storage Protection. In *Natural Crop Protection Based on Local Farm Resources in the Tropics and Subtropics*; Agrecol, Verlag Josef Margraf: Langen, Germany, 1986; 80–167.

9. Elwell, H.; Maas, A. Natural Pest and *Disease Control;* Natural Farming Network: Harare, Zimbabwe, 1995; 3–128.
10. http://www.communityipm.org/ (accessed January 2001).
11. Poswal, A.; Williamson, S. Off the 'Treadmill': Cotton IPM in Pakistan. In *Pesticides News*; 1998 June, 40, 12–13.
12. Myers, D.; Stolton, S. *Organic Cotton: From Field to Final Product*; Intermediate Technology and the Pesticides Trust: London, 1999; 1–120.
13. Farmer First-Field Schools are a Key to IPM Success. In *Growing Food Security: Challenging the Link between Pesticides and Access to Food;* Dinham, B., Ed.; Pesticide Action Network U.K.: London, 1996; 87–88.
14. Tiffen, M.; Mortimore, M.; Gichuki, F. Management and Managers. In *More People, Less Erosion—Environmental Recovery in Kenya Part III;* John Wiley and Sons: Chichester, U.K., 1994; 131–226.
15. Pretty, J. *Regenerating Agriculture: Policies and Practice for Sustainability and Self-Reliance;* (op.cit 2), Earthscan Publications: London, 1995; 1–25.
16. In FAO. *Rice and the Environment: Production Impact, Economic Costs and Policy Implications*, Committee on Commodity Problems, Intergovernmental Group on Rice, Seville, Spain, May 14–17, 1996. FAO: Rome, 1996; 1–13. CCP:RI 96/CRS1.

56

Pest Management: Ecological Aspects

David J. Horn

Introduction

Every animal and plant population is surrounded by an interactive biotic and physical environment, and these enormously complex interactions often stabilize population densities to reduce the probability of a pest outbreak. However, ecological interactions are often disrupted, simplified, or overridden in managed ecosystems, reducing the impact of naturally-occurring pest population regulation and leading to outbreaks. Understanding how ecological interactions may either cause or prevent pest outbreaks can lead to crop and landscape management activities that achieve relative stability of pest populations below damaging levels without resorting to widespread and environmentally disruptive intervention. Increasing environmental complexity can enhance pest management. Intelligent environmental management with due regard for the place of a pest within a complex and interconnected ecosystem reduces pest outbreaks when attention is given to 1) increasing species diversity within cropping or landscaped ecosystems, 2) crop rotations and use of short-term, rapidly-maturing cultivars, 3) reducing field size and encouraging intervening areas of uncultivated and undisrupted vegetation, 4) reduced tillage and tolerance of weedy backgrounds, and 5) increased genetic diversity within the crop.

Pest Population Dynamics and Species Diversity

Apparently simple ecosystems such as annual agricultural crop monocultures are complex, and the relative impact of alternate crops, weeds, natural enemies, competitors, and associated organisms on pests may be highly variable. Pest populations and their effective environments are constantly changing in space and time, and events impacting one population at one location and time interval usually do not duplicate events in the same population at another time and place. Despite ecosystem complexity, seasonal agricultural crops that are periodically disrupted due to harvesting and tilling may never achieve a steady, sustained state typical of later stages in ecological succession. Annual or more frequent disruption selects for pests that can locate and exploit resources quickly and efficiently. Colonizing species of plants and insects display rapid dispersal and an ability to increase numbers quickly when suitable habitat is located. The ancestors of many crop plants are typical of early stages of ecological succession, as are their associated insect pests. Conventional agriculture invites early- successional species that are likely to undergo uncontrolled population outbreaks. Populations of such pests are rarely in equilibrium at any given place and time, but are maintained by a loose balance of colonization and extinction over a large geographical area. Equilibrium is more likely to be reached in populations occupying longer-lasting ecosystems such as orchards and forests. Population fluctuations in these more complex ecosystems are partly buffered by the complex interactions within food webs, so there is less likelihood of outbreak of any particular pest species.

Species diversity is formally measured as an index combining numbers and proportion of each species present in an ecosystem. A species diversity index reflects the number of links in a food web, and the overall stability of any ecosystem is partly a function of the number of interactions among plant, pests, natural enemies, and pathogens. Ecologists and pest managers debate over whether there exists a direct relationship between species diversity and stability of individual populations within an ecosystem. It is often presumed that pest outbreaks are suppressed in more complex (and therefore more diverse) ecosystems. The so-called "diversity–stability hypothesis" holds that ecological communities with a higher species diversity are more stable because outbreaks of pest species are ameliorated by the checks and balances and alternative pathways that exist within a large and integrated food web. Evidence supporting this view comes from experiments that mix several plant species with the primary host of a specialist herbivore, resulting in reduced populations of the specialist herbivore. This observation is termed the *resource concentration hypothesis,* which holds that insect herbivores are more likely to locate and to remain on hosts growing in dense or pure stands, and the most specialized species frequently attain highest densities in ecosystems with low plant species diversity. Biomass becomes concentrated in a few species, with a concomitant decrease in species diversity of herbivores in monocultures. Increases in herbivore populations in crop monocultures generally result from higher rates of colonization and reproduction along with reductions in dispersal, predation, and parasitism. As species diversity increases in agroecosystems, more internal links result within food webs and these links promote greater stability, resulting in fewer pest outbreaks. Structural diversity is an important physical component of overall diversity; for instance, cropping systems with taller plants (such as corn among beans and squash) present more physical space to arthropods and this increases the variety of prey and provides greater shelter for predators.

Monoculture and Polyculture

Monoculture is the planting of a single species of crop plant, which often results in increased populations of specialist herbivores, a result consistent with the resource concentration hypothesis. Polyculture, the planting of more than one crop species in the same local area (often the same field), may reduce impact of herbivorous pests because the presence of a variety of plants disrupts orientation of specialist herbivores to their hosts. For example, cabbage flea beetles and cabbage aphids that locate their hosts via specific chemical cues (such as the alkaloid sinigrin) are less effective in locating their host plants when these are intermingled with a variety of other plant species. The result is lower populations of the flea beetles and aphids. Local movement of cucumber beetles and lady beetles is enhanced when cucumbers are interplanted with corn and beans as compared with these insects' movement in monocultures, where they tend to remain on individual plants. Numbers of specialist herbivores (cabbage aphids, diamondback moth, and imported cabbageworm) on collards planted in weedy backgrounds are lower than populations of these same herbivores on collards planted against bare soil or plastic mulch. This influence of weeds intensifies once the weeds become as tall as the collards, effectively allowing the collards to escape herbivory by concealment among the weeds. The frequent replacement of one crop by another in crop rotation maintains populations of specialist herbivores below damaging levels. Field crop producers in the midwestern U.S. can prevent the increase of corn rootworm populations by rotating from corn to soybeans every 2 or 3 years.

Open and Closed Ecosystems

Ecosystems may be considered to be *open* (subsidized) or *closed* depending on the amount of nutrient and energy exchange with ecosystems outside themselves. Open ecosystems are characterized by regular input of energy and nutrients, followed by removal of a large proportion of nutrients. In maize fields in the midwestern U.S. heavy importation of mineral fertilizer occurs at planting and subsequent energy is input when the crop is tilled and pesticides are applied. Most of the nutrients in a maize field are removed at harvest, as yield or crop residue. In landscaping towns and suburbs we fertilize (providing input) and rake leaves and remove mowed grass (exporting nutrients) to maintain a pleasing appearance.

The assemblage of species within a maize field is artificial, with novel interspecific associations. Many of the major insect pests are of exotic origin, and the association with maize is relatively recent. For instance, the European corn borer arrived in North America around 1910, before which maize (native to Mesoamerica) had no ecological association with the corn borer. It may take many years for native natural enemies to expand their host or prey range to include exotic organisms. Most anthropogenic agroecosystems and landscaped ecosystems are artificial assemblages and open ecosystems. The species assemblage in planned landscapes often includes a preponderance of exotic species.

In a closed ecosystem, such as a deciduous forest, most nutrient movement remains localized with little import and export. For example, the nutrients and energy in the forest canopy fall to the ground as leaves and frass, or leaves are converted into caterpillars, which in turn are eaten by insectivorous birds, predatory insects, and parasitic wasps. There is little "leakage" of nutrients from closed ecosystems into surrounding environments. The species assemblages of many closed ecosystems have been associated for millennia, resulting in multiple trophic links and close ecological associations among mostly native species. Such ecosystems are relatively "immune" to invasion by exotic species (although there are exceptions, such as the successful invasion of the gypsy moth into the forests of eastern North America).

Examples

The Mexican bean beetle overwinters in hedgerows and along field edges, so that soybean fields nearest overwintering sites are likely to become infested earlier and bean beetle populations subsequently will be higher. Soybeans located near bush and pole beans (which are more suitable hosts for the bean beetle) are also likely to develop economically damaging infestations earlier. Natural enemies often move from unmanaged field edges and nearby hedgerows and forests into adjacent farm fields and the nature of this movement may be very important to local suppression of pests. Many studies have shown that there are increased numbers and activity of natural enemies near field borders when there is sufficient natural habitat to provide cover and alternate prey and hosts, as well as food in the form of nectar and pollen. This function of wild border areas significantly enhances biological control.

Weedy vegetation in or near crop fields support a diverse fauna, including natural enemies of pests on the crop plants. This depends on the species of weeds present; for instance, if the weeds were particularly attractive to aphids and their natural enemies, this enhances aphid control on a commercial crop. Weeds such as pigweed *(Amaranthus),* lambs quarters *(Chenopodium),* and shepherd's purse (*Capsella*) when heavily infested with aphids serve as "nurseries" for production of aphid predators and parasitoids, which move onto neighboring crop plants when these are located near these weeds.

Growing several crops in the same space reduces pest problems relative to monocultures of the same species. When blackberries are planted among grapevines in central California, the parasitoid *Anagrus epos* attacks the eggs of both grape leafhopper and the leafhopper *Dikrella cruentata* on blackberry. By encouraging blackberries between alternate grape arbors, a constant supply of eggs of both leafhoppers are available to the parasitoid, which persists in populations high enough to bring the grape leafhopper under biological control.

Floral undergrowth in orchards provides resources to adult parasitic wasps and flies and increases parasitism of phytophagous insects (particularly Lepidoptera) on the trees. The presence of nectar and pollen along with alternate prey is particularly favorable to populations of generalist predators such as the lady beetle *Coleomegilla maculata* resulting in lower aphid populations. In relay cropping, two (or more) different crops are grown on the same area in successive seasons. The seasonal change from one crop to another prevents the increase of specialist pests especially if the crops are evolutionarily distantly related (such as legumes and grasses). When soybeans are relay cropped after winter wheat, pests of soybeans are less abundant than when soybeans are cropped alone.

Planting and harvesting corn and beans in alternating plots rather than solid monocultures reduces pest numbers on both crops, and this management approach is widely practiced in traditional agriculture. Where alfalfa can be grown throughout the year it is possible to harvest on a 3–4 week rotation when half the field is cut in strips. Natural enemies of the alfalfa weevil, alfalfa caterpillar, and aphids are conserved in the regrowth, and there are alternative food sources and hiding places for these predators and parasitoids all year, so they are always present to suppress pest populations below damaging levels. By planting alfalfa adjacent to cotton, control of *Lygus* bugs is achieved by allowing increase of natural enemies of *Lygus* in the alfalfa. These natural enemies move into the cotton and control *Lygus* bugs there.

Bibliography

1. Altieri, M.A. *Biodiversity and Pest Management in Agroecosystems*; Food Products Press: New York, 1994; 185.
2. Collins, W.W.; Qualset, C.O. *Biodiversity in Agroecosystems*; CRC Press: Boca Raton, FL, 1999; 334.
3. Horn, D.J. *Ecological Approach to Pest Management*; Guilford Press: New York, 1988; 285.
4. National Research Council. *Ecologically Based Pest Management: New Solutions for a New Century*; National Research Council: Washington, DC, 1996; 160.

57

Pest Management: Legal Aspects

Michael T. Olexa
and Zachary
T. Broome

Introduction

To promote public health, personal safety, and environmental protection, pesticides and their use are extensively regulated. Pesticide regulation is primarily decided at the federal level and enforced by the Environmental Protection Agency (EPA). Although the states generally copy the federal laws, some state laws are more restrictive than federal laws. Both federal and state laws provide for criminal prosecution and can impose penalties such as fines or imprisonment. In addition, common law actions are used to regulate pesticide use. Common law actions are civil claims brought by private citizens based on an allegation of improper pesticide production or use. When a pesticide manufacturer or user causes personal injury or property damage, the victim initiates a civil action to recover financial compensation for the harm.

This entry will detail the federal regulatory framework of pesticides, including product registration and methods of enforcement. This entry will also address the conflicts between pesticide regulation and other federal environmental law, and the current preemption guideline for conflicts between federal and state regulations. Finally, this entry will address various common law remedies available as private forms of pesticide regulation, to ensure pesticide users and manufacturers act reasonably and limit potential harm to humans and the environment.

Statutory Law and Regulation

Statutory laws are the formal acts of federal and state legislatures. Although statutory laws often provide general directions, technical details are assigned to a regulatory agency. The agency rules specify how the statute passed by Congress will be implemented and enforced by the agency. These rules have the force of law. Regulatory agencies implement their rules by requiring permits or licenses, and enforce those rules through civil and criminal penalties.[1]

Federal Law

Overview

Pesticides were first subject to federal laws by the Insecticide Act of 1910.[2] This law protected farmers from sellers of adulterated or misbranded pesticide products. Following a surge in the development of new pesticides during the Second World War, Congress repealed the Insecticide Act of 1910 and enacted the Federal Insecticide, Fungicide, and Rodenticide Act (FIFRA) of 1947.[3] FIFRA broadened federal control of pesticides by requiring the United States Department of Agriculture (USDA) to register any pesticide before its introduction into interstate commerce. In 1970, Congress transferred the administration of FIFRA to the newly created EPA. Thereafter, Congress enacted the 1972 Federal Environmental Pesticide Control Act (FEPCA) to address public concerns about pesticides' environmental impact and to ensure greater pesticide use regulation.[4] As a result, federal policy shifted from regulation of pesticides for reasonably safe use in agriculture to regulation of pesticides to prevent unreasonable risks to people and the environment. Subsequent amendments to FIFRA (1988, etc.), have clarified the EPA's duties and responsibilities.[5]

The Federal Food, Drug, and Cosmetic Act (FDCA), and 1954 Miller Amendments, required the government to establish tolerances (maximum allowable pesticide residue limits) for all pesticides used on food and feed crops.[6] The 1996 Food Quality Protection Act (FQPA) amended the FDCA and FIFRA, especially the process of establishing tolerances for pesticide residues in food and feed.[7] For example, the EPA now uses a single standard to evaluate pesticide residues on raw and processed foods. FQPA represents the single largest shift in federal pesticide policy and process ever undertaken. The FQPA established a new safety standard to be met when establishing tolerances for a reasonable certainty of no harm from aggregate exposure. It also required all existing tolerances to be reassessed under the new standard and required the EPA to make an explicit determination that tolerances for residues in food are safe for infants and children.[7]

The heart of the regulatory scheme for pesticide application, storage, and disposal is the FIFRA provision making it unlawful to "… use any registered pesticide in a manner inconsistent with its labeling."[8] Thus, courts consider pesticide label instructions legislative regulations having the force of law. For pesticide users, labeling is the primary basis for enforcement of FIFRA. Hence, every person using a pesticide has a legal obligation to read and follow all label instructions attached to the product and all product usage directions contained in any printed materials mentioned on the label.

Pesticide Product Registration

The EPA retains primary oversight for providing guidance to major pesticide producers for what data are needed before the product is federally registered. Some states also conduct a review of scientific data before registering a pesticide in their state. State agencies generally identify product registration violations during their inspections and refer them to EPA for enforcement.[5] However, states also have the ability to regulate pesticide use with state-level registration, although the registration only applies within the regulating state.[9]

The federal government controls which pesticides are put on the market by requiring pesticide registration under FIFRA.[10] Unless exempt, all pesticides must be registered with the EPA administrator before being distributed or sold.[11] A pesticide will not be registered by EPA unless the agency's review of the registration data shows the pesticide will perform its intended function without "unreasonable adverse effects on the environment."[12] Pesticides are registered for general use or restricted use, depending upon the potential for adverse side effects.[13] If EPA determines that a pesticide generally will not have adverse effects on the environment or injure the applicator, then the pesticide is safe for general use.[14] This means that the product can be purchased by an unlicensed applicator. If EPA determines that a pesticide requires additional care or specific training in use, then the pesticide is defined as a restricted use product (RUP).[15] RUPs are subject to additional regulations and limited to use by a certified and licensed applicator.[16] Although many pesticides predated 1947, and were therefore exempt from registration, FIFRA was amended to require the reregistration of all pesticides "containing any active ingredient contained in any pesticide first registered before November 1, 1984."[17]

Most states require the federally regulated products to be registered in each state as well. A state also may regulate the registration of EPA-registered pesticides for additional distribution or use within that state, so long as the registered state use does not violate FIFRA.[18] For instance, a state can make it illegal to use a product registered by EPA, or limit the use of a registered pesticide. Unless the EPA administrator has denied the state's registered use, state registration of pesticides is equivalent to federal registration, but only within that state. State registrations are subject to disapproval by the administrator, who may also suspend a state's registration authority if the state issues registrations that violate either FIFRA or FDCA.[19]

Not all pesticides are regulated by the EPA. If the EPA administrator determines that a pesticide does not require regulation, then the administrator, by regulation, may exempt that pesticide from the registration requirements.[20] The pesticides that do not require regulation are natural substances and minimum-risk pesticides, such as cedar or geraniol, as well as substances treated with pesticides, pheromones, and biological preservatives.[21]

Pesticide Use Enforcement

Although EPA does have enforcement power, in many cases EPA delegates enforcement of pesticide use violations to the states, under cooperative agreements. Each state's agencies are responsible for administering pesticide regulations under state and federal statutes. States occasionally refer pesticide use violation cases to the EPA for enforcement.[22]

Enforcement of pesticide regulations is based on a hierarchy of state and federal agencies, beginning with the EPA Office of Pesticide Programs (OPP). OPP is the EPA entity responsible for registering all pesticides and formulating the label requirements that restrict pesticide use.[23] EPA in Washington works through 10 regional field offices (Regions) of EPA, which serve as liaisons between the states and OPP. The Regions work with the states to develop cooperative agreements, examine cases, answer questions, and conduct joint inspections.

Although EPA is the ultimate arbiter of pesticide regulation, states have the primary responsibility to enforce those regulations.[24] Each state has a designated lead agency responsible for meeting the OPP goals. The lead agencies investigate all major pesticide use sites, such as agricultural and urban operations, to ensure that registered pesticides are used according to the labels approved by OPP. The lead agencies are also responsible for investigating sites of pesticide misuse and taking appropriate enforcement information.[24]

Some Native American tribes have pesticide lead agencies that function the same as state lead agencies, and those tribes work with their EPA regions to develop pesticide field programs. Although the tribal enforcement programs are created in the same manner and with the same goals as state programs, OPP and the regions must treat the process with a high degree of deference to a tribe's sovereign treaty rights.[25]

If a pesticide user or manufacturer violates any FIFRA regulation, EPA and the state can stop the sale or use of the pesticide, seize the pesticide, and seek civil and criminal penalties.[26] If a pesticide manufacturer or distributor violates FIFRA, EPA can seek a $5000 civil penalty for each offense, or a criminal penalty of a $50,000 fine or 1-year jail term.[27] If a pesticide user violates FIFRA, EPA can seek a $1000 civil penalty for each offense, or a criminal penalty of a $1000 fine or 30-day jail term.[27] The most drastic enforcement procedures, in regard to pesticide manufacturers, are the cancellation or modification of a pesticide registration. Once a pesticide is registered under FIFRA, the applicant is not guaranteed permanent registration or classification. If EPA determines that a registered pesticide, or its labeling, does not comply with FIFRA or causes unreasonable adverse effects on the environment, the administrator may cancel the pesticide's registration, change its classification, or alter the label.[28]

Enforcement of Tolerances for Pesticides in Food

One important consideration in regulating pesticide use is the level of pesticide residue left on agricultural products that can be consumed by humans. The Food and Drug Administration (FDA) monitors raw and processed commodities for compliance with residue tolerances. The USDA monitors meat, milk, and eggs for residue tolerance compliance. The FDA and USDA programs cover both domestically produced and imported commodities. A few states, such as California and Florida, have additional residue monitoring and enforcement programs, but all states have the authority to enforce federal regulations against food located within each state.[29] States cannot impose regulatory limits on pesticide residue, however, unless the residue limits are identical to the federal levels or the state successfully petitioned the federal government for authorization to change the regulation.[30]

Under the FDCA, introducing adulterated food into interstate commerce is prohibited.[31] A claim for violating the FDCA can be brought in an appropriate federal district court, where the court can punish the guilty party with injunctive relief, criminal prosecution, or seizure.[32] The key for pesticide regulation is whether food containing pesticide residue constitutes adulterated food. Normally, a food is adulterated if it contains any poisonous substance or chemical.[33] Pesticide chemical residues, however, are specifically exempted from blanket regulation.[34] If the EPA has established a tolerance level for a pesticide, or exempted a pesticide from regulation, then the food product is not considered adulterated.[35] The EPA administrator can only establish a pesticide tolerance if consumption of the residue poses little risk to the consumer.[36] The consumption is safe so long as "there is a reasonable certainty that no harm will result from aggregate exposure to the pesticide chemical residue."[36] Once the EPA has established the tolerance level for a pesticide residue, the FDA enforces compliance because noncompliance residue that exceeds established tolerances constitutes food adulteration.

The FDA not only monitors U.S. pesticide residue but also develops cooperative agreements with foreign countries that export agricultural products into the United States.[37] These agreements require imported food to meet the FDCA standards for pesticide tolerance.[37]

While the EPA determines if the amount of pesticide residue found in the food is of "negligible risk" or below tolerable levels, there is a danger from consuming certain pesticide residues that the EPA has deemed carcinogenic.[38] Consumption of pesticide residue is also dangerous because of the pesticides that act as endocrine- disrupting chemicals (EDCs).[39] Evidence has established a link "between EDCs and decreased sperm counts; breast, testicular, and prostate cancer; and neurological disorders."[39] In fact, many of the chemicals found in pesticide residue have been banned in the European Union because of the danger to humans.[40]

Pesticide Regulation Interaction with Other Federal Regulations

One problem for the EPA is registering pesticides that conflict with other federal environmental law the EPA is charged with enforcing, especially the Endangered Species Act (ESA) and the Clean Water Act (CWA).[41]

Under section nine of the ESA, all persons, including federal agencies, are prohibited from causing the "take" of any endangered species.[42] While pesticides may threaten endangered species, EPA did not have a formal process to evaluate the potential impact of pesticides on threatened and endangered species, but EPA had never been liable for registering the pesticides.[43] In 1989, however, the Eighth Circuit held the EPA liable for approving the registration of a pesticide that later harmed a protected species.[44] According to the Eighth Circuit, an action solely based on pesticide regulation ordinarily should be brought under FIFRA, but a claim based on pesticide harm to an endangered species could be brought under the ESA.[45] Because EPA did not dispute the distributed pesticide caused the death of an endangered species, killing an endangered species is a taking, and pesticide distribution cannot legally occur without an EPA registration, the Eighth Circuit held EPA liable for a taking.[46]

Similarly, section seven of the ESA requires all federal agencies to conserve endangered species, which usually requires a consultation with the Fish and Wildlife Service or National Marine Fishery Service to ensure agency actions will not "jeopardize the continued existence of any endangered species."[47] In 2004, the Ninth Circuit ruled EPA violated section seven of the ESA by failing to obtain a "jeopardy" consultation before registering pesticide uses around listed salmon habitat.[48] EPA argued that any cancellation or modification of a pesticide's use must conform to FIFRA, and FIFRA's standard for registration and cancellation already accounted for an effect on listed species.[49] According to the Ninth Circuit, however, FIFRA does not exempt EPA from the ESA regulations, since the two statutes have different purposes, and EPA must comply with the ESA if its registration of pesticides will affect listed species.[49] In addition, EPA has an ongoing duty to ensure all registered pesticides do not violate the ESA, because EPA always has the discretion to cancel a pesticide's registration.[49]

EPA also has to consider the impact that registered pesticides may have on waterways, to ensure the pesticides will not violate the CWA.[50] Under the CWA, any discharge of a pollutant into the nation's waters requires the discharging entity to obtain a National Pollutant Discharge Elimination System (NPDES) permit.[51] The pesticide manufacturer claimed that EPA did not include a requirement for a NPDES permit in the FIFRA registration process, and since water-related usage restrictions were placed on the pesticide as part of the FIFRA registration, the manufacturers did not need additional permit- ting.[52] According to the Ninth Circuit, however, the CWA and FIFRA have different and complementary purposes, so the statutes should be treated as distinct entities.[52] FIFRA's objective is to protect human health and prevent environmental harm from pesticides through cost–benefit analysis, while the CWA's objective is to restore and maintain waterways by limiting pollution discharge.[52] Therefore, the court ruled that pesticide users are not exempt from CWA permitting requirements because they obtain FIFRA registration.[52]

This dual responsibility creates a tension for EPA, which is responsible for establishing water quality standards that dictate the permit requirements for a CWA permit.[53] Until recently, EPA had a Final Rule that exempted all FIFRA-compliant pesticides from NPDES permitting when those pesticides were put into water, or ran off property into water, as part of the pesticide's intended use.[53] The Sixth Circuit rejected this analysis, citing the CWA requirement that no pollutant could be discharged into any navigable water unless EPA determined the discharge would not cause undue harm to water quality and issued the discharger a NPDES permit.[54] Although the court agreed FIFRA-compliant chemical pesticides did not require a NPDES permit if put into water to work as a pesticide, because the pesticide was not being discharged as a waste and EPA only approves pesticides that "will not generally cause unreasonably adverse effects on the environment," the court cautioned that any introduced pollutant discharged with excess chemicals or residue required a NPDES permit.[55] In addition, all biological pesticides, regardless of whether or not it is excess pesticide or residue, always require NPDES permits because the CWA requires a permit for all biological material, not just waste.[56] While the court rejected EPA's Final Rule exempting pesticides from NPDES permitting, the court noted EPA and state authorities may grant general permits that allow for the discharge of a specific pollutant or type of pollutant across an entire region.[57] Once EPA or a state agency issues such a general permit, pesticide dischargers do not have to meet any additional requirements to comply with the CWA.[57] This issue is currently being worked on between EPA and the states, as they try to implement a national and state plan.

Preemption

Because of the interaction between state and federal regulations, preemption is a fundamental concern for pesticide regulation. Under FIFRA, the states are allowed to regulate the sale and use of federal pesticides, but precluded from passing additional regulations on pesticide labels.[58] This preemption applied to both state regulations and common law tort claims that "directly or indirectly attacked the adequacy of the warnings on the EPA approved pesticide label."[59]

Background

When a manufacturer submits an application to register a new pesticide, the manufacturer must include a proposed label, and the EPA will only register the pesticide if the data support the registration and its label complies with the statute's misbranding prohibition. A pesticide is "misbranded" if its label contains a statement that is "false or misleading" or lacks adequate instructions or warnings.[60] Although preemption was originally restricted to state regulations, in 1992, the federal courts began applying preemption to common law tort claims.[61] Although the Supreme Court never ruled on FIFRA-specific preemption, after *Cipollone,* most state and federal courts would reject any common law claim that was premised on "failure to warn."[62] A common test for preemption, which led to *Bates v. Don Agrosciences, LLC,* was whether it was reasonably foreseeable that a tort claim against a pesticide manufacturer would "induce it [the manufacturer] to alter its product label."[63] If the litigation would give the manufacturer a "strong incentive" to change the pesticide label, then the claim was preempted.[63]

Current Law

In *Bates V. Don Agrosciences, LLC,* the U.S. Supreme Court held that FIFRA's express preemption clause did not preempt state common law claims merely because an adverse judgment on the claim might impel a manufacturer to change a product label.[64] According to the Court, "private remedies that enforce federal misbranding requirements would seem to aid, rather than hinder, the functioning of FIFRA" because the claims serve as catalysts for evaluating harm and pesticide performance.[65] Although the Court acknowledged the lack of uniformity in national FIFRA regulation, the Court maintained that Section 136v still prevents states from imposing labeling requirements different from FIFRA.[66]

Applying *Bates* to future pesticide litigation and regulation will be necessarily fact specific. In *Bates,* the Court parsed the plaintiffs' tort claims against the pesticide manufacturer into two categories, design defect and failure to warn.[67] The Court held that design defect claims, formerly preempted under the "inducement" test, were not preempted because a successful claim would incentivize a manufacturer to change its label or ingredients, but the claim placed no requirement on the labeling.[67] The plaintiffs' fraud and negligent failure-to-warn claims, however, were premised on common law rules that qualify as "requirements for labeling and packaging," but the Court established that Section 136v(b) only prohibits state law labeling and packaging requirements that are "in addition to or different from" the labeling and packaging requirements under FIFRA.[66] Therefore, any preemption analysis begins with the plaintiff's claim and "calls for an examination of the elements of the common law duty at issue."[68] After *Bates,* only tort claims that impose liability for a labeling requirement that is in addition to or different from EPA labeling requirements would be preempted.[59]

Common Law

Common law actions do not depend upon statutes for their authority. Instead, common law arises from the generalized legal duty individuals in a law-abiding society owe to one another. Every adult person is obligated to a certain duty of care for the personal and property rights of others. A violation of this obligation can become a basis for a common law action. Common law theories generally encountered in actions resulting from pesticide use include negligence, trespass, nuisance, and strict liability.[68]

Negligence

Negligence is a legal standard applied to an individual who fails to act in a reasonably prudent manner.[69] To establish a negligence claim, the plaintiff must prove that the defendant owed the plaintiff a duty, the defendant breached that duty, the breach was the direct and proximate cause of harm to the plaintiff, and plaintiff's harm actually resulted in compensable damages.[69] Negligence actions can be used as regulatory tools against both pesticide users and the pesticide producers, to ensure pesticides are used reasonably to prevent damage to human health and the environment.[70]

A pesticide user can be liable for negligence if the user fails to exercise a reasonable duty of care in applying a toxic pesticide, and others are injured as a direct result of the negligent behavior.[70] A pesticide manufacturer can be liable for negligence if the manufacturer's label is defective, or if the manufacturer's label constitutes a failure to warn, and a person is injured because of the improper label.[67] The common law duty for pesticide use or production generally stems from the industry-established standard of care. However, because pesticides are heavily regulated at the state and federal level, statutes or administrative regulations generally define the applicable standard of care in a jurisdiction. These regulations also create a cause of action based on negligence per se, so any violation of a law regulating the manufacture, label, or use of a pesticide that results in harm is negligence.[71] In making a negligence claim for pesticide use, res ipsa loquitur can be a successful argument in cases where a pesticide, used as intended, was the direct cause of an injury.[72] Res ipsa can be difficult to prove, however, since many factors other than negligence of a manufacturer or user could cause or contribute to such injuries.

In negligence cases based on pesticide use, breach of duty and causation often relate to the same action—the pesticide dispersal. Pesticide users have a duty to act carefully when applying pesticides, so as to avoid applying pesticide on unintended targets. If a pesticide user does not breach that duty, then it is unlikely the pesticide caused a plaintiff's injury. To prove causation, the plaintiff must prove exposure to the pesticide and have an expert testify that exposure to the chemicals in the pesticide caused plaintiff's injury.[73] Generally, an expert can establish causation by using clinical studies where a certain level of pesticide caused certain chemical reactions, such that plaintiff's alleged exposure to that pesticide would likely result in plaintiff's injury. Experts can also establish causation by diagnosing the plaintiff's symptoms, and then eliminating all alternatives except the pesticide exposure.[70]

In actions based on pesticide use, damages are usually related to personal injury or property damage. Personal injury damages result from exposure to the toxic chemicals in pesticides and usually relate to health problems with the lungs and respiratory systems, skin rashes, swelling, discoloration, and lesions, as well as nervous system disorders, such as numbness or headaches, confusion, and memory loss.[70] Agricultural pesticide users often sue when a product malfunctions, so property damages are usually for harm to currently growing crops, although in certain cases, damages can be sought for the loss of property utility, if the pesticide causes property to lose agricultural value.[70] The same injury and damage rules apply to livestock and other "farmed" animals.[70]

Trespass

Trespass is an unauthorized entry onto the property of another by a person or thing that causes damage.[74] Pesticide application can result in liability for trespass if the pesticide, its residue, or container becomes deposited on another's land (through dumping, drift, runoff, incineration, or other means) and causes substantial damage to the property. Although trespass normally requires intent for liability, if the pesticide reached a plaintiff's property because the pesticide user behaved negligently or used the pesticide in an abnormally dangerous manner, then there is no requirement to prove intent.[75] The invasion element of pesticide occurs by either airborne drift or migration through soil or water.[76]

Nuisance

A nuisance is substantial interference with another's use and enjoyment of land.[77] A nuisance lawsuit requires no physical invasion, only a substantial interference with the possessor's enjoyment of land. Pesticide use resulting in offensive odors can be grounds for a nuisance suit. In almost all states, state "right-to-farm" statutes provide limited protection from nuisance actions. The farmer's "right-to-farm" defense is limited to nuisance actions.[78] Nuisance claims can be brought as either private or public actions, although most pesticide-based claims are for private nuisance, because a plaintiff cannot collect damages for a public nuisance claim.[79]

A private nuisance is a nontrespassory invasion of another's interest in the private use and enjoyment of land.[80] Generally, property owners bring private nuisance claims because some property use is prevented by a neighboring property owner's use of pesticides.[81] A private nuisance claim is often used to enjoin a defendant from continuing a certain action, although damages are often sought as compensation for harm already done by the defendant. Because nuisance is a common tort, and has been regularly applied to instances of pesticide drift, most courts will not bar a provable claim for nuisance. In addition, when the plaintiff is unable to prove which pesticide user out of several caused his or her particular harm, courts often allow the plaintiff to proceed on joint liability, so the defendants share the damages burden. However, a nuisance claim requires an ongoing pattern of pesticide application, so a single pesticide use is not actionable under this particular theory.[81]

A nuisance suit against a farmer for the use of pesticides often violates state right-to-farm laws.[78] Right-to-farm laws protect farmland in a variety of ways, including reducing farmer exposure to liability for common law claims based on farming operations.[78] One protection is the coming to the nuisance defense, which prevents a plaintiff from moving near a farm then claiming harm from an ongoing farm activity.[81] The coming to the nuisance defense only protects the farmer from reasonable actions, however, so using pesticides in a negligent manner could result in liability. In addition, coming to the nuisance only protects farm activities that were ongoing before the plaintiff's arrival. If the farmer did not apply pesticides until after the plaintiff's arrival, then the farmer could be subject to nuisance liability.[81]

Strict Liability

Some states hold pesticide applicators absolutely responsible for their pesticide application activities, regardless of fault, without a showing of negligence. This is known as strict liability.[82] Strict liability is normally associated with inherently dangerous or ultrahazardous activities. Whether an activity is inherently dangerous, and therefore subject to strict liability, usually depends on the degree of risk, likelihood of serious harm, inability to eliminate risk, commonness of usage, appropriateness of activity to area, and its value to the community.[83] Most misuse of pesticide claims are brought under the theory of negligence, but a few states allow strict liability claims against the aerial spraying of pesticides.[81]

A common strict liability claim is for a manufacturing defect in the pesticide, where some contaminant in the pesticide caused damage during the pesticide use.[84] Because FIFRA registration requires a manufacturer to register the exact chemical makeup of a pesticide product, the presence of any other chemical in the pesticide is a statutory violation that exposes the manufacturer to strict liability.[85] To bring a strict liability claim against a pesticide manufacturer, a plaintiff must be able to prove that he or she suffered injuries as a result of the use of a pesticide, the defendant was engaged in the manufacture and sale of the harmful pesticide products, the plaintiff purchased the particular pesticide that caused the alleged injuries, the pesticide had not been contaminated after leaving the manufacturer, and the pesticide was unreasonably dangerous to any user.[86]

Regulatory Trends

For the foreseeable future, current and emerging societal issues will continue to stimulate regulatory action. EPA's regulatory focus is found in the National Pesticide Field Program, which establishes EPA's guideline for action to meet the statutory requirements to protect human health and the environment. Currently, EPA regulations focus on worker safety, water quality, and endangered species protection, and promoting pesticide stewardship.[87]

EPA regulation of worker safety is done to minimize occupational risks from pesticides, as outlined in the Worker Protection Standard. The potential danger of pesticide exposure is especially prevalent in agricultural operations, because workers have high occupational exposure to pesticides. In large-scale commercial agriculture operations, workers are often in such close contact with pesticides that they develop serious health problems.[88] Regulating worker safety also gives regulatory bodies an additional "watchdog" group of on-site workers trained in appropriate pesticide manufacture and application, as EPA hopes that the educated workers will report pesticide misuse to mitigate health risks.[89]

Another regulatory trend is reducing environmental risks from registered pesticides, in order to reduce EPA liability under federal environmental protection statutes. Courts have repeatedly held that EPA has a duty to comply with complementary environmental statutes when issuing FIFRA registrations, so two of the focal points for EPA regulation are compliance with the ESA and CWA water quality controls.[89]

Finally, EPA promotes better performance regulation as a part of an incentive plan to encourage agricultural pesticide users to move away from high-risk pesticides. EPA is currently collecting funds for grants and collaborating with the USDA to establish the best environmentally and socially sound pesticide use policies, and EPA plans to implement those policies by giving incentives to pesticide users. To establish which users are successful, and deserve grants, EPA is continually altering alternative regulations to establish baseline performance standards.[89]

Conclusion

To promote public health, personal safety, and environmental protection, pesticides and their use are extensively regulated. Pesticide regulation is primarily decided at the federal level, and enforced by the EPA. Although the states generally copy the federal laws, some state laws are more restrictive than federal laws. Both federal and state laws provide for criminal prosecution and can impose penalties such as fines or imprisonment. In addition, common law actions are used to regulate pesticide use. Common law actions are civil claims brought by private citizens based on an allegation of improper pesticide production or use. When a pesticide manufacturer or user causes personal injury or property damage, the victim can initiate a civil action to recover financial compensation for the harm.

Regulating the manufacture and use of pesticides is necessary to protect human health and the environment, but those regulations require enforcement. Although EPA penalties are a deterrent, the specters of cancelled registration or toxic tort litigation offer successful alternatives, especially in light of the Supreme Court's relaxation of the limitations on state tort claims. In addition, the recent bout of environmental litigation compelling increased EPA enforcement of pesticide regulations has clearly impacted EPA's model of enforcement, which will hopefully lead to overall social benefit.

Acknowledgments

Special thanks to Fred Whitford, coordinator of Purdue Pesticide Programs, and Dr. Fred Fishel (University of Florida/Institute of Food and Agricultural Sciences) for reviewing this entry.

References

1. Olexa, M.T.; Leviten, A., Eds. *Circular 1139: Farm and Ranch Handbook of Florida Solid and Hazardous Waste Regulation;* University of Florida Cooperative Extension Service, 1999; 10.
2. *The Insecticide Act of 1910,* Pub. L. No. 6–152, 36 Stat. 331, 1910.
3. *Fungicide and Rodenticide Act of 1947,* Pub. L. No. 80–104, 61 Stat. 163, 1947.
4. *The Federal Environmental Pesticide Control Act of 1972,* Pub. L. No. 92–516, 86 Stat. 973(codified at 7 U.S.C. § 136, 1988; see also 7 U.S.C. §§ 135–135y, as amended 7 U.S.C. §§ 136 et seq., 2009.
5. Olexa, M.T.; Kubar, S.; Cunningham, T.; Meriwether, P., Eds. *Bulletin 311: Laws Governing Use and Impact of Agricultural Chemicals: An Overview*; University of Florida Cooperative Extension Service, 1995; 7.
6. 21 U.S.C. §§ 301 et seq. seq., 2009.
7. *Food Quality Protection Act of 1996,* Pub. L. No. 104–170, 110 Stat. 1489, 1996.
8. 7 U.S.C. § 136 j(a)(2)(G), 2009.
9. 7 U.S.C. §§ 136v (a), (b), (c)(1), 2009.
10. 7 U.S.C. §§ 136 et seq., 2009.
11. 7 U.S.C. §§ 136a (a)–(b), 2009.
12. 7 U.S.C. § 136a (c)(5)(C), 2009.
13. 7 U.S.C. § 136a (d)(1) (A–C), 2009.
14. 7 U.S.C. § 136a (d)(1) (B), 2009.
15. 7 U.S.C. § 136a (d)(1) (C), 2009.
16. 7 U.S.C. § 136a-1 (a), 2009.
17. 7 U.S.C. § 136v (a), 2009.
18. 7 U.S.C. § 136v (c)(1), 2009.
19. 7 U.S.C. § 136v (c)(3–4), 2009; see also 40 CFR §162 (D) (regulations governing state pesticide registration authority and suspension, state registration procedures, and the disapproval of state registrations).
20. 7 U.S.C. § 136w (b), 2009; 40 CFR § 152.20, 2010 (pesticides regulated by another federal agency); 40 CFR § 152.25, 2010 (pesticides that don't require regulation).
21. 40 CFR § 152.25, 2010.
22. Olexa, M.T.; Kubar, S.; Cunningham, T.; Meriwether, P., Eds. *Bulletin 311: Laws Governing Use and Impact of Agricultural Chemicals: An Overview*; University of Florida Cooperative Extension Service, 1995; 11.
23. EPA. *Pesticide Field Programs: Ensuring Protection through Partnerships,* 2006; 5–6.
24. EPA. *Pesticide Field Programs: Ensuring Protection through Partnerships,* 2006; 16.
25. 7 U.S.C. §§ 136a (a)–(b), 2009.
26. 7 U.S.C. §§ 136k–l, 2009.
27. 7 U.S.C. § 136l, 2009.
28. 7 U.S.C. § 136d, 2009.
29. Olexa, M.T.; Kubar, S.; Cunningham, T.; Meriwether, P., Eds. *Bulletin 311: Laws Governing Use and Impact of Agricultural Chemicals: An Overview;* University of Florida Cooperative Extension Service, 1995; 14–16. [pubmed]; 21 U.S.C. § 337, 2009.
30. 21 U.S.C § 346a (n)(3–4), 2009.
31. 21 U.S.C. § 331 (a), 2009.
32. 21 U.S.C. §§ 332–334, 2009.
33. 21 U.S.C. § 342 (a)(1–2), 2009.
34. 21 U.S.C. § 342 (a)(2)(A–B), 2009; 21 U.S.C. § 346a (a) (2009).
35. 21 U.S.C. § 346a (a)(1) and (4), 2009.
36. 21 U.S.C. § 346a (b)(2)(A), 2009.

37. 21 U.S.C. § 1402 (a), 2009.
38. Eubanks, W.S., II. A rotten system: Subsidizing environmental degradation and poor public health with our nation's tax dollars. Stan. Environ. Law J. **2009**, 213, 277.
39. Watnick, V.J. Our toxics regulatory system and why risk assessment does not work: Endocrine disrupting chemicals as a case in point. Utah Law Rev. **2004**, 1305, 1307–1309, 1310–1311.
40. Pollan, M. *The Omnivore's Dilemma: A Natural History of Four Meals;* Penguin Press: New York, 2006; 178.
41. Angelo, M.J. The killing fields: Reducing the casualties in the battle between U.S. Species Protection Law and U.S. Pesticide Law, Harvard Environ. Law Rev. **2008**, *32,* 95, 97–98.
42. 16 U.S.C. § 1538 (a)(1), 2009; 16 U.S.C. § 1532 (19), 2009.
43. Angelo, M.J. The killing fields: Reducing the casualties in the battle between U.S. Species Protection Law and U.S. Pesticide Law, Harvard Environ. Law Rev. **2008**, *32*, 95,98 –101, 111.
44. *Defenders of Wildlife v. Administrator, EPA,* 882 F.2d 1294 (8th Cir. 1989).
45. *Defenders of Wildlife v. Administrator, EPA,* 882 F.2d 1294, 1296–1299 (8th Cir. 1989).
46. *Defenders of Wildlife v. Administrator, EPA,* 882 F.2d 1294, 1301 (8th Cir. 1989).
47. 16 U.S.C. § 1536 (a)(1–2), 2009.
48. *Wash. Toxics Coal. v. EPA,* 413 F.3d 1024, 1031–1032 (9th Cir. 2005) *cert. denied,* 546 U.S. 1090, 2006.
49. *Wash. Toxics Coal. v. EPA,* 413 F.3d 1024, 1031–1033 (9th Cir. 2005).
50. *Headwaters, Inc. v. Talent Irrigation Dist.,* 243 F.3d 526, 532 (9th Cir. 2001).
51. 33 U.S.C. § 1311 (a), 2009.
52. *Headwaters, Inc. v. Talent Irrigation Dist.,* 243 F.3d 526, 531–532 (9th Cir. 2001).
53. *National Cotton Council of America v. EPA,* 553 F.3d 927 (6th Cir. 2009).
54. *National Cotton Council of America v. EPA,* 553 F.3d 927, 929–932 (6th Cir. 2009); 33 U.S.C. §§ 1311 (a) and 1342 (2006).
55. *National Cotton Council of America v. EPA,* 553 F.3d 927, (6th Cir. 2009).
56. *National Cotton Council of America v. EPA,* 553 F.3d 927, (6th Cir. 2009).
57. *National Cotton Council of America v. EPA,* 553 F.3d 927, 931 (6th Cir. 2009).
58. 7 U.S.C. § 136v (a) and (b), 2009.
59. Mandler, J.P. FIFRA preemption in the wake of *Bates v. Dow.* Nat. Resour. Environ. **2005**, *20*, 64.
60. 7 U.S.C. § 136 (q)(1) and (2), 2009.
61. *Cipollone v. Liggett Group, Inc.,* 505 U.S. 504 (1992) *cited by* Bates v. Dow Agrosciences LLC, 544 U.S. 431, 441, 2005.
62. *Bates v. Dow Agrosciences LLC,* 544 U.S. 431, 441–442, 2005.
63. *Bates v. Dow Agrosciences LLC,* 544 U.S. 431, 436, 2005.66.
64. *Bates v. Dow Agrosciences LLC,* 544 U.S. 431, 443–445, (2005) (jury verdict is not a requirement to change, merely motivation).
65. *Bates v. Dow Agrosciences LLC,* 544 U.S. 431, 451, 2005.
66. *Bates v. Dow Agrosciences LLC,* 544 U.S. 431, 452, 2005 (pre-emption of any statutory or common-law rule that would impose a labeling requirement divergent from FIFRA, but not of state rules fully consistent with federal requirements).
67. *Bates v. Dow Agrosciences LLC,* 544 U.S. 431, 445–446, 2005.
68. Whitford, F.; Olexa, M.T.; Thornburg, M.; Gunter, D.; Ward, J.; Lejurne, L.; Harrison, G.; Becovitz, J., Eds. *Pesticides and the Law: A Guide to the Legal System;* Purdue University Cooperative Extension System: West Lafayette, Indiana, 1996; 12.
69. Restatement (Third) of Torts: Liab. Physical Harm § 6, 2010.
70. Payne, A. Causes of action for damages from pesticides. Causes of Action 2d **2009**, *39,* 579.
71. *Obendorf v. Terra Hug Spray Co., Inc,* 188 P.3d 834 (Idaho, 2008).
72. *Farm Services, Inc. v. Gonzales,* 756 S.W.2d 747 (Tex. App. Corpus Christi 1988) (doctrine requires defect); Eaton Fruit Co. v. California Spray-Chemical Corp., 445 P.2d 437, 440 (Ariz. 1968) (plaintiff

must establish defendant's exclusive control of the pesticide, as well as proper preparation and application of the pesticide); Stone's Farm Supply, Inc. v. Deacon, 805 P.2d 1109, 1114 (Colo. 1991) (supplier sold farmer chemical containing substances unsuitable for farmers' crops).

73. *Rhoten v. Dickson,* 223 P.3d 786 (Kan. 2010); Terry v. Caputo, 875 N.E.2d 72 (Ohio 2007).
74. Keeton, et al., Eds. *Prosser and Keeton on the Law of Torts,* 5th Ed.; Section 16; West Publishing Co.: St. Paul, Minnesota, 1984; 67.
75. Restatement (Second) of Torts § 165, 2010.
76. *Fisher v. Ciba Specialty Chemicals Corp.,* 245 F.R.D. 539, 69 Fed. R. Serv. 3d 119 (S.D. Ala. 2007).
77. Keeton, et al. Eds. *Prosser and Keeton on the Law of Torts*, 5th Ed. Section 87 1984; 622.
78. Fischer, J.; Olexa, M.T.; Leviten, A.; Saju, A., Eds. *Circular 1224: Handbook of Florida Agricultural Laws;* University of Florida Cooperative Extension Service: Gainesville, Florida, 1999; 4.
79. Restatement (Second) of Torts § 821B (2010) (public nuisance requires unreasonable interference with right common to the public); *see also* 58 Am. Jur. 2d, Nuisances § 39 (2010); 58 Am. Jur. 2d, Nuisances §§ 42–47 (private nuisance is an invasion of interest in private use and enjoyment of land).
80. Restatement (Second) of Torts § 821D, 2010.
81. Grossman, M.R. Biotechnology, property rights, and environment. Am. J. Comp. Law **2002**, *50,* 215.
82. Keeton, et al., Eds. *Prosser and Keeton on the Law of Torts,* 5th Ed. Section 75 1984; 536–537.
83. Restatement (Second) of Torts § 520, 2010.
84. Restatement (Third) of Torts: Products Liability § 2.
85. In re DuPont-Benlate Litigation, 859 F. Supp. 619 (D.P.R. 1994).
86. Am. L. Prod. Liab. 3d § 110:1, 2010.
87. EPA. *Pesticide Field Programs: Ensuring Protection through Partnerships,* 2006; 8.
88. Eubanks, W.S. II. A rotten system: Subsidizing environmental degradation and poor public health with our nation's tax dollars. Stan. Environ. Law J. **2009**, *28,* 213, 276.
89. EPA. *Pesticide Field Programs: Ensuring Protection through Partnerships,* 2006; 9–14.

VI

ENT: Environmental Management Using Environmental Technologies

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

58

Acid Sulfate Soils: Formation

Martin C. Rabenhorst, Delvin S. Fanning, and Steven N. Burch

Introduction

Soils containing sulfide minerals that have not yet been oxidized through acid sulfate weathering are referred to as potential acid sulfate soils (potential acid SS). The processes involved in the formation and accumulation of sulfide minerals in soils leading to the formation of potential acid SS will be discussed first. Subsequently, processes related to the oxidation of sulfides in the formation of active acid SS will be examined. The extent of acid SS worldwide has been estimated to be approximately 12–15 MHa.[1]

Potential Acid Sulfate Soils and Sulfide Mineral Formation—Sulfidization

Biogeochemistry of Sulfide Mineral Formation

Several factors are required for sulfate reduction. These include a source of sulfate, a source of oxidizable carbon, reducing conditions and the presence of sulfate reducing bacteria.[2] Any of these components could theoretically limit sulfate reduction. In saturated soil or sedimentary environments where the required factors are present, heterotrophic microbes utilize sulfate as an electron acceptor that becomes reduced to sulfide according to Eq. (1).

$$\mathrm{SO_4^{2-}} + 10\mathrm{H}^+ + 8_e^- \rightarrow H_2\mathrm{S} + 4H_2\mathrm{O} \qquad (1)$$

Sulfate

Provided that the other required factors are met, the quantity of sulfate may limit the rate of sulfate reduction. Gold-haber and Kaplan[3] reported sulfate reduction to be independent of concentration when sulfate levels are above 10 mM (320 mg/L). Work by Haering,[4] in Chesapeak Bay, indicated that sulfate levels may begin to limit sulfur accumulation in marsh soils when levels drop below 1 mM (32 mg/L). Some degree of sulfate reduction will continue as long as sulfate is present at minimal levels (>5–20 μM, 0.16–0.6 mg/L).[5] Because sulfate-reducing bacteria are better able to complete for electron-donating substrates than are methane-generating bacteria, methanogenesis is of minimal significance so long as sulfate levels are above 0.03–0.4 mM.[6] Therefore, sulfate reduction dominates in brackish systems. In freshwater, sulfate reduction may become overshadowed by methanogenesis as sulfate is depleted.

Oxidizable Organic Carbon

The oxidation of organic matter provides the energy microorganisms need to facilitate sulfate reduction. Plant materials rich in labile components are more easily decomposed than humified soil organic matter or peat. In sediments low in organic matter, sulfate reduction may be limited by the paucity of oxidizable carbon. This can be demonstrated in thin sections from mineral horizons in tidal marsh soils where iron sulfide minerals have accumulated in pores occupied by decaying plant roots (Figure 1). The intimate association of pyrite with the decomposing organic minerals, and its near absence from the surrounding soil matrix suggests that organic matter is limiting the formation of sulfide.[7]

Reducing/Saturated Conditions

Because diffusion of gases through saturated soils and sediments is very slow, oxygen becomes depleted under saturated conditions and microbes which utilize other electron acceptors become active. Nitrate, Mn(IV) and Fe(III) are so utilized as the environment becomes progressively reduced (Figure 2). If the conditions permit the entry of oxygen, then redox potentials may never become sufficiently low to foster sulfate reduction. More typically, diffusion of oxygen into a saturated soil or sediment is sufficiently slow, and if other necessary factors are present, sulfate reduction will occur. Figure 2 illustrates that pH, as well as E_h, must be specified in assessing sulfur phase equilibria. For example, as the pH increases from 5 to 7, the minimum E_h at which sulfate reduction is expected decreases from approximately −50 to −200 (based on a SO_4^{2-} concentration of 10 mM and a pH_2S of 0.0001 atm).

Sulfate-Reducing Bacteria

Some 15 genera of bacteria have been recognized as sulfate reducers including *Desulfovibrio, Desulfotomaculum,* and *Desulfobacter.*[8,9] These organisms thrive under strongly reducing conditions,

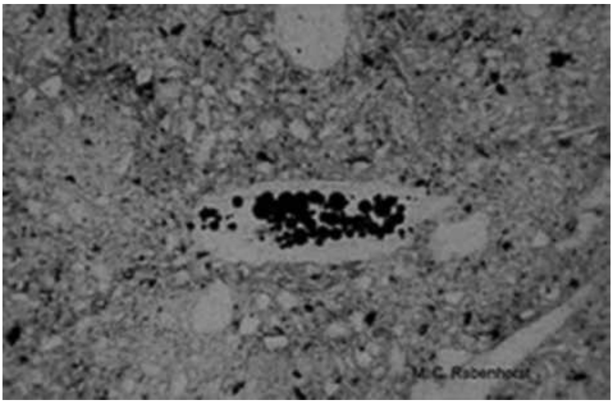


FIGURE 1 Micrograph of a thin section from the mineral (Cg) horizon of a tidal marsh soil illustrating accumulation of pyrite framboids in the channel occupied by decaying plant roots. Plane polarized light; frame length 1.2 mm.

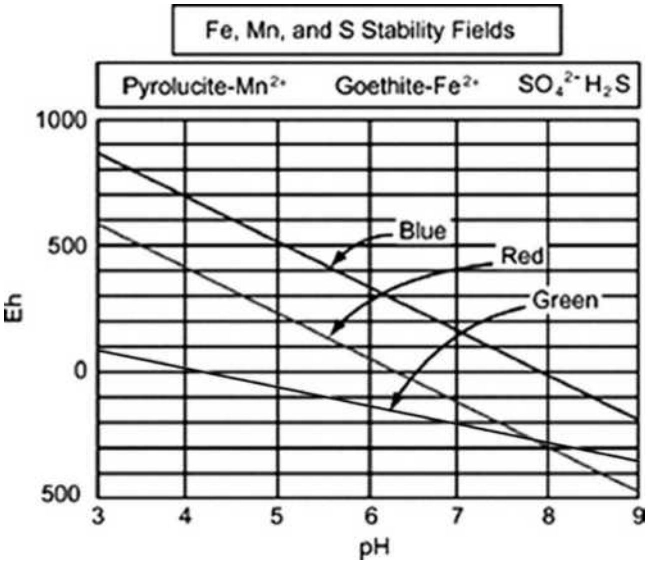


FIGURE 2 pe-pH diagram illustrating location of stability fields for redox sensitive components. The sulfate-sulfide lines (Green) is based on a $(SO)_4^{2-}$ concentration of 10 mM and a pH_2S of 0.001 atm. (Blue line separates the pyrolucite-Mn^{2+} stability fields; red line separates the goethite-Fe^{2+} stability fields.)

but many are able to persist in aerobic conditions for significant periods of time. Thus, if the other factors necessary for sulfate reduction are present, sulfate reducing bacteria will also become active.

As with most heterotrophic bacteria, rates of sulfate reduction are temperature dependent. Optimum temperature for most sulfate reducers is 30–40°C,[8] and the rate of sulfate reduction generally increase with temperature across this range. Some groups of sulfate reducers are thermophyllic and can function at temperature up to 85°C. Thus, in tropical coastal wetlands, sulfate reduction occurs all year round. In higher latitudes, where soil and sediment temperatures may approach biological zero, rates may become very slow during winter.

Reactive Iron

Once formed, sulfide is available to form a variety of minerals provided there is adequate reactive iron present. Most of the iron enters coastal environments as iron oxides sorbed to the surface of clay and silt particles. When iron oxides in the sediments and marsh soils become reduced to Fe(II), they can form iron sulfide minerals. While monosulfide species may form first [Eq. (2)], and minerals such as greigite (Fe_3S_4) may persist in recent sediments, disulfide forms such as pyrite are energetically more stable and will form at the expense of the monosulfides.

$$\mathrm{FeS_2} \text{ and } \left(+\mathrm{S_x^{y-}}, \text{loss of } \mathrm{e^-}, \text{or} + \mathrm{H_2S}\right) \tag{2}$$

$$\rightarrow \mathrm{FeS_2} + \text{various} \tag{3}$$

Mechanisms for pyrite formation may follow several possible pathways including 1) reaction of monosulfide with polysulfide; 2) partial oxidation of monosulfide; and 3) reaction of monosulfides with H_2S[10] (Eq. 3). Sulfide itself has the ability to reduce Fe(III) to Fe(II) on the surface of iron oxides.[11] Pyrite can occur either as small (<2 μm) individual crystals or as spherical clusters of crystals called framboids. In low organic mineral sediments, reactive iron is usually present in excess, resulting in a low degree of pyritization.[12] However, in organic-rich soils iron may limit the accumulation of sulfide minerals, and the degree of pyritization is generally high. This has been demonstrated experimentally in salt marsh Histosols.[13]

Environments of Sulfide Formation and Accumulation

It is clear that in environments which provide a source of oxidizable carbon and sulfate and which are sufficiently saturated to enhance reducing conditions, sulfate reducing bacteria will generate sulfide. If reactive iron is present, then solid phase ferrous minerals will accumulate. This process of *sulfidization*[14] is shown schematically in Figure 3. The obvious settings for these processes are coastal marine and brackish environments, where sulfate is abundant. Under permanently submersed conditions, detrital carbon is added by flora and fauna to the sediment. In shallow water settings (<3 m) where various pedogenic processes are at work, these accumulated sediments have been recognized as subaqueous soils[15] and are classified to reflect the sulfide components.

The soils of coastal marshes (in temperate environments) and mangroves (in tropical settings) also are ideal for sulfide formation and accumulation. The high primary productivity of these ecosystems (up to 3 kg m^{-2} yr^{-1} in marshes and up to 5 kg m^{-2} yr^{-1} in mangroves)[16] makes these an exceptionally good environment for sulfate reduction. Such soils may contain up to 20–30 g/kg of pyrite sulfur, and estimates of pyrite S accumulation rates in estuarine marshes are as high as 7 g m^{-2} yr^{-1}.[17]

Sulfate reduction can occur in other settings, so long as a source of sulfate is available. While generally small, atmospheric deposition of sulfate may be enough to induce sulfate reduction in the sediments of some interior freshwater lakes. Sulfate reduction has also been documented in prairie potholes where sulfate has apparently been contributed by the weathering of sulfur-bearing shales.[18]

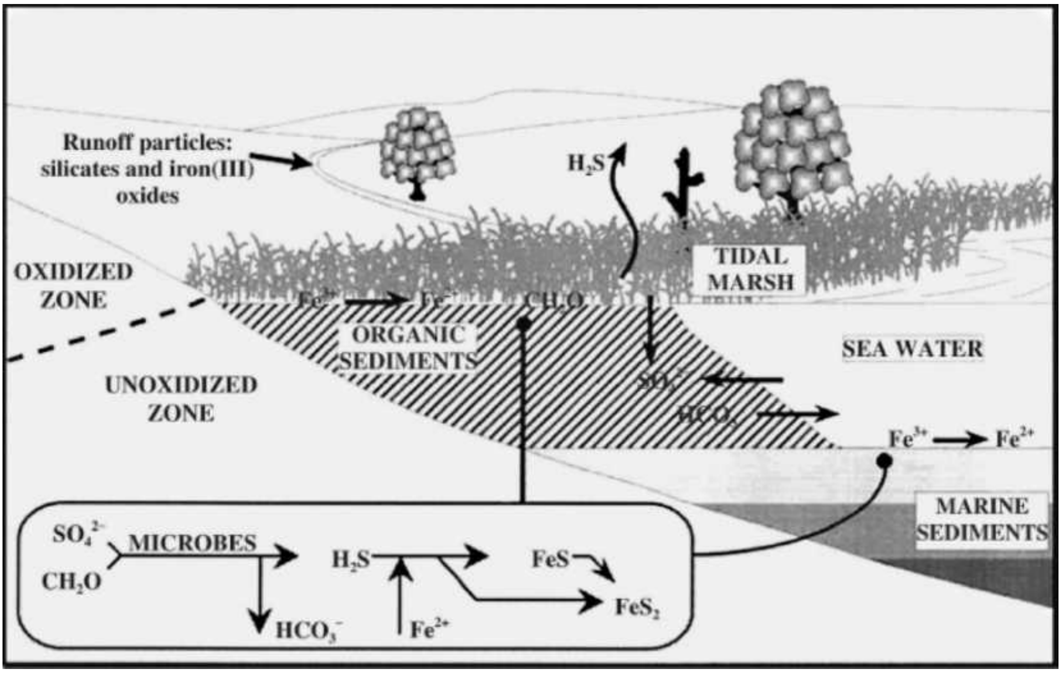


FIGURE 3 Schematic diagram illustrating the generalized process of sulfidization which leads to the formation of iron sulfide minerals and potential acid SS.
Source: Fanning et al.[26]

Formation of Active Acid Sulfate Soils—Sulfuricization

Chemistry of Sulfide Oxidation

Sulfides begin to oxidize once they are exposed to more oxidizing conditions. This occurs most often as a result of such human activities as drainage or dredging of sulfide bearing soils or sediments, or the mining of sulfide bearing coal, but may also occur due to tectonic uplift or oceanic regression. Under humid or moist aerobic conditions, sedimentary sulfide minerals can oxidize chemically[19] but this is a slow process, probably due to particular rate-limiting reactions. Various microorganisms are adapted to oxidize sulfides either directly through sulfur transformations or by facilitating (catalyzing) such rate-limiting reactions as the oxidation of Fe(II) to Fe(III).[20] While there are many possible intermediate reactions in the oxidation of pyrite, the overall reaction is summarized in Eq. (4). One mole of pyrite eventually yields two moles of sulfuric acid and a mole of iron hydroxide.

$$\mathrm{FeS_2} + 3\frac{3}{4}\mathrm{O_2} + 3\frac{1}{2}\mathrm{H_2O} \rightarrow 2H_2\mathrm{SO_4} + Fe(OH)_3 \tag{4}$$

The oxidation of pyrite proceeds along two fronts (Figure 4). First the S is oxidized (through intermediates) to sulfate yielding sulfuric acid and the remaining Fe(II). The generated Fe(II) sulfate salts are very soluble and potentially mobile. Secondly Fe(II) is oxidized to Fe(III), which when hydrolyzed produces additional acid. At high pH, the oxidation of sulfide is accomplished with oxygen, but under low pH conditions sulfide is oxidized by Fe(III). Microorganisms such as *Thiobacillus ferrooxidans* facilitate this reaction by oxidizing Fe(II) to Fe(III). For more details refer to Gagliano and Bigham.[21]

Other Aspects of Sulfuricization and Properties of Acid Sulfate Soils

Sulfuricization is the overall process by which sulfide bearing minerals are oxidized, minerals are weathered by the sulfuric acid produced and new mineral phases are formed from the dissolution products.[14,21] When $CaCO_3$ minerals are present, the sulfuric acid reacts with them to form the mineral gypsum, according to Eq. (5).

$$\begin{aligned} &\mathrm{CaCO_3} + \mathrm{H_2SO_4} + \mathrm{H_2O} \\ &\rightarrow CaSO_4 \cdot 2H_2\mathrm{O} + \mathrm{CO_2} \end{aligned} \tag{5}$$

As long as sufficient $CaCO_3$ is present, the pH is prevented from becoming very low and the soil does not become acid. When insufficient acid neutralizing minerals are present, the oxidation of pyrite in soils will lower the pH. The pH of active SS commonly drops to below four and in extreme can go below two. As iron is oxidized and hydrolyzed, various iron minerals form in the soil including ferrihydrite, schwestmannite and goethite. If the soil pH falls below four while maintaining an oxidizing environment ($E_h > 400\,\mathrm{mV}$) then the mineral jarosite ($KFe_3(SO_4)_2(OH)_6$) can form.[23] Because jarosite forms under conditions of high E_h and very low pH, which can only develop from the generation of sulfuric acid, it is considered a diagnostic mineral for acid SS.[24]

Jarosite has been reported in soils, which are not extremely acid and which may even contain carbonates.[22] These are interpreted to be “postactive” acid SS, meaning that earlier in their pedogenic history, they had undergone acid sulfate weathering. Subsequently, the soil pH has risen due to weathering of silicate minerals or addition of eolian carbonates. Because the redox potential has remained strongly oxidized, the jarosite has persisted as a metastable species. A recent review of acid sulfate soils including a discussion of the modeling of associated processes was recently completed.[25]

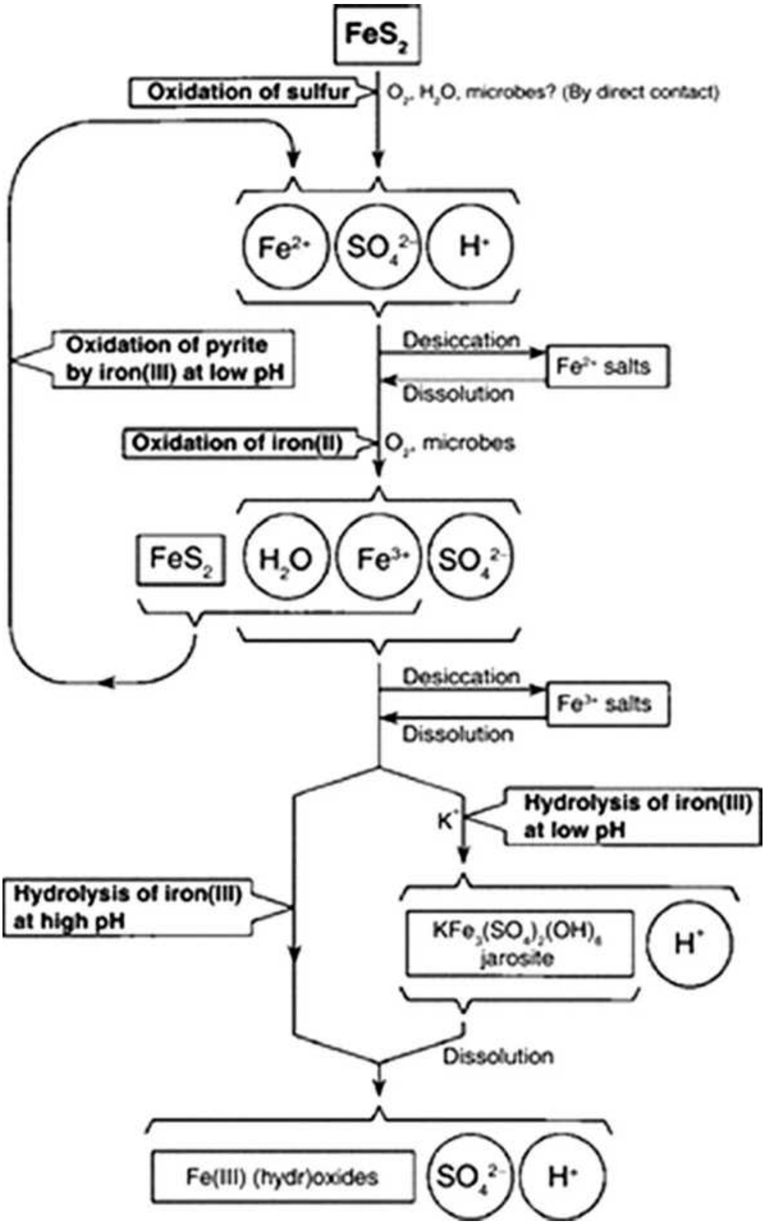


FIGURE 4 Schematic diagram illustrating the generalized process of sulfuricization which involves the oxidation of iron sulfide minerals and the production of acidity and the formation of new sulfate and other minerals.
Source: Fanning et al.[26]

References

1. Andriesse, W. Acid sulfate soils: Global distribution. In *Encyclopedia of Soil Science*; Lal, R., Ed.; Marcel Dekker: New York, 2001.
2. Rabenhorst, M.C.; James, B.R.; Magness, M.C.; Shaw, J.N. Iron removal from acid mine drainage in wetlands by optimizing sulfate reduction. In *The Challenge of Integrating Diverse Perspectives in Reclamation*. Proc. Am. Soc. Surf. Mining Reclam; Spokane: Washington, 1993; 678–684.

3. Goldhaber, M.B.; Kaplan, I.R. Controls and consequences of sulfate reduction rates in recent marine sediments. Soil Sci. **1975**, *119,* 42–55.
4. Haering, K.C. *Sulfur Distribution and Partitionment in Chesapeake Bay Tidal Marsh Soils*; M.S. thesis University of Maryland: College Park, MD, 1986; 172 pp. 5.
5. Ingvorsen, K.; Zehnder, A.J.B.; Jorgensen, B.B. Kinetic of sulfate and acetate uptake by delulfobacter postgatei. Appl. Environ. Microbiol. **1984**, *47*, 403–408.
6. Smith, D.W. Ecological actions of sulfate-reducing bacteria. In The *Sulfate-Reducing Bacteria: Contemporary Perspectives*; Odom, J.M., Singleton, R., Jr., Eds.; Springer: New York, 1993; 161–188.
7. Rabenhorst, M.C.; Haering, K.C. Soil micromorphology of a Chesapeake Bay Tidal Marsh: implications for sulfur accumulation. Soil Sci. **1989**, *147*, 339–347.
8. Fauque, G.D. Ecology of sulfate-reducing bacteria. In *Sulfate-Reducing Bacteria*; Barton, L.L., Ed.; Plenum Press: New York, 1995; 217–241.
9. Postgate, J.R. The sulphate-reducing bacteria (Ch. 7). In *Ecology and Distribution*, 2nd Ed.; Cambridge University Press: London, 1984; 107–122.
10. Rickard, D.; Schoonen, M.A.A.; Luther, G.W., III Chemistry of iron sulfides in sedimentary environments. In *Geochemical Transformations of Sedimentary Sulfur*; Vairavamurthy, M.A., Schoonen, M.A.A., Eds.; American Chemical Society: Washington, DC, 1995; 168–193.
11. Ghiorse, W.C. Microbial reduction of manganese and iron. In *Biology and Anaerobic Microorganisms*; Zehnder, A.J.B., Ed.; John Wiley and Sons: New York, 1988; 305–331.
12. Griffin, T.M.; Rabenhorst, M.C. Processes and rates of pedogenesis in some Maryland Tidal Marsh soils. Soil Sci. Soc. Am. J. **1989**, *53*, 862–870.
13. Rabenhorst, M.C. Micromorphology of induced iron sulfide formation in a Chesapeake Bay (USA) Tidal Marsh. In *Micromorphology: A Basic and Applied Science*; Douglas, L.A., Ed.; Elsevier: Amsterdam, 1990; 303–310.
14. Fanning, D.S.; Fanning, M.C.B. *Soil Morphology, Genesis, and Classification*; John Wiley and Sons: New York, 1989; 395 pp.
15. Demas, G.P.; Rabenhorst, M.C. Subaqueous soils: pedogenesis in a submersed environment. Soil Sci. Am. J. **1999**, *63*, 1250–1257.
16. Mitch, W.J.; Gosselink, J.G. *Wetlands,* 2nd Ed.; Van Nostrand Reinhold: New York, 1993; 722 pp.
17. Hussein, A.H.; Rabenhorst, M.C. Modeling of sulfur sequestration in coastal marsh soils. Soil. Sci. Soc. Am. J. **1999**, *63*, 1954–1963.
18. Arndt, J.L.; Richardson, J.L. Geochemistry of hydric soil salinity in a recharge-throughflow-discharge Prairie-Pothole wetland system. Soil. Sci. Soc. Am. J. **1989**, *53*, 848–855.
19. Borek, S.L. Effect of humidity of pyrite oxidation. In *Environmental Geochemistry of Sulfide Oxidation*; Alpers, C.N., Blowes, D.W., Eds.; American Chemical Society: Washington, DC, 1994; 31–44.
20. Nordstrom, D.K. Aqueous pyrite oxidation and the consequent formation of secondary iron minerals. In *Acid Sulfate Weathering*; Kittrick, J.A., Fanning, D.S., Hossner, L.R., Eds.; Soil Sci. Soc. Am. Spec. Pub. No. 10: Madison, WI, 1982; 37–56.
21. Gagliano, W.B.; Bigham, J.M. Acid mine drainage. In *En-cylopedia of Soil Science*; Lal, R., Ed.; Marcel Dekker: New York, 2001.
22. Carson, C.D.; Fanning, D.S.; Dixon, J.B. Alfisols and ultisols with acid sulfate weathering features in Texas. In *Acid Sulfate Weathering*; Kittrick, J.A., Fanning, D.S., Hossner, L.R., Eds.; Soil Sci. Soc. Am. Spec. Pub. No. 10: Madison, WI, 1982; 127–146.
23. van Breeman, N. Genesis, Morphology, and classification of acid sulfate soils in coastal plains. In *Acid Sulfate Weathering;* Kittrick, J.A., Fanning, D.S., Hossner, L.R., Eds.; Soil Sci. Soc. Am. Spec. Pub. No. 10: Madison, WI, 1982; 95–108.
24. Fanning, D.S. Sulfate and Sulfide Minerals. In *Encyclopedia of Soil Science;* Lal, R., Ed.; Marcel Dekker: New York, 2001.

25. Ritsema, C.J.; van Mensvoort, M.E.F.; Dent, D.L.; Tan, Y.; van den Bosch, H.; van Wijk, A.L.M. Acid sulfate soils. In *Handbook of Soil Science*; Sumner, M.E., Ed.; CRC Press: Boca Raton, 1999; G121–G154.
26. Fanning, D.S.; Rabenhorst, M.C.; Burch, S.N.; Islam, K.R.; Tangren, S.A. Sulfides and sulfates. In *Soil Mineralogy with Environmental Applications;* Dixon, J.B., Schulze, D.G., Eds.; Soil Science Society of America Book Series #7: Madison, WI, 2002; 229–260.

59

Erosion and Sediment Control: Vegetative Techniques

Samson D. Angima

Introduction

Soil detachment and erosion occur when soil is disturbed by either human activity or natural conditions such as extreme weather. As a result of this, the soil is being moved from one point to another by either wind or water. Practices such as road construction, suburban and industrial developments, stream channel and other types of construction on sloping lands, inadequate drainage facilities, poor grading practices, deforestation, cultivation on sloping lands, and general lack of adequate planning by land users cause soil erosion. Consequences of soil erosion include water and air pollution, reduced land productivity from loss of topsoil and nutrients, and degradation of the environment. Soil erosion by water is controlled mainly by earthworks and engineered constructions such as terracing aimed at reducing the slope and at collecting and storing moisture while reducing runoff to acceptable limits. Vegetative barriers are grass, shrubs, or small trees grown in close rows that can be used to control both water and wind erosion by providing protection from soil and dislodging sources such as rainfall, and offering a semipermeable barrier to erosion agents resulting in soil deposition. The resultant vegetation also shields the soil surface from overland flow and decreases the erosive capacity of water flow by reducing its velocity.

Vegetative Materials for Soil Erosion and Sedimentation Control

Vegetative cover is essential for the design and stabilization of many structural erosion-control devices (Figure 1). Proper vegetative cover provides excellent erosion protection and sedimentation control. Vegetative barriers are planted in close rows along contours on slopes to intercept water runoff, or are planted perpendicular to the direction of wind to retard wind movement, resulting in soil deposition. Plant roots and lateral stems provide a structure to hold soil particles in place. These features also

FIGURE 1 Vegetative barriers/strips hold the terraces intact for agricultural production in Kabale, Uganda. (Courtesy of Samson Angima.)

improve the soil's physical properties and increase infiltration rate, thereby decreasing runoff. Plant transpiration reduces soil moisture levels, which increases soil absorption capabilities. Not every plant material can serve as a soil erosion-control agent. Those plants that do possess some bioengineering characteristics in both the root and the shoot systems that encompass both living plants and organic materials as construction elements for erosion control. Such properties include noncompetition with adjacent crops or fruit trees for moisture or nutrients, a rooting system that reaches deep down to anchor the plant and also extract leached nutrients from the subsoil, and shade tolerance. The plants must be perennials with high seed vigor, have the ability to increase soil organic matter and reduce surrounding soil bulk density to allow faster water infiltration, and have strong woody stalks to withstand pressure from erosive agents.

Vegetative materials used for soil erosion control include grass species, legumes, trees, shrubs, vetches, and sods. Common grass species that bunch during growth include tall fescue, perennial ryegrass, orchardgrass, timothy, switchgrass, weeping lovegrass, deertongue, and big bluestem. Bunching legumes include birdsfoot trefoil and serecia lespedeza. Cereals such as winter wheat, winter rye, spring oats, sudangrass, and Japanese millet are also used. Sod-forming species include grasses such as redtop, fine fescue, Kentucky bluegrass, smooth bromegrass, and legumes such as crownvetch and flatpea.[1] Shrubs and small trees that are used as barriers in erosion control include *Calliandra calothyrsus, Sesbania sesban, Leucaena leucocephala, Gliricidia sepium, Cassia siamea, Eucalyptus* spp., *Casuarina* spp., *Acacia* spp., *Azadirchta indiaca,* and *Grevillea robusta*.[2] Tree species selection depends on local landuse and climatic conditions that favor establishment.

Mass Planting of Vegetative Cover

Mass planting of vegetative cover is done where land that is susceptible to erosion is converted from cropland to vegetative cover such as in the conservation reserve program (CRP). In this program, trees as well as sod are planted in strategic places, where they will provide maximum protection to the soil resource, and the landowner agrees to leave the areas under vegetation for a given period of time for land stabilization to occur. The vegetative cover also reduces water pollutants such as fertilizer nutrients, pesticides, and herbicides in the runoff water, increases oxygen levels, reduces greenhouse gases and evaporation rates, and provides shade and buffers against high winds.[2]

Vegetative Strips

Vegetative strips constitute different types of filter strips that reduce runoff velocity and provide differing degrees of filtering action depending on the species used. Larger soil particles tend to settle out readily, leaving only clay particles suspended and thereby reducing pollutant transfer to ponds, rivers, and lakes. Filter strips can be classified as grass strips, vegetated waterways, filter strip terraces, buffer and riparian strips, and settling basins.[3]

Grass Strips

Grasses are by far the most important vegetative materials used to control erosion. Bands of grass about 1–2 m in width are planted along the contours and spaced every 10–100 m depending on the slope (Figure 2). Prior to planting, the land is roughened by disking, harrowing, or raking, and then limed, fertilized, and seeded with mixtures of adapted grasses and legumes to enhance a good stand.[4] Vetiver grass (*Vetiveria zizanioides*), napier grass (*Pennisetum purpureum*), and stiff grass (*Miscanthus sinensis*) are the most commonly used species.[2] These are upright, tufted, deep-rooted, and very dense grasses that are by far the most important in erosion control, particularly in tropical countries. Vetiver is a bunch grass with very rapid growth in warm and moist conditions. It grows to more than 2 m in height and has a remarkably dense and vertical rooting system, growing deep (3–5 m). Napier grass is a tall perennial reaching over 3 m high. It is resistant to drought, and grows at altitudes up to 2400 m. Stiff grass has dense roots and coarse stems that withstand erosive agents. These grass hedges present a virtually impenetrable barrier through which soil can hardly pass and only water, with much reduced velocity, passes.[5] As silt builds up behind the grass, other grass shoots arise from the nodes above the deposited silt to form a natural terrace. Weeds and undesirable foreign grasses are unable to penetrate through a well-established grass hedge. Studies carried out in Mississippi, using stiff grass (*M. sinensis*) on no-till and conventional tillage, [6] showed that grass strips help reduce sediment losses by up to 88% on conventional tillage and 57% on no-till cotton plots (Table 1). Other grass species include Buffalo grass (*Buchloe dactyloides*), grama or mesquite grasses (genus *Bouteloua*), switch grass (*Panicum virgatum*), and *Phalaris aquatica*.

Vegetated Waterways

Vegetated waterways are channels or waterways that transfer runoff from a higher to a lower elevation over a short distance without allowing erosion to occur (Figure 3). In addition to dissipating flow energy,

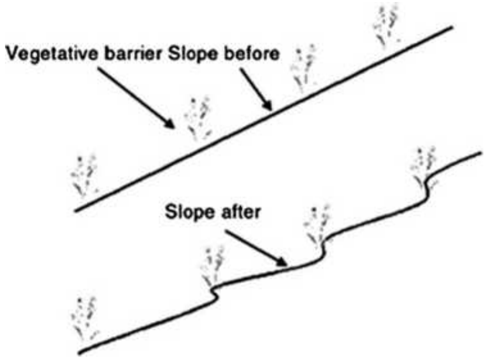


FIGURE 2 Schematic sketch of vegetative barriers illustrating expected change in land slope over time resulting from tillage and erosion deposition process.

TABLE 1 Sediment Loss with and without Grass Strip (*M. sinensis*) under No-Till and Conventional Tillage Cotton at 5% Slope

	Sediment loss (t/ha)				
Period	No-Till with Grass Strip	No-Till without Grass Strip	Conventional Tillage with Grass Strip	Conventional Tillage without Grass Strip	No-Till with Winter Wheat Cover
1992	2.6	4.5	12.3	60.3	2.9
1993	0.9	1.6	5.4	21.9	1.2
1994	3.2	9.6	18.6	63.2	1.9
Average	2.2	5.2	12.1	48.5	2.0

Source: McGregor et al.[6]

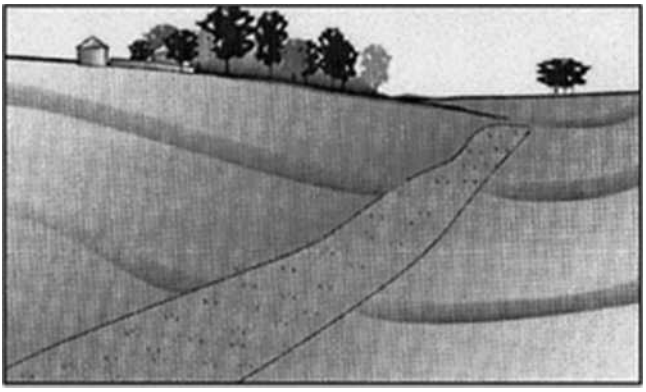


FIGURE 3 Grass waterways are the most common type of vegetative filter strip.
Source: Regehr.[3]

some structures also act to retain soil. They are especially effective in arresting gully development, a situation that might need both structure(s) and some vegetated channel. They have many uses in comprehensive conservation plans, but they primarily collect and concentrate flows and then safely transport the water to major drainage systems. Dense vegetation is used to minimize the area required, i.e., the protective action of the vegetation permits higher flow velocities and thus smaller waterway cross-sections. Most vegetated waterways run directly down a slope; however, they can also be constructed somewhat across the slope as diversions or sometimes just to reduce channel slope. Vegetated channels are not used where continuous flow occurs because the vegetation will die out. Tillage near a waterway is accomplished in a direction across the waterway. Sodforming, cool-season grasses such as smooth brome or western wheatgrass are used in grass waterways. Grass waterways are usually designed to carry runoff of a 24 hr storm of the intensity that happens once every 10 years. In areas with prolonged water flows, high water tables, or seepage problems, a rock-lined center is added.[1]

Filter Strip Terraces

Filter strip terraces are strips of grass sod, legumes, and other vegetation on the contour that surface water runoff crosses as it runs downhill. They serve as an alternative to earthen terraces but do not have a channel to conduct water along the contour, as earthen terraces do. Filter strip terraces are excellent removers of sediment, pesticides, organic matter, and other pollutants. They are better than grass waterways because water enters the strip uniformly and over a wide area.[3] The width and type of vegetation established in the filter area are determined by site conditions including soil type, land slope, and type of runoff entering the filter.

Small trees and shrubs are also used as conservation hedges in erosion control. These are planted close together and are periodically pruned to maintain height while the cut branches are inserted upslope of the hedge to trap more sediment. Trees that have been used successfully in erosion control, especially in agroforestry systems in the tropics, include *C. calothyrsus, S. sesban, L. leucocephala, G. sepium,* and *C. siamea*. Trees used for windbreaks include *Eucalyptus* spp., *Casuarina* spp., *Acacia* spp., Neem tree (A. *indiaca*), and *G. robusta*[7] In the temperate zone, thorny hedge plants include barberry, Osage orange, buckthorn, and hawthorn. Evergreen hedge plants are box, privet, azalea, yew, arborvitae, rhododendron, mountain laurel, and holly.[8] Decorative deciduous shrubs often used are lilac, forsythia, mock orange, spiraea, euonymus, and viburnum. *Rosa rugosa* can be planted along the highway embankments, and rows of poplars, hemlocks, and other trees can be used as shelterbelts. Vegetative or biological measures may include log bundles anchored to the stream bank or the planting of herbaceous or woody plants, which can withstand high velocity flow, while the roots form a protective net for the soil.

Buffer Strips and Riparian Strips

Buffer strips at lower elevations of fields (Figure 4) and Riparian strips along stream banks, ponds, and lakes (Figure 5), intercept surface runoff water from crop fields. Buffer strips may constitute ordinary grassed fencerows or strips of grasses, shrubs, and trees lining hillsides or banks of rivers. Runoff water must flow in a shallow, even layer across the buffer strip to remove sediments. Most common grasses used in buffer strips are Bluestem and Indiangrass. Riparian strips are planted so that surface and subsurface runoff must filter through them before it reaches a pond, lake, or stream.[1] The body of water can be permanent or temporary. Riparian strips can also be placed next to wetlands, such as marshy or swampy areas, and additional vegetation can be placed uphill if excessive amounts of sediments enter the waters.

Settling Basins

Settling basins are constructed around inlets to tile-outlet terraces. These are important in reducing soil sediment loads and also act as setback zones for herbicides such as Atrazine. The basins are designed to retain water for up to 24 hr, giving most of the larger sediments time to settle out.[3] These are common where tile lines have been installed.

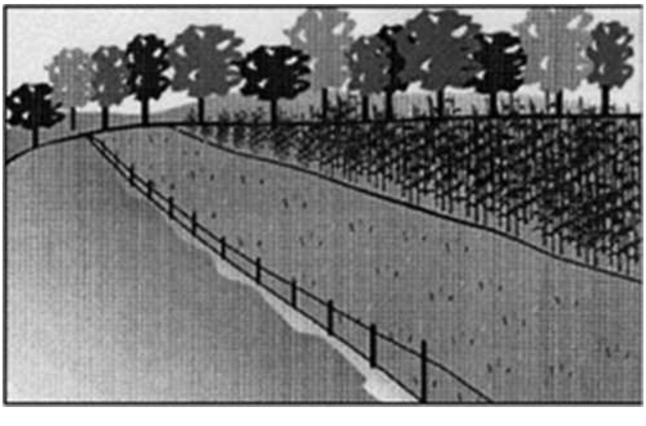


FIGURE 4 Buffer strips at lower elevations of fields intercept surface runoff water from crop fields.
Source: Regehr.[3]

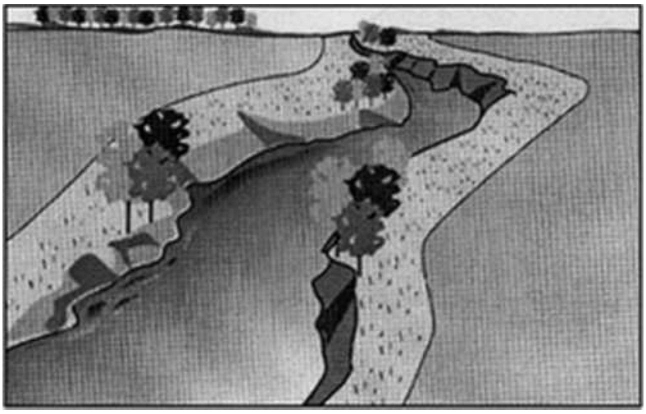


FIGURE 5 Riparian strips along stream banks intercept surface runoff water from crop fields.
Source: Regehr.[3]

Conclusions

Vegetative techniques for erosion and sedimentation control offer a wide range of tools for use by a wide range of land users especially those that cannot afford engineered or earthen structures. However, for these tools to work well in erosion prevention, they have to be regularly maintained and landowners need to understand that they require time to effectively control enough sediment that can gradually change slope. Research is adding more species of grass and/or shrubs that can be multipurposely used to control erosion and at the same time provide other uses beneficial to the land owner. When used well in conjunction with best management practices, vegetative techniques can substantially contribute to overall water and wind erosion control.

References

1. Landschoot, P. *Erosion Control and Conservation Plantings on Non Cropland*; Penn State College of Agricultural Sciences Extension Publication, 1997.
2. Young, A. Agroforestry for soil conservation. In *Agroforestry for Soil Conservation*; CAB International: Wallingford, U.K., 1989; 197–229.
3. Regehr, D.L.; Devlin, D.L.; Barnes, P. *Using Vegetative Strips in Crop Fields*; Kansas State University, 1996. http://www.oznet.ksu.edu.
4. Dabney, S.M.; Liu, Z.; Lane, M.; Douglas, J.; Zhu, J.; Flanagan, D.C. Landscape benching from tillage erosion between grass hedges. Soil Till. Res. **1999**, *51,* 219–231.
5. Dabney, S.M.; McGregor, K.C.; Meyer, L.D.; Grissinger, F.H.; Foster, G.R. Vegetative barriers for runoff and sediment control. In *Integrated Resources Management and Landscape Modification for Environmental Protection;* Mitchell, J.K., Ed.; Amer. Soc. of Agri. Eng.: St. Josephs, 1993; 60–70.
6. McGregor, K.C.; Dabney, S.M.; Johnson, J.R. Runoff and soil loss from plots with and without stiff-grass hedges. Trans. Amer. Soc. of Agri. Eng. **1999**, *42* (2), 361–368.
7. Angima, S.D.; O'Neill, M.K.; Omwega, A.; Stott, D.E. Use of tree/grass hedges for soil erosion control in the Central Kenyan highlands. J. Soil Water Conserv. **2000**, *55* (4), 478–482.
8. Nair, P.K.N. *An Introduction to Agroforestry;* Kluwer Academic Publishers: Boston, 1993; 325–343.

60

Precision Agriculture: Engineering Aspects

Joel T. Walker,
Reza Ehsani, and
Matthew O. Sullivan

Introduction

Information technology is playing an increasingly important role in today's agricultural production systems of all sizes, commodities, and management philosophies. Precision agriculture[12,14,17] or site-specific management is an information- based management technique that has the potential to improve profitability[7] and reduce the environmental impact[19] of crop production. It also has the potential to improve the quality and nutrient content of the product. Precision agriculture, rather than the "one-size-fits-all" management strategy, provides for differential treatment of selected areas of a production field, called management zones, based upon expectation of increased yield, profit, or some other agronomic goal.[2–4,18] Management zones may be selected for differential treatment based upon various documented differences such as soil type, soil fertility or pH, yield history, presence of weeds, insects, or diseases, or other measures for which a differential treatment helps the producer achieve a selected goal. The ability to provide differential treatment to management zones, also called site-specific management, depends upon availability of both proper equipment and effective treatment algorithms.

What Makes It Possible?

Precision agriculture techniques have been made possible by the advent of global positioning systems (GPS) and high speed computer processing. GPS provides real-time location information to a computer that, from stored information, determines the current management zone, selects appropriate treatment for that management zone, and controls mechanisms to provide the treatment. Figure 1 is a graphic representation of the precision agriculture paradigm. GPS provides position information for a variety of data gathering processes or for control of site-specific treatments. Information of various types (shown as layers) may be used in analysis of yield results or to develop an application map to control site-specific treatments. The whole system taken together is often called precision agriculture or site-specific agriculture. Note that a feedback loop is implied where results of the previous growing season (yield) become part of the information that influences current treatment practices. Given the proper

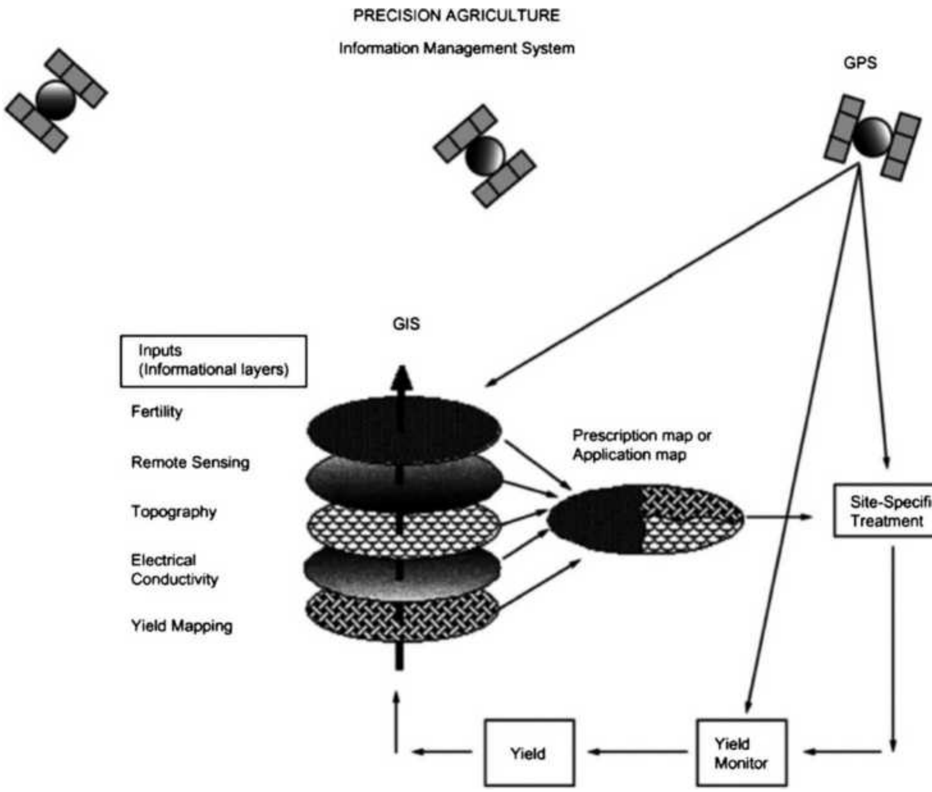


FIGURE 1 Paradigm for precision agriculture.

treatment algorithms, current treatment practices may optimize the goal parameter. Maximum yield is not necessarily the best goal because the cost of treatments required to achieve that yield may be greater than the increased crop value.[4,18]

GPS consists of a minimum of 24 satellites orbiting the Earth and sending signals to a local receiver for which location is desired.[9] Each and satellite broadcasts encoded information with particular timing. By measuring the time a signal travels (at the speed of light) to reach the receiver, the distance from a satellite to the receiver may be calculated. Determining distances from four or more satellites of known location may establish the receiver's location (latitude, longitude, and elevation). Even simple, inexpensive receivers are capable of accuracy better than 15 m, close enough to return to a favorite fishing spot. With specialized transmissions containing information to correct for known errors (called differential signals), accuracy better than 1 m may be achieved. Specialized local transmitters make possible real-time kinematic (RTK-GPS) systems with accuracy of better than 1 cm. RTK systems are used in surveying, guidance, and where the fine precision may justify a relatively high equipment cost.

High-speed computer processing systems have also played an important role in the advent of precision agriculture. Precision agriculture requires collection and storage of data, decision-making computation, and controlling of equipment by computers operating at billions of operations per second. GPS locations are recorded in realtime by the computer. Digital maps of field conditions and parameters are carried in memory or storage media. Digital data such as digital still or video images of weeds, insects, or disease damage, soil properties, crop spectral reflectance, or climactic conditions may also be collected in real-time by the computer. Using such information, the computer may work through a response model to decide what actions should be taken with current equipment. For example, a precision agriculture capable planter may be able to adjust planting rate and depth or change the seed variety on-the-go. Thus, the computer may decide that areas with a selected soil type and relative elevation will get a reduced population (investment of seeds) because less yield is expected. The computer may dictate that another soil type with a high yield potential and high existing moisture content will get a different variety planted to a shallower depth. Across an entire field, many combinations of the controlled variables will be chosen to optimize the desired result (yield, profit, or other goal). A large field would require constant decisions and adjustment of the equipment that would not be possible without computerized systems. Finally, volumes of data may be produced by precision agriculture techniques and computer methods are being developed to extract useful information; and models form this data[5,6,10,11,15,16] and present it to the public in readily accessible form.[13]

Sensing for Precision Agriculture

Real-time sensors will be important to many precision agriculture systems.[8] A real-time sensor provides data in a nearly constant stream as the machine traverses the field. For example, a camera may provide digital images from which the presence of weeds may be determined. A sprayer may then be directed to spray only where weeds are present. Organic matter and moisture levels in soil affect the performance of some herbicides. Sensors that measure organic matter and/or moisture on-the-go allow for optimum rates of chemical application—adequate to control the weeds, but no more than necessary to preserve environmental quality.

One of the most popular real-time sensors measures grain yield on combines.[12] Yield sensors are also available or are being developed for a variety of crops such as cotton, potatoes, tree fruits, and strawberries. Yield sensors provide a measure of yield over a whole field. Areas of the field with unusually high or low yield may be identified and corrective action (subject to some predictive model created by the producer or by computer) can be taken for the following year. This one aspect of precision agriculture has created a lot of activity in adjusting soil drainage and fertility and crop management decisions such as the relative value of fertilizer or chemical inputs.

For many agronomic parameters, site-specific soil sampling is more practical, either because a real-time sensor is not yet available, or because the spatial variation of the parameter is more gradual and can be estimated with a few site samples. Soil samples to determine fertility are common. The values of pH (acidity/alkalinity), nitrogen, potassium, and phosphorus are particularly important in prediction of yield levels. Soil type is another parameter that is often determined by a site visit.

Remote sensing is becoming increasingly important to precision agriculture. Early data that came from random satellite observations had low resolution and limited value. Now specially equipped satellites and aircraft may be hired to collect specific crop or soil information. Crop growth or health may be deduced from this data. Certain wavelengths of the electromagnetic spectrum are particularly helpful. The visible light frequencies provide some information. Unhealthy crops tend to reflect more yellow and red light. Various frequencies in the infrared range also have been correlated with plant health and soil moisture conditions. Multi-spectral systems provide data on the visible spectrum (often three primary colors) and a limited range of the infrared. More sophisticated equipment, called hyper-spectral, can provide data from a much broader range of the spectrum and from narrower sample bands. This type of data offers greater opportunity to correlate specific crop or soil conditions to measured spectral data.

Other types of remote sensing are becoming available. Light detection and ranging (LIDAR), for example, is a laser ranging technique that may be used to measure topography and plant height simultaneously.[1] Currently applied to the forest industry and relatively expensive, this method holds promise for rapid feedback of information on crop growth problems to the producer, perhaps allowing solution of the problem before yield is permanently affected. For example, nematodes in soybeans are a common cause of reduced vigor and yield if left untreated. Conventional crop scouting or remote sensing may not detect small areas of infestation or provide feedback in time to take corrective action. With regular LIDAR imaging, a producer could detect and treat such trouble spots in a timely and efficient manner.

Geographic Information Systems

Geographic information systems (GIS) techniques are an integral part of precision agriculture. Basically, GIS is a storage system for geographically referenced digital data. Many of the data types discussed above can be digitized (if not already in digital form) and referenced to specific locations in a field. Each parameter or variable then may be represented as a layer of information. Geographic position of the information matches the position on a map of the field. The value of a parameter may be represented on a map as a shade of gray or a color. So a GIS information layer representing soil type can be drawn in the physical shape of the field with patches of color—each color representing a different soil type. Many such layers may exist in a GIS dataset for a particular field. Depending upon the need, these layers may be viewed as overlays—simultaneous presentation of several variables in one field. As many layers of information are added to the GIS system, it becomes apparent that human vision is inadequate to detect the important patterns. Computers, however, are infinitely more able to "see" data patterns with a potential to produce an economic advantage.

Future

The future of precision agriculture depends on economic results. Can the cost of additional equipment and operations be more than offset by increased economic return and value of environmental protection? It seems possible that all agricultural operations could be monitored and recorded, linked by digital transmissions to databases containing weather, remote sensing, and historical data, and controlled through a general model of the crop's predicted response to specific inputs. Such a system may not only control specific agricultural operations, but also be involved in scheduling, ordering seed, fertilizer, and supplies, providing records to regulating agencies, and feeding valuable information back into the research system for further refinement of the control model. Already, manufacturers are producing prototype agricultural machines that perform without an operator. These machines are guided by GPS, are controlled by computer, and have onboard sensors to detect obstructions or people in the way. The advancement of agriculture will come with technology—technology to feed the world.

References

1. Ackermann, F. Airborne laser scanning—present status and future expectations. ISPRS J. Photogram. Remote Sens. **1999**, *54* (2–3), 64–67.
2. Adams, M.L.; Cook, S.; Corner, R. Managing uncertainty in site-specific management: what is the best model? Precis. Agric. **2000**, *2*, 39–54.
3. Boote, K.J.; Jones, J.W.; Pickering, N.B. Potential uses and limitations of crop models. Agron. J. **1996**, *88*, 704–716.
4. Bullock, D.S.; Bullock, D.G. From agronomic research to farm management guidelines: a primer on the economics of information and precision technology. Precis. Agric. **2000**, *2*, 71–101.
5. Data Mining for Site-Specific Agriculture; http://www.gis.uiuc.edu/cfardatamining/ (accessed April 16, 2001).
6. Data Mining Puts Geographic Information to Work for Public, Private Sectors; http://csd.unl.edu/csd/resource/Vol.13/datamine.htm (accessed April 16, 2001).
7. Erickson, K., Ed.; *Precision Farming Profitability;* Agricultural Research Programs; Purdue University: West Lafayette, IN, 2000.

8. Gaultney, L.D. *Prescription Farming Based on Soil Property Sensors***.** ASAE Paper No. 89–1036, ASAE Annual International Meeting; ASAE: St. Joseph, MI, 1989.
9. Kennedy, M. The Global Positioning System and GIS: An Introduction; Ann Arbor Press: Chelsea, MI, 1996.
10. Kerschberg, L.; Lee, S.W.; Tischer, L.A. Methodology and Life Cycle Model for Data Mining and Knowledge Discovery for Precision Agriculture; Final Report, Center for Information Systems Integration and Evolution, George Mason University: Fairfax, VA, 1997.
11. Lazarevic, A.; Fiez, T.; Obradovic, Z.A. *Software System for Spatial Data Analysis and Modeling*, Report to INEEL University Research Consortium Project No. C94–175936.
12. Morgan, M.; Ess, D. *The Precision-Farming Guide for Agriculturalists;* John Deere Publishing: Moline, IL, 1997.
13. Ohio State Precision Agriculture; http://www.ag.ohiostate.edu/~precisfm/ (accessed April 16, 2001).
14. Pierce, F.J., Sadler, E.J., Ed.; The State of Site-Specific Management for Agriculture. *ASA/CSSA/SSSA;* Madison, WI, 1997.
15. Precision Farming Support; http://www.umac.org/farming/precision.html (accessed April 16, 2001).

16 Regressionand Emerging Data Mining Tools Modeling; http://www.gis.uiuc.edu/cfardatamining/APRReport.htm (accessed April 16, 2001).

17 Sonka, S.T., Ed. Precision Agriculture for the 21st Century: Geospatial and Information Technologies in Crop Management; National Academy Press: Washington, DC, 1998.

18 Swinton, S.M.; Lowenberg-Deboer, J. *Site-Specific Management Guidelines—Profitability of Site-Specific Farming;* Publication SSMG-3, Phosphorus and Potash Institute, 1998.

19 Wang, X.; Tim, U.S. Problem-Solving Environment for Evaluating Environmental and Agronomic Implications of Precision Agriculture;http://www.esri.com/library/userconf/proc00/professional/papers/PAP102/p102.htm (accessed April 16, 2001).

61

Sodic Soils: Reclamation

Jock Churchman

Introduction

Two main mechanisms are involved in the reclamation of sodic soils: 1) improvement in water flow; and 2) displacement of exchangeable Na ions.

Improvement in Water Flow

Sodic soils show a low permeability to water. Their permeability can be improved quite rapidly by increasing electrolyte concentrations[1,2] (the "electrolyte effect"). Ions in solution tend to inhibit the swelling and dispersion of fine clay particles by compressing the electric double layers of adjacent particles (Figure 1). The volume around the particle surfaces within which repulsion takes place is reduced, ultimately to zero.

It has been found[2] that there is no swelling and dispersion when the total electrolyte concentration exceeds a critical level, i.e., the "threshold electrolyte concentration" that depends on soil properties.[1,3]

Displacement of Exchangeable Na Ions

The deleterious effects of exchangeable sodium reflect the more extensive hydration of the sodium ion than that of other common exchangeable cations on soil particles (Figure 2). Exchangeable sodium, which forms an ionic association with the clay surface, causes hydration, dissociation of particles, and hence swelling and dispersion. By contrast, exchangeable calcium forms polar covalent bonds with the clay surface.[4] Calcium shows only limited hydration and dissociation of particles, and therefore causes only limited dispersion and swelling.[4] Na^+ on exchange sites is replaced by Ca^{2+} for the longer term remediation of sodic soils.

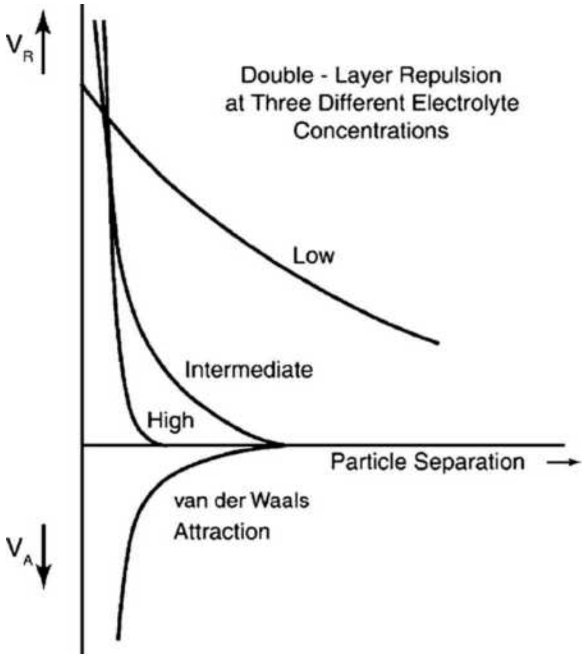


FIGURE 1 Repulsive and attractive energy as a function of particle separation at three electrolyte concentrations.
Source: van Olphen.[19]

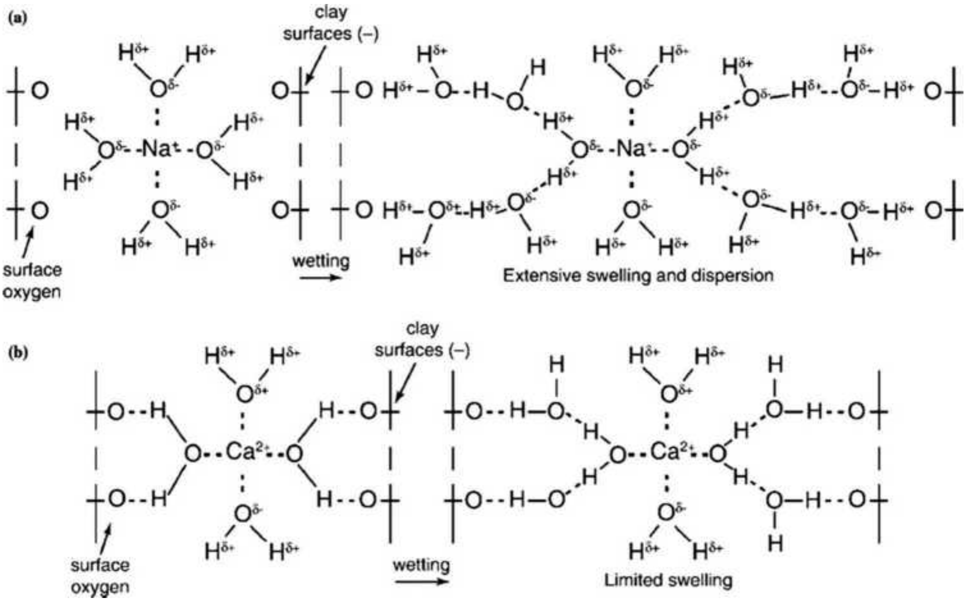


FIGURE 2 Schematic representation of the effect of wetting and nature of bonding between cations and clay surfaces on swelling and dispersion: (a) Na^+ aquo ion linking clay particles by ionic bonding, (b) Ca^{2+} aquo ion linking clay particles by polar covalent bonding. Water molecules are linked to cations by hydrogen bonding.
Source: Rengasamy and Sumner.[4]

Chemical Améliorants

Gypsum

Soluble sources of calcium ions are the most suitable sodic soil ameliorants. While gypsum is used most often, not all sources of gypsum are similarly effective or similarly suitable for each specific problem from sodicity. It was found[1,5] that only "by-product" gypsum, as a surface application, could prevent crusting. Mined gypsum was too slowly soluble for the purpose.[1] Gypsum is a by-product of many industrial processes.[5] Furthermore, gypsums from different mines can differ quite markedly in purity[6] They can also differ in dissolution rate. A finer particle size generally provides a greater solubility.[5] As well, sodium salts associated with gypsum (Table 1)[6] decrease its efficiency for amelioration.[7]

Lime

Lime has also been used as an ameliorant for sodic soils.[6,8,9] While generally less soluble than gypsum, lime may provide a useful incidental source of calcium ions when used to raise the pH of acid sodic soils. It may provide a cheaper alternative to gypsum or else be used in association with gypsum to extend the lifetime of the calcium reserves.[6] Because of its increased solubility in the presence of sodium salts, it may be more effective than gypsum for the reclamation of saline sodic soils. Lime also contributes less to the salt load than gypsum.[9]

Reclamation of Calcareous Sodic Soils

Calcareous sodic soils have an abundance of Ca, but it is present as $CaCO_3$, hence is insoluble at the high pH of these soils. Their sodicity is often overcome through acid additions,[10,11] either directly, or indirectly. Additions of elemental sulfur, pyrite (FeS_2), iron and aluminum sulfates,[10,12] and also organic matter[13] each provide acid through microbial activity for indirect acidification. Plant growth can also help dissolve native calcium carbonate.[12]

Miscellaneous Chemical Ameliorants

Sodic soils may be reclaimed with soluble sources of Ca that include waste materials, e.g., acidic cottage cheese whey,[1] acid resin by-products of the oil industry, and also $CaCl_2$ as an industrial by-product.[3] Addition of polymers can also increase the permeability of soils.[1]

Role of Biology and Organic Matter

Effects of Plant Growth

Rice culture ameliorates sodic soils through the buildup of carbon dioxide and hence dissolution of calcium carbonate, among other effects.[10] Among crops, rice is particularly tolerant to sodicity, and some grasses are also highly tolerant[10,12] (Table 2).

Ameliorants are readily leached out of soils, so losing the benefit of the electrolyte effect. In addition, Na^+ from sodium salts often present in soils can replace exchangeable Ca^{2+}. Therefore long-term reclamation probably requires increased plant growth, hence enhanced biological activity[13] and stabilization of soil structure. This occurs partly through transient binding agents, e.g., roots, hyphae, polysaccharides, and hydrophobic binding[13] and partly through associations between soil organic matter and Ca^{2+} in the soil system. Organic matter has a stronger affinity for Ca than for Na,[13] so is most effective when used with gypsum or lime to help restore a stable structure to soils.[13] There is a resulting increase in stable pores for the transmission of water and nutrients, root growth, and soil stability.

TABLE 1 Composition of Some Mined Gypsum

Sample	Fe_20_3 (%)	MnO (%)	TiO_2 (%)	CaO (%)	k_2O (%)	SO3 (%)	P_2O_5 (%)	SiO_2 (%)	AI_2O_3 (%)	MgO (%)	Na_2O (%)	ZrO_2 (%)	Sr (%)	SUM (%)	OrgC (%)	$CaCO_3$ (%)
Plant life	0.48	0.010	0.10	25.7	0.260	30.6	0.057	18.9	1.16	0.47	0.18	0.001	0.271	78.1		
Neindorf	1.54	0.021	0.19	23.7	0.662	28.9	0.066	17.6	3.32	1.64	0.28	0.001	0.300	78.3	0.41	1.8
Salt lake	0.03	0.003	0.02	33.5	0.011	29.7	0.051	0.20	0.07	0.009	0.05	0.001	0.241	73.8	0.53	3.4
Austral Pacific	0.26	0.006	0.04	22.6	0.70	28.5	0.047	30.2	0.56	0.04	0.03	0.001	0.115	82.4	0.07	0.1
Agra unsieved	0.15	0.004	0.03	31.2	0.093	39.0	0.050	4.50	0.61	0.06	0.03	0.001	0.341	76.0		
Agra sieved (A)	0.12	0.010	0.03	31.0	0.067	39.5	0.049	3.90	0.40	0.04	0.03	0.01	0.845	75.9	0.02	0.1
Agra sieved (B)	0.11	0.012	0.03	31.0	0.067	39.4	0.053	3.80	0.28	0.06	0.03	0.001	0.847	75.6	0.1	
Cresco	0.15	0.004	0.02	31.0	0.089	39.2	0.071	4.20	0.37	0.07	0.04	0.001	0.357	75.5	0.03	0.1
Top gypsum	0.32	0.006	0.10	23.4	0.340	28.5	0.057	26.3	1.12	0.35	0.19	0.001	0.194	80.9		
Jomoco (A)	0.46	0.010	0.08	28.0	0.237	33.9	0.286	11.8	1.02	0.68	0.17	0.001	0.403	77.0	n.d.	n.d
Jomoco (B)	0.44	0.007	0.07	27.9	0.232	33.7	0.277	12.4	1.01	0.67	0.14	0.001	0.402	77.2	n.d.	n.d.

Source: Naidu.[61]

TABLE 2 Relative Tolerance of Crops to Sodicity

ESP Range	Crops[a]
10–15	Safflower, mash, pea, lentil, pigeon-pea, curd bean
16–20	Bengal gram, soybean
20–25	Groundnut, cowpea, onion, pearl millet
25–30	Linseed, garlic, guar
30–50	Indian mustard, wheat, sunflower, berseem, hybrid napier, guinea grass
50–60	Barley, sesbania, saftal, panicums
60–70	Rice, para grass
70+	Karnal, rhodes, and bermuda grasses

[a] Yields are about 50% of the potential yields in the respective sodicity ranges.
Source: Gupta and Abrol[10] and Oster and Jayawardene.[12]

Effects of Additions of Organic Matter

Due to its stabilizing effect on soil structure, organic matter can also help to reclaim sodic soils when added to soils, even though additions of organic matter can sometimes also enhance their dispersion, with resulting physical problems.[13] The different effects vary with types of soils and also of organic matter.[13] The addition of readily decomposable organic matter tends to reduce the pH of calcareous sodic soils, and hence aid their reclamation.[13]

Strategies for the Reclamation of Sodic Soils

Reclamation without the Addition of Ameliorants

Reclamation can be achieved slowly without ameliorants by many cycles of irrigation and cropping.[10] As well, underlying calcareous or gypsiferous layers may be incorporated into upper parts of the soil profile through deep plowing.[3]

Optimal Supply of Ameliorants

The effectiveness of gypsum can be increased by its placement in cracks for their stabilization so that they remain open to conduct water, air, and plant nutrients when soils become wet.[1] Water flow to subsoil layers, and also aeration, hence higher crop yields, were achieved by incorporating gypsum in narrow tilled slots.[12] More generally, gypsum is often incorporated into subsoil layers by tillage.

Ameliorants may also be used in association with other additions. Only a little gypsum applied with acidifying nitrogenous fertilizers led to effective reclamation of a saline-sodic soil.[1] This combination confers immediate benefits through a higher electrolyte concentration and also medium-term benefits through an increased supply of exchangeable Ca ions. Enhanced plant growth from fertilizer use provides long-term benefits.

Supply of Water for Reclamation

Ameliorants are only effective when dissolved. A supply of water and its flow through the profile are required for effective amelioration. Irrigation and/or rainfall may supply water but the quality of water available for irrigation can vary. Gypsum can be supplied in electrolyte-free irrigation water, as in parts of California.[12] Saline-sodic irrigation water, including water commonly derived from groundwater sources, may exacerbate the deleterious effects of sodicity when its application during the dry season is followed by rains, as can occur in Israel.[12] In this case, (by-product) gypsum is applied to the surface prior to the rainy season.[12] Adequate drainage to beyond the root zone should accompany irrigation in order to leach displaced electrolyte but this should not cause increased salinity in underlying groundwater.[11]

As long as there is provision for the disposal of the salt water, dilution with salt water[3,11] is also useful in sodic soil reclamation. Its immediate effect is to increase the hydraulic conductivity of sodic soils by the electrolyte effect, but long-tern reclamation will occur provided it also contains sufficient Ca (Ca : total cations ≥ 0.3). This can be added as gypsum.

Rate of Supply of Ameliorants

All additives increase costs so their optimal use is required. The cost of transporting gypsum often limits application rates. Recommendations from field trial results of the most effective application rates for ameliorants are necessarily specific to soil type and climate, among other factors. Nonetheless, both simulations[9,14,15] and laboratory studies[16] have shown that the solubility of gypsum increases as the proportion of exchangeable sodium on the soil increases. They have also shown that gypsum is more effective when mixed throughout the soil than when applied to the soil surface alone. Simulations[9] have shown that reclamation is primarily limited to the depth interval in which gypsum was applied (Figure 3A). They have also shown (Figure 3B) that the electrical conductivity EC of the solution within and below the gypsum-amended layer decreases as the exchangeable sodium percentage (ESP) within that layer decreases.

Nonetheless, calculations of gypsum requirements have been based on 1) exchange of Na by Ca, rather than the attainment of a sufficiently high electrolyte concentration EC to maintain a low permeability; and 2) assumptions of chemical equilibrium. The gypsum requirements for the achievement of a suitable permeability by the electrolyte effect also depend on the inherent EC of the soil solution.[11] Because mineral dissolution can maintain a high EC throughout most of the soil profile, gypsum often needs to be added to the soil-surface only, mainly to prevent crusting. Assumptions of equilibrium, while strictly invalid, have probably been useful for the calculation of gypsum requirements because of low flow rates of percolating solutions and high surface areas of gypsum particles.[10] Models of soil and water flow, e.g., UNSATCHEM[16] which consider equilibrium and kinetic expressions for ion exchange and the dissolution and precipitation of minerals and also ameliorants, may enable more robust calculations of requirements for ameliorants.

Ultimately, the main purpose of reclaiming sodic soils is to obtain the maximum possible improvement of yield of the particular crop in relation to the cost of ameliorant applied. Work in the Australian

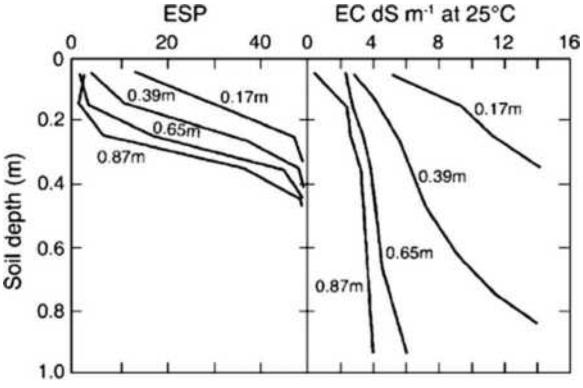


FIGURE 3 Computer-model results (ESP and EC) for reclamation of a soil (initial ESP = 50; CEC = 200 $\text{mmol}_c\ \text{kg}^{-1}$ with gypsum and water (EC = 0). Numbers next to each line are depths of applied water.
Source: Oster and Frenkel.[9]

sugar industry[17] has established a quantitative relationship between percentage increase in ESP and consequent loss in sugarcane yield. Sugarcane yield decreased by 1.5–2.1 ha^{-1} for every 1% increase in ESP.[17] This relationship forms the basis of a cost–benefit analysis for calculating gypsum application rates from the ESP, the quality of the gypsum used, the quality of irrigation water and economic factors. These include the cost of gypsum and its application, the price for the product (sugarcane) and the discount rate.[18] Clearly there are opportunities for more practical calculations of this kind to be carried out for other crops and in other soil and climatic environments.

References

1. Sumner, M.E. Sodic soils: new perspectives. Austr. J. Soil Res. **1993**, *31* (6), 683–750.
2. Quirk, J.P.; Schofield, R.K. The effect of electrolyte concentration on soil permeability. J. Soil Sci. **1955**, *6* (2), 163–178.
3. Loveday, J. Amendments for reclaiming sodic soils. In *Soil Salinity Under Irrigation: Processes and Management*; Shainberg, I., Shalhevet, J., Eds.; Springer: Berlin, 1984; 353–374.
4. Rengasamy, P.; Sumner, M.E. Processes involved in sodic behaviour. In *Sodic Soils;* Sumner, M.E., Naidu, R., Eds.; Oxford University Press: New York, 1998; 35–50.
5. Shainberg, I.; Sumner, M.E.; Miller, W.P.; Farina, M.P.W.; Pavan, M.A.; Fey, M.V. Use of gypsum on soils: a review. Adv. Soil Sci. **1989**, *9,* 1–112.
6. Naidu, R.; Merry, R.H.; Churchman, G.J.; Wright, M.J.; Murray, R.S.; Fitzpatrick, R.W.; Zarcinas, B.A. Sodicity in south Australia. a review. Austr. J. Soil Res. **1993**, *31* (6), 911–930.
7. Rengasamy, P.; Olsson, K.A. Irrigation and sodicity. Austr. J. Soil Res. **1993**, *31* (6), 821–837.
8. Levy, G.J.; Shainberg, I.; Miller, W.P. Physical properties of sodic soils. In *Sodic Soils;* Sumner, M.E., Naidu, R., Eds.; Oxford University Press: New York, 1998; 77–94.
9. Oster, J.D.; Frenkel, H. The chemistry of the reclamation of sodic soils with gypsum and lime. Soil Sci. Soc. Am. J. **1980**, *44* (1), 41–45.
10. Gupta, R.K.; Abrol, I.P. Salt-affected soils: their reclamation and management for crop production. Adv. Soil Sci. **1990**, *11*, 223–288.
11. Oster, J.D.; Shainberg, I.; Abrol, I.P. Reclamation of salt-affected soil. In *Soil Erosion, Conservation and Rehabilitation;* Agassi, M., Ed.; Marcel Dekker: New York, 1996; 315–352.
12. Oster, J.D.; Jayawardene, N.S. Agricultural management of sodic soils. In *Sodic Soils;* Sumner, M.E., Naidu, R., Eds.; Oxford University Press: New York, 1998; 125–147.
13. Nelson, P.N.; Oades, J.M. Organic matter, sodicity and soil structure. In *Sodic Soils;* Sumner, M.E., Naidu, R., Eds.; Oxford University Press: New York, 1998; 51–75.
14. Tanji, K.K.; Deverel, S.J. Simulation modeling for reclamation of sodic soils. In *Soil Salinity Under Irrigation: Processes and Management;* Shainberg, I., Shalhevet, J., Eds.; Springer: Berlin, 1984; 238–251.
15. Frenkel, H.; Gerstl, Z.; Alperovitch, N. Exchange induced dissolution of gypsum and the reclamation of sodic soils. J. Soil Sci. **1989**, *40* (3), 599–611.
16. Suarez, D.L.; Šimunek, J. UNSATCHEM: unsaturated water and solute transport model with equilibrium and kinetic chemistry. Soil Sci. Soc. Am. J. **1997**, *61,* 1633–1646.
17. Nelson, P.; Ham, G. Exploring the response of sugar cane to sodic and saline conditions through natural variation in the field. Field Crops Res. **2000**, *66* (3), 245–255.
18. Nelson, P.; Fitzgerald, T.; Swan, G.; Brennan, L. *Gypsy, Discounted Cash Flow Analysis for Application of Gypsum to Sodic Soils Under Sugarcane: Manual;* CRC for Sustainable Sugar Production: Townsville, Australia, 2000.
19. van Olphen, H. *An Introduction to Clay Colloid Chemistry;* Wiley: New York, 1963.

62

Tillage Erosion: Terrace Formation

Seth M. Dabney and
Dalmo A.N. Vieira

Introduction

All tillage moves soil laterally as well as vertically. While a few implements such as a moldboard plow or a one-way disk plow throw soil to only one side, most tillage implements—including tandem disks, chisel plows, harrows, and cultivators—throw soil to both sides. With such implements, a tillage operation along the contour moves some soil uphill, more soil downhill, and still more soil along the contour in the direction of travel.[1–3] Within an undulating field, each tillage operation flattens the topography as soil is progressively removed from convexities and is deposited in depressions and other concavities.

Field boundaries interrupt soil fluxes caused by tillage and can result in the formation of two distinct classes of terraces. On sloping land with untilled strips, tillage moves soil toward a down-gradient untilled strip and away from an up-gradient untilled strip. This leads to aggradation of soil upslope of untilled strips and degradation of soil downslope of them and causes the gradual formation of bench terraces. The speed of bench terrace formation is greater where the spacing between the untilled strips is narrower, where the initial slope is steeper, and where tillage is more intense. The other class of terraces formed at field boundaries is due to local ridges and furrows. Depending on tillage tool design and operation, some soil often leaves the tilled zone in the form of clods deposited along field borders. The combination of clods thrown from the tilled area and dead furrows left behind creates small terrace channels at the margins of tilled areas that can intercept and alter runoff flow patterns. The impact of these small gradient terraces on hydrology is greatest on gently sloping lands. This entry reviews research into terrace formation by tillage, identifies ways in which these terraces can be advantageous to soil and water conservation, and discusses problems that have been recognized with tillage terraces.

Ancient Lynchets

Archeologists use the term "lynchets" to refer to soil banks that are believed to be remnants of ancient agricultural activities. These ancient soil banks have been studied intensively for over 100 years. Lynchets dating from the bronze age, iron age, and medieval periods are all generally recognized to be the morphological response on a hillslope to the presence of field boundaries in cultivated landscapes.[4] While medieval lynchets tend to be oriented along the slope, older "Celtic fields" were often square and bounded by lynchets on all sides,[5] and some lynchets are oriented up and down the slope.[6] Lynchets form at all field boundaries whether bounded by untilled grass strips (balks), hedge rows, paths, fences, or ditches.[6,7] Frequently, lynchets contain piles of stones at their core. These stones may have marked property boundaries or may simply represent convenient disposal locations as stones were removed from fields.

Contemporary Lynchets

Contemporary lynchet formation has been observed and described around the globe.[7–23] The presence of a physical boundary such as field borders, fences, hedge rows, or vegetative barriers creates a local interruption of the soil movement created by tillage, resulting in morphological changes in the vicinity of the boundary. Erosion or deposition occurs depending on the orientation of terrain slopes with respect to the boundary. When the tilled terrain slopes away from the boundary, the net soil translocation in the downhill direction creates erosion near the boundary, while soil deposition occurs where the terrain slopes toward the boundary. Localized erosion also occurs wherever tillage implements first engage the soil, and deposition occurs at locations where tillage tools are removed from the ground and turned around. The aggrading side of a field boundary is often termed a "positive lynchet," while the degrading side (the upslope edge of a tilled area) is termed a "negative lynchet" (Figure 1). Positive lynchets have deeper, more fertile, and more productive soils than those on negative lynchets, where subsoil may be exposed.[12,19,21,22,24]

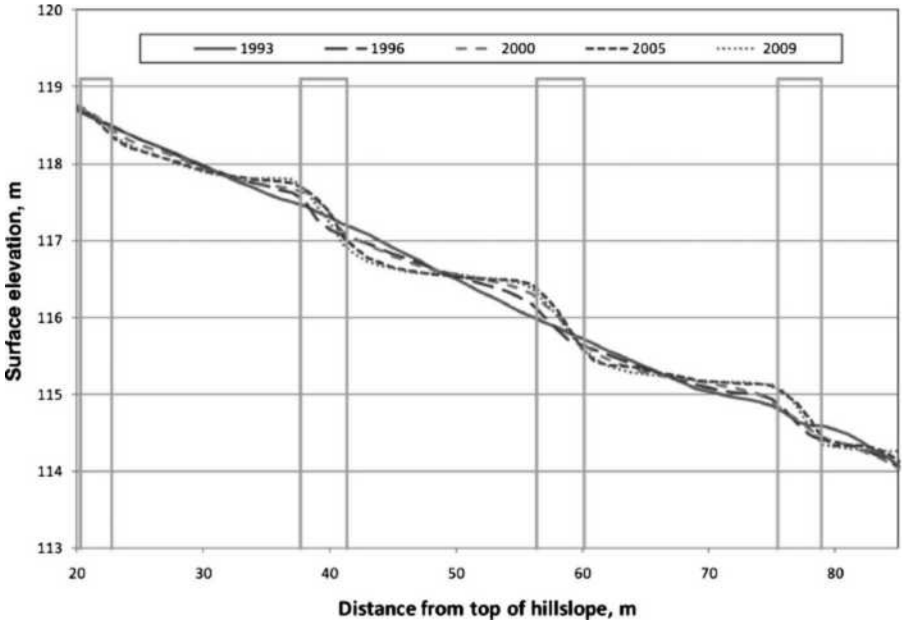


FIGURE 1 Patterns of erosion and deposition created by tillage and water erosion and deposition in a field near Coffeeville, Mississippi, United States. Tillage translocation leads to erosion downslope of each grass hedge, creating the negative lynchets, while deposition due to tillage and water transport reinforces the positive lynchet.

A modern example of lynchet formation is provided by a field subdivided by grass hedges planted on the contour in North Mississippi, United States, that was surveyed periodically over 16 years (Figure 1). Tillage between the hedges led to erosion in the downslope side of each hedge as soil was displaced in the downhill direction by tillage, and the hedges caused deposition of sediment transported by water and tillage on their upslope side. Lynchets developed rapidly once contour tillage with disks and chisel plows began between the grass hedges.[25,26] Changes in elevation reflected the combined influences of tillage and water erosion. At each grass hedge, lynchets of up to 0.8 m formed during a period of 16 years, and the average slope of field was reduced from 7.2%, when the hedges were first established, to 3.7%.[27]

Hand tillage with a hoe differs from that associated with draft animals or tractors because soil usually moves in a direction opposite of the workers' travel. Because hoeing is hard work, it is common for soil to move mainly downhill, with work starting at the bottom of a field and proceeding upslope.[19] Again, the net result is gradual terrace formation.

Soil erosion by water generally increases with increasing slope length because of an accumulation of runoff. This should lead to the development of concave hillslope profiles with little change in elevation near hilltops.[28] In contrast, the quantity of soil moved by tillage is independent of field size and leads to rapid soil loss at upslope field boundaries. Since erosion rates are generally considered in terms of mass per unit area, the amount of soil "eroded" by tillage is much greater on narrow fields than on long slopes.[19,23,29] Thus, the more strips into which a field is divided, the greater the contribution of tillage translocation to terrace formation, and the more rapid terrace formation will be.

Engineered Terraces

Engineered terraces are designed to manage runoff from a certain contributing area. Therefore, when a field is to receive more than one terrace, it is common practice to start at the top of the field and work down so that the bottom terrace does not get overloaded by a storm that might occur before upslope terraces are completed. While specialized equipment such as bulldozers and scrapers are generally used for terrace construction, ordinary farm machinery can also be used to form broad-base terraces,[30] and proper tillage operations are needed as part of routine maintenance to preserve functionality of all terraces.[8]

Bench Terraces

For deep soils, tillage translocation can be used to reduce the cost and increase the farmability of bench terraces. On irregular fields, starting at the bottom of the slope may facilitate efficient development of straight parallel terrace systems. First, a bulldozer pushes up soil from below into a terrace with extra height where it crosses existing gullies.[31,32] A tile outlet near the thalweg drains impounded runoff, leaving sediment trapped in the low area. When the thalweg area is sufficiently filled (up to 5 m of fill in 3 years), further sediment is cut off by construction of the next upslope terrace, parallel to the first. In one case, downhill moldboard plowing between terraces accounted for 50% of the soil moved, reducing a slope from 14% to 4% over a 15 years period. Similar to the natural development of lynchets that occurs when tillage is performed between strips of permanent vegetation, bench terraces can be developed by soil moved by tillage operations. Tillage smoothes the benches, and the vegetation conserves soil and water while benches gradually develop.

A similar process occurs when fields are tilled between parallel vegetated strips, as described with reference to lynchet formation. The vegetation conserves soil and water while benches gradually develop. This process has been formalized by the U.S. Department of Agriculture–National Resources Conservation Service (USDA-NRCS) in the vegetative barrier national conservation practice standard.[33] The vertical interval between vegetated strips should not exceed 2 m, and the planned vegetated back slope should not be steeper than 1 on 2, horizontal to vertical, which is consistent with historical practice.[8]

Gradient Terraces

Gradient terraces are ridge-and-channel systems that reduce erosion by intercepting surface runoff and redirecting it to a stable outlet at reduced velocities. Tillage adjacent to vegetated strips may create small berms that form when soil is thrown beyond the implement width into and against the vegetation and deposited next to it and within it, later coalescing into a stable ridge. If the tillage boundary remains unchanged, subsequent operations increase the berm height until soil cannot be thrown atop the berm. When the vegetated strip deviates from the contour, these tillage induced channels can act as gradient terraces and divert an appreciable fraction of runoff.

A microtopography survey was conducted at 28 locations along the upslope edge of several grass strips in a field with 6%–9% slope steepness,[25] farmed with tandem disks and chisel plows. At every location, asymmetrical triangular channels (Figure 2) formed during 7 years of commercial farming. The hydraulic radius of the smallest channel was 0.04 m. At 0.2% grade and Manning's n = 0.025, this smallest channel could carry 25 mm/hr of runoff from a 15 m–wide cropped strip more than 100 m long. Since the distance between local depressions in the field is usually less than 200 m, these tillage-induced channels could divert most of the fields' runoff to these depressions, where a stable outlet could be provided. Reducing downslope runoff accumulation interrupts slope length and reduces soil erosion by water.

A more detailed study showed that when tillage was conducted with tandem disks and chisel plows immediately adjacent to the vegetation, berms quickly formed along grass hedges planted close to the contour at the bottom of experimental plots with 22.1 m–long, 5% slopes.[27] Four years after tillage adjacent to hedges had begun, berms of triangular shape attained an average height of 0.13 m (Figure 3), forming a shallow channel in the upslope side of the vegetation that diverted runoff. Observations indicated when no soil berm was present, most (>95%) runoff passed through grass hedges planted on 0.3% gradient from a true contour. Once tillage berms were formed, however, runoff from small storms no longer crossed the vegetation but was diverted to flow along the upslope side of the berm and grass hedge. The resulting berms diverted more than 70% of runoff from runoff events less than 10 mm/d and more than 55% for events less than 80 mm/d.[34]

Tillage berms may or may not form, depending on the design and adjustment of tillage tools, which affect how soil is displaced to the sides. When soil is thrown beyond the implement's width, a line of deposition is created; if soil is moved away from the boundary, a dead furrow is formed. The presence of a grass hedge or other fixed boundary ensures that the position of the first pass near the boundary

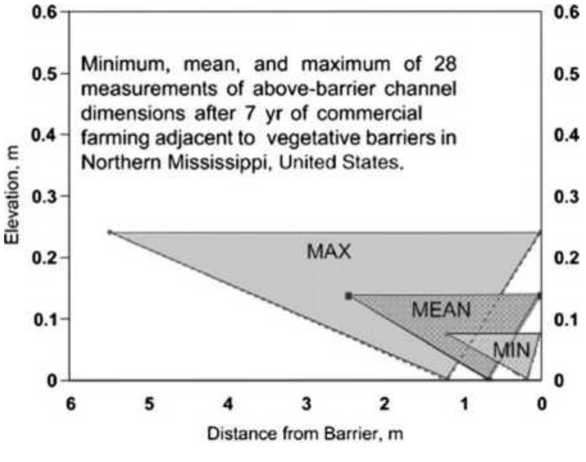


FIGURE 2 Typical channel cross sections observed after several years of tillage with disks and chisel plows. When vegetated strips deviate slightly from the contour, small channels can act as small gradient terraces and divert an appreciable fraction of the runoff.

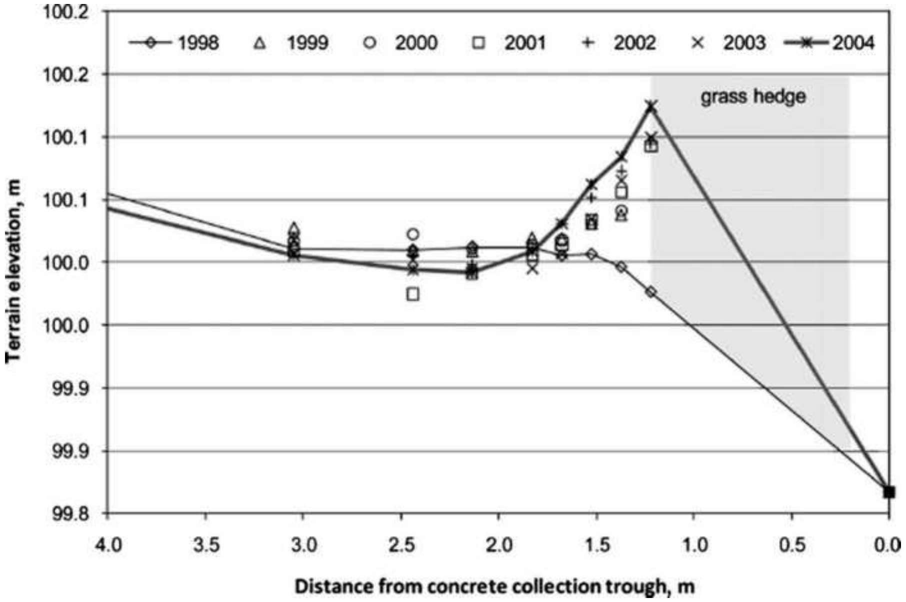


FIGURE 3 Evolution of topography in the lowest 4 m of research plots with grass hedges and an average contributing area steepness of 5%. Chisel and disk tillage was kept a distance away from hedges from 1996 through 2000 and was moved adjacent to the hedges from 2001 through 2004.
Source: Vieira and Dabney.[27]

is maintained, facilitating the formation of berms or furrows. Tillage implement type and speed, slope steepness, and the superposition of sequential tillage operations determine the formation of berms and their dimensions. On flat lands, the creation of dead furrows and ridges associated with the layout of lands used to till the fields makes the tillage patterns clearly visible in light detection and ranging topographic data[35] and can greatly alter surface hydrology.

Terrace Benefits and Problems

Bench terraces facilitate contour farming operations. On very steep slopes, mechanical cultivating is difficult, and contour plowing can be dangerous due to the risk of overturning. The formation of bench terraces reduces the slope steepness and therefore makes contour tillage easier and safer. When tractor or animal power is limited, it may be easier to plow or hoe downhill.[11,29] However, when sufficient power is available, it is most efficient to plow on the contour because a uniform load can be matched to a tractor or team of animals.

Contour tillage reduces runoff and soil erosion, thereby conserving plant nutrients and making more water available to grow crops.[36] On the other hand, large fertility differences can develop between the negative and positive lynchets.[12,19,24] This reduces uniformity of crop growth where fertilizer and liming amendments are not available. In these situations, farmers often cut soil from the terrace back slope and distribute it on the degraded negative lynchet. If the upper field is owned by a different landowner, an entire terrace may be undermined in this fashion.[12] Manuring, fertilization with seaweed, and liming with marl were ancient responses to similar problems.[37]

While tillage berms acting as gradient terraces can improve soil conservation where runoff is diverted to a stable outlet, they can have negative consequences on the effectiveness of some conservation practices. For example, if berms keep a significant fraction of runoff water from entering a grassed waterway, erosion on the margin of the waterway may be increased. Tillage berms may keep water from entering a vegetated filter strip on very flat fields (Figure 4), which may preclude the filter from improving water

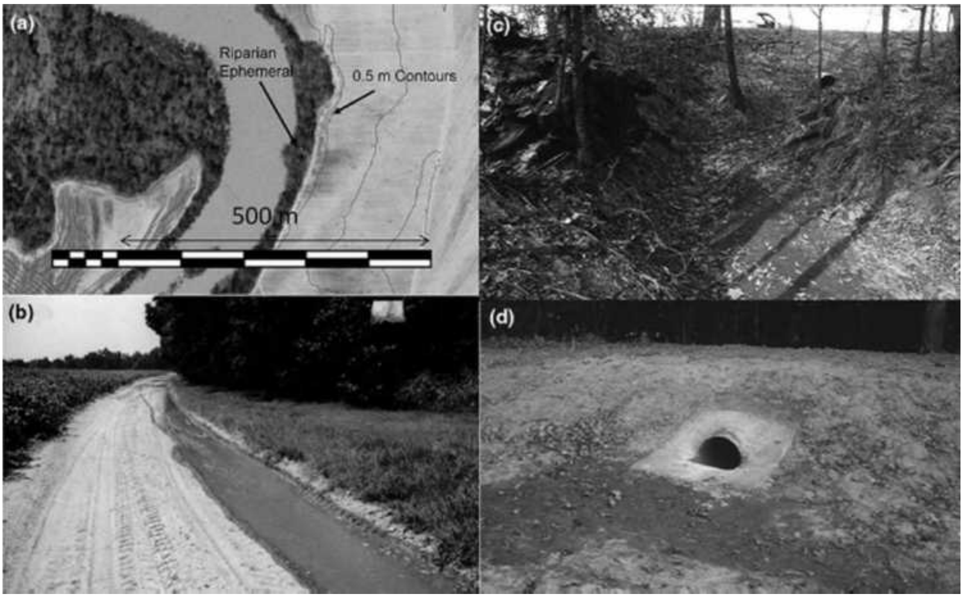


FIGURE 4 Runoff reaching (a) a filter strip/riparian forest buffer surrounding an oxbow lake in the Mississippi Delta was (b) redirected by a small tillage berm to (a and c) a riparian ephemeral flow channel. Storm flow through this channel was controlled by (c and d) a slotted inlet pipe and associated earthen pad, which created backwaters that allowed the riparian buffer to function (**Source:** Dabney, Moore, and Locke.[38])

quality from small runoff flows. This problem is the main reason why the USDA-NRCS restricts the establishment of filter strips to fields with an upslope contributing area steepness greater than 1%.[39] Tillage berms are also implicitly recognized in the USDA-NRCS contour buffer strip standard[40] that states, “If sediment accumulates just below the upslope edge of the buffer strip to a depth of 6 inches or more … relocate the buffer/cropped strip interface location.”

Conclusions

Tillage erosion has come to be recognized as an important contributor to the evolution of agricultural landscapes and, in some cases, the degradation of soil resources. With each tillage operation, a tilled area becomes flatter, and some soil leaves the tilled zone in the form of clods deposited along field borders. Over time, these clods, along with stones removed from the tilled area and sediment deposited by runoff, coalesce to form terraces or lynchets. If a field is divided into several tilled strips, then the consequence will be the gradual development of bench terraces, unless the untilled areas are periodically relocated. Also, tillage-induced berms located at the upslope edge of the untilled areas may act as small gradient terraces that alter runoff flow patterns. On flat lands, tillage berms may dominate surface drainage patterns. The formation of such bench and gradient terraces due to agricultural tillage operations may or may not be desirable but can hardly be avoided if field boundaries on sloping lands are fixed for long periods of time.

References

1. Lindstrom, M.J.; Nelson, W.W.; Schumacher, T.E.; Lemme, G.D. Soil movement by tillage as affected by slope. Soil Tillage Res. **1990,** *17* (3–4), 255–264.
2. Lobb, D.A.; Kachanoski, R.G.; Miller, M.H. Tillage translocation and tillage erosion in the complex upland landscapes of southwestern Ontario, Canada. Soil Tillage Res. **1999**, (3–4), 189–209.

3. Van Oost, K.; Govers, G.; De Alba, S.; Quine, T.A. Tillage erosion: A review of controlling factors and implications for soil quality. Prog. Phys. Geogr. **2006**, *30* (4), 443–466.
4. Fisher, P.F. Pedogenesis within the archaeological landscape at South Lodge Camp, Wiltshire, England. Geoderma, **1983**, *29,* 93–105.
5. Fowler, P.J.; Evans, J.G. Plough-marks, lynchets and early fields. Antiquity **1967**, *41,* 289–301.
6. Smith, R.T. Early agriculture and soil degradation. In *The Effect of Man on the Landscape: The Highland Zone*; Evans, J.G., Limey, S., Cleere, H., Eds.; CBA Res. Rep. No. 11, 1975, 27–37.
7. MacNab, J.W. British strip lynchets. Antiquity **1965**, *39,* 279–290.
8. USDA-Soil Conservation Service. *A Manual of Conservation of Soil and Water,* Agriculture Handbook No. 61; U.S. Department of Agriculture: Washington, DC, 1954; 79–107.
9. Papendick, R.I.; Miller, D.E. Conservation tillage in the Pacific Northwest. J. Soil Water Conserv. **1977**, *32,* 49–56.
10. Abujamin, S.; Abdurachman, A.S. Contour grass strips as a low cost conservation practice. Food and Fertilizer Technology Center Extension Bulletin No. 221; ASPAC: Taipei City, Taiwan, Republic of China, 1985; 1–7.
11. Quine, T.A.; Walling, D.E.; Zhang, X.; Wang, Y. Investigation of soil erosion on terraced fields near Yanting, Sichuan Province, China, using cesium-137. In *Erosion, Debris Flows and Environment in Mountain Regions*, Proceedings of the Chengdu Symposium, July 1992; IAHS Publ. no. 209, 155–168.
12. Lewis, L.A. Terracing and accelerated soil loss on Rwan-dian steeplands: A preliminary investigation of the implications of human activities affecting soil movement. Land Degrad. Rehabil. **1992**, *3*, 241–246.
13. Govers, G.; Vandaele, K.; Desmet, P.J.; Poessen, J.; Bunte, K. The role of tillage in soil redistribution on hillslopes. Eur. J. Soil Sci. **1994**, *45,* 469–478.
14. Revel, J.C.; Guiresse, M. Erosion due to cultivation of calcarious clay soils on the hillsides of south west France. I. Effect of former farming practices. Soil Tillage Res. **1995**, *35,* 147–155.
15. Morgan, R.P.C.; Rickson, R.J. Water erosion control. In *Slope Stabilization and Erosion Control: A Bioengineering Approach;* Morgan, R.P.C., Rickson, R.J., Eds.; E & FN Spon: London, 1995; 133–190.
16. Kiepe, P. Cover and barrier effect of *Cassia siamea* hedgerows on soil conservation in semi-arid Kenya. Soil Tech. **1996**, *9,* 161–171.
17. McEntee, M.A. Colluvial processes and soil variation at field boundaries in County Down. Ir. Geogr. **1998**, *31* (1), 55–69.
18. Thapa, B.B.; Cassel, D.K.; Garrity, D.P. Ridge tillage and contour natural grass barrier strips reduce tillage erosion. Soil Tillage Res. **1999**, *51,* 341–356.
19. Turkelboom, F.; Poesen, J.; Ohler, I.; Ongprasert, S. Reassessment of tillage erosion rates by manual tillage on steep slopes in northern Thailand. Soil Tillage Res. **1999**, 245–259.
20. Van Oost, K.; Govers, G.; Desmet, P. Evaluating the effects of changes in landscape structure on soil erosion by water and tillage. Landscape Ecol. **2000**, *15,* 577–589.
21. Nyssen, J.; Haile, M.; Moeyersons, J.; Poesen, J.; Deckers, J. Soil and water conservation in Tigray (northern Ethiopia): The traditional daget technique and its integration with introduced techniques. Land Degrad. Dev. **2000**, *11*, 199–208.
22. Follain, S.; Minasny, B.; McBrantney, A.B.; Walter, C. Simulation of soil thickness evolution in a complex agricultural landscape at fine spatial and temporal scales. Geoderma **2006**, *33*, 71–86.
23. Dercon, G.; Govers, G.; Poesen, J.; Sánchez, H.; Rombaut, K.; Vandenbroeck, E.; Loaiza, G.; Deckers, J. Animalpowered tillage erosion assessment in the southern Andes region of Ecuador. Geomorphology **2007**, *87,* 4–15.
24. Dercon, G.; Deckers, J.; Govers, G.; Poesen, J., Sánchez, H.; Vanegas, R.; Ramírez, M.; Tacuri, E.; Loaiza, G. Spatial variability in crop response under contour hedgerow systems in the Andes region of Ecuador. Soil Tillage Res. **2006**, *86,* 15–26.

25. Dabney, S.M.; McGregor, K.C.; Meyer, L.D.; Grissinger, E.H.; Foster, G.R. Vegetative barriers for runoff and sediment control. In *Integrated Resources Management and Landscape Modification for Environmental Protection;* Mitchell, J.K., Ed.; ASAE: St. Joseph, MI, 1993; 60–70.
26. Dabney, S.M.; Liu, Z.; Lane, M.; Douglas, J.; Flanagan, D.C. Landscape benching from tillage erosion between grass hedges. Soil Tillage Res. **1999**, *51,* 219–231.
27. Vieira, D.A.N., Dabney, S.M. Modeling edge effects of tillage erosion. Soil Tillage Res. **2011**, *111*, 197–207.
28. Meyer, L.D.; Kramer, L.A. Erosion equations predict land slope development. Agric. Eng. **1969**, *50,* 552–523.
29. Poesen, J.; van Wesemael, B.; Govers, G.; Martinez-Fernandez, J.; Desmet, P.; Vandaele, K.; Quine, T.; Degraer, G. Patterns of rock fragment cover generated by tillage erosion. Geomorphology **1997**, *18,* 183–197.
30. Samsel, L.G. Building and maintaining terraces with ordinary farm machinery. Agric. Eng., **1943**, *24,* 337–338, 342.
31. Jacobson, P. New methods of bench terracing steep slopes. Trans. ASAE **1963**, *6*, 257–258, 261.
32. Jacobson, P. Improvements in bench terracing. In the Proceedings of the 60th Annual Meeting of the ASAE, Saskatoon, Saskatchewan, Canada, June 27–30, 1967; American Society of Agricultural Engineering: St. Joseph, MI; 1–11.
33. USDA-NRCS. Vegetative Barrier, Code 601; 2010, available at https://efotg.sc.egov.usda.gov/references/public/CO/CO601_SOW.pdf (accessed February 2011).
34. Dabney, S.M.; Wilson, G.V.; McGregor, K.C.; Vieira, D.A.N. Runoff through and upslope of contour switchgrass hedges. Soil Sci. Soc. Am. J. **2012**, *76,* 210–219.
35. Li, S.; Lobb, D.A.; Tiessen, K.H.D. Modeling tillage-induced morphological features in cultivated landscapes. Soil Tillage Res. **2009**, *103,* 33–45.
36. Wadsworth, R.; Swetnam, R. Modeling the impact of climate warming on the landscape scale: Will bench terraces become economically and ecologically viable structures under changed climate? Agric. Ecosyst. Environ. **1998**, *68,* 27–39.
37. Bell, M. The prehistory of soil erosion. In *Past and Present Soil Erosion: Archaeological and Geographical Perspectives;* Bell, M., Boardman, J., Eds.; Oxbow Monograph 22, Oxbow Books: Oxford, U.K., 1992; 21–35.
38. Dabney, S.M.; Moore, M.T.; Locke, M.A. Integrated management of in-field, edge-of-field, and after-field buffers. J. Am. Water Resour. Assoc. **2006**, *42* (1), 15–24.
39. USDA-NRCS. Filter Strip, Code 393; 2010, available at https://efotg.sc.egov.usda.gov/references/public/MN/393mn.pdf (accessed February 2011).
40. USDA-NRCS. Contour Buffer Strips, Code 332; 2010, available at https://efotg.sc.egov.usda.gov/references/public/MN/332mn.pdf (accessed February 2011).

VII

NEC: Natural Elements and Chemicals Found in Nature

63

Erosion by Wind-Driven Rain

Gunay Erpul, L. Darrell Norton and Donald Gabriels

Introduction

Wind-driven rain can be described as raindrops falling through a wind field at an angle different from the vertical under the effects of both gravitational and drag forces. Schematic representation of wind-driven rain incidental to a sloping soil surface is given in Figure 1.

A study of wind-driven rain erosion is an attempt to investigate the combined effect of wind and rain on soil erosion processes in situations where wind and rain occur at the same time.

Effects of Wind on Physical Characteristics of Rains

Raindrop Size Distribution

In the assessment of the distribution of small-simulated raindrops in a wind tunnel, Erpul, Gabriels, and Janssens[1] obtained a narrower raindrop distribution and observed a distinct increase in mean drop diameter under wind-driven rain compared to windless rain. Collisions between small drops occurred more frequently as a result of their greater number per unit volume in air leading to an increase in mean drop size. For large drops, however, this would not occur, as large drops are less stable and the wind caused some of them to break up into smaller drops.[2,3] Basically, the effect of wind on raindrop size distributions is a potentially important effect that needs to be considered when estimating the rainfall erosivity.

Raindrop Impact Energy

Wind-driven raindrops gain some degree of horizontal velocity, which increases their resultant impact velocity. The kinetic energy load of the rainfall may be expected to change as a result of increased velocity and the altered size of the raindrops. The exponential relationship between the horizontal wind

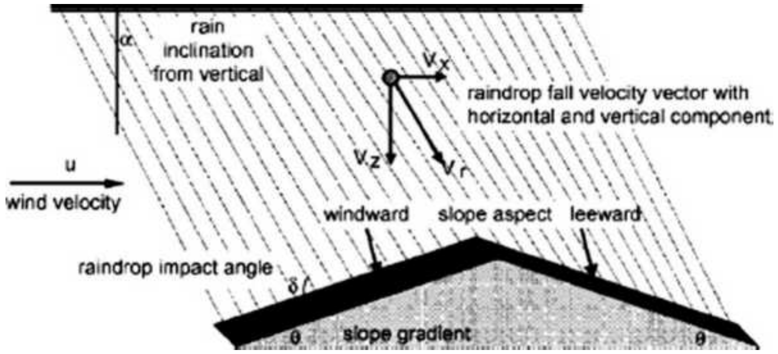


FIGURE 1 Schematic representation of wind-driven rain with an angle from vertical and incident on sloping soil surface.

velocity and the kinetic energy of raindrops was found in natural rains.[4] The effect of the wind on the horizontal component of small raindrops would be greater, so a greater percentage increase in kinetic energy would be expected for small raindrops than large raindrops.[5]

Raindrop Impact Angle

Wind-driven raindrops strike the soil surface with an angle deviated from the vertical because of their horizontal and vertical velocities. This inclination depends on the magnitude of wind velocity. In mid-latitudes rain mostly falls at considerable inclination from the vertical, and resultant angles of 40–70° have been found in rains driven by wind velocity of 10 m/s.[6] Gabriels et al.[7] reported that the mean angles of rain inclination were 52°, 66°, and 67° for the simulated rains in a wind tunnel driven by 6, 10, and 12 m/s of wind velocities, respectively. For these rains, a median drop size was approximately 1.50 mm. Little is known about the physical impact of raindrops on a soil in situations where this impact is not vertical. It is also not known if inclined raindrops have stronger erosive effects than vertical ones. The extent and magnitude of the rain-splash detachment increased as the angle of deviation increased within the range of 5–30°;[8] whereas, with greater deviations, the impact angle could be so small that raindrops would hardly hit the soil surface. Especially when impact angles were less than 30°, the rain-splash detachment rate decreased greatly (Table 1).

Raindrop Impact Frequency

The distribution and the intensity of rain on sloping surfaces differ depending on wind direction and velocity. In fact, the angle of rain incidence ($\boldsymbol{\pi}$), which is a function of rain inclination ($\boldsymbol{\alpha}$), slope gradient ($\boldsymbol{\theta}$), and slope aspect, determines the rain intensity in wind-driven rains. As an example, a windward facing slope can receive two times more rain intensity than a leeward facing slope, or even exceed it in extreme cases for rain inclinations of 40–70°.[6] When the rain inclination and the slope gradient increase, the discrepancies in the rain intensity between wind slope and leeward slope become greater (Table 1).

Effects of Wind on Rainsplash Detachment and Transport

Rainsplash Detachment

Similar to effects on rain characteristics, wind movement and velocities can have a profound effect on some aspects of the soil erosion process. When wind accompanies rain, rainsplash detachment tends to increase owing to the increased energy of raindrops. Pedersen and Hasholt[4] obtained a better

TABLE 1 The Effect of Wind on Rain Energy, Rain Intensity, and Rainsplash Detachment and Transport of a Silt Loam Soil

		Windward							Leeward				
u (m/sec)	KE (J/m2/mm)	α (°)	θ (°)	π (°)	I (mm/hr)	D (g/m^2/sec)	X$_c$ (m)	qs (g/m/ec)	π (°)	I (mm/hr)	D (g/m^2/sec)	X$_c$ (m)	qs (gm/sec)
0	2.23	0	4	4	143	0.23	0.35	0.08	4	161	0.47	0.34	0.16
			9	9	141	0.32	0.38	0.12	9	172	0.75	0.35	0.26
			11	11	132	0.41	0.45	0.18	11	180	0.98	0.36	0.35
6	4.37	52	4	48	92	0.41	0.61	0.25	56	126	0.82	0.73	0.60
			9	44	106	0.51	0.61	0.31	61	107	0.28	0.74	0.21
			11	41	112	0.77	0.62	0.48	63	97	0.30	0.70	0.21
10	13.20	66	4	62	125	2.01	1.26	2.53	70	90	0.78	1.12	0.87
			9	53	137	2.01	1.37	2.75	75	61	0.27	1.05	0.28
			11	55	141	2.91	1.40	4.07	78	50	0.29	0.95	0.28
12	17.94	67	4	63	93	2.26	1.92	4.34	71	66	0.94	1.66	1.56
			9	59	114	4.39	1.86	8.17	76	42	0.28	1.50	0.42
			11	56	120	5.47	1.88	10.28	78	32	0.24	1.17	0.28

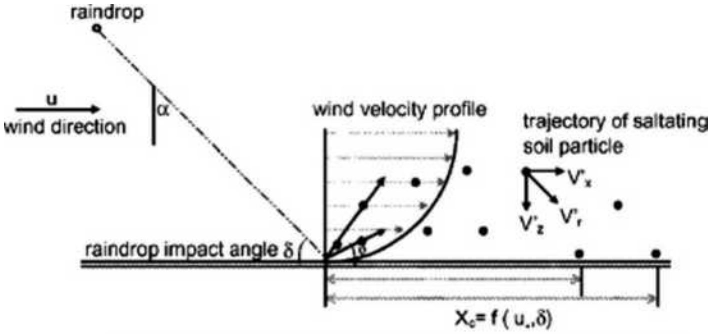
u, horizontal wind velocity; KE, kinetic energy [the exponential relationship found between KE and u is $E(u) = 1.9723\ e^{0.1796u}$ (the relationship between columns 1 and 2)]; α, rain inclination from vertical; *θ*, slope gradient; π, angle of rain incidence calculated using rain inclination, slope gradient, and slope aspect by cosine law: $\cos(\pi) = \cos(\alpha - \theta) = \cos \alpha \cos \theta + \sin \alpha \sin \theta$ for windward slopes and $\cos(\pi) = \cos(\alpha + \theta) = \cos \alpha \cos \theta - \sin \alpha \sin \theta$ for leeward slopes; I, rain intensity; *D*, rainsplash detachment rate; X_c, mean rainsplash distance calculated by center of gravity of mass distribution curves; qs, rainsplash transport estimated by $q_s = DX_c$.

correlation between the rainsplash detachment and erosivity indexes, when wind velocity was taken into account in the kinetic energy calculation. However, the wind not only increased the raindrop impact energy but also altered the angle of rain incidence, resulting in the variations in the raindrop impact frequency and the raindrop impact angle. In other words, raindrop impact energy, impact frequency, and impact angle determine the magnitude of rainsplash detachment under wind-driven rainfall. The influence of each factor on the process is not exclusive and not clearly distinguished, because all are closely related and each factor is a function of the wind speed. The highest rainsplash rate occurred with the highest impact frequency and the highest impact angle at a given raindrop impact energy under wind-driven rains (Table 1).

Rainsplash Transport

Wind, as well as slope and overland flow, is another possible factor capable of transporting detached particles by raindrop impact. It is possible to find similarities between the movement and trajectory of sand saltation by wind and the movement and trajectory of soil particles by rainsplash under wind-driven rainfall. In saltation, the hitting sand grains, once ejected, initiate the motion of uplift,[9] whereas hitting raindrops on the soil surface with an angle initiate a jumping movement of soil particles in wind-driven rainfall.[10] A threshold is given by impacting raindrops, and wind does not account for the upward movement except that it changes energy, frequency, and angle of raindrop impact. Once soil particles are entrained in the splash droplets that have risen into the air by raindrop impact, wind velocity gradient will transport these particles. It is found that, together with wind velocity gradient, the impact angle is also playing a significant role in determining the extent of the process. At the same wind speed, longer mean rainsplash distances (X_c) were observed as the impact angle became greater (Table 1). There was a very discernible decrease in X_c in leeward slopes, which was mostly associated with the impact angles of less than 30°. These results showed that the raindrop impact angle might determine the soil particle ejection angle.

As mentioned above, the amount of soil particles to be transported by the wind will depend upon the raindrop impact energy, impact frequency, and impact angle. Subsequently, the soil particle ejection angle and the wind velocity gradient will determine the extent of the process (Figure 2).



V'_x: the horizontal component of the ejected soil particle velocity;
V'_z: the vertical component of the ejected soil particle velocity;
V'_r: resultant fall velocity of the ejected soil particle;
φ: soil particle ejection angle;
X_c: mean rainplash distance;
u_*: wind shear velocity;

FIGURE 2 Rainsplash transport: raindrop-induced and wind-driven splash trajectories of soil particles falling through a wind profile.

Conclusions

Wind could make significant changes in physical characteristics of rains and hence in soil erosion processes. The rainsplash detachment and transport under wind-driven rainfall would differ from that under windless rain. The rainsplash transport could be a significant process to the extent that it may not be negligible in accurately predicting water erosion, and this process could result in a net transportation in the prevailing wind direction.

References

1. Erpul, G.; Gabriels, D.; Janssens, D. Assessing the drop size distribution of simulated rainfall in a wind tunnel. Soil Till. Res. **1998**, *45,* 455–463.
2. Lyles, L.; Disrud, L.A.; Woodruff, N.P. Effects of soil physical properties, rainfall characteristics, and wind velocity on clod disintegration by simulated rainfall. Soil Sci. Soc. Am. Proc. **1969**, *33* (2), 302–306.
3. Disrud, L.A.; Lyles, L.; Skidmore, E.L. How wind effects the size and shape of raindrops. Agric. Eng. **1969**, *50* (10), 617.
4. Pedersen, H.S.; Hasholt, B. Influence of wind speed on rainsplash erosion. Catena **1995**, *24,* 39–54.
5. Erpul, G.; Gabriels, D.; Janssens, D. The effect of wind on size and energy of small simulated raindrops: a wind tunnel study. Int. Agrophys. **2000**, *14,* 1–7.
6. Sharon, D. The distribution of hydrologically effective rainfall incident on sloping ground. J. Hydrol. **1980**, *46,* 165–188.
7. Gabriels, D.; Tack, K.; Erpul, G.; Cornelis, W.M.; Norton, L.D.; Biesemans, J. Effect of wind-driven rain on splash detachment and transport of a silt loam soil: a short slope wind tunnel experiment. Proceedings of the International Workshop on Technical Aspects and Use of Wind Tunnels for Wind-Erosion Control, and Combined Effect of Wind and Water on Erosion Processes, Ghent, Belgium, Nov 17–18, 1997; Gabriels, D., Cornelis, W.M., Eds.; I.C.E. Special Report No. 1998/1, International Center for Eromology, University of Ghent: Belgium, 1997; 87–93.
8. Van Heerden, W.M. Splash Erosion as Affected by the Angle of Incidence of Raindrop Impact Unpublished. Ph.D. Thesis, Purdue University: Lafayette, IN, 1961.
9. Bagnold, R.A. *The Physics of Blown Sand and Desert Dunes;* 1973 Reprint; Chapman and Hall: London, 1941; 265.
10. De Lima, J.L.M.P.; Van Dick, P.M.; Spaan, W.P. Splash-saltation transport under wind-driven rain. Soil Technol. **1992**, *5*, 151–166.

64

Organic Matter: Global Distribution in World Ecosystems

Wilfred M. Post

Introduction

Globally, the amount of organic matter in soils, commonly represented by the mass of carbon, is estimated to be 1200–1500 Pg C (1 Pg C = 10^{15} g carbon) in the top 1 m of soil.[1,2] This is 2–3 times larger than the amount of organic matter in living organisms in all terrestrial ecosystems.[1] The exact ratio between living and dead organic matter in terrestrial ecosystems varies, depending on the ecosystem. The amount of carbon stored in soil is determined by the balance of two biotic processes—the productivity of terrestrial vegetation and the decomposition of organic matter. Each of these processes has strong physical and biological controlling factors. These include climate; soil chemical, physical, and biological properties; and vegetation composition. Interactions among these controlling factors are of particular importance. These biological and physical factors are the same as the ones that influence the above ground structure and composition of terrestrial ecosystems, so there are strong correspondences between soil organic matter content and ecosystem type.

Organic Matter Inputs

Quantity

The amount of carbon stored in soils is to a great extent determined by the rate of organic matter input through litterfall, root exudates, and root turnover. The main factors that influence vegetation production are suitable temperatures for photosynthesis, available soil moisture for evapotranspiration, and rates of CO_2 and H_2O exchange. Dry and/or cold climates support low vegetation production rates

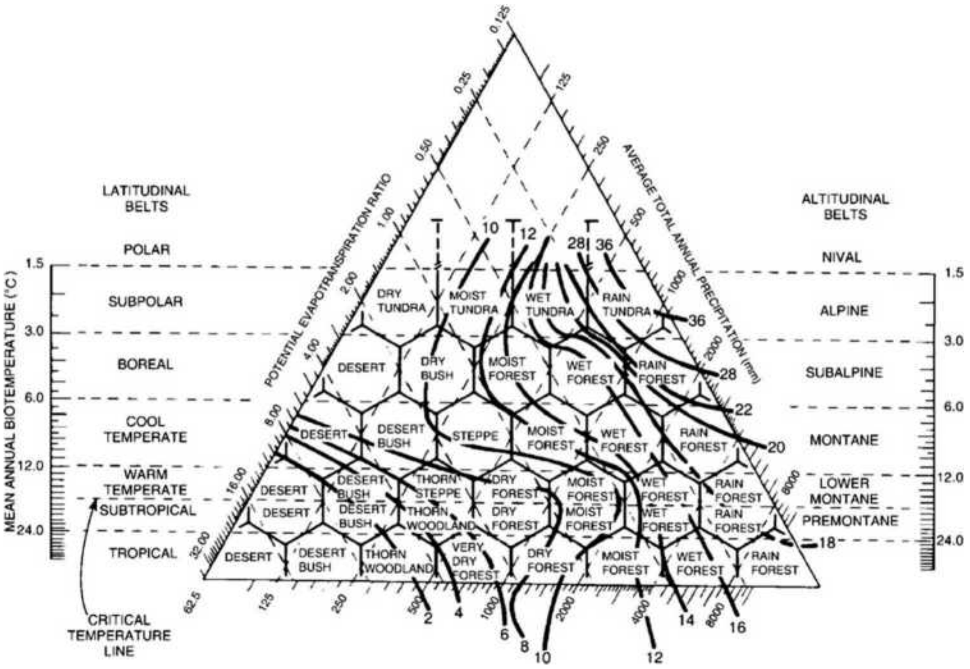


FIGURE 1 Contours of soil carbon density (kg m^{-2}) plotted on Holdridge diagram[3] for world life-zone classification. Values of biotemperature and precipitation uniquely determine a life zone and associated vegetation. Contour lines for mean soil carbon content in the surface meter of soil are determined from data derived from over 3000 soil profiles.[4,5]

and soils under such climates have low organic matter contents. Where climates are warm and moist, vegetation production is high and soil organic matter contents are correspondingly high. Figure 1 shows the striking correspondence between soil organic matter content and general climate measurements that results from the relationship between vegetation production and suitable moisture and temperature conditions.

Vegetation production depends not only on climate but also on nutrient supply from decomposition and geochemical weathering. Walker and Adams[6] hypothesized that the level of available phosphorus during the course of soil development is the primary determinant of terrestrial net primary production. Numerous workers have examined this hypothesis. Tiessen, Stewart, and Cole[7] and Roberts, Stewart, and Bettany[8] found that available phosphorus explained about one-fourth of the variance in soil organic matter in many different soil orders. The relationship between phosphorus and carbon is strongest during the aggrading stage of vegetation–soil system development.[9] Initially, the production of acidic products by pioneer vegetation promotes the release of phosphorus by weathering of parent material. Organic matter builds up in the soil, increasing the storage of phosphorus in decomposing organic compounds. Nitrogen fixing bacteria populations, which depend on a supply of organic carbon and available phosphorus, can grow to meet ecosystem demands for nitrogen. Plant growth is enhanced by this increasing nitrogen and phosphorus cycling, resulting in increased rates of weathering. This process continues until the vegetation is constrained by other factors affecting phosphorus availability: Leaching losses become larger than the weathering inputs;[10] or an increasing fraction of the phosphorus becomes unavailable by adsorption or precipitation with secondary minerals;[11] or nitrogen availability (denitrification or leaching is affected) reaching or exceeding nitrogen inputs and fixation.[12] In mature soils, net primary production is more likely to be limited by nitrogen. Availability of other nutrients that are largely derived from parent materials, such as most base cations, may also influence soil organic matter accumulation during early soil development.[13] Soils derived from base cation rich volcanic parent materials (Andisols) have much higher carbon contents on average than soils from other parent materials.[4]

Species Composition

Biotic factors, in particular plant species composition, also affect soil organic matter dynamics. Production and decomposition rates are to some degree controlled by species composition. Each terrestrial plant species produces different amounts and chemical compositions of leaves, roots, branches, and wood of varying decomposability. This range of decomposability may be summarized by the lignin and nitrogen content of the organic material.[14,15] Litter decay rate is inversely related to C : N and lignin: N ratios and positively related to N content. Species with tissues that have low nutrient or high lignin content produce litter that is slow to decay. Nitrogen is made available to plants during the decomposition process. Nitrogen is a limiting element for productivity in most terrestrial ecosystems so the rate at which it is released during decomposition is an important factor in ecosystem production. Thus, the interactions between processes regulating plant populations and their productivity and microbial processes regulating nitrogen availability result in some of the observed variation in soil carbon and nitrogen storage.[16–19]

Placement

The deeper that fresh detritus is placed in the soil, the slower it decomposes. This is a result of declining decomposer activity and increased protection from oxidation with depth in the soil. Prairies have a somewhat lower productivity than forests and produce no slowly decomposing woody material. Nevertheless, prairies have a very high soil organic matter content because prairie grasses allocate twice as much production to belowground roots and tillers than to aboveground leaves.[20] The result is high soil organic matter contents with a uniform distribution in the upper 1 m of soil (Figure 2). In contrast, a spruce-fir forest contains 50 percent of its soil organic matter in the top 10 cm. There are interesting exceptions to the rule that above-/belowground plant allocation determines soil organic matter distribution patterns in soil. Tropical moist forest soils show a uniform depth distribution similar to the depth distribution of temperate grasslands, however, in tropical forests this is largely due to a long-term accumulation of recalcitrant organic materials at lower depths in the soil rather than increased allocation to roots. Alpine tundra soils support a largely herbaceous flora but show a similar depth distribution as forest soils because of inhibition of surface litter decomposition by low temperatures and high water saturation.

Decomposition

Climate

Organic matter decay rates can be related to environmental parameters such as temperature and soil moisture. Climatic indices that correlate well with decay rates include plant moisture and temperature indices,[21,22] linear combinations of temperature and rainfall,[23] and actual evapotranspiration.[15] Warm temperatures and available soil moisture enhance microbial, and micro-and macro-invertebrate activity. These environmental conditions are also correlated with plant production. As a result, the amount of organic matter present in soil is highest in vegetation types with the highest rates of organic matter production. These are ones found in the warm, moist climate regions. The contours of soil carbon density displayed in Figure 1 reflect the balance of input by vegetation production and loss from decomposition imposed by climate. Soil carbon content increases from lower left to the upper right in Figure 1 as the temperature decreases in the cool temperate, boreal, and sub-polar life zones and as precipitation increases in the warm temperate, subtropical, and tropical life zones.

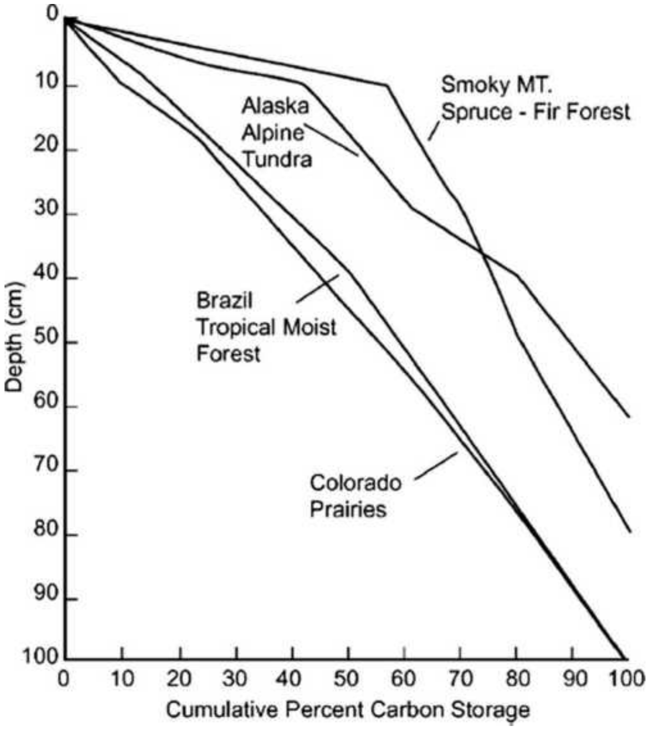


FIGURE 2 Cumulative carbon storage as a function of depth for four ecosystems. Refer text for explanation of these patterns.
Source: Zinke et al.[4]

The combined influence of temperature and precipitation is presented by the third axis of the Holdridge diagram (Figure 1) as the ratio of potential evapotranspiration (PET) to annual precipitation. When this ratio is less than 1.0, rainfall exceeds PET and vice versa. Life zones bordering the line with the PET is equal to precipitation (PET ratio = 1.0) have soil carbon contents around 10 kg m^{-2} except in warm temperate and subtropical zones where strong seasonality limits production, but decomposition conditions are favorable for most of the year. Soil carbon content increases as the PET ratio decreases indicating that productivity increases faster than the rate of decomposition with increasing moisture availability.

Organic Matter Quality

On global scale, climate may be the most important factor controlling decay rates, but within a given region, substrate chemistry is the more important factor.[15,24,25] Decay rate is often negatively related to substrate C : N ratio. Litter C : N is initially much greater than microbial C : N but approaches microbial C : N as the microbes release the carbon as CO2 while taking up nitrogen (nitrogen immobilization). The further the initial litter C : N is from microbial C : N, the slower the decay rate. Lignin content or lignin: N ratios may be better predictors of decay rates because lignin itself is difficult to decompose, and it shields nitrogen and other more easily degraded chemical fractions from microbes. Concise and simple models of decay rate are based on a combination of chemical and climatic indices.

The effect of litter quality on soil organic matter content is most dramatically expressed in Podzols (Spodosol in the United States Department of Agriculture classification). These occur over large areas in boreal zones dominated by evergreen conifers, but often occur in other regions on shallow or sandy soils. Low nitrogen content of organic matter inputs and cool temperatures reduce decomposition and soil animal activity. As a result, large surface organic matter accumulations occur over a thin A horizon. Low temperatures combined with leaching of organic acids result in podsolization as the predominant soil-forming process. Leaching of iron, aluminum oxides, and organic matter result in a distinct E horizon near the surface where these materials are removed and deposited in the B horizon. If the surface organic layers are included, these soils can have substantial organic matter contents, exceeding the expected amount for the climate conditions. Batjes[2] gives an average value for Podzols of 24.2 kg m^{-2} for the surface meter which is considerably above the mean for most other soil types (see Table 1).

Significant Physical and Chemical Influences

There are several notable exceptions to the climate-based explanation of variation in soil carbon content. There are two in particular that have lower rates of decomposition and therefore higher accumulations of organic matter than expected (Table 1). These include Histosols due to hydrological conditions and Andisols due to parent material chemical effects.

TABLE 1 Mean Organic Carbon Contents (kg m^{-2}) by FAO–UNESCO Soil Units to 1 m Depth

Soil Unit	Mean C (kg m^{-2})
Acrisols	9.4
Cambisols	9.6
Chernozems	12.5
Podzoluvisols	7.3
Ferrasols	10.7
Gleysols	13.1
Phaeozems	14.6
Fluvisols	9.3
Kastanozems	9.6
Luvisols	6.5
Greyzems	19.7
Nitosols	8.4
Histosols	77.6
Podzols	24.2
Arenosols	3.1
Regosols	5.0
Solonetz	6.2
Andisols	25.4
Vertisols	11.1
Planosols	7.7
Xerosols	4.8
Yermosols	3.0
Solochaks	4.2

These soil units generally span a wide range of climate conditions and therefore present a different view of soil organic matter content based on additional soil factors. In particular, the high C content of Podzols, Histosols, and Andisols is apparent. Refer text for additional explanation of biological, chemical and physical factors responsible.

Source: Batjes.[2]

Histosols

In landscape positions where water accumulates at or above the surface of the soil for an appreciable part of the growing season, decomposition can be reduced to such an extent that large amounts of undecomposed organic matter can accumulate. This soil type is called a Histosol and can be found in any region in wetlands where decomposition is restricted. The soil-surface of mature or old-growth boreal forests over shallow water tables are often covered with *Sphagnum* moss which may also lead to development of Histosols. Histosols with the largest areas and thickest accumulations occur in lowland tundra where a mixture of sedges, lichens, and mosses grow at the northern limit of vegetation in the northern hemisphere. Production, decomposition, and evaporation are limited by low temperatures and water-saturated soils. In these cold regions, deeper layers may freeze and not become thawed during the short growing season (permafrost). As a result, Histosols have carbon contents over 70 kg m^{-2} in the surface meter (Table 1). Some regions have been accumulating organic matter since the last glacial period without any substantial decomposition. Histosols in such regions may be several meters thick and contain over 250 kg C m^{-2}.[2] Globally it is estimated that boreal and sub-arctic Histosols contain 455 PgC that has accumulated during the postglacial period.[26]

Andisols

Andisols form on young volcanic stone (basalt lava) rich in nutrients and alkaline. Andisols are weakly weathered soils associated with pyroclastic parent materials that are rich in allophane, ferrihydrite, and other minerals that readily form complexes with humus molecules. These chemical constituents provide conditions promoting high vegetation production and also the retention of organic matter in soil. As a result, Andisols typically have higher soil carbon contents (25.4 kg m^{-2}, Table 1) than soils with the same environmental conditions but different parent materials.

Conclusions

Over long periods of time, organic matter in soils is the result of climatic, biological, and geological factors. These factors are not independent. In particular there exists a strong relationship between climate and vegetation type. In Figure 1, the Holdridge climate based life zones have names that depict the dominant vegetation of climates. Jobbágy and Jackson[27] provide a summary of soil data based on biomes that demonstrates similar soil carbon distribution as that based on climate (Table 2).

TABLE 2 Mean Organic Carbon Content (kg m^{-2}) by Biome to 1 m Depth

Biome	Mean C (kg m^{-2})
Boreal forest	9.3
Crops	11.2
Deserts	6.2
Sclerophyllous shrubs	8.9
Temperate deciduous forest	17.4
Temperate evergreen forest	14.5
Temperate grassland	11.7
Tropical deciduous forest	15.8
Tropical evergreen forest	18.6
Tropical grassland/savanna	13.2
Tundra	14.2

Note: Biome classification is based on Whittaker.[28]

Source: Jobbágy and Jackson.[27]

Over shorter periods of time soil carbon varies with vegetation disturbances and changes in land use patterns that affect rates of organic matter input and its decomposition. Various land uses result in very rapid declines in soil organic matter from the native condition.[29–32] Losses of 50% in the top 20 cm and 30% for the surface 100 cm are average. Much of this loss in soil organic carbon can be attributed to erosion, reduced inputs of organic matter, increased decomposability of crop residues, and tillage effects that decrease the amount of physical protection to decomposition. Evidence from long-term experiments suggest that C losses due to oxidation and erosion can be reversed with soil management practices that minimize soil disturbance and optimize plant yield through fertilization. These experimental results are believed to apply to large regions and that organic matter is being restored as a result of establishment of perennial vegetation, increased adoption of conservation tillage methods, efficient use of fertilizers, and increased use of high yielding crop varieties.[33,34] Additionally, when agricultural land is no longer used for cultivation and allowed to revert to natural vegetation or replanted to perennial vegetation, soil organic carbon can accumulate by processes that essentially reversing some of the effects responsible for soil organic carbon losses initially—from when the land was converted from perennial vegetation—and return them to typical amounts for the climate, vegetation, landscape position, and parent material conditions.[35,36]

Acknowledgments

Work sponsored by U.S. Department of Energy, Carbon Dioxide Research Program, Environmental Sciences Division, Office of Biological and Environmental Research and performed at Oak Ridge National Laboratory (ORNL). ORNL is managed by UT-Battelle, LLC, for the U.S. Department of Energy under contract DE-AC05-00OR22725.

References

1. Post, W.M.; Peng, T.-H.; Emanuel, W.R.; King, A.W.; Dale, H.; DeAngelis, D.L. The global carbon cycle. American Scientist 1990, *78,* 310–326.
2. Batjes, N.H. Total carbon and nitrogen in the soils of the world. European Journal of Soil Science 1996, *47,* 151–163.
3. Holdridge, L.R. Determination of world plant formations from simple climatic data. Science 1947, *105,* 367–368.
4. Zinke, P.J.; Stangenberger, A.G.; Post, W.M.; Emanuel, W.R.; Olson, J.S. *Worldwide Organic Soil Carbon and Nitrogen Data;* ORNL/TM-8857; Oak Ridge National Laboratory: Oak Ridge, TN, 1984.
5. Post, W.M.; Pastor, J.; Zinke, P.J.; Stangenberger, A.G. Global patterns of soil nitrogen storage. Nature 1985, *317,* 613–616.
6. Walker, T.W.; Adams, A.F.R. Studies on soil organic matter: I. Influence of phosphorus content of parent materials on accumulations of carbon, nitrogen, sulfur, and organic phosphorus in grassland soils. Soil Science 1958, *85,* 307–318.
7. Tiessen, H.J.; Stewart, W.B.; Cole, C.V. Pathways of phosphorus transformations in soils of differing pedogenesis. Soil Science Society of America Journal 1984, *48*, 853–858.
8. Roberts, T.L.; Stewart, J.W.B.; Bettany, J.R. The influence of topography on the distribution of organic and inorganic soil phosphorus across a narrow environmental gradient. Canadian Journal of Soil Science 1985, *65,* 651–665.
9. Anderson, D.W. The effect of parent material and soil development on nutrient cycling in temperate ecosystems. Biogeochemistry 1988, *5*, 71–97.
10. Jenny, H. *The Soil Resource;* Springer: Berlin, 1980.
11. Walker, T.W.; Syers, J.K. The fate of phosphorus during pedogenesis. Geoderma 1976, *15,* 1–19.
12. Schlesinger, W.H. *Biogeochemistry: An Analysis of Global Change;* Academic: New York, 1991.

13. Torn, M.S.; Trumbore, S.E.; Chadwick, O.A.; Vitousek, P.M.; Hendricks, D.M. Mineral control of soil organic carbon storage and turnover. Nature 1997, *389,* 170–173.
14. Aber, J.D.; Melillo, J.M. Nitrogen immobilization in decaying hardwood leaf litter as a function of initial nitrogen and lignin content. Canadian Journal of Botany 1982, *58*, 416–421.
15. Meentemeyer, V. Macroclimate and lignin control oflitter decomposition rates. Ecology 1978, *59*, 465–472.
16. Zinke, P.J. The pattern of influence of individual trees on soil properties. Ecology 1962, *42,* 130–133.
17. Wedin, D.A.; Tilman, D. Species effects on nitrogen cycling: a test with perennial grasses. Oecologia 1990, *84,* 433–441.
18. Hobbie, S.E. Effects of plant species on nutrient cycling. Trends in Ecology and Evolution 1992, 7, 336–339.
19. Hobbie, S.E. Temperature and plant species control over litter decomposition in Alaskan tundra. Ecological Monographs 1996, *66*, 503–522.
20. Sims, P.L.; Coupland, R.T. *Grassland Ecosystems of the World: Analysis of Grasslands and Their Uses*; Coupland, R.T., Ed.; Cambridge University Press: Cambridge, 1979.
21. Olson, J.S. Energy storage and the balance of producers and decomposers in ecological systems. Ecology 1963, *44*, 322–331.
22. Fogel, R.; Cromack, K. Effect of habitat and substrate quality on douglas fir litter decomposition in western Oregon. Canadian Journal of Botany 1977, *55*, 1632–1640.
23. Pandey, V.; Singh, J.S. Leaf-litter decomposition in an oak– conifer forest in himalaya: the effects of climate and chemical composition. Forestry 1982, *55*, 47–59.
24. Flanagan, P.W.; VanCleve, K. Nutrient cycling in relation to decomposition and organic matter quality in tiaga ecosystems. Canadian Journal of Forest Research 1983, *13,* 795–817.
25. McClaugherty, C.A.; Pastor, J.; Aber, J.D.; Melillo, J.M. Forest litter decomposition in relation to soil nitrogen dynamics and litter quality. Ecology 1984, *66*, 266–275.
26. Gorham, E. Northern peatlands: role in the carbon cycle and probable responses to climatic warming. Ecological Applications 1991, *1*, 182–195.
27. Jobbágy, E.G.; Jackson, R.B. The vertical distribution of organic carbon and its relation to climate and vegetation. Ecological Applications 2000, *10* (2), 423–436.
28. Whittaker, R.H. *Communities and Ecosystems;* MacMillan: London, 1975.
29. Jenny, H. *Factors of Soil Formation;* McGraw-Hill: New York, 1941.
30. Davidson, E.A.; Ackerman, I.L. Changes in soil carbon inventories following cultivation of previously untilled soils. Biogeochemistry 1993, *20,* 161–193.
31. Mann, L.K. Changes in soil carbon after cultivation. Soil Science 1986, *142,* 279–288.
32. Schlesinger, W.H. Changes in soil carbon storage and associated properties with disturbance and recovery. In *The Changing Carbon Cycle: A Global Analysis*; Trabalka, J.R., Reichle, D.E., Eds.; Springer: New York, 1985.
33. Buyanovsky, G.A.; Wagner, G.H. Carbon cycling in cultivated land and its global significance. Global Change Biology 1998, *4*, 131–142.
34. Lal, R.; Kimble, J.M.; Follett, R.F.; Cole, C.V. *The Potential of U.S. Cropland to Sequester Carbon and Mitigate the Greenhouse Effect;* Ann Arbor Press: Ann Arbor, MI, 1998.
35. Post, W.M.; Kwon, K.C. Soil carbon sequestration and land-use change: processes and potential. Global Change Biology 2000, *6,* 317–328.
36. Silver, W.L.; Ostertag, R.; Lugo, A.E. The potential for carbon sequestration through reforestation of abandoned tropical agricultural and pasture lands. Restoration Ecology 2000, *8* (4), 394–407.

65

Permafrost

Douglas Kane
and Julia Boike

Introduction

On an extended south-to-north transect at northern latitudes, a transition from ground that never experiences seasonal freezing, to those that occasionally freeze during the winter, to those that freeze every year, to those that may remain frozen for an extended time could be encountered. The topic of discussion here is those surface soils and deeper geologic layers that remain at or below freezing for a duration of two years or more and how they impact people living in this environment. Such frozen ground, both unconsolidated and bedrock, are commonly referred to as permafrost.[1] Although at or below the freezing point of bulk water (0°C), the term "permafrost" neither implies that water is present or that water, if present, is frozen. In fact, it is possible for significant amounts of water to remain unfrozen in permafrost; this is also true for water in seasonal frost.

Permafrost Characteristics

Spatially extensive permafrost can be found in Russia and Canada as far south as 45°N, and even farther south on the elevated Tibetan Plateau and Himalayan Mountains.[2] Approximately 25 million km^2 of permafrost exist in the northern hemisphere. In the higher latitudes, permafrost is continuous under the land surfaces. At intermediate latitudes permafrost is discontinuous or sporadic. Legget[3] reported that 20% of the land surface of the world is underlain by permafrost. More than 50% of Russia and Canada are underlain by permafrost. Alaska has continuous permafrost in the northern 1/3 of the state and discontinuous permafrost in the rest of the state, excluding the coastal areas from the Aleutian Islands to southeastern Alaska (Figure 1). In the southern hemisphere, permafrost distribution is confined to Antarctica and high alpine or mountainous regions. Isolated permafrost is common at higher elevations, and evidence of past permafrost is common in areas that no longer have permafrost. The thickness of permafrost can vary from a thin lens of less than 1 m to greater than 1000 m (Figure 1). Permafrost can also be found in coastal areas at the bottom of seas.

Permafrost ground is interesting to the engineer and scientist because the medium is usually composed of two solids (porous medium and ice) and two fluids (air and liquid water). The more components that are present in a mixture, the more difficult it is to predict the medium's response to an input of energy or mass. The amount of unfrozen water in saturated ground is strictly a function of the grain

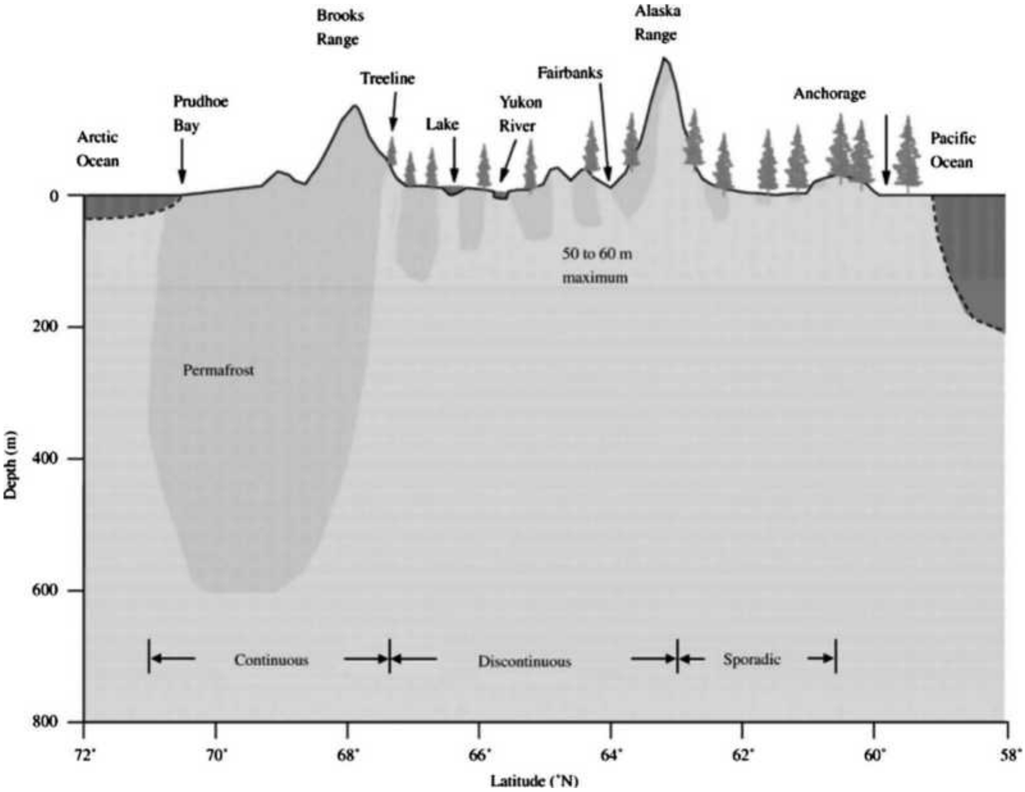


FIGURE 1 Schematic of permafrost distribution in Alaska on a south-to-north transect.

size (more specifically surface area) and the freezing temperature. A fine-grained soil such as clay has a very high surface area relative to coarse-grained soils such as sand; in frozen ground, this translates into much higher unfrozen water contents at the same temperature. The unfrozen water found in permafrost exists as a film of water around each soil particle. It is via these unfrozen films that water moves in permafrost. Frozen clay can have as much as 5%–7% unfrozen water by volume at –15°C. As the temperature decreases the amount of unfrozen water also decreases. Thermal and hydraulic properties of permafrost are quite variable and depend upon the percentages of the various ground components. Most heat transfer in permafrost is by conduction and can be modeled by Fourier's law, although simpler methods have been developed.

A typical temperature profile of permafrost appears in F2. Since there is a geothermal flux outward from the center of the earth, this heat has to be successfully transferred to the ground surface or the permafrost will warm and melt. In order to maintain the thermal integrity of the permafrost, the soils above the top of the permafrost table must completely freeze during the winter so there is a continuously decreasing thermal gradient along which the geothermal heat can be transferred to the surface by conduction (Figure 2).

The seasonally thawed soil layer at the ground surface that goes through freezing/thawing annually and mantles the permafrost is called the active layer (Figure 2). This layer acts as a buffer to heat and mass transfer to the permafrost. A typical active layer in the continuous permafrost zone would typically thaw to a maximum depth of 60 cm. The top 15–25 cm of these soils are generally composed of organic material, with the deeper soils being mineral. Organic soils are good thermal insulators and, when coupled with snow cover during the winter months, they minimize heat loss from the ground. The thermal balance of the permanently frozen ground is maintained by heat loss to the atmosphere that occurs at high latitudes over the extended winter months.

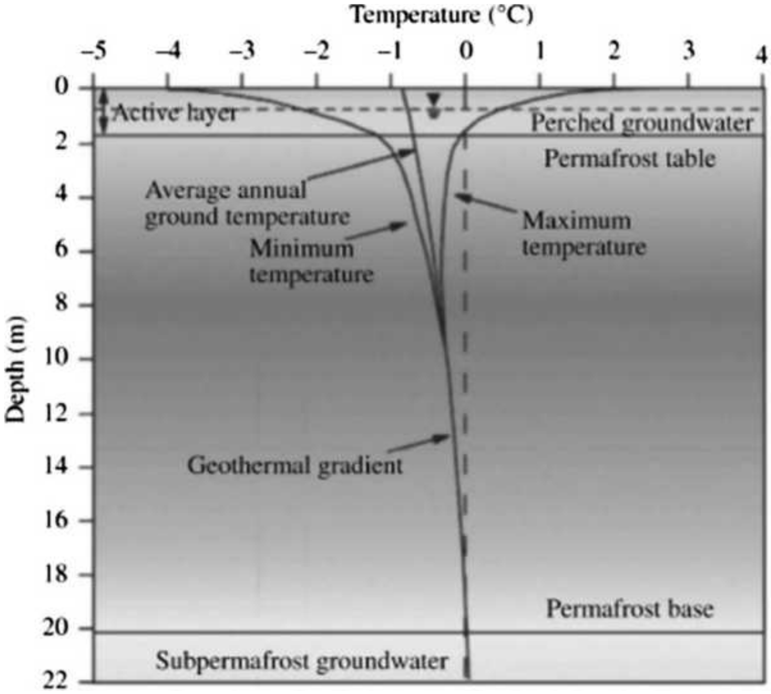


FIGURE 2 Typical temperature profile for permafrost, with annual variation indicated near the ground surface.

Permafrost is usually considered to be impermeable to water movement. This is usually a good assumption for short durations of days or even a few months. For longer periods of time, the redistribution of water within permafrost can be appreciable. During snowmelt and major rainfall events, considerable water enters the active layer. Most of this water resides in the organic soils as the mineral soils are usually already near saturation. Since permafrost has relatively low hydraulic conductivity, there is no hydraulic connection between the perched water above the permafrost (suprapermafrost groundwater) and the subpermafrost groundwater below the permafrost (Figure 2). In continuous permafrost, the subsurface hydrology is confined to the active layer. For areas of discontinuous permafrost, the subsurface hydrology is a combination of shallow flow over the frozen ground and deeper flow around and under it.

Surface Energy Balance

Any time the ground surface is disturbed, the surface energy balance that sustains permafrost is upset. This generally results in warming of the permafrost and thickening of the active layer. Much sporadic and discontinuous permafrost is maintained at temperatures just below freezing; climatic warming of just a few degrees would result in the melting of this frozen ground and warming of colder permafrost. Also where permafrost exists, surface disturbances such as removing vegetation or surface soils and ponding of water are sufficient to alter the surface energy balance and cause permafrost degradation. Permafrost generally appears where the mean annual surface temperature is at least a few degrees Celsius below freezing; for instance, Hay River, NWT, Canada has sporadic, shallow permafrost with a mean annual temperature of –3.4°C.

In areas of discontinuous permafrost, the south-facing slopes are generally permafrost-free (Figure 1) while north-facing slopes have permafrost; the east- and west-facing slopes and valley bottoms may or may not have permafrost, depending upon subtle site-specific conditions that impact the energy balance (vegetation, slope, moisture content, etc.). Because it took thousands of years for permafrost to develop, deep permafrost is a particularly good recorder of past climates, and changes in the temperature profile will reflect these impacts. Many general atmospheric circulation models (GCM) predict variable global warming in the high latitudes, and in some areas there is already field evidence of atmospheric warming[4] and permafrost warming.[5] To sustain permafrost, the geothermal heat radiating out from the core of the earth must be removed; this only occurs if the active layer is completely frozen during winter.

Living with Permafrost

Permafrost influences many facets of everyday living that are often taken for granted in warmer climates. Permafrost negatively impacts both the quality and quantity of the available groundwater and limits the potential for infiltration of wastewater. As a result, housing units with individual water and wastewater systems must be designed with innovative options to replace the traditional groundwater well and septic system for wastewater treatment. Utilities that are typically buried are often placed in aboveground utilidors. Also as roadways age, they become very bumpy as differential movement of the road surface occurs.

The significant property of permafrost from an engineering viewpoint is the ice content. The amount of ice can range from essentially none in well-drained ground to values that exceed the porosity of poorly drained ground. Within the permafrost, segregated ice with minimal entrained soil exists as ice lenses and wedges. As long as the permafrost stays frozen, the ground is relatively stable. However, thawing of ice-rich permafrost causes ground to settle or subside. This thaw settlement compromises the structural integrity of buildings, roadways, pipelines, and other structures built on it. For areas where construction must take place on frozen ground, special efforts—for example the use of insulation and thermosyphons—are necessary to maintain the structural integrity of the permafrost. Thermosyphons are vertical devices for removing heat from the ground, and they are used extensively under buildings and pipelines to maintain below freezing conditions in the ground.

Soil freezing causes additional engineering problems near the ground surface when strong thermal gradients induce water movement from warm ground to cold ground through unfrozen films of water on the surfaces of soil particles, regardless of whether it is frozen or not. This process causes "frost heave," as it results in an accumulation of ice in the form of lenses that can cause the ground to expand and deform. Surface heaving becomes significant over time periods of several years, and it is important for those engineering structures with design lives of tens of years to compensate for this additional stress.

Conclusions

Permafrost and seasonally frozen ground are very extensive in the high latitudes of Earth. Also, resource development and population in these areas is increasing. Initial efforts directed at building an infrastructure (roads, airports, water and sewer distribution, etc.), on permafrost assumed that the frozen ground was in thermal equilibrium. It is now obvious that because of climatic change that the permafrost is not in thermal equilibrium. Instead, it is warming in many areas and this needs to be considered in the present design criteria for all engineering structures built on permafrost. Warming, with subsequent thawing of permafrost, will affect the surface energy budget, water and gas fluxes (potential for release of greenhouse gases), and vegetation; hence, there is a direct feedback with the climate system. Just as it took a long time for permafrost to develop, it will take a long time for it to completely melt; but the consequences of thawing permafrost can impact our daily lives in the first or second season of melting.

References

1. Muller, S.W. *Permafrost or Permanently Frozen Ground and Related Engineering Problems.* U.S. Engineers Office, Strategic Engineering Study, Special Report No. 62, 1943; 136 pp.
2. Brown, J.; Ferrians, O.J., Jr.; Heginbottom, J.A.; Melnikov, E.S. *Circum-Arctic Map of Permafrost and Ground-Ice Conditions*; U.S. Dept. of Interior, Geological Survey, Map CP-45, 1997.
3. Legget, R.F. Permafrost research. Arctic **1954**, *7* (3/4), 153–158.
4. Serreze, M.C.; Walsh, J.E.; Chapin, F.S.; Osterkamp, T.; Dyurgerov, M.; Romanovsky, V.; Oechel, W.C.; Morison, J.; Zhang, T.; Berry, R.G. Observational evidence of recent change in the northern high-latitude environment. Climate Change **2000**, *46,* 159–207.
5. Lachenbruch, A.H.; Marshall, B.V. Changing climate: geothermal evidence from permafrost in the Alaskan Arctic. Science **1986**, *234,* 689–696.

66

Salt-Affected Soils: Plant Response

Anna Eynard,
Keith D. Wiebe,
and Rattan Lal

Introduction

Salt-affected soils can be defined as soils on which the growth of most crop plants is limited by an excess of easily soluble salts. Salts are considered easily soluble when they are more soluble in water than gypsum. Salts may include chlorides, sulfates, carbonates and bicarbonates of sodium, potassium, magnesium, and calcium. The salt concentration in the soil solution is usually measured by the electrical conductivity of the soil saturation extract (EC). A soil is saline if the EC of the extract is higher than 4 dS/m.[1] However, plant growth may be severely affected or inhibited at much lower levels of EC. Limiting EC values change with the ionic composition of the soil solution and with the plant species and variety grown. Sodic soils are salt-affected at EC <4 dS/m (Table 1). In sodic soils, the high sodium content relative to other cations is the main factor affecting plant growth. The amount of sodium is usually expressed as exchangeable sodium percentage (ESP) or sodium adsorption ratio (SAR). The critical level of ESP for growth reduction depends on many interacting factors and may vary from <6% to >15%. There is a continuum of plant responses to the continuous increase of salt stress with increasing salt or sodium concentrations in the soil. Saline and sodic conditions, although very different, are often both implicated in adversely affecting plant performance so that terms such as salinity, saline, and salinization are generically referred to soil conditions and processes leading to salt and sodium stresses on plants.[2]

TABLE 1 General Classification of Salt-Affected Soils

Soil Class	EC (dS/m)	SAR	ESP (%)
Nonsaline nonsodic	<4	<13	<15
Saline	>4	<13	<15
Saline–sodic	>4	>13	>15
Sodic	<4	>13	>15

Extent and Distribution

Salt-affected soils are most common in arid climates with evapotranspiration greater than precipitation for at least part of the year (aridic, ustic, or xeric soil moisture regimes), but they are present at any latitude and altitude. Salinization by improper irrigation management was a major cause of degradation of the Nile Delta, the Mesopotamian Plain, and the valley of the Yangtze and the Hwang Ho.[3] Estimates of the extent of soil salinity and sodicity problems vary from 5% to 10% of the world land area. Recent estimates of FAO-AGL[4] indicate that >800 million ha are salt-affected, i.e., >6% of the world land area (Figure 1). Salinization is a serious hazard in San Joaquin Valley, United States; Murray-Darling Basin, Australia; Euphrates Basin, Pakistan; Indo-Gangetic Basin, India; and Aral Sea Basin, Central Asia.[5] Global land area degraded by salt-affected soils may be 1 billion hectare (Bha),[6,7] and the world may be losing as much as 2000 ha/day.[5] Worldwide 20% of the cultivated land area and 33% of the irrigated lands are degraded because of salt-induced problems.[8,9] It is also estimated that salinized area may be increasing at the rate of 10% annually. Fifty percent of arable land may be salinized by 2050.[8] Crop growth and agronomic productivity in salt-affected soils is adversely affected[10] because of osmotic and drought stress,[11] degradation of soil structure especially in sodic soils, and alteration in soil enzymatic activity.[12] Salinity problems also affect greenhouse crops, mine spoils, and disposal areas. Part of the yield loss occurs on irrigated lands, which are otherwise highly productive. In the irrigation district of Ciudad Juarez Valley in Mexico, the productivity of 70% of the lands is hindered by salinity. On average, 25% of irrigated Mexican lands are salt-affected. The impact of yield losses is magnified at a local scale because salt-affected soils are not uniformly distributed, threatening the economic welfare of some regions and countries. In Bangladesh, 24% of the total land area is salt-affected with a rapid expansion

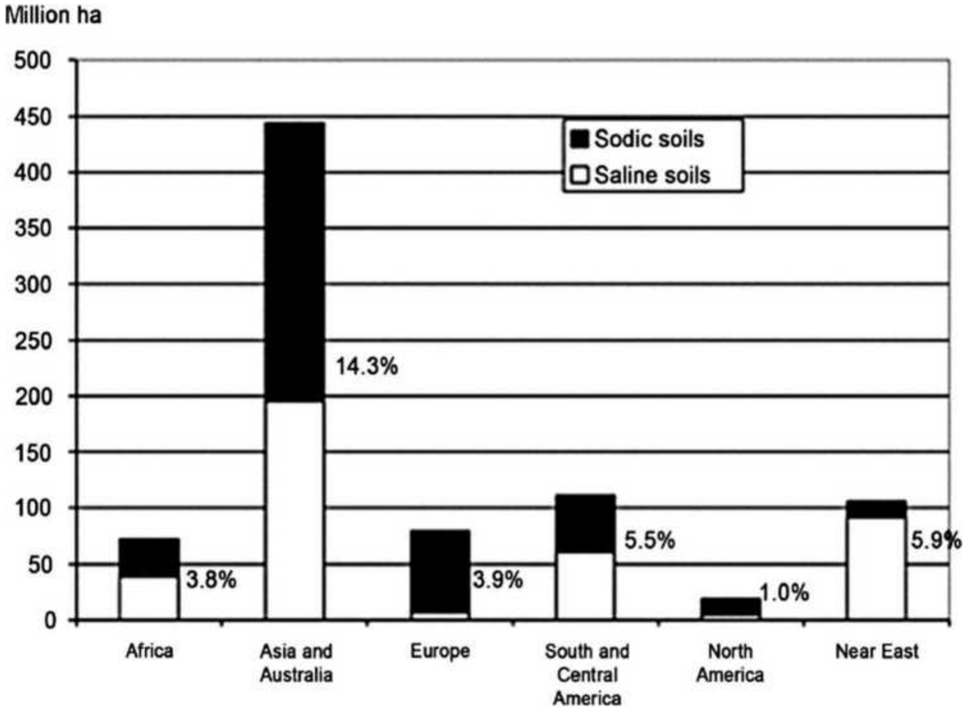


FIGURE 1 Distribution of salt-affected soils. The percentage of salt-affected soils on the total land surface is reported for each region of the world.
Source: FAO-AGL.[4]

during the last quarter of the 20th century from <1 million ha to >3 million ha. In Australia, almost 20% of the land is salt-affected.[4]

Salt-affected soils not currently used for crop production (e.g., salt flats) are a potential source of salts for salinization of surrounding fields. According to Oldeman, Hakkeling, and Sombroek,[13] the annual rate of loss of agricultural land by salinization, alkalinization, and waterlogging is about 1.5 million ha. Previous estimates reported 10 million ha of agricultural land loss per year.[14] The rate of increase is likely to vary widely across regions and from year to year, but all data suggest an increasing extent of salt-affected lands, especially where the natural resources are scarce and the growing human population increases pressure on soils. Changes in the global environment are likely to increase yield losses consequent to soil salinity. The predicted increase of the sea level from thermal expansion of seawater ranges from 15 to >50 cm by the year 2100.[15] The rise of seawater is likely to worsen salinity problems from tidal inundation of coastal lands. On the other hand, there is no evidence to suggest that elevated atmospheric CO_2 would increase the level of salinity suitable for plant growth.[16] Increased CO_2 is expected to increase plant growth only on nonsaline soils, magnifying yield losses due to soil salinity.

Crop Tolerance

Soil salinity and sodicity limit the potential area of growth of sensitive crops. All plants are sensitive to salts at some concentration, but the limiting concentration varies with plant species and variety. Crop tolerance is defined in relation to the level of salinity causing yield losses. High salt concentration may lead to plant death and no yield. Decreased growth occurs above a soil critical salt or sodium concentration (tolerance threshold). Crop yield reductions in salt-affected soils result from alteration of various metabolic processes in plants under salt stress. Negative effects of excess of salts in the soil solution include increased osmotic pressure limiting water uptake (physiological drought), abnormal pH and ionic competition limiting nutrient uptake, and ionic toxicity. Total salinity effects always depend on the specific soil ionic composition. The negative response of plants to low water potential may prevail in saline soils, while single ion toxicity or nutritional imbalance may be particularly severe in sodic soils. Soil structural impedance of plant growth is common in the absence of sufficient Ca^{2+}. Structural problems derive from crusting, formation of compacted layers, poor infiltration, and poor permeability to water and air. In waterlogged soils, anaerobic conditions may enhance salinity stresses.

In all plants growing on saline media, a part of the metabolic energy is diverted to ion transport, synthesis of organic osmolytes (contributing to osmotic adjustment to low water potential), and ion compartmentation preventing ion toxicity.[2] Osmotic adjustments in salt-tolerant plants need to follow seasonal changes in soil salinity and water availability. In general, plants that are more salt-tolerant tend to grow more slowly at low salinity levels than less-tolerant plants. Broadly adaptable plants can produce good yields where strong temporal changes of soil salinity occur in the soil. But such plants, tolerant to a wide range of salinities, perform less well at any salinity than less adaptable plants at their optimal salinity.[8] Reduced yield in tolerant plants can be related to (1) greater allocation of organic C in the roots of tolerant plants at the expense of leaf area, (2) decreased use of the solar radiation, and (3) low transpiration rate. Soil salinity interacts with climate and other edaphic factors in determining yield. Reduced availability of water in the soil is concomitant to higher evaporative demand in the leaves. Adaptation to soil salinity is a reduction in the capacity for photosynthetic carbon assimilation consequent to the necessity of minimizing evapotranspiration. In general, crops are less sensitive to salinity in glasshouse conditions than outdoors. At low soil fertility levels, plants may appear more tolerant to salinity than at high fertility levels if salinity is not the limiting factor of yield. On the other hand, nutrient deficiencies induced by salinity and sodicity can further reduce low yields due to low fertility.

In natural conditions, soil salinity varies with depth and time. Plants extract water from the least saline parts of the soil surrounding the roots. Plants may be able to tolerate high salinity levels in the root zone if occurring only for short periods of time during the growing season. Soil salinity may peak

TABLE 2 Classification of Crop Salt Tolerance Based on the Values of EC above Which Yield Loss Begins

Relative Tolerance	Critical EC (dS/m)	Examples of Crops
Sensitive	<1.3	Beans
Moderately sensitive	1.3–3.0	Rice, corn
Moderately tolerant	3.0–6.0	Wheat
Tolerant	6.0–10.0	Barley, cotton
Unsuitable	>10.0	Most crops

Source: Ayers and Wescot.[17]

TABLE 3 Classification of Crop Exchangeable Sodium Tolerance. Growth Is Usually Already Affected at Lower Values of ESP and Crop Production Is Usually Excluded at ESP>30

Relative Tolerance	ESP (%)	Examples of Crops
Sensitive	<15	Beans, corn
Semitolerant	15–40	Rice, wheat
Tolerant	>40	Barley, cotton

Source: Ayers and Wescot.[17]

at different stages of the crop, when plants may be more or less sensitive. Consequently, the critical point for yield reductions varies with duration of the salt stress and stage of development. Growth can be inhibited at any stage of the biological cycle. Severe reductions of yields can be due to low germination and limited early plant establishment. Yield reductions can be caused by reduced vegetative growth and/or by perturbation of the reproductive phases. Developmental shifts of relative salt tolerance are common during plant life and vary with cultivar and environment. Often, qualitative yield reductions enhance total yield losses.

Irrigated drylands are at a high risk of salinization,[18] and there are a wide range of reclamation measures.[19,20] A common classification of crop salinity tolerance distinguishes five categories (Table 2). Sodium is not considered a limiting factor in the above classification.

A classification of sodium tolerance of crops consists of three groups (Table 3). Sodium-tolerance tables are usually based on nutritional responses in absence of soil structural degradation, which generally excludes crop production at ESP >30.[9] These classifications are based on the agronomic criterion of comparing the relative yield of each crop on saline media to its yield on normal nonsaline media. Published tolerance tables provide a reference of relative salt tolerance of crops under irrigation without local climatic specifications. Such data refer to duration of salinity from late seedling stage to maturation with irrigation, fertilization, and pest control maintained optimal for each crop species and variety at the time of the experiment. Tolerance tables may apply to individual cultivars in the absence of interactions between soil salinity and other environmental factors.

Conclusions

Salt-affected soils decrease crop productivity at local and global scales. The extent of soil salt problems is likely to increase if no measures are taken to limit soil salinization and sodication. Management and reclamation of salt-affected soils require a combination of agronomic practices, including hydraulic, mechanical, amending, and cropping practices.[4] The selection of proper management practices is relative to the soil, water, crop, and human available resources.

Improved crop salt tolerance can contribute to the remediation of soil salt problems because:

- Extends the choice of crops that can grow at each soil salinity level
- Allows irrigation using more saline water
- Increases soil organic matter and improves soil structure.

Consequently, tolerant crops may assist in establishing a proper water balance and reduce the need and cost of reclamation measures such as leaching.

Information on crop response to salinity and sodicity needs to be updated in relation to changes of plant materials, agronomic practices, and climatic conditions. A range of values in the continuum of salt and sodium stresses would be better suited than a single critical value to represent the tolerance or sensitivity of a plant since the intensity of salt stress depends on many other factors determining yields.

List of Abbreviations

EC	electrical conductivity of the soil saturation extract in dS/m
ESP	exchangeable sodium percentage in %
SAR	sodium adsorption ratio = $Na^+/([Ca^{2+} + Mg^{2+}]/2)^{1/2}$, where concentrations are in mmol/L

References

1. SSSA. *Glossary of Soil Science Terms*. Soil Science Society of America: Madison, WI, 1997; 134 pp.
2. Läuchli, A., Epstein, E. Plant responses to saline and sodic conditions. In *Agricultural Salinity Assessment and Management*, Tanji, K.K., Ed.; ASCE Manuals and Reports No. 71; ASCE: New York, 1990; 113–137.
3. Szabolcs, I. Salt buildup as a factor of soil degradation. In Methods for Assessment of Soil Degradation, Lal, R., Blum, W.H., Valentine, C., Stewart, B.A., Eds.; Boca Raton, FL: CRC Press, 1998; 253–264.
4. FAO-AGL. *Global Network on Integrated Soil Management for Sustainable Use of Salt-Affected Soils*. FAO-AGL, Land and Plant Nutrition Management Service, 2000; http://www.fao.org/ag/AGL/ag11/spush.
5. Collins, T. *World Losing 2000 Hectare of Fair Soil Daily to Salt-Induced Degradation*. INWEH-UNU, Hamilton, Canada, 2014.
6. Yensen, N.P., Halophyte uses for the twenty-first century. In Ecophysiology of High Salinity Tolerant Plants, Khan, M.A., Weber, D.J., Eds.; Dordrecht: Springer, 2008; 367–396.
7. Metternicht, G.I., Zinck, J.A. Remote sensing of soil salinity: potentials and constraints. *Remote Sens. Environ*. 2003; 85: 1–20.
8. Patel, B.B., Patel Bharat, B., Dave, R.S. Studies on infiltration of saline-alkali soils of several parts of Mehsana and Patan districts of north Gujarat. *J. Appl. Technol. Environ. Sanitation*. 2011; 1(1): 87–92.
9. Jamil, A., Riaz, S., Ashraf, M., Foolad, M.R. Gene expression profiling of plants under salt stress. *Crit. Rev. Plant Sci*. 2011; 30(5): 435–458.
10. Paul, D. Osmotic stress adaptions in rhizobacteria. *J. Basic Microbiol*. 2012; 52: 1–10.
11. Munns, R. Genes and salt tolerance: bringing them together. *New Phytol*. 2005; 167: 645–663.
12. Seckin, B., Sekmen, A.H., Turkan, I. An enhancing effect of exogenous mannitol on the antioxidant enzyme activities in roots of what under salt stress. *J. Plant Growth Regul*. 2009; 28: 12–20.
13. Oldeman, L.R., Hakkeling, R.T.A., Sombroek, W.G. *Second Revised Edition. World Map of the Status of Human-Induced Soil Degradation. An Explanatory Note*. Wageningen, NL: International Soil Reference and Information Center, 1991; 35 pp.
14. WCED. *Our Common Future*. Oxford: Oxford University Press, 1987; 383 pp.
15. Warrick, R.A., Le Prevost, C., Meier, M.F., Oerlemans, J., Woodworth, P.L. Changes in sea level. In Climate Change 1995, Houghton, J.T., Meira Filho, L.G., Callander, B.A., Harris, N., Kattenberg, A., Maskell, K., Eds.; Cambridge: Cambridge University Press, 1996; 363–405.

16. Ball, M.C., Sobrado, M.A. Ecophysiology of mangroves: challenges in linking physiological processes with patterns in forest structure. *Physiological Plant Ecology: The 39th Symposium of the British Ecological Society*, Held at the University of York, 7–9 September 1998. Scholes, J.D., Barker, M.G., Press, M.C., Eds.; Oxford: Blackwell Science, 1999; 331–346.
17. Ayers, R.S., Wescot, D.W. Water Quality for Agriculture; FAO Irrigation and Drainage Paper 29, Rev. 1; Rome: FAO, 1989; 174 pp.
18. Squires, V.R., Glenn, E.P. Salinization, desertification, and soil erosion. In The Role of Food, Agriculture, Forestry and Fisheries in Human Nutrition 2011, Squires, V.R., Ed., Vol. III, Oxford: UNESCO EOLSS Publishers, 2004; 102–123.
19. Oster, J.D., Shainberg, I., Abrol, I.P. Reclamation of salt-affected soils. In *Agriculture Drainage, Agronomy Monograph* 38; Madison, WI: ASA, 1999; 659–692.
20. Gupta, R.K., Abrol, I.P. Salt-affected soils: their reclamation and management for crop production. In Advances in Soil Science, Lal, R., Stewart, B.A. Eds.; New York, NY: Springer. 1990; 223–288.

VIII

PRO: Basic Environmental Processes

67

Agricultural Runoff

Matt C. Smith,
David K. Gattie, and
Daniel L. Thomas

Introduction

Agricultural runoff is surface water leaving farm fields as a result of receiving water in excess of the infiltration rate of the soil. Excess water is primarily due to precipitation, but it can also be due to irrigation and snowmelt on frozen soils. In the early 20th century, there was considerable concern about erosion of farm fields due to rainfall. The concern was primarily related to the loss of valuable topsoil from the fields and the resulting loss in productivity. With the passage of the Federal Water Pollution Control Act Amendments of 1972, the potential for pollution of surface water features such as rivers and lakes due to agricultural runoff was officially recognized and an assessment of the nature and extent of such pollution was mandated.[1,2]

Agricultural runoff is grouped into the category of nonpoint source pollution because the potential pollutants originate over large, diffuse areas and the exact point of entry into water bodies cannot be precisely identified (see *Pollution: Point Sources,* p. 2190). Non-point sources of pollution are particularly problematic in that it is difficult to capture and treat the polluted water before it enters a stream. Point sources of pollution such as municipal sewer systems usually enter the water body via pipes and it is comparatively easy to collect that water and run it through a treatment system prior to releasing it into the environment. Because of the non-point source nature of agricultural runoff, efforts to minimize or eliminate pollutants are, by necessity, focused on practices to be applied on or near farm fields themselves. In other words, we usually seek to prevent the pollution rather than treating the polluted water.

Due to the great successes made in treating polluted water from point sources such as municipal and industrial wastewater treatment plants, the relative significance of pollution from agricultural runoff has increased. Agricultural runoff is now considered to be the primary source of pollutants to the streams and lakes in the United States. It is also the third leading source of pollution in U.S. estuaries.[3] The water pollutants that occur in agricultural runoff include eroded soil particles (sediments), nutrients, pesticides, salts, viruses, bacteria, and organic matter.

Agricultural Runoff Quantity

Agricultural runoff occurs when the precipitation rate exceeds the infiltration rate of the soil. Small soil particles that have been dislodged by the impact of raindrops can fill and block soil pores with a resulting decrease in infiltration rate throughout the duration of the storm. As the excess precipitation builds up on the soil surface it flows in thin layers from higher areas of the field towards lower areas. This diffuse surface runoff quickly starts to concentrate in small channels called rills. The concentrated flow will generally have a higher velocity than the flow in thin films over the surface. The concentrated flow velocity may become rapid enough to cause scouring of the soil that makes up the channel sides and bottom. The dislodged soil particles can then be carried by the flowing water to distant locations in the same field or be carried all the way to a receiving water body. If the quantity of flow and the velocity of flow are large enough, the rills can grow so large that they cannot be easily repaired by typical earth moving machinery. When this happens, the rill has become a gulley.

The quantity of runoff from agricultural fields is not usually listed explicitly as a concern separate from the quality of the runoff. However, it should be considered because it transports the pollutants and can cause erosion of receiving streams due to excessive flows. If less runoff is allowed to leave a field, there is less flow available to transport pollutants to the stream. Also, if more water is retained on the field, there is likely to be a corresponding reduction in the amount of supplemental water that will need to be added through irrigation. Runoff quantity varies significantly due to factors such as soil type, presence of vegetation and plant residue, physical soil structures such as contoured rows and terraces, field topography, and the timing and intensity of the rainfall event.

Some agricultural practices increase the infiltration capacity of the soil while other practices can result in decreases. The presence of vegetation and plant residues on a field reduce runoff due to improving and maintaining soil infiltration capacity. Actively growing plants also reduce the amount of water in the soil due to evapotranspiration, thus making more room for infiltrating water to be stored in the soil profile. Bare soils increase runoff because there is nothing except the soil surface to absorb the energy of the falling raindrops. The rain, therefore, dislodges soil particles that will tend to seal the surface and reduce infiltration.

Soil Erosion and Associated Pollutants

One of the primary pollutants in agricultural runoff is eroded soil. In 1975, 223 million acres of cropland produced 3700 million tons of eroded sediments or an average of 17 tons of soil lost per acre of cropland per year (see various *Erosion* entries, pp. 967–1103). It is estimated that cropland, pasture, and rangeland contributed over 50% of the sediments discharged to surface waters in 1977.[4] As noted above, the energy of raindrops can dislodge and transport soil particles. In the aquatic environment the eroded soil is called sediment. There are several concerns related to excessive sediments in aquatic systems. These include loss of field productivity, habitat destruction, reduced capacity in reservoirs, and increased dredging requirements in shipping channels.

Eroded sediments represent a loss of fertile topsoil from the field, which can reduce the productivity of the field itself. Soil formation is an extremely slow process occurring over periods ranging from decades to centuries.[5] Possible results to a grower from excessive erosion of their fields include increasing fertilizer and water requirements, planting more tolerant crops, and possibly abandoning the field for agricultural production.

A second concern is that many of these sediments are heavy and will settle out in slow moving portions of streams or in reservoirs. The settled sediments can dramatically alter the ecology of the streambed. Aquatic plants, insects, and fish all have specific requirements related to composition of the streambed for them to live and reproduce.[6] Sediments in reservoirs reduce the volume of the reservoir available to store water. This may result in reduced production of hydroelectric power, reduced water availability for municipal supply, interference with navigation and recreation, and increased dredging requirements to maintain harbor navigability.

Another concern with eroded sediments is that they can transport other pollutants into receiving waters. The plant nutrient phosphorus, for example, is most often transported from the fields where it was applied as fertilizer by chemically bonding to clay minerals. Many agricultural pesticides also bond to eroded clays and organic matter. Once these chemicals have entered the aquatic ecosystem, many processes occur that can result in the release of the pollutants from their sediment carriers. Phosphorus, when released, can contribute to the eutrophication of lakes and reservoirs (see the entry *Eutrophication,* p. 1115). Pesticides and their degradation products can be toxic to aquatic life and must be removed from municipal water supplies.

Erosion from animal agriculture such as feedlots and pastures can also result in the transport of sediments composed of animal manures (see the various *Manure Management* entries, pp. 1680–1695). These sediments can transport significant quantities of potential pathogens (viruses and bacteria). The animal manures are primarily organic in nature and can serve as a food source for natural bacteria in the receiving water. When these naturally occurring bacteria begin to utilize the organic matter in this way they may lower or deplete the water of dissolved oxygen as they respire and multiply. This use of oxygen by aquatic bacteria is known as biochemical oxygen demand (BOD). High levels of BOD can reduce stream oxygen level to the point that fish and other organisms that require dissolved oxygen suffer, die, or relocate, when possible, to more suitable habitats.[6]

Dissolved Pollutants

Agricultural runoff can carry with it many pollutants that are dissolved in the runoff water itself. These may include plant nutrients, pesticides, and salts. Since these pollutants are dissolved in the runoff, control measures are most often aimed at reducing the volume of runoff leaving an agricultural field, or making the pollutants less available to be dissolved into the runoff water.

One of the major pollutants of concern in agricultural runoff is the plant nutrient nitrogen. Nitrogen is a relatively cheap component of most fertilizers and is necessary for plant growth. Unfortunately, nitrogen in the form of nitrate is highly soluble in water. Thus nitrate can be easily dissolved in runoff water. Just as it does in an agricultural field, nitrogen can promote growth of aquatic vegetation. Excess nitrogen and phosphorus in runoff can lead to the eutrophication of lakes, reservoirs, and estuaries (see the entry *Eutrophication,* p. 1115). Nitrogen in the form of ammonia can be dissolved into runoff from pastures and feedlots. Ammonia is toxic to many aquatic organisms, thus it is important to minimize ammonia in runoff.[7]

Many agriculturally applied pesticides are also soluble in water. They can be dissolved in runoff and transported into aquatic ecosystems where there is a potential for toxic effects. These pesticides must also be removed from drinking water supplies and, if concentrations are high or persistent, such treatment can be difficult and expensive. Stable, persistent pesticides can bioaccumulate in the food chain with the result that consumers of fish from contaminated waters might be exposed to higher concentrations than exist in the water itself.[8]

Runoff from agricultural fields can contain significant concentrations of dissolved salts. These salts originate in precipitation, irrigation water, fertilizers and other agricultural chemicals, and from the soil minerals. Plants generally exclude ions of chemicals that they do not need. In this way, dissolved salts in irrigation water, for example, can be concentrated in the root zone of the growing crop. Runoff can redissolve these salts and transport them into aquatic ecosystems where some, naturally occurring selenium for example, can be toxic to fish and other wildlife.[9]

Transport of fertilizers and pesticides from their point of application can result in significant environmental costs. This transport, or loss from the field, can also have significant negative economic impacts on the grower. Fertilizers lost from the field are not available to promote crop growth. Agricultural chemicals lost from the field, likewise, are not available to protect the plants from pests and diseases. In both cases the grower is paying for expensive inputs and paying to apply them. It is always in the growers' and the environment's best interests, therefore, to keep agricultural chemicals in the field where they are needed and where they were applied.

Control of Agricultural Runoff

One of the most direct methods of controlling pollution by agricultural runoff is to minimize the potential for runoff to occur. Other methods can be employed to reduce the amounts of sediments and dissolved chemicals in runoff. As a whole, management practices designed to minimize the potential for environmental damage from agricultural runoff are called best management practices (BMPs), (see the entry *Nutrients: Best Management Practices,* p. 1829). Many times, practices aimed at controlling one aspect of agricultural runoff are also effective at reducing other components. This is due to the interrelationships between runoff volume, erosion, transport, dissolution, and delivery.

Maintaining good soil tilth and healthy vegetation can minimize runoff. This will promote increased infiltration and a resultant decrease in runoff. Other management practices such as terracing, contour plowing, and using vegetated waterways to convey runoff can result in decreased quantities of runoff by slowing the water leaving the field and allowing more time for infiltration to occur. Construction of farm ponds to receive runoff can result in less total runoff from the farm, lowered peak rates of runoff, and storage of runoff for use in irrigation or livestock watering.[2]

Control of water pollution by the mineral and organic sediments and associated chemicals in agricultural runoff is most effectively achieved by reducing erosion from the field. The primary method of reducing erosion is by maintaining a vegetative or plant residue cover on the field at all times or minimizing areas of the field that are bare. Techniques utilized to accomplish these tasks include conservation tillage, strip tillage, and the use of cover crops (see the entry *Erosion Control: Tillage and Residue Methods*, p. 1081). Additional measures that can be employed at the edge of the field, or off-site include vegetative filter strips and farm ponds.

Methods to control the loss of nitrogen and other plant nutrients from cropland include applying nitrogen in the quantity required by the crop and at the time the crop needs it (see the entry *Nutrients: Best Management Practices*, p. 1829). This requires multiple applications and can be difficult for tall crops. For this reason, most, or all, of the nitrogen required by the crop is often applied at planting. Nitrogen fertilizers have often been applied based on general recommendations for the type of crop to be grown. Since nitrogen fertilizers are relatively inexpensive, growers have tended to over apply rather than under apply. Soil tests can tell a grower how much nitrogen is already in the soil and how much needs to be applied for a specific crop. Efforts have been made to make the nitrogen less soluble by changing the form of nitrogen applied to the field so that it becomes available to the plants (and, thus available for loss in runoff) more slowly.[10]

One method of controlling the loss of agricultural chemicals is to minimize their solubility in water. Another is to minimize their use through programs such as integrated pest management (IPM) where some crop damage is allowed until it reaches a point that it becomes economically justified to apply pesticides.[11] And a third approach is to make the chemicals more easily degraded so that they do their job and then degrade into other, less harmful, chemicals so that they do not stay around long enough to be influenced by runoff-producing rainfall events.

Conclusion

Agricultural runoff is one of the leading causes of water quality impairment in streams, lakes, and estuaries in the United States. It can transport large quantities of sediments, plant nutrients, agricultural chemicals, and natural occurring minerals from farm fields into receiving water bodies. In many cases the loss of these substances from the field represent an economic loss to the grower as well as a potential environmental contaminants. There are many methods by which the quantity of agricultural runoff can be reduced. Many of these methods are referred to generically as BMPs. Adoption of BMPs can also improve the quality (reduce contaminant concentrations) of the runoff that does leave the farm. By reducing the quantity and improving the quality of agricultural runoff, it will be possible to improve the water quality in our streams, river, lakes, and estuaries.

References

1. U.S. Environmental Protection Agency. *EPA Releases Guidelines for New Water Quality Standards*; 2002; http://www.epa.gov/history/topics/fwpca/02.htm (accessed July 2002).
2. Stewart, B.A.; Woolhiser, D.A.; Wischmeier, W.H.; Caro, J.H.; Frere, M.H. *Control of Water Pollution from Cropland, Volume II—An Overview,* EPA-600/ 2–75-026b; U.S. Environmental Protection Agency: Washington, DC, 1976.
3. U.S. Environmental Protection Agency. *Nonpoint Source Pollution: The Nation's Largest Water Quality Problem;* 2002; http://www.epa.gov/OWOW/NPS/facts/point1.htm (accessed July 2002).
4. Leeden, Van der. *The Water Encyclopedia;* Lewis Publishers: Chelsea, MI, 1990.
5. Foth, H.D. *Fundamentals of Soil Science,* 8th Ed.; John Wiley and Sons, Inc.: New York, 1990.
6. Gordon, N.D.; McMahon, T.A.; Finlayson, B.L. *Stream Ecology: An Introduction for Ecologists*; John Wiley and Sons Inc.: New York, 1992.
7. Abel, P.D. *Water Pollution Biology,* 2nd Ed.; Taylor and Francis, Inc.: Bristol, 1996.
8. U.S. Environmental Protection Agency. *The Persistent Bioaccumulators Project;* 2002; http://www.epa.gov/chemrtk/persbioa.htm (accessed July 15, 2002).
9. U.S. Geological Survey. *Public Health and Safety: Element Maps of Soils;* http://minerals.cr.usgs.gov/gips/na/0elemap.htm#elemap (accessed July 15, 2002).
10. Owens, L.B. Impacts of soil N management on the quality of surface and subsurface water. In *Soil Process and Water Quality;* Lal, R., Stewart, B.A., Eds.; Lewis Publishers, Inc.: Boca Raton, FL, 1994.
11. U.S. Department of Agriculture. *National Integrated Pest Management Network;* 2002; http://www.reeusda.gov/agsys/nipmn/ (accessed July 15, 2002).

68

Desertification

David Tongway
and John Ludwig

Introduction

The term *desertification* was coined in the 19708 to graphically represent the state of the Sahelian lands, on the southern margin of the Sahara Desert. This was a period when major drought accompanied by big increases in the human population served to cause the desert margins to apparently move into formerly more productive land.[1,2] The image of an encroaching desert is powerful and evocative and resulted in major international efforts to understand and deal with the problem. Since that time, the notion of desertification has been reworked in the light of additional information and improved conceptual frameworks to the extent that the desert is no longer seen as inexorably increasing in size, nor restricted to the Sahelian fringe of the Sahara.[3–5] Most rangeland areas in the world have suffered some sort of degradation due to the impact of disturbance regimes, and recent reviews[6] have shown the process to be not at all restricted to hot deserts or areas of high population density. This is not to deny, however, the major effects on the human populations using these lands, and no doubt, much human hardship has been endured. This entry describes a process whereby the degree of desertification can be assessed and then used to design restoration activities appropriate to local biophysical and socioeconomic constraints.

Desertification Redefined

This entry focuses on the biophysical aspects of desertification. Traditionally, easy to measure structural and functional attributes of vegetation, such as species composition and productivity, were the means by which desertification was initially assessed. These methods served to show the effect of desertification but did not provide a predictive understanding of how to combat it. However, recent advances in landscape ecology and restoration technology[7,8,10] have led to generic and practical approaches to study the basic nature and reversibility of desertification. Principally, this involves treating the affected landscapes as biophysical systems, comprised of sequences of processes and feedback loops and summarized in a conceptual framework (Figure 1).[8,9]

In this framework, vital resources such as water, nutrients, and topsoil are transported, utilized and cycled in space and time, and the processes are driven by processes such as runoff/runon, erosion/deposition, and plant litter de- composition.[8,11,12] Landscapes are said to be “functional,” or

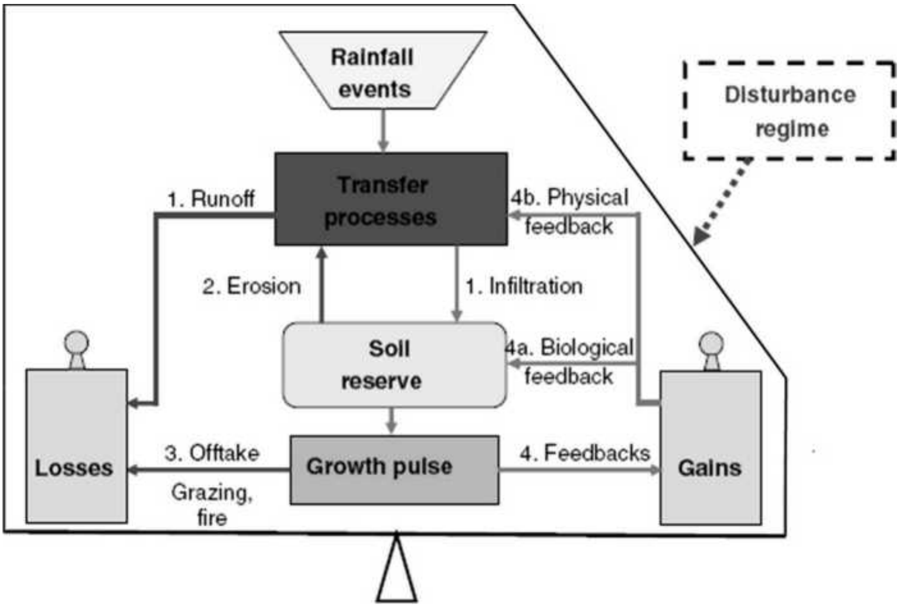


FIGURE 1 A conceptual framework summarizing landscape functioning. Numbers refer to the recommended sequence of assessing practical function. The disturbance regime shown is generic, and it may impinge on a number of landscape processes at the same time.
Source: A modification of Figure 2 in Tongway and Ludwig[9] (reproduced with permission).

non-desertified, if resources are substantially retained within the system and utilization and cycling processes are efficient. "Dysfunctional" or desertified landscapes are characterized by the depletion of the stock of some vital resources and the continued flow of these resources out of the system. This mind-set emphasizes the system attributes of processes acting in space, over time, in relation to applied stress and disturbance, rather than just changes in lists of species, or yields of marketable commodities. The role of vegetation as a significant regulator of energy and resources is integral to this approach.[13] Desertification should be viewed as a continuum ranging from slight to severe, rather than as a simple step function (Figure 2).

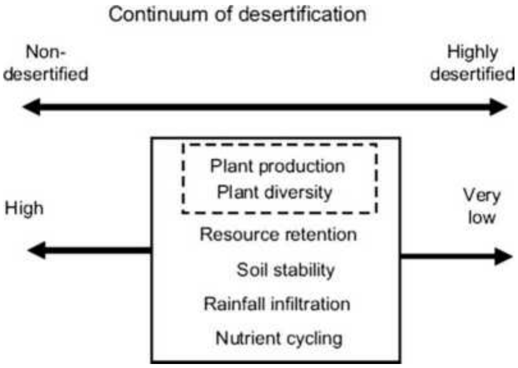


FIGURE 2 Desertification as a continuum. The four new biophysical parameters (bold) are added to the two existing desertification descriptors (dotted box) to locate any given landscape on the continuum.

Assessing the Degree of Desertification

If field sites are characterized according to "resource regulation" capacity, not only can the degree of desertification be assessed, but also the critical pathway of resource loss may be identified. Tongway and Hindley[14] and Tongway and Ludwig[15] have designed and implemented monitoring programs to quickly provide information about biophysical processes and edaphic properties related to plant habitat favorability at both landscape and plot or patch scales. Typically, the initial analysis examines the fate of rainfall into infiltrated water and runoff water. The data gathered need an interpretational framework to facilitate generic application. Graetz and Ludwig[16] proposed that system response to desertification be represented by a four-parameter sigmoidal or logistic curve. The curve form acknowledges upper and lower asymptotic plateaux, at the non- and highly desertified ends of the spectrum, respectively, and a gradual transition between those plateaux. This approach permits questions about whether a system was "fragile," i.e., easily made dysfunctional, with low restoration potential or "robust," or rather able to withstand stress and disturbance with only low attenuation of biophysical processes (Figure 3). Importantly, this curve type enables thresholds and milestones to be predicted and quantified using field indicators.[17]

Procedures to Reverse Desertification

Rehabilitation and restoration of desertified landscapes, under the functional biophysical system mind-set, require that processes that accumulate resources be reintroduced or augmented, thus providing a rational procedure in the repair and functional recuperation of desertified landscapes. The approach explicitly seeks to retain vital resources by ecological processes.[18–20] Once the analysis of resource regulation system has been completed, and the most affected process identified, remedial efforts can begin. For example, Rhodes and Ringrose-Voase[20] deduced that ponding water for extended periods on clay soils with modest swell/shrink properties would eventually result in soils with high infiltration and water store through sequences of swelling and shrinking processes permitting infiltration into greater soil depths. Recolonization and establishment of plants then began spontaneously, eventually cycling organic matter, so that open friable soils developed, colonized by soil macrofauna that further improved soil properties. Tongway and Ludwig[18,19] used piles of branches arranged on the contour in gently sloping country to trap water, sediment, and organic matter to effect substantial improvements to a range of soil properties, permitting perennial grasses to self-establish. In each of these cases, an analysis of the different underlying causes facilitated the selection of the most appropriate techniques to reverse

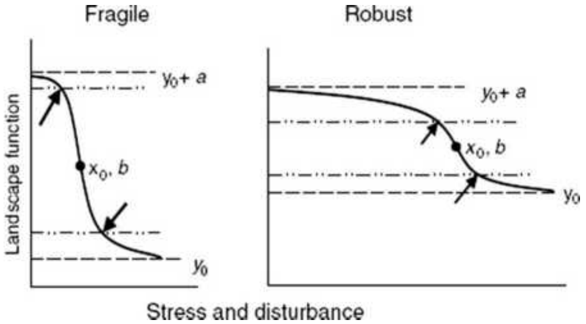


FIGURE 3 Examples of response curves for fragile and robust landscapes. The initial response of landscape function to stress and/ or disturbance is markedly different. A fragile landscape deteriorates with low applied stress and has a much lower base value (y_0) than the more robust landscape. Four-parameter sigmoid curves of the form $y = (y_0 + a)/1 + e^{(-x-x_0)/b}$ provide four practical values reflecting the nature of the landscape. Critical thresholds (arrows) for each index of desertification can be determined by curve analysis.

the observed desertification. Attempting to revegetate desertified areas by simply reseeding without understanding both the current and required edaphic properties needed for the desired vegetation mix frequently results in unexplained failure. In some instances, where the system function is close to the lower asymptote (Figure 3), simple treatments such as exclusion from grazing will be too slow or ineffective and active intervention may be needed to improve one or more functional processes.

Monitoring Rehabilitation

It is important to monitor the progress of processes set in train by the rehabilitation activities. Essentially, the degradation curves such as those in Figure 3 need to be driven "in reverse." The landscape assessment procedure proposed by Tongway and Hindley and Tongway and Ludwig[15,21] can also be used to follow the trajectory of improvement in ecosystem functioning. The procedures use simple, rapidly acquired indicators of processes of resource regulation. Data recording biota establishment and development are included in this procedure and interpreted in terms of the rising plane of delivery of goods and services to the whole system over time. It is important that monitoring should provide accurate information quickly and at low cost. Remotely sensed data, related to landscape function, is a cost-effective procedure,[22] and new products such as Google Earth may, in the future, provide synoptic assessment of restoration trends at coarse scales. The effect of rare stochastic events such as fires or storms may need to be assessed to see if the resultant stress and/or disturbance has set the system back beyond a critical threshold or not.

Conclusion

We have described an ecosystem-function-based set of data gathering processes by which the fundamental nature of desertification can be assessed and combated, using a framework that characterizes the biophysical status of the affected system. This can be simply expressed as "assessing the regulation of vital resources in space and time." In deploying the procedure, community groups can be easily instructed to "read the landscape." The procedure enables the user to design or adapt restoration or rehabilitation technologies appropriate to the problem at hand because of the predictive understanding acquired, rather than using a "recipe" from another type of landscape elsewhere. The information-gathering procedure can be adapted to a wide range of bioclimatic situations because it deals with the basic processes controlling the availability of vital resources to biota.

References

1. UNEP. *United Nations Conference on Desertification: An Overview;* United Nations Environment Program: Nairobi, Kenya, 1977.
2. UNEP. *General Assessment of Progress in the Implementation of the Plan of Action to Combat Desertification 1978–84*; United Nations Environment Program: Nairobi, Kenya, 1984.
3. Arnalds, O. Desertification: An appeal for a broader perspective. In *Rangeland Desertification;* Arnalds, O., Archer, S., Eds.; Kluwer Academic: Dordrecht, 2000; 5–15.
4. Chen, Z.; Xiangzhien, L. In *People and Rangelands: Building the Future*, Proceedings of the VI International Rangeland Congress, Townsville, Australia, July 19–23, 1999; Eldridge, D., Freudenberger, D., Eds.; VI International Rangeland Congress: Townsville, 1999; 105–107.
5. Tongway, D.; Whitford, W. In *People and Rangelands: Building the Future*, Proceedings of the VI International Rangeland Congress, Townsville, Australia, July 19–23, 1999; Eldridge, D., Freudenberger, D., Eds.; VI International Rangeland Congress: Townsville, 1999; 89–142.
6. Archer, S.; Stokes, C. Stress, disturbance and change in rangeland ecosystems. In *Rangeland Desertification*; Arnalds, O., Archer, S., Eds.; Kluwer Academic: Dordrecht, 2000; 17–38.

7. Breedlow, O.A.; Voris, P.V.; Rogers, L.E. Theoretical perspective on ecosystem disturbance and recovery. In *Shrub-Steppe: Balance and Change in a Semi-Arid Terrestrial Ecosystem;* Rickard, W.H., Rogers, L.E., Vaughn, B.E, Liebetrau, S.F., Eds.; Elsevier: New York, 1988; 257–269.
8. Ludwig, J.A; Tongway, D.J. A landscape approach to rangeland ecology. In: *Landscape Ecology Function and Management: Principles from Australia's Rangelands;* Ludwig, J., Tongway, D., Freudenberger, D., Noble, J., Hodgkinson, K., Eds.; CSIRO: Melbourne, Australia, 1997; 1–12.
9. Tongway, D.; Ludwig, J. *Restoring Disturbed Landscapes: Putting Principles into Practice*; Island Press: Washington DC, 2010.
10. Whisenant, S.G. *Repairing Damaged Wildlands: A Process- Oriented, Landscape-Scale Approach;* Cambridge Univ. Press: Cambridge, England, 1999; 312 pp.
11. Jorgenson, S.E.; Mitsch, W.J. Ecological engineering principles. In *Ecological Engineering: An Introduction to Ecotechnology;* Mitsch, W.J., Jorgensen, S.E., Eds.; John Wiley and Sons: New York, 1989; 21–37.
12. Schlesinger, W.H.; Reynolds, J.F.; Cunningham, G.L.; Hueneke, L.F.; Jarrel, W.M.; Virginia, R.A.; Whitford, W.G. Biological feedbacks in global desertification. Science **1990**, *247,* 1043–1048.
13. Farrell, J. The influence of trees in selected agroecosystems in Mexico. In *Agroecology: Researching the Ecological Basis for Sustainable Agriculture*; Gliessman, S.R., Ed.; Springer-Verlag: New York, 1990; 167–183.
14. Tongway, D.J.; Hindley, N.L. *Landscape Function Analysis: Methods for Monitoring and Assessing Landscapes, with Special Reference to Minesites and Rangelands;* CSIRO Sustainable *Ecosystems:Canberra,2004,availableathttp://www.csiro.au/services/EcosystemFunctionAnalysis.* htm.
15. Tongway, D.; Ludwig, J. *Restoring Disturbed Landscapes: Putting Principles into Practice*; Island Press: Washington DC, 2010.
16. Graetz, R.D.; Ludwig, J.A. A method for the analysis of piosphere data applicable to range assessment. Aust. Rangeland J. **1978**, *1,* 126–136.
17. Tongway, D.; Hindley, N. *Ecosystem Function Analysis of Rangeland Monitoring Data*; 2000, available at http://www.nlwra.gov.au/atlas.
18. Tongway, D.J.; Ludwig, J.A. Rehabilitation of semi-arid landscapes in Australia. I. Restoring productive soil patches. Restor. Ecol. **1996**, *4,* 388–397.
19. Ludwig, J.A.; Tongway, D.J. Rehabilitation of semi-arid landscapes in Australia. II. Restoring vegetation patches. Restor. Ecol. **1996**, *4*, 398–406.
20. Rhodes, D.W.; Ringrose-Voase, A.J. Changes in soil properties during scald reclamation. J. Soil Conserv. Serv. NSW **1987**, *43,* 84–90.
21. Tongway, D.; Hindley, N. Assessing and monitoring desertification with soil indicators. In *Rangeland Desertification;* Arnalds, O., Archer, S., Eds.; Kluwer Academic: Dordrecht, 2000, 89–98.
22. Kinloch, J.E.; Bastin, G.N; Tongway, D.J. (2000). Measuring landscape function in chenopod shrublands using aerial videography. In the Proceedings of the 10th Australasian Remote Sensing and Photogrammetry Conference, Adelaide, Australia 21–25 August 2000, 480–491; Causal Productions: Adelaide, South Australia.

69

Desertification: Prevention and Restoration

Claudio Zucca,
Susana Bautista,
Barron Orr, and
Franco Previtali

Introduction

This entry provides a brief overview of the most significant approaches to prevent and reverse land degradation in drylands, with a special emphasis on conceptual frameworks and on methods to monitor and evaluate their impacts.

The entry first discusses the main underlying concepts, in light of recent developments on the conceptual framework of desertification. The methodological developments related to integrated and participatory evaluation are then presented. Finally, the implementation of mitigation and restoration programs is addressed, with particular reference to the constraints and risk factors and to the practical lessons learned in the field.

Desertification: The Underlying Concepts

Desertification is defined here as "land degradation in drylands resulting from various factors, including climatic variations and human activities," in conformity with the UNCCD (United Nations Convention to Combat Desertification).[1] Before the entry into force of the UNCCD in 1996, the term "desertification" had been given a number of different definitions.[2–6] The UNCCD definition has been and still is the subject of scientific debate; the related evolving concepts have been reviewed by a number of papers.[7–12]

The Millennium Ecosystem Assessment (MA) defines desertification as a persistent reduction in the provision of ecosystem services over an extended period.[13,14] The scientific discussion promoted in 2009 by the UNCCD in view of the First Scientific Conference of its CST (Committee on Science and Technology), building on the MA definition, resulted in a proposal to redefine desertification as "an end state of the process of land degradation; this process is expressed by a persistent reduction or loss of biologic and economic productivity of lands that are under use by people."[15] The ongoing discussion also focuses on the integrated sets of indicators needed for monitoring and assessing desertification, and on the related conceptual frameworks that would help scientists, practitioners, and policy makers organize, use, and communicate the results of that monitoring.[16–18]

While the definition of desertification continues to evolve, ecosystem services are increasingly seen as a unifying supporting concept. The MA states that "desertification results from a long-term unbalance between demand for and supply of ecosystem services" and that measurement of persistent reduction in the capacity of ecosystems to supply services provides a robust and operational way to quantify land degradation and desertification. The ecosystem services framework is increasingly thought to provide a basis to assess and value the impacts of land change and degradation, as well as the effects of the actions aimed at reversing it.[19–21]

Desertification manifests itself through different forms and processes in different ecosystems and socioeconomic contexts. Its direct causes are many and can be generally ascribed to different forms of land mismanagement, such as overgrazing, deforestation, overuse of irrigation, and non-resource-conservative agriculture and forest practices.[1,7,17] These are at the origin of the major land degradation processes that are globally affecting the provision of ecosystem services, including water and wind erosion, soil salinization, loss of vegetation cover and diversity, and degradation of the hydrological cycles.[13]

No satisfying estimates of the global extent and severity of desertification are available thus far; however, the new World Atlas of Desertification, which is based on multiscale integrated sets of indicators, provides global datasets that can be used to identify local or regional areas of concern.[22] Assessment and monitoring of desertification, as well as of the impact of the prevention and restoration interventions, still constitute a major research challenge[15] and for this reason are a primary focus of this entry.

Prevention and Reversal

Policy and management responses to desertification can be grouped under two major classes: prevention and reversal.[13] The boundaries between these ones are vague, as in practice they form a continuum of potential prevention, mitigation, and restoration actions, to be adapted to particular sites and dynamics through adaptive management approaches (Figure 1). Through integrated land use planning, the optimal spatial mix of prevention and reversal responses can be identified, so that in net terms, desertification can be controlled. This approach, which was endorsed by the country Parties of the UNCCD in 2017, is known as land degradation neutrality (LDN).[23–25]

Prevention actions can be considered as avoidance approaches either through proactive management efforts or through changes in land use and management currently leading to desertification.

Prevention, however, is not enough to address the challenges posed by desertification. Vast areas of drylands are already severely degraded, with reduced plant cover and species diversity, falling

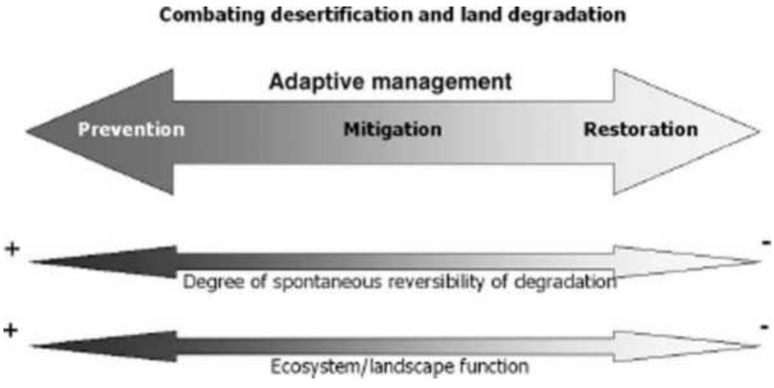


FIGURE 1 Continuum of actions to combat desertification and land degradation.

productivity, depleted or eroded soils, and very low potential for spontaneous recovery of ecosystem functions, even when degradation forces are no longer stressing the system. Many of these systems have changed at a level at which restoration is the only viable option to recover ecosystem services that have been lost.[26–28]

Examples of prevention and mitigation actions include measures to improve water management and agricultural practices. These are often referred to as soil and water conservation (SWC) or sustainable land management (SLM) practices. The FAO–LADA (Food and Agriculture Organization of the United Nations/Land Degradation Assessment in Drylands)[29] project and WOCAT (World Overview of Conservation Approaches and Technologies)[30] provide a framework for classifying and evaluating SLM actions. In view of the needs of the UNCCD, SLM has been recently defined as "land managed in such a way as to maintain or improve ecosystem services for human wellbeing, as negotiated by all stakeholders."[21]

Examples include long-term crop rotations with cereals/legumes; more efficient use of fertilizers; improvements in water-use efficiency; conservation-minded tillage methods; water-harvesting and water storage techniques; measures that protect soils from erosion, salinization, and other forms of soil degradation; improved crop–livestock integration, combining livestock rearing and cropping to allow a more efficient recycling of nutrients within the agricultural system; and in situ conservation of genetic resources and better resource use with efficient germplasm.[31] Creating viable livelihood alternatives, including the creation of economic opportunities in urban centers, can also help reduce current pressures on drylands.

For extremely degraded lands, rehabilitation and restoration approaches often involve the improvement in the quantity and/or quality of vegetation cover through, for example, the establishment of seed banks, reintroduction of selected species, control of invasive species, and reforestation programs.

Desertification is driven by a combination of proximate causes and underlying forces, including their interactions and feedback; these vary from region to region and change over time.[32] Approaches and strategies to prevent and reverse desertification need therefore to address the dynamic causal patterns and multiplicity of actors, factors, and scales involved. In general, developing the appropriate engagement between scientific and local environmental knowledge is critically important for efforts to prevent and reverse land degradation.[33] Desertification is framed within multiscale, coupled human–environmental dynamics, and so must be the approaches for desertification prevention and reversal.[34,35] The relationships between land degradation and its causal agents are non-linear and complex. Degradative and aggradative trajectories commonly exhibit thresholds and rapid shifts, as well as hysteresis dynamics, where the trajectories of degradation and recovery differ.[36–38] Understanding and monitoring these relationships are critical in the design of strategies to combat desertification.

In addition, socioeconomic conditions impose limitations on the technology and inputs available. Therefore, the approaches to combat desertification should incorporate both current conditions and scenario projections of socioeconomic and environmental constraints and opportunities.[13]

Finally, there is growing evidence that land degradation in desertification-prone areas is likely to increase with climate change.[22] Desertification is linked to biodiversity loss and global climate change through the regulation of water and carbon fluxes. Therefore, interlinkages in policy formulations aimed at combating desertification, mitigating the effects of climate change, and conserving biodiversity must be beneficially exploited by developing multifunctional strategies that address the three global environmental goals.

Implementing Solutions in the Field: Lessons Learned

Despite the availability of technological, institutional, and even financial resources, efforts to combat desertification often fail because of poor implementation. A list of "lessons learnt" is presented below to summarize a range of major constraints and risk factors that can hinder the successful implementation of the interventions, while highlighting the lessons learned and necessary improvements. Some of those issues stem from the concepts of adaptive management and multiscale human–environmental dynamics as introduced in the previous section. Others are more related to the issues of local participation and integrated assessment that will be discussed in the following section. Points 1 and 2 deal with the quality of the technical design and its degree of adaptation to local conditions and knowledge. Points 3–7 are related to the ability of the projects to cope with the long-term human–environmental dynamics. Points 8–11 are connected to the quality of the participatory processes implemented by the projects, while the last two points are linked to major, common organizational constraints.

1. Lack of awareness of actual land conditions. Sometimes, people (including decision makers) living in degraded areas do not realize that their land still maintains productive potential or perhaps are unaware of how that potential can be exploited. Preliminary land surveys should be done to support project design. These baseline assessments are necessary to guide subsequent actions. In some cases, "no-action" options could be considered. In the case of degraded rangelands, simple and cost-effective "self-learning" tests based on grazing exclusion and rotation can be proposed to local communities as a means to demonstrate the effects of pressure mitigation and sustainable management. Such tests allow hands-on learning and practical experience, and are much less risky than the direct introduction of often expensive "all or nothing" rehabilitation programs.
2. Schematic approach. Sometimes, large programs are extensively implemented by adopting schematic approaches, which are not able to adapt to specific local land characteristics and the needs of local people. In other cases, they target geographic areas where they are not necessary or inadvertently lead to negative side effects. Implementation protocols offering multiple technical solutions and integrating local stakeholder input on the perceived benefits (and unintentional consequences) are needed.
3. Lack of long-time planning in restoration. Often the long-term dynamics of the "restored human–environmental systems" are not fully considered. This is particularly relevant when the interventions are based on introducing fast-growing, income-generating alien species that may require a "renaturalization" strategy in the long term to balance ecological and social sustainability.
4. Inability to cope with natural crises. Especially in projects aimed at increasing plant cover, a poor or delayed wet season or recurrent droughts may cause the loss of part or all of the investment in a project. For example, this may happen in the case of fodder shrubs plantations, when farmers cannot avoid early grazing due to drought and unavailability of alternative feeding resources. Emergency/contingency funds, quick diagnostic and intervention mechanisms, and flexibility in project duration are necessary to mitigate against this risk.
5. Socioeconomic and demographic dynamics. Some projects have experienced labor shortfalls due to the out-migration of young people. In contrast, the return of people onto land previously "closed for restoration" may cause unsustainable pressure. Addressing economic constraints and associated demographic pressures such as migration prior to project implementation can reduce such kind of risks.
6. Market drivers. The dynamics of international and local market prices may completely and quite rapidly negate achievements produced by years of conservation programs.
7. Contrasting policies. A restoration initiative may be useful in practical terms and yet be overwhelmed by unrelated policies. A common example is farming incentives that come in conflict with the goals of mitigation programs, if not accompanied by adequate guidelines.
8. Passive community participation. Community participation should be strongly based on responsible awareness and sharing of project objectives, and participation in its planning, implementation, and evaluation. Cases where, for example, sectoral administration goals are implemented through prior agreement with land users (e.g., "I let you do on my land") may require or even force farmer action but, in the end, may not represent or address the key concerns or needs of local people.

9. Uncertain community commitment. Stakeholder engagement that fails to bring all parties to the table can have dire consequences. Comprehensive, balanced, and approved representation is not easy to obtain, but it is essential for success. The key question often raised is, "Who is committing on behalf of whom?" The commitment should include a community contribution or investment to cover the implementation costs, be linked to final results, and be based on taking ownership and responsibility after a project ends, rather than simply for the completion of individual tasks. Community commitment should be clearly defined in a way that can cope with changing community priorities.
10. Institutional commitment. Complex projects, in which the implementation is based on the support of local administration staff, need a formal and clear institutional commitment. This may take time and very rarely can be established before project approval: Projects should have an inception phase to set agreements.
11. Lack of transparency in the engagement process of the ultimate project beneficiaries. This is especially true (but not only) when the involvement of individual farmers or other land users in projects is not mediated by the community. Inappropriate or unrepresentative participation can lead to failure and loss of credibility. Transparent and objective selection mechanisms must be understood by all and be followed rigorously by project implementers.
12. Unrealistic project duration. Very often, the most common investment period imposed by donors (3–5 years) is too short to allow for adequate biophysical response and/or socioeconomic adaptation that would assure success. Of particular concern is enough time for effective "inception or learning phases" for community members to deal with the necessity to adapt to changing conditions and needs, to cope with unforeseen events, and most of all to monitor and assess impacts and sustainability.
13. Spending and reporting difficulties. Spending rules, procurement procedures, "exotic" rules imposed by the donors, or strict local rules, not well known by the local project management, often lead to delays and underperformance. More in general, the lack of flexibility can exacerbate the effects of most of the above-mentioned difficulties.

Finally, our capacity to design effective projects is often undermined by a lack of integrated assessment of the progress and success of the previous projects. Projects that are adequately implemented, monitored, and evaluated generate impact, and useful lessons that can feed into successful scaling strategies.[39]

Integrated Evaluation of Prevention and Restoration Actions

The development of integrated biophysical and socioeconomic analytical methods for evaluating progress and success, along with a framework for knowledge sharing and transfer, is crucial to combating desertification.[34] Furthermore, monitoring and evaluation are needed to demonstrate the benefits of sustainable dryland management, establish cost-effective thresholds for the various management alternatives, and identify priority areas where actions could be most effective.

In recent years, there have been a number of initiatives to develop common and comprehensive methodologies for assessing and evaluating the effectiveness of management and restoration programs.[40–44] These approaches focus on indicators that relate to ecosystem integrity and services, and human well-being (socioeconomic and cultural variables). Irrespective of the biophysical or socioeconomic attributes assessed, the selected indicators should be relevant, be sensitive to variations of environmental stress, have the capacity to respond to stress in a predictable manner, but also be simple and measurable with a reasonable level of effort and cost.[45,46]

Soil conditions and vegetation cover and composition are the most common metrics used for evaluating mitigation and restoration actions.[47,48] However, used in isolation, these indicators cannot always reflect how well an ecosystem is functioning. During the last decade, a variety of functional assessment approaches that relate to the spatial pattern of vegetation have been proposed.[49–51]

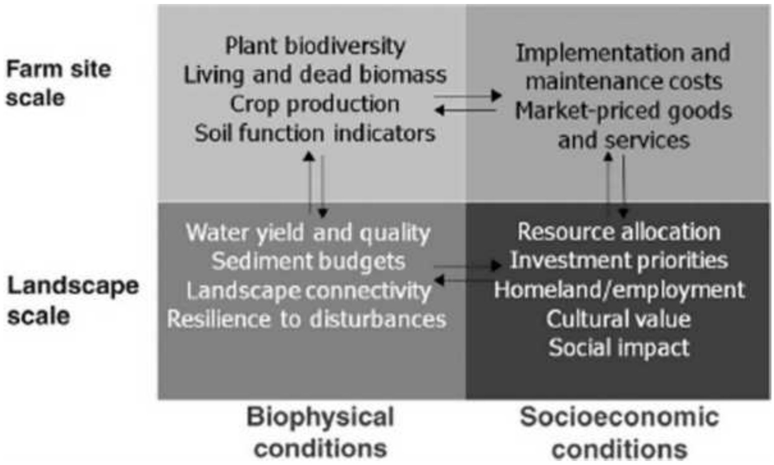


FIGURE 2 Example of a multiscale integrated framework and indicators for evaluating mitigation and restoration actions.

Some of these functional assessment methods also incorporate properties relative to the soil surface condition.[52,53] The theoretical framework for these approaches considers that landscapes occur along a continuum of functionality from highly patchy systems that conserve all resources to those that have no vegetation patches and leak all resources.[54] For semiarid ecosystems, it has been hypothesized that vegetation patchiness could be used as a signature of imminent transitions[55] and that changes in patch-size distributions may be a warning signal for the onset of desertification.[56]

Evaluation frameworks must account for the cross-scale and social–ecological interactions affecting the response of degraded lands to mitigation/restoration actions.[13] A multiscale approach is always advisable. Farm- and project-scale assessments focus on local resources and productivity, and a private economic valuation perspective (market-priced goods and services), while landscape- and program-level indicators address environmental benefits and public/social welfare considerations (Figure 2).

Because of the large spatial and temporal variability of ecosystems, particularly in drylands, it is critical to focus on "slow variables,"[57] so the assessment of long-term changes, and of the sustainability of land management, is not confused by short-term variations in land and socioeconomic conditions.[34] Assessment methods range from simple, qualitative assessments based on field observations to relatively complex protocols based on quantitative measurements of critical ecosystem and landscape attributes and socioeconomic surveys.[42] The development and accessibility of remote-sensing (RS) products have led some international bodies to recommend the integrated use of RS-based geospatial information with ground-based observations to assess land degradation.[13,15,58,59]

Participatory Approaches

While substantive progress has been made in developing more standardized and more relevant environmental assessment and monitoring approaches that reflect human–ecological interactions, the adoption of evaluation results at the local level remains challenging.[35,60] Though some suggest that local interests or even national policies outside the control of evaluators may partially explain this, environmental assessments tend to be independent, unilateral, and top-down, with results being delivered post-assessment. Stakeholder engagement is rare, or all-too-often limited to a period of "public comment" immediately before and/or after the assessment. In the same way that adaptive governance and ecosystem-based management approaches require local knowledge and continual stakeholder engagement, so too must the corresponding monitoring and evaluation if the goal is for results to be truly embraced and used.[61–63] Adaptive, ecosystems-based management is data intensive and requires a commitment to a variety of data sources, including local knowledge, which in turn necessitates more attention to knowledge integration and methods of analysis at different scales. Adoption of best practices that in theory come from a well-executed evaluation of a given desertification mitigation or restoration action can fail due to poor or limited communication and knowledge exchange among the involved actors. Essentially ignoring local knowledge increases the risk of missing essential key local factors, opportunities, or constraints.

Engaging stakeholders from the outset of an environmental assessment and maintaining the interaction throughout the process can result in the integration of local people and their perceptions into management, planning, and evaluation, helping develop feelings of ownership and representation while giving voice to locals in the process.[33,64,65] The potential benefits go beyond the assessment itself. Analysis of environmental conflict resolution processes suggests that ensuring all parties are at the table and are effectively engaged is directly correlated with often-sought outcomes like reaching an agreement, the quality of agreement, and improved working relationships among parties.[66,67]

An additional benefit to a participatory approach to the evaluation and the exchange of ideas among stakeholders, including researchers, is the learning that takes place. The collective self-reflection through interaction and dialogue among the diverse set of stakeholders involved with or affected by environmental challenge and the assessment of associated responses can result in the coproduction of knowledge.[68–70] Social groups that develop a shared understanding of a challenge can build up the experience necessary to improve linkages between knowledge and the environment, cope with change, and enhance adaptation because social learning helps solidify knowledge systems made up of the relevant sets of actors, networks, or organizations.[71] While the majority of discussion about the benefits of encouraging social learning has been focused on improving adaptive management, it is clear that the wealth of information explored during an evaluation suggests that an assessment period is an ideal time to encourage stakeholder interaction and knowledge exchange. In this sense, the evaluation itself becomes a tool for outreach and inreach, where land users, natural resource managers, and scientists all stand to learn from and potentially benefit from each other's insights. The process has the potential to empower individuals, build relationships, expand networks, and thereby enhance the relevance and impact of decision making.

Conclusion

Desertification is one of the major global environment and sustainable development challenges. It affects the livelihoods of millions of people, threatening human well-being in drylands.[13] That risk, when considered relative to the foreseeable impacts of climate change, stands to grow significantly in the future, with estimates suggesting that as many as 50 million people will be in peril of physical displacement in the next ten years.[72]

This entry underscores a range of conceptual and practical issues influencing the design and implementation of interventions, with a special emphasis on conceptual frameworks and associated methods to monitor and evaluate impacts.

Lessons learned in the field highlight several issues threatening the success of prevention/reversal actions. The capacity to design and implement effective restoration actions and other countermeasures to desertification is often undermined by a lack of assessment of the outcomes of previous projects.

The assessment of actions is complex and requires conceptual tools such as human–environmental frameworks for integrated assessment and participatory approaches to foster social learning. These should become the basis for the development of multifunctional mitigation strategies and the formulation of interlinked policies to address desertification, climate change, and biodiversity. The development of participatory approaches in the assessment of interventions is critical to capacity building and knowledge exchange. Encouraging social learning and a true sense of ownership is essential for successful adaptive management.

Other aspects, lying beyond the scope of the present entry, are crucial to combating desertification.

Primarily, the assessment of intervention programs should involve the policy makers to promote the adoption of sustainable rural development policies and to counteract the socioeconomic and policy-driven dynamics of desertification.[73] As a second priority, future interventions should be increasingly oriented to income-generating actions to strengthen social and economic sustainability in concert with environmental sustainability.[74] Finally, it is important to keep in mind that practices and interventions can be considered as "good" or "best" only with reference to their suitability in relation to specific human–environmental contexts. In this regard, to develop strategies tuned to the changing land features, land evaluation techniques could be updated based on the integrated assessment principles, as suggested by the new "anthroscape" concept.[75]

References

1. UNEP. *United Nations Convention to Combat Desertification in Those Countries Experiencing Serious Drought and/ or Desertification, Particularly in Africa*; UNEP: Geneva, 1994.
2. Aubreville, A. *Climats, forets et Désertification de l'Afrique tropicale*; Société d'éditions géographiques maritimes et coloniales: Paris, 1949.
3. UNEP. *Draft Plan of Action to Combat Desertification*. Document A/CONF.74/L.36; UNEP: Nairobi, 1977.
4. Dregne, H. *Desertification of Arid Lands*; Harwood Academic Publisher: London, 1983.
5. FAO/UNEP. Provisional Methodology for Assessment and Mapping of Desertification; FAO: Rome, 1983.
6. UNEP. *Status of Desertification and Implementation of the United Nations Plan of Action to Combat Desertification*; UNEP: Nairobi, 1991.
7. Thomas, D.S.G.; Middleton, N.J. *Desertification; Exploding the Myth*; John Wiley & Sons: Chichester, 1994.
8. Enne, G.; Zucca, C. *Desertification Indicators for the European Mediterranean Region. State of the Art and Possible Methodological Approaches*; ANPA: Rome, 2000.
9. Arnalds, O. Desertification: An appeal for a broader perspective. In *Rangeland Desertification*; Arnalds, O., Archer, S., Eds.; Springer: Berlin, 2000.
10. Le Houérou, H.N. Man-made deserts: Desertization processes and threats. *Arid Land Res. Manage.* 2002, 16 (1), 1–36.
11. Eswaran, H.; Lal, R.; Reich, P.F. Land degradation: An overview. In *Response to Land Degradation*; Bridges, E.M., Hannam, I.D., Oldeman, L.R., Pening de Vries, F.W.T., Scherr, S.J., Sompatpanit, S., Eds.; Science Publishers, Inc.: Enfield, NH, 2001; 20–35.
12. Herrmann, S.M.; Hutchinson, C.F. The changing contexts of the desertification debate. *J. Arid Environ.* 2005, 63 (3), 538–555.
13. MEA, Millennium Ecosystem Assessment. *Ecosystems and Human Well-Being: Desertification Synthesis*; Island Press: Washington, DC, 2005.
14. Safriel, U.N. The assessment of global trends in land degradation. In Climate and Land Degradation; Sivakumar, M.V.N., Ndiang'ui, N., Eds.; Springer-Verlag: Berlin, 2007; 1–38.
15. DSD. *Integrated Methods for Assessment and Monitoring Land Degradation Processes and Drivers (Land Quality)*. White paper of the DSD Working Group 1, version 2. DSD (Dryland Science for Development Consortium); JRC: Ispra, 2009.
16. Sommer, S.; Zucca, C.; Grainger, A.; Cherlet, M.; Zoug-more, R.; Sokona, Y.; Hill, J.; Della Peruta, R.; Roehrig, J.; Wang, G. Application of indicator systems for monitoring and assessment of desertification from national to global scales. *Land Degrad. Dev.* 2011, 22 (2), 184–197.
17. Zucca, C.; Della Peruta, R.; Salvia, R.; Sommer, S.; Cherlet, M. Towards a world desertification atlas. Relating and selecting indicators and datasets to represent complex issues. *Ecol. Indic.* 2012, 15, 157–170.

18. Orr, B.J. *Scientific Review of the UNCCD Provisionally Accepted Set of Impact Indicators to Measure the Implementation of Strategic Objectives 1, 2 and 3*; White Paper—Version 1. UNCCD: Bonn, 2011, available at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.454.420&rep=rep1&type=pdf (accessed October 2019).
19. Costanza, R.; d'Arge, R.; de Groot, R.S.; Farber, S.; Grasso, M.; Hannon, B.; Limburg, K.; Naeem, S.; O'Neill, R.V.; Paruelo, J.; Raskin, R.G.; Sutton, P.; van den Belt, M. The value of the world's ecosystem services and natural capital. *Nature* 1997, 387 (6630), 253–260.
20. de Groot, R.S.; Wilson, M.A.; Boumans, R.M.J. A typology for the classification, description and valuation of ecosystem functions, goods and services. *Ecol. Econ.* 2002, 41 (3), 393–408.
21. DSD. *Monitoring and Assessment of Sustainable Land Management*. White paper of the DSD Working Group 2, version 2. DSD (Dryland Science for Development Consortium); JRC: Ispra, 2009.
22. Cherlet, M.; Hutchinson, C.; Reynolds, J.; Hill, J.; Sommer, S.; von Maltitz, G., Eds. *World Atlas of Desertification*; Publication Office of the European Union: Luxembourg, 2018.
23. Orr, B.J.; Cowie, A.L.; Castillo Sanchez, V.M.; Chasek, P.; Crossman, N.D.; Erlewein, A.; Louwagie, G.; Maron, M.; Metternicht, G.I.; Minelli, S.; Tengberg, A.E.; Walter, S.; Welton, S. *Scientific Conceptual Framework for Land Degradation Neutrality. A Report of the Science-Policy Interface*; United Nations Convention to Combat Desertification (UNCCD): Bonn, 2017.
24. Cowie, A.L.; Orr, B.J.; Castillo Sanchez, V.M.; Chasek, P.; Crossman, N.D.; Erlewein, A.; Louwagie, G.; Maron, M.; Metternicht, G.I.; Minelli, S.; Tengberg, A.E.; Walter, S.; Welton, S. Land in balance: The scientific conceptual framework for Land Degradation Neutrality. *Environ. Sci. Policy*, 2018, 79, 25–35.
25. Chasek, P.; Akhtar-Schuster, M.; Orr, B.J.; Luise, A., Rakoto Ratsimba, H.; Safriel, U. Land degradation neutrality: The science-policy interface from the UNCCD to national implementation. *Environ. Sci. Policy* 2019, 92, 182–190.
26. Bell, R.W. Land restoration: Principles. In *Encyclopedia* of Soil Science, 2nd Ed.; Taylor and Francis: Abingdon, 2006; 978–981.
27. Hobbs, R. Restoration ecology. In *Encyclopedia* of Soil Science, 2nd Ed.; Taylor and Francis: Abingdon, 2006; 1513–1515.
28. Tongway, D.; Ludwig, J. Desertification: Reversal. In Encyclopedia of Soils Science, 2nd Ed.; Taylor & Francis: Abingdon, 2006; 465–467.
29. FAO/LADA Land Degradation Assessment in Drylands. Methodology and Results, available at http://www.fao.org/3/a-i3241e.pdf (accessed October 2019).
30. WOCAT. Where the Land Is Greener—Case Studies and Analysis of Soil and Water Conservation Initiatives Worldwide; Centre for Development and Environment (CDE): Bern, 2007.
31. Marques, M.J.; Schwilch, G.; Lauterburg, N.; Crittenden, S.; Tesfai, M.; Stolte, J.; Zdruli, P.; Zucca, C.; Petursdottir, T.; Evelpidou, N.; Karkani, A.; Asli-Yilmazgi, Y.; Panagopoulos, T.; Yirdaw, E.; Kanninen, M.; Rubio, J.L.; Schmiedel, U.; Doko, A. Multifaceted impacts of sustainable land management in Drylands; *A Rev. Sustain.* 2016, 8 (2), 177.
32. Geist, H.J.; Lambin, E.F. Dynamic causal patterns of desertification. *BioScience*, 2004, 54 (9), 817–829.
33. Reed, M.S.; Dougill, A.J.; Taylor, M.J. Integrating local and scientific knowledge for adaptation to land degradation: Kalahari rangeland management options. *Land Degrad. Dev.* 2007, 18 (3), 249–268.
34. Reynolds, J.F.; Stafford Smith, D.M.; Lambin, E.F.; Turner, B.L., II; Mortimore, M.; Batterbury, S.P.J.; Downing, T.E.; Dowlatabadi, H.; Fernandez, R.J.; Herrick, J.E.; Huber-Sannvald, E.; Leemans, R.; Lynam, T.; Maestre, F.T.; Ayarza, M.; Walker, B. Global desertification: Building a science for dryland development. *Science* 2007, 316 (5826), 847–851.
35. Liu, J.; Dietz, T.; Carpenter, S.R.; Folke, C.; Alberti, M.; Redman, C.L.; Schneider, S.H.; Ostrom, E.; Pell, A.N.; Lubchenco, J.; Taylor, W.W.; Ouyang, Z.; Deadman, P.; Kratz, T.; Provencher, W. Coupled human and natural systems. *Ambio* 2007, 36 (8), 639–649.

36. Scheffer, M.; Carpenter, S.R. Catastrophic regime shifts in ecosystems: Linking theory to observation. *Trends Ecol. Evol.* 2003, 18 (12), 648–656.
37. Suding, K.N.; Gross, K.L.; Houseman, G.R. Alternative states and positive feedbacks in restoration ecology. *Trends Ecol. Evol.* 2004, 19 (1), 46–53.
38. Suding, K.N.; Hobbs, R.J. Threshold models in restoration and conservation: A developing framework. *Trends Ecol. Evol.* 2009, 24 (5), 271–279.
39. Thomas, R.J.; Reed, M.; Clifton, K.; Appadurai, A.N.; Mills, A.J.; Zucca, C.; Kodsi, E.; Sircely, J.; Haddad, F.; von Hagen, C.; Mapedza, E.; Wolderegay, K.; Shalander, K.; Bellon, M.; Le, Q.B.; Mabikke, S.; Alexander, S.; Leu, S.; Schlingloff, S.; Lala-Pritchard, T.; Mares, V.; Quiroz, R. A framework for scaling sustainable land management options. *Land Degrad. Dev.* 2018, 29 (10), 3272–3284.
40. SER (Society for Ecological Restoration, Science & Policy Working Group). The SER Primer on Ecological Restoration, available at http://www.ser.org/ (accessed July 2010).
41. Adeel, Z.; King, C. Development of an Assessment Methodology for Sustainable Development of Marginal Drylands. *Proceedings of the Third Project Workshop for Sustainable Management of Marginal Drylands (SUMAMAD)*, Djerba, Tunisia, December 11–15, 2004; UNESCO-MAB: Paris, 2005; 13–22.
42. Bautista, S.; Alloza, J.A.; Evaluation of forest restoration projects. In Land Restoration to Combat Desertification: Innovative Approaches, Quality Control and Project Evaluation; Bautista, S.; Aronson, J.; Vallejo, V.R., Eds.; CEAM Foundation: Valencia, 2009, 47–72.
43. Bautista, S.; Orr, B.J.; Alloza, J.A.; Vallejo, V.R. Evaluation of the restoration of dryland ecosystems in the northern Mediterranean: Implications for practice. In Water in Arid and Semi-arid Zones. Advances in Global Change Research; Courel, M.F.; Schneier-Madanes, G., Eds.; Springer: Dordrecht, 2010; 295–310.
44. Ward, S.C. Restoration: Success and completion criteria. In Encyclopedia of Soil Science, 2nd Ed.; Taylor & Francis, 2006; 1516–1520.
45. Dale, V.H.; Beyeler, S.C. Challenges in the development and use of ecological indicators. *Ecol. Indic.* 2001, 1 (1), 3–10.
46. Jorgersen, S.E.; Xu, F.-L.; Salas, F.; Marques, J.C. Application of indicators for the assessment of ecosystem health. In Handbook of Ecological Indicators for Assessment of Ecosystem Health; Jorgensen, S.E., Costanza, R., Xu, F.-L., Eds.; Lewis Publishers, Inc.: Boca Raton, FL, 2005; 5–66.
47. Costantini, E.A.C.; Branquinho, C.; Nunes, A.; Schwilch, G.; Stavi, I.; Valdecantos, A.; Zucca, C. Soil indicators to assess the effectiveness of restoration strategies in dryland ecosystems. *Solid Earth* 2016, 7, 397–414.
48. Ruiz-Jaen, M.C.; Aide, T.M. Restoration success: How is it being measured? *Restor. Ecol.* 2005, 13 (3), 569–577.
49. Bastin, G.N.; Ludwig, J.A.; Eager, R.W.; Chewings, V.H.; Liedloff, A.C. Indicators of landscape function: Comparing patchiness metrics using remotely sensed data from rangelands. *Ecol. Indic.* 2002, 1 (4), 247–260.
50. Ludwig, J.A.; Bastin, G.N.; Chewings, V.H.; Eager, R.W.; Liedloff, A.C. Leakiness: A new index for monitoring the health of arid and semiarid landscapes using remotely sensed vegetation cover and elevation data. *Ecol. Indic.* 2007, 7 (2), 442–454.
51. Mayor, A.G.; Bautista, S.; Small, E.E.; Dixon, M.; Bellot, J. Measurement of the connectivity of run-off source areas as determined by vegetation pattern and topography: A tool for assessing potential water and soil losses in drylands. *Water Resour. Res.* 2008, 44, W10423.
52. Tongway, D.J.; Hindley, S. Landscape Function Analysis: Procedures for Monitoring and Assessing Landscapes; CSIRO Publishing: Brisbane, 2004.
53. Herrick, J.E.; van Zee, J.W.; Havstad, K.M.; Whitford, W.G. *Monitoring Manual for Grassland, Shrubland and Savanna Ecosystems*. USDA-ARS Jornada Experimental Range, Las Cruces, New Mexico. University of Arizona Press: Tucson, 2005.

54. Ludwig, J.A.; Tongway, D.J. Viewing rangelands as landscape systems. In Rangeland Desertification; Arnalds, O., Archer, S., Eds.; Kluwer Academic Publishers: Dordrecht, 2000; 39–52.
55. Rietkerk, M.; Dekker, S.C.; de Ruiter, P.C.; van de Koppel, J. Self-organized patchiness and catastrophic shifts in ecosystems. *Science* 2004, 305 (5692), 1926–1929.
56. Kéfi, S.; Rietkerk, M.; Alados, C.L.; Pueyo, Y.; Papanas-tasis, V.P.; El Aich, A.; de Ruiter, P.C. Spatial vegetation patterns and imminent desertification in Mediterranean arid ecosystems. *Nature* 2007, 449 (7159), 213–217.
57. Carpenter, S.R.; Turner, M.G. Hares and tortoises: Interactions of fast and slow variables in ecosystems. *Ecosystems* 2000, 3 (6), 495–497.
58. ICCD/COP8/CST. Report of the Fifth Meeting of the Group of Experts of the Committee on Science and Technology. Addendum: Methodologies for the Assessment of Desertification at Global, Regional and Local Levels. ICCD/COP(8)/CST/2/Add.6, 2008.
59. Zucca, C.; Wu, W.; Dessena, L.; Mulas, M. Assessing the effectiveness of land restoration interventions in drylands by multitemporal remote sensing – A case study in Ouled Dlim (Marrakech, Morocco). *Land Degrad. Develop.* 2015, 26, 80–91.
60. Turner, B.L., II; Kasperson, R.D.; Matson, P.A.; McCarthy, J.J.; Corell, R.W.; Christensen, L.; Eckley, N.; Kasperson, J.X.; Luers, A.; Martello, M.L.; Polsky, C.; Pulsipher, A.; Schiller, A. A framework for vulnerability analysis in sustainability science. *Proc. Natl. Acad. Sci. U.S.A.* 2003, 100 (14), 8074–8079.
61. Slocombe, D.S. Defining goals and criteria for ecosystem-based management. *Environ. Manage.* 1998, 22 (4), 483–493.
62. Folke, C.; Hahn, T.; Olsson, P.; Norberg, J. Adaptive governance of social–ecological systems. *Annu. Rev. Environ. Resour.* 2005, 30, 441–473.
63. Curtin, R.; Prellezo, R. Understanding marine ecosystem based management: A literature review. *Mar. Policy* 2010, 34 (5), 821–830.
64. Johannes, R.E. Integrating traditional ecological knowledge and management with environmental assessment. In Traditional Ecological Knowledge: Concepts and Cases. International Program on Traditional Ecological Knowledge; Inglis, J., Ed.; Canadian Museum of Nature: Ottawa, Canada, 1993.
65. Brush, S.B.; Stabinsky, D. Valuing Local Knowledge: Indigenous People and Intellectual Property Rights; Island Press: Washington, DC, 1996.
66. Emerson, K.; Orr, P.J.; Keys, D.L.; McKnight, K.M. Environmental conflict resolution: Evaluating performance outcomes and contributing factors. *Conflict Resolut. Q.* 2009, 27 (1), 27–64.
67. Orr, P.J.; Emerson, K.; Keys, D.L. Environmental conflict resolution practice and performance: An evaluation framework. *Conflict Resolut. Q.* 2008, 25 (3), 283–301.
68. Bautista, S.; Llovet, J.; Ocampo-Melgar, A.; Vilagrosa, A.; Mayor, A.G.; Murias, C.; Vallejo, V.R.; Orr, B.J. Integrating knowledge exchange and the assessment of dryland management alternatives – A learning-centered participatory approach. *J. Environ. Manag.* 2017, 195 (1), 35–45.
69. Pahl-Wostl, C.; Hare, M. Processes of social learning in integrated resources management. *J. Commun. Appl. Soc.* 2004, 14 (3), 193–206.
70. Schusler, T.M.; Decker, D.J.; Pfeffer, M.J. Social learning for collaborative natural resource management. *Soc. Nat. Resour.* 2003, 16 (4), 309–326.
71. Steins, N.A.; Edwards, V.M. Platforms for collective action in multiple-use common-pool resources. *Agric. Hum. Values* 1999, 16 (3), 241–255.
72. Rechkemmer, A. Societal impacts of desertification: Migration and environmental refugees. In Facing Global Environmental Change: Environmental, Human, Energy, Food, Health and Water Security Concepts; Brauch, H., BerghofStiftung, K., Eds.; Springer: Berlin, 2009.
73. Zdruli, P.; Trisorio Liuzzi, G., Eds. Managing Natural Resources through Implementation of Sustainable Policies. *Proceedings of the MEDCOASTLAND Euro-Mediterranean Conference*, Beirut, Lebanon, Jun. 25–30, 2006; IAM: Bari, 2007.

74. Zdruli, P.; Trisorio Liuzzi, G., Eds. Determining an Income-Product Generating Approach for Soil Conservation Management. *Proceedings of the MEDCOASTLAND International Seminar*, Marrakech, Morocco, Feb. 12–16, 2004; IAM: Bari, 2005.
75. Eswaran, H.; Berberoglu, S.; Cangir, C.; Boyraz, D.; Zucca, C.; Özevren, E.; Yazici, E.; Zdruli, P.; Dingil, M.; Dönmez, C.; Akça, E.; Çelik, I.; Watanabe, T.; Koca, Y.K.; Montanarella, L.; Cherlet, M.; Kapur, S. The anthroscape approach in sustainable land use. In Sustainable Land Management—Learning from the Past for the Future; Kapur, S., Eswaran, H., Blum, W.E.H., Eds.; Springer-Verlag: Berlin Heidelberg, 2011.

70

Erosion

Dennis C. Flanagan

Introduction

Soil erosion is the detachment or breaking away of soil particles from a land surface by some erosive agent, most commonly water or wind, and subsequent transportation of the detached particles to another location. Usually, erosion occurs when a fluid (air or water) moves into and/or across a soil surface. Fluid and sediment particle impact forces, shear forces, and turbulence act to detach and lift soil into the fluid flow that then transports the particles away (Figure 1). The force of gravity moves detached soil particles downward, while cohesive forces between soil particles resist detachment and transport. Physical and chemical dispersion can disrupt cohesion and break soil aggregates into smaller and more easily transported particles. At some time and location away from the initial point of detachment, the sediment particles will eventually move back down to a state of rest on a soil surface, in a process known as deposition or sedimentation.

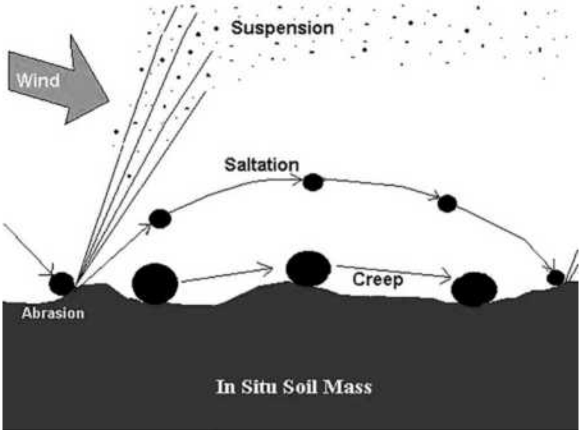


FIGURE 1 Soil erosion by wind, showing the three modes of movement (creep, saltation, and suspension).

Erosion by Wind

Erosion by wind occurs when wind speed exceeds a certain critical or threshold value. Soil particles can be detached and moved through suspension, saltation, or creep (Figure 1). Suspension usually lifts the smallest soil particles (clays, silts, organic matter) so high into the air mass that they are easily kept in motion and can travel for long distances. Soil particles that move by creep are larger sand grains and aggregates that stay in contact with the soil surface at almost all times—their motion is often through rolling and bouncing. Saltating soil particles are usually moderate in size and, once detached, move in trajectories up into the air and then back down to the soil surface. Saltating particles often cause further detachment through abrasion by striking the soil surface with sufficient momentum to dislodge additional soil particles from the in situ soil mass.

Erosion by Water

The most common types of soil erosion by water are sheet and rill erosion on upland areas, channel and gully erosion in small watersheds, and stream channel and bank erosion in larger catchments. Sheet erosion is caused by the action of raindrops (Figure 2) and shallow overland flows that remove a relatively uniform depth (or sheet) of soil. Because of the uniform nature of the soil loss, it is often difficult to detect and gauge the extent of damage caused by sheet erosion. On the other hand, rill erosion occurs in well-defined and visible flow concentrations or rills (Figure 3). Soil detachment in rills is large because of flow shear stress forces acting on the wetted perimeter of the rill channel (Figure 4). Once detached, larger sediment particles move as bedload, rolling and bouncing down slope with the flow, and are almost always in contact with the soil (or bed) surface. Smaller sediment particles (silts and clays) are much easier to transport and travel in the rill channels as suspended load. Rills are also the major pathways for transporting away sediment that is detached by sheet erosion (also known as interrill detachment).

By definition, rill channels are small enough to be obliterated by tillage and will not reform in exactly the same location. As one moves from smaller hillslopes to larger fields and watersheds, additional erosion processes come into play, because of the increasing amounts of runoff water. Gullies are incised erosion channels that are larger than rills and form in regions of large runoff flow concentration. Ephemeral gullies are a common type of erosion feature in many fields (Figure 5). They are small enough to be tilled over but will re-form in the same location owing to the convergent topography in small catchments. Runoff flows from large events can erode down through tilled soil layers, until a nonerodible layer is reached, and then the ephemeral gully channel will widen and soil detachment will decrease. Classical

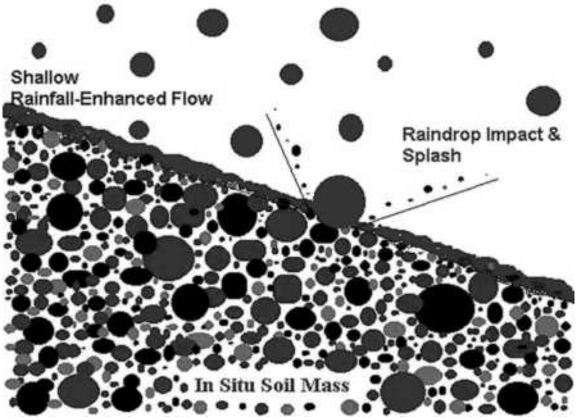


FIGURE 2 Soil detachment by raindrop impact and shallow flow transport.

FIGURE 3 Rill erosion, which is caused by concentration of flowing water, forms easily recognizable regions of detachment on a soil surface. (Courtesy M. Huhnke, Oklahoma State University.)

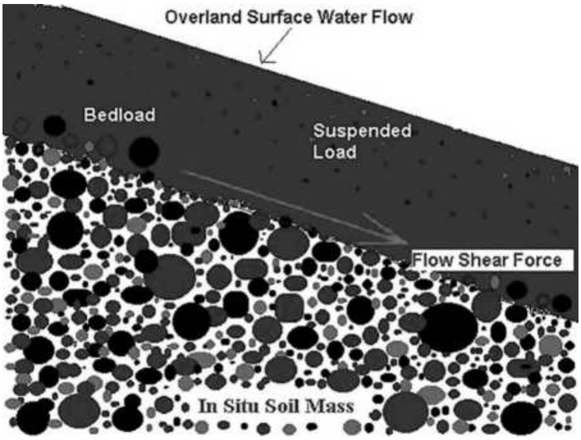


FIGURE 4 Soil detachment and transport in rills are largely because of flow shear forces.

gullies are larger erosion features that cannot normally be tilled across (Figure 6). The physical processes in classical gullies include other factors such as headcutting, seepage, sidewall sloughing, and clean-out of fallen sidewall materials.

As the size of watersheds increases further, and streams increase in size and become perennial (because of subsurface water flows from springs and aquifers), the erosion processes in play change as well. Stream and channel erosion at these larger scales can include scouring of the channel beds as well as the contributions from the channel banks. Areas in streams may be in states of degradation, in which active detachment is lowering the level of the channel bed, or they may be in states of aggradation, in which sediment deposition is raising the bed level.

Gravity-Induced Erosion

There are also less frequent but more extreme forms of gravity-induced erosion on steep slopes from saturated soils that can be exacerbated by events such as earthquakes. Large masses of land can slowly or rapidly slide down hills when the cohesive forces holding them in place fail (landcreep, landslide, debris flow, etc.).

FIGURE 5 Typical ephemeral gully located in a soybean field in Indiana.

FIGURE 6 Classical gullies in the Loess Plateau of China. Terrace farming is being used to stabilize some of the hillslopes.

These types of erosion events typically occur when large rainfall or snowmelt water depths saturate soil profiles and weaken their resistance to slip.

Erosion Assessment

Erosion is a serious problem within the United States and throughout the world. In 2015, the United States Department of Agriculture (USDA)-Natural Resources Conservation Service estimated that about 1.7 billion tons of soils are lost each year from nonfederal rural croplands because of sheet and rill erosion by water and erosion by wind.[1] Also, over 20% of cropland in the United States is eroding at excessive rates. These estimates are on the low side because erosion of other types (e.g., gully) and at other locations (urban lands, federal lands) were not included in this inventory. Throughout the world, FAO estimates that 16% of the total land area (21,960,000 km^2 of 134,907,000 km^2) is subject to significant risk of soil erosion.[2] In Asia, South America, and Africa, soil erosion rates are highest at an estimated average of 30–40 t/ha/yr, while in Europe and North America average rates are somewhat lower at about 17 tons/ha/yr.[3,4] A sustainable rate of soil loss (rate of soil loss is equal to rate of soil formation) is thought to be about 1 t/ha/yr.[3]

Erosion assessment can be a difficult task to perform in the field, and monitoring of soil lost and transported by wind or water can be expensive and prone to measurement errors. Gullies are easy to recognize, while soil lost to sheet and rill erosion is hard to gauge. Sheet and rill erosion may be occurring on hill slopes that are adjacent to a gully and may actually contribute more sediment to runoff water than the gully itself. Visual assessment of rates of wind erosion losses can be even more difficult to perform. Mathematical equations or sets of equations have been developed and used since the mid-1900s to estimate the rates of soil loss caused by wind ("Wind Erosion Equation"[5]) or water ("Universal Soil Loss Equation"[6,7]). More recently, computer models are being applied to simulate soil erosion processes and to estimate detachment, transport, and deposition of sediment.[8,9,10,11]

Erosion Impacts

Erosion has a range of impacts, both on-site and off-site. Soil loss removes fertile topsoil, organic matter, and nutrients, thus decreasing the tilth, water-holding capacity, and general productivity of a soil for on-site agricultural production. Regions of detachment can expand to dislodge and remove small crop seedlings, while regions of deposition can bury and kill small plants. In the case of wind erosion, the erosion process can damage fragile young seedlings through abrasion of plant tissue. When excessive detachment occurs, such as is the case with gully erosion, whole sections of fields may be destroyed or may become inaccessible to farmers and their equipment. Eroded sediment can cause a number of off-site problems, including deposition along windbreaks, ditches, and waterways. The deposited sediment may require costly dredging and removal operations. Nutrients and pesticides associated with sediments can also contaminate air and water bodies. Erosion by wind can cause massive dust storms that blind drivers and cause accidents, and sand particles can abrade and damage painted surfaces on buildings and vehicles. Some recent estimates are that the cost of combined on-site and off-site effects from soil erosion in the United States is as high as $44 billion per year.[3]

Erosion Control and Soil Conservation

Many nations have created government agencies or organizations to specifically deal with soil erosion problems and to interact with landowners to get conservation practices implemented on the landscape. In the United States, the Natural Resources Conservation Service assists in the implementation of soil conservation practices on agricultural lands, the Forest Service manages sediment delivery from forests and timber harvest roads, and the Bureau of Land Management manages soil loss on range and grazing lands. The Department of Defense is responsible for managing erosion and off-site sediment delivery from lands that it uses for military training activities. However, in some countries, efforts to address and minimize erosion problems are nonexistent or severely limited because of the poor economic conditions, failure to recognize the erosion threat, and/or the extreme magnitude of the soil erosion.

A variety of soil conservation methods are available that can be applied on a landscape to minimize erosion problems caused by wind or water. Wind erosion can be controlled through the use of windbreaks, crop residues, and tillage to induce significant surface roughness. Control procedures for erosion by water need to be determined, based upon the types of active erosion processes. For example, if sheet and rill erosion is a major problem, then some type of conservation tillage practice that leaves large amounts of crop residues intact on the soil surface may be appropriate. However, if the water erosion problem is because of the large amounts of surface runoff concentrating in a field and forming an ephemeral gully, then crop residues may not be adequate; instead, permanent vegetative cover in a grass waterway may be necessary, along with appropriate engineering structures (e.g., drop-box). Erosion prediction models can be used to assist in selecting and designing appropriate conservation practices.

Through the use of proper conservation planning and application of appropriate soil conservation methods, most erosion problems can be minimized or eliminated. This is critically important if the soil resource is to be preserved for continuous use in food and fiber production for current and future generations.

Conclusions

Erosion is a natural process of soil detachment and removal that can be greatly influenced through human activities (agriculture, construction, timber-harvesting, etc.). Use of proper erosion prediction technology and appropriate erosion control methodologies is critical if we are to sustain the soil resource for use by future generations.

References

1. U.S. Department of Agriculture. *Summary Report: 2015 National Resources Inventory;* Natural Resources Conservation Service: Washington, DC, and Center for Survey Statistics and Methodology, Iowa State University: Ames, Iowa, 2018; Available at: https://www.nrcs.usda.gov/Internet/FSE_DOCUMENTS/nrcseprd1422028.pdf (accessed April 2019).
2. Food and Agricultural Organization of the United Nations. *World Soil Resources Report 90;* FAO Land and Water Development Division: Rome, 2000; Available at: http://www.fao.org/3/a-x7126e.pdf (accessed April 2019).
3. Pimental, D.; Harvey, C.; Resosudarmo, P.; Sinclair, K.; Kurz, D.; McNair, M.; Crist, S.; Shpritz, L.; Fitton, L.; Saffouri, R.; Blair, R. Environmental and economic costs of soil erosion and conservation benefits. *Science* **1995**, *267,* 1117–1123.
4. Barrow, C.J. *Land Degradation*; Cambridge University Press: Cambridge, 1991.
5. Woodruff, N.P.; Siddoway, F.H. A wind erosion equation. *Proc. Soil Sci. Soc. Am.* **1965**, *29,* 602–608.
6. Wischmeier, W.H.; Smith, D.D. *Predicting Rainfall Erosion Losses—A Guide to Conservation Planning*; Agriculture Handbook No. 537; USDA: Washington, DC, 1978.
7. Renard, K.G.; Foster, G.R.; Weesies, G.A.; McCool, D.K.; Yoder, D.C. *Predicting Soil Erosion by Water: A Guide to Conservation Planning with the Revised Universal Soil Loss Equation (RUSLE)*; Agriculture Handbook No. 703; USDA: Washington, DC, 1997; 384 pp.
8. Flanagan, D.C.; Nearing, M.A., Eds. *USDA-Water Erosion Prediction Project (WEPP) Hillslope Profile and Watershed Model Documentation;* NSERL Report No. 10; National Soil Erosion Research Laboratory, USDA-Agricultural Research Service: West Lafayette, IN, 1995; 298 pp.; Available at: https://www.ars.usda.gov/midwest-area/west-lafayette-in/national-soil-erosion-research/docs/wepp/wepp-model-documentation/ (accessed April 2019).
9. Flanagan, D.C.; Gilley, J.E.; Franti, T.G. Water Erosion Prediction Project (WEPP): Development history, model capabilities, and future enhancements. *Trans. ASABE* **2007**, *50* (5), 1603–1612.
10. Hagen, L.J. A wind erosion prediction system to meet user needs. *J. Soil Water Conserv.* **1991**, *46* (2), 106–111.
11. Wagner, L.E. A history of wind erosion prediction models in the United States Department of Agriculture: The Wind Erosion Prediction System (WEPS). *Aeolian Res.* **2013**, *10*, 9–24.

71

Erosion by Water: Erosivity and Erodibility

Peter I.A. Kinnell

Introduction

Conceptually, rainfall erosivity is the capacity of rain to produce erosion, whereas soil erodibility is the susceptibility of the soil to be eroded. Historically, the terms erosivity and erodibility were originally associated with the *R* and *K* factors in the Universal Soil Loss Equation (USLE),

$$A = R\ K\ L\ S\ C\ P \tag{1}$$

where *A* is the long-term (e.g., 20 years) annual average soil loss per unit area from sheet and rill erosion, *R* is the rainfall (erosivity) factor defined as the average annual value of the product of the total storm kinetic energy (*E*) and the maximum 30-min intensity (I_{30}, twice the maximum amount of rain that falls in any 30-min period during a storm), *K* is the soil (erodibility) factor, *L* is the slope length factor, *S* is the slope gradient factor, *C* is the crop (vegetation) and crop management factor, and *P* is the conservation support practice factor.[1] Numerically, soil erodibility is the mass of soil eroded per unit of the erosive index. This means that numerical values of *K* can only be used when *R* is as it was originally defined as the average annual value of the product of *E* and I_{30}.

The USLE was designed to predict sheet and rill erosion from field-sized areas and only *R* and *K* have units. The *L*, *S*, *C*, and *P* factors each have values of 1.0 for the so-called "unit" plot, a bare fallow area 22.1 m long on a 9% slope with cultivation up and down the slope. Consequently, the soil loss for the "unit" plot (A_1) is given by

$$A_1 = R\ K \tag{2}$$

and, for any other situation,

$$A = A_1\ L\ S\ C\ P \tag{3}$$

Although, traditionally, K is calculated by dividing A_1 by R, K can be perceived as acting similarly to the regression coefficient in the direct relationship between event soil loss on the unit plot (A_{1e}) and EI_{30}. However, regression analysis of the relationships between storm soil losses on the unit plot and EI_{30} tends to produce regression coefficients that may differ appreciably from those calculated using the traditional method. Consequently, the traditional approach should always be used in determining K values from runoff and soil loss plot data.

To reduce the need to run long-term experiments to determine K values for soils where K is unknown, Wischmeier[2] developed a nomograph for determining K from soil properties for soils with less than 70% silt in the United States. Alternatively, K values in customary U.S. units for soils where the nomograph can be used may be obtained using

$$K = \left(2.1\ X_1^{1.14}10^{-4}\left(12 - X_2\right) + 3.25\left(X_3 - 2\right) + 2.5\left(X_4 - 3\right)\right)/100 \tag{4}$$

where X_1 is % silt multiplied by 100 – % clay, X_2 is % organic matter, X_3 is the soil structure code used in the U.S. soil classification, and X_4 is the profile permeability code. A number of other equations have been developed for soils at various geographic locations (e.g., El-Swaify,[3] Young and Mutchler,[4] Loch,[5] and Zhang et al.[6]), but Eq. (4) is frequently used outside the United States without being validated for the soils involved. Division of the right-hand side of Eq. (4) by 7.59 will yield K values in SI units of t h /MJ/mm.

The two-staged mathematical approach shown by Eqs. (2) and (3) results from the fact that the USLE is an empirically based model. It was developed in the 1960s and 1970s from more than 10,000 plot years of data. Mathematically, the Revised Universal Soil Loss Equation (RUSLE[7]) uses Eqs. (2) and (3) in the same way as the USLE, but changes were made to how some of the factors in the model are calculated. Originally, in the USLE, the events used to calculate R were restricted to those that produced more than 12.5 mm of rain or at least 6.25 mm of rain in 15 min. That rule was abandoned in the RUSLE when R values were calculated for the western part of the United States because it was argued that the rule had no appreciable effect on R values. Yu[8] noted that changing the threshold to zero increased the R factor by 5% in the tropical region of Australia. However, the abandonment of the rule in the RUSLE failed to recognize that that rule had been put in place as a means of discounting storms that tended to be nonerosive because they produced no runoff.

Variants of the USLE

Eq. (2) operates on the assumption that a direct linear relationship exists between event erosion (A_e) and EI_{30}. Although this assumption is appropriate at some geographic locations, it is not appropriate at others (Figure 1). Soil measured as a loss from the plots used to develop the USLE was discharged in runoff. When runoff occurs, the amount of soil discharged (Q_s) in the runoff can be considered to be the product of the amount of water discharged (Q_w) and the sediment concentration (c_s), the amount of soil per unit quantity of water,

$$Q_s = Q_w c \tag{5}$$

It follows from Eq. (5) that, in the USLE, the sediment concentration is directly proportional to EI_{30} divided by runoff amount. However, it has been observed[10] that the sediment concentration for an event at some geographic locations is correlated to the product of the kinetic energy per unit quantity of rain and I_{30} (Figure 2). This results in an event erosivity index (R_e) that is given by

$$R_e = Q_{Re} EI_{30} \tag{6}$$

where Q_{Re} is the runoff ratio (runoff amount divided by rainfall amount) for the event. This index is more effective in accounting for event soil losses at some locations where the EI_{30} index does not work

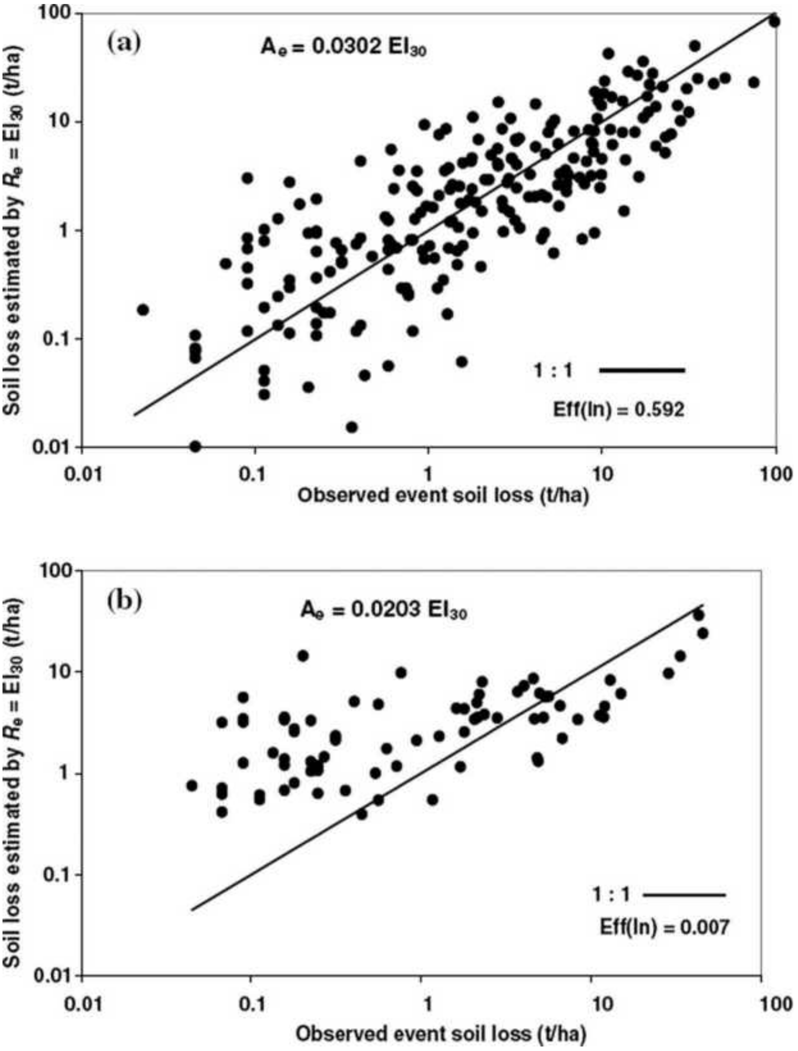


FIGURE 1 The relationships between event erosion (A_e) and EI_{30} obtained on bare fallow plots in the United States: (a) Plot 5 Experiment 3 at Holly Springs, Mississippi, and (b) Plot 5 Experiment 1 at Morris, Minnesota. Eff(ln) is the Nash–Sutcliffe[9] efficiency factor for logarithmic transforms of the data.

well (Figure 3). Because R_e is not equal to EI_{30}, USLE K factor values cannot be used in conjunction with Eq. (6). Soil erodibilities associated with Eq. (6) are calculated dividing A_1 by the average annual sum of the product of Q_{Re} and EI_{30}.

A number of other event erosivity indices have been proposed. Williams[11] proposed a modification of the USLE to predict event sediment yield (SY_e):

$$SY_e = 11.8\left(q_e q_p\right)^{0.56} K\ L\ S\ C_e\ P_e \tag{7}$$

where q_e is the volume of runoff (m^3) for the event; q_p is the peak flow rate (m^3/sec); K, L, and S are standard USLE factors; and C_e and P_e are event C and P factors, respectively. This model is commonly known as the Modified Universal Soil Loss Equation (MUSLE). It should be noted the value of 11.8 in

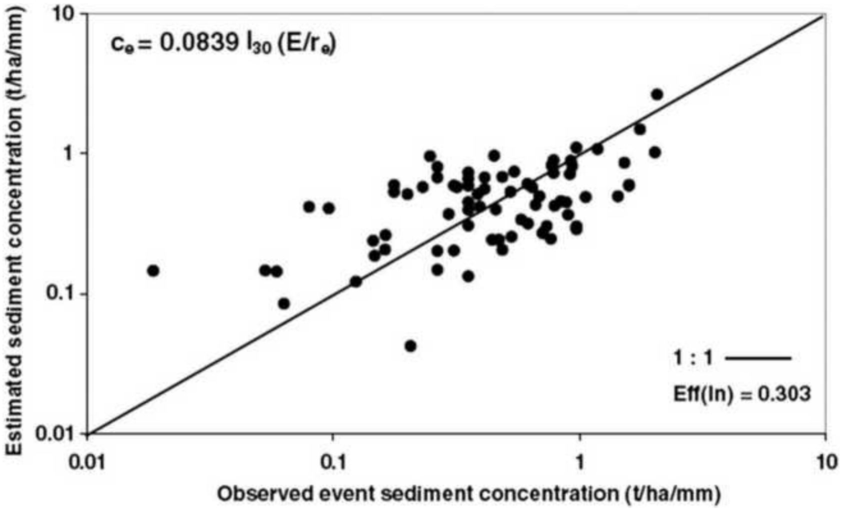


FIGURE 2 The relationship between event sediment concentration (c_e) and the product of I_{30} and the kinetic energy per unit quantity of rain (E/r_e) for Plot 5 Experiment 1 at Morris, Minnesota.

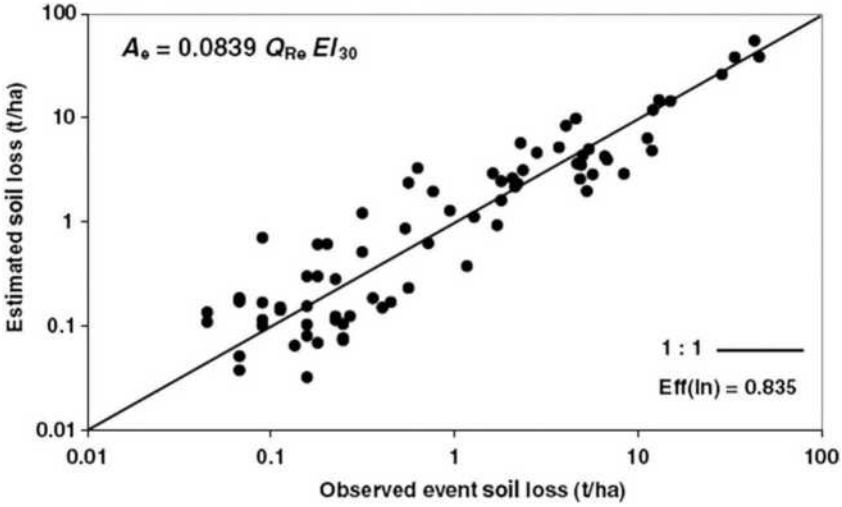


FIGURE 3 The relationship between event erosion and the product of event runoff ratio (Q_{Re}) and EI_{30} for Plot 5 Experiment 1 at Morris, Minnesota.

Eq. (7) was generated for the specific situation where Williams[11] obtained measured data. Consequently, the general applicability of this value is questionable. In Agricultural Policy/Experimental eXtender Model (APEX),[12]

$$\mathrm{SY_e} = X_e K\ L\ S\ C_e\ P_e[\mathrm{RKOF}] \tag{8}$$

X_e is selected from

$$X_e = EI_{30} \tag{9a}$$

$$X_e = 1.586\left(q_e q_{pe}\right)^{0.56} \mathrm{DA}^{0.12} \tag{9b}$$

$$X_{\mathrm{e}} = 0.65\ EI_{30} + 0.45\left(q_{\mathrm{e}}q_{\mathrm{pe}}\right)^{0.33} \tag{9c}$$

$$X_{\mathrm{e}} = 2.5\left(q_{\mathrm{e}}q_{\mathrm{pe}}\right)^{0.5} \tag{9d}$$

$$X_{\mathrm{e}} = 0.79\left(q_{\mathrm{e}}q_{\mathrm{pe}}\right)^{0.65}\ \mathrm{DA}^{0.009} \tag{9e}$$

$$X_{\mathrm{e}} = b_5 q_{\mathrm{e}}^{\ b6}\ q_{\mathrm{e}}^{\ b7}\ \mathrm{DA}^{b8} \tag{9f}$$

where DA is drainage area expressed in hectares, *b*6–*b*8 are user-selected coefficients,[12] and RKOF is the coarse fragment factor as defined by Simanton et al.[13] However, Eqs. (7) and (8) with Eqs. (9b–9f) all use USLE *K* factor values and not ones that are associated with the different erosivity indices used and so do not conform with the mathematical modeling rules upon which the USLE model is based. Consequently, the MUSLE and APEX models are not valid variations of the USLE. In addition, any model that uses event erosivity index values that are calculated using runoff from a vegetated area or any area that is not cultivated up and down the slope will violate the mathematical rules if it uses USLE *C* and *P* factor values.

Mathematical models like the USLE and its derivatives are largely designed to aid management decisions and operate at a level where spatial and temporal variations in the various forms of erosion (splash erosion, sheet erosion, rill erosion, interrill erosion) are not considered in any appreciable detail. However, rill erosion is driven by flow energy, while sheet and interrill erosion are associated more closely with rainfall kinetic energy. In order to better account for this, Onstad and Foster[14] used the equation

$$R_{\mathrm{e}} = 0.5EI_{30} + 0.5\alpha\left(q_{\mathrm{e}}q_{\mathrm{p}}\right)^{0.333} \tag{10}$$

Eq. (10) indicates that the sediment concentration for an event (c_{e}) in the data set considered by Onstad and Foster[14] is given by

$$c_{\mathrm{e}} = 0.5\left[EI_{30}/q_{\mathrm{e}}\right] + 0.5\left[\alpha\left(q_{\mathrm{p}}\right)^{0.333}/q_{\mathrm{e}}^{\ 0.666}\right] \tag{11}$$

The value of α in Eq. (10) was set so that the long-term average value of R_{e} produced by Eq. (10) was the same as the long-term average value produced when $R_{\mathrm{e}} = EI_{30}$. Given that *K* is, by definition, the amount of soil loss per unit of *R* when $R_{\mathrm{e}} = EI_{30}$, this enabled Eq. (10) to be used with USLE *K* values to predict soil loss from bare fallow areas. Although Williams et al.[15,16] indicated that Onstad and Foster[14] was the source of Eq. (9c), there is no provision in APEX to do the same and ensure that USLE *K* values can be used.

Bagarello et al.[17] observed that

$$R_{\mathrm{e}} = \left(Q_{\mathrm{Re}}EI_{30}\right)^{\beta} \tag{12}$$

with $\beta > 1.0$ applied to soil losses from a number of simultaneously operating plots of different length (λ) established at the experimental station of Sparacia, Sicily. Subsequent analysis [Bagarello et al.[18]] established that $\beta > 1.0$ appeared to be the result of an interaction between event runoff (Q_{e}) and slope length (λ) on sediment concentration so that

$$R_{\mathrm{e}} = Q_{\mathrm{Re}}EI_{30}Q_{\mathrm{e}}^{\ 0.0207\lambda} \tag{13}$$

when λ is in meters.

Arguably, from a physical viewpoint, since slope length and gradient influence the erosive stress, they can be considered to be factors that influence erosivity, but the USLE model is not designed to model the physical processes themselves. It is designed to predict rainfall erosion based on climate, soil, topography, crops, and management factors. Normally, in the USLE model, the effect of slope length on soil loss is expressed through the L factor, but in the case of Eq. (13), L would remain at 1.0 irrespective of variations in slope length. The positive effect of runoff on sediment concentration may have been associated with the fact that the data were collected on a 15% slope. Rainfall erosion on much lower slope gradients is not associated with as high an erosion stress from runoff as likely to have been the case at the Sparacia site. In fact, the RUSLE makes provision for the cushioning effect of water depth on the energy of raindrop impact on low slopes.

Because, as noted above, R_e is not equal to EI_{30}, USLE K factor values cannot be used in conjunction with Eqs. (12) and (13). Soil erodibilities associated with Eq. (12) are calculated dividing A_1 by the average annual sum of $(Q_{Re}\, EI_{30})^{\beta}$. β varies between geographic locations in Italy, and consequently, the units for the associated soil erodibilities vary geographically. That is inconsistent with the concept of erodibility in the USLE modeling system. However, Kinnell (2018) showed that Eq. (12) can be rewritten as

$$R_e = a_1 \left(Q_{Re} EI_{30}\right)^{\beta} \tag{X1}$$

where a_1 is the ratio of the average annual value of $(Q_{Re}\, EI_{30})$ to the average annual value of $(Q_{Re}\, EI_{30})^{\beta}$. a_1 and β are related mathematically to each other, a_1 being equal to 1.0 when β equals 1.0, less than 1.0 when β is greater than 1.0, and greater than 1.0 when β is less than 1.0. When Eq. (X1) is used, soil erodibilities determined by dividing A_1 by average annual value of $(Q_{Re}\, EI_{30})$ apply at all geographic locations irrespective of the value of β.

Rainfall Kinetic Energy

Data on storm rainfall kinetic energy are not measured at many geographic locations, and often, kinetic energy values are obtained indirectly. If data on storm rainfall intensities are available, then it is common for storm kinetic energies to be estimated from rainfall kinetic energy–intensity relationships. These relationships are based on raindrop-size data collected during rainfall events at a geographic location assumed to have rainfall characteristics that are consistent with those at the geographic location of interest. Various techniques have been employed to do this, and these have generated rainfall intensity–kinetic energy relationships which are often used at locations that lie outside the climatic zones where the measurements were originally made. In the USLE, the relationship recommended between the kinetic energy per unit quantity of rain (e) and rainfall intensity (I) for the United States is expressed by

$$\begin{gathered} e = 916 + 331 \log_{10} I, \\ \text{for } I < 3 \ \text{ in./h} \end{gathered} \tag{14a}$$

$$\begin{gathered} e = 1074, \\ \text{for } I \geq 3 \ \text{ in./h} \end{gathered} \tag{14b}$$

where e is in units of foot-tons per acre per inch and I is in inches per hour, following analysis of drop-size data collected by Law and Parsons.[19] In the RUSLE, Eq. (14) is replaced by

$$e = 1099\left(1 - 0.72 \exp\left[-1.27I\right]\right) \tag{15}$$

whose metric equivalent is

$$e_m = 0.29\left(1 - 0.72 \exp\left[0.05I_m\right]\right) \tag{16}$$

where e_m has units of megajoules per hectare per millimeter and I_m has units of millimeters per hour. Eqs. (15) and (16) use the mathematical form proposed by Kinnell.[20,21] Yu[8] observed that replacing

Eq. (14) by Eq. (15) reduced the R factor by about 10% in the tropical region of Australia. A number of other mathematical equations have been reviewed by van Dijk et al.[22] Some of these produce negative values of rainfall kinetic energy at low rainfall intensities. In RUSLE2 (USDA, 2013),

$$e_{\mathrm{m}} = 0.29\left(1 - 0.72\ \exp\left[0.08 I_{\mathrm{m}}\right]\right) \tag{X2}$$

Both Eq. (16) and Eq. (X2) produce response curves that increase non-linearly from low values at low intensities with e_{m} values remaining close to 0.29 MJ/ha/mm for intensities beyond 75–100 mm/h, but Eq. (X2) gives higher values of e_{m} below 75 mm/h. Eq. (X2) was adopted because it produced results that were more consistent with the USLE equation (Nearing et al., 2017). However, short-term values measured during rainstorms vary greatly about the values produced by Eq. (X2) or any other rainfall intensity-e_{m} relationship. Also, although Eqs. (15), (16), and (X2) were developed for use in the United States, they have been used in other parts of the world without validation. In reality, storm kinetic energies calculated from rainfall intensity–kinetic energy relationships are often not close to actual storm kinetic energies but are, in effect, numerical values that are biased toward high-intensity rainfall at the expense of low-intensity rainfall.

In many geographic locations, there is a lack of rainfall intensity data so that it is not possible to determine storm kinetic energies using rainfall kinetic energy–intensity relationships. Event or daily rainfall amounts are more commonly recorded. One approach that has been used in a number of geographic areas such as Canada,[23] Finland,[24,25] Italy,[26] and Australia[27] considers that EI_{30} is related to a power of event rainfall amount (X_{p}),

$$EI_{30} = a_1 X_{\mathrm{P}}^{b1} \tag{17}$$

where a_1 and b_1 are empirical constants. The value of a_1 may show seasonal variation.[25,28] Given that, in the context of the criteria associated with the USLE/RUSLE model, daily rainfall provides a reasonable proxy for storm rainfall amount[29]; Eq. (17) provides a practical approach to extending observed EI_{30} values to areas where appropriate rainfall intensity data are lacking.

Although determining R is a primary requirement for modeling soil loss using the USLE approach, temporal variations in erosivity during the year are required in order to account for the effects of crops and crop management on erosion. $C = 1$ when soil is bare, and $C = 0$ when vegetation completely protects the soil against erosion. Consequently, how erosivity and the protective effect of crops vary during the year are factors that are taken into account in determining C for various agricultural systems. In RUSLE2, a factor known as "erosivity density" is used (USDA, 2013). Erosivity density is EI_{30} divided by storm rainfall amount. While rainfall storm rainfall amounts are highly variable in time and space, erosivity densities are more readily mapped in time and space than EI_{30} itself. Because storm energy is highly influenced by the amount of rain energy that occurs when I_{30} is measured, storm erosivity densities tend to be highly influenced by variations in I_{30}.

More Process-Based Models

Although the USLE/RUSLE is the most widely used method of predicting soil losses from the land worldwide, rainfall erosion results from various forms of erosion (splash erosion, sheet erosion, rill erosion, interrill erosion) that are driven by different forces so that there is no absolute measure of either rainfall erosivity or soil erodibility. Consequently, more process-based models have been developed in order to predict the contributions of the various forms of erosion more directly. Often, the forms vary in a topographic sequence with splash erosion dominating erosion at the upper end of a slope and sheet erosion further down before areas of rill and interrill erosion. Particles detached at the top of the slope may be transported by a number of different transport mechanisms before being finally discharged from the eroding area. The Water Erosion Prediction Program in the United States generated the WEPP model, a more process-based model than the USLE/RUSLE designed to model the spatial contributions of rill and interrill erosion in agricultural landscapes.[30] Flow shear stress (τ) is used as the erosivity factor in rill erosion model,

$$D_r = k_r(\tau - \tau_c)(1 - q_{sr}/T_r) \tag{18}$$

where D_r is rill detachment, k_r is the rill erodibility factor, τ_c is the critical shear stress that has to be exceeded before detachment occurs, q_{sr} is the sediment load in the flow, and T_{cr} is the sediment load at the transport limit. For erosion to occur, particles must be plucked from within the soil surface where they are held by cohesion and interparticle friction. Detachment is the term used to refer to this process. For detachment in rills to occur, the flow must have shear stress that exceeds τ_c. Also, for erosion to occur, detached particles must be transported away from the site of detachment. Flows are known to have a limited capacity to transport soil material and that limit is represented by T_r in Eq. (18). Consequently, the term $1-q_{sr}/T_r$ causes $D_r = 0$ when the sediment load in the rill equals the transport limit. As a result, D_r may vary along the length of a rill. Interrill erosion contributes to q_{sr} so that rill erosion may be completely suppressed if the discharge of sediment from the interrill areas is high enough.

Originally, the erosivity factor in the WEPP interrill model was assumed to be the square of rainfall intensity (I) so that

$$D_i = k_i I^2 \tag{19}$$

where D_i is interrill detachment and k_i is the interrill soil erodibility factor. A series of rainfall simulation experiments[31] were undertaken to determine k_i values for soils in the United States. Subsequent analysis[32] of the data generated by these experiments established that it was more appropriate to use

$$D_i = k_{iq} q_i I \tag{20}$$

where q_i is the runoff rate from the interrill area and k_{iq} is the interrill erodibility factor. The k_i values determined using Eq. (19) cannot be used in Eq. (20).[32]

D_i is determined using data on the amounts of soil material discharged from interrill areas. The use of the term "detachment" in relation to D_i is not absolutely correct because the transport processes involved in moving soil material over interrill areas are not 100% efficient. Also, detachment varies with raindrop size and velocity but, in using Eq. (20), that effect is ignored. However, provision is made in WEPP to take account of differences in drop-size distribution associated with different rainfall simulators in experiments designed to determine values of k_{iq}.

Eqs. (18) and (20) provide the basic equations that enable WEPP to model the contributions of rill and interrill erosion to soil loss more directly than when the USLE/RUSLE model is used but, in WEPP, the spatial distribution of rill and interrill areas needs to be well defined. WEPP was originally designed to operate in situations where ridge tillage is used. With ridge tillage, interrill erosion is well defined as being on the ridge slopes and rill erosion as being in the furrows. In other situations, the distributions of rill and interrill areas are more difficult to specify. In rangelands, broad areas of surface water flow occur more commonly than rills so that splash and sheet erosion dominate. An alternative to the WEPP cropland interrill erosion model,

$$D_{SS} = k_{SS} I^{1.05} q_{SS}{}^{0.59} \tag{21}$$

where D_{ss} is the "detachment" associated with splash and sheet erosion, k_{ss} is the erodibility factor, and q_{ss} is the runoff rate from the associated eroding surface, has been developed for use in rangelands.[33]

EUROSEM is another erosion model that is more process based than the USLE/RUSLE. In EUROSEM, detachment by raindrop impact (D_r) is modeled using

$$D_{\mathrm{r}} = \frac{k_D}{\rho_{\mathrm{s}}} KE \exp(-zh) \tag{22}$$

where k_D is an index of detachability that is determined experimentally, ρ_{s} is particle density, KE is the rainfall kinetic energy of the raindrops impacting the ground, z is a factor that varies with soil texture, and h is the mean depth of the water layer on the soil surface. Detachment by flow is modeled using

$$D_{\mathrm{f}} = \beta w\ v_{\mathrm{s}}(C_{\mathrm{T}} - C) \tag{23}$$

where β is an index of detachability, w is flow width, v_{s} is particle settling velocity, C is the sediment concentration, and C_{t} is the sediment concentration at the transport capacity of the flow.

The coefficients used in Eq. (22) result from measurement of soil material transported by splash under ponded conditions in the experiments of Torri et al.[34] where increases in water depth were observed to reduce splash erosion. The effect of water depth on splash erosion results from (1) dissipation of raindrop energy in the water layer and (2) the effect of water depth on splash trajectories. Considering that splash does not transport 100% of the material detached by raindrops impacting water, Eq. (22) does not actually model detachment. Also, erosion by rain-impacted flows where sediment is transported by rolling, saltation, and suspension in the flow is the real focus of Eq. (22), and the effect of flow depth on erosion by rain-impacted flows is quite different from its effect on splash erosion.[35]

As a general rule, the values of soil erodibility factors have to be determined experimentally, although there are cases where they are predicted from soil properties. In areas where detachment results from raindrop impact, erodibility is affected by modification of the soil surface generated by the impacts. Raindrop impacts on surfaces not covered by water may break soil aggregates and generate surface crusts that affect soil erodibility. In areas where soil crusts occur, particles are held more tightly within the soil surface than in areas that are not crusted and this reduces detachment. With splash erosion, loose particles sit and wait on the soil surface between drop impacts. The transport efficiency of splash increases with slope gradient but, over time, a layer of loose particles builds up on the soil surface, and energy has to be expended in moving them before detachment can occur. Consequently, this layer provides a degree of protection against detachment. When raindrops impact a soil surface covered by water, the protection provided by loose particles sitting on the surface is in addition to that provided as the result of the dissipation of raindrop energy in the water layer.

Although splash erosion may dominate erosion for considerable periods during a rainfall event, rain-impacted flows are usually more important in moving soil material across the soil surface because the transport mechanisms in rain-impacted flows are much more efficient than splash transport. In rain-impacted flows, particles move across the soil surface by rolling, saltation, and complete suspension. Shallow low-velocity flows often do not have the capacity to cause particles to move by rolling and saltation by themselves, but rolling and saltation can be stimulated to occur when raindrops impact the soil surface through the flow. Under these circumstances, each rolling or saltation event is of limited duration and is associated with individual raindrop impacts. Figure 4 illustrates how raindrop and flow factors influence the detachment and transport processes associated with the erosion of fine particles, silt, and sand by rainfall. Particles moving by raindrop-induced rolling and saltation move across the surface at rates that depend on raindrop size, impact frequency, particle size, and density and the velocity of the flow. Consequently, depending on the rain and flow conditions, particles in rain-impacted flows travel across the surface at virtual velocities that vary from near zero to the velocity of the flow. In effect, particles are winnowed from the soil surface at various rates. Large gravel particles may not move at all and so, over time, soils that have high gravel contents may become highly resistant to erosion through the formation of erosion "pavements." Particles moving by raindrop-induced rolling and saltation sit on the soil surface between drop impacts and so provide a degree of protection against detachment as described in the case of splash erosion. As a result, soil resistance to erosion may vary considerably during a rainfall event. If H is the degree of protection provided by the loose material, then all the material

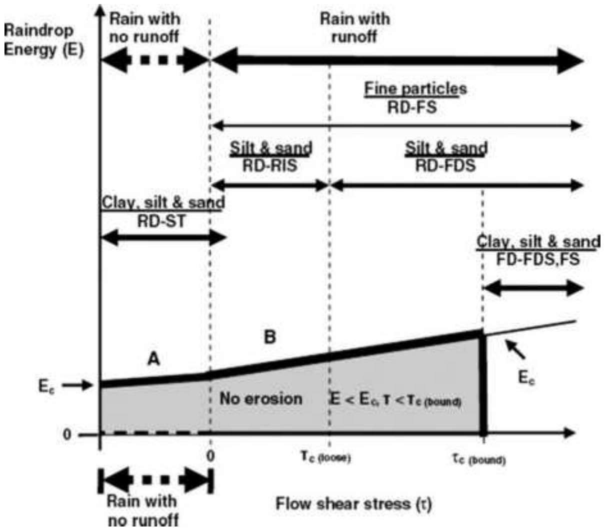


FIGURE 4 Schematic of how variations in raindrop kinetic energy and flow shear stress affect the detachment and transport of silt, sand, and fine particles during rainfall erosion. E_c is the critical raindrop kinetic energy required to cause raindrop detachment (RD). Its variation during rain with no runoff signifies changes in resistance to detachment caused by, for example, the development of a soil crust. Its variation with flow shear stress signifies the cushioning effect of increasing water depths. τ_c (bound) is the critical shear stress required to cause flow detachment (FD). τ_c (loose) is the critical shear stress required to cause flow-driven saltation (FDS). FS is continuous suspension in the flow. RIS is raindrop-induced saltation. ST is splash transport.

discharged from the soil surface comes from the layer of loose material when $H = 1.0$. Consequently, it follows from Eq. (20) that, for any given rainfall event,

$$D_{\mathrm{i}} = \left(k_{\mathrm{iq.U}}\left(1 - H\right) + Hk_{\mathrm{iq.P}}\right)q_{\mathrm{i}}I \tag{24}$$

where $k_{\mathrm{iq\cdot U}}$ is the interrill erodibility factor when no loose material exists on the surface and $k_{\mathrm{iq.P}}$ is the soil erodibility factor when loose material completely protects the soil surface from detachment. As noted earlier, detachment of soil particles from a cohesive surface varies with cohesion and interparticle friction so that factors such as the development of surface crusts can cause $k_{\mathrm{iq\cdot U}}$ to vary with time. The soil erodibility term in Eq. (24) is $k_{\mathrm{iq\cdot U}}(1\text{–H}) + H\, k_{\mathrm{iq.P}}$, and values of k_{iq} obtained in rainfall simulator experiments like those undertaken by Elliot et al.[31] lie between $k_{\mathrm{iq\cdot U}}$ and $k_{\mathrm{iq.P}}$. Where exactly they do lie is unknown because H is unknown.

Effect of Particle Travel Rates on Sediment Composition and Erodibility

As noted above, erodibilities have units of soil loss per unit of the erosivity index used in the model that is being considered in the analysis of the data. Factors such as cohesion, particle size, and aggregate stability have been observed to influence these erodibilities. Data on the physical and chemical nature of the soil involved may, in some cases, be used to predict erodibility, but often little attention is given to the composition of the sediment discharged in experiments undertaken to determine erodibility.

The composition of the soil transported from an eroding area by rain-impacted flow during an experiment tends to be finer than the original soil (e.g., Meyer et al.,[36] Miller and Baharuddin,[37] Palis,[38] and Parsons[39]), and particle travel rate is one of the factors influencing sediment composition. Fast-moving

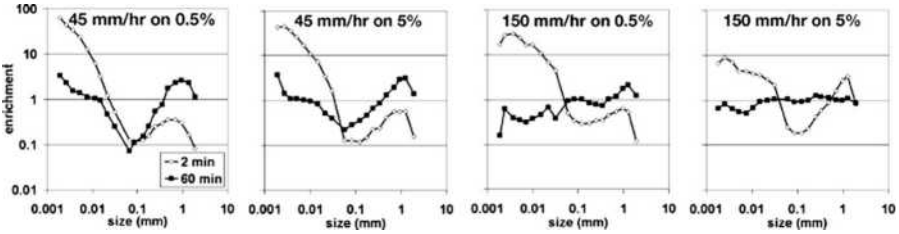


FIGURE 5 Enrichment factor curves for sediment discharged at 2 and 60 min when beds of sand were eroded by rain-impacted flows in experiments undertaken by Walker et al.[40] using 2.7 mm raindrops.
Source: Extracted from Figure 3 of Walker et al.[40]

particles detached at the top of an eroding area arrive at the downstream boundary well before slow-moving particles detached at the same place at the same time. Consequently, once sediment transport in rain-impacted flow starts, the fine material dominates the early discharge of sediment, but the composition of the sediment becomes coarser with time as more of the slower-moving particles reach the downstream boundary. If the particles are stable, then at the steady state, the composition of the sediment discharge must be the same both physically and chemically as the original soil. This is demonstrated by the results of laboratory experiments on erosion of sand by rain-impacted flows undertaken by Walker et al.[39] In these experiments, 3-m-long sloping beds of sand were eroded for 1 hour by rain-impacted flows generated by artificial rain at three rainfall intensities (45, 100, and 150 mm/h). The data shown in Figure 5 were generated by rain made up of a single drop size (2.7 mm) falling on rain generated flows over beds of sand with two different slope gradients (5% and 0.5%). The least erosive situation is the case where 45 mm/h rain fell when the slope gradient was 0.5%. The most erosive situation is the case where 150 mm/h rain fell when the gradient was 5%. The enrichment factor is the ratio of the proportion of the material in the sediment discharge to the proportion in the original material. In all cases, the discharge of sediment was dominated by fine material at 2 min and, in some cases, coarse material at 1 h (Figure 5). In the most erosive situation (150 mm/h rain falling on 5% slope), the sediment composition at 1 hour was close to the composition for the steady state, the composition that occurs when the

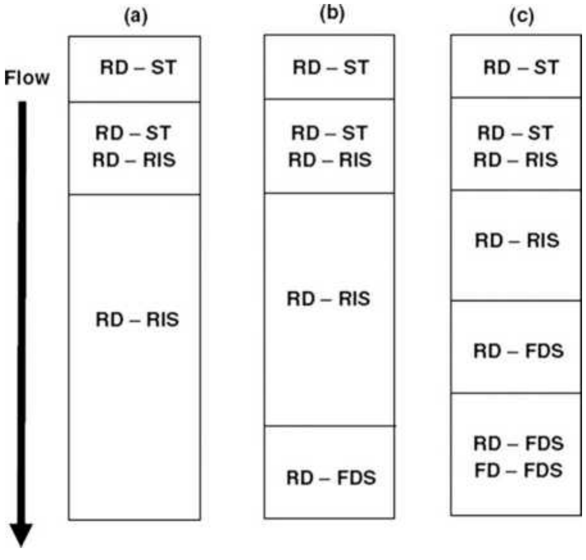


FIGURE 6 Schematic representation of how detachment and transport processes associated with rainfall erosion may vary spatially on planar inclined bare soil surfaces.

enrichment factor for all particle sizes equals 1.0. Sediment composition generated by the rain-impacted flows varies with the intensity of the rain, the slope gradient, slope length, and time because particles of different size and density travel at different rates. The differential rate of transport of particles in rain-impacted flows has consequences with respect to erodibility because the actual area contributing to the soil loss does not stabilize until the slowest-moving particles detached at the farthest point from the downslope boundary at the start of the rainfall event are discharged. This fact is seldom considered when experiments are undertaken to determine erodibility.

Eq. (24) is essentially targeted at situations where raindrop-induced saltation (RIS) controls sediment discharge in situations like that illustrated in Figure 6a. Particles of silt and sand may travel over the soil surface in more than one mode before being discharged. For example, as may be perceived from Figure 6b, they may leave the point of detachment traveling in splash (ST), then move further downslope by RIS, and finally by flow-driven saltation (FDS) as flow energy increases down the slope. Transitions between RIS and FDS may vary in time and space during a rainfall event and will depend on the intensity of the rain, the infiltration characteristics of the soil, the length of the slope, and the

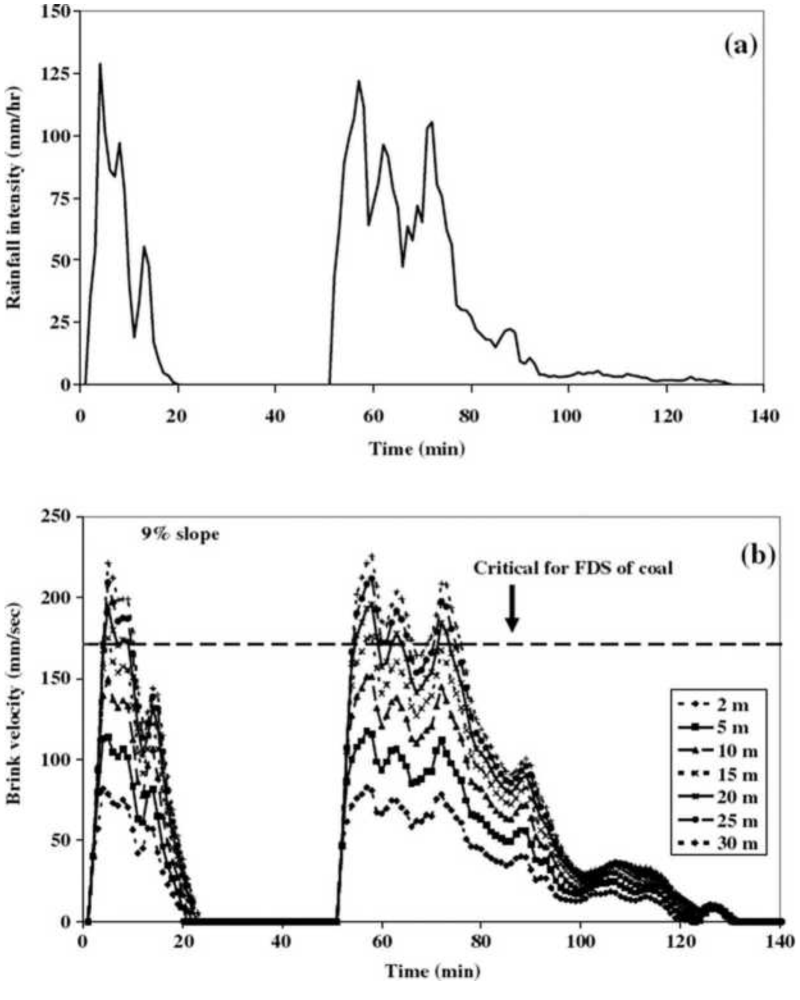


FIGURE 7 Brink (downstream boundary) flow velocities (b) for bare soil of various lengths inclined at 9% produced by the rainfall-runoff model described by Moore and Kinnell[41] and the rainfall intensities recorded during a rainfall event at the Ginninderra Experiment Station, Canberra, Australia (a).

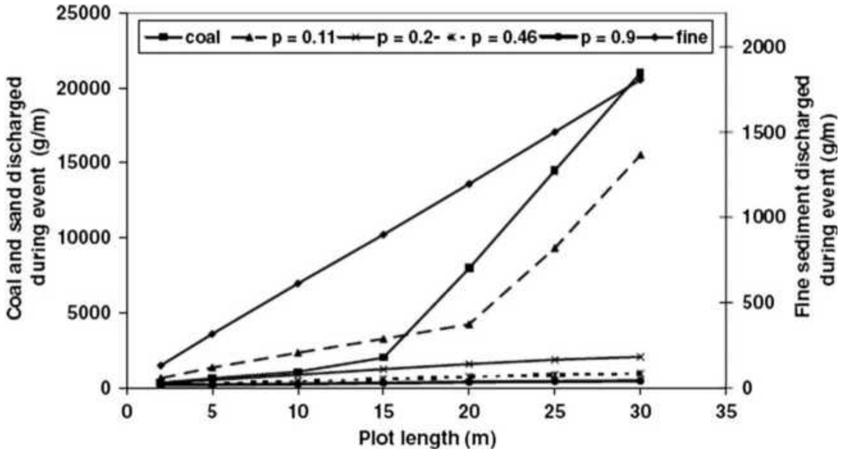


FIGURE 8 Amounts of coal, sand, and fine particles discharged for the rainfall and runoff conditions shown in Figure 7 when a mechanistic model of erosion by rain-impacted flow was used by Kinnell.[42]

slope gradient. The effect of the transition can have an appreciable effect on both soil loss and the composition of the sediment discharged. Figure 7b shows modeled flow velocities at the brink (discharge boundary) of planar bare soil surfaces of various length on a 9% slope resulting from the rainfall event shown in Figure 7a. Figure 8 shows the loss of materials of various size and density from those bare soil surfaces when a mechanistic model of erosion by rain-impacted flow was used by Kinnell.[42] For the surfaces up to 15 m in length, particles larger than 0.1 mm in size traveled over the whole length by RIS. Under these conditions, more of the 0.11 mm sand was lost during the event than the 0.46 mm coal. On the 20-m-long area, the 0.46 mm coal, which had been traveling slower than the 0.11 sand on shorter areas, traveled for a short period of time by FDS during the high-intensity bursts of rain, and, as a result, more of the 0.46 mm coal was lost than the 0.11 mm sand. On areas 25 m long and more, FDS also contributed to the discharge of the 0.11 mm sand, but the amount lost was always less than the amount of the 0.46 mm coal. Little regard is given to the effect of temporal and spatial changes in transport mechanism when experiments are undertaken to determine values for soil erodibility factors in sheet and interrill erosion areas even though these changes may have a major impact on the amount of soil lost from an area.

Conclusion

Although conceptually, rainfall erosivity is the capacity of rain to produce erosion, whereas soil erodibility is the susceptibility of the soil to be eroded, the factors controlling the erosive stress applied to the soil surface and the factors influencing the resistance of the soil to them vary in time and space in complex ways. In all existing predictive models of rainfall erosion, numerous simplifications and assumptions have to be made for practical reasons, and, as a result, these models do not have the capacity to deal with such complexity. Consequently, erosivity and erodibility values are specific to the model in which they are parameterized and to the scale that the model operates within.

References

1. Wischmeier, W.C.; Smith, D.D. *Predicting rainfall erosion losses—A guide to conservation planning*; Agricultural Handbook No. 537; US Department of Agriculture: Washington, DC, 1978.
2. Wischmeier, W.C. A soil erodibility nomograph for farmland and construction sites. *J. Soil Water Conserv.* **1971**, *26*, 189–193.

3. El-Swaify, S.A.; Dangler, E.W. Erodibilities of selected tropical soils. In Soil Erosion: Prediction and Control; Foster, G.R., Ed.; Soil Conservation Society of America: Ankeny, IA, 1976; 105–114.
4. Young, R.A.; Mutchler, C.K. Erodibility of some Minnesota soils. *J. Soil Water Conserv.* **1997**, *32*, 180–182.
5. Loch, R.J.; Slater, B.K.; Devoil, C. Soil erodibility (km) values for some Australian soils. *Aust. J. Soil Res.* **1998**, *36*, 1045–1056.
6. Zhang, K.L.; Shu, A.P.; Xu, X.L.; Yang, Q.K.; Yu, B. Soil erodibility and its estimation for agricultural soils in China. *J. Arid Environ.* **2008**, *72*, 1002–1011.
7. Renard, K.G.; Foster, G.R.; Weesies, G.A.; McCool, D.K.; Yoder, D.C. *Predicting soil erosion by water: A guide to conservation planning with the Revised Universal Soil Loss Equation (RUSLE)*; U.S. Department of Agriculture Agricultural Handbook. No. 703; US Department of Agriculture: Washington, DC, 1997.
8. Yu, B. A comparison of the R-factor in the Universal Soil Loss Equation and the revised Universal Soil Loss Equation. *Trans. ASAE* **1999**, *42*, 1615–1620.
9. Nash, J.E.; Sutcliffe, J.E. River flow forecasting through conceptual models. Part 1—A discussion of principles. *J. Hydrol.* **1970**, *10*, 282–290.
10. Kinnell, P.I.A. Event soil loss, runoff and the Universal Soil Loss family of models: A review. *J. Hydrol.* **2010**, *385*, 384–397.
11. Williams, J.R. Sediment-yield prediction with universal equation using runoff energy factor. In Present and Prospective Technology for Predicting Sediment Yield and Sources, *Publ. ARS-S-40.* US Dept. Agric.: Washington, DC, 1975; 244–252.
12. Williams, J.W.; Izaurralde, R.C.; Steglich, E.M. Agricultural Policy/Environmental Extender Model Theoretical Documentation, *BRC Report # 2008–17*; Blackland Research and Extension Center: Temple, TX, 2008.
13. Simanton, J.R.; Rawizt, E.; Shirley, E.D. Effects of rock fragments on erosion of semi-arid rangeland soils. In Erosion and Productivity of Soils Containing Rock Fragments; Krai, D.M., Hawkins, S.L., Eds.; Soil Science Society of America: Madison, WI, 1984; 65–67.
14. Onstad, C.A.; Foster, G.R. Erosion modelling on a watershed. *Trans. ASAE* **1975**, *18*, 288–92.
15. Williams, J.R.; Jones, C.A.; Dyke, P.T. A modelling approach to determining the relationship between erosion and productivity. *Transactions of the American Society of Agricultural Engineers* **1984**, *27*, 129–144.
16. Williams, J.R.; Jones, C.A.; Dyke, P.T. The EPIC model and its application. In *Proceedings of International Symposium on Minimum Data Sets for Agrotechnology Transfer, March 21–26, 1983.* ICRISAT Center: India, 1984; 111–121.
17. Bagarello, V.; Ferro, V.; Giordano, G. Testing alternative erosivity indices to predict event soil loss from bare plots in Southern Italy. *Hydrol. Process.* **2010**, *24*, 789–797. doi:10.1002/hyp.7538.
18. Bargarello, V.; Di Stefano, C.; Ferro, V.; Kinnell, P.I.A.; Pampalone, V.; Porto, P.; Todisco, F. Predicting soil loss on moderate slopes using an empirical model for sediment concentration. *J. Hydrol.* **2011**, *400*, 267–273.
19. Law, J.O.; Parsons, D.A. The relation of raindrop size to intensity. *Trans., Am. Geophys. Union* **1943**, *24*, 452–460.
20. Kinnell, P.I.A. Rainfall intensity-kinetic energy relationships for soil loss prediction. *Soil Sci. Soc. Am. J.* **1981**, *45*, 153–155.
21. Kinnell, P.I.A. Rainfall energy in Eastern Australia: Intensity-kinetic energy relationships for Canberra. *A.C.T. Aust. J. Soil Res.* **1987**, *25*, 547–553.
22. van Dijk, A.I.J.M.; Bruijnzeel, L.A.; Rosewell, C.J. Rainfall intensity-kinetic energy relationships: A critical literature appraisal. *J. Hydrol.* **2002**, *261*, 1–23.
23. Bullock, P.R.; de Jong, E.; Kiss, J.J. An assessment of rainfall erosion potential in southern Saskatchewan, from daily rainfall records. *Can. Agric. Eng.* **1987**, *29*, 109–115.

24. Rekolainen, S.; Posch, M. Adapting the CREAMS model for Finnish conditions. *Nord. Hydrol.* **1993**, *24*, 309–322.
25. Posch, M.; Rekolainen, S. Erosivity factor in the Universal Soil Loss Equation estimated from Finnish rainfall data. *Agric. Sci. Finl.* **1993**, *2*, 271–279.
26. Bagarello, V.; D'Asaro, F.V. Estimating single storm erosion index. *Trans. ASAE* **1994**, *37*, 785–791.
27. Yu, B.; Rosewell, C.J. An assessment of a daily rainfall. Erosivity model for New South Wales. *Aust. J. Soil Res.* **1996**, *34*, 139–152.
28. Richardson, C.W.; Foster, G.R.; Wright, D.A. Estimation of erosion index from daily rainfall amount. *Trans. ASAE* **1983**, *26*, 153–160.
29. Hoyes, N.; Waylen, P.R.; Jaramillo, A. Seasonal and spatial patterns of erosivity in a tropical watershed of the Colombian Andes. *J. Hydrol.* **2005**, *314*, 177–191.
30. Flanagan, D.C.; Nearing, M.A., Eds. *USDA-Water Erosion Prediction Project: Technical documentation*, NSERL Rep. No.10; National Soil Erosion Research Laboratory: West Lafayette, IN, 1995.
31. Elliot, W.J.; Liebenow, A.M.; Laflen, J.M.; Kohl, K.D. *A compendium of soil erodibility experiments*. Publication No. 3. USDA-ARS National Soil Erosion Research Laboratory: West Lafayette, IN, 1989.
32. Kinnell, P.I.A. Interrill erodibilities based on the rainfall intensity-flow discharge erosivity factor. *Aust. J. Soil Res.* **1993**, *31*, 319–332.
33. Wei, H.; Nearing, M.A.; Stone, J.J.; Guertin, D.P.; Spaeth, K.E.; Pierson, F.B.; Nichols, M.H.; Moffett, C.A. A new splash and sheet erosion equation for rangelands. *Soil Sci. Soc. Am. J.* **2009**, *73*, 1386–1392.
34. Torri, D.; Sfalanga, M.; Del Sette, F. Splash detachment: Runoff depth and soil cohesion. *Catena* **1987**, *14*, 149–155.
35. Kinnell, P.I.A. Raindrop impact induced erosion processes and prediction: A review. *Hydrol. Process.* **2005**, *19*, 2815–2844.
36. Meyer, L.D.; Harmon, W.C.; McDowell, L.L. Sediment sizes eroded from crop row sideslopes. *Trans. ASAE* **1980**, *23*, 891–898.
37. Miller, W.P.; Baharuddin, M.K. Particle size of interrilleroded sediments from highly weathered soils. *Soil Sci. Soc. Am. J.* **1987**, *51*, 1610–1615.
38. Palis, R.G.; Okwach, G.; Rose, C.W.; Saffigna, P.G. Soil erosion processes and nutrient loss. I. The interpretation of enrichment ratio and nitrogen loss in runoff sediment. *Aust. J. Soil Res.* **1990**, *28*, 623–639.
39. Parsons, A.J.; Abrahams, A.D.; Luk, S.-H. Size characteristics of sediment in interrill overland flow on a semiarid hill-slope, Southern Arizona. *Earth Surf. Process. Landforms* **2006**, *16*, 143–152.
40. Walker, P.H.; Kinnell, P.I.A.; Green, P. Transport of a noncohesive sand mixture in rainfall and runoff experiments. *Soil Sci. Soc. Am. J.* **1978**, *42*, 793–801.
41. Moore, I.D.; Kinnell, P.I.A. Kinematic overland flow—Generalization of Rose's approximate solution, Part II. *Hydrology* **1987**, *82*, 351–362.
42. Kinnell, P.I.A. The influence of raindrop induced saltation on particle size distributions in sediment discharged by rain-impacted flow on planar surfaces. *Catena* **2009**, *78*, 2–11.

Bibliography

Kinnell, P.I.A. Determining soil erodibility for the USLE-MM erosion model. *Catena* **2018**, *163*, 424–426.

Nearing, M.A.; Yin, S.; Borelli, P.; Polyakov, V.O. Rainfall erosivity: An historical review. *Catena* **2017**, *157*, 357–362.

USDA. Draft Science Documentation: Revised Universal Soil Loss Equation Version 2 (RUSLE2). 2013. https://www.ars.usda.gov/ARSUserFiles/60600505/RUSLE/RUSLE2_Science_Doc.pdf

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Index

Page numbers followed by f and t indicate figures and tables, respectively

SECOND EDITION

ENVIRONMENTAL MANAGEMENT HANDBOOK

VOLUME IV

Managing Water Resources and Hydrological Systems

edited by

Brian D. Fath

Sven E. Jørgensen

CRC Press
Taylor & Francis Group

Managing Water Resources and Hydrological Systems

Environmental Management Handbook, Second Edition

Edited by
Brian D. Fath and Sven E. Jørgensen

Volume 1
Managing Global Resources and Universal Processes

Volume 2
Managing Biological and Ecological Systems

Volume 3
Managing Soils and Terrestrial Systems

Volume 4
Managing Water Resources and Hydrological Systems

Volume 5
Managing Air Quality and Energy Systems

Volume 6
Managing Human and Social Systems

Managing Water Resources and Hydrological Systems

Second Edition

Edited by

Brian D. Fath and Sven E. Jørgensen

Assistant to Editor

Megan Cole

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

Cover photo: Forsand, Norway, N. Fath

Second edition published 2021
by CRC Press
6000 Broken Sound Parkway NW, Suite 300, Boca Raton, FL 33487-2742

and by CRC Press
2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

First edition published by CRC Press 2013

CRC Press is an imprint of Taylor & Francis Group, LLC

ISBN: 978-1-138-34266-8 (hbk)
ISBN: 978-1-003-04504-5 (ebk)

Typeset in Minion
by codeMantra

Contents

SECTION I APC: Anthropogenic Chemicals and Activities

SECTION II COV: Comparative Overviews of Important Topics for Environmental Management

SECTION III CSS: Case Studies of Environmental Management

SECTION IV DIA: Diagnostic Tools: Monitoring, Ecological Modeling, Ecological Indicators, and Ecological Services

SECTION V ELE: Focuses on the Use of Legislation or Policy to Address Environmental Problems

SECTION VI ENT: Environmental Management Using Environmental Technologies

SECTION VII NEC: Natural Elements and Chemicals Found in Nature

SECTION VIII PRO: Basic Environmental Processes

Preface

Given the current state of the world as compiled in the massive Millennium Ecosystem Assessment Report, humans have changed ecosystems more rapidly and extensively during the past 50 years than in any other time in human history. These are unprecedented changes that need certain action. As a result, it is imperative that we have a good scientific understanding of how these systems function and good strategies on how to manage them.

In a very practical way, this multivolume *Environmental Management Handbook* provides a comprehensive reference to demonstrate the key processes and provisions for enhancing environmental management. The experience, evidence, methods, and models relevant for studying environmental management are presented here in six stand-alone thematic volumes, as follows:

VOLUME 1 – Managing Global Resources and Universal Processes
VOLUME 2 – Managing Biological and Ecological Systems
VOLUME 3 – Managing Soils and Terrestrial Systems
VOLUME 4 – Managing Water Resources and Hydrological Systems
VOLUME 5 – Managing Air Quality and Energy Systems
VOLUME 6 – Managing Human and Social Systems

In this manner, the handbook introduces in the first volume the general concepts and processes used in environmental management. The next four volumes deal with each of the four spheres of nature (biosphere, geosphere, hydrosphere, and atmosphere). The last volume ties the material together in its application to human and social systems. These are very important chapters for a wide spectrum of students and professionals to understand and implement environmental management. In particular, features include the following:

- The first handbook that demonstrates the key processes and provisions for enhancing environmental management.
- Addresses new and cutting-edge topics on ecosystem services, resilience, sustainability, food–energy–water nexus, socio-ecological systems, etc.
- Provides an excellent basic knowledge on environmental systems, explains how these systems function, and gives strategies on how to manage them.
- Written by an outstanding group of environmental experts.

Since the handbook covers such a wide range of materials from basic processes, to tools, technologies, case studies, and legislative actions, each handbook entry is further classified into the following categories:

APC: Anthropogenic chemicals: The chapters cover human-manufactured chemicals and their activities

COV: Indicates that the chapters give comparative overviews of important topics for environmental management

CSS: The chapters give a case study of a particular environmental management example
DIA: Means that the chapters are about diagnostic tools: monitoring, ecological modeling, ecological indicators, and ecological services
ELE: Focuses on the use of legislation or policy to address environmental problems
ENT: Addresses environmental management using environmental technologies
NEC: Natural elements and chemicals: The chapters cover basic elements and chemicals found in nature
PRO: The chapters cover basic environmental processes.

Volume 4, *Managing Water Resources and Hydrological Systems*, has extensive coverage in over 80 entries of water supply, water treatment, wetlands, lakes, and other natural water systems. New entries cover the evolution of water management, with application of optimization tools, and the innovative move toward integrating the energy and water nexus. Case studies include the Aral Sea, Chesapeake Bay, Baltic Sea, and Yellow River to name a few. Policy implications regarding wetland conservation, use of remote sensing and GIS, and agricultural water use are included.

Brian D. Fath
Brno, Czech Republic
December 2019

Editors

Brian D. Fath is Professor in the Department of Biological Sciences at Towson University (Maryland, USA) and Senior Research Scholar at the International Institute for Applied Systems Analysis (Laxenburg, Austria). He has published over 180 research papers, reports, and book chapters on environmental systems modeling, specifically in the areas of network analysis, urban metabolism, and sustainability. He has co-authored the books *A New Ecology: Systems Perspective* (2020), *Foundations for Sustainability: A Coherent Framework of Life–Environment Relations* (2019), and *Flourishing within Limits to Growth: Following Nature's Way* (2015). He is also Editor-in-Chief for the journal *Ecological Modelling* and Co-Editor-in-Chief for *Current Research in Environmental Sustainability*. Dr. Fath was the 2016 recipient of the Prigogine Medal for outstanding work in systems ecology and twice a Fulbright Distinguished Chair (Parthenope University, Naples, Italy in 2012 and Masaryk University, Czech Republic in 2019). In addition, he has served as Secretary General of the International Society for Ecological Modelling, Co-Chair of the Ecosystem Dynamics Focus Research Group in the Community Surface Modeling Dynamics System, and member and past Chair of Baltimore County Commission on Environmental Quality.

Sven E. Jørgensen (1934–2016) was Professor of environmental chemistry at Copenhagen University. He received a doctorate of engineering in environmental technology and a doctorate of science in ecological modeling. He was an honorable doctor of science at Coimbra University (Portugal) and at Dar es Salaam (Tanzania). He was Editor-in-Chief of *Ecological Modelling* from the journal inception in 1975 until 2009. He was Editor-in-Chief for the *Encyclopedia of Environmental Management* (2013) and *Encyclopedia of Ecology* (2008). In 2004, Dr. Jørgensen was awarded the Stockholm Water Prize and the Prigogine Medal. He was awarded the Einstein Professorship by the Chinese Academy of Sciences in 2005. In 2007, he received the Pascal Medal and was elected a member of the European Academy of Sciences. He has published over 350 papers, and has edited or written over 70 books. Dr. Jørgensen gave popular and well-received lectures and courses in ecological modeling, ecosystem theory, and ecological engineering worldwide.

Contributors

Matthew Agarwala
Bennett Institute for Public Policy
University of Cambridge
Cambridge, United Kingdom

Shaikh Ziauddin Ahammad
School of Civil Engineering and Geosciences
Newcastle University
Newcastle, United Kingdom

Claude Amiard-Triquet
French National Center for Scientific Research (CNRS)
University of Nantes
Nantes, France

Javier Barragan
Department of Agroforestry Engineering
University of Lleida
Lleida, Spain

Lindsay Beevers
Lecturer in Water Management
School of the Built Environment
Heriot Watt University
Edinburgh, United Kingdom

Jerry M. Bigham
School of Environment and Natural Resources
Ohio State University
Columbus, Ohio

Marek Biziuk
Chemical Faculty
Department of Analytical Chemistry
Gdansk University of Technology
Gdansk, Poland

David L. Bjorneberg
Northwest Irrigation and Soil Research Laboratory
Agricultural Research Service (USDA-ARS)
U.S. Department of Agriculture
Kimberly, Idaho

Virginie Bouchard
School of Natural Resources
Ohio State University
Columbus, Ohio

Vince Bralts
Agricultural and Biological Engineering
Purdue University
West Lafayette, Indiana

Carl R. Camp, Jr.
Agricultural Research Service (USDA-ARS)
U.S. Department of Agriculture
Florence, South Carolina (Retired)

Kenneth L. Campbell
Agricultural and Biological Engineering Department
University of Florida
Gainesville, Florida

Kristi Denise Caravella
Florida Atlantic University
Boca Raton, Florida

Jesus Carrera
Technical University of Catalonia (UPC)
Barcelona, Spain

Dipankar Chakraborti
School of Environmental Studies
Jadavpur University
Calcutta, India

Bin Chen
State Key Joint Laboratory of Environmental Simulation and Pollution Control
School of Environment
Beijing Normal University
Beijing, China

Alexander H.-D. Cheng
Department of Civil Engineering
University of Mississippi
Oxford, Mississippi

Tørben Røjle Christensen
Climate Impacts Group
Department of Ecology
Lund University
Lund, Sweden

Matthew Cochran
School of Natural Resources
Ohio State University
Columbus, Ohio

Gretchen C. Coffman
Department of Environmental Science
University of San Francisco
San Francisco, California

Robin Kundis Craig
Attorneys' Title Professor of Law and Associate Dean for Environmental Programs
Florida State University College of Law
Tallahassee, Florida

Abhijit Das
Vijoygarh Jyotish Ray College
University of Calcutta
Kolkata, India

Antara Das
School of Environmental Studies
Jadavpur University
Kolkata, India

Bhaskar Das
School of Environmental Studies
Jadavpur University
Kolkata, India

Jorge A. Delgado
Soil Management and Sugar Beet Research Unit
Agricultural Research Service (USDA-ARS)
U.S. Department of Agriculture
Fort Collins, Colorado

Jan Dolfing
School of Civil Engineering and Geosciences
Newcastle University
Newcastle, United Kingdom

Pay Drechsel
International Water Management Institute (IWMI)
Colombo, Sri Lanka

Rathindra Nath Dutta
Department of Dermatology
Institute of Post Graduate Medical Education and Research
SSKM Hospital
Kolkata, India

Anna Ekberg
Department of Ecology
Lund University
Lund, Sweden

Guy Fipps
Agricultural Engineering Department
Texas A&M University
College Station, Texas

Gary W. Frasier
U.S. Department of Agriculture (USDA)
Fort Collins, Colorado

Martin V. Frey
Department of Soil Science
University of Stellenbosch
Matieland, South Africa

Wendy B. Gagliano
Clark State Community College
Springfield, Ohio

Agniezka Gałuszka
Division of Geochemistry and the Environment
Institute of Chemistry
Jan Kochanowski University
Kielce, Poland

Anja Gassner
Institute of Science and Technology
University of Malaysia-Sabah
Kota Kinabalu, Malaysia

David W. Graham
School of Civil Engineering and Geosciences
Newcastle University
Newcastle, United Kingdom

Timothy C. Granata
Department of Civil and Environmental Engineering and Geodetic Science
Ohio State University
Columbus, Ohio

Joyeeta Gupta
Department of Human Geography
Planning and International Development
Amsterdam Institute of Social Science Research
University of Amsterdam
Amsterdam, the Netherlands

and

IHE Institute for Water Education in Delft
Delft, the Netherlands

Robert W. Hill
Biological and Irrigation Engineering Department
Utah State University
Logan, Utah

Glenn J. Hoffman
Biological Systems Engineering
University of Nebraska–Lincoln
Lincoln, Nebraska

Sven Erik Jørgensen
Section of Environmental Chemistry
Copenhagen University
Copenhagen, Denmark

Sandeep Joshi
Shrishti Eco-Research Unit (SERI)
Pune, India

Douglas L. Karlen
U.S. Department of Agriculture (USDA)
Ames, Iowa

Subhankar Karmakar
Center for Environmental Science and Engineering (CESE)
Indian Institute of Technology Bombay
Mumbai, India

Gregory Kiker
University of Florida
Gainesville, Florida

Andrey G. Kostianoy
P.P. Shirshov Institute of Oceanology
Russian Academy of Sciences
Moscow, Russia

David P. Kreutzweiser
Canadian Forest Service
Natural Resources Canada
Sault Sainte Marie, Ontario, Canada

Atif Kubursi
Department of Economics
McMaster University
Hamilton, Ontario, Canada

Freddie L. Lamm
Northwest Research-Extension Center
Kansas State University
Colby, Kansas

Jean-Claude Lefeuvre
Laboratory of the Evolution of Natural and Modified Systems
University of Rennes
Rennes, France

Cheng Li
Guangdong Key Laboratory of Agricultural Environment Pollution Integrated Control
Guangdong Institute of Eco-Environmental Science and Technology
Guangzhou, China

Zhuoran Liang
Hangzhou Meteorological Bureau
Hangzhou, China

K.F. Andrew Lo
Department of Natural Resources
Chinese Culture University
Taipei, Taiwan

Richard Lowrance
Agricultural Research Service (USDA-ARS)
U.S. Department of Agriculture
Tifton, Georgia

Mahesh L. Makey
University of California
Oakland, California

Kyle R. Mankin
Department of Biological and Agricultural Engineering
Kansas State University
Manhattan, Kansas

Rudolf Marloth
San Diego State University
San Diego, California

J.F. Martin
Department of Food, Agricultural, and Biological Engineering
Ohio State University
Columbus, Ohio

Jocilyn Danise Martinez
University of South Florida
Tampa, Florida

Matgorzata Michalska
Institute of Maritime and Tropical Medicine
Gdynia, Poland

Zdzistaw M. Migaszewski
Division of Geochemistry and the Environment
Institute of Chemistry
Jan Kochanowski University
Kielce, Poland

Meenakshi Mukherjee
School of Environmental Studies
Jadavpur University
Kolkata, India

Subhas Chandra Mukherjee
Department of Neurology
Medical College
Kolkata, India

Rafael Munoz-Carpena
University of Florida
Gainesville, Florida

Stephen D. Murphy
Faculty of Environment
University of Waterloo
Waterloo, Ontario, Canada

O.M. Musthafa
Center for Pollution Control and Environmental Engineering
Pondicherry University
Pondicherry, India

Daria Nikitina
Department of Earth and Space Sciences
West Chester University of Pennsylvania
West Chester, Pennsylvania

J. Pablo Ortiz-Partida
Union of Concerned Scientists
Cambridge, Massachusetts

Lloyd B. Owens
U.S. Department of Agriculture (USDA)
Coshocton, Ohio

Shyamapada Pati
Department of Obstetrics and Gynaecology
Calcutta National Medical College
Kolkata, India

Zaneta Polkowska
Gdansk University of Technology
Gdansk, Poland

Manzoor Qadir
Institute for Water, Environment and Health (UNU-INWEH)
Hamilton, Ontario, Canada

Quazi Quamruzzaman
Dhaka Community Hospital
Dhaka, Bangladesh

Anna Rabajczyk
Independent Department of Environment Protection and Modeling
Jan Kochanowski University of Humanities and Sciences
Kielce, Poland

Mohammad Mahmudur Rahman
School of Environmental Studies
Jadavpur University
Kolkata, India

Liqa Raschid-Sally
International Water Management Institute (IWMI)
Colombo, Sri Lanka

James D. Rhoades
Agricultural Salinity Consulting
Riverside, California

Mark Robinson
Center for Ecology and Hydrology
Wallingford, United Kingdom

Tarit Roychowdhury
School of Environmental Studies
Jadavpur University
Calcutta, India

Clayton Rubec
Center for Environmental Stewardship and Conservation
Ottawa, Ontario, Canada

D.W. Rycroft
Department of Civil and Environmental Engineering
Southampton University
Southampton, United Kingdom

Khitish Chandra Saha
School of Environmental Studies
Jadavpur University
Kolkata, India

A.V. Shanwal
Department of Soil Science
Chaudhary Charan Singh Haryana Agricultural University
Hisar, India

Paul K. Sibley
School of Environmental Sciences
University of Guelph
Guelph, Ontario, Canada

S.P. Singh
National Bureau of Soil Survey and Land Use Planning
Indian Agricultural Research Institute
New Delhi, India

Jeff Skousen
Division of Plant and Soil Sciences
West Virginia University
Morgantown, West Virginia

Bogdan Skwarzec
Faculty of Chemistry
University of Gdansk
Gdansk, Poland

Sean M. Smith
Ecosystem Restoration Center
Maryland Department of Natural Resources
Annapolis, Maryland

Roy F. Spalding
Water Science Laboratory
University of Nebraska–Lincoln
Lincoln, Nebraska

B.A. Stewart
Dryland Agriculture Institute
West Texas A&M University
Canyon, Texas

Piotr Szefer
Department of Food Sciences
Medical University of Gdansk
Gdansk, Poland

Thomas L. Thurow
Department of Renewable Resources
University of Wyoming
Laramie, Wyoming

Ralph W. Tiner
National Wetlands Inventory Program
U.S. Fish and Wildlife Service
Hadley, Massachusetts

Francine van den Brandeler
Department of Human Geography
Planning and International Development
University of Amsterdam
Amsterdam, the Netherlands

George Vance
Department of Renewable Resources
University of Wyoming
Laramie, Wyoming

Alejandra Virgen-Urcelay
University of British Columbia
Vancouver, Canada

Victor de Vlaming
Aquatic Toxicology Laboratory
University of California—Davis
Davis, California

Jan Vymazal
Faculty of Environmental Sciences
Department of Landscape Ecology
Czech University of Life Sciences
Prague, Czech Republic

Ivan A. Walter
Ivan's Engineering, Inc.
Denver, Colorado

Saige Wang
State Key Joint Laboratory of Environmental Simulation and Pollution Control
School of Environment
Beijing Normal University
Beijing, China

Ynuzhang Wang
Academy of Yellow River Conservancy Science
Zhengzhou, China

Dennis Wichelns
International Water Management Unit
Colombo, Sri Lanka

Brent Wootton
Center for Alternative Wastewater Treatment
Fleming College
Lindsay, Ontario, Canada

Pai Wu
College of Tropical Agriculture and Human Resources
University of Hawaii
Honolulu, Hawaii

Colin N. Yates
Centre for Research and Innovation
Fanshawe College
London, Ontario, Canada

Honglin Zhong
Department of Geographical Sciences
University of Maryland
College Park, Maryland

Yifei Zhu
Gemune LLC
Fremont, California

Zixi Zhu
Henan Institute of Meteorology
Zhengzhou, China

I

APC: Anthropogenic Chemicals and Activities

1

Aquatic Communities: Pesticide Impacts

David P. Kreutzweiser and
Paul K. Sibley

Introduction

A biotic community can be defined as an assemblage of plant or animal species utilizing common resources and cohabiting a specific area. Examples could include a fish community of a stream, an insect community of a forest pond, or a phytoplankton community of a lake. Interactions among species provide ecological linkages that connect food webs and energy pathways, and these interconnections provide a degree of stability, or balance, to the community. Community balance can be described as a state of dynamic equilibrium in which species and their population dynamics within a community remain relatively stable, subject to changes through natural adjustment processes. Toxic effects of pesticides can disrupt these processes and linkages and thereby cause community balance upsets. For example, this can occur when a pesticide has a direct impact on a certain species in a community and reduces its abundance while other unaffected species increase in abundance in response to the reduced competition for food resources or increased habitat availability. Some of the best examples of pesticide impacts on biological communities are found in freshwater studies. Freshwater aquatic communities are usually contained within distinct boundaries or systems, and this generates a high degree of connectivity among species, thereby increasing their susceptibility to pesticide-induced disturbances at the community level.

We examine traditional and developing methods for measuring pesticide impacts on freshwater communities, with emphasis on recent improvements in risk assessment approaches and analyses, and provide some examples for illustration. We then describe some advances in impact mitigation strategies and discuss some ongoing issues pertaining to understanding, assessing, and preventing pesticide impacts including probabilistic risk assessment (PRA), population and ecological modeling, and pesticide interactions with multiple stressors. The integration of improved risk assessment and mitigation approaches and technologies together with information generated from the numerous impact studies available will provide a sound scientific basis for decisions around the use and regulation of pesticides in and near water bodies.

Measuring Impacts on Aquatic Communities

Changes in aquatic communities can be measured directly in water bodies by a number of quantitative and qualitative sampling methods. Descriptions of those methods can be found in any up-to-date text or handbook (e.g., Hauer and Lamberti[1]). Measurements can be in terms of community structure (species composition) or community function (a measurable ecosystem process attributable to a biotic community that causes a change in condition) and can include both direct and indirect effects.[2,3] Community structure is a measure of biodiversity in its most general sense, that is, the number of species or other taxonomic units and their relative abundances. Some community functions are referred to as environmental or ecosystem services. Examples include organic matter breakdown and nutrient cycling that is largely mediated by microbial communities, or water uptake, filtration, and flood control mediated by shoreline plant communities.[4] Both community structure (biodiversity) and function (ecosystem services) are being increasingly valued by society and global economies,[5,6] and therefore sustaining healthy aquatic communities will be an important driver of pesticide impact mitigation efforts.

Detecting impacts of pesticides typically involves repeated sampling and a comparison of community attributes among contaminated and uncontaminated test units over time, or across a gradient of pesticide concentrations. The test units can range from petri dishes to natural ecosystems, with a trade-off between experimental control in small test units and environmental realism in field-level testing and whole ecosystems.[7] In an effort to incorporate both experimental control and environmental realism in pesticide impact testing, the use of microcosms or model ecosystems for measuring impacts on aquatic communities has increased over the past couple of decades.[8,9] Model ecosystems for community-level pesticide testing can be quite simple at lower-trophic levels such as with microbial communities (e.g., Widenfalk et al.[10]) but will necessarily be more complex for testing higher-order biological communities (e.g., Wojtaszek[11]). Regardless of the test units, an important consideration for measuring pesticide impacts will be an assessment of the duration of impact or rate of recovery. A rapid return to pre-pesticide or reference (nopesticide) community condition will reduce the long-term ecological consequences of the pesticide disturbance.[12]

Traditional measures of community-level impacts have focused on structure and have usually been expressed in terms of single-variable indices such as species richness, diversity, or abundance. These indices are useful descriptors of community structure but suffer from the fact that they reduce complex community data to a single summary metric and may miss subtle or ecologically important changes in species composition across sites or times. Over the last couple of decades, ecotoxicologists have increasingly turned to multivariate statistical techniques for analyzing community response data.[13] A variety of multivariate statistical techniques and software are available and are usually considered superior for the analysis of community data because they retain and incorporate the spatial and temporal multidimensional nature of biological communities.[14] This includes various ordination techniques that can provide graphical representation of spatiotemporal patterns in community structure in which points that lie close together in the ordination plot represent communities of similar composition (richness, abundance), while communities with dissimilar species composition are plotted further apart.

Figure 1 illustrates the use of an ordination plot generated by nonmetric multidimensional scaling for detecting differences among aquatic insect communities in four control and eight insecticide-treated streams. These data have been adjusted for illustrative purposes but are based on real invertebrate community responses to an insecticide in outdoor stream channels.[15] At both concentrations of the insecticide, the community structure of stream insects clearly shifted away from the natural community composition in control streams as depicted by the separation of treated streams (T1 and T2) from controls (C) in the ordination bi-plot. The plot also illustrates that the variability among treated streams (relative distance between points) was greater than that among control streams, that the low-concentration streams (T1) and high-concentration streams (T2) tended to separate along axis 1, and that the T2 streams were further removed from controls than the T1 streams, indicating a differential response by the insect communities to the two test concentrations. Canonical correspondence analysis

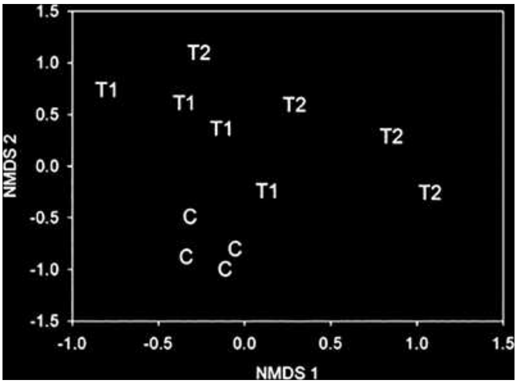


FIGURE 1 Ordination by nonmetric multidimensional scaling of aquatic insect communities in stream channels. Each point represents the community structure of control channels (C) and channels treated with a neem-based insecticide at a low (T1) and high (T2) concentration.
Source: Adapted from Kreutzweiser et al.[15]

and redundancy analysis have also been commonly used to assess aquatic community responses to pesticide contamination.[16,17] A useful refinement of an ordination technique for detecting and interpreting pesticide impacts on aquatic communities is principal response curves (PRCs).[18] PRC is derived from redundancy analysis, and time-dependent responses in the treatments are expressed as deviations from the control or reference system allowing for clear visualization of pesticide effects.

Assessing Risk of Pesticide Impacts on Aquatic Communities

The likelihood or risk of harmful effects on aquatic communities from exposure to pesticides will depend on the exposure concentration, bioavailability, exposure duration, rate of uptake, inherent species sensitivities, community composition, and other community attributes. All of these must be measured, estimated, modeled, or predicted to derive an assessment of risk to aquatic communities for any given pesticide. Formalized risk assessment frameworks and guidelines for pesticides have been developed in the United States,[19] the European Union,[20] Canada,[21] and elsewhere and can be consulted for detailed information on the various components of a risk assessment. In brief, pesticide risk assessments typically include the following phases: 1) defining the problem by determining the pesticide use patterns and developing conceptual models and hypotheses around how it is expected to behave, the anticipated exposure regimes, the kinds of organisms that are likely to be at risk, the community or entity that is to be protected, and the level of protection that will be acceptable; 2) developing the measurement endpoints for assessing risk of harm by establishing which response measurements are relevant and applicable, and how the measurements will be made; 3) outlining the risk assessment process by specifying the kinds of data to be used and how they will be derived including simulation modeling, empirical laboratory, microcosm or field testing, their appropriate spatial and temporal scales, and their statistical analyses; 4) applying the risk assessment by running models or collecting data, completing analyses, summarizing outputs, and providing risk estimates; 5) conducting risk communication and management by answering questions posed in the problem formulation, suggesting risk mitigation strategies if necessary, and communicating those to appropriate users; and 6) conducting follow-up monitoring to evaluate the success of mitigation strategies and to implement adaptive management to address deficiencies if or when necessary.[22,23]

Traditionally, pesticide risk assessments have relied on standardized, single-species toxicity tests to predict effects on communities, the underlying assumption being that protecting the most sensitive species will protect whole communities. In this case, the selection and relevance of test species are critically important to a successful and meaningful risk assessment.[24] However, the accuracy and relevance of estimating the potential risk to aquatic communities can be greatly improved by consideration of specific species or community attributes. In particular, attribute information can improve the ecological relevance and predictive capabilities of conceptual models and the generation of hypotheses in the risk assessment process. Insofar as these attributes affect exposure, sensitivity, or both, they can increase or decrease risk beyond what could be determined from toxicity estimates or species sensitivity distributions alone.

Behavioral attributes can elevate the risk of pesticide effects on species by increasing the likelihood of intercepting the stressor. For example, young-of-the-year bluefish (*Pomatomus saltatrix*) typically feed in estuaries during their early life stages where agricultural runoff can elevate concentrations of pesticides in food items. This feeding behavior can result in bioaccumulation and in adverse effects such as reduced migration, overwinter survival, and recruitment success in fish communities.[25] Incorporating this kind of information into conceptual models and risk hypotheses will generate more realistic risk assessments. In addition, behavioral attributes themselves can be relevant measurement endpoints if the pesticide mode of action indicates risk of sublethal behavioral effects at expected concentrations. For example, some pesticides have been shown to impair the ability to capture prey in fish[26] and the ability to avoid predators in zooplankton.[27] These types of adverse effects can disrupt trophic linkages and reduce survival or reproduction, thus impacting community balance.

Inclusion of life history information into conceptual models and risk hypotheses can also refine and improve the risk assessment process. Life history strategies can influence a species susceptibility to a stressor through effects on a population's resilience or ability to recover from disturbance.[28] Different species exposed to the same pesticide and experiencing similar levels of effect in terms of population declines do not necessarily recover at the same rates when recovery is dependent on reproduction or dispersion. Populations of organisms with short regeneration times (e.g., several generations per year) and/or high dispersal capacity have higher likelihood of recovery from pesticide-induced population declines than those with longer regeneration periods and limited dispersal capacity. These differential life history strategies and their influences on community response and recovery from pesticide effects have been demonstrated empirically (e.g., van den Brink et al.[29] and Kreutzweiser et al.[30]) and through population modeling.[31] These community balance upsets could not have been predicted from screening-level toxicity data or from species sensitivity data; thus, inclusion of life history information in conceptual models can improve risk hypotheses and direct the assessment to focus on species at higher risk owing to specific life history strategies.

Life history attributes can also influence the risk of pesticide effects through differential life-stage sensitivity or susceptibility. Early life stages are often (but not exclusively) more sensitive to pesticides than later stages. An organism's life stage can also influence its susceptibility to a pesticide by increasing or decreasing the likelihood of intercepting the stressor. If a contaminant is present in the environment at effective concentrations during a period in which the particular life stage of a species is present, then the risk to that species is increased. For some amphibians, aquatic (larval) stages could be at higher risk of direct and indirect effects of pesticides than their terrestrial (adult) life stages when their larval stage coincides with pesticide contamination of water bodies.[32] Thus, while a species sensitivity and geographical distribution may indicate potential risk, the life-stage information coupled with pesticide use pattern, timing, or fate information may indicate little likelihood of exposure to the pesticide and the risk assessment can be adjusted accordingly.

Functional attributes may also be important for refining or improving pesticide risk assessments. Protection goals for populations and communities often include the safeguarding of critical biological processes or ecosystem function. Measuring ecosystem function integrates responses of component populations and can be a relevant measurement endpoint when species loss affects ecosystem function such as energy transfer and organic matter cycling.[33] However, most ecosystems are complex and it may not be clear which functional attributes are critical for sustaining ecological processes or the extent to which they can sustain changes in structural properties (e.g., population levels, diversity) without adversely affecting ecosystem function. Neither is it clear if functional endpoints are more or less sensitive than structural endpoints for detecting ecosystem disturbance. Some studies investigating the relationship between species diversity and ecosystem function have indicated that ecosystems can tolerate some species loss because of functional redundancy.[34] Functional redundancy is thought to occur when several species perform similar functions in ecosystems such that some may be eliminated with little or no effect on ecosystem processes. Others have suggested that redundant species are required to ensure ecosystem resilience to disturbance as a form of biological insurance, especially at large spatial scales.[35]

Given these discrepancies, measurement endpoints based on functional attributes are not typically used in pesticide risk assessments because it is generally accepted that protection of community structure will protect ecosystem function. However, when specific functional attributes can be identified and are known or suspected to be at risk from a pesticide, they can be included in the data requirements for a risk assessment. An example would be the risk of adverse effects on leaf litter decomposition (a critical ecosystem function in forest soils and water bodies) posed by a systemic insecticide for control of wood-boring insects in trees.[36] In that case, the protection goal was maintaining leaf litter decomposition, the community at risk was decomposer invertebrates feeding on leaves from insecticide-treated trees, and the selection of test species was directed to a specific functional group because of the unique route of exposure to decomposer organisms identified in the risk hypotheses.

Some Examples of Pesticide Impacts on Aquatic Communities

A few examples will serve to illustrate how pesticides can cause disruptions to aquatic communities. DeNoyelles et al.[37] reviewed studies into pesticide impacts on aquatic communities and reported that herbicides like atrazine, hexazinone, and copper sulfate were directly toxic to most species of phytoplankton (waterborne algae). After herbicide applications, reductions in phytoplankton caused secondary reductions in herbivorous zooplankton, resulting from a depleted food source for the zooplankton. They further showed that direct adverse effects on phytoplankton can also cause disruptions to the bacterial-based energy pathways by reducing carbon flow from phytoplankton to bacteria, and ultimately to grazing protozoans and zooplankton. Boyle et al.[38] found that applications of the insecticide diflubenzuron to small ponds reduced populations of several aquatic invertebrate species. This in turn resulted in indirect effects on algae (increased productivity because of release from grazing pressure by the invertebrates) and on juvenile fish populations (reduced production because of limited invertebrate prey availability). George et al.[39] used a novel approach to predict effects of pesticide mixtures on zooplankton communities and then tested the predictions in outdoor microcosms. Responses among zooplankton populations within the community differed, depending on the pesticide mixture, and those differences appeared to reflect the relative susceptibilities among specific taxa within groups. Cladocerans declined but were less sensitive than copepods to a chlorpyrifos-dominated mixture, while rotifers actually increased after application in response to release from competition or predation pressures.

Kreutzweiser et al.[40] applied a neem-based insecticide to forest pond enclosures and measured effects on zooplankton community structure, respiration, and food web stability. Significant concentration-dependent reductions in numbers of adult copepods were observed, but immature copepods and cladocerans were unaffected (Figure 2). There was no evidence of recovery of adult copepods within the sampling season. During the period of maximal impact (about 4 to 9 weeks after the applications), total plankton community respiration was significantly reduced, and this contributed to significant concentration-dependent increases in dissolved oxygen and decreases in specific conductance. The reductions in adult copepods resulted in negative effects on zooplankton food web stability through elimination of a trophic link and reduced interactions and connectance.

Van Wijngaarden et al.[41] evaluated the responses of aquatic communities in indoor microcosms to a suite of pesticides used for bulb crop protection. At pesticide concentrations equivalent to 5%

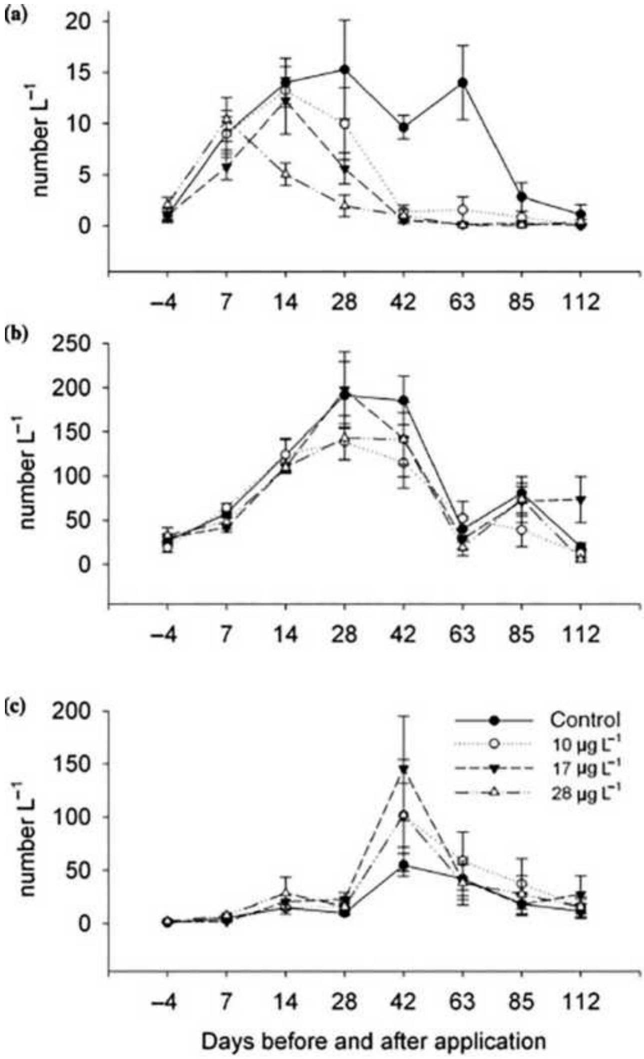


FIGURE 2 Mean abundance (±1 SE, n = 5) of (a) adult copepods, (b) immature copepods, and (c) cladocerans in natural pond microcosms (controls) and microcosms treated at three different rates of a neem-based insecticide
Source: Taken from Kreutzweiser et al.[40]

spray drift deposition, zooplankton taxa within communities showed significant changes relative to non-treated controls, reflecting taxon-specific sensitivities. Some copepods and rotifers in particular showed significant declines for at least 13 weeks, while many other rotifers and cladocerans were unaffected or increased weeks, while many other rotifers and cladocerans were unaffected or increased. Several macroinvertebrate taxa were negatively affected, and this contributed to significant declines in leaf litter decomposition among treated microcosms. The herbicide asulam was among the suite of pesticides, and it induced significant reduction of the macrophyte *Elodea nuttallii*. This in turn caused significant changes in water chemistry (decreases in dissolved oxygen and pH, increases in alkalinity and specific conductance) and increases in phytoplankton biomass from decreased competition for nutrients. Increased phytoplankton and reduced zooplankton predators combined to support higher abundance of less sensitive zooplankton taxa. The authors point out that most of these effects were not measurable at more realistic rates of spray drift deposition.

Relyea and Hoverman[42] investigated impacts of the insecticide malathion on aquatic communities in microcosms designed to mimic a simple aquatic food web that can be found in ponds and wetlands. The insecticide generally reduced zooplankton abundance, and these reductions stimulated increases in phytoplankton, decreases in periphyton (attached algae), and decreases in growth of frog tadpoles. While invertebrate predator survival was not affected, amphibian prey survival increased with insecticide concentration, apparently the result of insecticide-induced impairment of predation success by the invertebrates. Overall, the study demonstrated that realistic concentrations of an insecticide can interact with natural predators to induce large changes in aquatic community balance.

Reducing Risk of Pesticide Impacts on Aquatic Communities

For pesticides applied to crops and forests, exposure to aquatic communities can be minimized by the implementation of vegetated spray buffers or setbacks to intercept off-target spray drift and runoff.[43] Pesticide runoff can be further reduced by using formulations that are less prone to wash-off, leaching, and mobilization. Recent advances in spray drift reduction and improved spray guidance systems can also significantly reduce the off-target movement of pesticides to water bodies.[44] Examples include new technologies in map-based automated boom systems for row crops[45] and Geographical Information System (GIS)-based landscape analysis for predicting off-target pesticide movement.[46]

The risk of adverse effects on aquatic communities may also be decreased by intentional selection and use of pesticides that are inherently safer to the environment. This would include so-called reduced-risk pesticides that are bioactive compounds usually with unique modes of action and derived from microbial, plant, or other natural sources. These are generally thought to be less persistent and toxic to non-target organisms than conventional synthetic pesticides.[47] Examples include the bacteria-derived insecticide *Bt (Bacillus thuringiensis),* the plant-derived insecticide neem, and the microbe-derived herbicide phosphinothricin. However, Thompson and Kreutzweiser[48] caution that it cannot be assumed that this group of pesticides is inherently safer or more environmentally acceptable than synthetic counterparts and that full environmental risk evaluations must be conducted to ensure their environmental safety.

These types of technologies combined with the use of non-pesticide approaches to pest management form the basis of integrated pest management (IPM) strategies. IPM strategies are those in which the judicious use of pesticides is only one of several concurrent methods to control or manage losses from pest damage. This can include the use of natural enemies and parasites, biological control agents, insect growth regulators, confusion pheromones, sterile male releases, synchronizing with weather patterns known to diminish pest populations, and cultivation methods and crop varieties to improve conditions for natural enemies or degrade conditions for pest survival.[49] Increasing the use of IPM approaches can reduce reliance on pesticides and thus reduce the risk of pesticide impacts overall.

Recent Advances and Outstanding Issues

Pesticide risk assessments and risk reductions have recently been advanced in terms of ecological realism and effectiveness through some developing methods and techniques. Traditional risk assessments have estimated hazards from pesticides by comparing the expected environmental concentration (often predicted from worst-case scenarios) to the toxic threshold for the most sensitive test species. When the expected concentration is higher than the toxicity threshold, the pesticide is considered to have potential for environmental effects. These so-called hazard or risk quotient approaches are still widely used in pesticide risk assessment and regulation, but more recently, PRA and probabilistic hazard assessment (PHA) approaches are being adopted. In these approaches, pesticide exposure levels and the likelihood of toxic effects are estimated from probability distributions based on all reliable data available.[50] In PRA, exposure and effects distributions are developed from modeling or measurements in laboratory, microcosm, or field studies and used to improve the accuracy and relevance of the estimated likelihood

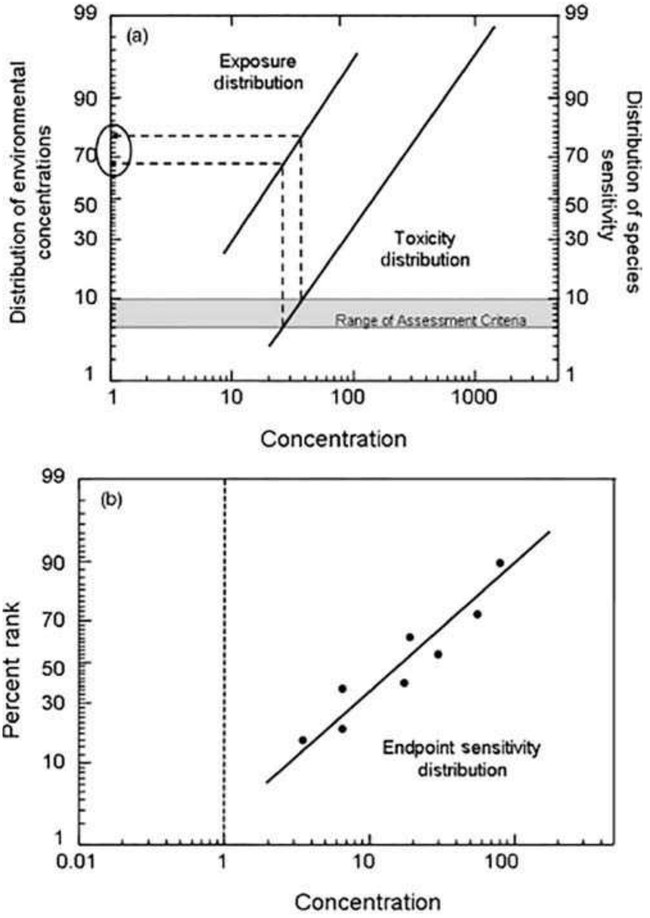


FIGURE 3 Schematic illustrating the principle of PRA (a) and PHA (b). PRA is based on a comparison of exposure and effects distributions using a predetermined criterion typically in the range of 5%-10% (shaded area and dashed lines in panel A) to determine the probability of exceeding the criterion (ellipse on y-axis); PHA is based on a comparison of an endpoint-derived sensitivity distribution within a test species to a threshold value such as a hazard quotient (dashed line in panel B).

of environmental effects compared to the traditional worst-case (hazard/risk quotient) approach (e.g., Solomon[51]). In PHA, a distribution approach is also used, except that the probability of hazard is estimated from distributions built on the relative sensitivity of interspecies endpoints rather than species sensitivity itself.[52] Figure 3 illustrates the principles of PRA (Figure 3a) and PHA (Figure 3b). Regardless of the approach, one important aspect of PRA that is ongoing is the development and use of uncertainty analysis to quantify variability and uncertainty in exposure and effects estimates. Characterizing and quantifying uncertainty will provide more meaningful risk assessments and improved decision making for minimizing potential risk of pesticide impacts in or near water.[53]

Efforts at incorporating population or ecological modeling into pesticide risk assessments have also improved their accuracy and relevance for predicting, and therefore mitigating, risk of harm to aquatic communities.[54] The use of ecological models to incorporate a suite of factors including lethal and sublethal effects and their influences on the risks to organisms, populations, or communities can provide useful insights into receptor/pesticide interactions and can thereby improve risk assessments and direct mitigation measures. Population models that account for differential demographics and population growth rates within communities have been shown to provide more accurate assessments of potential pesticide impacts on populations and communities than what conventional lethal concentration estimates can provide.[55] Ecological and population modeling combined with pesticide exposure modeling and case-based reasoning (drawing on past experience or information from similar chemical exposures) can provide further refinements and improve risk assessment for aquatic communities.[56] Another recent advancement in ecological modeling to predict pesticide effects is the use of trait-based information such as organism morphology, life history, physiology, and feeding ecology in risk assessments.[57] This approach includes some of the functional attributes and concepts described above in the section on "Assessing Risk of Pesticide Impacts on Aquatic Communities" and has the advantage of formally expressing communities as combinations of functional traits rather than as groups of species, thereby yielding a more meaningful description of community structure and function. Taken together, these modeling approaches that incorporate probability distributions, toxicological sensitivities, population dynamics, ecological information, and functional trait attributes can be integrated into improved risk assessments that will inform mitigation and prevention strategies for pesticide use.[58]

Two additional issues that present challenges to pesticide risk assessment and mitigation are pesticide mixtures and the combined or cumulative effects of multiple stressors on pesticide impacts. Pesticides frequently occur as mixtures in aquatic systems, particularly in agricultural regions, and methods to assess and/or predict pesticide mixture toxicity under laboratory conditions have been relatively well developed. However, there are still large uncertainties associated with the prediction of pesticide mixture toxicity, and additional studies are needed to evaluate the performance of mixture models when evaluating community-level endpoints and toxicity thresholds over long-term exposures.[59] Secondly, whereas most pesticide assessment data are derived from tests or experiments under controlled or semicontrolled environmental conditions, pesticides in natural environments may interact with a number of other natural or human-caused stressors that can substantially alter the likelihood and magnitude of pesticide impacts.[60] Other stressors could include overarching effects of climate change that can influence water temperature and quality; land use activities that result in chemical, sediment, and nutrient pollution of waterways; and biotic interactions with invasive species in aquatic communities. A number of studies have examined the combined effects of a pesticide with other stressors, but they have usually been single stressor effects tested at the single-species level. Examples of studies that examined combined effects include pesticide interactions with water temperature,[61] pH,[62] dissolved organic matter,[63] UV radiation,[64] predators,[65] competitors,[66] food availability,[67] elevated sediments,[68] and other chemical stressors.[69] However, potential multiple stressors and their interactions with pesticides can be myriad and testing or extrapolating to community-level impacts is onerous at best. Sorting out and mitigating pesticide impacts from among these multiple stressors continues to be a challenge, and the suggestion by Laskowski et al.[70] to include studies of toxicant interactions with a range of environmental conditions in risk assessments seems warranted.

Conclusions

Because of the high degree of connectivity among species in an aquatic community, pesticides pose a risk of harm to the community stability or balance. The community structure can be altered by direct effects, indirect effects, or both, and this can cause disruptions to the interactions and linkages among species and to their ecological function. This risk of harm will depend on exposure concentration, bioavailability, exposure duration, rate of uptake, species sensitivities, community composition, and other community attributes. Recent advances in pesticide risk assessment for aquatic communities have improved the ecological relevance and predictive capabilities for determining, and thus mitigating, potential harmful impacts. Pesticide impacts on aquatic communities can be minimized by the use of improved application technologies to reduce application rates and to decrease off-target movement to water bodies. Potential impacts can be further minimized through the selection and use of pesticides that are demonstrated to be inherently safer to the environment and through the application of IPM strategies. Given the preponderance of pesticide impact studies in freshwater aquatic ecosystems, the improved risk assessment frameworks and regulatory requirements for pesticide evaluations, and the recent advances in mitigation technologies, many decisions around the use of pesticides can be made on a sound scientific basis rather than on misinformed perceptions or politically driven agendas. Integrated, science-based pest management strategies including the prudent use of appropriate pesticides will contribute to ensuring the sustainability of aquatic communities in areas subjected to pest management programs.

References

1. Hauer, F.R.; Lamberti, G.A. *Methods in Stream Ecology,* 2nd Ed.; Elsevier: Amsterdam, 2006.
2. Fleeger, J.W.; Carman, K.R.; Nisbet, R.M. Indirect effects of contaminants in aquatic ecosystems. Sci. Tot. Environ. **2003**, *317,* 207–233.
3. Rohr, J.R.; Crumrine, P.W. Effects of an herbicide and an insecticide on pond community structure and processes. Ecol. Appl. **2005**, *15,* 1135–1147.
4. Daily, G.C. Introduction: What are ecosystem services? In *Nature's Services: Societal Dependence on Natural Ecosystems*; Daily, G.D., Ed.; Island Press: Washington, DC, 1997; 1–10.
5. Anderson, J.; Gomez, W.C.; McCarney, G.; Adamowicz, W.; Chalifour, N.; Weber, M.; Elgie, S.; Howlett, M. *Ecosystem Service Valuation, Market-based Instruments and Sustainable Forest Management: A Primer*; Sustainable Forest Management Network: Edmonton, Alberta, 2010.
6. Bayon, R.; Jenkins, M. 2010. The business of biodiversity. Nature **2010**, *466,* 184–185.
7. Sibley, P.K.; Chappel, M.J.; George, T.K.; Solomon, K.R.; Liber, K. Integrating effects of stressors across levels of biological organization: Examples using organophosphorus insecticide mixtures in field-level exposures. J. Aquat. Eco-syst. Stress Recovery **2000**, 7, 117–130.
8. Campbell, P.J.; Arnold, D.J.S.; Brock, T.C.M.; Grandy, N.J.; Heger, W.; Heimbach, F.; Maund, S.J.; Streloke, M. *Guidance Document on Higher-Tier Aquatic Risk Assessment for Pesticides;* SETAC Europe: Brussels, Belgium, 1999.
9. Kennedy, J.H.; LaPoint, T.W.; Balci, P.; Stanley, J.K.; Johnson, Z.B. Model aquatic ecosystems in ecotoxicological research: Considerations of design, implementation, and analysis. In *Handbook of Ecotoxicology,* 2nd Ed.; Hoffman, D.J.; Rattner, B.A.; Burton, G.A., Jr.; Cairns, J., Jr., Eds.; Lewis Publishers: Boca Raton, Florida, 2003; 45–74.
10. Widenfalk, A.; Svensson, J.M.; Goedkoop, W. Effects of the pesticides captan, deltamethrin, isoproturon, and pirimi-carb on the microbial community of a freshwater sediment. Environ. Toxicol. Chem. **2004**, *23,* 1920–1927.
11. Wojtaszek, B.F.; Buscarini, T.M.; Chartrand, D.T.; Stephenson, G.R.; Thompson, D.G. Effect of Release® herbicide on morality, avoidance response, and growth of amphibian larvae in two forest wetlands. Environ. Toxicol. Chem. **2005**, *24*, 2533–2544.
12. Barnthouse, L.W. Quantifying population recovery rates for ecological risk assessment. Environ. Toxicol. Chem. **2004**, 23, 500–508.
13. Maund, S.; Chapman, P.; Kedwards, T.; Tattersfield, L.; Matthiessen, P.; Warwick, R.; Smith, E. Application of multivariate statistics to ecotoxicological field studies. Environ. Toxicol. Chem. **1999**, *18,* 111–112.
14. Clarke, K. R.; Warwick, R.M. Change in Marine Communities: An Approach to Statistical Analysis and Interpretation, 2nd Ed.; PRIMER-E: Plymouth, U.K., 2001.
15. Kreutzweiser, D.P.; Capell, S.S.; Scarr, T.A. Community-level responses by stream insects to neem products containing azadirachtin. Environ. Toxicol. Chem. **2000**, *19,* 855–861.
16. Frieberg, N.; Lindstrom, M.; Kronvang, B.; Larsen, S.E. Macroinvertebrate/sediment relationships along a pesticide gradient in Danish streams. Hydrobiologia **2003**, *494,* 103–110.

17. Berenzen, N.; Kimke, T.; Schulz, H.K.; Schulz, R. Macroinvertebrate community structure in agricultural streams: Impact of run-off-related pesticide contamination. Ecotoxicol. Environ. Saf. **2005**, *60,* 37–46.
18. van den Brink, P.J.; ter Braak, C.J.F. Principal response curves: Analysis of time-dependent multivariate responses of biological community to stress. Environ. Toxicol. Chem. **1999**, *18,* 138–148.
19. USEPA. *Guidelines for Ecological Risk Assessment*; United States Environmental Protection Agency, Risk Assessment Forum: Washington, DC, 1998.
20. EUFRAM. *Introducing Probabilistic Methods into the Ecological Risk Assessment of Pesticides, Version 6*; European Framework for Risk Assessment of Pesticides (EUFRAM): York, U.K., 2005.
21. Delorme, P.; Francois, D.; Hart, C.; Hodge, V.; Kaminski, G.; Kriz, C.; Mulye, H.; Sebastien, R.; Takacs, P.; Wandel-maier, F. *Final Report for the PMRA Workshop: Assessment Endpoints for Environmental Protection*; Environmental Assessment Division, Pest Management Regulatory Agency, Health Canada: Ottawa, Ontario, 2005.
22. Suter, G.W.; Barnthouse, L.W.; Bartell, S.M.; Mill, T.; Mackay, D.; Patterson, S. *Ecological Risk Assessment*; Lewis Publishers: Boca Raton, Florida, 1993.
23. Reinert, K.H.; Bartell, S.M.; Biddinger, G.R., Eds. *Ecological Risk Assessment Decision-Support System: A Conceptual Design;* SETAC Press: Pensacola, Florida, 1998.
24. Maltby, L.; Blake, N.; Brock, T.C.M.; van den Brink, P.J. Insecticide species sensitivity distributions: Importance of test species selection and relevance to aquatic ecosystems. Environ. Toxicol. Chem. **2005**, *24,* 379–388.
25. Candelmo, A.C.; Deshpande, A.; Dockum, B.; Weis, P.; Weis, J.S. The effect of contaminated prey on feeding, activity, and growth of young-of-the-year bluefish, *Pomato-mus saltatrix,* in the laboratory. Estuaries Coasts **2010**, *33,* 1025–1038.
26. Baldwin, D.H.; Spromberg, J.A.; Collier, T.K.; Scholz, N.L. A fish of many scales: Extrapolating sublethal pesticide exposures to the productivity of wild salmon populations. Ecol. Appl. **2009**, *19,* 2004–2015.
27. Pestana, J.L.T.; Loureiro, S.; Baird, D.J.; Soares, A.M.V.M. Pesticide exposure and inducible antipredator responses in the zooplankton grazer, *Daphnia magna* Straus. Chemo-sphere **2010**, *78,* 241–248.
28. Stark, J.D.; Banks, J.E.; Vargas, R.I. How risky is risk assessment: The role that life history strategies play in susceptibility of species to stress. Proc. Natl. Acad. Sci. U. S. A. **2004**, *101,* 732–736.
29. van den Brink, P.J.; Hattink, J.; Bransen, F.; van Donk, E.; Brock, T.C.M. Impact of the fungicide carbendazim in freshwater microcosms. II. Zooplankton, primary producers and final conclusions. Aquat. Toxicol. **2000**, *48*, 251–264.
30. Kreutzweiser, D.P.; Back, R.C.; Sutton, T.M.; Pangle, K.L.; Thompson, D.G. Aquatic mesocosm assessments of a neem (azadirachtin) insecticide at environmentally realistic con-centrations—2: Zooplankton community responses and recovery. Ecotoxicol. Environ. Saf. **2004**, *59,* 194–204.
31. Wang, M.; Grimm, V. Population models in pesticide risk assessment: Lessons for assessing population-level effects, recovery, and alternative exposure scenarios from modeling a small mammal. Environ. Toxicol. Chem. **2010**, *29,* 1292–1300.
32. Brodman, R.; Newman, W.D.; Laurie, K.; Osterfeld, S.; Lenzo, N. Interaction of an aquatic herbicide and predatory salamander density on wetland communities. J. Herpetol. **2010**, *44*, 69–82.
33. Rosenfeld, J.S. Functional redundancy in ecology and conservation. Oikos **2002**, *98,* 156–162.
34. Lawton, J.H.; Brown, V.K. Redundancy in ecosystems. In *Biodiversity and Ecosystem Function*; Schulze, E.D., Mooney, H.A., Eds.; Springer: New York, 1993; 255–268.
35. Naeem, S.; Li, S. Biodiversity enhances ecosystem stability. Nature **1997**, *390,* 507–509.
36. Kreutzweiser, D.P.; Good, K.P.; Chartrand, D.T.; Scarr, T.A.; Thompson, D.G. Are leaves that fall from imidacloprid-treated maple trees to control Asian longhorned beetles toxic to non-target decomposer organisms? Journal of Environmental Quality **2008**, *37,* 639–646.

37. deNoyelles, F., Jr.; Dewey, S.L.; Huggins, D.G.; Kettle, W.D. Aquatic mesocosms in ecological effects testing: Detecting direct and indirect effects of pesticides. In *Aquatic Mesocosm Studies in Ecological Risk Assessment*; Graney, R.L., Kennedy, J.H., Rodgers, J.H., Jr., Eds.; Lewis Publishers: Boca Raton, 1994; 577–603.
38. Boyle, T.P.; Fairchild, J.F.; Robinson-Wilson, E.F.; Haver-land, P.S.; Lebo, J.A. Ecological restructuring in experimental aquatic mesocosms due to the application of diflubenzuron. Environ. Toxicol. Chem. **1996**, *15,* 1806–1814.
39. George, T.K.; Liber, K.; Solomon, K.R.; Sibley, P.K. Assessment of the probabilistic ecological risk assessment-toxic equivalent combination approach for evaluating pesticide mixture toxicity to zooplankton in outdoor microcosms. Archives of Environmental Contamination and Toxicology **2003**, *45,* 453–461.
40. Kreutzweiser, D.P.; Sutton, T.M.; Back, R.C.; Pangle, K.L.; Thompson, D.G. Some ecological implications of a neem (azadirachtin) insecticide disturbance to zooplankton communities in forest pond enclosures. Aquat. Toxicol. **2004**, *67,* 239–254.
41. van Wijngaarden, R.P.A.; Cuppen, J.G.M.; Arts, G.H.P.; Crum, S.J.H.; van den Hoorn, M.W.; van den Brink, P.J.; Brock, T.C.M. Aquatic risk assessment of a realistic exposure to pesticides used in bulb crops: a microcosm study. Environ. Toxicol. Chem. **2004**, *23,* 1479–1498.
42. Relyea, R.A.; Hoverman, J.T. Interactive effects of predators and a pesticide on aquatic communities. Oikos **2008**, *117,* 1647–1658.
43. Zhang, X.; Liu, X.; Zhang, M.; Dahlgren, R.A.; Eitzel, M. A review of vegetated buffers and a meta-analysis of their mitigation efficacy in reducing nonpoint source pollution. J. Environ. Qual. **2010**, *39,* 76–84.
44. van de Zande, J.C.; Porskamp, H.A.; Michielsen, J.M.; Holterman, H.J.; Juijsmans, J.M. Classification of spray applications for driftability to protect surface water. Aspects Appl. Biol. **2000**, 57, 57–64.
45. Luck, J.D.; Zandonadi, R.S.; Luck, B.D.; Shearer, S.A. Reducing pesticide over-application with map-based automatic boom section control on agricultural sprayers. Trans. Am. Soc. Agric. Biol. Eng. **2010**, 53, 685–690.
46. Pfleeger, T.G.; Olszyk, D.; Burdick, C.A.; King, G.; Kern, J.; Fletcher, J. Using a geographical information system to identify areas with potential of off-target pesticide exposure. Environ. Toxicol. Chem. **2006**, *25,* 2250–2259.
47. PMRA. *Regulatory Directive DIR2002–02: The PMRA Initiative for Reduced Risk Pesticides;* Pest Management Regulatory Agency, Health Canada Information Services: Ottawa, Ontario; 2002, availabe at http://www.hc-sc.gc.ca/pmra-arla/english/pdf/dir/dir2002–02-e.pdf. (accessed September 2010).
48. Thompson, D.G.; Kreutzweiser, D.P. A review of the environmental fate and effects of natural "reduced-risk" pesticides in Canada. In *Crop Protection Products for Organic Agriculture: Environmental, Health, and Efficacy Assessment, ACS Symposium Series 947*; Felsot, A.S., Racke, K. D., Eds.; American Chemical Society: Washington, DC, 2007; 245–274.
49. van Emden, H. Integrated pest management. In *Encyclopedia of Pest Management*; Pimentel, D., Ed.; Marcel Dekker Inc.: New York, 2002; 413–415.
50. Solomon, K.R.; Takacs, P. Probabilistic ecological risk assessment using species sensitivity distributions. In *Species Sensitivity Distributions in Ecotoxicology*; Posthuma, L., Suter, G.W., Traas, T.P., Eds.; Lewis Publishers: Boca Raton, Florida, 2002; 285–314.
51. Solomon, K.R.; Baker, D.B.; Richards, R.P.; Dixon, K.R.; Klaine, S.J.; La Point, T.W.; Kendall, R.J.; Weisskopf, C.P.; Giddings, J.M.; Giesy, J.P.; Hall, L.W.; Williams, W.M. Ecological risk assessment of atrazine in North American surface waters. Environ. Toxicol. Chem. **1996**, *15,* 31–76.
52. Hanson, M.L.; Solomon, K.R. New technique for estimating thresholds of toxicity in ecological risk assessment. Environ. Sci. Technol. **2002**, *36,* 3257–3264.

53. Warren-Hicks, W.J.; Hart, A., Eds. *Application of Uncertainty Analysis to Ecological Risks of Pesticides*; CRC Press: Boca Raton, Florida, 2010.
54. Thorbek, P.; Forbes, V.E.; Heimbach, F.; Hommen, U.; Thulke, H.; van den Brink, P.; Wogram, J.; Grimm, V. *Ecological Models for Regulatory Risk Assessments of Pesticides;* SETAC Press: Pensacola, Florida, 2010.
55. Stark, J.D.; Banks, J.E. Population-level effects of pesticides and other toxicants on arthropods. Annu. Rev. Entomol. **2003**, *48,* 505–519.
56. van den Brink, P.J.; Roelsma, J.; van Nes, E.H.; Scheffer, M.; Brock, T.C.M. PERPEST model: A case-based reasoning approach to predict ecological risks of pesticides. Environ. Toxicol. Chem. **2002**, *21,* 2500–2506.
57. Baird, D.J.; Baker, C.J.O.; Brua, R.B.; Hajibabaei, M.; McNicol, K.; Pascoe, T.J.; de Zwart, D. Towards a knowledge infrastructure for traits-based ecological risk assessment. Integr. Environ. Assess. Manage. **2010**, online DOI 10.1002/ieam.129 (accessed September 2010).
58. van den Brink, P.J. Ecological risk assessment: From bookkeeping to chemical stress ecology. Environ. Sci. Technol. **2008**, 42, 8999–9004.
59. Beldon, J.B.; Gilliom, R.J.; Lydy, M.J. How well can we predict the toxicity of pesticide mixtures to aquatic life? Integr. Environ. Assess. Manage. **2007**, 3, 364–372.
60. Heugens, E.H.W.; Hendricks, A.J.; Dekker, T.; van Straalen, N.M.; Admiraal, W. A review of the effects of multiple stressors on aquatic organisms and analysis of uncertainty factors for use in risk assessment. Crit. Rev. Toxicol. **2001**, *31,* 247–285.
61. Lydy, M.J.; Lohner, T.W.; Fisher, S.W. Influence of pH, temperature and sediment type on the toxicity, accumulation and degradation of parathion in aquatic systems. Aquat. Toxicol. **1990**, 17, 27–44.
62. Howe, G.E.; Marking, L.L.; Bills, T.D.; Rach, J.J.; Mayer, F.L. Jr. Effects of water temperature and pH on toxicity of terbufos, trichlorfon, 4-nitrophenol and 2,4-dinitrophenol to the amphipod *Gammarus pseudolimnaeus* and rainbow trout (*Oncorhynchus mykiss*). Environ. Toxicol. Chem. **1994**, *13,* 51–66.
63. Yang, W.; Spurlock, F.; Liu, W.; Gan, J. Effects of dissolved organic matter on permethrin bioavailability to *Daphnia* species. J. Agric. Food Chem. **2006**, *54,* 3967–3972.
64. Puglis, H.J.; Boone, M.D. Effects of technical-grade active ingredient vs. commercial formulation of seven pesticides in the presence or absence of UV radiation on survival of green frog tadpoles. Arch. Environ. Contam. Toxicol. **2010**, online DOI 10.1007/s00244–010–9528-z (accessed September 2010).
65. Sandland, G.J.; Carmosini, N. Combined effects of a herbicide (atrazine) and predation on the life history of a pond snail, *Physa gyrina*. Environ. Toxicol. Chem. **2006**, *25*, 2216–2220.
66. Davidson, C.; Knapp, R.A. Multiple stressors and amphibian declines: Dual impacts of pesticides and fish on yellowlegged frogs. Ecol. Appl. **2007**, *17,* 587–597.
67. Barry, M.J.; Logan, D.C.; Ahokas, J.T.; Holdway, D.A. Effect of algal food concentration on toxicity of tow agricultural pesticides to *Daphnia carinata*. Ecotoxicol. Environ. Saf. **1995**, *32,* 273–279.
68. Wu, Q.; Riise, G.; Pflugmacher, S.; Greulich, K.; Steinberg, C.E.W. Combined effects of the fungicide propiconazole and agricultural runoff sediments on the aquatic bryophyte *Vesicularia dubyana*. Environ. Toxicol. Chem. **2005**, *24*, 2285–2290.
69. Boone, M.D.; Bridges, C.M.; Fairchild, J.F.; Little, E.E. Multiple sublethal chemicals negatively affect tadpoles of the green frog, *Rana clamitans*. Environ. Toxicol. Chem. **2005**, *24,* 1267–1280.
70. Laskowski, R.; Bednarska, A.J.; Kramarz, P.E.; Loureiro, S.; Schell, V.; Kudlek, J.; Holmstrup, M. Interactions between toxic chemicals and natural environmental factors— A meta-analysis and case studies. Sci. Tot. Environ. **2010**, *408*, 3763–3774.

2

Coastal Water: Pollution

Piotr Szefer

Introduction

The anthropogenic activity of man in coastal regions and even in areas located far inland is responsible for generating a huge amount of pollutants that are transported to marine ecosystems directly or by means of coastal watersheds, rivers, and precipitation from air.

Therefore, water pollution is a key global problem that has threatened marine organisms, including edible ones, and marine life in general.

There are two types of water pollutants, i.e., point source and nonpoint source.

The point source type is attributable to harmful contaminants released directly to the aquatic environment while nonpoint source delivers pollutants indirectly to the site of their approach.

The former one is a single, well-localized source, e.g., directly discharging sewage or industrial waste to the sea, whereas in the latter, the source of pollution is not well defined. Examples of such nonpoint source are agricultural runoff and windblown debris. Nonpoint sources are considered to be much more difficult to control and regulate as compared to point source pollutants.

The following are the classifications of other sources of pollution in coastal waters:

- Discharge of sewage and industrial waste
- Exploration and exploitation of the seabed
- Accidental pollution by oil and other pollutants from the land via air and other routes

Among these sources of pollutants dominate those connected with the discharge of municipal sewage and industrial wastes into coastal or estuarine regions, especially in the case of their inadequate treatment to remove persistent and harmful compounds. However, natural (and not anthropogenic) phenomena (e.g., volcanoes, storms, algae blooms, earthquakes, and geysers) could also be responsible for polluting aquatic systems. Their influence causes crucial changes in the ecological status of aquatic ecosystems.

The following are factors that determine the severity of pollutants:[1]

- Chemical structure
- Concentration
- Persistence

Independent on their sources in the water, pollutants may be classified as those for which the environment has some or little/no absorptive capacity. They are named “stock pollutants” (e.g., persistent synthetic chemicals, non-biodegradable plastics, and heavy metals).[2]

Most marine pollutants have land origin. They are often transported via rivers from agricultural sources and also via atmospheric trajectory. A lot of pollutants may be taken up by various compartments (biotic and abiotic) of aquatic environments; some of them could be biomagnified along the successive members of the food chain. A good example of having such ability to biomagnify is mercury. Such biomagnification could have negative effect on the quality of the water and hence on the health of the plants, animals, and humans whose lives depend on the quality of aquatic environments.

It should be emphasized that coastal areas are generally damaged from pollution, resulting in considerable impact on commercial coastal and marine fisheries. The pollution problem is very complex because of its interactions, interconnectedness, and uncertainty.[3–5] Pollutants, independent of their origin (e.g., air, water, land), enter the ocean, whether earlier or later.[3] Spatial distribution patterns of contamination concentrations exhibit a trend of their increase during transition from the south to the northern part of all oceans, i.e., to areas neighboring with both industrial centers and concentration of main pollution sources.[6]

The following are considered major pollutants:[6–11] fertilizers, pesticides, and agrochemicals; domestic and municipal wastes and sewage sludge; oil and ship pollution; trace elements; radionuclides; organic compounds; plastics; sediments; eutrophication and algal bloom; biological pollution; noise pollution; and light.

Heavy Metals and Metalloids

In contrast to organic pollutants, e.g., Polichlorinated Biphenyls (PCBs), heavy metals occur as natural elements of particular abiotic and biotic components of continental and aquatic ecosystems. They are present at a natural background level in rocks, soils, sediments, water, and biota. Human industrial and agricultural activities result in the elevation of this natural level to sometimes significantly higher values.

Typical metal concentrations are generally observed in open waters of marine ecosystems, although these remote regions can be affected by elevated levels of trace elements of anthropogenic or volcanic origin. For instance, the atmosphere affects the oceans and continental matter facilitating metal fluxes between these two compartments. Therefore, the atmosphere is a very important component as it makes it possible to transport metals that are natural in origin into distances far from their sources, e.g., from areas closest to forest fires as well as windblown dust, vegetation, and sea aerosols.[12,13] These sources are responsible for contributing metals to the lower troposphere, and therefore, their transport is associated to local and regional wind patterns, in contrast to specific sources such as volcano eruption, which can be responsible for injecting particulate metals not only into the troposphere but also into the stratosphere. In the latter case, particulate metals can be transported long distances under the appropriate circumstances.[13] Another example of long-distance transport of metals from their sources is dust carried from the Sahara Desert resulting in deposition of Fe, Mn, Al, and trace elements across the Mediterranean Sea, Atlantic Ocean, and Caribbean Sea.[13,14] Therefore, wind-driven dust transports particulate metals far offshore, in contrast to riverine flux carrying greater pole of mineral components from continental material to the coastal waters. These metals are promptly deposited to bottom sediments or taken up by biota, especially by phytoplankton in the surface waters, and transferred next to the food chain, recycles or settled to the bottom.[13]

The mass of metal of anthropogenic origin emitted to ecosystems is now equal to or greater than the mass introduced to the natural cycle on a global scale.[13] Some metals, e.g., Pb, Hg, and Cd, owing to their great toxicity, pose a high health risk; therefore, great attention has been paid to estimate their inputs to marine ecosystems, particularly to coastal waters. For instance, the largest masses of Pb are emitted to the atmosphere during processing of the metal (smelters) or from combustion-related sources like motor vehicles. Lead from motor vehicle exhaust was identified not only in the atmosphere but also in remote surface waters as well as in remote terrestrial areas. A ban on Pb usage in vehicle fuel has resulted in the effective reduction of metal inputs to the ecosystem since the 1970s. The decline in atmospheric Pb detected over a time scale of 10 years (from 1979 to 1989) because of the reduction of leaded gasoline in western European countries should be reflected in decreasing Pb levels in surficial water and biota. In fact, the temporal negative trend for cod in the Baltic Sea seems to support this argumentation.[15]

According to Mason et al.,[16] preindustrial fluxes and reservoirs of Hg pose one-third of its fluxes in the civilization era. It is suggested that modern emission of Hg to the atmosphere increased considerably, even 4 to 5 times, due to the human activities. The extremely elevated levels of Hg are frequently associated with Hg mining.

As has been reported, ca. 300 metric tons of dissolved Cd annually enter the oceans from rivers while ca. 400 to 700 metric tons of dissolved and particulate Cd are annually deposited to the oceans from the atmosphere.[17,18] It is estimated that human activities have contributed to increased Cd inputs to the ocean by 60% in the 1980s. It is also found that the higher proportion of land deposition of Cd is associated to the rapid removal of this metal from the atmosphere near inputs of air pollution. A substantial pole of Cd transported by rivers is deposited in estuaries and continental margins of oceans. It is found that increased concentrations of Cd occur locally, reflecting its mosaic contamination, especially near mining and industrial point sources—not managed.

A significant fraction of Zn entering the oceans is derived from atmospheric deposition.[18] Soils and sediments are main natural reservoirs of Zn. Zinc, like Cd, is not distributed evenly across the Earth's surface, since its increased concentrations occur locally in the vicinity of increased inputs, i.e., specific points of source inputs.

The concentrations of many trace elements, e.g., Pb, Cd, Hg, Cu, Zn, Se, and As in coastal and estuarine waters, especially in highly industrialized areas, are generally significantly greater than those in open oceanic waters.[13,18–20] The waters of harbors and marinas around the world contain variable concentrations of tributyltin (TBT),[13] but its extremely high levels may be characteristic for marinas in southwest England.[21]

Human industrial and agricultural activities affect inputs of several metals to reservoir/reservoirs and hence increase their concentrations, even sometimes very dramatically, above natural background levels. There are numerous examples of worldwide events leading to serious contamination of coastal waters by heavy metals and metalloids.[15] Therefore, relationships between man and ecosystem health have been explored, especially in relation to perturbed ecosystems. This includes the pollution status of coastal regions harmed by some catastrophes, large-scale pollution, environmental accidents and episodes, etc. High risk groups consume extremely high quantities of trace metals present in specific assortment of seafood or offal and it concerns seaside populations. Marine fish and shellfish may be the dominant dietary sources of Hg for local populations.[22–25] A notable example of aquatic pollution by a toxic metal is the Minamata incident, commencing in 1953 and resulting in fish, shellfish, and bird mortalities in waters of the partially landlocked Minamata Bay.[15] Dogs, pigs, and especially cats were also victims of this incident. By the end of 1974, 107 of 798 officially verified patients had died. According to Tomiyasu et al.,[26] the sediments from the Minamata Bay contained levels of Hg that highly exceeded its background level. Among other incidents resulting in the release of Hg compounds to the environment, the most significant ones happened in the 1960s and early 1970s in Sweden, Canada and the United States, northern Iraq, Guatemala, Pakistan, and Ghana.[27–31] MeHg in aquatic ecosystems, especially those that bioaccumulated in fish, is a major public health problem all over the world.[32] Its levels in the hair of fishermen represent the critical group for dietary exposure. For instance, the concentrations of Hg (total and MeHg) in the hair of fishermen from Kuwait were 2 times higher than the "normal" level according to the World Health Organization.[33] Biomass burning in tropical forests also seems to have contributed significantly to the Hg input to the atmosphere. Approximately 31% of the Hg concentrations were associated with the vegetation fire component.[34] It is postulated (based on long-range air mass trajectory analyses) that Hg occurs in the Amazon basin over two main routes: to the South Atlantic and to the Tropical Pacific, over the Andes.[34]

Global emission flux estimates exhibited that biomass burning could be major contributor of heavy metals and black carbon to the atmosphere.[15] It is estimated that savannah and tropical forest biomass burning could emit huge amounts of Cu, Zn, and black carbon to the atmosphere, corresponding to 2%, 3%, and 12%, respectively, of the global level of these elements.[35]

The toxic effects of TBT were first indicated towards the end of the 1970s in Arcachon Bay, France, as the "TBT problem."[36,37] The release of TBT (from antifouling paints) to the area resulted in shell abnormalities and reduced growth and settlement in oysters, *Crassostrea gigas,* cultured in the vicinity of marinas. In much polluted water, oyster production was severely affected by the absence of reproduction, resulting in a strong decline in the marketable value of the remaining stock.[37] Imposex, i.e., the development of male sexual characteristics in female marine mesogastropoda and neogastropoda caused by TBT pollution, is a widespread phenomenon concerning several coastal species and, more recently, offshore species as well.[23,38,39] Subsequent regulations in 1990 that prohibit the use of TBT-based antifoulants on vessels less than 25 m in length have been highly effective in reducing TBT levels in coastal waters. However, larger vessels have continued the release of TBT, and major harbors remained pollution hot spots.[40] The Organotin Antifouling Paint Control Act restricted in the United States the use of TBT paints to vessels greater than 25 m in length.[41] The voluntary stoppage of TBT production in January 2001 by major U.S. and European manufacturers resulted in the decline of its presence in marine biota, but TBT paint is still being used in most Asian countries. The International Maritime Organization (IMO) imposed an international ban for the use of organotin compounds in antifouling treatments on ships longer than 25 m. The target is to prohibit their application starting 2003 and to require the removal of TBT from ships' hulls by January 1, 2008.[41,42]

The extensive flooding, especially occurred in river area of former or operating metalliferous mining can be responsible for wide-spreading of heavy metals and metalloids far distance from pollution source. An example of such environmental events is the flooding of the Severn catchment (United Kingdom) in January 1998.[43]

Radionuclides

Physicochemical aspects and applications of radioactivity in the environment were extensively presented in a book by Valcovic.[8] There are numerous papers reporting on problems resulting from radionuclide pollution and their sources in different ecosystems.[15,44] One of the first low- level emissions of radioactivity took place in the Hanford reactors (Columbia River, Washington, United States), which released radionuclides (mainly ^{60}Co, ^{51}Cr, and ^{65}Zn) to its environs from 1940 to 1971.[28] The nuclear reactors in Cumbria (northwest England) have also been responsible for discharging quantities of radioisotopes, i.e., ^{144}Ce, ^{137}Cs, ^{95}Nb, ^{106}Ru, and ^{95}Zr, to the marine environment. Although these emissions have been diminished recently, discharges from nuclear power stations such as Sellafield (formerly named Windscale) could still be identified, even in distances far away from their source.[45,46] Significant quantities of artificial radionuclides (^{137}Cs, ^{134}Cs, ^{90}Sr, ^{99}Tc) have been transported to the North Atlantic and Arctic from Sellafield, together with measurable amounts of Pu and Am.[46–48] The nuclear reprocessing plant at La Hague in France emitted ^{137}Cs and $^{239+240}$Pu to the environment, although this plant mainly supplies ^{129}I and ^{125}Sb.[28,48,49] Besides the expected emission of radionuclides from nuclear and reprocessing facilities, significant quantities of radioisotopes contaminate aquatic and terrestrial environments from either nuclear weapons testing or nuclear reactor accidents.[15] For instance, the thermonuclear detonation that took place in 1954 at Bikini Atoll resulted in the contamination of a large area of the Marshall Islands. A number of atmospheric tests (520 in total) were mostly carried out in the Northern Hemisphere, including eight underwater tests, with a total yield of 542 Mt. Moreover, there have been a total of 1352 underground tests with a total yield of 90 Mt.[50]

A number of nuclear incidents were concernedly noted, including those affecting the crew of the Japanese fishing vessel "Fukuru Maru".[28] Plutonium released from the Kyshtym accident in the Urals has been much probably detected in deep basins of the Arctic Ocean.[51] In 1968, an aircraft from the U.S. Strategic Air Command crashed near the Thule Airbase in NW Greenland, releasing to the marine environment ca. 1 TBq $^{239+240}$Pu.[52] As a consequence, marine sediments as well as benthic organisms, i.e., bivalves, shrimps, and sea stars, have been contaminated by Pu, although their levels rapidly decreased.[53] A number of American and Russian nuclear submarines have been lost in the world's oceans. For instance, the Soviet Komsomolets submarine sank at a depth of 1700 m at Bear Island in the eastern part of the Norwegian Sea. The estimated radioactivity in the wreck was 2.8 PBq ^{90}Sr and 3 PBq ^{137}Cs.[54] Some nuclear powered satellites can incidentally be sources of radioactivity. They can burn up in the upper atmosphere, resulting in the contamination of the ocean. For instance, such an accident happened in 1964 when a SNAP-9A nuclear power generator containing 0.6 PBq ^{238}Pu aboard a U.S. satellite re-entered the atmosphere in the Southern Hemisphere. The estimated ^{238}Pu/$^{239+240}$Pu ratio in this region was higher than that in the ocean water from the Northern Hemisphere.[55,56]

Sea dumping was carried out since the late 1940s to mid-1960s mainly by the United States in the Atlantic Ocean and Pacific Ocean as well as by the United Kingdom in the Northeast Atlantic Ocean.[56] In 1967, an international operation was initiated by the former European Nuclear Energy Agency that contributed to the deposition of ca. 0.3 PBq solid waste at a depth of 5 km in the eastern Atlantic Ocean. Other international operations were continued until 1982 when ca. 0.7 PBq α activity, 42 PBq β activity, and 15 PBq tritium activity have been dumped in the North Atlantic.[57] It has been assessed that the radiological impact of the NEA (former European Nuclear Energy Agency) dumping activities resulted in some releases of Pu from the dumped waste.[15] This source would be responsible for only a part of the total body burden radioactivity in local benthic organisms, e.g., sea cucumbers; the remainder has been attributed to fallout.[58] According to Consortium for Risk Evaluation with Stakeholder Participation (CRESP) evaluation, the individual dose of a critical group consuming seafood such as molluscs from the Antarctic Ocean was estimated to be 0.1 μSv yr^{-1}, in effect labeling ^{239}Pu and ^{241}Am as critical radionuclides. The indefinite collective dose to the world's population coming from sea dumping was estimated at 40,000 manSv with predominance of ^{14}C and ^{239}Pu.[56,58]

U.S. weapons production facilities account for a large fraction of radiocaesium discharges during the 1950s.[15] A striking incident occurred at Chernobyl in the former USSR where an explosion of a reactor core of the nuclear plant took place in April 1986. The Baltic countries and a large part of central and western Europe have been contaminated principally by ^{131}I, ^{134}Cs, and ^{137}Cs.[28,59] It is found that a significant part of the activity fell over the European marginal seas from which the Baltic Sea was the most affected by contamination.[56,60] It has been mainly responsible for additional inflow of the radioactive contaminants to the Northeast Atlantic Ocean.[56] Due to the Chernobyl accident, significant levels of ^{137}Cs were also found in the Black Sea. The outflow from this Sea has been the major source of additional ^{137}Cs in the Mediterranean Sea.[56] In the summer of 1987, the Chernobyl-derived ^{137}Cs was also detected in surficial waters of the Greenland Sea, Norwegian Sea, and Barents Sea as well as in the west coast of Norway and the Faroe Islands. According to Aarkrog,[56] the total Chernobyl ^{137}Cs input to the world's oceans was relatively significantly smaller than that estimated for nuclear weapons fallout because of the tropospheric nature of this accident that has contaminated the surrounding European continental areas.[15]

After the 2011 Tōhoku earthquake and tsunami, the radiation effects from the Fukushima Daiichi nuclear disaster resulted in the release of radioactive isotopes from the crippled Fukushima Daiichi

Nuclear Power Plant. The total amount of ^{131}I and ^{137}Cs released into the atmosphere has been estimated to exceed 10% of the emissions from the Chernobyl disaster. Large amounts of radioactive isotopes have also been released into the Pacific Ocean.[61]

Organic Compounds

The high lipophilicity of many persistent organic pollutants (POPs) enhances their bioconcentration/biomagnification, resulting in potential health hazards on predators at higher trophic levels, including humans. These xenobiotics occur widely in coastal waters and oceans from the Arctic to the Antarctic and from intertidal to abyssal. It should be emphasized that most of these compounds exist at a very low concentration level, and hence, their threat to marine biota is still not well recognized. However, it is well known that exposure to extremely low levels of halogenated hydrocarbons, e.g., PCBs, Dichlorodiphenyltrichloroethane (DDT), and TBT, may disrupt the normal metabolism of sex hormones in fish, birds, and marine mammals. Moreover, sublethal effects of these organic chemicals over long-term exposure may result in serious damage to marine populations since some of these POPs may impair reproduction functions of organisms while others may show carcinogenic, mutagenic, or teratogenic activity.[6] Some of the effects of these compounds have been reported by Goldberg.[62] For instance, very low levels of TBT (as endocrine disruptor) cause a significant disruption in sex hormone metabolism, resulting in the malformation of oviducts and suppression of oogenesis in female whelks, e.g., *Nucella lapillus*.[63] As a consequence, sex imbalance leads to species decline if not species extinction in some field populations.[64] Butyltins may be responsible for mass mortality events of bottlenose dolphins in Florida through suppression of the immune system.[65] Trace environmental levels of other compounds like chlorinated hydrocarbons, organophosphates, and diethylstilbestrol may be responsible for significant endocrine disruption and reproductive failure in different groups of animals, i.e., marine invertebrates, fish, birds, reptiles, and mammals.[6] For instance, high levels of DDT, PCBs, and organochlorines in the Baltic Sea significantly reduced the hatching rates of the fish-eating whitetailed eagle *(Haliaeetus albicilla)* in the 1960s and the 1970s.[66] Another example of the toxic impact of POPs is organochlorine contamination in different cetacean species dependent upon their diet, sex, age, and behavior. Many of these compounds, as endocrine disruptors, reduce reproduction and/or suppress immune function. DDT and PCBs are known as compounds affecting steroid reproductive hormones and can increase mammalian vulnerability to bacterial and viral diseases. Jepson et al.[67] reported a statistically significant relationship between elevated PCB level and infectious disease mortality of harbor porpoises *(Phocoena phocoena).*

The assessment and monitoring of existing and emerging chemicals in the European marine and coastal environment have been overviewed based on numerous, most recent worldwide references.[5] From this report, the extensive range of chemicals that are capable of disrupting the endocrine systems of animals can be categorized into the following: environmental estrogens (e.g., bisphenol A, methoxychlor, octylphenol, and nonylphenol), environmental anti-estrogens (e.g., dioxin, endosulfan, and tamoxifen), environmental anti-androgens [e.g., dichlorodiphenyldichloroethylene (DDE), procymidone, and vinclozolin], chemicals that reduce steroid hormone levels (e.g., fenarimol and ketoconazole), chemicals that affect reproduction primarily through effects on the central nervous system (e.g., dithiocarbamate pesticides, and methanol), and chemicals with multiple mechanisms of endocrine action (e.g., phthalates and TBT). There is a high level of international concern regarding developmental and reproductive impacts on marine organisms from exposure to endocrine- disrupting chemicals. This is the case for "new" substances such as alkylphenols; there is also renewed interest for some "old" organochlorines such as DDT and its metabolites. Brominated flame retardants (BFRs), particularly the brominated diphenyl ethers (BDEs) and hexabromocyclododecane (HBCD), have been detected in the European marine environment. It has been reported that the input of BDEs into the Baltic Sea through atmospheric deposition now exceeds that of PCBs by almost a factor of 40. BDEs are found in fish from various geographic regions. This resulted from the long-range atmospheric transport and deposition of these substances.[5] HBCD was detected in liver and blubber samples from harbor seals and harbor porpoises from the Wadden Sea and the North Sea. It is found that environmental concentrations of these BFRs in Japan and South China increased significantly during the last decades. PBDE levels in marine mammals and sediments from Japan, after showing peak concentrations in the 1990s, appear to have leveled off in recent years. Furthermore, in recent years, HBCD concentrations in marine mammals from Japanese waters appear to exceed those of PBDEs, presumably reflecting the increasing use of HBCDs over PBDEs. Pentabromotoluene (PBT) and Decarbomodiphenyl (DBDPE), for example, have been found in Arctic samples remote from sources of contamination. It is an indication of their potential for long-range atmospheric transport, showing a tendency for accumulation in top predators. Polymeric BFRs may be a source of emerging brominated organic compounds to the environment. Medium- and short-chain chlorinated paraffins (SC- CPs) are ubiquitous in the environment and tend to behave in a similar way to POPs. They have been found in water as well as in fish and marine mammals.[5]

Perfluorinated compounds (PFCs), namely, perfluorooctane sulfonate (PFOS), have been detected in marine mammals.[5] They are globally distributed anthropogenic contaminants. PFCs, such as PFOS, have been industrially manufactured for more than 50 years and their production and use have increased considerably since the early 1980s. The main producer of PFOS voluntarily ceased its production in 2002. Furthermore, the large-scale use of PFOS has been restricted. PFOS has been used in many industrial applications such as fire-fighting foams and consumer applications such as surface coatings for carpets, furniture, and paper. PFCs are released into the environment during the production and use of products containing these compounds. About 350 polyfluorinated compounds of different chemical structures are known.[5] The most widely known are PFOS (C8F17SO3) and perfluorooctanoic acid (PFOA; C8F15O2), which are chemically stable and thus may be persistent (substance dependent). PFCs do not accumulate in lipid but instead accumulate in the liver, gallbladder, and blood, where they bind to proteins. PFCs have been detected worldwide, including the Arctic Ocean and Antarctic Ocean, in almost all matrices of the environment. High concentrations of PFCs have been found in marine mammals.[5] A screening project in Greenland and the Faroe Islands indicated high biomagnification of PFCs, with elevated concentrations in polar bear liver. A time trend study (1983–2003) showed increasing concentrations for all PFCs for ringed seals from East Greenland. In the United Kingdom, a study on stranded and by-catch harbor porpoise liver (1992 and 2003) found PFOS at up to 2420 pg kg^{-1} wet weight. There is a decreasing trend going from south to north.[5]

Antifouling paint booster biocides were recently introduced as alternatives to organotin compounds in antifouling products.[5] These replacement products are generally based on copper metal oxides and organic biocides. Commonly used biocides in today's antifouling paints are as follows: Irgarol 1051, diuron, Sea-Nine 211, dichlofluanid, chlorothalonil, zinc pyrithione, TCMS (2,3,3,6-tetra- chloro-4-methylsulfonyl) pyridine, TCMTB [2-(thiocyanomethylthio) benzothiazole], and zineb. It has been reported that the presence of these biocides in coastal environments around the world is a result of their increased use (notably in Australia, the Caribbean, Europe, Japan, Singapore, and the United States). For example, Irgarol 1051, the Irgarol 1051 degradation product GS26575, diuron, and three diuron degradation products [1-(3-chlorophenyl)-3,1-di- methylurea (CPDU), 1-(3,4-dichlorophenyl)-3-methylu- rea (DCPMU), and 1-(3,4-dichlorophenyl)urea (DCPU)] were all detected in marine surface waters and some sediments in the United Kingdom. Risk assessments indicate that the predicted levels of chlorothalonil, Sea-Nine 211, and dichlofluanid, in contrast to Irgarol 1051, in marinas represent a risk to marine invertebrates. Finally, non-eroding silicone-based coatings can effectively reduce fouling of ship hulls and are an alternative to biocidal and heavy- metal-based antifouling paints. Although polydimethylsiloxanes (PDMSs) are unable to bioaccumulate in marine organisms and their soluble fractions have low toxicity to marine biota, undissolved silicone oil films or droplets can cause physical–mechanical effects such as trapping and suffocation of organisms.[5]

Human and veterinary pharmaceuticals are designed to have a specific mode of action, affecting the activity of, e.g., an enzyme, ion channel, receptor, or transporter protein.[5] Clotrimazole, dextropropoxyphene, erythromycin, ibuprofen, propranolol, tamoxifen, and trimethoprim were detected in U.K. coastal waters and in U.K. estuaries. Concentrations of some pharmaceutical compounds are effectively reduced during their passage through a tertiary wastewater treatment works, while others are sufficiently persistent to end up in estuaries and coastal waters.[5] Compared with mammalian and freshwater organisms, there is a lack of experimental data on the impacts of pharmaceuticals in marine and estuarine species. However, there is experimental evidence that selected pharmaceuticals have the potential to cause sublethal effects in a variety of organisms. It has been concluded that antibiotic substances in marine ecosystems can pose a potential threat to bacterial diversity, nutrient recycling, and removal of other chemical pollutants. Although data on the occurrence of pharmaceuticals and antibiotics in the marine environment are becoming more available, the true extent of the potential risks posed by this group of contaminants cannot, at present, be assessed, mainly due the lack of effect data.[5]

Several studies showed that among personal care products (PCPs), synthetic musks (nitromusks, polycyclic musks, and macrocyclic musks) are widespread in marine and freshwater environments and bioaccumulate in fish and invertebrates.[5] There were identified products such as benzotriazole organic UV filters, namely, UV-320 [2-(3,5- di-i-butyl-2-hydroxyphenyl)benzotriazole], UV-326 [2-(3- i-butyl-2-hydroxy-5-ethylphenyl)-5-chlorobenzotriazole], UV-327 [2,4-di-*t*-butyl-6-(5-chloro-2*H*-benzotriazol-2-yl) phenol], and UV-328 [2-(2*H*-benzotriazol-2yl)-4,6-di-*t*- pentylphenol]. Their relatively high concentrations were found in marine organisms collected from waters of western Japan. There are indications that marine mammals and seabirds accumulate UV-326, UV-328, and UV-327. Benzotriazole UV filters were also detected in surface sediments from this area. The results suggest a significant bioaccumulation of UV filters through the marine food webs and a strong adsorption to sediments. Although a full risk assessment of some of these has been performed (e.g., musks), for most PCPs, there is little data on their occurrence and their effects in the marine environment.[5]

Biological Pollution

Eutrophication and Algal Bloom

Nutrient loadings in coastal waters cause direct responses such as changes in chlorophyll, primary production, macro- and microalgal biomass, sedimentation of organic matter, altered nutrient ratios, and harmful algal blooms. The indirect responses of nutrient loadings are responsible for changes in benthos biomass, benthos community structure, benthic macrophytes, habitat quality, water transparency, sediment organic matter, sediment biogeochemistry, dissolved oxygen, mortality of aquatic organisms, food web structure, etc. Moreover, increase in phytoplankton biomass and attributing decrease in transparency and light intensity limit growth of submerged vascular plants.[6,68] Generally speaking, eutrophication leads to major changes in qualitative and quantitative species composition, structure, and function of marine communities over large areas. As for phytoplankton communities, such changes are connected with an increase in biomass and productivity.[69] For instance, a general shift from diatoms to dinoflagellates, as well as dominance of small-size nanoplankton (microflagellates, coccoids), has been reported. Similar trends were observed in the case of zooplankton communities, indicating replacement of herbivorous copepods by small- size zooplankton.[70,71] Some examples of consequences of eutrophication have been reported based on worldwide references.[15] The harmful deoxygenation of water giving rise to fish kills was producing nutrient-derived large mats of macroalgae in the Peel-Harvey Estuary, Western Australia.[72] Similar events took place in the northern Adriatic Sea where diatom blooming in summer resulted in the production of mucilage, affecting tourism in northeastern Italy and reducing fish catch.[28,73,74] Insufficient water exchange and increasing production of organic matter during this century caused depletion of O_2 in all deep waters of the Baltic Proper.[15] It resulted in devastating consequences for marine biota, leading to the replacement of O_2 by H_2S in these bottom waters.[75] Although eutrophication generally leads to an increase in fish productivity, it can also cause negative environmental changes in fish populations. Fish such as cod and plaice are threatened by O_2 depletion in Baltic deep basins, causing decreasing fish catch in Koge Bay in the Sound.[75]

The blooms of blue-green algae as well as *Nodularia* produce a toxic peptide hepatoxin under particular conditions, which can pass through the food web, affecting top consumers, e.g., man. The toxin is responsible for the degeneration of liver cells, promoting tumors and causing death from hepatic hemorrhage.[75] Paralytic shell poisoning (PSP) and/or ciguatera has/have been identified predominantly in the subtropical and tropical zones such as Australia[76–80] and especially in other Indo-Pacific regions, e.g., India, Thailand, Indonesia, Philippines, and Papua New Guinea.[81,82] Principal toxic dinoflagellate species, i.e., *Pyrodinium bahamense* var. *compressa,* killed many fish and shellfish from these regions.[15] The consumption of seafood in the Indo-Pacific area posed considerable public health problems.[28] The significant PSP incidences also took place in temperate zones. For instance, in May 1968, a poisoning episode affected 78 persons inhabiting Britain after consumption of soft tissue of the blue mussel *Mytilus edulis.*[83] Another dinoflagellate-poisoning event again happened in northeast England in the summer of 1990, possibly attributed to a specific combination of elevated nutrient inputs from rivers and exceptionally warm weather conditions, which could be favorable for algae growing.[28]

It has been reported that anthropogenically derived atmospheric N deposition to the North Atlantic Ocean was strictly responsible for harmful algal bloom expansion.[84] This event concerned especially the Eastern Gulf of Mexico, U.S. Atlantic coastal waters, the North Sea, and the Baltic Sea.[84–95] Expanding blooms of the noxious dinoflagellate *Alexandrium tamarense* have been observed along the Northeast U.S. Atlantic coastline.[84,92] There are numerous examples of specific harmful algal bloom expansions in coastal and off-shore waters in case of significant atmospheric deposition of N, e.g., in the North Sea, Adriatic Sea, Western Mediterranean Sea, and Baltic Sea.[84,96] Great attention has been paid to toxic hypoxia- inducing dinoflagellate blooms in the North Sea and the Western Baltic.[84] In the summer of 1991, a very extensive bloom of *Nodularia spumigena* in the open Baltic Sea and along the southern and southeastern Swedish coasts was observed. Dogs' mortalities caused by toxic *Nodularia* blooms have been observed in Denmark, Gotland, and the Swedish coastal waters.[15] In other Baltic areas, horses, cows, sheep, pigs, cats, birds, and fish also suffered from this event. *Nodularia* blooms have caused human health problems such as stomach complaints, headaches, eczema, and eye inflammation.[75] In the Skagerrak and Kattegat, harmful algal bloom expansion of toxic algae species such as *Prorocentrum, Dinophysis, Dichtyocha, Prymnesium,* and *Chrysochromulina* has taken place.[88] The recent blooms mostly killed pelagic organisms and the phyto- and zoobenthic organisms.

Invasive Species

The impacts of introduction and invasion of species throughout the world have recently been identified. There are an increasing number of reports that document this phenomenon taking place in coastal, estuarine, and marine waters.[6] For instance, the Chinese mitten crab *(Eriocheir sinensis),* as invasive species, now inhabits coastal regions in northwestern Europe, and it has caused damage to flood defense walls by burrowing, affecting local community structure.

Worldwide fish species introduction is connected with various consequences.[97] It has been pointed out that many aquaculture species are recently genetically modified. Such modified populations are frequently released and mixed with the natural populations and are breeding with them. It causes biological pollution from a molecular level to community and ecosystem levels. An example of such events is the flooding in Central Europe that caused the release of hybrid and modified fish like sturgeon *(Acipenser* spp.) from aquaculture installations.[98] The local populations of fish are generally not resistant to the pathogenic organisms carried by the introduced species and vice versa. Therefore, deliberate genetic selection and breeding for a long time may have numerous consequences in the aquaculture unit itself as well as the loss of the natural stock for numerous species in a global scale.[6,98,99]

Fertilizers and Pesticides

Agricultural activity as an important pollution source has contributed to significant enrichment of nutrients (mainly ammonium ion and nitrates) in coastal marine waters. It is found[100] that wastes, manures, and sludges provide soils with significantly more hazardous substances as compared to fertilizers for achieving the equivalent plant nutrient content. The worldwide use of fertilizers, including organic fertilizers like manure, is huge. In the case of intensively monocultivated areas, a relatively small number of pesticides have been widely used in spite of their variety.[6]

The large mass of pesticide residues is accumulated in the environment since they are not rapidly degradable. The total global DDT production from the 1940s to 2004 was estimated as ca. 4.5 Mt.[101] Duursma and Marchand[102] estimated the world production of DDT to be ca. 2.8 Mt, of which 25% is assumed to be released to the ocean. According to Shahidul Islam and Tanaka,[6] the total emission of DDT through agricultural applications amounts to 1030 kt between 1947 and 2000. Organochlorine pesticides (OCPs) originating mostly in temperate and warmer areas of the world can be transported to coastal waters and even via atmospheric long-range transport and ocean currents to the Arctic. Owing to their bioaccumulative abilities (as lipophilic compounds) and biomagnification along the sequential trophic levels of the food chain, pesticides are classified as one of the most destructive agents for marine organisms. As a consequence, their very high levels can be observed among top predators, including man. Their toxic effects to marine organisms are often complex because they may be associated with the combination of exposure to pesticides and other POPs with environmental stresses such as eutrophication and pathogens.[6]

Sewage Effluents

Sewage effluents contain industrial, municipal, and domestic wastes; animal remains; etc. The huge amounts of these effluents generated in big cities are transported by drainage systems into rivers or other aquatic systems, e.g., coastal waters. It is estimated that the annual production of sewage amounts to ca. 1.8×10^8 m^3 for a population of 800,000. This load is equivalent to an annual release of 3.6×10^3 tons of organic matter.[6] Sewages pose significant effects on coastal marine ecosystems because they contain POPs (heavy metals/trace elements, organic pollutants) as well as viral, bacterial, and protozoan pathogens and organic substances subjected to bacterial decay. In case of such bacterial activity, the content of oxygen in water is reduced, resulting in the destruction of proteins and other nitrogenous compounds. Releasing hydrogen sulfide and ammonia exhibits toxic activity to marine biota, even at low levels. As for pathogens, domestic sewage released to coastal waters contains such harmful pathogens as *Salmonella* spp., *Escherichia coli, Streptococcus* sp., *Staphylococcus aureus*, *Pseudomonas aeruginosa*, the fungi *Candida*, and viruses such as enterovirus, hepatitis, poliomyelitis, influenza, and herpes.[6] Different bacteria and viruses can be transferred to some representatives of marine fauna, e.g., marine mammals.

Oils

The recently observed increase in tanker operations and oil use as well as marine tanker catastrophes has been responsible for the presence of excessively large amounts of oil spillage in coastal and marine ecosystems. It is estimated that ca. 2.7 million tons of oil pollution enter the ocean each year. The tanker accidents between 1967 and 2007 released ca. 4.5 million tons of oil to seawater. Notable examples of ecological catastrophes are the huge spill from a drilling platform in Gulf of Mexico (Mexico) in 1979 and the Deepwater Horizon drilling rig explosion in the Gulf of Mexico (United States) (April 20 to July 15, 2010), resulting in massive amounts of oil in the gulf. Another similar example took place during the Persian Gulf War in 1991, where ca. 2 million tons of oil was spilled, resulting to the death of many species of marine biota.[7,103,104] Therefore, oil pollution poses serious adverse effects on aquatic environment and marine organisms represented different trophic levels from primary producers to the top predators.[6]

Although aerial and flying birds (e.g., gulls, gannets) are not seriously exposed to oil toxicity, birds that spend most of their time in contact with oil on the water surface (e.g., ducks, auks, divers, penguins) are at greater risk of oil toxicity. According to Smith,[105] the annual release of hydrocarbon can range from 0.6 to 1 million tons. Coastal refineries can be an important source of oil pollution since millions of gallons of crude oil and its fractions are processed and stored there. During their operation, pollutants are continuously released by way of leakages, spills, etc.

Marine Debris and Plastics

Marine debris, especially plastics, is one of the most pervasive pollution problems. Nets, food wrappers, bottles, resin pellets, etc., have serious impacts on humans and marine biota. Medical and personal hygiene debris can enter coastal water through direct sewage outflows, posing a serious threat to human health and safety. Contact with water contaminated with these pollutants and pathogens (e.g., *E. coli*) can result in infectious hepatitis, diarrhea, bacillary dysentery, skin rashes, typhoid, and cholera.[106]

There are numerous reviews devoted to an important topic such as pollution by marine debris.[106–110] Entanglement in marine debris such as nets, fishing line, ropes, etc., can hamper an organism's mobility, prevent it from eating, inflict wounds, and cause suffocation or drowning. It was estimated that 136 marine species have been involved in entanglement incidents, including some species of seabirds, marine mammals, and sea turtles.[111] The decline in the population of the northern sea lion *(Eumetopias jubatus*), endangered Hawaiian monk seal (*Monachus schauinslandi*), and northern fur seal has been explained by entanglement of young specimens in lost or discarded nets and packing bands.[112] Abandoned fishing gear, e.g., fishing net, can contribute to catching and killing marine animals. This process called ghost fishing or ghost net can kill a huge number of commercial species.[108] An example of another serious pollution problem is ingestion of debris by marine animals. Plastic pellets and plastic shopping bags can be swallowed and lodged in animals' throats and digestive tracts, causing some animals to stop eating and slowly starve to death.[106] According to the U.S. Marine Mammal Commission,[111] ingestion incidents concerned 111 species of seabirds, 26 species of marine mammals, and 6 species of turtles. For instance, plastic cups were found in the gut of some species of fish from British coastal waters; the ingested cups were eventually responsible for their deaths.[112] Even Antarctic and sub-Antarctic seabirds, e.g., Wilson's storm-petrel *(Oceanites oceanicus)* and white-faced storm-petrel *(Pelagodroma marina),* are at risk for this ingestion hazard.[112–115] It is reported that the proportion of plastic debris among litter increases with distance from source because it is transported more easily as compared to a denser material like glass or metal and because it lasts longer than other low-density materials (paper). Floating plastic articles (material less dense than water, e.g., polyamide, polyterephthalate, polyvinyl chloride) pose a global problem because they can contaminate even the most remote islands.[107,116] Drift plastics can increase the range of some marine organisms or introduce unwanted and aggressive alien taxa species into an environment. It could be risky to littoral, intertidal zones, and the shoreline.[112,117] There is also potential danger to marine ecosystems from the accumulation of plastic debris (material more dense than water) on the seafloor. Such bottom accumulation of plastic can inhibit the gas exchange between overlying waters and the pore water. This process can result in hypoxia or anoxia in the benthic fauna, altering the makeup of life on the sea bottom.[6] Another threat is connected with potential entanglement and ingestion hazards for pelagic and benthic animals.[62,112,118] Plastic can adsorb and concentrate some pollutants in coastal waters, including PCBs, DDE, nonylphenyl, and phenanthrene. It has been reported that these sorbed POPs could subsequently be released if the plastics are ingested.[109,110] For instance, PCBs in tissues of great shearwaters *(Puffinus gravis)* were derived from ingested plastic debris.[119]

Noise Pollution

In recent years, the marine biota has been affected by noise pollution. Natural sources of underwater noise may be physical and biological in character. Physical sources include wind, waves, rainfall, thunder and lighting, earthquake-generated seismic energy, and the movement of ice. Biological sources include marine mammal vocalizations and sounds produced by fish and invertebrates.[120,121]

Anthropogenic sound sources can be grouped into six categories, namely, shipping, seismic surveying, sonars, explosions, industrial activity, and miscellaneous.[122] Vessel traffic significantly contributes to underwater noise, mainly at low frequencies. Commercial shipping vessels generate noise mainly in areas confined to ports, harbors, and shipping lanes.[122] In contrast to wide geographic distribution of shipping industry, the oil and gas industry activities have taken place along continental margins in specific worldwide areas. Such resources exploration activities have been typically observed in shallow waters less than 200 m in depth. Other activities, in spite of their geographically widespread range, are also confined to near-shore coastal regions, namely, pile driving, dredging, operation of land- and ocean-based wind power turbines, power plant operations, and typical harbor and shipyard activities.[120] Offshore wind turbines may have significantly contributed noise to the underwater ecosystem bearing in mind that the relatively recent growth in offshore wind development has increased. It has been suggested that marine mammals may be indirectly affected by noise from offshore wind turbines, e.g., prey fish avoiding the sound source as well as the masking of marine mammals' mating and communication calls. On the other hand, a number of mass stranding of marine mammals, especially whales, found on worldwide beaches may be associated with the use of concurrent military sonar.[120] Another example of noise pollution affecting marine animals is continued exposure to anthropogenic noise pressure in vital sea turtle habitats, resulting in potential impact on its behavior and ecology. Brown shrimp exposed to higher pressure levels of noise in experimental area exhibited increased aggression, higher mortality rates, and significant reduction in their food uptake, growth, and reproduction. Sound exhibits measurable damage to sensory cells in the ears of fish.[123]

Light Pollution

A remarkable recent interest concerns the introduction of light to the coastal zone and nearshore environment. It is estimated that at least 3351 cities in the coastal zones all over the world are illuminated. It is expected that artificial light will be continuously intensified not only by population growth but also by dramatically increasing the number of locations of high-intensity artificial light. According to the United Nations World Tourism Organization (UN- WTO), there were ca. 900 million international tourist arrivals all over the world.[9] Tourist visits to beaches cause light pollution along the coastline since tracking the movement of population over time by research using satellite imagery showed that wherever human population density increases, the use of artificial light at night also increases. Living organisms are mostly sensitive to changes in the quality and intensity of natural light in the ecosystem. For instance, for algae and seaweeds, photosynthetic activity is highly dependent on available light, i.e., different cycles in natural light intensity and quality.[9] Light pollution takes place when biota is exposed to artificial light, especially in coastal areas, resulting in damaging effects on marine species in seas. The behavior, reproduction, and survival of marine invertebrates, amphibians, fish, and birds have been influenced by artificial lights. Light pollution disrupts the migration patterns of nocturnal birds and can result in hatchling sea turtles to head inland, away from the sea, which could be eaten by predators or run over by cars.[124] Ecological effects of light pollution concern disruption of predator–prey relationship. For instance, artificial light disturbs natural vertical migrations of zooplankton in the water column in accordance with the day-night cycle when natural light helps to reduce their predation by fish and other animals.[125]

Conclusion

The anthropogenic activity of man in coastal regions and even in offshore areas is responsible for emission of a huge amount of pollutants that are transported to marine ecosystems directly or by means of coastal watersheds, rivers, and precipitation from air. A lot of pollutants may be taken up by various compartments, i.e., biotic and abiotic, of aquatic environments and some of them could be biomagnified along the successive members of the food chain. Therefore, water pollution could have a negative effect on the quality of the water and hence on the health of the plants, animals, and humans whose lives depend on the quality of aquatic environments. Coastal areas are generally damaged from pollution, resulting in considerable impact on commercial coastal and marine fisheries.

There are numerous examples of worldwide events leading to serious contamination of coastal waters by persistent pollutants. Therefore, these areas have been extensively explored, especially in relation to perturbed ecosystems by heavy metals, radionuclides, POPs, oils, etc.

Elevated levels of nutrients in coastal waters resulted in eutrophication and proliferation of toxic algal blooms. The recently observed increase in tanker operations and oil use as well as marine tanker catastrophes has been responsible for the presence of excessively large amounts of oil spillage in coastal and marine ecosystems. Marine debris, especially plastics, is one of the most pervasive pollution problems. Marine pollutants are generally present in increased concentrations in the enclosed seas and coastal areas than in the open seawaters. Spatial distribution patterns of contamination concentrations exhibit a trend of their increase during transition from the south to the northern part of all oceans, i.e., in areas neighboring with industrial centers and concentration of main pollution sources.

References

1. *Pollutant,* 2011, available at http://en.wikipedia.org/wiki/Pollution (accessed 2011).
2. *Pollutant*, 2011, available at http://en.wikipedia.org/wiki/Pollutant#a-Stock_pollutants (accessed 2011).
3. Williams, C. Combating marine pollution from land-based activities: Australian initiatives. Ocean Coastal Manage. **1996**, *33*, 87–112.
4. Falandysz, J.; Trzosinska, A.; Szefer, P.; Warzocha, J.; Draganik, B. The Baltic Sea, especially southern and eastern regions. In *Seas at the Millennium: An Environmental Evaluation, Vol. I: Europe, The Americas and West Africa;* Sheppard, C.R.C., Ed.; Pergamon, Elsevier: Amsterdam, 2000; 99–120.
5. Albaiges, J.; Bebianno, M.J.; Camphuysen, K.; Cronin, M.; de Leeuw, J.; Gabrielsen, G.; Hutchinson, T.; Hyl- land, K.; Janssen, C.; Jansson, B.; Jenssen, B.M.; Roose, P.; Schulz-Bull, D.; Szefer, P. *Monitoring Chemical Pollution in Europe's Seas-Programmes, Practices and Priorities for Research;* Position Paper 16, Marine Board—European Science Foundation: Ostend, 2011.
6. Shahidul Islam, Md.; Tanaka, M. Impacts of pollution on coastal and marine ecosystems including coastal and marine fisheries and approach for management: A review and synthesis. Mar. Pollut. Bull. **2004**, *48,* 624–649.
7. Laws, E.A. *Aquatic Pollution;* John Wiley and Sons: New York, 2000.
8. Valcovic, V. Radioactivity in the Environment—Physicochemical Aspects and Applications; Elsevier: Amsterdam, 2000.
9. Depledge, M.H.; Godard-Codding, C.A.J.; Bowen, R.E. Light pollution in the sea. Mar. Pollut. Bull. **2010**, *60,* 1383–1385.
10. Sheppard, C.R.C., Ed. Seas in the Millennium: An Environmental Evaluation, Vol. I: Europe, The Americas and West Africa; Pergamon, Elsevier: Amsterdam, 2000.
11. Tanabe, S., Ed. *Mussel Watch—Marine Pollution Monitoring in Asian Waters*; Center for Marine Environmental Studies, Ehime University: Japan, 2000.
12. Nriagu, J.O. Global inventory of natural and anthropogenic emissions of trace metals to the atmosphere. Nature **1979**, *279,* 409–411.
13. Luoma, S.N.; Rainbow, P.S. Metal Contamination in Aquatic Environments: Science and Lateral Management; Cambridge University Press: Cambridge, 2008.
14. Prospero, J.M. African dust in America. Geotimes. **2001**, 24–27, available at http://www.rsmas.miami.edu/assets/pdfs/mac/fac/Prospero/Publications/Prospero_Geotimes_African%20Dust_2001.pdf (accessed 2011).

15. Szefer, P. Metals, Metalloids and Radionuclides in the Baltic Sea Ecosystem; Elsevier Science B.V.: Amsterdam, 2002.
16. Mason, R.P.; Fitzegerald, W.F.; Morel, F.M.M. The biogeochemical cycling of elemental mercury: Anthropogenic influences. Geochim. Cosmochim. Acta **1994**, *58,* 3191–3198.
17. Jickells, T. Atmospheric inputs of metals and nutrients to the oceans: Their magnitude and effects. Mar. Chem. **1995**, *48,* 199–214.
18. Neff, J.M. Bioaccumulation in Marine Organisms: Effects of Contaminants from Oil Well Produced Water; Elsevier: Amsterdam, 2002.
19. Bryan, G.W.; Gibbs, P.E. Impact of low concentrations of tributyltin (TBT) on marine organisms: A review. In *Metal Ecotoxicology: Concepts and Applications;* Newman, M.C.; McIntosh, A.W., Eds.; Lewis Publishers: Ann Arbor, MI, 1991; 323–361.
20. Kabata-Pendias, A.; Mukherjee, A.B. *Trace Elements from Soil to Human;* Springer: Berlin, 2007.
21. Bryan, G.W.; Langston, W.J. Bioavailability, accumulation and effects of heavy metals in sediments with special reference to United Kingdom estuaries: A review. Environ. Pollut. **1992**, *76,* 89–131.
22. US EPA (US Environmental Protection Agency). *Health Effects Assessment of Mercury;* Environmental Criteria and Assessment Office: Cincinnati, Ohio, 1984.
23. dos Santos, M.M.; Vieira, N.; Santos, A.M. Imposex in the dogwhelk *Nucella lapillus* (L.) along the Portuguese coast. Mar. Pollut. Bull. **2000**, *40,* 643–646.
24. Gray, J.E.; Theodorakos, P.M.; Bailey, E.A., Turner, R.R. Distribution, speciation, and transport of mercury in stream- sediment, stream-water, and fish collected near abandoned mercury mines in southwestern Alaska, USA. Sci. Total Environ. **2000**, *260,* 21–33.
25. Maurice-Bourgoin, L.; Quiroga, I.; Chincheros, J.; Courau, P. Mercury distribution in waters and fishes of the upper Madeira rivers and mercury exposure in riparian Amazonian populations. Sci. Total Environ. **2000**, *260,* 73–86.
26. Tomiyasu, T.; Nagano, A.; Yonehara, N.; Sakamoto, H.; Rifardi; Oki, K; Akagi, H. Mercury contamination in the Yat- sushiro Sea, south-western Japan: Spatial variations of mercury in sediment. Sci. Total Environ. **2000**, *257,* 121–132.
27. Förstner, U.; Wittmann, G.T.W. *Metal Pollution in the Aquatic Environment,* 2nd Ed.; Springer-Verlag: Berlin, 1983.
28. Phillips, D.J.H.; Rainbow, P.S. *Biomonitoring of Trace Aquatic Contaminants;* Elsevier Science Publishers Ltd.: London, 1993.
29. Akagi, H.; Malm, O.; Kinjo, Y.; Harada, M.; Branches, F.J.P.; Pfeiffer, W.C.; Kato, H. Methylmercury pollution in the Amazon, Brazil. Sci. Total Environ. **1995**, *175,* 85–95.
30. Harada, M. Characteristics of industrial poisoning and environmental contamination in developing countries. Environ. Sci. **1996**, *4,* 157–169.
31. Harada, M.; Nakachi, S.; Cheu, T.; Hamada, H.; Ohno, Y.; Tsuda, T.; Yanagida, K.; Kizaki, T.; Ohno, H. Monitoring of mercury pollution in Tanzania: Relation between head hair and health. Sci. Total Environ. **1999**, *227,* 249–256.
32. Wheatley, B.; Wheatley, M.A. Methylmercury and the health of indigenous peoples: A risk management challenge for physical and social sciences and for public health policy. Sci. Total Environ. **2000**, *259,* 23–29.
33. Al-Majed, N.B.; Preston, M.R. Factors influencing the total mercury and methyl mercury in the hair of the fishermen of Kuwait. Environ. Pollut. **2000**, *109,* 239–250.
34. Artaxo, P.; Calixto de Campos, R.; Fernandes, E.T.; Martins, J.V.; Xiao, Z.; Lindquist, O.; Fernandez-Jimenez, M.T.; Maenhaut, W. Large scale mercury and trace element measurements in the Amazon basin. Atmos. Environ. **2000**, *34,* 4085–4096.
35. Yamasoe, M.A.; Artaxo, P.; Miguel, A.H.; Allen, A.G. Chemical composition of aerosol particles from direct emissions of vegetation fires in the Amazon Basin: Water- soluble species and trace elements. Atmos. Environ. **2000**, *34,* 1641–1653.

36. Alzieu, C. TBT detrimental effects on oyster culture in France—Evolution since antifouling paint regulation. In Proceedings of Oceans 86 Conference Record. Organotin Symposium; 4 Institute of Electrical and Electronics Engineers: New York, 1986; 1130–1134.
37. Alzieu, C. Environmental impact of TBT: The French experience. Sci. Total Environ. **2000**, *258,* 99–102.
38. Shim, W.J.; Kahng, S.H.; Hong, S.H.; Kim, N.S.; Kim, S.K.; Shim, J.H. Imposex in the rock shell, *Thais clavigera,* as evidence of organotin contamination in the marine environment of Korea. Mar. Environ. Res. **2000**, *49,* 435–451.
39. Hung, T.-C.; Hsu, W.-K.; Mang, P.-J.; Chuang, A. Organo- tins and imposex in the rock shell, *Thais clavigera*, from oyster mariculture areas in Taiwan. Environ. Pollut. **2001**, *112,* 145–152.
40. Evans, S.M.; Nicholson, G.J. The use of imposex to assess tributyltin contamination in coastal waters and open seas. Sci. Total Environ. **2000**, *258*, 73–80.
41. Batt, J.M. *The world of organotin chemicals: Applications, substitutes, and the environment*; 2006, available at http://www.ortepa.org/WorldofOrganotinChemicals.pdf (accessed 2011).
42. Rumengan, I.F.; Ohji, M.; Arai, T.; Harino, H.; Arfin, Z.; Miyazaki, N. Contamination status of butyltin compounds in Indonesian coastal waters. Coastal Mar. Sci. **2008**, *32,* 116–126.
43. Zhao, Y.; Marriott, S.; Rogers, J.; Iwugo, K. A preliminary study of heavy metal distribution on the floodplain of the River Severn, U.K. by a single flood event. Sci. Total Environ. **1999**, *243/244,* 219–231.
44. Skwarzec, B. *Radiochemia Srodowiska i Ochrona Radiolo- giczna* (in Polish), Environmental Radiochemistry and Radiological Protection; Wydawnictwo DJ sc.: Gdansk, 2002.
45. ISSG. *The Irish Sea: An Environmental Review;* Report of the Irish Sea Study Group; Liverpool University Press: Liverpool, 1990.
46. Kershaw, P.J.; McCubbin, D.; Leonard, K.S. Continuing contamination of North Atlantic and Arctic waters by Sellafield radionuclides. Sci. Total Environ. **1999**, *237/238,* 119–132.
47. Aarkrog, A.; Dahlgaard, H.; Hansen, H.; Holm, E.; Hall- stadius, L.; Rioseco, J.; Christensen, G. Radioactive tracer studies in the surface waters of the northern North Atlantic including the Greenland, Norwegian and Barents Seas. Rit. Fiskideildar **1985**, *9*, 37–42.
48. Kershaw, P.J.; Baxter, A.J. The transfer of reprocessing wastes from north-west Europe to the Arctic. Deep Sea Res. **1995**, *42,* 1413–1448.
49. Förstner, U. Inorganic pollutants, particularly heavy metals in estuaries In *Chemistry and Biogeochemistry of Estuaries;* Olausson, E., Cato, I., Eds.; John Wiley & Sons: New York, 1980; 307–348.
50. UNSCEAR. *Sources and effects of ionizing radiation;* Report to the General Assembly by the United Nations Scientific Committee on the Effects of Atomic Radiation; United Nations: New York, 1993.
51. Beasley, T.M.; Cooper, L.W.; Grebmeier, J.M.; Orlandini, K.; Kelley, J.M. Fuel reprocessing Pu in the Arctic Ocean Basin: evidence from mass spectrometry measurements. In Proc. Conf. on Environmental Radioactivity in the Arctic, Oslo, August 1995.
52. Aarkrog, A.; Dahlgaard, H.; Nilsson, K.; Holm, E. Studies of plutonium and americum at Thule, Greenland. Health Phys. **1984**, *46,* 29–44.
53. Smith, J.N.; Ellis, K.M.; Aarkrog, A.; Dahlgaard, H.; Holm, E. Sediment mixing and burial of the ^{239,240}Pu pulse from the 1968 Thule, Greenland nuclear weapons accident. J. Environ. Radioact. **1994**, *25,* 135–159.
54. Joint Russian–Norwegian Expert Group. Radioactive Contamination of Dumping Sites for Nuclear Wastes in the Kara Sea. Results from the 1993 Expedition; Norwegian Radiation Protection Authority: Østerås, 1994.
55. National Academy of Sciences. *Radioactivity in the Marine Environment;* National Academy of Sciences: Washington, DC, 1971.
56. Aarkrog, A. A retrospect of anthropogenic radioactivity in the global marine environment. Radiat. Prot. Dosim. **1998**, *75*, 23–31.

57. Commission of the European Communities. The Radiological Exposure of the Population of the European Community from Radioactivity in North European Marine Waters— Project MARINA, Report EUR 12483; EU, 1989.
58. NEA. Co-ordinated Research and Environmental Surveillance Programme Related to Sea Disposal of Radioactive Waste, CRESP Final Report, 1981–1995; OECD Paris, 1996.
59. INSAG. *Post Accident Review Meeting on the Chernobyl Accident;* Summary Report; International Atomic Energy Agency: Vienna, 1986.
60. WHO. Health hazards from radiocaesium following the Chernobyl nuclear accident. Report on a WHO Working Group. J. Environ. Radioact. **1989**, *10,* 257–259.
61. Radiation effects from Fukushima Daiichi nuclear disaster, available at http://en.wikipedia.org/wiki/Radiation_effects_from_Fukushima_Daiichi_nuclear_disaster (accessed 2011).
62. Goldberg, E.D. Emerging problems in the coastal zone for the 21th century. Mar. Pollut. Bull. **1995**, *31,* 152–158.
63. Gibbs, P.E. Oviduct malformation as a sterilising effect of tributyltin-induced imposex in *Ocenebra erinacea* (Gastropoda: Muricidae). J. Molluscan Stud. **1996**, *62,* 403–413.
64. Cadee, G.C.; Boon, J.P.; Fischer, C.V.; Mensink, B.P.; Tjabbes, C.C. Why the whelk *Buccinum undatum* has become extinct in the Dutch Wadden Sea. Neth. J. Sea Res. **1995**, *34,* 337–339.
65. Jones, P. TBT implicated in mass dolphin deaths. Mar. Pollut. Bull. **1997**, *34,* 146.
66. HELCOM. Batlic Sea Environment Proceedings No. 64B. Third Periodic Assessment on the State of the Marine Environment of the Baltic Sea, 1989–93 Background Document. Helsinki Commission, Baltic Marine Environment Protection Commission, 1996; 252.
67. Jepson, P.D.; Bennet, P.M.; Allchin, C.R.; Law, R.J.; Kui- ken, T.; Baker, J.R.; Rogan, E.; Kirkwood, J.K. Investigating potential associations between chronic exposure to polychlorinated biphenyls and infectious disease mortality in harbour porpoises from England and Wales. Sci. Total Environ. **1999**, *243–244,* 339–348.
68. Cloern, J.E. Our evolving conceptual model of the coastal eutrophication problem. Mar. Ecol. Prog. Ser. **2001**, *210,* 223–253.
69. Riegman, R. Nutrient-related selection mechanisms in marine plankton communities and the impact of eutrophication on the plankton food web. Water Sci. Technol. **1995**, *32,* 63–75.
70. Kimor, B. Impact of eutrophication on phytoplankton composition. In *Marine Coastal Eutrophication*; Vollenweider, R.A., Marchetti, R., Vicviani, R., Eds.; Elsevier: Amsterdam, 1992; 871–878.
71. Zaitsev, Y.P. Recent changes in the trophic structure of the Black Sea. Oceanography **1992**, *1,* 180–189.
72. Birch, P.B.; Forbes, G.G.; Schofield, N.J. Monitoring effects of catchment management practices on phosphorus loads into the eutrophic Peel-Harvey Estuary, Western Australia. Water Sci. Technol. **1986**, *18,* 53–61.
73. Justic, D. Long-term eutrophication of the Northern Adriatic Sea. Mar. Pollut. Bull. **1987**, *18,* 281–284.
74. Degobbis, D. Increased eutrophication of the Northern Adriatic Sea. Second act. Mar. Pollut. Bull. **1989**, *20,* 452–457.
75. Forsberg, C. *Eutrophication of the Baltic Sea;* The Baltic Sea Environment: Uppsala, Sweden, 1993; 32 pp.
76. Gillespie, N. Ecological and epidemiological aspects of ciguatera fish poisoning. In Proc. of the Red Tide Workshop, Cronulla, June 18–20, 1984; Australian Department of Science: Canberra.
77. Hallegraeff, G.M.; Sumner, C. Toxic plankton blooms affect shellfish farms. Aust. Fish. **1986**, *45,* 15–18.
78. Holmes, P.R.; Lam, C.W.Y. Red tides in Hong Kong waters—Response to a growing problem. Asian Mar. Biol. **1985**, *2*, 1–10.

79. Phillips, D.J.H. Monitoring and control of coastal water quality. In *Pollution in the Urban Environment, POLMET 85;* Chan, M.W.H., Hoare, R.W.M., Holmes, P.R., Law, R.J.S., Reed, S.B., Eds.; Elsevier Applied Science Publishers: London, 1985; 559–565.
80. Morton, B.S. Pollution of the coastal waters of Hong Kong. Mar. Pollut. Bull. **1989**, *20,* 310–318.
81. Maclean, J.L. Indo-Pacific red tides, 1985–1988. Mar. Pollut. Bull. **1989**, *20,* 304–310.
82. Maclean, J.L.; White, A.W. Toxic dinoflagellate blooms in Asia: A growing concern. In *Toxic Dinoflagellates;* Anderson, D.M., White, A.W., Baden, D.G., Eds.; Elsevier: New York, 1985; 517–520.
83. Ayres, P.A. Mussel poisoning in Britain with special reference to paralytic shellfish poisoning. Environ. Health **1975**, *July,* 261–265.
84. Paerl, H.W., Whitall, D.R. Anthropogenically-derived atmospheric nitrogen deposition, marine eutrophication and harmful algal bloom expansion: Is there a link? Ambio **1999**, *28,* 307–311.
85. Paerl, H.W. Enhancement of marine primary productivity by nitrogen enriched rain. Nature **1985**, *315,* 747–749.
86. Paerl, H.W. Coastal eutrophication in relation to atmospheric nitrogen deposition: Current perspectives. Ophelia **1995**, *41,* 237–259.
87. Anderson, D.M. Toxic algal blooms and red tides. A global perspective. In *Red Tides: Biology, Environmental Science and Toxicology;* Okaichi, T., Anderson, D.M., Nemoto, T., Eds.; Elsevier Science Publishing Co., Inc.: New York, 1989; 11–21.
88. Aksnes, D.L.; Aure, J.; Furnes, G.K.; Skjoldal, H.R.; Sae- tre, R. Analysis of the *Chrysochromulina polylepis* bloom in the Skagerrak. Environmental conditions and possible causes. Bergen Scientific Centre Publication 1989, No. BSC 89/1.
89. Tester, P.A.; Stumpf, R.P.; Vukovich, F.M.; Fowler, P.K.; Turner, J.T. An expatriate red tide bloom: Transport, distribution, and persistence. Limnol. Oceanogr. **1991**, *36,* 1053–1061.
90. Buskey, E.J.; Stockwell, D.A. Effects of a persistent "brown tide" on zooplankton populations in the Laguna Madre of South Texas. In *Toxic Phytoplankton Blooms in the Sea,* Proc. 5th Intern. Conf. on Toxic Marine Phytoplankton; Elsevier, 1993; 659–665.
91. Hallegraeff, G.M. A review of harmful algal blooms and their apparent global increase. Phycologia **1993**, *32,* 79–99.
92. Anderson, D.M.; Kulis, D.M.; Doucette, G.J.; Gallagher, J.C.; Balech, E. Biogeography and toxic dinoflagellates in the genus *Alexandrium* from the northeastern United States and Canada. Mar. Biol. **1994**, *120,* 467–478.
93. ECOHAB. *The Ecology and Oceanography of Harmful Algal Blooms;* A National Research Agenda, US N.S.F./ N.O.O.A. Publication; Woods Hole Oceanographic Inst.: Mass., USA, 1995.
94. Howarth, R.W.; Billen, G.; Swaney, D.; Townsend, A.; Ja- worski, N.; Lajtha, K.; Downing, J.A.; Elmgren, R.; Caraco, N.; Jordan, T.; Berendse, F.; Freney, J.; Kudeyarov, V.; Murdoch, P.; Zhu, Z.-L. Regional nitrogen budgets and riverine N and P fluxes for the drainages to the North Atlantic Ocean: Natural and human influences. Biogeochemistry **1996**, *35,* 75–139.
95. Prospero, J.M.; Barret, K.; Church, T.; Detener, F.; Duce, R.A.; Galloway, J.N.; Levy, H.; Moody, J.; Quinn, P. Atmospheric deposition of nutrients to the North Atlantic basin. Biogeochemistry **1996**, *35*, 27–73.
96. Paerl, H.W. Coastal eutrophication and harmful algal blooms: Importance of atmospheric deposition and groundwater as "new" nitrogen and other nutrient sources. Limnol. Oceanogr. **1997**, *42,* 1154–1165.
97. Mills, E.L.; Holeck, K.T. Biological pollutants in the Great Lakes. Clearwaters **2001**, *31* (1), available at http://www. nywea.org/clearwaters/pre02fall/311010.html (accessed 2011).
98. Elliott, M. Biological pollutants and biological pollution— An increasing cause for concern. Mar. Pollut. Bull. **2003**, *46,* 275–280.

99. FAO. Precautionary Approach to Fisheries. Part 1. Guidelines on the Precautionary Approach to Capture Fisheries and Species Introductions; Food and Agriculture Organization of the United Nations: Rome, 1995.
100. Joly, C. Plant nutrient management and the environment. In *Prevention of Water Pollution by Agriculture and Related Activities*, Proceedings of the FAO Expert Consultation, Santiago, Chile, October 20–23, 1992; Water Report 1,FAO: Rome.
101. Li, Y.F.; Macdonald, R.W. Sources and pathways of selected organochlorine pesticides to the Arctic and the effect of pathway divergence on HCH trends in biota: A review. Sci. Total Environ. **2005**, *342,* 87–106.
102. Duursma, E.K.; Marchand, M. Aspects of organic marine pollution. Oceanogr. Mar. Biol. Ann. Rev. **1974**, *12,* 315–431.
103. *Oil spill;* 2011, available at http://en.wikipedia.org/wiki/ Oil_spill (accessed 2011).
104. *Oil spills and disasters;* 2011, available at http://www.info- please.com/ipa/A0001451.html (accessed 2011).
105. Smith, N. The problem of oil pollution of the sea. Adv. Mar. Biol. **1970**, *8,* 215–306.
106. Sheavly, S.B.; Register, K.M. Marine debris and plastics: Environmental concerns, sources, impacts and solutions. J.Polym. Environ. **2007**, *15,* 301–305.
107. Ryan, P.G.; Moore, C.J.; van Franeker, J.A.; Moloney, C.L. Monitoring the abundance of plastic debris in the marine environment. Philos. Trans. R. Soc. **2009**, *364,* 1999–2012.
108. Moore, C.J. Synthetic polymers in the marine environment: A rapidly increasing. Environ. Res. **2008**, *108,* 131–139.
109. Barnes, K.A.; Galgani, F.; Thompson, R.C.; Barlaz, M. Accumulation and fragmentation of plastic debris in global environments. Philos. Trans. R. Soc. **2009**, *364,* 1985–1998.
110. Teuten, E.L.; Saquing, J.M.; Knappe, D.R.U.; Barlaz, M.A.; Jonsson, S.; Björn, A.; Rowland, S.J.; Thompson, R.C.; Galloway, T.S.; Yamashita, R.; Ochi, D.; Watanuki, Y.; Moore, C.; Viet, P.H.; Tana, T.S.; Prudente, M.; Boonyatumanond, R.; Zakaria, M.P.; Akkhavong, K.; Ogata, Y.; Hirai, H.; Iwasa, S.; Mizukawa, K.; Hagino, Y.; Imamura, A.; Saha, M.; Takada, H. Transport and release of chemicals from plastics to the environment and to wildlife. Philos. Trans. R. Soc. B **2009**, *364,* 2027–2045.
111. US Marine Mammal Commission. *Marine Mammal Commission Annual Report to Congress;* Effects of Pollution on Marine Mammals: Bethesda, MD, 1996.
112. Derraik, J.G.B. The pollution of the marine environment by plastic debris: A review. Mar. Pollut. Bull. **2002**, *44*, 842–852.
113. Slip, D.J.; Green, K.; Woehler, E.J. Ingestion of anthropogenic articles by seabirds at Macquarie Island. Mar. Orni- thol. **1990**, *18,* 74–77.
114. Van Frakener, J.A.; Bell, P.J. Plastic ingestion by petrels breeding in Antarctica. Mar. Pollut. Bull. **1988**, *19,* 672–674.
115. Bourne, W.R.P.; Imber, M.J. Plastic pellets collected by a prion on Gough Island, Central South Atlantic Ocean. Mar. Pollut. Bull. **2001**, *13,* 20–21.
116. Mato, Y.; Isobe, T.; Takada, H.; Kanehiro, H.; Ohtake, C.; Kaminuma, T. Plastic resin pellets as a transport medium for toxic chemicals in the marine environment. Environ. Sci. Technol. **2001**, *35*, 318–324.
117. Gregory, M.R. Plastics and South Pacific Island shores: Environmental implications. Ocean Coastal Manage. **1999**, *42,* 603–615.
118. Hess, N.A.; Ribic, C.A.; Vining, I. Benthic marine debris, with an emphasis in fishery-related items, surrounding Kodiak Island, Alaska 1994–1996. Mar. Pollut. Bull. **1999**, *38,* 885–890.
119. Ryan, P.G.; Connell, A.D.; Gardener, B.D. Plastic ingestion and PCBs in seabirds: Is there a relationship? Mar. Pollut. Bull. **1988**, *19,* 174–176.
120. Firestone, J.; Jarvis, C. Response and responsibility: Regulating noise pollution in the marine environment. J. Int. Wildl. Law Policy **2007**, *10,* 109–152.

121. Hatch, L.T.; Wright, A.J. A brief review of anthropogenic sound in the oceans. Int. J. Comp. Psychol. **2007**, *20*, 121–133.
122. National Research Council (NRC). *Ocean Noise and Marine Mammals;* National Academy Press: Washington, DC, 2003.
123. Samuel, Y.; Morreale, S.J.; Clark, C.W.; Greene, C.H.; Richmond, M.E. Underwater, low-frequency noise in coastal sea turtle habitat. J. Acoust. Soc. Am. **2005**, *117*, 1465–1472.
124. Gallaway, T.; Olsen, R.N.; Mitchell, D.M. The economics of global light pollution. Ecol. Econ. **2010**, *69*, 658–665.
125. Gliwicz, Z.M. A lunar cycle in zooplankton. Ecology **1986**, *67*, 883–897.

3

Groundwater: Mining Pollution

Jeff Skousen and
George Vance

Introduction

Mining activities can impact the quantity, quality, and usability of groundwater supplies. Underground mining for coal by longwall or room and pillar mining methods often interrupts and depletes groundwater, and can also alter its quality. Surface mining can enhance the introduction of surface water with dissolved solids into groundwater systems through fractures or other conduits. The type and nature of the mining activity, the disturbed geologic strata, and alteration of surface and subsurface materials will determine how groundwater supplies will be impacted. As waters contact and interact with disturbed geologic materials, constituents such as salts, metals, trace elements, and organic compounds become mobilized [1,2]. The dissolved substances can leach into deep aquifers and cause groundwater quality impacts [3]. In addition to concerns due to naturally occurring contaminants from disturbance activities, mining operations may also contribute to groundwater pollution from leaking underground storage tanks, improper disposal of lubricants and solvents, and contaminant spills. Blasting and hydraulic fracking activities can provide additional connection to surface water inputs, and underground injection of wastes can also occur during these operations [4].

In the United States, the Clean Water Act (CWA) and its subsequent amendments establish the authority for all water pollution control actions at the federal level [5] and regulate discharges into surface streams, wetlands, and oceans. Mining operations must acquire National Pollutant Discharge Elimination System (NPDES) permits for discharges to surface waters. Groundwater quality in the United States is regulated by the Safe Drinking Water Act (SDWA), which was originally enacted in 1974 and amended in 1996. The SDWA was passed to protect drinking water supplies by requiring discharges into groundwaters to meet the use standard or the ambient condition, whichever is of higher quality [6]. This is done by legislating maximum contaminant levels (MCLs) above which waters are considered unsafe for human consumption. The Office of Water within the Environmental Protection Agency provides guidance, specifies scientific methods and data collection requirements, and performs oversight for entities that supply drinking water including groundwater. Examples of some water contaminants with specified MCLs associated with mining activities are listed in Table 1 [7].

Because mining activities can result in poor-quality groundwaters, enforcement of regulations is needed to minimize and/or eliminate potential problems. The Surface Mining Control and Reclamation

TABLE 1 Selected Contaminants in Drinking Waters That May Be Influenced by Mining Activities [7]

Contaminant	MCL (mg/L)	MCLG
	Inorganics	
Arsenic	0.010	0
Cadmium	0.005	0.005
Chromium	0.1	0.1
Copper	LV	1.3
Cyanide	0.2	0.2
Fluoride	4.0	4.0
Lead	LV	0
Mercury	0.002	0.002
Nitrate (NO_3-N)	10	10
Selenium	0.05	0.05
Sulfate	500	500
	Radionuclides	
Radium	5 pCi/L	0
Uranium	30 ug/L	0
	Organics	
Benzene	0.005	0
Carbon tetrachloride	0.005	0
Pentachlorophenol	0.001	0
Toluene	1	1
Xylenes	10	10
	Microbiological	
Total coliforms	LV	0
Viruses	LV	0

MCL, Maximum contaminant levels permissible for a contaminant in water that is delivered to any user of a public water system; MCLG, Maximum contaminant level goals of a drinking water contaminant that is protective of adverse human health effects and which allows for an adequate margin of safety; LV, Lowest value that can be achieved using the best available technology.

Act (SMCRA) of 1977 identifies policies and practices for mining and reclamation to minimize water quality impacts [8]. SMCRA requires that specific actions be taken to protect the quantity and quality of both on- and off-site groundwaters. All mines are required to meet either state or federal groundwater guidelines, which are generally related to priority pollutant standards described in the CWA and SDWA.

Groundwater Resources

Groundwater resources are the world's third largest source of water behind oceans (97%) and glaciers (2%), and represent 0.6% of the earth's water content [9]. Approximately 53% of the US population uses groundwater as a drinking water source, but this percentage increases to almost 97% for rural households. In areas of low rainfall, weathering and translocation of dissolved constituents are relatively slow compared to high rainfall areas. For example, only 12% of precipitation will recharge underground water supplies in a dry coal mining area like Gillette, Wyoming, while almost 47% of precipitation was available for recharge in coal mining areas of Tennessee [10]. Transport of contaminants from surface

TABLE 2 Important Hydrogeological Characteristics of a Site That Determine Groundwater Quantity and Quality

Geological
Type of water-bearing unit or aquifer (rock type, overburden).
Thickness and areal extent of water-bearing units and aquifers.
Type of porosity (primary, such as intergranular pore space, or secondary, such as bedrock discontinuities, e.g., fracture or solution cavities).
Presence or absence of impermeable units or confining layers.
Depths to water tables; thickness of vadose zone.
Permeability and connectivity to other voids or conduits.
Hydraulic
Hydraulic properties of water-bearing unit or aquifer (hydraulic conductivity, transmissivity, storability, permeability, dispersivity).
Pressure conditions (confined, unconfined, leaky confined).
Groundwater flow directions (hydraulic gradients, both horizontal and vertical), volumes (specific discharge), rate (average linear velocity).
Recharge and discharge areas.
Groundwater or surface water interactions; areas of groundwater discharge to surface water or vice versa.
Seasonal variations of groundwater conditions.
Groundwater Use
Existing or potential underground sources of drinking water.
Existing or near-site use of groundwater.

and subsurface environments to groundwaters is generally accelerated as the amount of percolating water increases.

Infiltrating water moves through the vadose zone (unsaturated region) into groundwater zones (saturated region). The upper boundary of the groundwater system (e.g., water table) fluctuates depending on the amount of water received or removed from the groundwater zone. Groundwater movement is a function of hydraulic gradients and hydraulic conductivities, which represent the combined forces with which water moves as a function of gravitational, osmotic, and pressure forces and the permeability of geologic strata. Groundwater moves faster in coarse-textured materials and where hydraulic gradients are high. Aquifers are groundwater systems that have sufficient porosity and permeability to supply enough water for specific purposes. For an aquifer to be useful, it must be able to store, transmit, and yield sufficient amounts of good-quality water. Important hydrogeological characteristics of a site that determine groundwater quantity and quality are listed in Table 2.

Groundwater Contaminants

Several types of substances affect groundwater quality [1,11]. Water contaminants include inorganic, organic, and biological materials. Some have a direct impact on water quality, while others indirectly cause physical, chemical, or biological changes that make the water unsuitable for its designated use. Substances that degrade groundwaters include nutrients, salts, heavy metals, trace elements, and organic chemicals, as well as contaminants such as radionuclides, carcinogens, pathogens, and petroleum wastes (Table 3, [12]). Several types of organic chemicals entering groundwaters are less dense than water and tend to move to and along the surface of the water table. Changes can also occur in groundwaters due to temperature fluctuations and odors. Some groundwaters near coal seams contain natural organic substances (such as dissolved methane gas) and synthetic organic chemicals. Methane gas can be extracted from coal beds where underground and surface mining operations are projected,

TABLE 3 Different Classes of Groundwater Pollutants and Their Causes [12]

Water Pollutant Class	Contributions
Inorganic chemicals	Toxic metals and acidic substances from mining operations and various industrial wastes
Organic chemicals	Petroleum products, pesticides, and materials from organic wastes industrial operations
Infectious agents	Bacteria and viruses
Radioactive substances	Waste materials from mining and processing of radioactive substances or from improper disposal of radioactive isotopes

and this extraction can alter methane gas concentrations in groundwaters [10]. Organic contamination may also result from leaking gas tanks, oil spills, or runoff from equipment-serving areas. In these cases, the source of the contamination must be identified and removed. Gasoline, diesel, or oil-soaked areas should be immediately excavated and disposed of by approved methods.

The chemistry of groundwaters and potential levels of naturally occurring contaminants are related to (1) groundwater hydrologic conditions, (2) mineralogy of the mined and locally impacted geological materials, (3) mining operations (e.g., extent of disturbed materials and its exposure to atmospheric conditions), and (4) time. Movement of metal contaminants in groundwaters varies depending on the chemical of concern. Solubility considerations include metals such as cobalt, copper, nickel, and zinc being more mobile than silver and lead, and gold and tin being even less mobile [1]. As conditions such as pH, redox, and ionic strength change over time, dissolved constituents in groundwaters may decrease due to adsorption, precipitation, and chemical speciation reactions and transformations.

Acid mine drainage (AMD) is the most prevalent groundwater quality concern at inactive and abandoned surface and underground mine sites. If geologic strata containing reduced S minerals (e.g., pyrite (FeS_2)) are exposed to weathering conditions, such as when pyritic overburden materials are brought to the surface during mining activities and then reburied, high concentrations of sulfuric acid (H_2SO_4) can develop and form acid waters with pH levels below 2 [2]. Neutralization of some or all of the acidity produced during the oxidation of reduced S compounds can occur when carbonate minerals in proximity to the acid-producing materials dissolve [3]. Neutralization can also occur when silicate minerals dissolve, but sometimes high levels of potentially toxic metals such as Al, Cu, Cd, Fe, Mn, Ni, Pb, and Zn may be released. For example, mining of coal in the Toms Run area of northwestern Pennsylvania resulted in groundwater contamination by AMD containing high concentrations of Fe and sulfate (SO_4) that leached into the underlying aquifer through joints, fractures, and abandoned oil and gas wells.

The Gwennap Mining District in the United Kingdom contained numerous mines that operated over several centuries to extract various mineral resources. One of these mines, the Wheal Jane metal mine in Cornwall, extracted ores that included cassiterite (Sn-containing mineral), chalcopyrite (Cu), pyrite (Fe), wolframite (tungsten, W), arsenopyrite (arsenic, As), in addition to smaller deposits of Ag, galena (Pb), and other minerals. After closure in the early 1990s, extensive voids remaining in the Wheal Jane mine that contained oxidized and weathered minerals were flooded. Initial groundwater quality was poor with a pH of 2.9 and a total metal concentration of 5000 mg/L, which contained elevated levels of Fe, Zn, Cu, and Cd. Water quality worsened with depth, and at 180 m, the groundwater had a pH of 2.5 and a metal concentration of 7000 mg/L. Treatment of discharge waters originating from the mine involves an expensive process that will continue long term to preserve environmental quality in surface and groundwaters in the region. A similar situation occurred when a Zn mine in southwestern France was closed. In this case after flooding, discharge mine waters had a solution pH near neutral, but the water still contained high concentrations of Zn, Cd, Mn, Fe, and SO_4.

Within the Coeur D' Alene District of Idaho at the Bunker Hill Superfund site, groundwater samples were found to contain high concentrations of Zn, Pb, and Cd [13]. The contamination originated from the leaching of old mine tailings deposited on a sand and gravel aquifer. When settling ponds were constructed to catch the runoff from the tailings, water from the ponds infiltrated into the aquifer and caused an increase in metal concentration in the local groundwater system [14].

Gold mining operations have used cyanide as a leaching agent to solubilize Au from ores, which often contain arsenopyrite (As, Fe, and S) and pyrite [1]. Unfortunately, cyanide, in addition to being toxic on its own, is a powerful nonselective solvent that solubilizes numerous substances that can be environmental contaminants. These ore waste materials are often stored in tailing ponds and, depending on the local geology and climate, the cyanide present in the tailings can exist as free cyanide (CN^-, HCN); inorganic compounds containing cyanide (NaCN, $HgCN_2$); metal-cyanide complexes with Cu, Fe, Ni, and Zn; and/or the compound CNS. Because cyanide species are mobile and persistent under certain conditions, a large potential exists for trace element and cyanide migration into groundwaters. For example, a tailings dam failure resulted in cyanide contamination of groundwater at a gold mining operation in British Columbia, Canada [1].

Arsenic and uranium (U) contamination has resulted from extensive mining and smelting of ores containing various metals (Ag, Au, Co, Ni, Pb, and Zn) and/or nonmetals (As, P, and U). Arsenic-contaminated groundwaters have been a source of surface recharge and drinking water supplies. At one site, a nearby river had As levels 7 and 13 times greater than the recommended national and local drinking water standards, respectively [1]. Arsenic is known as a carcinogen and has been the contributing cause of death to humans in several parts of the world that rely on As-contaminated drinking water [11]. Waters from dewatering a U mine in New Mexico had elevated levels of U and radium (Ra) activities as well as high concentrations of dissolved Mo and Se, which were detected in stream water 140 km downstream from the mine.

Groundwater Analysis

Both the remediation and prevention of groundwater contamination by nutrients, salts, heavy metals, trace elements, organic chemicals (natural and synthetic), pathogens, and other contaminants require the evaluation of the composition and concentration of these constituents either *in situ* or in groundwater samples [2,10]. Monitoring may require the analysis of physical properties, inorganic and organic chemical compositions, and/or microorganisms according to well-established protocols for sampling, storage, and analysis [15]. For example, if groundwater will be used for human or animal consumption, the most appropriate tests would be nitrate-nitrogen (NO_3-N), trace metals, pathogens, and organic chemicals. Several common constituents measured in groundwaters are listed in Table 4. However, other tests can be conducted on waters including tests for hardness, electrical conductivity (EC), chlorine, radioactivity, water toxicity, and odors [16].

Recommendations based on interpretation of the groundwater test results should be related to the ultimate use of the water [2]. The interpretation and recommendation processes may be as simple as determining that a drinking water well exceeds the established MCLs for NO_3-N and recommending the

TABLE 4 Groundwater Quality Parameters and Constituents Measured in Some Testing Programs [16]

Physical	Metals and Trace Elements	Nonmetallic Constituents	Organic Chemicals	Microbiological Parameters
Conductivity Salinity Sodicity Dissolved solids Temperature Odors	Al, Ag, As, Ca, Cd, Cr, Cu, Fe, Mg, Mn, Na, Ni, Pb, Se, Sr, Zn	pH, acidity, alkalinity, dissolved oxygen, carbon dioxide, bicarbonate, B, Cl, CN, F, I, ammonium, nitrite, nitrate, P, Si, sulfate	Methane Oil and grease Organic acids Volatile acids Organic C Pesticides Phenols Surfactants	Fecal coliforms Bacteria Viruses

well should not be used as a drinking water source or that a purification system be installed. However, interpretations of most groundwater analyses can be quite complicated and require additional information for proper interpretation. If a contaminant exceeds an acceptable concentration, all potential sources contributing to the pollution and pathways by which the contaminant moves must be identified. In many cases, multiple groundwater contaminants are present at different concentrations. Because the interpretation of water analyses is a complex process, recommendations should be based on a complete evaluation of the water's physical, chemical, and biological properties. Integrating water analyses into predictive models that can assess the effects of mining activities on water quality is needed in the long term to determine the most effective means to preserve and restore water quality.

Strategies for Remediating Contaminated Groundwaters

Mine sites that have been contaminated generally contain mixtures of inorganic and/or organic constituents, so it is important to understand these multi-component systems in order to develop remediation strategies. Therefore, a proper remediation program must consider identification, assessment, and correction of the problem [17,18]. Identification of a potential problem site requires that the past history of the area and activities that took place are known, or when a water analysis indicates a site has been contaminated. Assessment addresses questions such as (1) what is the problem, (2) where and to what extent is the problem, and (3) who and what is affected by the problem. Afterward, a remediation action plan must be developed that will address the specific problems identified. A remediation action program may require that substrata materials (e.g., backfill) and groundwater be treated.

If remedial action is considered necessary, then three general options are available: (1) containment, (2) *in situ* treatment, or (3) pump-and-treat method (Figure 1). The method(s) used for the containment of contaminants are beneficial for restricting contaminant transport and dispersal. Of the remediation techniques, *in situ* treatment measures are the most appealing because they generally do less surface damage, require a minimal amount of facilities, reduce the potential for human exposure to contaminants, and when effective, reduce or remove the contaminant so that the groundwater can be utilized again [18]. *In situ* remediation can be achieved by physical, chemical, and/or biological techniques. Biological *in situ* techniques used for groundwater bioremediation can either rely on the indigenous (native) microorganisms to degrade organic contaminants or on amending the groundwater environment with specialized microorganisms (bioaugmentation). The pump-and-treat method, however, is one of the more commonly used processes for remediating contaminated groundwaters [17]. With the pump-and-treat methods, the contaminated waters are pumped to the surface where one of the

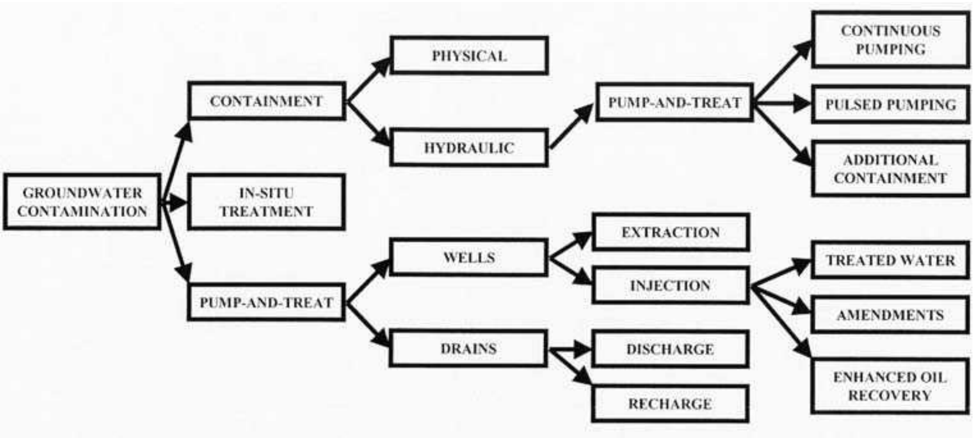


FIGURE 1 Remedial options to consider if cleanup of contaminated groundwater is deemed necessary.

many treatment processes can be utilized. A major consideration in the pump-and-treat technology is the placement of wells, which is dependent on the contaminant and site characteristics (see Table 2). Extraction wells are used to pump the contaminated water to the surface where it can be treated and re-injected or discharged. Injection wells can be used to re-inject the treated water, water containing nutrients and other substances that increase the chances for chemical alteration or microbial degradation of the contaminants, or materials for enhanced oil recovery.

Treatment techniques can be grouped into three categories, namely, physical, chemical, and biological methods [2,18].

Physical methods include several techniques. Adsorption methods physically adsorb or trap contaminants on various types of resins. Separation treatments include physically separating contaminants by forcing water through semipermeable membranes (e.g., reverse osmosis). Flotation, or density separation, is commonly used to separate low-density organic chemicals from groundwaters. Air and steam stripping can remove volatile organic chemicals. Isolation utilizes barriers placed above, below, or around sites to restrict movement of the contaminant. Containment systems should have a permeability of 10^{-7} cm/s or less.

Chemical methods are also numerous. Chemical treatment involves addition of chemical agent(s) in an injection system to neutralize, immobilize, and/or chemically modify contaminants. Extraction (leaching) of contaminants uses one of the several different aqueous extracting agents such as an acid, base, detergent, or organic solvent miscible in water. Oxidation and reduction of groundwater contaminants are commonly done using air, oxygen, ozone, chlorine, hypochlorite, and hydrogen peroxide. Ionic and nonionic exchange resins can adsorb contaminants, thus reducing their leaching potential.

Biological methods for contaminant remediation are less extensive than physical and chemical techniques. Land treatment is an effective method for treating groundwaters by applying the contaminated waters to lands using surface, overland flow, or subsurface irrigation. Activated sludge and aerated surface impoundments are used to precipitate or degrade contaminants present in water and include both aerobic and anaerobic processes. Biodegradation is one of the several biological-mediated processes that transform contaminants, and it utilizes vegetation and microorganisms.

References

1. Ripley, E.A., R.E. Redmann, and A.A. Crowder. 1996. *Environmental Effects of Mining*. St. Lucie Press, Delray Beach, FL. 356 pp.
2. Pierzynski, G.M., J.T. Sims, and G.F. Vance. 2005. *Soils and Environmental Quality*. 3rd Edition. CRC Press, Inc., Boca Raton, FL. 584 pp.
3. Prokop, G., P. Younger, and K. Roehl. 2003. Groundwater management in mining areas. *Proceedings of the 2nd IMAGE TRAIN Advanced Study Course*. Available at http://www.umweltbundesamt.at/fileadmin/site/publikationen/CP035.pdf.
4. U.S. Environmental Protection Agency. 2016. EPA's study of hydraulic fracturing for oil and gas and its potential impact on drinking water resources. Available at https://www.epa.gov/hfstudy.
5. U.S. Congress. 1977. Clean Water Act. Public Law 95-217. Available at https://www.epa.gov/laws-regulations/summary-clean-water-act.
6. U.S. Congress. 1996. Safe Drinking Water Act and Amendments. Public Law 104-182. Available at https://www.epa.gov/laws-regulations/summary-safe-drinking-water-act.
7. U.S. Environmental Protection Agency. 2002. Potential impacts of hydraulic fracturing of coalbed methane wells on underground sources of drinking water. Available at http://permanent.access.gpo.gov/lps21800/www.epa.gov/safewater/uic/cbmstudy.html.
8. U.S. Congress. 1977. Surface Mining Control and Reclamation Act. Public Law 95-87. Available at https://www.osmre.gov/lrg/docs/SMCRA.pdf.
9. National Groundwater Association (NGWA). 2019. Information on earth's water. Available at https://www.ngwa.org/what-is-groundwater/About-groundwater/information-on-earths-water.

10. National Academy of Sciences. 1990. Surface Coal Mining Effects on Ground Water Recharge. Committee on Ground Water Recharge in Surface-Mined Areas, Water Science and Technology Board, National Research Council. 170 pp. Available at http://www.nap.edu/catalog.php?record_id=1527.
11. Manahan, S.E. 2017. *Environmental Chemistry*. 10th Edition. Lewis Publishers, Chelsea, MI.
12. Nielsen, D.M. 2005. *Practical Handbook of Groundwater Monitoring*. 2nd Edition. Lewis Publishers, Boca Raton, FL.
13. Todd, D.K., D.E.O. McNulty. 1975. *Polluted Groundwater*. Water Information Center, Inc., Huntington, NY.
14. National Research Council. 2005. *Superfund and Mining Megasites: Lessons learned from the Coeur D' Alene River Basin*. The National Academies Press, Washington, DC. https://doi.org/10.17226/11359.
15. Rice, E.W., R.B. Baird, and A.D. Eaton (Eds). 2017. *Standard Methods for the Examination of Water and Wastewater*. 23rd Edition. American Public Health Association, Washington DC.
16. U.S. Environmental Protection Agency. 2009. Drinking Water Contaminants. Available at https://www.epa.gov/sdwa/drinking-water-regulations-and-contaminants
17. Interstate Technology & Regulatory Council. 2005. Overview of Groundwater Remediation Technologies for MTBE and TBA. Available at https://www.itrcweb.org/GuidanceDocuments/MTBE-1.pdf.
18. Hyman, M.H. 1999. Groundwater and Soil Remediation, pp. 684–712. In: Meyers, R.A., (ed.), *Encyclopedia of Environmental Pollution and Cleanup*. Volume 1. John Wiley & Sons, Inc., New York, NY.

4

Groundwater: Nitrogen Fertilizer Contamination

Lloyd B. Owens and
Douglas L. Karlen

Introduction

Groundwater is widely used for domestic and public water supplies, particularly in rural areas.[1] When it resurfaces, groundwater also becomes surface water that ultimately flows through streams and lakes to the oceans. Generally, groundwater is quite pure unless it contains natural contaminants such as high iron, sulfur, or possibly an intrusion of saltwater if the aquifer is located close to an ocean or estuary. Such contaminants usually result from the geological formations through which the groundwater source is flowing or residing in. Another groundwater contaminant, nitrate-nitrogen (NO_3-N), can originate naturally but is most often associated with areas where human intervention and the management practices being used have significantly increased the amount of NO_3-N that is available for leaching as water moves through soil and the underlying geology into groundwater aquifers.

Why is NO_3-N in groundwater a problem? When water with high nitrate (NO_3) concentrations is consumed by humans, it can cause several adverse health problems. One of the most common is known as “blue baby syndrome” or methemoglobinemia, an illness that arises when an infant’s blood is unable to carry enough oxygen to body cells and tissue. Consumption of high-NO_3 water has also been associated with increased levels of nitrosamine and some types of cancer. High levels of NO_3-N in streams, lakes, and oceans can stimulate excess growth of plants and bacteria, which upon death and decay subsequently deplete much of the oxygen in water. This causes fish kills and “dead zones,” such as the hypoxic area in the Gulf of Mexico.

Elevated levels of NO_3-N in groundwater have been reported in many parts of the world.[2,3] Although this can occur naturally, leaching of N fertilizer is often a major factor responsible for high NO_3-N concentrations in groundwater. This occurs because nitrogen (N) is an essential plant nutrient, and to achieve optimum crop yields, the amount of N that becomes available through natural cycling must be supplemented with additional N from either inorganic (i.e., fertilizer) or organic (i.e., manure or legume) sources. When the total amount of N provided through natural and supplemental sources exceeds that removed by harvested portions of the crop, excess N can accumulate in the soil profile and become available for leaching into subsurface drainage lines or directly into groundwater aquifers. The challenge of providing adequate plant-available N without creating an excess that is available for leaching has been studied for decades in attempts to determine the economically optimum N level (EON). Such studies have been conducted not only for agronomic crops (e.g., maize, wheat, cotton, and rice) but also for turf grasses and containerized horticultural crops, which usually receive frequent and high N fertilizer rates. Despite these efforts to identify EON rates, there are some situations when there will still be a positive yield response to N fertilizer at rates exceeding the amount of N that the plants will utilize. When this occurs, excess NO_3-N accumulates in the soil profile and becomes available for leaching, thus resulting in NO_3-N concentrations that exceed the maximum contaminant levels (MCLs) for drinking water (i.e., 10 mg L^{-1}). Many approaches can be used to reduce the NO_3 leaching potential. This includes reducing overall N fertilizer inputs; applying N in split applications, coinciding with the time when plants can use the N most efficiently; using slow-release N fertilizers; and growing cover crops to capture part of the residual N accumulating in the soil profile before it can be leached to the groundwater. This entry focuses on groundwater and how it is impacted by NO_3-N leaching. Several factors contributing to leaching are addressed and supported by accompanying references, but space does not permit this to be a comprehensive literature review. Our primary emphasis is given to alternate management practices for row crops, grasslands/turf, and containerized horticultural crops that can be used to reduce potential N loss to groundwater resources.

Why N in Groundwater Is a Problem

Human Health Impacts

Groundwater is a major source of water for human consumption. When contaminated, the N is usually in the form of NO_3, which can pose major human health concerns at high levels, especially for infants. The link between high NO_3 in polluted water and serious blood changes in infants was first reported in 1945. From 1947 to 1950, 139 cases of methemoglobinemia were reported, including 14 deaths in Minnesota alone. In response to this documented threat to human health, a MCL standard has been set stating that NO_3 in excess of 45 mg L^{-1} (10 mg L^{-1} NO_3-N) is considered hazardous to human health.[4]

Even though the number of reported cases of methemoglobinemia has greatly decreased, recent studies indicate possible adverse impacts on human health at NO_3 levels below MCL.[5–7] Occurrences of bladder and ovarian cancer[8] and non-Hodgkin's lymphoma[9] have been linked to people with long-term exposure to public water supplies with NO_3-N concentrations of 2 to 4 mg L^{-1}.

Environmental Impacts

As part of the hydrologic cycle, a portion of what is classified as groundwater, especially shallow groundwater, resurfaces to feed streams, rivers, reservoirs, and eventually, estuaries and oceans. Nutrients and pollutants in the groundwater are thus transported in surface waters as well. A major challenge associated with groundwater contamination is that its flow is generally very slow. It can take years or decades for water to move through an aquifer from recharge areas to discharge areas and for land management changes/chemical applications to be reflected in "downstream" groundwater quality.[1,10,11] Green et al.[12] concluded that current fertilizer management practices in the United States will likely affect regional groundwater quality for decades to centuries.

Over the past 40 years, there has been an eightfold increase in the use of synthetic N fertilizers.[13] This has led to increased NO_3-N contamination in both groundwater and surface water bodies.[2,14,15] High NO_3-N levels can cause excess plant and bacterial growth in aquatic systems. Subsequently, the decay of this organic matter can deplete much of the oxygen in the water, causing fish kills and dead zones to appear. Phosphorus receives much of the attention with regard to eutrophication in fresh waters because it often is the limiting nutrient, but as water systems become more brackish, N often becomes the limiting factor.[16] Well-known examples of this situation include hypoxia in the Gulf of Mexico,[17] Danish coastal waters,[18] the western Indian continental shelf,[19] and the Changjiang Estuary in the East China Sea.[20] Hypoxia occurs when the concentration of dissolved oxygen is less than 2 mg L^{-1}. Nitrogen contributions to the Gulf of Mexico, and other large bodies of water, come from multiple sources, including surface runoff, subsurface drainage, and resurfacing of groundwater. There are several sources of N, e.g., fertilizer, animal manures, septic tanks, atmospheric deposition, land application of treated wastewater and biosolids, and mineralization of organic matter that can contribute to NO_3-N contamination. For example, in a 960 km^2 basin in Florida, it was estimated that fertilizer applied to cropland, lawns, and pine stands contributed 51% of the annual N load to groundwater in the basin.[21]

Agricultural Practices Contributing to Groundwater NO_3

Row Crops

High levels of NO_3-N in subsurface drainage from row crops, especially corn (*Zea mays* L.), are well documented. Nitrate-nitrogen concentrations in tile drainage from silt loam soils in Iowa that were fertilized for either continuous corn or corn grown in rotation with soybean [*Glycine max* (L.) Merr.] were reported to exceed 10 mg L^{-1} more than two decades ago.[22–24] Similar findings have been reported for clay loam soil in Minnesota,[25–27] silty clay loam in Illinois,[28] silt loam in Indiana,[29] silt loam and silty clay loam in Ohio,[30] and clay over silty clay loam and fine sand over clay in Ontario, Canada.[31] Analyses of subsurface water collected with monolith lysimeters[32,33] and ceramic porous-cup samplers[34,35] are in agreement with these findings.

Nitrate-nitrogen concentrations in tile drainage have been studied frequently because of the widespread use of artificial drainage throughout the U.S. Corn and Soybean Belt and the relative ease of collecting water samples. The majority of NO_3-N moves in the subsurface water during the late autumn, winter, and early spring recharge period[28,29,32] in this region. There are several factors influencing the amount of N exported via tile drainage, including timing, rate, source, and area receiving N fertilization.[28] Variable weather patterns, especially rainfall amounts, have a major influence on the amount of drainage that occurs. Increasing subsurface drain spacing can decrease NO_3-N losses,[25,29] although the NO_3-N concentration in the drainage water may change very little.[29] Furthermore, even though the increased spacing may reduce NO_3-N losses through the drainage water, it probably increases NO_3-N losses in seepage below the drains. Model simulation studies show that reducing N fertilization rates will have much greater impact for reducing NO_3-N losses than changing tile drain spacing or depth.[25]

There are many management factors contributing to the leaching loss of NO_3-N. Nitrogen is an essential plant nutrient that exhibits easily recognizable visual symptoms (e.g., yellow or light green plants, slow growth and development, decreased yield) when plant-available supplies are low. Also, until recently, N fertilizer was relatively inexpensive, and therefore, to reduce the risk of encountering a deficiency, it was often applied at rates in excess of crop need to ensure that inadequate N would not limit crop yield. Another factor is the difficulty in synchronizing N application with crop need because of the narrow window of time that producers may have if weather conditions are not optimum. As a result, it has been shown conclusively that there is a direct relationship between NO_3-N loss by leaching and N application rates that exceed crop needs.[34,36,37] Soil NO_3-N in excess of current plant needs can result from overapplication of N fertilizer or manure or from not accurately accounting for residual N from previous years, mineralization of organic N, or decomposition of legume crops. The latter is especially true following a dry year because if water limits crop growth, the plant will not utilize as much N as when a "normal" amount of water is available.[27] When this occurs, there is an increased amount of residual N to begin the next cropping season. Nitrogen management is therefore very difficult, and even when EON practices are used, it is not unusual to find NO_3-N concentrations in subsurface water that exceed the 10 mg L^{-1} MCL standard for drinking water.[34–36] This leads to the conclusion that optimum corn production will likely result in elevated NO_3-N concentrations in groundwater.[36] Stated in another manner, to achieve NO_3-N concentrations in groundwater that are less than the MCL, N fertilizer rates will need to be below the level associated with normal crop production recommendations.[37]

Nitrogen fertilizer management is also difficult within irrigated production systems, as evidenced by high NO_3-N concentrations found in subsurface water beneath sprinkler-irrigated crops in Spain[38] and Nebraska.[39] High NO_3-N concentrations also occurred in North Dakota[40] with sprinkler-irrigated corn and intermittent soybean and potato (*Solanum tuberosum* L.). The different crops did not directly affect NO_3-N concentrations. The most important factor leading to increased NO_3-N moving through the root zone and into groundwater was the amount of fall residual NO_3-N in the soil profile. Flood-irrigated wheat in Arizona[41] also produced high NO_3-N concentrations in groundwater, and even with best management practices (BMPs) for flood irrigation, NO_3-N concentrations in groundwater in excess of the MCL can be expected.

Animal manures are generally land applied for both disposal and their fertilizer nutrient value. This is especially true for organic production systems. High rates of manure application can result in high groundwater NO_3-N concentrations,[42,43] but even at low rates, manure management can be very difficult. For example, when liquid swine manure was applied at rates to meet the N requirements for either continuous corn or the corn phase of a corn/soybean rotation in Iowa, NO_3-N concentrations in tile flow from manure and urea ammonium NO_3 treatments consistently exceeded the MCL for drinking water. Furthermore, NO_3-N concentrations from the manure treatment exceeded those from the fertilizer treatments.[42] In another Iowa study, swine manure applications to both corn and soybean in a corn/soybean rotation were compared with application to only the corn phase of the rotation.[44] Average flow-weighted NO_3-N concentrations and leaching losses to subsurface drainage water were more than 50% greater when manure was applied to both corn and soybean. This response was credited to greater total N application when manure was applied every year compared with every other year. The studies also identified some of the challenges associated with using manure as a nutrient source. These include variability in manure composition and the importance of knowing the composition at the time of application.[43] Because of these factors, knowing the quantity of manure-applied N is much more difficult than for mineral fertilizer. Therefore, manure is often applied at a target rate to meet crop P requirements, with additional N subsequently being applied using another source.

Although leaching of NO_3-N dominates N loss to surface and groundwater, movement of dissolved organic N (DON) has also been recognized for more than 100 years.[45] Numerous studies have shown that DON leaching from forest ecosystems can be substantial, but DON leaching from agricultural soil has received little attention. Significant amounts of DON leaching were reported for agricultural soils in England,[46] and for cropped soils in Germany, DON leaching accounted for 6–21% of the total N flux.[47] In Ohio, DON accounted for 32–37% of the total N leaching from corn/soybean rotations,[48] with the primary source of DON in agricultural soils being crop residues and soil organic matter.[45] Just as for mineral N dynamics,[46] several factors, including leaching, mineralization, immobilization, and plant uptake, also affect DON leaching.

Grasslands/Turf

Because of the animal component, NO_3-N leaching in grazed grassland is very complex,[49,50] and even on highly fertilized pastures, much of the N loss has been attributed to excreta.[51,52] Considerable NO_3-N leaching can occur from the feces and urine in a management-intensive grazing system[53] and can result in a greater impact on water quality than moderate N fertilization.[54] A New Zealand study showed that when subsurface drainage occurred immediately following intensive grazing, drainage water had higher total N concentrations as a result of total organic N and NH_4-N from direct drainage or preferential flow of cattle urine.[55] However, those elevated concentrations were short lived and contributed a relatively small amount to the total N loss. Studies in England,[56,57] the Netherlands,[58] and the eastern United States[59,60] have shown that NO_3-N concentrations in subsurface water are often greater than the MCL when >100 kg N ha^{-1} was applied annually to grazed grasslands. In grazed bahiagrass (*Paspalum notatum* Fluegge) pastures in Florida that were fertilized with 76.5 kg N ha^{-1} (the recommended rate), NO_3-N leaching was not harmful to water quality.[61] Other processes, including accumulation of fertilizer N during drought or release of N from decaying plant material following tillage or chemically killing the sod in preparation for reseeding, may affect N leaching from pastures more than the loss from small areas affected by urine.[51] Once again, management is the key as demonstrated by a field experiment in the Netherlands showing that cattle slurry can replace some or most of the mineral N fertilizer for cut grassland on wet sandy soils without increasing NO_3-N leaching if the slurry is applied during the growing season and at rates that do not exceed crop uptake.[62] This was not the case on dry sandy soil, perhaps because denitrification in the wet soil minimized the amount of N available for leaching. Nevertheless, long-term leaching risk may increase as soil organic N increases unless mineral N applications are decreased accordingly.[61] In some nongrazed systems, NO_3-N leaching from highly fertilized systems has been reported to be very low, e.g., 29 kg N ha^{-1} lost from ryegrass (*Latium perenne* L.) receiving 420 kg N ha^{-1}.[63]

Fertilized turf, whether for home lawns or golf courses, is another N source with potential environmental issues. Annual applications up to 244 kg N ha^{-1} to turfgrass on sandy loam soils in Rhode Island did not appear to pose a threat to drinking water aquifers,[64] but overwatering did increase N loadings to bays and estuaries in coastal areas. A literature review by the Horsley Witten Group[65] concluded that a 20% loading rate to groundwater was an adequate estimate for modeling N leaching in the Cape Cod, Massachusetts, area. The excess N movement was most prevalent with late-summer N applications. Several studies have shown that late-fall applications of N fertilizer to turf result in higher N leaching rates during the late fall and winter.[66–69] Some late-season (October through December) N fertilization recommendations are for the purpose of maintaining/improving turf color, but increased N leaching from such applications indicated they should be managed to achieve acceptable water quality and not to maximize turf color.[67,70] Slow-release formulations of N fertilizer have also shown great reductions in the rate of N leaching.[66,69] The differences were even greater during a high-precipitation year than during a low-or normal-precipitation year,[69] especially with high autumn precipitation.[68] Irrigation management is critical, as evidenced by studies showing that increasing irrigation from 70% to 140% replacement of daily pan evaporation increased N leaching by almost 400%[71] and that flushing porous golf greens with high rates of irrigation can leach elevated rates of N.[72] Grass species is an important factor influencing N leaching, with less occurring beneath grasses with greater aboveground biomass and deeper root systems, e.g., bentgrass species (*Agrostis* spp.). High-quality grass could also reduce fertilizer N leaching under sand-based putting greens.[73]

Turfgrass is most vulnerable for NO_3-N leaching during establishment. Even though this period represents a small period of the average turf's life,[65] the soil disturbance, limited root biomass and N uptake, and tendency of turf managers to overfertilize during this period cause the NO_3-N leaching losses to exceed the MCL.[74,75] A mature turf may have actual N requirements below recommended levels. Research in Michigan on a 10-year-old Kentucky bluegrass (*Poa pratensis* L.) compared annual applications of 245 kg N ha^{-1} (49 kg N ha^{-1} per application) and 98 kg N ha^{-1} (24.5 kg N ha^{-1} application).[76] With resultant NO_3-N leachate concentrations often greater than 20 mg L^{-1} and below 5 mg L^{-1}, respectively, the conclusion was that applying high rates of water-soluble N to mature turfgrass should be avoided to minimize the potential for NO_3-N leaching. Returning clippings to the turfgrass ecosystem reduced the N fertilizer requirement by 25–60%, with the reduction increasing as the time after establishment increased from 10 to more than 50 years.[77]

Grass sod has the capacity to use large amounts of N, with 85–90% of fertilizer N being retained in the turf-soil ecosystem.[78] Roots and thatch can represent a large N pool that becomes available for mineralization and subsequent leaching if disturbed.[79] Reseeding and sod establishment within 2 mo of "turfdeath" can stabilize this N pool.[78] High rates of NO_3-N leaching can occur at very high N fertilizer rates, e.g., 450 kg N ha^{-1} per year. Although most NO_3-N leaching occurred during autumn and winter, it was the accumulation of all N fertilizer applications and not just the autumn application[68] that determined the actual loss. Excess soil NO_3 in the fall is the driving force that causes leaching, regardless of N source or time of application. Therefore, high rates of N application to turf should be avoided in the fall because it can result in high NO_3-N leaching rates. A significant portion of the N leached can be in the organic form.[71] A survey of several golf courses across the United States indicated that NO_3-N concentrations above the MCL occurred in only 4% of the samples,[80] with most of these being attributed to prior agricultural land use. Pollution of groundwater by NO3-N leaching from N-fertilized turf should be minimal with good management, which includes consideration of soil texture, N source, rate and timing, and irrigation/rainfall.[81]

Containerized Horticultural Crops

Although the acreage for containerized horticultural crops is small compared with row crops or grasslands, the production intensity is great, and "hot spots" of potential NO_3-N leaching can develop. Assuming 80,000 pots ha^{-1} for a typical foliage plant nursery and using a soluble granular fertilizer, over 650 kg N ha^{-1} could be lost annually through leaching.[82] During a 10-week greenhouse study of potted flowers, average NO_3-N in the leachate ranged from 250 mg NL^{-1} to 450 mg NL^{-1}.[83] Irrigation and fertilization are critical management components of this intensive industry. Trickle irrigation has been shown to move less water and leach less N than overhead irrigation,[84] but precipitation could nullify this difference.[85] Increasing the irrigation rate, e.g., from 1 to 2 cm day^{-1}, increased the amount of water lost and N leached[86] even though N concentrations were decreased. The use of controlled-release fertilizers (CRFs) is one practice that can significantly reduce leaching losses.[82,87] Even with the use of CRF, NO_3-N concentrations can be high in leachate[88]; if the use of CRF is combined with large irrigation volumes, NO_3-N can move into the soil profile beneath containers.[89] Water management is crucial, as illustrated by one experiment that compared a low leaching fraction (0.0 to 0.2 of the irrigation water) with a high leaching fraction (0.4 to 0.6). The lower leaching fraction reduced irrigation volume, effluent volume, and NO_3-N in the effluent by 44%, 63%, and 66%, respectively.[90] However, these gains in efficiency resulted in a 10% loss in total plant growth (shoots and roots). Thus, establishing an acceptable balance among the level of plant growth, water and nutrient use efficiencies, and the potential environmental impacts becomes a management decision. Although beyond the scope of this entry, vegetable crops also have a high N demand but low apparent N recovery, as illustrated by sweet peppers,[91] which can leave large amounts of N in the soil and residues at harvest.

Practices That Can Mitigate NO_3 Leaching from Agriculture

Nitrate contamination of groundwater is not caused by any single factor, because non-agricultural and agricultural practices contribute to the problem. Fertilizer management decisions (i.e., rates, formulations, timing, etc.) are an agricultural contributing factor, but so are tillage, crop selection, soil organic matter levels, and drainage.[92] Temperature and precipitation patterns also combine with these factors. Likewise, various strategies and approaches are needed to reduce NO_3-N loss to groundwater. This includes using appropriate N fertilizer rates, proper timing, soil testing and plant monitoring, nitrification inhibitors, cover crops, diversified crop rotations, and reduced tillage[92] as well as various combinations of these and other management practices.

Testing, Timing, Rates of Application, and Nitrification Inhibitors

Preplant N tests or pre-side-dress N tests[93] can assess soil N levels from cover crops and help provide adequate N credits for legume or fertilizer carryover from prior crops. For example, in Iowa, use of the late-spring NO_3 test reduced fertilizer N application[94] and resulted in up to a 30% decrease in NO_3-N in discharge waters compared with using traditional fall application. Testing plant tissue for N concentration can determine in-season crop N status and be used to guide supplemental N fertilizer applications. Calibrated chlorophyll meter readings have been correlated to plant N status[95–97] and used to increase N fertilizer efficiency and reduce leaching.[96] Splitting N applications also has the potential to reduce the total application rate and to apply the fertilizer N when it is needed by the crop. However, one disadvantage of applying N fertilizer after initial crop growth is that weather conditions may create a very small window of opportunity for side-dress N applications.

Nitrification inhibitors used with ammoniacal N sources can slow the rate of oxidation to NO_3-N and thus decrease the amount of NO_3-N available for leaching,[98] especially with fall N applications. In a Minnesota study, NO_3-N losses in subsurface drainage from a corn/soybean rotation were reduced 14% by spring N application compared with late-fall anhydrous ammonia application and 10% by late-fall N application using the nitrification inhibitor nitrapyrin.[99] The use of nitrapyrin with spring-applied N showed no further reduction of N losses. However, even with these improved practices, it may be necessary to reduce the N fertilization rate to below the EON level to achieve NO_3-N concentrations in groundwater that are below the MCL[37] for drinking water.

Winter Cover Crops, Diversified Crop Rotations, and Reduced Tillage

Winter cover crops have been shown to be an effective strategy for reducing NO_3-N leaching.[100–103] A variety of crops including annual grasses, cereals, and legumes have been used as cover crops with varying degrees of success—often depending upon the specific soil and climatic pattern of cropping sequence being used. Winter rye (*Secale cereale* L.) is the most common cover crop used to reduce NO_3-N leaching following corn or corn-soybean rotations in the United States.[100,101,104–106] Winter cover crops have also significantly reduced NO_3-N leaching for broccoli (*Brassica oleracea* L.) crops[107] and potato (*S. tuberosum* L.)–based rotations[108] in the United States. For potato, spring wheat (*Triticum aestivum* L.), sugar beet (*Beta vulgaris* L.), and oat (*Avena sativa* L.) rotation in the Netherlands, adding cover crops helped to decrease NO_3- N concentrations in leachate to near or below the European Union standard (11.3 mg L^{-1}).[109] Care needs to be exercised with long-term cover crops because if they are disturbed, some of the accumulated N may become mineralized and actually increase NO_3-N leaching.[110] Even in short-term situations, the efficacy of cover crops to reduce NO_3-N leaching is relatively low when considering the entire crop succession, and N saved by the cover crop generally does not increase N utilization by the next crop.[111] Deep-rooted cover crops may help capture N leached to deeper soil layers,[112] and rapid establishment of a deeprooting system is one factor influencing the efficacy of cover crops for reducing NO_3-N leaching.[113] Although sometimes cover crops are considered to be a BMP for reducing NO_3-N leaching,[114] this is not always true. For example, on the Delmarva Peninsula in the mid-Atlantic United States, a rye winter cover crop following corn did not reduce NO_3-N leaching, presumably because the existing corn crop did not allow the rye to be seeded early enough in the autumn. Similarly, owing to the restricted time for cover crop growth in a Wisconsin study, winter rye did not utilize significant amounts of fertilizer N from the previous crop residues or soil.[115] However, in temperate regions with mild winters that favor long growing seasons and winter N mineralization, cover crops have been demonstrated as a valuable tool to reduce NO_3-N leaching.[116]

Diversifying cropping systems to include perennial crops is another strategy that will probably reduce NO_3-N leaching.[117] Compared with annual crops, perennials have an extended period of growth and therefore greater N utilization. However, having an adequate market for satisfactory economic return on such crops can be problematic. Fortunately, the emerging biofuel industry provides a potential market that may make such cropping systems more profitable.

Reduced tillage is a BMP for reducing soil erosion, but it is not usually an option for N management. Tillage promotes N mineralization from soil organic matter, which decreases soil quality and makes more N available for oxidation and subsequent leaching. Although water movement through the soil profile tends to be greater with no-till than conventional or limited tillage, NO_3-N concentrations are usually lower with no-till. Therefore, the amount of NO_3-N leaching can be either similar or less with no-till.[118] Overall, differences are small, and it is often concluded that tillage has less impact on NO_3-N leaching than factors such as crop rotation.[92,117,119]

Use of Alternate Grassland/Turf Management

Demonstrated management options to reduce NO3-N leaching from grasslands include using grass–legume mixtures instead of highly fertilized grass[52,120]; ensuring total N input does not exceed 100 kg ha^{-1}[59,60]; coordinating fertilizer application rate and timing with other N sources (e.g., manure applications); avoiding excessive N application rates[121]; using irrigation, especially during dry periods, to encourage N uptake[52]; and integrating forage cutting and grazing for optimum management throughout the year. Concentrations of NO_3-N in groundwater associated with high-fertility, high-stocking-density grazing systems can also be reduced by continuous grazing or haying if external N inputs are reduced or eliminated.[122]

In areas where NO_3-N contamination from turf is a concern, late-summer N fertilizer applications should be reduced, and the amount of irrigation water should be limited.[64] Other management practices for reducing N leaching from turf include using slow-release fertilizers,[66,69] avoiding excess irrigation, and using grass species with greater aboveground biomass and deeper root systems.[73] Because the vulnerability for leaching during turf establishment, irrigation and fertilizer rates should be limited until the turf root system is well developed. Throughout the midwestern United States, excess soil NO_3-N in the fall is the driving force responsible for NO_3-N leaching. Therefore, using any or all of the management practices mentioned above should help avoid this buildup and thus help reduce NO_3-N leaching.

Conclusions

High NO_3-N concentrations in drinking water can pose human health problems, and when found in surface water bodies, they can create many environmental problems. Several factors have contributed to high groundwater NO_3-N concentrations, including the increased use of relatively inexpensive N fertilizer to minimize the yield risk associated with encountering nutrient deficiencies. The unintended consequence is that when N is applied at rates exceeding levels to which crops respond, the potential for excess N to accumulate in the soil and leach is increased. Thus, EON rates may often produce NO_3-N levels in groundwater that exceed MCLs. This has been measured with grain crops, grasslands, and turf, which receive high rates of fertilizer N, and nursery container crops.

There are several management practices that can help reduce excess N application. These include splitting applications of N so that it is most available when plants need it; using preplant, pre-sidedress, and plant tissue N tests to determine appropriate rates of application; applying slow-release fertilizer sources so that the N is available as plants need it instead of all at once; improving timeliness of irrigation to help plants take up N during drought periods and to avoid excess irrigation, which will reduce leaching; avoiding late-season fertilizer applications, especially common for turf; and using cover crops to capture excess N that remains in the soil profile after growth of the crop so that NO_3-N leaching is reduced.

Research challenges to improve N management and reduce NO_3-N leaching are multidisciplinary. Although *precision agriculture* is a term generally applied to using soil tests, yield results, and Geographic Information system (GIS) mapping to match fertilizer rates with yield responses in a field, it also has the potential to reduce NO_3-N leaching by encouraging the use of variable N rates instead of a constant rate for an entire field. Traditionally, the sole focus for agriculture has been on crop yield, but to simultaneously address groundwater quality, effects of climate change, and other emerging factors, interactions with other crops must also be included. Collectively, these practices can improve overall N use efficiency and have both economic and environmental benefits.

References

1. Debrewer, L.M.; Ator, S.W.; Denver, J.M. Temporal trends in nitrate and selected pesticides in mid-Atlantic ground water. J. Environ. Qual. **2008**, *37* (5), 296–308.
2. Rupert, M.G. Decadal-scale changes of nitrate in ground water of the United States, 1988–2004. J. Environ. Qual. **2008**, *37* (5), 240–248.
3. Spalding, R.F.; Exner, M.E. Occurrence of nitrate in groundwater—A review. J. Environ. Qual. **1993**, *22* (3), 392–402.
4. U.S. Public Health Service. *Public Health Drinking Water Standards,* U.S. Public Health Pub I. 956; U.S. Government Printing Office: Washington, DC, 1996.
5. Brender, J.D.; Olive, J.M.; Felkner, J.M.; Suarez, L.; Hendricks, K.; Marckwardt, W. Intake of nitrates and nitrites and birth defects in offspring. Epidemiology **2004**, *15* (4), 184.
6. Brender, J.D.; Olive, J.M.; Felkner, J.M.; Suarez, L.; Marckwardt, W.; Hendricks, K.A. Dietary nitrites and nitrates, nitrosatable drugs, and neural tube defects. Epidemiology **2004**, *15* (3), 330–336.
7. DeRoos, A.J.; Ward, M.H.; Lynch, C.F.; Cantor, K.P. Nitrate in public water systems and the risk of colon and rectum cancers. Epidemiology **2003**, *14* (6), 640–649.
8. Weyer, P.J.; Cerhan, J.R.; Kross, B.C.; Hallberg, G.R.; Kantamneni, J.; Breuer, G.; Jones, M.P.; Zheng, W.; Lynch, F. Municipal drinking water nitrate level and cancer risk in older women: The Iowa Women's Health Study. Epidemiology **2001**, *12* (3), 327–338.
9. Ward, M.H.; Mark, S.D.; Cantor, K.P.; Weisenburger, D.D. Drinking water nitrate and the risk of non-Hodgkin's lymphoma. Epidemiology **1996**, *7* (5), 465–471.
10. Bohlke, J.K. Groundwater recharge and agricultural contamination. Hydrogeol. J. **2002**, *10* (1), 153–179.
11. Puckett, L.J.; Hughes, W.B. Transport and fate of nitrate and pesticides: Hydrogeology and riparian zone processes. J. Environ. Qual. **2005**, *34* (6), 2278–2292.
12. Green, C.T.; Puckett, L.J.; Bohlke, J.K.; Bekins, B.A.; Phillips, S.P.; Kauffman, L.J.; Denver, J.M.; Johnson, H.M. Limited occurrence of denitrification in four shallow aquifers in agricultural areas of the United States. J. Environ. Qual. **2008**, 37 (3), 994–1009.
13. Mulvaney, R.L.; Khan, S.A.; Ellsworth, T.R. Synthetic nitrogen fertilizers deplete soil nitrogen: A global dilemma for sustainable cereal production. J. Environ. Qual. **2009**, *38* (6), 2295–2314.
14. Goolsby, D.A.; Battaglin, W.A. Long-term changes in concentrations and flux of nitrogen in the Mississippi River Basin, USA. Hydrol. Processes **2001**, *15* (7), 1209–1226.
15. Zhu, J.G.; Liu, G.; Han, Y.; Zhang, Y.L.; Xing, G.X. Nitrate distribution and denitrification in the saturated zone of paddy field under rice/wheat rotation. Chemosphere **2003**, *50* (6), 725–732.
16. Correll, D.L. The role of phosphorus in the eutrophication of receiving waters: A review. J. Environ. Qual. **1998**, *27* (2), 261–266.
17. Burkart, M.R.; James, D.E. Agricultural-nitrogen contributions to hypoxia in the Gulf of Mexico. J. Environ. Qual. **1999**, *28* (3), 850–859.
18. Conley, D.J.; Carstensen, J.; Aertebjerg, G.; Christensen, P.B.; Dalsgaard, T.; Hansen, J.L.S.; Josepson, A.B. Longterm changes and impacts of hypoxia in Danish coastal waters. Ecol. Appl. *17* (5), S165–S184.
19. Naqvi, S.W.A.; Jayakumar, D.A.; Narvekar, P.V.; Naik, H.; Sarma, V.V.S.S.; D'Souza, W.D.; Joseph, S.; George, M.D. Increased marine production of N_2O due to intensifying anoxia on the Indian continental shelf. Nature *408* (6810), 346–349.
20. Wei, H.; He, Y.; Li, Q.; Liu, Z.; Wang, H. Summer hypoxia adjacent to the Changjiang Estuary. J. Mar. Syst. *67* (3–4), 252–303.
21. Katz, B.G.; Sepulveda, A.A.; Verdi, R.J. Estimating nitrogen loading to ground water and assessing vulnerability to nitrate contamination in a large karstic springs basin, Florida. J. Am. Water Res. Assoc. **2009**, *45* (3), 607–627.

22. Baker, J.L.; Campbell, H.P.; Johnson, H.P.; Hanway, J.J. Nitrate, phosphorus, and sulfate in subsurface drainage water. J. Environ. Qual. **1975**, *4* (3), 406–412.
23. Baker, J.L.; Johnson, R.P. Nitrate-nitrogen in tile drainage as affected by fertilization. J. Environ. Qual. **1981**, *10* (4), 519–522.
24. Kanwar, R.S.; Cruse, R.M.; Ghaffarzadeh, M.; Bakhsh, A.; Karlen, D.L; Bailey, T.B. Corn-soybean and alternative cropping systems effects on NO_3-N leaching losses in subsurface drainage water. Appl. Eng. Agric. **2005**, *21* (2), 181–188.
25. Davis, D.M.; Gowda, P.R.; Mulla, D.J.; Randall, G.W. Modeling nitrate nitrogen leaching in response to nitrogen fertilizer rate and tile drain depth or spacing for southern Minnesota, USA. J. Environ. Qual. **2000**, *29* (5), 1568–1581.
26. Randall, G.W.; Huggins, D.R.; Russelle, M.P.; Fuchs, D.J.; Nelson, W.W.; Anderson, J.L. Nitrate losses through subsurface tile drainage in Conservation Reserve Program, alfalfa, and row crop systems. J. Environ. Qual. **1997**, *26* (5), 1240–1247.
27. Randall, G.W.; Iragavarapu, T.K. Impact of long-term tillage systems for continuous corn on nitrate leaching to tile drainage. J. Environ. Qual. **1995**, *24* (2), 360–366.
28. Gentry, L.E.; David, M.B.; Smith, K.M.; Kovacic, D.A. Nitrogen cycling and tile drainage nitrate loss in corn/soybean watershed. Agric. Ecosyst. Environ. **1998**, *68* (1–2), 85–97.
29. Kladivko, E.J.; Grochulska, J.; Turco, R.F.; Van Scoyoc, G.E.; Eigel, J.D. Pesticide and nitrate transport into subsurface tile drains of different spacings. J. Environ. Qual. **1999**, *28* (3), 997–1004.
30. Logan, T.J.; Schwab, G.O. Nutrient and sediment characteristics of tile effluent in Ohio. J. Soil Water Conserv. **1976**, *31* (1), 24–27.
31. Miller, M.H. Contribution of nitrogen and phosphorus to subsurface drainage water from intensively cropped mineral and organic soils in Ontario. J. Environ. Qual. **1979**, *8* (1), 42–48.
32. Cookson, W.R.; Rowatth, J.S.; Cameron, K.E. The effect of autumn applied^{15}N-labelled fertilizer on nitrate leaching in a cultivated soil during winter. Nutr. Cycling Agroecosyst. **2000**, *56* (2), 99–107.
33. Owens, L.B.; Malone, R.W.; Shipitalo, M.J.; Edwards, W.M.; Bonta, J.V. Lysimeter study on nitrate leaching from a corn-soybean rotation. J. Environ. Qual. **2000**, *29* (2), 467–474.
34. Andraski, T.W.; Bundy, L.G.; Brye, K.R. Crop management and corn nitrogen rate effects on nitrate leaching. J. Environ. Qual. **2000**, *29* (4), 1095–1103.
35. Steinheimer, T.R.; Scoggin, K.D.; Kramer, L.A. Agricultural chemical movement through a field-size watershed in Iowa: Subsurface hydrology and distribution of nitrate in groundwater. Environ. Sci. Technol. **1998**, *32* (8), 1039–1047.
36. Jemison, J.M., Jr.; Fox, R.H. Nitrate leaching from nitrogen-fertilized and manured corn measured with zero-tension pan lysimeters. J. Environ. Qual. **1994**, *23* (2), 337–343.
37. Lawlor, P.A.; Helmers, M.J.; Baker, J.L.; Melvin, S.W.; Lemke, D.W. Nitrogen application rate effect on nitrateitrogen concentration and loss in subsurface drainage for a corn-soybean rotation. Trans. ASABE **2008**, *51* (1), 83–94.
38. Diez, J.A.; Caballero, R.; Roman, R.; Tarquis, A.; Cartegena, M.E.; Vallejo, A. Integrated fertilizer and irrigation management to reduce nitrate leaching in central Spain. J. Environ. Qual. **2000**, *29* (5), 1539–1547.
39. Klocke, N.L.; Watts, D.G.; Schneekloth, J.P.; Davison, D.R.; Todd, R.W.; Parkhurst, A.M. Nitrate leaching in irrigated corn and soybean in a semi-arid climate. Trans. ASAE **1999**, *42* (2), 1621–1630.
40. Derby, N.E.; Casey, F.X.M.; Knighton, R.E. Long-term observations of vadose zone and groundwater nitrate concentrations under irrigated agriculture. Vadose Zone J. **2009**, *8* (2), 290–300.
41. Ottman, M.J.; Tickes, B.R.; Husman, S.H. Nitrogen-15 and bromide tracers of nitrogen fertilizer movement in irrigated wheat production. J. Environ. Qual. **2000**, 29 (5), 1500–1508.
42. Bakhsh, A.; Kanwar, R.S.; Karlen, D.L. Effects of liquid swine manure applications on NO_3-N leaching losses to subsurface drainage water from loamy soils in Iowa. Agric. Ecosyst. Environ. **2005**, *109* (1–2), 118–128.

43. Karlen, D.L.; Cambardella, C.A.; Kanwar, R.S. Challenges of managing liquid swine manure. Appl. Eng. Agric. **2004**, *20* (5), 693–699.
44. Bakhsh, A.; Kanwar, R.S.; Baker, J.L.; Sawyer, J.; Malarino, A. Annual swine manure applications to soybean under corn-soybean rotation. Trans. Am. Soc. Agric. Biol. Eng. **2009**, *52* (3), 751–757.
45. Van Kessel, C.; Clough, T.; van Groenigen, J.W. Dissolved organic nitrogen: An overlooked pathway of nitrogen loss from agricultural systems? J. Environ. Qual. **2009**, *38* (2), 393–401.
46. Murphy, D.V.; Macdonald, A.J.; Stockdale, E.A.; Goulding, K.W.T.; Fortune, S.; Gaunt, J.L.; Poulton, P.R.; Wakefield, J.A.; Webster, C.P.; Wilmer, W.S. Soluble organic nitrogen in agricultural soils. Biol. Fertil. Soils **2000**, *30* (5–6), 373–387.
47. Siemens, J.; Kaupenjohann, M. Contribution of dissolved organic nitrogen to N leaching from four German agricultural soils. J. Plant Nutr. Soil Sci. **2002**, *165* (6), 675–681.
48. Shuster, W.D.; Shipitalo, M.J.; Subler, S.; Aref, S.; McCoy,E.L. Earthworm additions affect leachate production and nitrogen losses in typical Midwestern agroecosystems. J. Environ. Qual. **2003**, 32 (2), 2132–2139.
49. Ball, P.R.; Ryden, J.C. Nitrogen relationships in intensively managed temperate grasslands. Plant Soil **1984**, *76* (1–3), 23–33.
50. Ryden, J.E.; Ball, P.R.; Garwood, E.A. Nitrate leaching from grassland. Nature **1984**, *311* (5981), 50–53.
51. Cuttle, S.P.; Scurlock, R.V.; Davies, B.M.S. A 6-year comparison of nitrate leaching from grass/clover and N fertilized grass pastures grazed by sheep. J. Agric. Sci., Cambridge **1998**, *131* (1), 39–50.
52. Whitehead, D.E. Leaching of nitrogen from soils. In *Grassland Nitrogen;* CAB International: Wallingford, 1995; 129–151.
53. Stout, W.L.; Fales, S.A.; Muller, L.D.; Schnabel, R.R.; Priddy, W.E.; Elwinger, G.F. Nitrate leaching from cattle urine and feces in northeast USA. Soil Sci. Soc. Am. J. **1997**, *61* (6), 1787–1794.
54. Decau, M.L.; Simon, J.C.; Jacquet, A. Nitrate leaching under grassland as affected by mineral nitrogen fertilization and cattle urine. J. Environ. Qual. **2004**, *33* (2), 637–644.
55. Houlbrooke, D.; Hanly, J.; Horne, D.; Hedley, M. Nitrogen losses in artificial drainage and surface runoff from pasture following grazing by dairy cattle. In *SuperSoil 2004,* Proceedings of the 3rd Australian New Zealand Soil Conference, Sydney, Australia, Dec 5–9, 2004; The Regional Institute Ltd., http://www.regional.org.au/au/asssi/, 8 pp.
56. Haigh, R.A.; White, R.E. Nitrate leaching from a small, underdrained, grassland, clay catchment. Soil Use Manage. **1986**, *2* (2), 65–70.
57. Roberts, G. Nitrogen inputs and outputs in a small agricultural catchment in the eastern part of the United Kingdom. Soil Use Manage. **1987**, *3* (4), 148–154.
58. Steenvoorden, J.H.A.M.; Fonck, H.; Oosterom, H.P. Losses of nitrogen from intensive grassland systems by leaching and surface runoff. In *Nitrogen Fluxes in Intensive Grassland Systems;* van der Meer, H.G., Ryden, J.E., Ennik, G.E., Eds.; Martinus Nijhoff Publ.: Dordrecht, The Netherlands, 1986; 85–97.
59. Owens, L.B.; Van Keuren, R.W.; Edwards, W.M. Nitrogen loss from a high-fertility, rotational pasture program. J. Environ. Qual. **1983**, *12* (3), 346–350.
60. Owens, L.B.; Edwards, W.M.; Van Keuren, R.W. Nitrate levels in shallow groundwater under pastures receiving ammonium nitrate or slow-release nitrogen fertilizer. J. Environ. Qual. **1992**, *21* (4), 607–613.
61. Sigua, G.C.; Hubbard, R.K.; Coleman, S.W.; Williams, M. Nitrogen in soils, plants, surface water and shallow groundwater in a bahiagrass pasture of Southern Florida, USA. Nutr. Cycling Agroecosyst. **2010**, *86* (2), 175–187.
62. Schroder, J.J.; Assinck, F.B.T.; Uenk, D.; Velthof, G.L. Nitrate leaching from cut grassland as affected by the substitution of slurry with nitrogen mineral fertilizer on two soil types. Grass Forage Sci. **2010**, *65* (1), 49–57.
63. Garwood, E.A.; Ryden, J.E. Nitrate loss through leaching and surface runoff from grassland: Effects of water supply, soil type and management. In *Nitrogen Fluxes in Intensive Grassland Systems;* van der Meer, H.G., Ryden, J.C., Ennik, G.E., Eds.; Martinus Nijhoff Publ.: Dordrecht, Netherlands, **1986**; 99–113.

64. Morton, T.G.; Gold, A.J.; Sullivan, W.M. Influence of overwatering and fertilization on nitrogen losses from home lawns. J. Environ. Qual. **1988**, *17* (1), 124–130.
65. Horsley Witten Group. *Evaluation of Turfgrass Nitrogen Fertilizer Leaching Rates in Soils on Cape Cod, Massachusetts*; MA Department of Environmental Protection: Hyannis, MA, 2009; 33 pp.
66. Guillard, K.; Kopp, K.L. Nitrogen fertilizer form and associated nitrate leaching from cool-season lawn turf. J. Environ. Qual. **2004**, *33* (5), 1822–1827.
67. Mangiafico, S.S.; Guillard, K. Fall fertilization timing effects on nitrate leaching and turfgrass color and growth. J. Environ. Qual. **2006**, *35* (1), 163–171.
68. Roy, J.W.; Parkin, G.W.; Wagner-Riddle, E. Timing of nitrate leaching from turfgrass after multiple fertilizer applications. Water Qual. Res. J. Canada **2000**, *35* (4), 735–752.
69. Petrovic, A.M. Nitrogen source and timing impact on nitrate leaching from turf. In *I International Conference on Turfgrass Management and Science for Sports Fields*, Athens, Greece, June 2–7, 2003. ISHS Acta Hortic. *661,* 427–432.
70. Mangiafico, S.S.; Guillard, K. Nitrate leaching from Kentucky bluegrass soil columns predicted with anion exchange membranes. Soil Sci. Soc. Am. J. **2007**, *71* (1), 219–224.
71. Barton, L.; Wan, G.G.Y.; Colmer, T.D. Turfgrass (*Cynodon dactylon* L.) sod production on sandy soils: II. Effects of irrigation and fertilizer regimes on N leaching. Plant Soil **2006**, *284* (1–2), 147–164.
72. Shuman, L.M. Normal and flush irrigation effects on nitrogen leaching from simulated golf greens in the greenhouse. Commun. Soil Sci. Plant Anal. **2006**, *37* (3–4), 605–619.
73. Pare, K.; Chantigny, M.H.; Carey, K.; Johnston, W.J.; Dionne, J. Nitrogen uptake and leaching under annual bluegrass ecotypes and bentgrass species: A lysimeter experiment. Crop Sci. **2006**, *46* (2), 847–853.
74. Geron, E.A.; Danneberger, T.K.; Traina, S.J.; Logan, T.J.; Street, J.R. The effects of establishment methods and fertilization practices on nitrate leaching from turfgrass. J. Environ. Qual. **1993**, *22* (1), 119–125.
75. Easton, A.M.; Petrovic, A.M. Fertilizer source effect on ground and surface water quality in drainage from turfgrass. J. Environ. Qual. **2004**, *33* (2), 645–655.
76. Frank, K.W.; O'Reilly, K.M.; Crum, J.R.; Calhoun, R.N. The fate of nitrogen applied to a mature Kentucky bluegrass turf. Crop Sci. **2006**, *46* (1), 209–215.
77. Qian, Y.L.; Bandaranayake, W.; Parton, W.J.; Mecham, B.; Harivandi, M.A.; Mosier, A.R. Long-term effects of clipping and nitrogen management in turfgrass on soil organic carbon and nitrogen dynamics: The CENTURY model simulation. J. Environ. Qual. **2003**, *32* (5), 1694–1700.
78. Bushoven, J.T.; Jiang, Z.; Ford, H.J.; Sawyer, E.D.; Hull, R.J.; Amador, J.A. Stabilization of soil nitrate by reseeding with ryegrass following sudden turf death. J. Environ. Qual. **2000**, *29* (5), 1657–1661.
79. Jiang, Z.; Bushoven, J.T.; Ford, H.J.; Sawyer, E.D.; Amador, J.A.; Hull, R.J. Mobility of soil nitrogen and microbial responses following the sudden death of established turf. J. Environ. Qual. **2000**, *29* (5), 1625–1631.
80. Cohen, S.; Svrjcek, A.; Durborow, T.; Barnes, N.L. Water quality impacts by golf courses. J. Environ. Qual. **1999**, *28* (3), 798–809.
81. Petrovic, A.M. The fate of nitrogenous fertilizers applied to turfgrass. J. Environ. Qual. **1990**, *19* (1), 1–14.
82. Broschat, T.K. Nitrate, phosphorus, and potassium leaching from container-grown plants fertilized by several methods. HortScience **1995**, *30* (1), 74–77.
83. McAvoy, RJ. Nitrate nitrogen movement through the soil profile beneath a containerized greenhouse crop irrigated with two leaching fractions and two wetting agent levels. J. Am. Soc. Hort. Sci. **1994**, *119* (3), 446–451.
84. Rathier, T.M.; Frink, C.R. Nitrate in runoff water from container grown juniper and Alberta spruce under different irrigation and N fertilization regimes. J. Environ. Hort. **1989**, *7* (1), 32–35.

85. Colangelo, D.J.; Brand, M.H. Nitrate leaching beneath a containerized nursery crop receiving trickle or overhead irrigation. J. Environ. Qual. **2001**, *30* (5), 1564–1574.
86. Million, J.; Yeager, T.; Albano, J. Consequences of excessive overhead irrigation on runoff during container production of sweet viburnum. J. Environ. Hort. **2007**, *25* (3), 117–125.
87. Million, J.; Yeager, T.; Albano, J. Effects of container spacing practice and fertilizer placement on runoff from overhead-irrigated sweet viburnum. J. Environ. Hort. **2007**, *25* (2), 61–72.
88. Newman, J.; Albano, J.; Merhaut, D.; Blythe, E. Nutrient release from controlled-release fertilizers in neutral-pH substrate in an outdoor environment: I. Leachate electrical conductivity, pH, and nitrogen, phosphorous, and potassium concentrations. HortScience **2006**, *41* (7), 1674–1682.
89. Colangelo, D.J.; Brand, M.K. Effect of split fertilizer application and irrigation volume on nitrate-nitrogen concentration in container growing area soil. J. Environ. Hort. **1997**, *15* (4), 205–210.
90. Tyler, H.H.; Warren, S.L.; Bilderback, T.E. Reduced leaching fractions improve irrigation use efficiency and nutrient efficacy. J. Environ. Hort. **1996**, *14* (4), 199–204.
91. Tei, F.; Benincasa, P.; Guiducci, M. Nitrogen fertilisation of lettuce, processing tomato and sweet pepper: Yield, nitrogen uptake and the risk of nitrate leaching. Welles- bourne, Warwick, United Kingdom. December 1999. Acta Hort. **1999**, *506,* 61–67.
92. Dinnes, D.L.; Karlen, D.L.; Jaynes, D.B.; Kaspar, T.C.; Hatfield, J.L.; Colvin, T.S.; Cambardella, C.A. Nitrogen management strategies to reduce nitrate leaching in tile-drained Midwestern soils. Agron. J. **2002**, *94* (1), 153–171.
93. Magdorff, F.R.; Ross, D.; Amadon, J. A soil test for nitrogen availability to corn. Soil Sci. Soc. Am. J. **1984**, *48* (6), 1301–1304.
94. Jaynes, D.B.; Dinnes, D.L.; Meek, D.W.; Karlen, D.L.; Cambardella, C.A.; Colvin, T.S. Using the Late Spring Nitrate Test to reduce nitrate loss within a watershed. J. Environ. Qual. **2004**, *33* (2), 669–677.
95. Schepers, J.S.; Francis, D.D.; Vigil, M.; Below, F.E. Comparison of corn leaf nitrogen concentration and chlorophyll meter readings. Commun. Soil Sci. Plant Anal. **1992**, *23* (17–20), 2173–2187.
96. Varvel, G.E.; Schepers, J.S.; Francis, D.D. Ability for in-season correction of nitrogen deficiency in corn using chlorophyll meters. Soil Sci. Soc. Am. J. **1997**, *61* (4), 1233–1239.
97. Hawkins, J.A.; Sawyer, J.E.; Barker, D.W.; Lundvall, J.P. Using relative chlorophyll meter values to determine nitrogen application rates for corn. Agron. J. **2007**, *99* (4), 1034–1040.
98. Owens, L.B. Nitrate leaching losses from monolith lysimeters as influenced by nitrapyrin. J. Environ. Qual. **1987**, *16* (1), 34–38.
99. Randall, G.W.; Vetsch, J.A. Nitrate losses in subsurface drainage from a corn-soybean rotation as affected by fall and spring application of nitrogen and nitrapyrin. J. Environ. Qual. **2005**, *34* (2), 590–597.
100. McCracken, D.V.; Smith, M.S.; Grove, J.H.; MacKown, E.T.; Blevins, R.L. Nitrate leaching as influenced by cover cropping and nitrogen source. Soil Sci. Soc. Am. J. **1994**, *58* (5), 1476–1483.
101. Rasse, D.P.; Ritchie, J.T.; Peterson, W.R.; Wei, J.; Smucker, A.J. Rye cover crop and nitrogen fertilization effects on nitrate leaching in inbred maize fields. J. Environ. Qual. **2000**, *29* (1), 298–304.
102. Zhou, X.; MacKenzie, A.F.; Madramootoo, C.A.; Kaluli, J.W.; Smith, D.L. Management practices to conserve soil nitrate in maize production systems. J. Environ. Qual. **1997**, *26* (5), 1369–1374.
103. Ball-Coelho, B.R.; Roy, R.C. Overseeding rye into corn reduces NO_3 leaching and increases yield. Can. J. Soil Sci. **1997**, *77*, 443–451.
104. Kaspar, T.C.; Jaynes, D.B.; Parkin, T.B.; Moorman, T.B. Rye cover crop and gamagrass strip effects on NO_3 concentration and load in tile drainage. J. Environ. Qual. **2007**, *36* (5), 1503–1511.
105. Li, L.; Malone, R.W.; Ma, L.; Kaspar, T.C.; Jaynes, D.B.; Saseendran, S.A.; Thorp, K.R.; Yu, Q.; Ahuja, L.R. Winter cover crop effects on nitrate leaching in subsurface drainage as simulated by RZWQM-DSSAT. Trans. ASABE **2008**, *51* (5), 1575–1583.
106. Strock, J.S.; Porter, P.M.; Russelle, M.P. Cover cropping to reduce nitrate loss through subsurface drainage in the northern U.S. corn belt. J. Environ. Qual. **2004**, *33* (3), 1010–1016.

107. Wyland, L.J.; Jackson, L.E.; Chaney, W.E.; Klonsky, K.; Koike, S.T.; Kimple, B. Winter cover crops in a vegetable cropping system on nitrate leaching, soil water, crop yield, pests and management costs. Agric. Ecosyst. Environ. **1996**, *59* (1–2), 1–17.
108. Weinert, T.L.; Pan, W.L.; Moneymaker, M.R.; Santo, G.S.; Stevens, R.G. Nitrogen recycling by nonleguminous winter cover crops to reduce leaching in potato rotations. Agron. J. **2002**, *94* (2), 365–372.
109. Vos, J.; van der Putten, P.E.L. Nutrient cycling in a cropping system with potato, spring wheat, sugar beet, oats and nitrogen catch crops. II. Effect of catch crops on nitrate leaching in autumn and winter. Nutr. Cycling Agroecosyst. **2004**, *70* (1), 23–31.
110. Hansen, E.M.; Djurhuus, J.; Kristensen, K. Nitrate leaching as affected by introduction or discontinuation of cover crop use. J. Environ. Qual. **2000**, *29* (4), 1110–1116.
111. Herrera, J.M.; Liedgens, M. Leaching and utilization of nitrogen during a spring wheat catch crop succession. J. Environ. Qual. **2009**, *38* (4), 1410–1419.
112. Kristensen, H.L.; Thorup-Kristensen, K. Root growth and nitrate uptake of three different catch crops in deep soil layers. Soil Sci. Soc. Am. J. **2004**, *68* (2), 529–537.
113. Herrera, J.M.; Feil, B.; Stamp, P.; Liedgens, M. Root growth and nitrate-nitrogen leaching of catch crops following spring wheat. J. Environ. Qual. **2010**, *39* (3), 845–854.
114. Ritter, W.F.; Scarborough, R.W.; Christie, A.E.M. Winter cover crops as a best management practice for reducing nitrogen leaching. J. Contam. Hydrol. **1998**, *34* (1), 1–15.
115. Bundy, L.G.; Andraski, T.W. Recovery of fertilizer nitrogen in crop residues and cover crops on an irrigated sandy soil. Soil Sci. Soc. Am. J. **2005**, *69* (3), 640–648.
116. Hooker, K.V.; Coxon, C.E.; Hackett, R.; Kirwan, L.E.; O'Keeffe, E.; Richards, K.G. Evaluation of cover crop and reduced cultivation for reducing nitrate leaching in Ireland. J. Environ. Qual. **2008**, *37* (1), 138–145.
117. Randall, G.W.; Mulla, D.J. Nitrate nitrogen in surface waters as influenced by climatic conditions and agricultural practices. J. Environ. Qual. **2001**, *30* (2), 337–344.
118. Kanwar, R.S.; Baker, J.L.; Baker, D.G. Tillage and split N-fertilization effects on subsurface drainage water quality and crop yields. Trans. ASAE **1988**, *31* (2), 453–461.
119. Zhu, Y.; Fox, R.H.; Toth, J.D. Tillage effects on nitrate leaching measured by pan and wick lysimeters. Soil Sci. Soc. Am. J. **2003**, *67* (5), 1517–1523.
120. Owens, L.B.; Edwards, W.M.; Van Keuren, R.W. Groundwater nitrate levels under fertilized grass and grass-legume pastures. J. Environ. Qual. **1994**, *23* (4), 752–758.
121. Jarvis, S.C. Progress in studies of nitrate leaching from grassland soils. Soil Use Manage. **2000**, *16* (s1), 152–156.
122. Owens, L.B.; Bonta, J.V. Reduction of nitrate leaching with haying or grazing and omission of nitrogen fertilizer. J. Environ. Qual. **2004**, *33* (4), 1230–1237.

5

Groundwater: Pesticide Contamination

Roy F. Spalding

Introduction

Trace concentrations of most of the commonly used pesticides have been confirmed in groundwaters of the United States. Since groundwater is the source of 53% of the potable water, the more toxic pesticides and their transformation products are a concern from the standpoint of human health. Others are a risk to the environment in areas where contaminated groundwater enters surface water. Through toxicological testing, the USEPA has established Maximum Contaminant Levels (MCLs) or lifetime Health Advisory Levels (HALs) for several pesticides (Table 1).

The EPA also has a separate list of unregulated compounds, including newly registered pesticides and their transformation products, such as acetochlor and alachlor ESA, that are presently being evaluated or being considered for toxicological evaluation. Based on the results of the EPAs National Pesticide Assessment,[2] 10.4% of 94,600 community systems contained detectable concentrations of at least one pesticide. Evaluation of these results led to an estimated 0.6% of rural domestic wells containing one or more pesticides above the MCL.

Pesticide Use

In the United States about 80% of pesticide usage is in agriculture. The remainder is used by industry, homeowners, and gardeners. About 500 million pounds of herbicide, 180 million pounds of insecticide, and 70 million pounds of fungicide were applied for agricultural purposes in 1993.[3] Several maps of the United States delineate usage patterns of several pesticides.[4] The majority of the triazine and amide herbicides are applied to fields in the north central corn belt states of Michigan, Wisconsin, Minnesota, Nebraska, Iowa, Illinois, Indiana, and Ohio. Commonly used organophosphorus insecticides are more heavily applied to fields in California and along the southeastern seaboard than in the northern corn belt. Carbamate and thiocarbamate pesticides are heavily used in potato growing areas of northern

TABLE 1 U.S. Maximum Contaminant Levels for Drinking Water

Organic Chemical Name	MCL (mg/L)	Organic Chemical Name	MCL (mg/L)	Organic Chemical Name	MCL (mg/L)
2,4,5-TP (Silvex)	0.05	Chlordane	0.002	Heptachlor	0.0004
2,4-D	0.07	Dalapon	0.2	Heptachlor Epoxide	0.0002
Alachlor	0.002	Dinoseb	0.007	Lindane	0.0002
Aldicarb	0.007	Diquat	0.02	Methoxychlor	0.04
Aldicarb sulfone	0.007	Endothall	0.1	Oxamyl (Vydate)	0.2
Aldicarb sulfoxide	0.004	Endrin	0.002	Picloram	0.5
Atrazine	0.003	Ethylene dibromide	0.00005	Simazine	0.004
Carbofuran	0.04	Glyphosate	0.7	Toxaphene	0.003
Carbon tetrachloride	0.005				

Source: U.S. Environmental Protection Agency.[1]

Maine, Idaho, the Delmarva Peninsula, and vegetable fields of California and the southeastern coastal states. Fungicide use is concentrated in high humidity and irrigated areas of the coastal states and to some extent along the Great Lakes and Mississippi River Valley. The fumigants carbon tetrachloride and ethylene dibromide (EDB) were used heavily in the past at grain storage elevators throughout the Midwest and elsewhere in the United States.

Associated Pesticide Behavior in Soils and Water

Although pesticide use is a dominant factor in groundwater contamination, leaching variability among pesticides exhibiting similar behaviors is striking and explains why several heavily used pesticides seldom if ever are detected in groundwater. In general, pesticides within a class have similar chemical characteristics upon which soil leaching predictions can be made based on persistence, solubility, and mobility. Pesticide class relationships with soils and water transport described in the following text are detailed in Weber.[5] Individual frequencies of groundwater pesticide detection, in parenthesis next to commonly used products, are calculated from the Pesticide Groundwater Data Base (PGWDB)[4] and the National Water Quality Assessment (NAWQA) database.[6] High frequencies of detection identify those pesticides with a disposition to leach.

Insecticides

Chlorinated hydrocarbons are one of the oldest chemical classes of insecticides. Some of the best-known compounds include aldrin, dieldrin, DDE, DDT, endrin, and toxaphene. Although banned since the 1960s, their extremely persistent nature precludes their detection in very trace quantities in groundwater of the upper Midwest. On the other hand, heavily used organophosphates like malathion, methylparathion, disulfoton, and others have been extensively surveyed during several groundwater monitoring studies and have not been detected. The organophosphate insecticides, parathion (not reported (NR), <1), % occurrence from PGWDB, % occurrence from NAQWA data. terbufos (<1, <1), fonofos (<1, <1), and chlorpyrifos (<1, <1), which are heavily used on corn and sorghum, were also seldom detected. Diazinon (1.1, 1.3), the common garden insecticide, is occasionally detected in groundwater. Generally, the organic phosphates are rapid degraders and are strongly retained on soils.

For the most part, carbamates and thiocarbamates are very sparingly soluble and exhibit low to moderate soil retention; however, a small number have high solubility and low soil retention. Most carbamates are characterized as having short longevities. Generally, pesticides in this group having half-lives of 30 days or more have the potential to leach. The thiocarbamates butylate (<1, <1) and

EPTC (2.6, <1) are extensively used in agriculture and have relatively short half-lives. Aldicarb (<1, <1) and carbofuran (14.7, <1) are at the high end for solubility and longevity in their class. Their metabolites have been frequently detected beneath high use crops, such as potatoes in the potato growing regions of the United States.

The pyrethroid insecticides have low solubilities, short half-lives, and high soil retentions that make them unlikely to leach. Yet, permethrin (<1, <1) is occasionally detected in very trace quantities in groundwater.

Fungicides and Fumigants

Fungicides are non-volatile organometallic compounds with low aqueous solubility that inhibit growth of actinomycetes and many fungi. The best-known fungicides zineb (not detected (ND), NR) and captan are zinc-based, and maneb (ND, NR) is manganese-based. Some, like bordeaux, are copper sulfate-based. Although their detection frequency is very low, fungicides have not been analyzed in many surveys.

Fumigants are very volatile halogenated compounds that generally are knifed below the soil surface. These compounds have high aqueous solubility and very low soil retention. The fumigants EDB and 1,3-dichloropropene have been frequently detected in the subsurface and in groundwater in high-use regions, such as California.[7] Ethylene dibromide and carbon tetrachloride were also used in grain storage facilities during the 1950s and 1960s. Spills, leaks, and improper handling resulted in 400 reported groundwater contamination sites in Kansas and Nebraska.

Herbicides

There are at least eight major chemical classes of herbicides. These include: quaternary N, basic, acidic, carboxylic acid, hydroxy and aminosulfonyl, amide and anilide, dinitroaniline, and phenylurea herbicides.[5] Several herbicide classes have similar behaviors with respect to soil and water.

Both quaternary N and dinitroaniline herbicides are very highly retained by soils and are not expected to be detected in groundwater. However, paraquat, pendimethalin, and trifluralin have been reported several times in groundwater. Their presence indicates that transport is dependent on factors not directly related to compound longevity, solubility, and mobility. Vertical transport by preferential flow through macropores is a commonly accepted mechanism used to explain these detections. In some instances, compounds have been described as preferentially transported attached to colloidal material.

Carboxylic, hydroxy, and aminosulfonyl acids, and thiocarbamate herbicides have very low to low soil retention and very short to moderate longevity. Thus, the more heavily used and persistent pesticides in these groups are the ones most generally detected in groundwater. They include the acids, dicamba (2.0, NR), picloram (2.5, <1), bromacil (1.8, 1.0), and dinoseb (1.4, <1).

Phenylurea herbicides have low to high soil retentions and short to moderate longevity. Linuron (16.7, <1) and diuron (<1, 1.9) are the most frequently detected in groundwater and both have moderately long half-lives ranging from 60 to 90 days.

Amide and anilide herbicides have low soil retention and short to moderate longevity. Several amide herbicides and their transformation products have been detected in groundwater. The commonly used amides in the Midwestern corn belt, namely alachlor (1.7, 2.7), metolachlor (<1, 12), propachlor (1.2, <1), and acetochlor (NR, <1), are the most frequent offenders because they are relatively persistent.

As suggested by the name, basic herbicides behave as bases. The group contains several subclasses including aniline, formamidine, imidazole, pyrimidine, thiadiazole, triazines, and triazole. Basic herbicides have low to high soil retention and very short to moderate longevity. Again, it is generally the most persistent and heavily used pesticides that are more frequently found in groundwater. The most frequently detected compounds in the group are the triazines, namely atrazine (5.6, 30), metribuzin (4.2, 1.9), cyanazine (2.0, 1.4), simazine (2.0, 14.8), and prometon (2.1, 11.6).

Groundwater Contamination

It stands to reason that there are generally good associations between pesticide use and their detection in groundwater. Since groundwater flows very slowly at rates normally ranging from 0.1 ft/day to 3 ft/day, pesticide sources are generally very near the monitored well. Thus, high frequencies of triazine and acetamide detections are reported in the states of the northern corn belt. More fungicides and fumigants were detected in warm humid states of California and Florida where vegetable and fruit crops dominate the landscape. In an analysis of the 20 NAWQAs for pesticides, frequencies of pesticide detection in groundwater were significantly related to the estimated amount of agricultural use within a 1km radius of the sampled site.[6] They also emphasized that pesticides were detected beneath both agricultural (60.4%) and urban areas (48.5%). Discontinued used pesticides have been detected numerous times in shallow aquifers.

In general, families of pesticides have similar chemical characteristics from which predictions have been made as to the product's potential for contamination of groundwater; however, differences in the leaching behavior of pesticides exhibiting similar chemistry can be appreciable and is the reason several heavily used pesticides are seldom, if ever, detected in groundwater.

Management of Point Sources of Groundwater Contamination

Important steps are being taken to reduce water quality pollution by pesticides occurring from spills and back siphoning events (point sources). Since it is easier to resolve point than non-point sources, laws have been enacted to eliminate contamination of surface water bodies, which may be in hydraulic contact with groundwater, from used pesticide containers and rinseate from chemical wash downs. Check valves are mandatory when pesticides are mixed and/or diluted and prevent backflow to groundwater. Soils at and adjacent to agrichemical supply facilities have been surveyed in several states and found to be highly contaminated with pesticide residues. The herbicides, atrazine, alachlor, metolachlor, cyanazine, and metribuzin are the worst offenders from the standpoint of pesticide mass in the soils at sites in Wisconsin and Illinois.[4] Many of these sites and those in other states are now involved in soil cleanups, which are designed to protect underlying groundwater from further pollution.

Management of Non-Point Sources of Groundwater Contamination

Normal farm chemical applications of pesticides are generally considered potential non-point sources of groundwater contamination because they are dispersed over large areas ranging from fields to watersheds. Management strategies are in place to reduce leaching of field applied chemicals.[8] These strategies vary from regulatory restrictions to outright bans on application in areas deemed more vulnerable to leaching. Integrated pest management, fostered by the office of pesticide management at the USEPA, is designed to reduce chemical applications. The practice of banding applications has reduced amounts applied. Both target more efficient pesticides and genetically engineered plants sensitive only to specific herbicidal action have been and are being developed. These new pesticides and pesticide-plant combinations require less chemical than in the past, and the altered plants allow for pest control with more environmentally sensitive chemicals. The USEPA has announced a plan to reduce the mass of applied chemicals from commonly used triazines and amides that are frequently detected in groundwater.

Irrigation Management

Irrigation practices can influence pesticide leaching. Atrazine was vertically transported deeper and faster when using flood rather than sprinkler irrigation.[19] Sprinkler systems allow for much more uniform and efficient water management practices than furrow irrigation, and recent studies have shown that they reduce chemical leaching.[10] In the Nebraska's Platte Valley[11] and in the Walnut Creek watershed in Iowa,[12] peak herbicide concentrations were strongly related to rapid flushing beneath drainage areas where surface water ponds during heavy rainfall events on the cropped fields. Application of excess irrigation water also was reported to increase herbicide leaching.[9,11]

Future Research

More research is necessary to evaluate the health risks of transformation products from heavily used pesticides that are frequently detected in groundwater. Research needs to focus on precision application of pesticides to specific field problem areas as a potential mechanism to reduce chemical application.

There is a need to evaluate the environmental cost/benefit of safer product replacements used in conjunction with genetically altered crops. As new products are registered to replace more persistent and mobile pesticides, long-term fate studies, including the monitoring of the transformation product impact, on groundwater quality are necessary.

References

1. U.S. Environmental Protection Agency. *Drinking Water Standards and Health Advisories,* EPA 822-B-00–001; USEPA Office of Water: Washington, DC, Summer, 2000.
2. U.S. Environmental Protection Agency. *National Survey of Pesticides in Drinking Water Wells: Phase I Report,* EPA 570/9–90-015; USEPA Office of Pesticide and Toxic Substance: Washington, DC, November, 1990.
3. Aspelin, A.L. *Pesticide Industry Sales and Usage, 1992 and 1993 Estimates;* Economic Analysis Branch Report 733-K-92-001; USEPA Office of Pesticide Programs, Biological and Economic Analysis Division, 1994; 1–37.
4. Barbash, J.E.; Resek, E.A. *Pesticides in Ground Water. Distribution, Trends, and Governing Factors*; Ann Arbor Press, Inc.: Chelsea, MI, 1996; 1–588.
5. Weber, J. Properties and behavior of pesticides in soils. In *Mechanisms of Pesticide Movement into Ground Water,* Honeycutt, R.C., Schabacker, D.J., Eds.; Lewis Publishers: London, 1994; 15–41.
6. Kolpin, D.; Barbash, J.E.; Gilliom, R.J. Pesticides in ground water of the United States, 1992–1996. Ground Water **2000**, *38* (6), 858–863.
7. Troiano, J.; Weaver, D.; Marade, J.; Spurlock, F.; Pepple, M.; Nordmark, C.; Bartkowiak, D. Summary of well water sampling in California to detect pesticide residues resulting from nonpoint-source applications. J. Environ. Qual. **2001**, *30* (2), 448–459.
8. Guyot, C. Strategies to minimize the pollution of water by pesticides. In *Pesticides in Ground and Surface Water*, Bőrner, H., Ed.; Springer: Berlin, 1994; 87–148.
9. Troiano, J.; Garretson, C.; Krauter, C.; Brownwell, J.; Huston, J. Influence of amount and method of irrigation water application on leaching of atrazine. J. Environ. Qual. **1993**, *22*, 290–298.
10. Spalding, R.F.; Watts, D.G.; Schepers, J.S.; Burbach, M.E.; Exner, M.E.; Poreda, R.J.; Martin, G.E. Controlling nitrate leaching in irrigated agriculture. J. Environ. Qual. **2001**, *30* (4), 1184–1194.
11. Spalding, R.F.; Watts, D.G.; Snow, D.D.; Cassada, D.A.; Exner, M.E.; Schepers, J.S. Herbicide loading to shallow ground water beneath Nebraska's management systems evaluation area. J. Environ. Qual. **2003**, *32,* 84–91.
12. Moorman, T.B.; Jaynes, D.B.; Cambardella, C.A.; Hatfield, J.L.; Pfeiffer, R.L.; Morrow, A.J. Water quality in walnut creek watershed: herbicides in soils, subsurface drainage and groundwater. J. Environ. Qual. **1999**, *28* (1), 35–45.

6

Lakes and Reservoirs: Pollution

Subhankar
Karmakar and
O.M. Musthafa

Introduction

Surface water is one of the most important natural resources in the world. It has been explicitly established that water of good quality is a fundamental element to sustainable socioeconomic development. It is the habitat for a large number of species and is a crucial component for metabolic activities of plants and animals. Aquatic ecosystems are endangered on a worldwide scale by a multitude of pollutants as well as damaging land-use or water-management practices. Some problems have been present for a long time but have only recently reached a critical level, while the rest are recently emerging. Oxygen balance in the aquatic systems is severely affected by organic pollution, which often results in severe pathogenic contamination. Enrichment of aquatic systems with nutrients from various origins, predominantly domestic sewage, agricultural runoff, and agro-industrial effluents, results in enhanced eutrophication, of which lakes and reservoirs are affected the most.

Lakes and reservoirs are the major resources of fresh surface water. They are larger and deeper than ponds and are not part of the ocean. Lakes and reservoirs are major resources as these hold about 90% of the world's fresh surface water and are the key freshwater resources for agriculture, fisheries, domestic, industrial, recreational, landscape entertainment, and energy production. Natural lakes are bodies of water, created by volcanic, tectonic, or glacial activity, whereas reservoirs are artificial impoundments. A lake is a relatively large lentic freshwater or saltwater body, which is localized in a basin surrounded by land. Natural lakes are generally found in mountainous areas, rift zones, and areas with ongoing glaciations. In some parts of the world, there are many lakes formed due to chaotic drainage patterns left over from the last ice age. They are the habitats of a variety of flora and fauna, making them a source of fish and a destination for migratory birds to reproduce or rest. A reservoir, which is known as an artificial lake, is constructed for the benefit of man's water needs, sometimes for one particular purpose, but more recently for multiple purposes. Reservoirs are different from lakes in many ways. They have usually larger drainage basins than lakes, and many are located in watersheds with extensive agricultural activities.

Direct contamination of surface waters with metals in discharges from mining, smelting, and industrial manufacturing is a long-lasting phenomenon. The emission of airborne pollutants has now reached such proportions that long-range atmospheric transport causes contamination, not only in the vicinity of industrial regions but also in more remote areas. Similarly, precipitation of acid rain occurs, when moisture in the atmosphere combines with gases such as sulfur dioxide, which are produced when fossil fuels are burnt. This may cause significant acidification of surface waters, especially lakes and reservoirs. Contamination of water by synthetic organic micropollutants and emerging contaminants results either from direct discharge into the surface waters through runoff or after transport through the atmosphere.

This entry briefly presents classification of lakes and reservoirs based on the flow of water in and out of the system and their utility, respectively, followed by the problems associated with lakes and reservoirs causing deterioration of water quality. A discussion on monitoring of water quality and various protective and restorative measures is also made through a literature survey, which may be useful to future research aspirants and water resource professionals in identifying economically and environmentally sustainable lake and reservoir management strategy.

Classification of Lakes and Reservoirs

All lakes are temporary over geologic time scales, as they will slowly fill in with sediments or spill out of the basin.[1] Water enters into lakes from a variety of sources such as seepage through groundwater storage, runoff from watershed, direct precipitation into the lake, and other surface waters bodies (like streams or rivers). Water may drain out from lakes through deep percolation to join groundwater table or through surface water flow and evaporation. Natural lakes can be classified into four major types based on how water enters and exits the lake. Water may enter into the lake through one source or multiple sources. The water quality of a lake and its biodiversity are significantly influenced by the type of lake. Depending upon the way of entrance and exit of water, lakes can be classified into four categories as shown in Figure 1.

Seepage lakes do not have a distinct inlet or an outlet, and occasionally overflow. The major sources of water are direct precipitation, surface runoff from the immediate drainage area, and seepage through groundwater storage, as seepage lakes are landlocked water bodies. Since seepage lakes are sensitive to groundwater levels and local rainfall patterns, water levels may fluctuate seasonally. These lakes may have a less diverse fishery as the direct source of water is not a flowing water body or stream. Seepage lakes also have a smaller drainage area, which may help to account for lower nutrient levels. *Spring lakes* have no distinct inlet, but do have an outlet. The major source of water for spring lakes is groundwater flowing into the bottom of the lake from inside and outside the immediate surface drainage area. *Groundwater drained lakes* have no inlet, but similar to spring lakes, these may have an uninterruptedly flowing outlet. Drained lakes are not groundwater fed and their principal sources of water are precipitation and direct drainage from the surrounding land. The water levels in drained lakes fluctuate frequently depending on the supply of water. Under severe conditions, the outlets from drained lakes may become intermittent. *Drainage lakes* have both an inlet and an outlet where the main water source is stream drainage. These lakes support fish populations that are not necessarily identical to the streams connected to them. Drainage lakes mostly have higher nutrient levels than natural seepage or spring lakes.[2] Depending on the utility, reservoirs can be classified into four classes as shown in Figure 2.

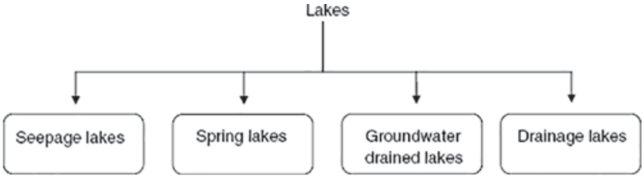


FIGURE 1 Classification of lakes.

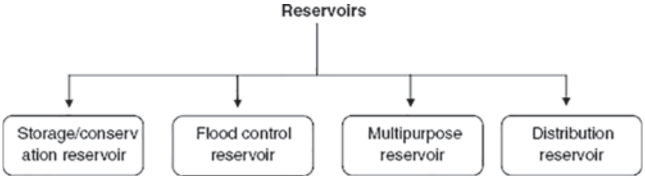


FIGURE 2 Classification of reservoirs.

Storage/conservation reservoirs retain excess water supplies during peak flows and release water gradually during low flows as and when needed. A *flood control reservoir* stores a portion of the flood flows so that it can minimize the flood peaks in the areas to be protected downstream. A *multipurpose reservoir* is meant for serving multiple purposes such as water supply, flood and soil erosion control, hydroelectric power generation, recreation, irrigation, etc. A *distribution reservoir* is connected with a network of primary water supply and is used to supply water to the end users according to fluctuations in demand over a short time period and serves as local storage in the case of emergency. Such reservoirs, thus, support the water treatment plants to work at a uniform rate and can store water when there is less demand and, thus, supply water during high-demand periods. Water quality in both lakes and reservoirs is influenced by many factors such as water body type, ecosystem characteristics, land use and land cover, and human activities.[3]

Freshwater/Saline Lakes

Most lakes hold freshwater, but some, particularly those where water cannot discharge via a river, can be salty. As a matter of fact, some lakes such as the Great Salt Lakes are saltier than the oceans. Lakes whose salinity content is more than 3 g/L are considered as saline lakes. They are prevalent and present on all regions, including Antarctica(e.g., Caspian Sea, Dead Sea, etc.). Though the inland saline water constitutes around 45% of the total inland water, a few deep lakes (mainly the Caspian Sea) occupy a significant volume of these saline waters. Salinity has a great influence on the freezing point of water, amount of dissolved oxygen, etc.[4]

Trophic Status

Lakes can be classified based on their trophic state as “eutrophic,” “mesotrophic,” and “oligotrophic.” The word “trophic” means nutrition or growth. A eutrophic lake is characterized by the presence of a high concentration of plant nutrients and associated excess plant growth. On the other hand, an oligotrophic lake is characterized by low nutrient concentrations and low plant growth. Mesotrophic lakes fall in between these two. The major factors that regulate the trophic status of a lake are as follows: 1) rate of nutrient supply; 2) climatic condition (sunlight, temperature, precipitation, lake basin turnover time, etc.), and 3) morphometry/shape of lake basin (mean and maximum depth, volume and surface area, watershed-to-lake surface area ratio, etc.).[5]

Problems Associated with Lakes and Reservoirs

Although more than three-fourths of the earth is occupied with water, only less than 0.3% (including surface water and groundwater) is available for human consumption.[6] The total area of the lakes on earth amounts to approximately 2.5×10^6 km^2 or 1.8% of the continental area, containing 1.2×10^5 km^3 of water. The 253 largest lakes (larger than 500 km^2) of the planet contain an estimated 78% of the world's unfrozen fresh surface water and thus represent an essential global life support system.[7] Though the amount of water in lakes and reservoirs is very small, because of its rapid renewal, these habitats are the primary sustainable supplier of freshwater for most regions. Since freshwater on the earth is scarce, sincere efforts are required for the conservation of existing water resources to ensure the availability of sufficiently good quality water. The surface water resources are more susceptible to pollution as compared to the groundwater resources due to the ease of access of pollutants and contaminants in the former.[8,9]

Most of the lakes and reservoirs around the world face environmental stress, and the appropriate functioning of various vital ecosystems is in danger. Due to the explosion in human population, supplementary demand-related damage is forced on lakes and reservoirs. Water levels become lower as a result of higher consumption by households and industries; a growing number of human population results in shrinking and altering their water resources; inappropriate use of the land, particularly hills and mountains, results in increased sedimentation on their basins. Ultimately, pollution from agricultural lands and from domestic and industrial sources may produce eutrophication, resulting in undesirable effects such as the presence of toxic algae, reduction of oxygen, and generation of foul odor. The proliferation of contaminants within lake and reservoir systems can deteriorate water quality significantly. In many regions, lake ecosystems have already degraded and restoration to desirable water quality needs enormous effort and high cost. Following the present status, none can predict for how long these resources can serve as renewable sources of pure surface water for domestic, industrial, and agricultural uses, or as sources of protein-rich food.[10] Lakes and reservoirs have a more vulnerable and complex ecosystem than rivers as they do not have a self purifying ability and, hence, readily accumulate pollutants. Because of their importance, their beauty, their religious and cultural significance, and their relative susceptibility to degradation, lakes and reservoirs require more concerted attention than is paid generally to river systems.[11] The next few paragraphs elaborate the major pollution problems of lakes and reservoirs.

Eutrophication

It can be defined as the process of enrichment of waters with plant nutrients that causes raw water quality loads, such as high primary production, low oxygen concentrations, and increased concentrations of hydrogen sulfide, carbon dioxide, dissolved iron, and manganese in the hypolimnion.[12] It is considered as one of the serious negative effects faced by the lentic water bodies such as lakes, ponds, and reservoirs and is one of the major water quality problems worldwide[13] and the most serious challenge for water management professionals in densely populated areas.[14] Water from eutrophic reservoirs may have meager taste, odor, and color, and some have high concentrations of naturally occurring organic compounds that may form trihalomethanes, which are carcinogenic and mutagenic and other by-products of disinfection.[15,16]

Seasonal pattern is prominently influenced by the availability of solar insolation and nutrients. Usually, the shallower the lake, the less the check of internal nutrient recycling by thermal stratification and light availability by critical depth as compared to mixing depth. During summer wind events, the lakes that are having an Osgood Index (OI)[17] less than 7 will show a strong tendency for mixing, as a result of which, nutrients from sediments/hypolimnion undergo recycling and enter into the photic zone. Such events entrain nutrients and cause summer blooms.[18] Thus, in highly dynamic lakes, the pattern of phytoplankton abundance/species composition differs with wind events.

Toxic Materials

It includes mainly heavy metals, toxin-producing microphytes, and pesticides from agricultural land. Toxic substances may enter water bodies directly as land runoff from urban streets and mining areas or as agricultural runoff including forestry drainage, discharge of inadequately treated sewage, and industrial effluents, and through deposition of airborne pollutants. Toxic substances can also be created in drinking water when chemicals from treatment plants interact with organic molecules in the raw water to form carcinogenic compounds such as trihalomethanes. A range of heavy metals such as mercury, arsenic, lead, and cadmium; chlorinated substances such as dichlorodiphenyltrichloroethane (DDT), dichlorodiphenyldichloroethylene (DDE), and polychlorinated biphenyls (PCBs); organic substances such as polyaromatic hydrocarbons (PAHs); Dieldrin; etc., create toxic conditions.

A large diversity of organic pollutants poses a serious threat to aquatic ecosystems. PCBs and 3,4-benzpyrene were confirmed as jeopardizing Lake Constance.[19] However, pesticide is the major group of contaminants in the top layers of sediments. The recognition of DDT as being harmful to human and animals served as the eye opener to serious consequences of various pesticides. PCBs are also considered as persistent organic pollutants and, thus, have huge potential hazards. The main sources of pesticides in lakes and reservoirs are 1) agriculture and forestry; 2) actions against aquatic weeds (e.g., water fern *Salvinia molesta, Eichhornia crassipes*); 3) actions against parasites and waterborne diseases (e.g., malaria and schistosomiasis); and 4) regulation of fish populations with rotenone (an active ingredient of derris, which is used as a fish poison for centuries). Atrazine, which is used for the protection of corn from weeds, is demonstrated as detrimental to human, and hence, it has recently been proscribed in many countries. Likewise, lindane, a persistent organochlorine insecticide, is toxic to fish in concentrations as low as 1 ppm. Pesticides are applied for curbing undesirable weeds such as water fern *Salvinia* sp. and *Azolla* sp., water hyacinth *E. crassipes,* water lettuce *Pistia stratiotes,* etc. Application of the pesticides should only be considered in cases where mechanical and biological approaches fail.[19]

Sedimentation

Sedimentation of lakes is a common phenomenon; in general, it takes place very slowly. Any process in a lake watershed that disturbs the soil can significantly hasten this process. Most of these activities are anthropogenic and comprised farming on fragile soils and on steep slopes, surface mining, and construction activities. Suspended sediment and sedimentation have many degrading effects on lakes and shoreline ecosystems. Peripheral wetlands can be completely covered by silt, eradicating their value as nutrient sinks, wave absorbers, nursery areas for fishers, and habitat. As silt settles in a lake, spawning areas are covered and lake volume is reduced, and this causes degradation in fishery production. As a result of reduced storage capacity, both the volume and extent of flooding can increase. In the case of hydroelectric amenities, generating capacity may reduce significantly. The increase in shoal areas as a result of sedimentation can enhance increased macrophyte growth and can interfere with recreational activities such as boating and fishing. Soil loss and suspended silt also contribute to eutrophication and lake contamination since the silt generally includes attached nutrients, herbicides, pesticides, and other chemicals. This increases water treatment costs and maintenance problems of water treatment plants.

Acidification

Atmospheric pollution by sulfur dioxide and nitrous oxides is the major cause of acidification in aquatic systems, through precipitation (dry or wet deposition) and, to a larger extent, through leaching from affected land. Other sources of acid deposits include industrial effluents and mining wastes.

Fish Depletion

Fish provides a substantial portion of animal protein consumed by humans. In tropical developing countries, 60% of the people depend on fish and 40% or more of total protein intake comes from fish.[20] The majority of the world's landed fish catch (87%) comes from marine areas. New developments in watersheds, such as dams and reservoirs to control the annual distribution of water, frequently cause large losses in floodplain fertility and species diversity. Floodplains are needed by many species of fish for reproduction and for refuge. The natural seasonal flooding and drying are signals used by some river-dwelling animals to begin to reproduce or migrate. Proper timing of water level drawdowns in upstream reservoirs is therefore important, especially to enhance spawning of certain species.

Stratification

Stratification is a significant feature influencing water quality in reasonably stagnant, deep waters, such as lakes and reservoirs, which occurs mainly because of the difference in temperature, leading to a variation in density. Occasionally, it can be due to the difference in solute concentrations. Water quality in various layers of stratified water body is subjected to different influences. Solar insolation will be more in the upper layer while the lower layer is physically detached from the atmosphere and may be in touch with decaying sediments that exert an oxygen demand. Because of these varying influences, the lower layer will usually have a reduced oxygen concentration relative to the upper layer. The anoxic condition thus produced will enhance the diffusion of constituents from sediments and form various compounds such as ammonia, nitrate, phosphate, sulfide, silicate, iron, and manganese.

During summer and spring, the surface layer of the water body in temperate regions becomes warmer and hence less dense. A resistance to vertical mixing is formed, because of the existence of warm water over cold water. The top warmer surface layer is known as the epilimnion and the colder water stuck underneath is the hypolimnion. There is a shallow zone called metalimnion or the thermocline, in between epilimnion and hypolimnion, where the temperature changes from warmer epilimnion to colder hypolimnion. Normally, wind-and surface-current-induced mixing is limited to epilimnion, while hypolimnion remains stagnant. The density difference between two layers (viz., epilimnion and hypolimnion) is diminished gradually when the weather becomes cooler. This will enhance the wind-induced vertical mixing between these two layers, which will result in the phenomenon known as "overturn," which can occur quite rapidly. The frequency of overturn and mixing governs predominantly on climate (temperature, solar insolation, and wind) and the characteristics of the lake and its surroundings (depth and exposure to wind). Lakes may be classified according to the frequency of overturn as follows:

- Monomictic: once a year—temperate lakes that do not freeze
- Dimictic: twice a year—temperate lakes that do freeze.
- Polymictic: several times a year—shallow, temperate, or tropical lakes
- Amictic: no mixing—arctic or high-altitude lakes with permanent ice cover, and underground lakes
- Oligomictic: Poor mixing—deep tropical lakes
- Meromictic: incomplete mixing—mainly oligomictic lakes but sometimes deep monomictic and dimictic lakes Because of the action of crosswind and the flow of water, thermal stratification is not seen in lakes, which have depths less than 10 m. In shallow tropical lakes, complete mixing occurs several times a year, whereas in very deep lakes, stratification may continue all over the year, even in tropical and equatorial regions. This stable stratification results in "meromixis."

In the case of tropical lakes, as a result of moderately constant solar insolation, seasonal changes in water temperature are small. The annual water temperature range is only 2–3°C at the surface and even less at depths greater than 30 m.[21] Winds and precipitation, both play a vital role in mixing. Because of the large difference in rainfall between wet and dry seasons, large variation in water level is seen in some tropical lakes. Such variations have a prominent influence on dilution and nutrient supply, which, in turn, affect algal blooms, zooplankton reproduction, and fish spawning. Wind speeds are usually greater during the dry season and evaporation rates are at their highest. The subsequent heat losses, combined with the turbulence caused by wind action, stimulate the process of mixing.

As far as recreation is concerned, reservoirs are as important as natural lakes but have surplus scopes for flood control, hydropower generation, and water supply. Although both lakes and reservoirs are subjected to silt, organic, and nutrient loadings, reservoirs usually having hefty watersheds and peculiar morphometric conformations are subjected to more water quality problems. Even though lakes and reservoirs have biotic and abiotic processes in common and similar habitats, they have some significant differences. They differ in their geologic history and setting, basin morphology, and hydrologic factors.[22,23]

Non-point source of water pollution generated by expanding agricultural production is considered as a major environmental threat to some lakes. Many chemicals in common agricultural use have a strong affinity for fine soil particles. When the latter erode, these chemicals are carried with them into surface waters. The soil itself is a problem when it accumulates in great quantity in lakes. Lake Pittsfield lost nearly a quarter of its volume to sedimentation in only 24 years. Transport and deposition of eroded materials as well as substances dissolved in runoff and attached to soil particles lead to negative impacts on agricultural land and the Three Gorges Reservoir including water quality decline,[24,25] which are generally thought to be caused by land use changes of converting forest resources to agriculture in watersheds. Conversion of cropland with a slope greater than 10° into forestland meets the reduction goal.[26]

Over recent decades, the water quality of lakes and reservoirs has been deteriorating rapidly due to external and internal pollution including that from the sediment, and the eutrophication phenomenon has become a more serious global threat. Some researchers suggest applying a plan of sediment dredging to this water body. After dredging, however, a vast amount of sediment would become solid pollutants containing high concentrations of heavy metals (mainly Hg and Cd), which would certainly cause secondary pollution. This pollution should not be ignored, and a further research is therefore needed on the elaborate restoration scheme for these precious drinking-water sources.[27] Major pollution issues of lakes and reservoirs with few representative past studies are tabulated in Table 1.

TABLE 1 Major Pollution Issues of Lakes and Reservoirs

Sl. No.	Pollution Issue	Effects	Representative Case Studies
1	Eutrophication	High primary production, low oxygen concentrations and increased concentrations of hydrogen sulfide, carbon dioxide, dissolved iron, and manganese in the hypolimnion	West Twin Lake, Ohio[28]
2	Sedimentation	Volume will be decreased, macrophytes growth may be enhanced	Lake Superior at Superior Harbor, Wisconsin[29]
3	Acidification	Adverse effects on the most sensitive aquatic species	Many Scandinavian lakes[11]
4	Toxic substances and heavy metal contamination	Bioaccumulation of these toxicants poses health hazards to all members of the food chain including humans	Lake Nainital, India[30]
5	Fish depletion	As the fishery plays a vital role in the supply of animal protein, its depletion affects the food security and economy of a significant portion of humans	Lake Victoria, Uganda[11]

Sources of Lakes and Reservoir Pollution

Lakes and reservoirs tend to collect not only sediments but also most of the pollutants that are washed into them, and thus they function, in part, as environmental sinks. Eroded soil dissolves in the water and fills in lake bottoms—this activity has significantly degraded lake ecosystems across the world. The sources of pollution can be mainly classified into two: 1) point sources and 2) non-point sources. A point source of pollution is a single identifiable localized source of air, water, thermal, noise, or light pollution.[31] In earlier days, control of "point source" nutrients and toxic contaminants was the principal focus of exertions for the protection and restoration of lakes and reservoirs, but nowadays, the substantial contaminant and nutrient sources to lakes and reservoirs are "non-point" type such as agricultural runoff, erosion from urban or deforested areas, surface mining, or atmospheric depositions. Most lake water is rich in nutrients that support growth of many aquatic macrophytes and algal blooms. Besides, water is contaminated with metals like chromium (Cr), copper (Cu), iron (Fe), manganese (Mn), nickel (Ni), lead (Pb), and zinc (Zn). High concentrations of these metals are also found in sediments, but it is found that the level of metal concentrations of lake varies considerably in different seasons.

In earlier days, we believed that lakes and reservoirs were enriched with nutrients and organic matter from "point" sources such as industrial discharges and wastewater treatment plant outfalls. However, for many lakes, non-point or diffuse nutrient loading, both internal and external, is found to be momentous. These non-point sources are challenging to assess and regulate,[32] and water quality in many lakes has remained deteriorated following diversion or treatment of point sources.

The main pollutants that enter through non-point sources are various forms of phosphorus and nitrogen such as total phosphorus, phosphate phosphorus, nitrate nitrogen, nitrite nitrogen, ammoniacal nitrogen, organic nitrogen, chlorine, sodium, calcium, and suspended solids. The exports of all constituents occur mostly during rainfall-or snowmelt-generated runoff events during the spring runoff period. It has been shown that one of the major reasons for the enrichment of lakes is the conversion of previously cropped land into agricultural production while conversion to forests has very little impact on enrichment.

The runoff of nutrients from old fields depends on the nutrient status of the soils and soil water. This nutrient status reflects the soil type, fertilizer and cropping practices prior to abandonment, number of years since abandonment, and the succeeding vegetation present at any particular point. Soil data identify the reservoir of nutrients available for runoff, provide a means to relate runoff to nutrient content of that particular soil, and provide the data necessary for design of management schemes to prevent release of nutrients. Farm lands that have been abandoned for 15 to 20 years are not major non-point sources of pollution.

Water Quality Monitoring

Water quality monitoring refers to the acquisition of quantitative and representative information on the physical, chemical, and biological characteristics of a water body over time and space.[33] It is a complex task, comprising all the activities to extract information with respect to the aquatic system. A variety of contaminants, in addition to a multitude of imprudent water quality management practices and destructive land uses, are currently threatening aquatic systems on a worldwide scale. In addition, it has been shown that water of good quality is a critical component for sustainable socioeconomic development.[34] The impact and behavior of contaminants in an aquatic ecosystem are complex and may involve adsorption-desorption, precipitation-solubilization, filtration, biological uptake, excretion, and sedimentation-suspension. Besides natural processes affecting water quality, there are also anthropogenic impacts, such as man-induced point and non-point sources, xenobiotic, and alteration of water quality due to unwise water use and river engineering projects (e.g., irrigation, damming, etc.).[35]

The degradation of water resources has increased the need for determining the ambient status of water quality, in order to provide an indication of changes induced by anthropogenic activities. To understand the process dynamics of a watershed, a well-designed water quality monitoring network identifies water quality problems while establishing baseline values for short-and long-term trend analysis. The need to evaluate observed water quality conditions and their suitability for the intended uses reflects a need for cost-effective and logistically practical water quality monitoring network design methods.

Types of Monitoring

Lake Sampling

It characterizes the water quality of the lake to identify status and trends. Two main types of lake monitoring are water-column sampling and near-shore (shallow water) sampling. Water-column sampling tries to quantify the overall response of the lake to contamination. The lake's trophic status, as designated by phosphorus, turbidity, chlorophyll-a, and dissolved oxygen, is of particular concern. Some invasive species, such as spiny water flea, can also be identified by water-column sampling. During seasons when the lake is stratified, the water-column sample should be sampled in both the epilimnion (the warm upper layer) and hypolimnion (lower layer of cold water).

Near-shore refers to the depth at which rooted plants can grow. Sampling can be adequately done at the end of a dock. Sampling for pathogens and pathogen indicators is important because of contact recreation such as swimming. Near-shore monitoring allows study of the lake bottom including sediment sampling for heavy metals, macroinvertebrates, and attached or rooted invasives such as zebra mussels and Eurasian watermilfoil.

Tributary Mass Load Sampling

It determines the tributary mass loads of water contaminants entering the lake. A significant portion of the lake's water pollution is brought by the tributaries flowing into it. Determination of tributary mass loads is particularly important for management of the lake's phosphorus and sediment problems.

Tributary Water Quality Sampling

It characterizes the water quality of tributaries to identify status and trends. Tributaries may be threatened by contaminants or stresses that affect the stream health but are not significantly detrimental to the lake. The tributaries are valued for recreation and aesthetics, drinking water, irrigation, and wildlife habitat and deserve protection.

Biological Integrity Sampling

It characterizes the long-term ecological health of the lake and tributaries. Ecological sampling is useful for detecting the effects of impairments that are not present at the time of sampling, for evaluating habitat health and for determining the biological integrity of surface waters. Ecological sampling may include bioassessments of fish and benthic macroinvertebrate communities, periphyton, and single-species monitoring (trout, salmon, and freshwater mussels are often used). Biological indices, a composite of different indicators, can be developed.

Citizen Monitoring

It encourages citizen participation in the measurement of watershed quality. To the extent that people care about the watershed's lands and waters, the watershed will be protected and enhanced for generations to come. One way to encourage such stewardship is through involvement of students and other citizens in water quality monitoring. Monitoring conducted by citizen volunteers increases public awareness and knowledge about water quality and its protection.

The design of efficient water quality monitoring network is essential for effective water management. To date, many water quality monitoring networks for surface freshwaters have been rather arbitrarily designed without a consistent or logical design strategy. Moreover, design practices in recent years indicate a need for cost-effective and logistically adaptable network design approaches.[36] Furthermore, the monitoring in a water quality management program is recognized as a statistical approach, so that both the assessment and the design problems can be addressed via a statistical method. In this view, the statistical methods have been found very efficient for redesigning and assessment of water quality monitoring networks (WQM).

The International Organization for Standardization (ISO) defines water quality monitoring as: "the programmed process of sampling, measurement and subsequent recording or signaling, or both, of various water characteristics, often with the aim of assessing conformity to specified objectives." This general definition can be differentiated into three types of monitoring activities that distinguish between long-term, short-term, and continuous monitoring programs:

- Long-term observation and standardized measurement of the aquatic environment for defining the water quality status and prediction of the trend are known as monitoring.
- Intensive programs for the measurement and observation of status of the aquatic environment for a specific purpose are known as surveys. These are of limited duration.
- Continuous measurement and observation of the aquatic environment for the management of the quality of the water and other operational activities are known as surveillance.

The constituents that decide the quality of the water are transported by water from the watershed. Therefore, we have to construct a perfect water budget, as it plays a key role in identifying a lake's (reservoir's) problem. By conducting a reconnaissance survey of water from the watershed, main tributaries can be selected. Since high flows are the chief segment of the water budget and huge volume influxes are followed by high concentration, continuous gauge recording is recommended for the determination of flow in major tributaries. From a successive record of inflow and outflow in the main tributaries, an annual water budget is constructed so that estimated inflows equal outflows with a correction for lake storage.

The water budget is formulated as:[37]

$$\mathrm{SF_i + GW + DP + WW = SF_o + EVP + EXF + WS + \Delta STOR}$$

where $\mathrm{SF_i}$ and $\mathrm{SF_o}$ are stream flow in and out, respectively; GW is groundwater in (includes deep and subsurface seepage); DP is direct precipitation on the lake surface; WW is wastewater, if any; EVP is evaporation; EXF is exfiltration; WS is removal for water supply, if any; and ASTOR is the change in lake volume.

Protective and Restorative Measures

Removal or treatment of direct input of wastewater, stormwater, or both constitutes the primary step in the restoration of water quality of eutrophic lakes and reservoirs, as these sources frequently contain comparatively high concentrations of phosphorus and nitrogen. For the realization of any long-term benefits from in-lake treatments, such external loadings should be reduced. In some cases, reduction of external loading is adequate to restore the water body (e.g., Lake Washington[38,39]), but in others, where internal loading of nutrients is significant, in-lake treatments may be indispensable to accomplish lake quality improvement (e.g., Lake Trummen[40]).

Advanced wastewater treatment (AWT) and diversions are two most commonly used techniques for the reduction of external inputs. In diversion, the treated sewage or industrial wastewater is carried away from the degraded water body to waters that are having high assimilative capacity, by the installation of interceptor lines. AWT involves the reduction of phosphorus concentration in wastewater effluent by using chemicals such as alum (aluminum sulfate), lime (calcium hydroxide), or ferric chloride. Stormwater runoff is the next dominant source of external enrichment. Even though the P concentration in stormwater is very low (2%–10%), and solubility is less than that of sewage effluent, such non-point sources can denote momentous contributions. P retention in wet detention basins and wetlands, rapid filtration through soil, and P removal in predetention basins are the principal P removal methods applied for runoff water. If internal loading of P is anticipated to hinder the recovery following primary treatments such as diversion or AWT, then supplementary in-lake treatments may be justified to accelerate reclamation. Lakes and reservoirs lose volume due to siltation. Sediment removal, together with land management and construction of device to trap silt, is an example of their restoration and protection. Some management approaches are institutional arrangement, formation and operation of lake association, sports fishery management, etc.

Eutrophication can be controlled if the phosphorus (P) concentrations in the water body are lowered to a level that will limit the growth of the algae. This can be achieved by diversion of external input, dilution, flushing, or a combination of these approaches. Where there is substantial loading reduction, comparatively augmented rate of algal flushing, and negligible recycling from sediment, in-lake phosphorus concentration can be reduced significantly and trophic state can be improved rapidly. Lake Washington is a good example for this approach.[40,41] However, for many lakes, internal phosphorus release sustained the lake's enriched trophic state and reinforced the state of continued eutrophication, in spite of the removal of a noteworthy fraction of external loading by diversion.[42,43] In those lakes, following nutrient diversion, supplementary in-lake treatment may be necessary, to avoid an extended eutrophic state.

Other methods are dilution and flushing, which are used interchangeably. Dilution involves the reduction in the concentration of the nutrients and a washout of algal cells, whereas flushing comprises only the latter. Where there is a high nutrient load, water should be diverted if possible for the low dilution rate to be most effective. This plan provides for a reduction in biomass primarily through nutrient limitation. If only moderate to high nutrient water is available, flushing may work well if the loss rate of cells is sufficiently great relative to the growth rate. Flushing rate on the order of 10%–15% per day will afford some control through washout. We can opt for the technique such as "phosphorus inactivation," which involves, usually, the application of salts of aluminum, such as aluminum sulfate (alum), sodium aluminate, etc., to precipitate phosphate as aluminum phosphate and thus to bind a significant fraction of P. Aluminum hydroxide floc, which is formed during this process, enhances the settling of the precipitate and it will continue to sorb and retain P in their molecular lattices.

Nutrient reduction mainly consists of processes such as source reduction in the watershed, issuance of fertilizer guidelines, setting up of shoreland buffer strips, and restriction of motorboats. *Biological control* involves processes such as use of bacteria for algae control, use of algae-eating fishes, biomanipulation, aquascaping, and bioscaping. *Lake aeration* comprises conventional aeration, solar powered aeration, wind-powered aeration, fountain aeration, and hypolimnetic aeration. *Addition of chemicals* includes use of barley straw, alum, buffered alum for sediment treatments, calcium compounds, liquid dyes, chlorine, and algicide[44] (Figure 3). Copper is an efficient algicide. Copper sulfate application, the typical treatment for algal problems for many decades, is often effective for short-term solution to a current algae problem, predominantly in water supply reservoirs. However, there is substantial

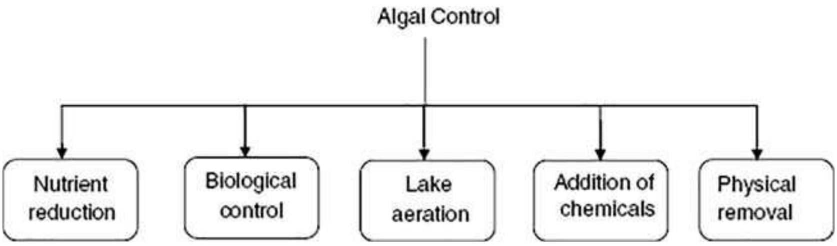


FIGURE 3 Major algal control measures.

TABLE 2 Major Restoration Measures for the Lakes and Reservoirs

Sl. No.	Restoration Measure	Advantage(s)	Drawback(s)
1	Diversion	Reduce the overall nutrient loading to the system	Presence of another water body (having high assimilative capacity) is required in the near vicinity
2	Advanced wastewater treatment	Control eutrophication by limiting nutrient concentration with suitable chemicals	Accumulation of chemicals in the system and high cost
3	Dilution and flushing	Control eutrophication by limiting the P concentration	Cannot accommodate high nutrient loading
4	Application of appropriate algicide (such as copper sulfate)	Control eutrophication by curbing algal growth	Its effects are found to be temporary, higher cost, there are major negative impacts to non-target organisms, and significant copper contamination of sediments can be possible
5	Sediment removal (dredging)	Sediment removal, together with land management can restore the lake, affected by siltation	High cost and, in some cases, may cause environmental damage

confirmation against the prolonged usage of this chemical. Its effects are found to be temporary, cost can be higher, there are major negative impacts to non-target organisms, and significant copper contamination of sediments can be possible. We can go for long-term and permanent options such as control of external and internal nutrient loading for the effective management of algal bloom. For a better water quality, and if the P is not estimated to reach an algal limiting level, an in-lake treatment for the control sediment should be established soon after external controls are in place.

Macrophytes includes all macroscopic aquatic flora, comprising macroalgae such as the stoneworts *Chara* and *Nitella*, aquatic liverworts, mosses, ferns, and flowering vascular plants. Aquatic plant management aims to curb annoyance species, to exploit the favorable features of plants in water bodies, and to reorganize plant communities. Its principal objective is the establishment of stable, diverse, aquatic plant communities comprising high percentages of desirable species. An exhaustive understanding of macrophytes biology is the foundation for evolving innovative management tactics. Continued research and development will advance our understanding of the relationship of aquatic plants to overall lake and reservoir quality and our capability to manage aquatic plant communities to preserve or improve that quality.[36,37] Table 2 shows major restoration measures for the lakes and reservoirs.

Eutrophication Model Framework

Phosphorus-loading models are often employed for the evaluation of eutrophication problems in lakes and reservoirs. In these models, phosphorus loading is linked to the average total phosphorus concentration in the lake water and to other indicators of water quality that are associated to algal bloom, such as chlorophyll and transparency. Physical and hydrologic features influence the response of the lake to phosphorus loading and thus these models take in to account various characteristics such as lake volume, average depth, flushing rate, etc. The underlying principles behind the eutrophication model are as follows: 1) phosphorus acts as the limiting nutrient for algal growth; 2) any change in the amount of phosphorus discharged into a lake over an annual or seasonal period will alter the average concentration of P in lakes and hence the extent of algal bloom; and 3) the capacity of the lake to adjust with the P loading, without causing algal blooms increases with the volume, depth, and the flushing rate of the lake. Models recapitulate these relationships in mathematical forms, based upon observed water quality responses of large numbers of lakes and reservoirs. Eutrophication models are pitched for the prediction of average status of water quality over a season or a year. Averaging is mainly done over three dimensions: 1) depth; 2) sampling stations; and 1) season.[45]

Conventions for the Protection of Lakes and Reservoirs

Lakes and Wetlands—Ramsar Convention, Iran, February 2, 1971

The Ramsar Convention (The Convention on Wetlands of International Importance) is an international treaty for the conservation and sustainable utilization of wetlands. The convention was developed and implemented by participating countries at a meeting in Ramsar, hosted by the Iranian Department of Environment, and came into force on December 21, 1975. The Ramsar List of Wetlands of International Importance presently comprises 1950 sites (known as Ramsar Sites) covering around 1,900,000 km^2 up from 1021 sites in 2000. The nation with the highest number of sites is the United Kingdom at 168; the nation with the greatest area of listed wetlands is Canada, with more than 130,000 km^2. Presently, there are 161 contracting parties, up from 119 in 1999 and from 21 initial signatory nations in 1971. Signatories meet every 3 years as the Conference of the Contracting Parties (COP); the first was held in Cagliari, Italy, in 1981. There is a standing committee, a scientific review panel, and a secretariat. The headquarters is located in Gland, Switzerland, shared with the International Union for Conservation of Nature (IUCN).[46]

UNECE Water Convention, Helsinki, March 17, 1992

The Convention on the Protection and Use of Transboundary Watercourses and International Lakes (Water Convention) is intended to strengthen national measures for the protection and ecologically sound management of transboundary surface waters and groundwater. The Parties to this convention are obliged to prevent, control, and reduce transboundary impact; use transboundary waters in a reasonable and equitable way; and ensure their sustainable management. Parties bordering the same transboundary waters shall cooperate by entering into specific agreements and establishing joint bodies. The Convention includes provisions on monitoring, research and development, consultations, warning and alarm systems, mutual assistance, and exchange of information, as well as access to information by the public.

Protocol on Water and Health, London, June 17, 1999

The Protocol on Water and Health aims to protect human health and well-being by better water management, including the protection of water ecosystems, and by preventing, controlling, and reducing water-related diseases. It is the first international agreement of its kind adopted specifically to attain an adequate supply of safe drinking water and adequate sanitation for everyone and effectively protect water used as a source of drinking water. Parties to the Protocol commit to set targets in relation to the entire water cycle.[47]

Conclusion

Many of the lakes and reservoirs are situated in developing countries or industrializing countries that only now identify the devastating economic and social impact that pollution, overfishing, and habitat degradation are having on their water resources. Once these natural systems are severely polluted, they often cannot be restored completely; they can at best only be improved to a level where they can meet basic functions, and society must bear the increased costs and risks to human health.

The reviewed case studies in this entry and elsewhere illustrate that contaminated lakes and reservoirs can be restored, at least partially. The entry presented a number of examples where this process has been undertaken effectively. From an economic and environmental point of view, it appears wise to prevent the occurrence or exacerbation of problems, rather than going for radical and expensive restorative actions. Ample evaluations of water resources should be conducted to understand the state of lakes and reservoirs, so that we can provide the essential means for the establishment of imperative water resource management goals and objectives.

All-inclusive watershed management programs that comprise significant water quality objectives precise to the hydraulic characteristics of lakes and reservoirs should be prepared. The differences between temperate and tropical lake situations need to be highlighted. During the formulation of management strategies, all exertion should be made to avert the discharge of toxic substances to lakes and reservoirs. Introduction of exotic and invasive species should be banned, unless there has been ample environmental assessment.

We should incorporate the "precautionary" approach and the "polluter pays" principle in the management of lakes and reservoirs, which can benefit from various economic tools and ample financial policies. Primary action areas for protection and restoration of lakes and reservoirs involve watershed assessment, watershed control measures, and best management practices, which comprise management of pollution from agriculture, silviculture, mining, industrial pollution, urban runoff, and other pollutant sources, predominantly for nutrients and persistent toxic pollutants. Management programs for lakes and reservoirs should be comprehensive in scope and watershed-wide in nature. If necessary, joint governance institutions, such as joint commissions, composed of high-level government officials, should be established to improve management of transboundary lake and reservoir basins. The formulation of legal instruments, treaties, and a hierarchy for making the decision are essential for the advancement of the water resources management. It is clear that the world's lake reservoirs are threatened. In many cases, lake degradation is so advanced that populations depending on them are in great danger. The scope, magnitude, and dimension of the problem demand an international cooperation as well as national exertions.

A logical and consistent design methodology that allows more efficient and effective data collection and, hence, more useful information extraction should be developed. Such an approach not only permits better water pollution control recommendations and better allocation of financial resources but also, ultimately, a better understanding of the ecosystems. To appreciate the challenges of designing water quality networks, it is crucial to clearly define objectives and identify statistically acceptable assumptions. Assumptions are an inherent part of the monitoring network design process, mainly due to the stochastic influences on water quality variables in the aquatic environment. The number and type of simplifying assumptions made or allowed are dependent upon network objectives. Furthermore, assumptions in monitoring network design should be made relative to water quality hydrologic principles, applicable statistics, information utilization, and budget constraints.

References

1. Bindler, R.; Renberg, I.; Brännvall, M.-L.; Emteryd, O.; El-Daoushy F. A whole-basin study of sediment accumulation using stable lead isotopes and flyash particles in an acidified lake, Sweden. Limnol. Oceanogr. **2001**, *46*, 178–188.
2. Available at http://www.wisconsinlakes.org/index.php/the-science-of-lakes/21-lake-types (accessed May 2012).
3. Available at http://www.unep.or.jp/ietc/publications/short_series/lakereservoirs-1/fwd.asp (accessed February 2011).
4. Available at http://ga.water.usgs.gov/edu/earthlakes.html (accessed February 2012).
5. Available at http://www.waterontheweb.org/under/lakeecology/16_trophicstatus.html (accessed February 2012).
6. Dodds, W.K. *Freshwater Ecology—Concepts and Environmental Applications, 1st Ed.;* Elsevier: New Delhi, India, *2006*.
7. Tilzer, M.M.; Bossard, P. Large lakes and their sustainable development. J. Great Lakes Res. **1992**, *18* (3), 508–517.
8. Papatheodorou, G.; Demopoulou, G.; Lambrakis, N. A long-term study of temporal hydrochemical data in a shallow lake using multivariate statistical techniques. Ecological Modell. **2005**, *193*, 759–776.

9. Zhao, Y.; Yang, Z.; Li, Y. Investigation of water pollution in Baiyangdian Lake, China. Procedia Environ. Sci. **2010**, 2, 737–748.
10. Yu, F.; Fang, G.; Ru, X. Eutrophication, health risk assessment and spatial analysis of water quality in Gucheng Lake, China. Environ. Earth Sci. **2010**, *59* (8), 1741–1748.
11. Dinar, A.; Seidl, P.; Olem, H.; Jordan, V.; Duda, A.; Johnson, R. *Restoring and Protecting the World's Lakes and Reservoirs*; World Bank Technical Paper Number 289, The World Bank: Washington, DC, 1995; 1–113.
12. Richards, R.P. The Lake Erie agricultural systems for environmental quality project: An introduction. J. Environ. Qual., **2002**, *31,* 6–16.
13. Zemenchik, R.A. Bio-available phosphorus in runoff from alfalfa, smooth bromegrass, and alfalfa-smooth bromegrass. J. Environ. Qual. **2002**, *31*, 280–286.
14. Ouyang, W.; Hao, F.H.; Wang, X.L.; Cheng H G. Nonpoint source pollution responses simulation for conversion cropland to forest in mountains by SWAT in China. Environ. Manage. **2008**, *41,* 79–89.
15. Cook, G.D.; Carlson, R.E. *Reservoir Management for Water Quality and THMPrecursor Control;* American Water Works Association Research Foundation: Denver, CO, 1989.
16. Cook, G.D.; Kennedy, R.H. Managing drinking water supplies. Lakes Reservoirs Manage. **2001**, *17,* 157–174.
17. Osgood, R.A. Lake mixes and internal phosphorous dynamics. Arch. Hydrobiol. **1988**, *113,* 629–638.
18. Larsen, D.P.; Schultz, D.W.; Malueg, K.W. Summer internal phosphorous supplies in Shagava Lake, Minnesota. Limnol. Oceanogr. **1981**, *26,* 740–753.
19. Jørgensen, S.E. *Lake and Reservoir Management,* Revised Ed.; Developments in Water Science; Elsevier, 2005; Vol. 54.
20. WRI (World Resources Institute). World Resources 1992–93. Washington, DC, 1994.
21. Bartram, J; Balance, R. *Water Quality Monitoring—A Practical Guide to the Design and Implementation of Freshwater Quality Studies and Monitoring Programmes;* Published on behalf of United Nations Environment Programme and the World Health Organization UNEP/WHO ISBN 0 419 22320 7 (Hbk) 0 419 21730 4 (Pbk), 1996.
22. Kennedy, R.H.; Thornton, K.W.; Ford, D.E. Characterization of the reservoir ecosystem. In *Microbial Processes in Reservoirs*; Gunnison, D., Ed.; Junk Publishers: The Hague, Netherlands, 1985; 27–38.
23. Kennedy, R.H. Consideration for establishing nutrient criteria for reservoirs. Lake Reservoir Manage. **2001**, *17,* 175–187.
24. Kurtz, D.A. *Long Range Transport of Pesticides;* Kurtz, Ed.; Lewis Publisher: Chelsea, MI, 1990; 440 pp.
25. Cotham, W.E; Bidleman, T.F. Estimating the atmospheric deposition of organochlorine contaminants to the Arctic. Chemosphere **1991**, 22 (1–2), 165–188.
26. Wolf, M.S; Toniolo, P.G. Environmental organochlorine exposure as potential etiologic factor in breast cancer. Environ. Health Perspect. **1995**, *103,* 141–145.
27. Wania, F.; Mackay, D. Global fractionation and cold condensation of low volatility of organochlorine compounds in Polar Regions. Ambio **1994**, 22, 10–18.
28. Cook, G.D.; Heath, R.T.; Kennedy, R.H.; MComas, M.R. The Effect of Sewage Diversion and Aluminium Sulfate Application on Two Eutrophic Lakes. USEPA-600/3–78-033, 1978.
29. Kostic, S.; Parker, G. Progradational sand-mud deltas in lakes and reservoirs. Part 1. Theory and numerical modeling. J. Hydraulic Res. **2003**, *41* (2), 127–140.
30. Ali, M.B.; Tripathi, R.D.; Rai, U.N.; Pal, A.; Singh, S.P. Physico-chemical characteristics and pollution level of lake Nainital (U.P., India): Role of macrophytes and phytoplankton in biomonitoring and peytoremediation of toxic metal ions. Chemosphere, **1999**, *39* (12), 2171–2182.
31. Liu, Y.; Islam, M.A.; Gao, J. Quantification of shallow water quality parameters by means of remote sensing. Prog. Phys. Geogr. **2003**, 27(1), 24–43.

32. Line, D.E.; Jennings, G.D.; McLaughlin, R.A.; Osmond, D.L.; Harman, W.A.; Lombardo, L.A.; Tweedy, K.L.; Spooner, J. Nonpoint sources. Water Environ. Res. **1999**, 71, 1054–1069.
33. Sanders, T.G.; Ward, R.C.; Loftis, J.C.; Steele, T.D.; Adrian, D.D.; Yevjevich, V. *Design of Networks for Monitoring Water Quality;* Water Resources Publications LLC: Highlands Ranch, CO, 1983.
34. Bartram, J.; Balance, R. *Water Quality Monitoring—A Practical Guide to the Design and Implementation of Freshwater Quality Studies and Monitoring Programmes;* Published on behalf of United Nations Environment Programme and the World Health Organization 1996, UNEP/WHO ISBN 0 419 22320 7 (Hbk) 0 419 21730 4 (Pbk).
35. Chapman, D., Ed. *Water Quality Assessments. A Guide to the Use of Biota, Sediments and Water in Environmental Monitoring;* Chapman and Hall: London, 1996.
36. Strobl, R.O.; Robillard, P.D. Network design for water quality monitoring of surface freshwaters: A review. J. Environ. Manage. **2008**, *87,* 639–648.
37. Cooke, G.D.; Welch, E.B.; Peterson, S.A.; Nichols. S.A. *Restoration and Management of Lakes and Reservoirs*, 3rd Ed.; CRC Taylor and Francis Group, New York, 2005.
38. Edmondson, W.T. *Trophic Equilibrium of Lake Washington.* USEPA-600/3–77–087, 1978.
39. Edmondson, W.T. Sixty years of Lake Washington: A curriculum vitae. Lake Reservoir Manage. **1994**, *10,* 75–84.
40. Bjork, S. *European Lake Rehabilitation Activities.* Inst. Limnol. Rept. University of Lund: Sweden, 1974.
41. Edmondson, W.T. Phosphorus, nitrogen, and algae in Lake Washington after diversion of sewage. *Science* **1970**, *169,* 690–691.
42. Cullen, P.; Fosberg, C. Experiences with reducing point sources of phosphorous to lakes. Hydrobiologia **1988**, *170,* 321–336.
43. Scheffer, M. *Ecology of Shallow Lakes;* Chapman and Hall: New York, 1998.
44. McComas, S. *Lake and Pond Management Guidebook*; Lewis Publishers, A CRC Press Company: U.S., 2003.
45. Moore, L.; Thornton, K. *The Lake and Reservoir Restoration Guidance Manual.* EPA 440/5–88-002 First Edition. Washington, DC, 1988.
46. Available at http://www.ramsar.org/cda/ramsar/display/main/main.jsp?zn=ramsar&cp=1_4000_0_ (accessed February 2012).
47. Available at http://www.unece.org/env/water/ (accessed February 2012).

7

Mines: Acidic Drainage Water

Wendy B. Gagliano
and Jerry M. Bigham

Introduction

What Is Acid Mine Drainage?

Acid mine drainage refers to metal-rich sulfuric acid solutions released from mine tunnels, open pits, and waste rock piles (Table 1). Similar solutions are produced by the drainage of some coastal wetlands, resulting in the formation of acid sulfate soils. Acid mine drainage typically yields pH values ranging from 2 to 4; however, extreme sites such as Iron Mountain, California, have produced pH values as low as −3.6.[1] Neutral to alkaline mine drainage is also common in areas where the surrounding geologic units contain carbonate rocks to buffer acidity (Table 1).

Why Is Acid Mine Drainage a Problem?

Landscapes exposed to acid mine drainage do not support vegetation and are susceptible to erosion. When acid mine drainage enters natural waterways, changes in pH and the formation of voluminous

TABLE 1 Summary of Mine Drainage Chemistry from 101 Bituminous Coal Mine Sites in Pennsylvania

	Range	Median	Mean
pH	2.7–7.3	5.2	3.6
Fe (mg/L)	0.16–512.0	43.0	58.9
Al (mg/L)	0.01–108.0	1.3	9.8
Mn (mg/L)	0.12–74.0	2.2	6.2
SO4 (mg/L)	120–2000	580.0	711.2

Unpublished data from C. Cravolta, III, 2001. USGS, Lemoyne, PA.

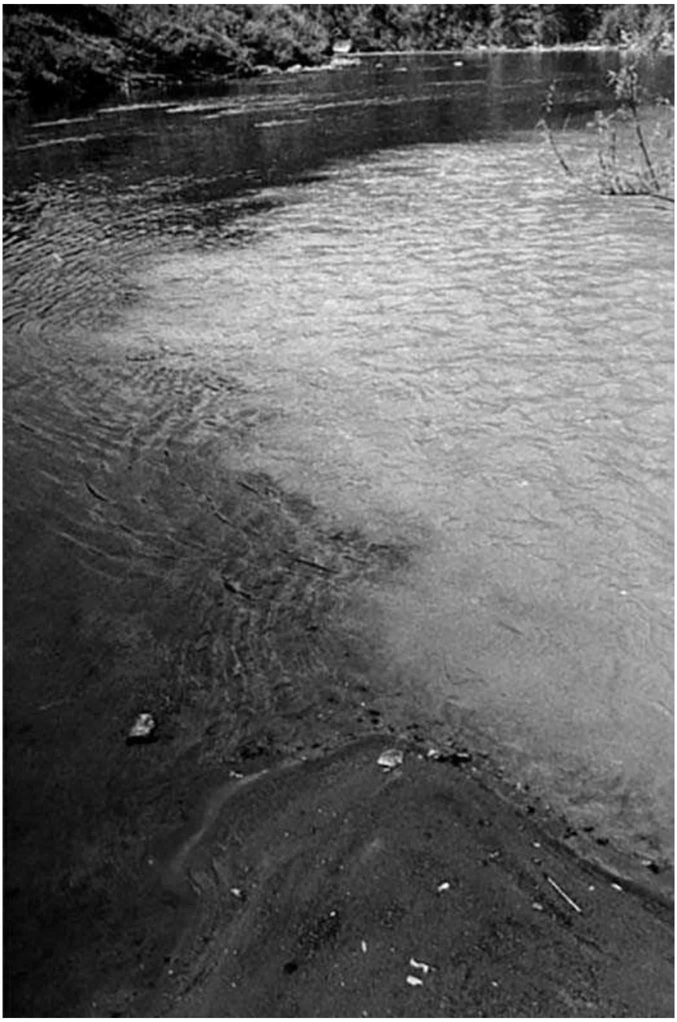


FIGURE 1

precipitates of metal hydroxides can devastate fish populations and other aquatic life (Figure 1). The corrosion of engineered structures such as bridges is also greatly accelerated. There may be as many as 500,000 inactive or abandoned mines in the United States, with mine drainage severely impacting approximately 19,300 km of streams and more than 72,000 ha of lakes and reservoirs.[2,3] Once initiated, mine drainage may persist for decades, making it a challenging problem to solve.

What Causes Acid Mine Drainage?

Mine drainage results from the oxidation of sulfide minerals such as pyrite (cubic FeS_2), marcasite (orthorhombic FeS_2), pyrrhotite ($Fe_{1-X}S$), chalcopyrite ($CuFeS_2$), and arsenopyrite ($FeAsS$). These minerals are commonly found in coal and ore deposits and are stable until exposed to oxygen and water. Their oxidation causes the release of metals and the production of sulfuric acid. This process can occur as a form of natural mineral weathering but is exacerbated by mining because of the sudden, large-scale exposure of unweathered rock to atmospheric conditions.

Mine Drainage Chemistry

Mine drainage is a complex biogeochemical process involving oxidation-reduction, hydrolysis, precipitation, and dissolution reactions as well as microbial catalysis.[1] The entire sequence is commonly represented by Reaction 1, which describes the overall oxidation of pyrite by oxygen in the presence of water to form iron hydroxide [$Fe(OH)_3$] and sulfuric acid.

$$FeS_{2(s)} + 3\frac{3}{4}O_{2(g)} + 3\frac{1}{2}H_2O_{(l)} \rightarrow Fe(OH)_{3(s)} + 2H_2SO_{4(aq)} \tag{1}$$

The actual oxidation process is considerably more complicated.

Pyrite and related sulfide minerals contain both Fe and S in reduced oxidation states. When exposed to oxygen and water, the sulfur moiety is oxidized first, releasing Fe^{2+} and sulfuric acid to solution (Reaction 2). The rate of oxidation is dependent on environmental factors like temperature, pH, Eh, and relative humidity, as well as mineral surface area and microbial catalysis.

$$FeS_{2(s)} + 3\frac{1}{2}O_{2(g)} + H_2O_{(l)} \rightarrow Fe^{2+}_{(aq)} + 2SO^{2-}_{4(aq)} + 2H^{+}_{(aq)} \tag{2}$$

Reaction 2 is most important in the initial stages of mine drainage generation and can be either strictly abiotic or mediated by contact with sulfur-oxidizing bacteria.[4] The Fe^{2+} released by pyrite decomposition is rapidly oxidized by oxygen at $pH > 3$ as per Reaction 3.

$$Fe^{2+}_{(aq)} + \frac{1}{4}O_{2(aq)} + H^{+}_{(aq)} \rightarrow Fe^{3+}_{(aq)} + \frac{1}{2}H_2O_{(l)} \tag{3}$$

If acidity generated by Reaction 2 exceeds the buffering capacity of the system, the pH eventually decreases. Below pH 3, Fe^{3+} solubility increases and a second mechanism of pyrite oxidation becomes important[5] (Reaction 4).

$$FeS_{1(2)} + 14Fe^{3}_{(aq)} + 8H_2O_{(l)} \rightarrow 15Fe^{2+}_{(aq)} + 2SO^{2-}_{4(aq)} + 16H^{+}_{(aq)} \tag{4}$$

In this case, pyrite is oxidized by Fe^{3+} resulting in the generation of even greater acidity than when oxygen is the primary oxidant. Pyrite decomposition is thus controlled by the rate at which Fe^{2+} is converted to Fe^{3+} at low pH.[6] At $pH < 3$, Fe^{2+} oxidation is very slow unless it is catalyzed by populations of iron-oxidizing bacteria like *Acidithiobacillus ferrooxidans* or *Leptospirillum ferrooxidans.* These acidophilic bacteria oxidize Fe^{2+} as a means of generating energy to fix carbon. In doing so, they supply soluble Fe^{3+} at a rate equal to or slightly greater than the rate of pyrite oxidation by Fe^{3+}.[5] Pyrite oxidation then regenerates Fe^{2+} (Reaction 4), creating a cyclic situation that leads to vigorous acidification of mine drainage water.

Mine Drainage Mineralogy

The hydrolysis of Fe^{3+} causes the precipitation of various iron minerals—generally represented as ($Fe[OH]_3$)—that are often the most obvious indicators of mine drainage contamination (Reaction 5).

$$Fe^{3+}_{(aq)} + 3H_2O_{(l)} \rightarrow Fe(OH)_{3(s)} + 3H^{+}_{(aq)} \tag{5}$$

These precipitates are yellow-to-red-to-brown in color and have long been referred to by North American miners as "yellow boy." The actual mineralogy of the precipitates is determined by solution parameters like pH, sulfate, and metal concentration and can vary both spatially and temporally. Some of the most common mine drainage minerals are goethite (α-FeOOH), ferrihydrite ($Fe_5HO_8.4H_2O$), schwertmannite ($Fe_8O_8[OH]_6SO_4$), and jarosite ($[H,K,Na]\ Fe_3[OH]_6[SO_4]_2$).[7]

Goethite is a crystalline oxyhydroxide that occurs over a wide pH range, is relatively stable, and may represent a final transformation product of other mine drainage minerals.[8] Ferrihydrite is a poorly crystalline ferric oxide that forms in higher pH (>6.5) environments. Schwertmannite is commonly found in drainage waters with pH ranging from 2.8 to 4.5, and with moderate to high sulfate contents. It may be the dominant phase controlling major and minor element activities in most acid mine drainage. Jarosite group minerals form in more extreme environments with $pH < 3$, very high sulfate concentrations, and in the presence of appropriate cations like Na and K.

Mine Drainage Microbiology

The most studied bacterial species in mine drainage systems belong to the genus *Acidithiobacillus* (formerly *Thiobacillus*)[9] Species like *Acidithiobacillus thiooxidans* and *A. ferrooxidans* are important to sulfur and iron oxidation in acid drainage; however, many other microorganisms may also be involved.[10] Bacteria have been found in close association with pyrite grains and may play a direct role in mineral oxidation, but they most likely function indirectly through oxidation of dissolved Fe^{2+} as described previously. In low pH systems (<3), *A. ferrooxidans* can increase the rate of iron oxidation as much as five orders of magnitude relative to strictly abiotic rates.[6]

Iron-oxidizing bacteria are chemolithotrophic, meaning they oxidize inorganic compounds, like Fe^{2+}, to generate energy and use CO_2 as a source of carbon. Iron oxidation, however, is a very low energy yielding process. It has been estimated that the oxidation of 90.1 mol of Fe^{2+} is required to assimilate 1 mol of C into biomass.[11] Thus, large amounts of Fe^{2+} must be oxidized to achieve even modest growth.

In addition to mediating iron oxidation, bacteria may play an additional role in mineral formation. Bacteria in mine drainage systems have been shown to be partially encrusted with mineral precipitates.[12] Bacterial cell walls provide reactive sites for the sorption of metal cations, which can accumulate and subsequently develop into precipitates, using the bacterial surface (living or dead) as a template.[13,14]

Environmental Impacts of Mine Drainage

Mine drainage is primarily released from open mine shafts or from mine spoil left exposed to the atmosphere. The drainage produced can have devastating effects on the surrounding ecosystem. Chemical precipitates can obstruct water flow, dramatically increase turbidity, and ruin stream aesthetics. Dissolved metals and acidity can also affect plant and aquatic animal populations.

Besides iron, Al is the most common dissolved metal in acid mine drainage. The primary source of Al is the acid dissolution of aluminosilicates found in soil, spoils, tailings deposits, and gangue material.[7] At high concentrations, Al can be toxic to plants, and colloidal aluminum precipitates can irritate the gills of fish, causing suffocation. Aluminum occurs as a dissolved species at low pH but rapidly hydrolyzes at about pH 5 to form felsöbanyáite $[Al_4(SO_4)(OH)_{10}\ 4H_2O]$ or gibbsite $[Al(OH)_3]$.[7,15] Aluminum precipitates are white in color but are readily masked by associated iron compounds.

Elevated levels of trace elements like As, Cu, Ni, Pb, and Zn may be released during the oxidation of sulfide minerals. These elements can play a role in mineralization processes by forming coprecipitates[16,17] but occur primarily as sorbed species.[18] Mine drainage precipitates can retain both anions and cations, depending on pH. While coprecipitation and sorption function to immobilize trace elements by removing them from solution, this effect may not be permanent. Dissolution of precipitates and shifts in pH can result in the release of sorbed species, providing a latent source of pollution.[19]

Dealing with Mine Drainage

Successful control of mine drainage usually involves elements of both prevention and treatment.

Prevention

Prevention techniques include sealing mine shafts, burying or submerging spoil piles, and adding bactericides to limit the function of iron-oxidizing bacteria. These techniques often have limited success. Sealing of mines is extremely difficult due to fractures and the permeability of surrounding rocks. Covering spoil with soil material can decrease the degree of sulfide oxidation by limiting exposure to oxygen, but establishment of a vegetative cover is necessary to prevent erosion from re-exposing the spoil. Inhibition of iron-oxidizing bacteria with bactericides can decrease sulfide oxidation and reduce metal mobility; however, reapplication is necessary and adequate distribution to all affected areas is difficult. In addition, target bacteria may develop resistance, and beneficial bacteria may be harmed.[20]

Treatment

Solution pH usually underestimates the total acidity of mine drainage. Total acidity is the sum of "proton acidity" and "mineral acidity" generated upon oxidation and hydrolysis of metals like Fe^{2+}, Fe^{3+}, Mn, and Al^{3+}.[21] The traditional approach to treatment of acid mine drainage involves neutralization of total acidity by the addition of alkaline agents like caustic soda (NaOH) or hydrated lime [$Ca(OH)_2$]. This method is effective in neutralizing acidity and precipitating dissolved metals; however, it requires continuous oversight and produces large amounts of waste sludge that require disposal. Newer remediation strategies focus on low-cost, sustainable methods for treatment of drainage waters. For example, limestone drains coupled with settling ponds or compost wetlands have shown some promise as passive remediation technologies.[22] In these systems, drainage is channeled through either oxic or anoxic limestone substrates to neutralize active acidity. Dissolved metals are then allowed to hydrolyze and precipitate in ponds or wetland cells. A major difficulty is the loss of reactive surface by armoring of limestone particles with precipitates of Fe and Al that eventually obstruct flow.

Compost wetlands are designed to stimulate the development of anaerobic microbial populations, particularly sulfate- reducing bacteria. The bacteria use the compost as an organic substrate and remove sulfate from solution, either by converting it to H_2S, which is lost to the atmosphere, or by forming insoluble iron sulfides (Reactions 6 and 7).

$$2CH_3CHOHCOO^-_{(aq)} + SO^{2-}_{4(aq)} \rightarrow 2CH_3COO^-_{(aq)} + 2HCO^-_{3(aq)} + H_2S_{(g)} \tag{6}$$

$$H_2S_{(s)} + Fe^{2+}_{(aq)} \rightarrow FeS_{(s)} + 2H^+_{(aq)} \tag{7}$$

Bicarbonate is formed as a by-product of sulfate reduction and functions to buffer acidity. These systems have also shown limited success in the field. The sulfate removal rates are usually low (<10%), and pH often remains unchanged or decreases within the wetland.[23]

References

1. Nordstrom, D.K.; Alpers, C.N. Geochemistry of acid mine waters. In *The Environmental Geochemistry of Mineral Deposits*; Plumlee, G.S., Logsdon, M.J., Eds.; Society of Economic Geologists Inc: Littleton, CO, 1999; Chapt. 6, 133–160.
2. Kleinmann, R.L.P. Acid mine drainage in the United States: Controlling the impact on streams and rivers. *4th World Congress on the Conservation of Built and Natural Environments;* Toronto, Ontario, 1989; 1–10.
3. Lyon, J.S.; Hilliard, T.J.; Bethel, T.N. *Burden of Guilt;* Mineral Policy Center: Washington, DC, 1993; 68.
4. Rojas, J.; Giersig, M.; Tributsch, H. Sulfur colloids as temporary energy reservoirs for *Thiobacillus ferrooxidans* during pyrite oxidation. Arch. Microbiol. **1995**, *163,* 352–356.

5. Nordstrom, D.K. Aqueous pyrite oxidation and the consequent formation of secondary iron minerals. In *Acid Sulfate Weathering;* Kittrick, J.A., Fanning, D.S., Hossner, L.S., Eds.; Soil Science Society of America: Madison, WI, 1982; 37–56.
6. Singer, P.C.; Stumm, W. Acidic mine drainage: The rate determining step. Science **1970**, *167,* 1121–1123.
7. Bigham, J.M.; Nordstrom, D.K. Iron and aluminum hydroxy- sulfates from acid mine waters. In *Sulfate Minerals: Crystallography, Geochemistry and Environmental Significance;* Alpers, C.N., Jambor, J.L., Nordstrom, D.K., Eds.; Reviews in Mineralogy and Geochemistry; The Mineralogical Society of America: Washington, DC, 2000; Vol. 40, 351–403.
8. Gagliano, W.B.; Brill, M.R.; Bigham, J.M.; Jones, F.S.; Traina, S.J. Chemistry and mineralogy of ochreous sediments in a constructed mine drainage wetland. Geochim. Cosmochim. Acta **2004**, *68,* 2119–2128.
9. Kelly, D.P.; Wood, A.P. Reclassification of some species of *Thiobacillus* to the newly designated genera *Acidithio- bacilus* gen. nov., *Halothiobacillus* gen. nov. and *Thermi- thiobacillus* gen. nov. Int. J. Syst. Evol. Microbiol. **2000**, *50,* 511–516.
10. Gould, W.D.; Bechard, G.; Lortie, L. The nature and role of microorganisms in the tailings environment. In *The Environmental Geochemistry of Sulfide Mine-Wastes*; Blowes, D.W., Jambor, J.L., Eds.; Mineralogical Association of Canada Short Course, 1994; Vol. 22, 185–200.
11. Ehrlich, H.L. *Geomicrobiology*, 3rd Ed.; Marcel Dekker, Inc.: New York, 1996.
12. Clarke, W.A.; Konhouser, K.O.; Thomas, J.C.; Bottrell, S.H. Ferric hydroxide and ferric hydroxysulfate precipitation by bacteria in an acid mine drainage lagoon. FEMS Microbiol. Rev. **1997**, *20,* 351–361.
13. Schultze-Lam, S.; Fortin, D.; Davis, B.S.; Beveridge, T.J. Mineralization of bacterial surfaces. Chem. Geol. **1996**, *132,* 171–181.
14. Konhauser, K.O. Diversity of bacterial iron mineralization. Earth-Sci. Rev. **1998**, *43,* 91–121.
15. Nordstrom, D.K. Mine waters: Acid to circumneutral. Elements **2011**, *7,* 393–398.
16. Carlson, L.; Bigham, J.M.; Schwertmann, U.; Kyek, A.; Wagner, F. Scavenging of As from acid mine drainage by schw- ertmannite and ferrihydrite: A comparison with synthetic analogues. Environ. Sci. Technol. **2002**, *36,* 1712–1719.
17. Hita, R.; Torrent, J.; Bigham, J.M. Experimental oxidative dissolution of sphalerite in the Aznalcóllar sludge and other pyritic matrices. J. Environ. Qual. **2006**, *35*, 1032–1039.
18. Winland, R.L.; Traina, S.J.; Bigham, J.M. Chemical composition of ochreous precipitates from Ohio coal mine drainage. J. Environ. Qual. **1991**, *20,* 452–460.
19. Lee, G.; Faure, G.; Bigham, J.M.; Williams, D.J. Metal release from bottom sediments of Ocoee Lake No. 3, a primary catchment area for the Ducktown Mining District. J. Environ. Qual. **2008**, *37,* 344–352.
20. Ledin, M.; Pedersen, K. The environmental impact of mine wastes—Roles of microorganisms and their significance in treatment of mine wastes. Earth-Sci. Rev. **1996**, *41,* 67–108.
21. Kirby, C.S.; Cravotta, C.A., III. Net alkalinity and net acidity I: Theoretical considerations. Appl. Geochem. **2005**, *20,* 1920–1940.
22. Cravotta, C.A., III; Trahan, M.K. Limestone drains to increase pH and remove dissolved metals from acidic mine drainage. Appl. Geochem. **1999**, *14,* 581–606.
23. Mitsch, W.J.; Wise, K.M. Water quality, fate of metals, and predictive model validation of a constructed wetland treating acid mine drainage. Water Resour. **1998**, *32,* 1888–1900.

8

Rivers and Lakes: Acidification

Agniezka Gałuszka
and Zdzistaw M.
Migaszewski

Introduction

Acidification of rivers and lakes was one of the major environmental issues in the second half of the 20th century. As a consequence of the release of large amounts of acidifying gases (SO_2 and NO_x) to the atmosphere, mostly from coal burning and vehicle exhausts, the pH of precipitation dropped below 5.6. Acid deposition was responsible for lowering of the pH of surface waters and soils, and in many regions of the world, it resulted in forest decline and fish kills.

This entry presents the problem of acidification of rivers and lakes caused by natural and anthropogenic factors that contribute to decrease in the pH of freshwaters. In the first section, a definition of acidification and characteristics of processes causing acidification are provided with a brief description of the metrics used for acidification measurements. Historical perspectives of the problem are also briefly described. In a subsequent section, natural and anthropogenic acidification of rivers and lakes is discussed in more detail. The final part of the entry summarizes the consequences of freshwater acidification and the attempts made to solve this issue.

What Is Acidification, and How Is It Measured?

The most frequently used parameter that characterizes the acidity or alkalinity of natural waters is the pH. It is defined as a negative logarithm of hydrogen (hydronium) ion activity. Typical pH values in lakes and rivers are in the 6–9 range. The term “acidification” is usually used for freshwater with a pH of about 5.0 (and alkalinity below 0).[1] Some authors suggest that the pH threshold for acidified water should be 5.65 because this value characterizes pure water in equilibrium with atmospheric CO_2.[2] Two types of acidification of lakes and rivers can be distinguished: 1) chronic or long term in case acidifying factor is permanent (on a decade– century timescale), for example, related to bedrock or soil characteristics; and 2) episodic or short term (on an hour– day timescale), when acidifying factor is temporal, for example, acidic metal-rich effluent, snowmelt runoff, heavy rainfalls etc.

Acidification of surface waters begins when the concentration of hydrogen ions exceeds base ions present in water. The bases are produced during hydrolysis of carbonate minerals (mostly calcite) that occur in rocks and soils. The difference between the sum of the cations of strong bases and the sum of the anions of strong acids is termed acid-neutralizing capacity (ANC). This parameter is widely used in modeling of freshwater acidification and in assessment of freshwater sensitivity to acidification (critical load concept). Its components include both strong (OH^-) and weak bases (e.g., carbonate species, aluminohydroxides, organic acids). Alkalinity expresses the ability of water to neutralize acids. It is often understood as a synonym of ANC only measured in filtered water samples, whereas ANC is measured in unfiltered samples.[3] The buffering capacity of natural waters is mostly governed by a bicarbonate buffering system. Gaseous CO_2, which is a product of weathering of carbonate rocks or decay of organic remains, dissolves in water and produces carbonic acid, which subsequently dissociates to form H^+, HCO_3^-, and CO_3^{2-} ions. These reactions remain in equilibrium:

$$CO_2 + H_2O \leftrightarrow H_2CO_3 \leftrightarrow HCO_3^- + H^+ \leftrightarrow CO_3^{2-} + 2H^+$$

Bicarbonate and carbonate ions are capable of removing H^+ from water, causing the reactions to shift to the left. The bicarbonate buffering system results in a relatively constant pH and is mainly responsible for the alkalinity of natural waters. Thus, the influence of acidifying compounds on freshwater pH is more complex and it depends on ANC of the catchment area. The concept of "critical load" was introduced to find catchments that are the most vulnerable to acidification. Critical loads, defined as "an ecological threshold or intolerance to the accumulation of a pollutant in an ecosystem,"[4] can be estimated precisely on the basis of deposition rates of acidifying compounds and freshwater chemistry (empirical approach). The critical loads for lakes is a pH of 6.0 and an ANC of 0.02 meq/L.[5] Waters that have low alkalinity values (in the range of 0–10 mg/L $CaCO_3$) are highly sensitive to acidification.[6]

Alternatively, information about local geologic makeup, soil, and the land use of a study area may be employed to assess the sensitivity of freshwater to acidification.[7] The response of diatom assemblages to acid deposition and calcium levels in waters can also be used to calculate critical loads.[8] The difference between deposition of acidifying compounds and critical load values is termed the exceedance of critical load, and its values are crucial in assessment of potential acidification of freshwaters in a given area.

Paleoenvironmental data and fly-ash particle analysis in dated sediment cores are applied to the study of the history of surface water acidification and allow one to address the question regarding the origin of acidification (anthropogenic or natural). Paleolimnological pH reconstruction is based on the assumption that diatom and chrysophyte microfossils in dated sediment core intervals reflect the pH of water at a time when these organisms thrived in it. Based on the diatom techniques, it is possible 1) to evaluate models of acidification; 2) to determine critical loads of acidity; 3) to study the extent of episodic acidification; and 4) eventually to monitor lakes and rivers that are recovering from acidification.[9] The results of paleolimnological studies have indicated that the pH of some lakes located in northern Europe, eastern Canada, and northeastern United States have dropped by 0.5–1.5 from the early 1800s to recent times.[10]

Historical Perspectives

Acid deposition has been studied since the onset of the industrial revolution when Robert Smith, a creator of the term "acid rain", published the first results of his study on the rainwater chemistry in Manchester, U.K.[10] Surprisingly, it took more than 100 years to recognize the influence of acid deposition on freshwater acidification.[11,12] Although the first studies on acid deposition were carried out in Europe, the impact of precipitation chemistry on freshwater ecosystems was originally identified in North America in the late fifties by Gorham.[11] In 1968, Svante Odén showed the relationship between sulfur emissions in continental Europe and sulfate deposition affecting acidification of lakes in Scandinavia.[12]

Fleisher and others[13] have distinguished three periods of research regarding acidification of the environment: 1) an inventory period (during the 1960s and 1970s); 2) a period of mechanism studies (late 1970s and 1980s); and 3) a synthesizing period (1990s). This sequence of scientific interests has been supplemented by Herrmann[14] who suggested that a consequence of detailed knowledge on this issue should trigger a period of implementation.

Lake and river acidification became one of the major environmental issues in the 1970s and 1980s, when most of the reports from regional monitoring surveys were published. Freshwater acidification was most intensively studied in Europe (the U.K. and Scandinavian countries including Norway, Sweden, and Finland) and North America (Canada and United States).

During the 1980s, the concept of critical load was applied to quantify ecosystem vulnerability to acidification. This approach was also widely used in an assessment of freshwater sensitivity to acidification, mostly in Europe, North America, and Asia.[6] Recent research activities are focused on the recovery of lakes and rivers from long term acidification in most susceptible to acidification areas throughout the world, with some constraints to their future.[15–17] Attention is also being given to the emerging acidification of freshwaters in developing countries that experience rapid industrialization (Asia, Africa, Latin America).[18,19]

Sources of Freshwater Acidification

Factors influencing freshwater acidification may be natural (Figure 1) or anthropogenic (Figure 2) in origin. The lowest pH values are recorded in inland waters influenced by natural sources of acidifying compounds, especially of geologic origin. However, it should be stressed that natural acidification is commonly overlapped or enhanced by anthropogenic influences. Hence, separating natural acidity from anthropogenic freshwater acidification due to their different effects on biota is reasonable.[20] An important difference between anthropogenic and natural acidification is that the former one is reversible, as shown by many studies conducted after decreasing SO_2 and NO_x emissions.[21]

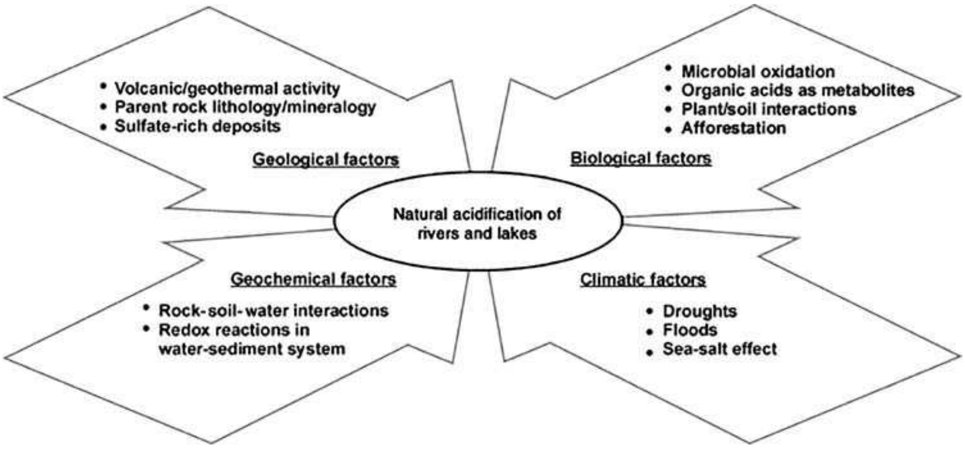


FIGURE 1 The most important natural factors influencing freshwater acidification.

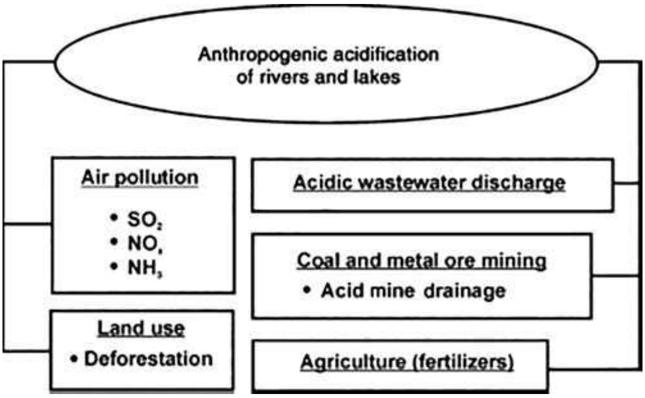


FIGURE 2 The most important anthropogenic factors influencing freshwater acidification.

Natural Sources of Freshwater Acidification

Geological Factors

Acidification of lakes and rivers as a consequence of natural processes in the environment is highly dependent on geologic factors, such as the presence of volcanic and geothermal activity and lithology/mineralogy of parent rocks. The volcanic/postvolcanic processes are responsible for formation of the world's mostly acidified surface waters; for example, the pH of caldera lake Kawah Ijen in western Java, Indonesia, is <0.3,[22] whereas the Banyupahit river, which originates from this lake, exhibits a pH in the range of about 0.7 to 3.3. Another example is the Kuril Islands (currently Russia) where many lakes and rivers are influenced by volcanic/postvolcanic activity. The main constituents of exhalations are sulfuric and hydrochloric acids. For example, the pH of Lake Usoriko (Japan) varies from 3.4 to 3.8.[23]

The extremely low pH values of inland waters are also caused by oxidation of pyrite and iron-bearing sulfide minerals (e.g., chalcopyrite, pyrrhotite, and marcasite). The mechanism of this process is natural and is known as "acid rock drainage" (ARD). It should be emphasized that ARD formation rate is usually accelerated by mining activities, and this process is called "acid mine drainage" (AMD), which will be further discussed in the next section. A series of reactions accompany the weathering of pyrite. Pyrite can undergo oxidation by two natural oxidants: 1) oxygen and 2) the even more effective ferric iron (Fe^{3+}), according to simplified reactions[24]:

$$FeS_2\left(\text{pyrite}\right) + 3.5O_2 + H_2O \rightarrow Fe^{2+} + 2SO_4^{2-} + 2H^+ \tag{1}$$

$$FeS_2 + 14Fe^{3+} + 8H_2O \rightarrow 15Fe^{2+} + 2SO_4^{2-} + 16H^+ \tag{2}$$

The first reaction is assisted by acidophilous bacterium species *Acidithiobacillus ferrooxidans*. In case of the second reaction, Fe^{3+} rapidly oxidizes pyrite in abiotic and anaerobic conditions. These two reactions (1 and 2) bring about a substantial drop in pH and an increase in concentrations of ferrous (Fe^{2+}) and sulfate ions in water. Subsequently, the Fe^{2+} undergoes oxidation and the dissolved Fe^{3+} supports the second reaction or may hydrolyze and ferric hydroxide precipitates according to the following simplified reactions:

$$Fe^{2+} + 1/4O_2 + H^+ \rightarrow Fe^{3+} + 1/2H_2O$$

$$Fe^{3+} + 3H_2O \rightarrow Fe\left(OH\right)_3 \downarrow 3H^+$$

This acidity has further been generated by complex processes of transformation of iron oxyhydroxysulfates and oxyhydroxides into goethite (α-FeOOH).

Although geologic processes lead to the formation of extremely low pH waters, their detrimental influence on the environment is usually restricted to relatively small areas. Besides, these factors occur in the environment for a long time, creating a unique ecological niche for acid tolerant organisms. Table 1 presents examples of surface inland waters acidified mostly by geological factors.

Geochemical Factors

Rock–Soil–Water Interactions The pH and concentrations of metals in surface waters are largely controlled by the chemistry and mineralogy of bedrock and soils. In areas where bedrock consists of igneous and metamorphic rocks (granites, gneisses, porphyries) or silicoclastic rocks, soils have a low buffering capacity, resulting in a naturally low pH of lakes and rivers. The slow weathering of rocks contributes particularly to increasing freshwater acidification. This process can be accelerated by construction activities or land melioration. The extent of exposure of mineralized rock formations to weathering at the surface, the amount of scattered sulfide minerals, and the presence or absence of buffering carbonate rocks or gangue minerals are decisive factors that affect the range of acidification of surface waters. Other geochemical factors having an influence on acidification are acidic metalliferous springs, whose waters have interacted with mineralized rock formations, and unmined mineral deposits.

Other soil properties can affect the pH of surface waters, mostly through ion-exchange reactions and their influence on water buffering capacity. Lakes and rivers located in areas with natural acidic soils are more susceptible to acidification.

Acid sulfate soils are soils abundant in pyrite and other sulfide minerals. These soils, formed under waterlogged conditions, when exposed to the air (e.g., by draining, excavation, cultivation), undergo chemical reactions that are responsible for sulfuric acid generation and subsequent acidification.

TABLE 1 Examples of Surface Inland Waters Acidified Mostly by Geological Factors

Name	Type	pH	Country	Reference
		Influence of volcanic/geothermal activity		
Kawah Ijen	Lake	<0.3	Indonesia	[22]
Banyupahit	River	0.7–3.3		
Usoriko	Lake	3.4–3.8	Japan	[23]
Caviahue	Lake	2.40–2.48	Argentina	[25]
Upper Rio Agrio	River	1.78		
Yugama	Lake	0.90–1.80	Japan	[26]
Popocatépetl	Lake	1.37–1.50	Mexico	[27]
Poás	Lake	0	Costa Rica	[28]
Kislyi Creek	River	2.45	Russia	[29]
Lesnaya	River	4.8		
		Influence of bedrock lithology/mineralogy or acid rock drainage		
Woods Lake	Lake	4.4–5.9	United States (Adirondack Mts)	[30]
Lake Härkälampi	Lake	3.8–7.3	Finland	[31]
Langedalstjenn	Lake	4.4	Norway	[32]
Isiurqutuuq	Lake	4.5	Canada	[33]
Snake River	River	3.0	United States, Colorado	[34]
Rio Tinto	River	<3.0	Spain	[35]

Pyrite-rich soils that are not exposed to weathering pose no hazard to the environment and are called "potential acid sulfate soils" or cat-clays.

Most of the acid sulfate soils formed during the last 10,000 years following the postglacial sea level rise and the eustatic uplift of landmass. The acid sulfate soils occur mostly in lowland coastal regions (Australia, southeastern Asia, Finland), but they may also be associated with other environmental compartments favoring pyrite formation and oxidation, such as freshwater wetlands or, as mentioned above, areas with sulfide-scattered rock formations. It is estimated that acidic sulfate soils encompass an area of about 20 million ha globally.[36] The impact of acidic sulfate soils on freshwaters is highlighted by leaching of oxidation products into streams and lakes. The increased acidity of both soils and waters intensifies the mobility of many elements creating potentially toxic environments for living organisms.

Although acidic sulfate soils influence surface water acidification, there are also other soil types that may contribute to this process. Red soils developed on silicate rocks in Okinawa Island, Japan,[37] are one example. Their pH is low (4.5–5.5) and surface waters in the red-soil-dominated areas exhibit a pH within the range of 4.9–5.8.

Redox Reactions in Water–Sediment Systems

Freshwater pH can be changed by redox processes in sediments. During oxidation reactions, H^+ ions are produced, which decreases the pH of water. Nitrogen and sulfur (as inorganic and organic species) and iron are the most important elements in sediment/water redox processes. The changes in their oxidation state are usually microbially mediated. The effective acid production capacity (APC_{eff}) is a measure of acidification caused by oxidation reactions in sediment suspension. It is defined as

$$APC_{eff} = V/W.\left(\left[H^+\right]_e - \left[H^+\right]_O\right)$$

where V is the suspension volume, W is the solid mass, and $[H^+]_{e,o}$ is the H^+ concentration before and after oxidation.[38] The oxidation reactions that are responsible for acid generation in aquatic systems include 1) oxidation of sulfur from hydrogen sulfide, atomic sulfur, metal sulfides, and organic compounds (R-SH); 2) oxidation of pyrite (FeS_2) and Fe^{2+} ions; and 3) oxidation of N from NH_4^+, NO^-_2 (nitrification), and organic species (R-NH_2) (ammonification).

Several important buffer reactions occur in sediments, which can consume excess hydrogen ions and contribute to an increased ANC. These reactions involve $CaCO_3$, Al_2O_3, and Fe_2O_3:

$$CaCO_3 + 2H^+ \rightarrow Ca^{2+} + H_2O + CO_2 \tag{5}$$

$$Al_2O_3 + 6H^+ \rightarrow 2Al^{3+} + 3H_2O \tag{6}$$

$$Fe_2O_3 + 6H^+ \rightarrow 2Fe^{3+} + 3H_2O \tag{7}$$

Table 2 presents examples of surface inland waters acidified mostly by geochemical factors.

Biological Factors

Metabolic processes of living organisms also can contribute to the lowering of surface water pH. This drop in the pH values may be caused by numerous oxidation reactions that are mediated by microorganisms and produce H^+ as described in the previous sections.

As suggested by Rosenqvist as early as 1978,[44] expansion of coniferous forests brought about a decrease in the pH of surface waters in Norway. Afforestation of acid sensitive areas may result in freshwater acidification due to the following factors: 1) increased evapotranspiration; 2) intensive uptake of

TABLE 2 Examples of Surface Inland Waters Acidified Mostly by Geochemical Factors

Name	Type	pH	Country	Reference
Cudgen Lake	Lake	2.5	Australia	[39]
Rocky Mouth Creek	Stream	<4.5	Australia	[40]
Esse River		5.9–6.8		[41]
Purmo River	Rivers	4.8		
Kronoby River		5.0	Finland	[42]
Kovjoki River		4.7		
Larsmo Lake	Lake	4.7–5.7		
Colour Lake	Lake	3.6–4.7	Canada	[43]

cations during the tree-stand growth; 3) increased filtering of gases and aerosols in the canopy and a subsequent increase of dry deposition; and 4) influx of organic acids from litter fall decay.[45] Acidifying effects of afforestation depends on the tree-stand maturity. After reaching a state of equilibrium between the standing and dead biomass, the forest-induced acidity decreases due to the lower uptake of cations from the soil by mature trees.[46]

Organic acids produced by living organisms are also responsible for acidification of freshwater. It has been estimated that organic acids may change the pH of lake waters having ANC values within the range of 0–50 μeq/L by 0.5–2.5 pH units.[47] On the other hand, organic acids prevent surface waters from further acidification due to their role as buffering substances for acids originating from acid deposition. Dissociation of humic and fulvic acids increases H^+ concentrations in waters leading to their pH decrease. The typical pH of lakes affected by organic acids generated by *Sphagnum* mosses is 3.3–4.5.[1] Table 3 presents examples of surface inland waters acidified mostly by biological factors.

Climatic Factors

The influence of climatic factors on freshwater acidification is mostly related to precipitation and temperature variations. Several processes are involved with acidification brought about by seasonal climatic changes. Water table lowering during long drought periods accelerates oxidation reactions and accompanying production of H^+ ions, especially in wetlands.[55] Water table rising during floods causes intensive water flow through surface soil horizons, which are usually more acidic as a result of higher organic acid concentrations and higher susceptibility to anthropogenic acidification in relation to underlying horizons. During winter, the ice-covered lake waters accumulate CO2 originating from respiration, organic decay, and rock weathering. Due to this process, the pH of waters in epilimnion

TABLE 3 Examples of Surface Inland Waters Acidified Mostly by Biological Factors

Name	Type	pH	Country	Reference
Valkea-Kotinen	Lake	5.0–5.4	Finland	[48]
Rio Negro	River	4.63–5.80	Brazil	[49]
Rio Daraá	River	3.91		
Dumai	River	4.3	Indonesia	[50]
Pebbleloggitch	Lake	4.3	Canada	[51]
Yuanyang	Lake	4.8–5.5	Taiwan	[52]
Grosse Fuchskuhle	Lake	4.2–6.1	Germany	[53]
Liepsalas	Lake	4.1–4.3		
Murmasts	Lake	3.7–3.9	Latvia	[54]
Pieslaistes	Lake	4.0–4.2		

is decreased. Precipitation waters are ionically deficient and have a very low ANC. Therefore, their discharges to surface waters during spring snowmelt floods or heavy storms bring about an episodic decrease in the pH of surface waters.

Occasional acidification of lakes and rivers in coastal areas is caused by deposition of marine aerosol salts during high wind periods, changing cation-exchange equilibria in soil. Cations present in seawater (Na^+ and Mg^{2+}) undergo cation-exchange reactions with hydrogen and aluminum ions, which are released from soil to water. This phenomenon is termed "sea-salt effect."[45]

Winds that carry $CaCO_3$-rich. dusts may increase ANC and pH of surface waters by mitigating the effect of freshwater acidification, as was reported for the influx of Saharan dust to southern Alps.[56] Table 4 presents examples of surface inland waters acidified mostly by climatic factors.

Anthropogenic Acidification of Rivers and Lakes

Air Pollution with Acidifying Compounds

Serious attention to the problem of freshwater acidification was given when it became clear that there was an influence of air pollutants on the formation of acid precipitation and subsequent changes in the pH of surface waters and soils. Atmospheric deposition of acidifying compounds from anthropogenic sources has greatly contributed to the river and lake acidification. Sulfur dioxide and nitrogen oxides undergo chemical transformations in the air, producing sulfuric and nitric acids, which are then deposited on the land surface, lowering the pH of surface waters. Oxidation of SO_2 may undergo gas-phase or aqueous-phase reactions, and the sulfuric acid produced may occur as liquid aerosols, or it may dissolve in clouds or rainwater. It is noteworthy that only a small fraction of acid precipitation is deposited directly in rivers and lakes; freshwaters are mostly impacted by surface runoff.

Most of the studies on the influence of acid deposition on acidification of soils and waters have been focused on SO_2 emissions derived from fossil fuel combustion and smelting of sulfide minerals, as well as on NO_x released from fossil fuel combustion and vehicle exhausts.[67] Before reduction of SO_2 emissions, acid precipitation was mostly attributed to sulfuric acid. Since the 1990s, nitrogen compounds have been regarded as a major contributor to lowering of the precipitation pH.[68] Hydrochloric acid also causes a decrease in the precipitation pH, although to a much lesser extent than sulfur and nitrogen oxides. Anthropogenic emissions of NH_3, derived mostly from livestock farming and agriculture, neutralize acids that occur in the atmosphere. However, after deposition, $NH4^+$ ions affect freshwater pH through nitrification reactions that produce H^+ ions and via uptake of NH_4^+ ions, which in turn

TABLE 4 Examples of Surface Inland Waters Acidified Mostly by Climatic Factors

Name	Type	pH	Country	Reference
		Induced by drought or flood		
Swan Lake	Lake	4.5	Canada	[57]
White pine Lake	Lake	4.1– 5.7	Canada	[58]
Hermanninlampi	Lake	4.0	Finland	[59]
Sink Beck	Pool	4.3	U.K.	[60]
Van Campens Brook	Stream	4.5	United States	[61]
Biwa	Lake	4.5– 6.5	Japan	[62]
		Sea-salt effect (pH values measured during episodes)		
Svela	Stream	4.45	Norway	[63]
Lake Terjevann	Lake	4.28	Norway	[64]
Espedal	River	4.6	Norway	[65]
Allt a'Mharcaidh	Stream	<5.5	U.K.	[66]

release H^+ to soil solution. The acidification caused by strong acids contained in precipitation occurs immediately, whereas the acidification brought about by NH_3/NH_4^+ deposition is delayed in time. It is interesting to note that the acidification of ecosystems from NH_3 emissions may be equal to that from SO_2 emissions or even greater than that from NO_x emissions.[69]

Acid Mine Drainage

Acidification of freshwaters can be triggered by weathering of pyrite and other iron-bearing sulfide minerals in metal and coal mining areas. This process is termed acid mine drainage (AMD). Mine waste piles, mineral ponds, underground mines, and lignite open pits can potentially form acidic and/or metal-rich effluents that can enter local streams and rivers, which in turn may eventually reach various inland water bodies. The mechanism of acid formation resulting from pyrite oxidation was previously described in the subsection "Geological Factors."

AMD was responsible for serious environmental disasters in many areas throughout the world where metal ores or coals were mined.[70] The acid mine waters lead to lowering of the pH combined with the increase of sulfates and toxic element levels in surface water and groundwater. The most acidified waters in the world, originating from abandoned metal ore mines, were found in the Iron Mountain mining area in Shasta County, northern California.[71] In this area, several mines extracted pyrite (for sulfuric acid production) and Ag, Au, Cu, Fe, and Zn sulfide ores from the 1860s to 1962. The abandoned mines contain approximately 600,000 m^3 of strongly acidic water having a pH of about 1 and concentrations of Zn, Cu, and Cd in the range of several grams per liter. Prior to remediation, extremely high loads of dissolved Cd, Cu, and Zn (about 300 tons/yr) were discharged to the Sacramento River. Several massive fish kills were recorded as a consequence of high runoff episodes. Subsequent to the closure of the mines, more than 20 such events have occurred, with the most severe one in 1967, when at least 47,000 trout were killed during a single week.[72] Table 5 presents examples of surface inland waters acidified mostly by anthropogenic factors.

Other Sources

Aside from the natural and anthropogenic factors that were described in the previous subsections, there are also numerous sources of surface water acidification that are manifested by small-scale pH variations or are related to very specific cases.

Acidic wastewaters are also generated by metal processing industries, electroplating, oil refineries, and other chemical industries. Some wastewaters are not acidic, but eventually will become so after they undergo microbially mediated transformations (e.g., brewery wastewaters rich in organic carbon). The knowledge of wastewater properties enables us to control and avoid their detrimental influence on the environment with the help of the proper treatment technology.

Land use is sometimes responsible for acidification of surface waters. One of the examples is deforestation, which causes increased sensitivity of lakes and streams to acidification due to the removal of

TABLE 5 Examples of Surface Inland Waters Acidified Mostly by Anthripogenic Factors

Name	Type	Minimal pH Value	Country	Reference
Big Moose	Lake	4.6	United States (Adirondack Mts.)	[73]
Spring Creek	Stream	2.5	United States (Iron Mountain, California)	[74]
Lohi	Lake	4.4	Canada	[75]
ML111	Mining lake	2.54	Germany (Lusatia)	[76]
Lake Orta	Lake	3.9	Italy	[77]
Lysevatten	Lake	4.5	Sweden	[78]
Vikedal	River	5.35	Norway	[79]
Gentil Sapin	Stream	4.16	France	[80]
Čertovo	Lake	4.3	Czech Republic	[81]

cations with harvested biomass and intensification of nitrification of soils.[2] The increased acidity of lakes and rivers may also occur in agricultural lands, where acidification may come from fertilizers.

There are some additional natural factors contributing to surface water acidification, which need mentioning. Acidification of lakes depends on the watershed location because headwater lakes are more susceptible to acidification than downstream ones. Lakes located in the headwaters have a lower ANC.[82] Large amounts of sulfur dioxide released during natural fires in exposed lignite beds in Smoking Hills (Canada) caused acidification from deposition of sulfuric acid that was produced in the atmosphere.[83]

Environmental Problems Caused by Acidification of Lakes and Rivers

Inland water acidification has serious consequences for living organisms and ecosystem functioning. The direct influence of low pH on aquatic biota includes 1) the loss of species, especially fish (below pH 4.5), zooplankton, phytoplankton, benthic invertebrates, and periphyton; and 2) decreased or inhibited reproduction of fish and amphibians.[84]

Biological response to acidification may be rapid (a short-term effect) or delayed in time (a long-term effect).[85] Rapid response is caused by episodic events that result in a sudden decrease in the water pH, such as spring floods, heavy storms, or discharge of acidic wastewaters. The long-term effects on aquatic biota are highlighted mostly by changes in the ecosystem structure, with some reductions in species diversity and disappearance of acidsensitive species and dominance of acid-tolerant ones. The reduced biodiversity of freshwater systems in turn has an influence on the balance in biogeochemical cycles[86] and causes adverse changes in food chains. Acidification of lakes and rivers affects geochemical processes, including 1) higher mobility of most metals (e.g., Al, Pb, Zn, Mg, Cd); 2) increased methylmercury production; and 3) decreased phosphorus availability.

Aluminum mobilization from soil minerals to waters is responsible for increased fish mortality. The lethal aluminum toxicity to aquatic animals is most often recorded when acidic water mixes with alkaline water and $Al(OH)_3$ precipitates on the gills and filtering apparatus.[87] Some authors regard the release of free aluminum ions, which represent a bioavailable form for aquatic biota, as the most detrimental effect of freshwater acidification.[2]

Although methylmercury production is reduced at a low pH, synthesis of dimethylmercury by base-catalyzed disproportionation contributes to conversion of dimethyl mercury to monomethyl mercury by acid hydrolysis:

$$(CH_3)_2\,Hg + H^+ \rightarrow CH_3Hg^+ + CH_4 \uparrow \tag{8}$$

As a consequence of this reaction, an easily available and very toxic form of mercury is released to acidified surface waters.[86]

Bioavailability of phosphorus in acidified surface waters is diminished, as this element is adsorbed on aluminum oxides and hydroxides and may be irreversibly retained in sediments.[88] Acidic freshwaters decrease bioavailability of essential elements, such as calcium and magnesium.[89]

The response of aquatic biota to natural and anthropogenic acidification is believed to be different.[90] The mechanism of this difference is not well understood and can be explained by biogeographical histories of the region or the long-term evolutional adaptation to natural acidification sources.[90,91]

Solutions to the Problem of Freshwater Acidification

The problem of freshwater acidification has stimulated actions to minimize its consequences, mostly through legislation and various international initiatives that allowed reduction of the levels of acidifying air pollutants. An important tool for rehabilitation of acidified waters has been offered by technology.

Every effort undertaken to reduce the risk of freshwater acidification, apart from its character, has been driven by progress in scientific research as well as by international cooperation.

Environmental Legislation

Environmental legislation, both on national and international scales, has led to substantial reduction of SO_2 and NO_x emissions in Europe and North America in recent decades and has ameliorated the effect of acid precipitation on acidification of rivers and lakes. One of the most important documents concerning the control of acidifying compounds is The Convention on Long-Range Transboundary Air Pollution (LRTAP) under the auspices of the United Nations Economic Commission for Europe.[92] This Convention has been extended by four Protocols that control emission of acidifying compounds: 1) the 1999 Protocol to Abate Acidification, Eutrophication, and Ground-Level Ozone, supplemented with the revised guidance document on ammonia; 2) the 1994 Protocol on Further Reduction of Sulphur Emissions; 3) the 1988 Protocol Pertaining to the Control of Nitrogen Oxides or Their Transboundary Fluxes; and 4) the 1985 Protocol on the Reduction of Sulphur Emissions or their Transboundary Fluxes by at least 30%. One of the initiatives directly regarding acidification of lakes and rivers is The International Cooperative Programme on Assessment and Monitoring of Acidification of Rivers and Lakes (ICP Waters), which is the action of The Working Group on Effects within LRTAP Convention. The ICP Waters initiative collects data that are essential for monitoring the effects of acid precipitation on surface waters from 18 European countries, United States, and Canada.

For European Union member states, the emission standards of SO_2 and NO_x from specific sources are regulated by adequate directives, for example, emission limit values for new plants and industrial facilities from the Large Combustion Plant Directive (88/609/EEC) or the Directives 98/70/EC and 1999/32/EC that establish the limits on the sulfur content of gas oil for stationary and mobile sources and for heavy fuel oil.[93] In Europe, present emissions of SO_2 are less than half of the 1980s levels and emissions of NOx have been reduced by about 20%.[21]

In the United States, reductions in SO_2 and NO_x emission are regulated by the Amendment to the Clean Air Act, Title IV: Acid Deposition Control, also known as the Acid Rain Program, which was implemented in 1990. This document imposed reductions in SO_2 and NO_x emissions from fossil fuel power plants. The emissions of SO_2 from power generating plants in the United States were reduced from 17.3 million tons in 1980 to 11.2 million tons in 2000 and those of NO_x from 6.0 to 5.1 million tons, respectively.[94]

Technology

Both environmental technology and engineering help to reduce the emissions of acidifying gases from industrial sources and facilitate the proper management of acidified freshwaters. The practice, which is most commonly used for rehabilitation of acidified lakes and rivers, is liming. Its purpose is to restore the water to its preacidification state, which is often reconstructed with paleolimnological techniques. Limestone powder can be dispersed on the whole surface of the lake or, alternatively, can be dosed onto the ice cover. It is suggested that liming of lakes and inlet streams should be done with terrestrial liming, which ameliorates the effect of reacidification of the littoral zone during the ice cover and the snowmelt events.[95] Liming was extensively applied in Scandinavian countries; for example, in Sweden, it was used for about 8000 lakes and 12,000 km of surface waters to achieve the following restoration goals: $pH>6$ and $ANC>100$ meq/L.[96] Although liming is very popular, there are some limitations in its application; for example, Wällstedt and Borg,[97] in their study on remobilization of metals from limed and nonlimed lake sediments, have shown that there is a risk of potential toxicity of Al, Cd, Mn, and Zn in lakes that have been limed. Moreover, some toxic anionic metalloids (especially arsenic), which are more mobile at a higher pH, may be released from the sediment to the water column in case of alkalization by lime.

The Swedish Forest Agency has recommended that wood ash (received from timber waste combusted for energy production) should be used to mitigate acidification and to supplement nutrient removal.[98] In this method, the wood ash is added to the soil cover. Treatment of acidified waters is also possible with the use of ecotechnologies, such as constructed wetlands or controlled eutrophication.[99] The employment of these methods has many advantages. They are cost-effective, they are based on natural processes, and they can be applied to reduce acidification originating from non-point pollution sources (e.g., AMD).

Conclusion

Lake and river acidification has been probably one of the most often discussed environmental issues in recent decades. Although natural acidic freshwaters occur throughout the world, the human-induced changes in water pH have aroused international concern, especially in countries where anthropogenic emissions of SO_2, NO_2, and NH_3 caused a substantial decrease in the freshwater pH due to their low ANC (Scandinavian countries, eastern Canada, northeastern United States). Acidic lakes and rivers have jeopardized aquatic biota, not only due to the low acid tolerance of living organisms but also because of the geochemical changes that have led to increased geo and bioavailability of toxic metals (Al, Zn, Mn, Pb, Hg, Cd) and reduction in the amount of bioavailable essential elements (Ca, Mg, P). The long-term international interest in the problem of acidification resulted in better understanding of the acidification processes and mechanisms and in well-established knowledge of the methods of monitoring and remediation and proper management of acidified rivers and lakes. As a result of international efforts in the abatement of SO_2 and NO_x emissions from industrial sources, many lakes and streams have already recovered from anthropogenic acidification. Both environmental legislation and technology have contributed to substantial improvement of the quality of surface inland waters. The lessons learned from almost 50 years of studies on acidification and its mitigation have shown that environmental pollution is a global problem, which can only be solved with cooperation of governments, scientists, and societies. Whether these lessons will protect us from similar problems in the future is an open question. Many authors claim that acidification will soon develop in rapidly developing countries of Asia, Africa, and Latin America. The experience gained from remediation of acidified lakes and rivers in the northern hemisphere may thus become invaluable for the global sustainable future.

References

1. Mattson, M.D. Acid lakes and rivers. In *Encyclopedia of Environmental Science*; Encyclopedia of Earth Sciences Series; Alexander, D.E., Fairbridge, R.W., Eds.; Kluwer Academic Publishers: Dordrecht, 1999; 6–9.
2. Norton, S.A.; Vesely J. *Acidification and Acid Rain*; Treatise on Geochemistry; Lollar B.S., Ed.; Pergamon: Oxford, 2005; 367–406.
3. Allen, D.J.; Castillo, M.M. *Stream Ecology: Structure and Function of Running Waters,* 2nd Ed.; Kluwer Academic Publishers: Boston, 2007.
4. O'Riordan, T. Critical load. In *Encyclopedia of Environmental Science*; Encyclopedia of Earth Sciences Series; Alexander, D.E., Fairbridge, R.W., Eds.; Kluwer Academic Publishers: Dordrecht, 1999; 102.
5. Elvingson, P.; Ågren, C. *Air and the Environment;* Swedish NGO Secretariat on Acid Rain: Goteborg, 2004.
6. Krzyzanowski, J.; Innes, J.L. Back to the basics—Estimating the sensitivity of freshwater to acidification using traditional approaches. J. Environ. Manage. **2010**, *91,* 1227–1236.
7. Hall, J.R.; Wright, S.M.; Sparks, T.H.; Ullyeti, J.; Allot, T.E.H.; Hornung, M. Predicting freshwater critical loads from national data on geology, soil and land use. Water Air Soil Pollut. **1995**, *85,* 2443–2448.

8. Battarbee, R.W.; Allot, T.E.H.; Juggins, S.; Kreiser, A.M.; Curtis, C.; Harriman, R. Critical loads of acidity to surface waters: An empirical diatom-based palaeolimnological model. Ambio **1995**, *25*, 366–369.
9. Battarbee, R.W.; Charles, B.F.; Bigler, C.; Cumming, B.F.; Renberg, I. Diatoms as indicators of surface-water acidity. In *The Diatoms: Aplications for the Environmental and Earth Sciences;* Smol, J., Stoermer, E., Eds.; Cambridge University Press: New York, 2010; 98–121.
10. Smith, R.A. On the air and rain of Manchester. Mem. Lit. Philos. Soc. Manchester **1852**, *10,* 207–217.
11. Gorham, E. The influence and importance of daily weather conditions in the supply of chloride, sulphate, and other ions to fresh waters from atmospheric precipitations. Philos. Trans. R. Soc. B **1958**, *247,* 147–178.
12. Odén, S. *The acidification of air and precipitation and its consequences in the natural environment*; Ecological Committee Bulletin No. 1. Swedish Natural Science Research Council, Stockholm. Translations Consultants, Ltd.: Arlington, Virginia, 1968.
13. Fleisher, S.; Andersson, G.; Brodin, Y.W.; Dickson, W.; Herrmann, J.; Muniz, I.P. Acid water research in Sweden— Knowledge for tomorrow? Ambio **1993**, *22*, 258–263.
14. Herrmann, J. Aluminum is harmful to benthic invertebrates in acidified waters, but at what threshold(s)? Water Air Soil Pollut. **2001**, *130,* 837–842.
15. Arseneau, K.M.A.; Driscoll, C.T.; Brager, L.M.; Ross K.A.; Cumming B.F. Recent evidence of biological recovery from acidification in the Adirondacks (New York, USA): A multiproxy paleolimnological investigation of Big Moose Lake. Can. J. Fish. Aquat. Sci. **2011**, *68,* 575–592.
16. Skjelkvále, B.L.; Evans, C.; Larssen, T.; Hindar, A.; Raddum, G.G. Recovery from acidification in European surface waters: A view to the future. Ambio **2003**, *32* (3), 170–175.
17. Helliwell, R.C.; Simpson, G.L. The present is the key to the past, but what does the future hold for the recovery of surface waters from acidification? Water Res. **2010**, *44* (10), 3166–3180.
18. Kuylenstierna, J.C.I.; Rodhe, H.; Cinderby, S.; Hicks, K. Acidification in developing countries: Ecosystem sensitivity and the critical load approach on a global scale. Ambio **2001**, *30* (1), 20–28.
19. Ye, X.; Hao, J.; Duan, L.; Zhou, Z. Acidification sensitivity and critical loads of acid deposition for surface waters in China. Sci. Total Environ. **2002**, *289,* 189–203.
20. Petrin, Z.; Englund, G.; Malmqvist B. Contrasting effects of anthropogenic and natural acidity in streams: A meta-analysis. Proc. R. Soc. B **2008**, *275*, 1143–1148.
21. Evans, C.D.; Cullen, J.M.; Alewell, C.; Kopácek, J.; Marchetto, A.; Moldan, F.; Prechtel, A.; Rogora, M.; Vesely, J.; Wright, R. Recovery from acidification in European surface waters. Hydrol. Earth Syst. Sci. **2001**, *5* (3), 283–297.
22. Löhr, A.J.; Bogaard, T.A.; Heikens, A.; Hendriks, M.R.; Sumarti, S.; Van Bergen, M.J.; Van Gestel, C.A.; Van Straalen, N.M.; Vroon, P.Z.; Widianarko, B. Natural pollution caused by the extremely acidic crater lake Kawah Ijen, East Java, Indonesia. Environ. Sci. Pollut. Res. Int. **2005**, *12* (2), 89–95.
23. Satake, K.; Oyagi, A.; Iwao, Y. Natural acidification of lakes and rivers in Japan: The ecosystem of Lake Usoriko (pH 3.4–3.8). Water Air Soil Pollut. **1995**, *85,* 511–516.
24. Nordstrom, D.K.; Alpers, C.N. Geochemistry of acid mine waters. In *The Environmental Geochemistry of Mineral Deposits*; Plumlee, G.S., Logsdon, M.J., Eds.; Rev. Econ. Geol. 6A: Littleton, CO, 1999; 133–160.
25. Pedrozo, F.; Kelly, L.; Diaz, M.; Temporetti, P.; Baffico, G.; Kringel, R.; Friese, K.; Mages, M.; Geller, W.; Woelfl, S. First results on the water chemistry, algae and trophic status of an Andean acidic lake system of volcanic origin in Patagonia (Lake Caviahue). Hydrobiologia **2001**, *452,* 129–137.
26. Takano, B.; Watanuki, K. Monitoring of volcanic eruptions at Yugama crater lake by aqueous sulfur oxyanions. J. Volcanol. Geotherm. Res. **1990**, *40,* 71–87.
27. Armienta, M.A.; De la Cruz-Reyna, S.; Macías, J.L. Chemical characteristics of the crater lakes of Popocatépetl, El Chichón, and Nevado de Toluca volcanoes, Mexico. J. Volcanol. Geotherm. Res. **2000**, *97,* 105–125.

28. Brantley, S.L.; Borgia, A.; Rowe, G.; Fernandez, J.F.; Reynolds, J.R. Poás volcano crater lake acts as a condenser for acid metal-rich brine. Nature **1987**, *330* (3), 470–472.
29. Chudaev, O.; Chudaeva, V.; Sugimori, K.; Kuno, A.; Matsuo, M. Geochemistry of recent hydrothermal systems of Mendeleev Volcano, Kuril Islands, Russia. J. Geochem. Explor. **2006**, *88*, 95–100.
30. April, R.; Newton, R. Influence of geology on lake acidification in the ILWAS watersheds. Water Air Soil Pollut. **1985**, *26*, 373–386.
31. Loukola-Ruskeeniemi, K.; Uutela, A.; Tenhola, M.; Paukola, T. Environmental impact of metalliferous black shales at Talvivaara in Finland, with indication of lake acidification 9000 years ago. J. Geochem. Explor. **1998**, *64* (1–3), 395–407.
32. Hindar, A.; Lydresen, E. Extreme acidification of lake in southern Norway caused by weathering of sulphide-containing bedrock. Water Air Soil Pollut. **1994**, *77*, 17–25.
33. Cameron, E.M.; Prévost, C.L.; McCurdy, M.; Hall, G.E.M.; Doidge, B. Recent (1930s) natural acidification and fish-kill in a lake that was an important food source for an Inuit community in northern Québec, Canada. J. Geochem. Explor. **1998**, *64,* 197–213.
34. Munk, L.A.; Faure, G.; Pride, D.E.; Bigham, J.M. Sorption of trace metals to an aluminum precipitate in a stream receiving acid rock-drainage; Snake River, Summit County, Colorado. Appl. Geochem. **2002**, *17* (4), 421–430.
35. Aguilera, A.; Manrubia, S.C.; Gómez, F.; Rodríguez, N.; Amils, R. Eukaryotic community distribution and its relationship to water physicochemical parameters in an extreme acidic environment, Río Tinto (Southwestern Spain). Appl. Environ. Microbiol. **2006**, *72* (8), 5325–5330.
36. Burton, E.D.; Bush, R.T.; Sullivan, L.A. Sedimentary iron geochemistry in acidic waterways associated with coastal lowland acid sulfate soils. Geochim. Cosmochim. Acta **2006**, *70*, 5455–5468.
37. Vuai, S.A.; Ishiki, M.; Tokuyama, A. Acidification of freshwaters by red soil in a subtropical silicate rock area, Okinawa, Japan. Limnology **2003**, *4*, 63–71.
38. Calmano, W.; von der Kammer, F.; Schwartz, R. Characterization of redox conditions in soils and sediments: Heavy metals. In *Soil and Sediment Remediation: Mechanisms, Technologies and Applications,* 1st Ed.; Lens, P., Grotenhuis, T., Malina, G., Tabak, H., Eds.; IWA Publishing: London, 2005; 102–120.
39. White, I.; Melville, M.D.; Wilson, B.P; Sammut, J. Reducing acidic discharges from coastal wetlands in eastern Australia. Wetlands Ecol. Manage. **1997**, *5*, 55–72.
40. Lin, C.; Wood, M.; Haskins, P.; Ryffel, T.; Lin, J. Controls on water acidification and de-oxygenation in an estuarine waterway, eastern Australia. Estuarine, Coastal Shelf Sci. *61* (1), 55–63.
41. Roos, M.; Åström, M. Seasonal and spatial variations in major and trace elements in a regulated boreal river (Esse River) affected by acid sulphate soils. River Res. Appl.2004 *21,* 351– 361.
42. Toivonen, J.; Österholm, P. Characterization of acid sulfate soils and assessing their impact on a humic boreal lake. J. Geochem. Explor. **2011**, *110* (2), 107–117, doi:10.1016/ j.gexplo.2011.04.003.
43. Johannesson, K.H.; Lyons, W.B. Rare-earth element geochemistry of Colour Lake, an acidic freshwater lake on Axel Heiberg Island, Northwest Territories, Canada. Chem. Geol. **1995**, *119,* 209– 223.
44. Rosenqvist, I. Alternative sources of acidification of river water in Norway. Sci. Total Environ. **1978**, *10,* 39– 49.
45. Larssen, T.; Holme, J. Afforestation, seasalt episodes and acidification—A paired catchment study in western Norway. Environ. Pollut. **2006**, *139,* 440– 450.
46. Emmett, B.A.; Reynolds, B.; Stevens, P.A.; Norris, D.A.; Hughes, S.; Gorres, J.; Lubrecht, I. Nitrate leaching from afforested Welsh catchment—Interactions between stand age and nitrogen deposition. Ambio **1993**, *22*, 386– 394.
47. Lydersen, E. Humus and acidification. In *Aquatic Humic Substances. Ecology and Biogeochemistry*; Hessen D.O., Tranvik L.J., Eds.; Springer: Berlin-Heidelberg, 1998; 63– 92.

48. Keskitalo, J.; Salonen, K.; Holopainen, A.-L. Long-term fluctuations in environmental conditions, plankton and macrophytes in a humic lake, Valkea-Kotinen. Boreal Environ. Res. **1998**, *3*, 251– 262.
49. Küchler, I.L.; Miekeley, N.; Forsberg, B.R. A contribution to the chemical characterization of rivers in the Rio Negro Basin, Brazil. J. Braz. Chem. Soc. **2000**, *3*, 286– 292.
50. Alkhatib, M.; Jennerjahn, T.C. Biogeochemistry of the Dumai River estuary, Sumatra, Indonesia, a tropical blackwater River. Limnol. Oceanogr. **2007**, *52* (6), 2410– 2417.
51. Kerekes, J.; Freedman, B.; Beauchamp, S.; Tordon, R. Physical and chemical characteristics of three oligotrophic lakes and their watersheds in Kejimkujik National Park, Nova Scotia. Water Air Soil Pollut. **1989**, *46* (1– 4), 99– 117.
52. Wu, J.-T.; Chang, S.-C.; Wang, Y.-S.; Wang, Y.-F.; Hsu, M.-K. Characteristics of the acidic environment of the Yuanyang Lake (Taiwan). Bot. Bull. Acad. Sin. **2001**, *42,* 17– 22.
53. Casper, P.; Chan, O.C.; Furtado, A.L.S.; Adams, D.D. Methane in an acidic bog lake: The influence of peat in the catchment on the biogeochemistry of methane. Aquat. Sci. **2003**, *65,* 36– 46.
54. Klavins, M.; Rodinov, V.; Druvietis, I. Aquatic chemistry and humic substances in bog lakes in Latvia. Boreal Environ. Res. **2003**, *8*, 113– 123.
55. Laudon, H. Recovery from episodic acidification delayed by drought and high sea salt deposition. Hydrol. Earth Syst. Sci. **2008**, *12,* 363– 370.
56. Rogora, M.; Mosello, R.; Marchetto, A. Long-term trends in the chemistry of atmospheric deposition in NW Italy: The role of increasing Saharan dust deposition. Tellus **2004**, *56B,* 426– 434.
57. Yan, N.D.; Keller, W.; Scully, N.M., Lean, D.R.S.; Dillon, P. Increased UV-B penetration in a lake owing to droughtinduced acidification. Nature **1996**, *381,* 141– 143.
58. Tranter, M.; Davies, T.D.; Wigington P.J.; Eshleman, K.N. Episodic acidification of freshwater systems in Canada— Physical and geochemical processes. Water Air Soil Pollut.1994 *72* (1– 4), 19– 39.
59. Holopainen, I.J. The effects of low pH on planktonic communities. Case history of a small forest pond in eastern Finland. Ann. Zool. Fenn. **1992**, *28,* 95– 103.
60. Tipping, E.; Smith, E.J.; Lawlor, A.J.; Hughes S.; Stevens P.A. Predicting the release of metals from ombrotrophic peat due to drought-induced acidification. Environ. Pollut. **2003**, *123* (2), 239– 253.
61. Stansley, W.; Cooper, G. An acidic snowmelt event in a New Jersey stream: Evidence of effects on an indigenous trout population Water Air Soil Pollut. **1990**, *53* (3– 4), 227– 237.
62. Fushimi, H.; Kawamura, T.; Iida, H.; Ochiai, M.; Nakajima, T.; Azuma, Y. Internal distribution of acid materials within snow crystals. Water Air Soil Pollut. **2001**, *130* (1– 4), 1709– 1714.
63. Hindar, A.; Henriksen, A.; Kaste, Ø.; TØrseth, K. Extreme acidification in small catchments in southwestern Norway associated with a sea salt episode. Water Air Soil Pollut.1994, *85* (2), 547–552.
64. Andersen, D.O.; Seip, H.M. Effects of a rainstorm high in sea-salts on labile inorganic aluminium in drainage from the acidified catchments of Lake Terjevann, southernmost Norway. J. Hydrol. **1999**, *224*, 64–79.
65. Teien, H.-C.; Salbu, B.; Heier, L.S.; Kroglund, F.R.; BjØrn, O. Fish mortality during sea salt episodes—Catchment liming as a countermeasure. J. Environ. Monit. **2005**, *7,* 989–998.
66. Bonjean, M.C.; Hutchins, M.; Neal, C. Acid episodes in the Allt a'Mharcaidh, Scotland: An investigation based on subhourly monitoring data and climatic patterns. Hydrol. Earth Syst. Sci. **2007**, *11* (1), 340–355.
67. Gorham, E. Acid precipitation and its influence upon aquatic ecosystems—An overview. Water Air Soil Pollut. **1976**, *6*, 457–481.
68. Psenner, R. Environmental impacts on freshwaters: Acidification as a global problem. Sci. Total Environ. **1994**, *143* (1), 53–61.
69. Galloway, J.N. Acid deposition: Perspectives in time and space. Water Air Soil Pollut. **1995**, *85,* 15–24.

70. *Lottermoser, B.G.* Mine wastes: Characterization, treatment and environmental impacts; *Springer: Berlin, Germany, 2007.*
71. Nordstrom, D.K.; Alpers, C.N.; Ptacek, C.J.; Blowers, D.W. Negative pH and extremely acid mine waters from Iron Mountain Superfund site, California. Environ. Sci. Technol. **2000**, *34* (2), 254–258.
72. Nordstrom, D.K.; Alpers, C.N. Negative pH, efflorescent mineralogy, and consequences for environmental restoration at the Iron Mountain Superfund site, California. Proc. Natl. Acad. Sci. U. S. A. **1999**, *96* (7), 3455–3462.
73. Charles, D.F.; Whitehead, D.R.; Engstrom, D.R.; Fry, B.D.; Hites, R.A.; Norton, S.A.; Owen, J.S.; Roll, L.A.; Schindler, S.C.; Smol, J.P.; Uutala, A.J.; White, J.R.; Wise, R.J. Paleolimnological evidence for recent acidification of Big Moose Lake, Adirondack Mountains, N.Y. (USA). Biogeochemistry **1987**, *3*, 267–296.
74. Nordstrom, D.K.; Alpers, C.N.; Coston, J.A.; Taylor, H.E.; McCleskey, R.B.; Ball, J.W.; Davis, J.A.; Ogle, S. Geochemistry, toxicity, and sorption properties of contaminated sediments and pore waters from two reservoirs receiving acid mine drainage. In Proceedings: Charleston, S.C., U.S. Geological Survey Water-Resources Investigations Report, 99–4018-A, Morganwalp, D.W., Buxton, H.T., Eds.; USGS: Denver, CO, 1999; 289–296.
75. Yan, N.D. Effects of changes in pH on transparency and thermal regimes of Lohi Lake, near Sudbury, Ontario. Can. J. Fish. Aquat. Sci. **1983**, *40* (5), 621–626.
76. Knöller, K.; Fauville, A.; Mayer, B.; Strauch, G.; Friese, K.; Veizer, J. Sulfur cycling in an acid mining lake and its vicinity in Lusatia, Germany. Chem. Geol. **2004**, *204* (3–4), 303–323.
77. Bonacina, C. Lake Orta: The undermining of an ecosystem. J. Limnol. **2001**, *60* (1), 53–59.
78. Renberg, I.; Hultberg, H. A paleolimnological assessment of acidification and liming effects on diatom assemblages in a Swedish lake. Can. J. Fish. Aquat. Sci. **1992**, *49* (1), 65–72.
79. Hesthagen, T. Fish kills of Atlantic salmon (*Salmo salar*) and brown trout (*Salmo trutta*) in an acidified river of SW Norway. Water Air Soil Pollut. **1986**, *30* (3–4), 619–628.
80. Dangles, O.J.; Guérold, F.A. Structural and functional responses of benthic macroinvertebrates to acid precipitation in two forested headwater streams (Vosges Mountains, northeastern France). Hydrobiologia 2000, *418,* 25–31.
81. Vesely, J.; Almquist-Jacobson, H.; Miller, L.M.; Norton, S.A.; Appleby, P.; Dixit, A.S.; Smol, J.P. The history and impact of air pollution at Certovo Lake, southwestern Czech Republic. J. Paleolimnol. **1993**, *8*, 211–231.
82. Kratz, T.K.; Webster, K.E.; Bowser, C.J.; Benson, B.J. Influence of landscape position on lakes in northern Wisconsin. Freshwater Biol. **1997**, *37,* 209–217.
83. Delleur, J.W. Hydrological response to acid rain. In *Monitoring to Detect Changes in Water Quality Series*, Proceedings of the Budapest Symposium, Budapest, Hungary July 1986; International Association of Hydrological Sciences Publication no. 157, Wallingford, U.K., 1986; 175–184.
84. Driscoll, C.; Lambert, K.F.; Chen, L. Acidic deposition: Sources and ecological effects. In *Acid in the Environment: Lessons Learned and Future Prospects*; Visgilio, G.R., Whitelaw, D.M., Eds.; Springer Science+Business Media: New York, 2007; 27–58.
85. Galloway, J.N. Acidification of the world: Natural and anthropogenic. Water Air Soil Pollut. **2001**, *130,* 17–24.
86. Wood, J.M. Effects of acidification on the mobility of metals and metalloids: An overview. Environ. Health Perspect. **1985**, *63,* 115–119.
87. Havas, M.; Rosseland, B.O. Response of zooplankton, benthos, and fish to acidification: An overview. Water Air Soil Pollut. **1995**, *85,* 51–62.
88. Kopàcek, J.; Ulrich, K.; Hejzlar, J.; Borovec, J.; Stuchlík, E. Natural inactivation of phosphorus by aluminum in atmospherically acidified water bodies. Water Res. **2001**, *35*, 3783–3790.
89. Scheuhammer, A.M. Effects of acidification on the availability of toxic metals and calcium to wild birds and mammals. Environ. Pollut. **1991**, *71* (2–4), 329–375.

90. Petrin, Z.; Laudon, H.; Malmqvist, B. Does freshwater macroinvertebrate diversity along a pH gradient reflect adaptation to low pH? Freshwater Biol. **2007**, *52,* 2172–2183.
91. Dangles, O.; Malmqvist, B.; Laudon, H. Naturally acid freshwater ecosystems are diverse and functional: evidence from boreal streams. Oikos **2004**, *104,* 149–155.
92. United Nations Economic Commission for Europe. Convention on Long-range Transboundary Air Pollution; United Nations: New York and Geneva, 1999.
93. Schöpp, W.; Posch, M.; Mylona, S.; Johansson, M. Longterm development of acid deposition (1880–2030) in sensitive freshwater regions in Europe. Hydrol. Earth Syst. Sci. **2003**, *7* (4), 436–446.
94. Chestnut, L.G.; Mills, D.M. A fresh look at the benefits and costs of the US acid rain program. J. Environ. Manage. **2005**, *77,* 252–266.
95. Henrikson, L.; Hindar, A.; Abrahamsson, I. Restoring acidified lakes: An overview. In *The Lakes Handbook. Volume 2: Lake Restoration and Rehabilitation;* O'Sullivan, P., Reynolds, C.S., Eds.; Blackwell Publishing: Oxford, 2005; 483–500.
96. Norberg, M.; Bigler, C.; Renberg, I. Monitoring compared with paleolimnology: Implications for the definition of reference condition in limed lakes in Sweden. Environ. Monit. Assess. **2008**, *146*, 295–308.
97. Wällstedt, T.; Borg, H. Effects of experimental acidification on mobilisation of metals from sediments of limed and nonlimed lakes. Environ. Pollut. **2003**, *126,* 381–391.
98. Aronsson, K.A.; Ekelund, N.G.A. Limnological effects on a first order stream after wood ash application to a boreal forest catchment in Bispgården, Sweden. For. Ecol. Manage. **2008**, *255*, 245– 253.
99. Ugochukwu, C.N.C.; Nukpezah, D. Ecotechnological methods as strategies to reduce eutrophication and acidification in lakes. Environmentalist **2008**, *28,* 137– 142.

9

Rivers: Pollution

Bogdan Skwarzec

Introduction

More than 97% of all water on Earth is salty, and most of the remaining 3% is frozen in the polar ice caps. The atmosphere, rivers, lakes, and underground stores hold less than 1% of all freshwater and this tiny amount has to provide the freshwater needed to support the earth's population. Most of the water for human consumption originates from riverine system. The quantity of this water is very important for human life because the concentration of chemical and biological pollutants in many rivers is high.[1,2] The big rivers of Asia are the most polluted rivers in the world. They contain three times as many bacteria from human waste as the global average and 20 times more lead than rivers in industrialized countries. Water from half of the tested sections of China's and Indian's seven major rivers was found to be undrinkable because of pollution.[3–7] Almost 40% of America's rivers are too polluted for fishing, swimming, or aquatic life. The Mississippi River—which drains nearly 40% of the continental United States, including its central farmlands—carries an estimated 1.5 million metric tons of nitrogen pollution into the Gulf of Mexico each year. The resulting hypoxic coastal dead zone in the Gulf each summer is about the size of Massachusetts.[8–12] Also, other major rivers in the world (Amazon, Niger, big Russian rivers, and Australian rivers) are polluted by trace elements and heavy metals.[13–18] The King River is Australia's most polluted river, suffering from a severe acidic condition related to mining operations.[19,20] In European Union countries, rivers are polluted with sewage or fertilizer. The Sarno is the most polluted river in Europe, featuring a nasty mix of sewage, untreated agricultural waste, industrial waste, and chemicals.[21]

Apart from anthropogenic pollutants, continental material from land is also transported to the marine ecosystem through rivers. The total flux of dissolved and suspended matter transported by rivers is estimated to be 20×10^{15} g yr^{-1}, i.e., 15.5×10^{15} and 4.5×10^{15} g yr^{-1} for solid and dissolved loads, respectively.[1] The total load of pollutants in the river ecosystems varies among several regions and depends on the population density, location of industry centers, and the abundance as well as intensity of the exploitation of natural resources. The anthropogenic sources of riverine pollutants are mainly industrial wastewater, leakage from products in use and those removed from service, natural degradation of pro-products, as well as pollution from different types of land use, e.g., fertilizing and mining.[22] Most freshwater pollution is caused by the addition of organic material, which is mainly sewage but can be food waste or farm effluent. Bacteria, viruses, or parasites cause diseases such as cholera, typhoid, schistosomiasis, dysentery, and other diarrheal diseases. Moreover, other microorganisms feed on organic matter and large populations quickly develop, using up much of the oxygen dissolved in the water.[23–27]

Normally, oxygen is present in high quantities, but even a small drop in the level can have a harmful effect on the river animals. Animals can be listed according to their ability to tolerate low levels of oxygen. In the following species of animals are listed in the order of the least tolerant to the most tolerant to low oxygen levels: stone-fly nymphs, mayfly nymphs, freshwater shrimps, freshwater hog lice, blood worms, tubifex worms, and rat-tailed maggots.[23]

Sources of River Pollution

The general important sources of water pollution that work together to reduce the overall river water quality are acid rain and industrial and agricultural discharge liquid waste products. Rain, as it falls through the air or drains from urban areas and farmland, absorbs contaminants. Serious incidents resulting from spillages or discharges of toxic chemicals are pollution events that make the news.[1]

Acid Rain

Rain, when falling through polluted air, absorbs some pollutants. The main pollutant gases are sulfur dioxide (SO_2) and nitrogen oxides (N_xO_y), which are formed when fuels are burned. They react with rainwater to form sulfuric and nitric acids. On reaching the ground, the acid liquid produces a number of effects. It can release harmful substances such as aluminum and heavy metals from the soil. These substances are normally present in an inert, harmless state; however, in acid conditions, they can turn into compounds that are poisonous to plant and animal life. When washed into lakes and streams, aluminum can kill small water creatures and fish. Particularly at risk is the dipper, a river bird that actually walks underwater to catch its insect food. If acid rain kills the insects, then the dippers will disappear in turn.[28,29]

Industrial Pollution

Factories use water from rivers to power or cool down machinery. Dirty water containing chemicals flows back in the river. Water used for cooling is warmer than the river itself. Raising the temperature of the water lowers the level of dissolved oxygen and upsets the balance of life in the water. Many industrial wastes discharged into river water are mixtures of chemicals that are difficult to treat. Some industrial wastes are so toxic that they are strictly controlled, making them an expensive problem to deal with. Some companies try to cut the costs of safely dealing with waste by illegally dumping chemicals at times and in places where they think they will not be caught.[30–32] Chemical waste products from industrial processes are sometimes accidentally discharged into rivers. Examples of such pollutants include cyanide and toxic heavy metals (zinc, lead, copper, cadmium, and mercury) as well as oil. These substances may enter the water in such high concentrations that fish and other animals are killed immediately. Sometimes, the pollutants enter a food chain and accumulate until they reach toxic levels, eventually killing birds, fish, and mammals.[1,33,34]

Agricultural Pollution

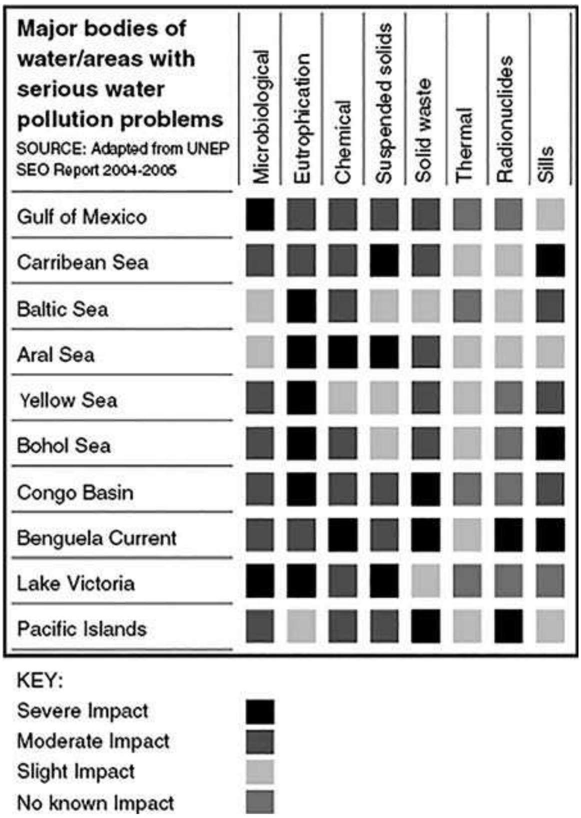
Farmers use fertilizers and pesticides to enhance crop growth. However, these fertilizers and pesticides can be washed through the soil by rain, to end up in rivers. If large amounts of fertilizers or farm waste drain into a river, the concentration of nitrate and phosphate in the water increases considerably.[35] Algae use these substances to grow and multiply rapidly, turning the water green. This massive growth of algae, called eutrophication, leads to pollution. When the algae die, they are broken down by the action of bacteria, which quickly multiply, using up all the oxygen in the water, leading to the death of many animals. Also, when organic farm wastes such as silage or liquid manure (slurry) escape into rivers, the amount of oxygen in the water is reduced.[36,37] Nitrate pollution problems occur when too much chemical fertilizer is applied to the land. The excess runs off and can find its way into drinking water sources, or can trickle into rivers and lakes. Some experts believe that high levels of nitrate in drinking water may pose a threat to health. A European directive states that drinking water should not contain more than 50 mg of nitrates per liter of water. In rivers, streams, ponds, and lakes, too much nitrate can create a “pea soup” effect. The water becomes clogged with fast-growing plant life such as algae and weeds.[36–39]

Radionuclides

Radionuclides and radioactive waste are stored in riverine water ponds next to nuclear power stations. Nuclear atmospheric weapons tests and the Chernobyl accident were sources of many artificial radionuclides in oceans.[40–44] The river runoff of radioactive substances is an important source of marine radionuclide pollution, especially for closed seas, estuaries, and coastal zones. For example, the annual runoff of radiocesium, radiostrontium, polonium, uranium, and plutonium is an important source of radionuclides in southern Baltic Sea. The total runoff of these radionuclides from rivers to the southern Baltic Sea was calculated to consist of about 3.3 TBq of ^{90}Sr, 5.2 TBq of ^{137}Cs, 0.7 TBq of ^{210}Po, 4.5 TBq of $^{234+238}$U, and 1.3 GBq of $^{239+240}$Pu.[45]

Removal of Pollutants in River Waters

Many types of pollution are discharged into rivers, and purification processes remove them at different speeds. Some heavy metals, for example, are removed relatively quickly because suspended clay and



Major bodies of water/areas with serious water pollution problems SOURCE: Adapted from UNEP SEO Report 2004-2005	Microbiological	Eutrophication	Chemical	Suspended solids	Solid waste	Thermal	Radionuclides	Sills
Gulf of Mexico	■	▩	▩	▩	▩	▨	▨	□
Carribean Sea	▩	▩	▩	■	▩	□	□	■
Baltic Sea	□	■	▩	□	□	▨	□	▩
Aral Sea	□	■	■	■	▩	□	□	□
Yellow Sea	▩	■	□	□	▩	□	▨	▩
Bohol Sea	▩	■	▩	□	▩	□	▨	■
Congo Basin	▩	■	▩	▩	■	▨	▨	▩
Benguela Current	▩	▩	■	▩	■	□	■	■
Lake Victoria	■	■	▩	■	□	▨	▨	▨
Pacific Islands	▩	□	▩	▩	■	□	■	□

KEY:

Severe Impact	■
Moderate Impact	▩
Slight Impact	□
No known Impact	▨

FIGURE 1 Impact of inflow of pollutants from riverine systems on ecological conditions of several coastal zones, estuaries, and closed seas.

Source: Grinning Planet, Water Pollution Facts, 26th July 2005, http://www.grinningplanet.com/2005/07-26/water-pollution-facts-article.htm.[46]

organic particles have a slight electric charge and adsorb the metal atoms. When the clay or organic particles settle out of the water, they take the metal atoms with them. Unfortunately, some pollutants are very persistent in the water and can accumulate downstream, causing great hazard.[1] Suspended solids in a moving body of water will settle out at various points or be carried within longer distances, depending on their size and the rate of the flow. The higher the amount of suspended solids is, the cloudier or more turbid the water becomes. Suspended matter can affect the amount of light entering the water, thus restricting the amount of photosynthesis that can occur, which consequently can limit the growth of plants. Small particles settling out in large amounts on the bottom of a water body can prevent some organisms from living there as well as prevent green plants from photosynthesizing.[19,21]

Conclusions

Rivers are contaminated by many natural and anthropogenic industrial pollutants. A lot of these pollutants are highly toxic to river organisms and to humans. The most important toxic chemicals in the seas are organic compounds (PCBs, DDT, PAHs, pesticides, furans, dioxins, phenol, oil), heavy metals, and radionuclides. Chemical and biological pollutants flow from big areas of river catchments to the seas and oceans. This runoff of toxic substances through river water is very dangerous, especially for coastal and estuarine zones and closed small seas. Figure 1 presents 10 of the most polluted areas and bodies of water in the world.[46]

References

1. Martin, J.-M.; Meybeck, M. The significance of the river input of chemical elements to the ocean. In *Trace Metals in Sea Water*; Wong, C.S., Boyle, E., Bruland, K.W., Burton, J.D., Goldberg, E.D., Eds.; Plenum Publishing Corporation, New York, 1983; 265–296
2. Martin, J.-M.; Meybeck, M. Elemental mass balance of material carried by major rivers. Mar. Chem. **1979**, *7*, 173–206.
3. Li, Y.-H.; Teraoka, H.; Yang, T.-S.; Chen, J.-S. The elemental composition of suspended particles from the Yellow and Yangtze rivers. Geochim. Cosmochim. Acta **1984**, *48,* 1561–1564.
4. Bhatt, S.D.; Pathak, J.K. Assessment of water quality and aspect of pollution in a stretch of river Gomati (Kumaun: Lesser Himalaya). J. Environ. Biol. **1992**, *13* (2), 113–126.
5. Tanizaki, Y.; Nagatasuka, S. Distribution and chemical behavior of 21 elements in river water in the Yakushima Island. Bull. Chem. Soc. Jpn. **1983**, *56,* 619–624.
6. Agarwal, H.C.; Mittal, P.K.; Menon, K.B.; Pillal, M.K.K. DTT residues in the River Jumuna in Delhi, India. Water Air Soil Pollut. **1986**, *28,* 89–104.
7. Bishwanath, G.; Nandini, B. Impact of informal regulation of pollution on water quality in rivers in India. J. Environ. Manage. **2004**, *73*, 117–130.
8. Newchurch, E.J.; Kahwa, I. A. Heavy metals in the Lower Mississippi River. J. Environ. Sci. Health **1984**, *19* (8), 973–988.
9. Presley, B.J.; Trefry, J.H. Heavy metal inputs to Mississippi delta sediments—A historical view. Water Air Soil Pollut. **1980**, *13* (4), 481–494.
10. Taylor, H.E.; Garbarino, J.R.; Brinton, T.I. The occurrence and distribution of trace metals in the Mississippi River and its tributaries. Sci. Total Environ. **1990**, *97/98,* 369–384.
11. Taylor, H.E.; Shiller, A.M. Mississippi River methods comparison study—Implications for water quality monitoring of dissolved trace elements. Environ. Sci. Technol. **1995**, *29,* 1313–1317.
12. Brinton, T.I.; Garbarino, J.R.; Peart, D.B.; Taylor, H.E.; Antweiler, R.C. Concentration and transport data for dissolved inorganic constituents in water collected during seven cruises on the Mississippi River and some of its tributaries, July 1987–June 1990, U.S. Geological Survey Open-File Report 94–524, 102, 1995.

13. Sholkovitz, E.R.; Price, N.B. The major element chemistry of suspended matter in the Amazon estuary. Geochim. Cosmochim. Acta **1980**, *44*, 163–171.
14. Sholkovitz, E.R. The geochemistry of rare earth elements in the Amazon River estuary. Geochim. Cosmochim. Acta **1993**, *57,* 2181–2190.
15. Vital, H.; Stattegger, K. Major and trace elements of stream sediments from the lowermost Amazon River. Chem. Geol. **2000**, *168,* 151–168.
16. Olajire, A.A.; Altenberger, R.; Kuster E.; Brack W. Chemical and ecotoxicological assessment of polycyclic aromatic hydrocarbon-contaminated sediment of the Niger Delta, Southern Nigeria. Sci. Total Environ. **2005**, *340,* 123–136.
17. Queirolo, F.; Stegen, S.; Mondaca, J.; Cortes, R.; Rojas, R.; Contreras, C.; Munoz, L.; Schwuger, M.J.; Ostapczuk, P. Total arsenic, lead, cadmium, copper and zinc in some salt rivers in the northern Andes of Antofagasta, Chile. Sci. Total Environ. **2000**, *255*, 85–95.
18. Orr, J.C.; Fabry, V.J.; Aumont, O.; Bopp, L.; Doney, S.C.; Feely, R.A.; Gnanadesikian, A.; Gruber, N.; Ishida, A.; Joos, F.; Key, R.M.; Lindsay, K.; Maier-Reimer, E.; Matear, R.; Monfray, Mouchet, A.; Najjar, R.G.; Plattner, G.K.; Rodgers, K.B.; Sabine, C.L.; Sarminto, J.L.; Schlitzer, Slater, R.D.; Totterdell, I.J.; Weirig, M.F.; Yamanaka, Y.; Yool, A., Anthropogenic ocean acidification over the twenty-first century and its impact on calcifying organisms. Nature **2005**, *437* (7059), 681–686.
19. Likens, G.E.; Keene, W.C.; Miller J.M; Galloway, J.N. Chemistry of precipitation from a remote, terrestrial site in Australia. J. Geophys. Res. **1987**, *92* (D11), 13,299–13,314.
20. Trefry, J.H.; Nelson, T.A.; Tocine, R.P.; Metz, S.; Vetter, T.W. Trace metal fluxes through the Mississippi River delta system. In *Contaminant Fluxes through the Coastal Zone;* Kullenberg, Ed.; Rapports et Proces-Verbaux des Reunions; Conseil International pour l'Exploration de la Mer, Copenhague, Denmark, 1986; Vol. 186, 277–288.
21. Smith, R.A.; Alexander, R.B.; Wolman, M.G. Water-quality trends in the Nation's rivers. Science **1987**, *235,* 1607–1615.
22. Meybeck, M. Atmospheric inputs and river transport of dissolved substances. Int. Assoc. Hydrol. Sci. Publ. **1983**, *141,* 173–192.
23. Meybeck, M.; Helmer, R. The water quality of rivers: From pristine stage to the global pollution. Palaeogeogr. Palaeoclimatol. Palaeoecol. (Global Planet. Change Sect.) **1989**, *75*, 283–309.
24. Morrison, T.E.; Fraser, R.J.; Smith, Pn.N.; Mahalingam,S.; Heise, M.T. Complement contributes to inflammatory tissue destruction in a mouse model of Ross River virus-induced disease. J. Virol. **2007**, *81* (10), 5132–43.
25. Harley, D.; Sleigh, A.; Ritchie S. Ross River virus transmission, infection, and disease: A cross-disciplinary review. Clin. Microbiol. Rev. **2001**, *14* (4), 909–932.
26. Knight, J. Meet the Herod bug. Nature **2001**, *412,* 12–14.
27. Breault, J.L. Candiru: Amazonian parasitic catfish. J. Wilderness Med. **1998**, *2* (4), 304–312.
28. Likens, G.E.; Driscoll, C.T.; Buso, D.C. Long-term effects of acid rain: Response and recovery of a forest ecosystem. Science **1996**, *272*, 244–246.
29. Galloway, J.N., Zhao, D.; Xiong, J.; Likens, G.E. Acid rain: A comparison of China, United States and a remote area. Science **1987**, *236,* 1559–1562.
30. Szefer, P. *Metals, Metalloids and Radionuclides in the Baltic Sea Ecosystem*; Elsevier: Amsterdam, the Netherlands, 2002; Vol. 5.
31. Laws, E.A. *Aquatic Pollution;* John Wiley and Sons, 2000.
32. Bruneau, L. Pollution from industries in the drainage area of the Baltic. Ambio **1980**, *9,* 145–152.
33. Billen, G.; Garnier, J.; Delinge, C.; Billen, C. Estimate of early-industrial inputs of nutrients to river system: Implication for coastal eutrophication. Sci. Total Environ. **1999**, *243/244*, 43–52.
34. Gribble, G.W. The Natural production of chlorinated compounds. Environ. Sci. Technol. **1994**, *28,* 310–319.

35. Horrigan, L.; Lawrence, R.S.; Walker, P. How sustainable agriculture can address the environmental and human health harms of industrial agriculture. Environ. Health Perspect. **2002**, *110* (5), 445–456.
36. Ambio. Special issues: Marine eutrophication, Ambio **1990**, *19*, 101–176.
37. Smith, V.H.; Tilman, G.D.; Nekola, J.C. Eutrophication: Impacts of excess nutrient inputs on freshwater, marine, and terrestrial ecosystems. Environ. Pollut. **1999**, *100* (1–3), 179–196.
38. Raimammake, A.; Pietilainen, O.P.; Rekolainen, S.; Kauppila, P.; Pitkanen, H.; Niemi, J.; Raateland, A.; Vuorenmaa, J. Trends of phosphorus, nitrogen, and chlorophyll *a* concentrations in Finnish rivers and lakes in 1975–2000. Sci. Total Environ. **2003**, *310*, 47–59.
39. Cole, J.J.; Peierls, B.L.; Caraco, N.F.; Pace, M.L. Nitrogen loading of rivers as a human-driven process. In *Humans as Components of Ecosystems*, McDonnell, Pickett, S.T.A., Eds.; Springer-Verlag: New York, 1993; 141–157.
40. Skwarzec, B. Radiochemia srodowiska i ochrona radiologiczna (in Polish), Environmental radiochemistry and radiological protection, Wydawnictwo DJ sc., Gdańsk, 2002.
41. More, W.S. Amazon and Mississippi River concentration of uranium, thorium, and radium isotopes. Earth Planet. Sci. Lett. **1967**, *2*, 231–234.
42. Spalding, R.F.; Sackett, W.M. Uranium in runoff from the Gulf of Mexico distributive province: Anomalous concentrations. Science **1972**, *175*, 629–631.
43. Sayles, F.; Livingston, H.; Panteleyev, G. The history and source of particulate ^{137}Cs and $^{239+240}$Pu, deposition in sediment of the Ob River Delta, Siberia. Sci. Total Environ. **1997**, *202*, 25–41.
44. Johnson-Pyrtle, A.; Scott, M.R.; Laing, T.E.; Smol, J.P. ^{137}Cs distribution and geochemistry of Lena River (Siberia) drainage basin lake sediments. Sci. Total Environ. **2000**, *255*, 145–159.
45. Skwarzec, B. Polonium, uranium and plutonium radionuclides in aquatic environment of Poland and southern Baltic. Baltic Coastal Zone **2009**, *13*, 127–166.
46. Grinning Planet, Water Pollution Facts, 26th July 2005, http://www.grinningplanet.com/2005/07-26/waterpollution-facts-article.htm.

10

Sea: Pollution

Bogdan Skwarzec

Introduction

Pollution is the release of harmful environmental contaminants, or the substances so released. Generally, the process needs to result from human activity to be regarded as pollution. Even relatively benign products of human activity are liable to be regarded as pollution, if they precipitate negative effects later on. The different forms of pollution are as follows: air pollution, water pollution, soil contamination, radioactive contamination, noise pollution, light pollution, and thermal pollution. Chemical pollutants can be natural substances (organic and mineral matter), or synthetic degradable or non-degradable substances (plastics, pesticides). They can be toxic or not with varying endurance. For example, in humans, polychlorinated biphenyls (PCBs) play a part in causing breast, lung, liver, and colon cancers. It can cause retardations in neurological development and growth. Persistent organic pollutants, drugs and their metabolites, and brominated organic compounds are concentrated in soft tissue of marine organisms and other filters, and they can develop typhoid, hepatitis, and other illnesses if they are ingested.[1–3]

Marine pollution occurs when harmful effects, or potentially harmful effects, can result from the entry into the ocean system of chemical, industrial, agricultural, and residential wastes, as well as invasive organisms. Marine pollution is, in the majority of cases, of land and atmospheric origin. Catastrophic accidents (shipping, platform) are also an important factor of marine pollution. Generally, sea pollutants are classified as having chemical and biological origin; however, toxic chemical substances are dominant. The most important chemical substances belong to the next group: trace and heavy toxic metals, metalo-organic compounds (tributyltin, TBT), nutrients, acid gases, radionuclides (especially artificial), radioactive waste, military toxic substances, and organic toxic substances (oil, pesticides, dioxins, furans, phenols, halogeno-organic compounds—especially DDT, PCBs, polychlorinated terphenyls [PCT], and their metabolites, as well as aromatic hydrocarbons polycyclic aromatic hydrocarbon, [PAHs]). Toxins can accumulate in the tissues of many species of aquatic life in a process called bioaccumulation.[1–3]

Many potentially toxic chemicals adhere to tiny particles that are taken up by marine organisms (plankton and benthos animals), most of which are either deposit or filter feeders (bivalvia). In this way, by bioconcentration of dissolved species, the toxins are concentrated upward within ocean food chains. When pesticides are incorporated into the marine ecosystem, they quickly become absorbed into marine food webs. Once in the food webs, these pesticides can cause mutations, as well as diseases, which can be harmful to humans as well as the entire food web. Toxic metals, metalloids, and radionuclides can also be introduced into marine food webs. These can cause a change to tissue matter, biochemistry, behavior, reproduction, and suppress growth in marine life. In addition, many animal feeds have a high fish meal or fish hydrolysate content. In this way, marine toxins can be transferred to land animals, and appear later in meat and dairy products. Pollution in marine environment is often classified as point source or non-point source pollution. Point source pollution occurs when there is a single, identifiable, and localized source of the pollution. An example is a direct discharge of sewage and industrial waste into the ocean. Pollution such as this occurs particularly in developing nations. Non-point source pollution occurs when the pollution comes from ill-defined and diffuse sources.[4,5]

Sources of Marine Pollution

There are many different ways to categorize and examine the inputs of pollution into our marine ecosystems. In principle, there are three main types of inputs of pollution into the ocean: direct discharge of waste into the oceans, runoff into the waters due to rain, and pollutants that are released from the atmosphere.[1,6]

Land and River Runoffs

Between 75% and 80% of marine pollution is caused by land, particularly agriculture, and 30% of this is from the atmosphere; however, around 12% of the pollution is caused by marine transport.[7] In South America, 98% of domestic wastewater ends up, untreated, in the sea. The countries along the Mediterranean Sea throw 50 million tons waste into it every year, and the Chinese throw 60 million tons of waste into the Yellow Sea daily.[8,9] Also, closed marine basins, such as the Baltic, Caspian, and Black Seas, are polluted every year with many million tons of chemical waste. Baltic fauna is practically decimated on half of the central Baltic's seabed as an effect of land-based pollution, especially nitrogenous.[1] In the Caspian Sea, 140 million tons of pollutants are poured every year in the northern part, especially that of the Volga delta. Also, the Black Sea is polluted by the Dniestr, Dniepr, Danube, and Don rivers. For eutrophication reason, 80% drop in fishing yield, only 6 species of fish can still be fished out of the 26 that could previously be commonly marked in this basin.[3] One important common path of entry by land and atmosphere contaminants to the sea are rivers.[9] The evaporation of water from oceans exceeds precipitation. The balance is restored by rain over the continents entering rivers and then being returned to the sea. Pollutants enter rivers and the sea directly from urban sewerage and industrial waste discharges, sometimes in the form of hazardous and toxic wastes (inorganic, radioactive, and organic). Most of the pollution is simply soil, which ends up in rivers flowing to the sea. More than 80% of all marine pollution comes from land-based activities, and many pollutants are deposited in estuarine and coastal waters. Here, the pollutants enter marine food chains, building up their concentrations until they reach toxic levels. It often takes human casualties to alert us to pollution and such was the case in Minamata Bay in Japan, when 649 people died from eating fish and shellfish contaminated with mercury and 3500 people had mercury poisoning. A factory was discharging waste containing mercury in low concentration into the sea and, as this pollutant passed through food chains, it became more concentrated in the tissues of marine organisms until it reached toxic levels.[10]

Surface water runoff contains pollutants from farming. Also, urban runoff and runoff from the construction of roads, buildings, ports, channels, and harbors can carry soil and particles laden with carbon, nitrogen, phosphorus, and minerals. This nutrient-rich water can lead to fleshy algae and phytoplankton thriving in coastal areas, known as algal blooms, which have the potential to create hypoxic conditions by using all available oxygen. Polluted runoff from roads and highways can be a significant source of water pollution in coastal areas.[11]

Oil and Ship Pollution

About 2.7 million tons of oil pollution enters our oceans each year. Of this, less than 10% is from natural seepage of oil from the ocean floor and sedimentary rock erosion. The remaining 90% comes from human activities. Offshore drilling, as a result of accident spills and other operations, accounts for only 2%. Spills from large tankers, which are reported worldwide, account for just over 5%.[3] Air pollution from cars and industry accounts for just over 13% of the total, as the hundreds of tons of hydrocarbons land in our oceans from particle fallout aided by rains, which washed the particles from air. Almost 4 times the amount of oil that comes from spills from large tankers spills, 19% is regularly released into the ocean from routine maintenance, which includes discharge as well as other ship operations. Oil is perhaps the most publicly recognized toxic pollutant. Tanker accidents between 1967 and 2007 spilled to the oceans nearly 4.5 million tons of oil. Now, from May 2010, the next million tons of oil has been delivered to the Gulf of Mexico, as a result of platform crash (BP accident). The biggest spill ever recorded took place during the 1979 drilling platform in the Gulf of Mexico as well as the 1991 Persian Gulf War, when about 2 million tons of oil was spilled. The effect of oil pollution on wildlife can be terrible. Many marine organisms (fishes, invertebrates, mammals, and birds) died.[3]

Ships can pollute waterways and oceans in many ways. Oil spills can have devastating effects. While being toxic to marine life, PAHs, the components in crude oil, are very difficult to clean up and last for years in the sediment and marine environment.[1,2] Discharge of cargo residues from bulk carriers can pollute ports, waterways, and oceans. In many instances, vessels intentionally discharge illegal waste despite foreign and domestic regulations prohibiting such actions. It has been estimated that container ships lose over 10,000 containers at sea each year (usually during storms). Ships also create noise pollution that disturbs natural wildlife, and water from ballast tanks can spread harmful algae and other invasive species. Also, ballast water taken up at sea and released in port is a major source of unwanted exotic marine life. The invasive marine species, native to the Black, Caspian, and Azov Seas, were transported to other basins (Baltic Sea, Great Lakes) via ballast water from a transoceanic vessel. Invasive species can take over once occupied areas, facilitate the spread of new diseases, introduce new genetic material, alter underwater seascapes, and jeopardize the ability of native species to obtain food.[12] Some strong toxic metalo-organic substances, especially TBT, are a source of marine pollution, as a result of commercial shipping.[13–16]

Atmospheric Pollution

Another important pathway of pollution occurs through the atmosphere.[17–21] Wind-blown dust and debris, including plastic bags, are blown seaward from landfills and other areas. Dust from the Sahara, moving around the southern periphery of the subtropical ridge, moves into the Caribbean and Florida during the warm season as the ridge builds and moves northward through the subtropical Atlantic. Dust can also be attributed to a global transport from the Gobi and Taklamakan deserts across Korea, Japan, and the Northern Pacific to the Hawaiian Islands.[22] Since 1970, dust outbreaks have worsened due to periods of drought in Africa. There is a large variability in dust transport to the Caribbean and Florida from year to year; however, the flux is greater during positive phases of the North Atlantic Oscillation.[23] Atmospheric particles containing chemical pollutants can be transported over thousands of kilometers, and airborne particles with heavy metals and radionuclides originating from anthropogenic sources can be detected even in remote areas.[24–26] The trace and heavy metals produced and emitted in Europe or North America are transported by air and deposited in the North Atlantic environment.[27,28] Also, organic toxic substances (especially pesticides) used in agriculture by farmers in warm regions (Africa and South Asia) are sublimated to air and transported to Arctic Ocean and accumulated in fish and mammals.[29]

Deep Sea Mining

Deep sea mining is a relatively new mineral retrieval process that takes place on the ocean floor; however, the complete consequences of full-scale mining operations are still unknown. Ocean mining sites are located usually around large areas of polymetallic nodules or active and extinct hydrothermal vents at about 1400–3700 m below the ocean's surface.[30] The vents create sulfide deposits, which contain precious metals such as silver, gold, copper, manganese, cobalt, and zinc.[31,32] The deposits are mined using either hydraulic pumps or bucket systems that take ore to the surface to be processed. As with all mining operations, deep sea mining raises questions about environmental damage to the surrounding areas. The removal of parts of the seafloor will result in disturbances to the benthic layer, increased toxicity of the water column, and sediment plumes from tailings. Removing parts of the seafloor disturbs the habitat of benthic organisms, possibly depending on the type of mining and location, causing permanent disturbances.[33] Among the deep sea mining products, sediment plumes could have the greatest impact. Plumes are caused when the tailings from mining (usually fine particles) are dumped back into the ocean, creating a cloud of particles floating in the water. Two types of plumes occur: near-bottom plumes and surface plumes.[30] Near-bottom plumes occur when the tailings are pumped back down to the mining site. The floating particles increase the turbidity, or cloudiness, of the water, clogging filter-feeding apparatuses used by benthic organisms. Surface plumes cause a more serious problem. Depending on the size of the particles and water currents, the plumes could spread over vast areas. The plumes could affect zooplankton and light penetration, in turn affecting the food web of the area.[30,34]

Acidification and Acid Rains

The ocean waters absorb carbon dioxide (CO_2), sulfur dioxide (SO_2) and trioxide (SO_3), and nitrogen oxides (N_xO_y) from the atmosphere. Because the atmospheric concentrations of these gaseous oxides are increasing, the oceans are becoming more acidic.[35] The potential consequences of ocean acidification are not fully understood; however, there are concerns that structures made of calcium carbonate may become vulnerable to dissolution, affecting corals and the ability of shellfish to form shells.[36,37] Oceans and coastal ecosystems play an important role in the global carbon cycle and have removed about 25% of the carbon dioxide emitted by human activities between 2000 and 2007. Rising ocean temperatures and ocean acidification means that the capacity of the ocean carbon sink will gradually become weaker. Also, the methane clathrate reservoirs, containing large amounts of the greenhouse gas methane, under sediments on the ocean floors, can potentially release the methane when oceanic water becomes warm. In 2004, the global inventory of ocean methane clathrates was estimated to occupy between 1 and 5 million cubic kilometers. This estimate corresponds to 500–2500 gigatonnes carbon (Gt C), and can be compared with the 5000 Gt C estimated for all other fossil fuel reserves.[38]

Radioactive Waste

Radionuclide and radioactive wastes are stored in marine water ponds next to nuclear power stations (electricity nuclear power and nuclear reprocessing factory) until it is considered safe for disposal. The waste is released directly into the sea about some kilometers from the coast. The ocean depth is used to dump high-level nuclear waste.[39,40] The liquid waste is sealed in glass, a process called vitrification, and stored steel canisters contain concrete. These containers are dumped in the sediment on the ocean floor. Also, nuclear atmospheric weapons test and the Chernobyl accident were sources of many artificial radionuclides in oceans.[24] Natural radionuclides in the environment can also be enhanced owing to human activity: industry, coal power, phosphate fertilizers in agriculture, and domestic and industrial sewage.[39] Moreover, coal mining is a source of huge amounts of waste containing large quantities of natural radionuclides, especially polonium, thorium, and uranium. During ashing of coal in power plants, some natural radionuclides are emitted to the atmosphere as a gas and radioactive dust, whereas the others stay as concentrated ash.[41,42] Also during production of phosphate fertilizers, about 10% of the initial of ^{226}Ra, 20% of uranium, and about 85% of ^{210}Po is found in the phosphogypsum waste.[43] Phosphogypsum wastes are often located near coasts of seas and enhance the concentration of soil groundwater and river water. For this reason, the surroundings of phosphogypsum waste are strongly polluted by natural radionuclides.[44,45] The runoff of radioactive substances are important sources of marine radionuclides pollution, especially for closed seas. For example, the total runoff of natural and artificial radionuclides from rivers to the southern Baltic Sea was calculated to comprise about 3.3 TBq of ^{90}Sr, 5.2 TBq of ^{137}Cs, 0.7 TBq of ^{210}Po, 4.5 TBq of $^{234+238}$U, and 1.3 GBq of $^{239+240}$Pu.[46] In turn, some basins of the North Sea (Irish Sea and French Channel) are strongly contaminated with artificial radionuclides (especially, radiocesium and plutonium) emitted from European nuclear reprocessing facilities in Sellafield and Cap de la Hague.[40] Radionuclides in the marine environment are strongly accumulated by biota, and the values of bioaccumulation factor for some radioactive elements (polonium, plutonium, americium) in sea algae, benthic animals, and fish are more than 5000.[47] Transuranic elements belong to the group of pollutants caused by human activity and are important from the radiological point of view owing to their high radiotoxicity, long physical lifetime, high chemical reactivity, and long residence in biological systems in the marine environment.[48]

Eutrophication

Eutrophication is an increase in chemical nutrients, typically compounds containing nitrogen or phosphorus, in an ecosystem. It can result in an increase in the ecosystem's primary productivity (excessive plant growth and decay), and further effects, including lack of oxygen and severe reductions in water quality and in fish and other animal populations.[11] The biggest culprits are rivers that empty into the ocean, and along with it many chemicals used as fertilizers in agriculture as well as waste from livestock and humans. An excess of oxygen-depleting chemicals in the water can lead to hypoxia and the creation of a dead zone.[49–51] Estuaries tend to be naturally eutrophic because land-derived nutrients are concentrated where runoff enters the marine environment in a confined channel. The World Resources Institute has identified 375 hypoxic coastal zones around the world, concentrated in coastal areas in Western Europe, the eastern and southern coasts of the United States, and East Asia, particularly in Japan. In the ocean, there are frequent red tide algae blooms that kill fish and marine mammals and cause respiratory problems in humans and some domestic animals when the blooms reach close to shore.[52–54] In addition to land runoff, atmospheric anthropogenic nitrogen can enter the open ocean. A study in 2008 found that this could account for around one-third of the ocean's external (non-recycled) nitrogen supply and up to 3% of the annual new marine biological production. It has been suggested that accumulating reactive nitrogen in the environment may have consequences as serious as putting carbon dioxide in the atmosphere.[54]

Plastic Debris

Marine debris is mainly discarded human rubbish that floats on, or is suspended in the ocean. Eighty percent of marine debris is plastic—a component that has been rapidly accumulating since the end of World War II. The mass of plastic in the oceans may be as high as 100 million metric tons. Discarded plastic bags, six-pack rings, and other forms of plastic wastes that finish up in the ocean constitute dangers to wildlife and fisheries. Aquatic life can be threatened through entanglement, suffocation, and ingestion. Fishing nets, usually made of plastic, can be left or lost in the ocean by fishermen. Known as ghost nets, they entangle fish, dolphins, sea turtles, sharks, dugongs, crocodiles, seabirds, crabs, and other creatures, restricting movement, causing starvation, laceration and infection, and, in those that need to return to the surface to breathe, suffocation. Many animals that live on or in the sea consume flotsam by mistake, as it often looks similar to their natural prey. Plastic debris, when bulky or tangled, is difficult to pass, and may become permanently lodged in the digestive tracts of these animals, blocking the passage of food and causing death through starvation or infection.[55] Plastics accumulate because they do not biodegrade in the way many other substances do. They will photodegrade on exposure to the sun, but they do so properly only under dry conditions, and water inhibits this process. In marine environments, photodegraded plastic disintegrates into ever smaller pieces while remaining polymers, even down to the molecular level. When floating plastic particles photodegrade down to zooplankton sizes, jellyfish attempt to consume them, and in this way the plastic enters the ocean food chain. Many of these long-lasting pieces end up in the stomachs of marine birds and animals, including sea turtles and black-footed albatross.[56,57] Plastic debris tends to accumulate at the center of ocean gyres. In particular, the Great Pacific Garbage Patch has a very high level of plastic particulate suspended in the upper water column. In samples taken in 1999, the mass of plastic exceeded that of zooplankton (the dominant animal life in the area) by a factor of 6.[57] Midway Atoll, in common with all the Hawaiian Islands, receives substantial amounts of debris from the garbage patch. When it comes to 90% plastic, this debris accumulates on the beaches of Midway where it becomes a hazard to the bird population of the island. Midway Atoll is home to two-thirds (1.5 million) of the global population of Laysan Albatross. Nearly all of these albatrosses have plastic in their digestive system and one-third of their chicks die.[58] Some plastic additives are known to disrupt the endocrine system when consumed; others can suppress the immune system or decrease reproductive rates. Floating debris can also absorb persistent organic pollutants from seawater, including PCBs, DDT, and PAHs.[58] Aside from toxic effects, when ingested, some of these are mistaken by the animal brain for estradiol, causing hormone disruption in the affected wildlife.[59]

Noise Pollution

Marine life can be susceptible to noise or sound pollution from sources such as passing ships, oil exploration seismic surveys, and naval low-frequency active sonar. Sound travels more rapidly and over larger distances in the sea than in the atmosphere. Marine animals, such as cetaceans, often have weak eyesight, and live in a world largely defined by acoustic information. This applies also to many deeper sea fish, which live in a world of darkness. Between 1950 and 1975, ambient noise in the ocean increased by about 10 dB (i.e., a tenfold increase).[3] Noise also makes species communicate louder, which is called the Lombard vocal response. Whale songs are longer when submarine detectors are on. If creatures do not "speak" loud enough, their voice can be masked by anthropogenic sounds. These unheard voices might be warnings, finding of prey, or preparations of net bubbling. When one species begins to speak louder, it will mask other species' voices, causing the whole ecosystem to eventually speak louder. Undersea noise pollution is like the death of a thousand cuts. Each sound in itself may not be a matter of critical concern, but taken altogether, the noise from shipping, seismic surveys, and military activity is creating a totally different environment than that existed even 50 years ago. That high level of noise is bound to have a hard, sweeping impact on life in the sea.[3,60]

Conclusions

The seas and oceans are contaminated with many natural and anthropogenic industrial pollutants. A lot of them are strongly toxic to the organisms and to humans. Apart from plastic, a very particular problem are toxins in the marine environment that do not disintegrate rapidly enough. The most important toxins in the seas are organic compounds (PCBs, DDT, PAHs, pesticides, furans, dioxins, phenol, oil), heavy metals, and radioactive waste. Heavy metals (especially, mercury, lead, nickel, and cadmium) are toxic (in the oxidized form, as ions) or poisonous at low concentrations. These toxins can accumulate in the tissues of many aquatic species. In particular, accumulation of toxins in benthic animals and fish in contaminated estuarine areas and closed seas is very dangerous, due to the high position of these organisms in the food chain and the strong toxicity of the chemicals. Moreover, surface runoff of pesticides to the marine environment can genetically alter the gender of snails and fish species, transforming male into female fish. Finally, the marine environment is an important part of the natural environment consisting of people and human activities, which in the last century have been responsible for production and deposition of many toxins in the oceans and seas.

References

1. Szefer, P. *Metals, Metalloids and Radionuclides in the Baltic Sea Ecosystem*; Elsevier: Amsterdam, the Netherlands, 2002; Vol. 5.
2. Falandysz, J.; Trzosinska, A.; Szefer, P.; Warzocha, P.; Draganic, B. The Baltic Sea, especially southern and eastern regions (Ch. 7). In *Seas in the Millennium: An Environmental Evaluation, Vol. I: Europe, The Americas and West Africa*; Sheppard, C.R.C., Ed.; Pergamon, Elsevier: Amsterdam, 2000; 99–120.
3. Laws, E.A. *Aquatic Pollution*; John Wiley and Sons: New York, 2000.
4. Dowrd, B.M.; Press, D.; Los Huertos, M. Agricultural nonpoint sources: Water pollution policy: The case of California's central coast. Agric. Ecosyst. Environ. **2008**, *128* (3), 151–161.
5. Carpenter, S.R.; Caraco, R.F.; Cornell, D.F.; Howarth, R.W.; Sharpley, A.N.; Smith V.V. Nonpoint pollution of surface waters with phosphorus and nitrogen. Ecol. Appl. **1998**, *8*, 559–568.
6. Forstner, U.; Wittmann G.T.W. *Metal Pollution in the Aquatic Environment*, 2nd Ed.; Springer-Verlag: Berlin, 1983.
7. Daoji, L.; Dag, D. Ocean pollution from land-based sources: East China Sea. Ambio **2004**, *33* (1/2), 107–113.
8. Ahn, Y.H.; Hong, G.H.; Neelamani, S.; Philip, L.; Shanmugam, P. Assessment of levels of coastal marine pollution of Chennai city, southern India. Water Resour. Manage. **2006**, *21* (7), 1187–1206.
9. Li, Y.-H.; Teraoka, H.; Yang, T.-S.; Chen J.-S. The elemental composition of suspended particles from Yellow and Yangtze Rivers. Geochim. Cosmochim. Acta **1984**, *48*, 1561–1564.
10. Mance, G. Pollution Threat of Heavy Metals in Aquatic Environments; Elsevier Applied Science: London, 1987.
11. Ambio. Special issues: Marine eutrophication, Ambio, **1990**, *19*, 101–176.
12. Pimentel, D.; Zuniga, R.; Morrison, D. Update on the environmental and economic coast associate with alieninvasive species in the United States. Ecol. Econ. **2005**, *52*, 273–288.
13. Strandenes, S.P. The second order effects on commercial shipping of restrictions on the use of TBT. Sci. Total Environ. **2000**, *258*, 111–117.
14. Tanabe, S. Butyltin concentration in marine mammals. A review. Mar. Pollut. Bull. **1999**, *39*, 62–72.
15. Alzieu, C. Environmental impact of TBT: The French experience. Sci. Total Environ. **2000**, *258*, 99–102.
16. Evans, S.M. Tributyltin pollution: The catastrophe that never happened. Mar. Pollut. Bull. **1999**, *38*, 629–636.
17. Cambray, R.S.; Jefferies, D.F.; Topping, G. The atmospheric input of trace elements to the North Sea. Mar. Sci. Commun. **1979**, *5*, 175–194.
18. Spengler, J.D.; Sexton K. Indoor air pollution: A public health perspective. Science (New Ser.) **1983**, *221* (4605), 9–17.
19. Nriagu, J.O. A global assessment of natural sources of atmospheric trace metals. Nature **1989**, *338*, 47–51.
20. Pacyna, J.M. Estimation of the atmospheric emissions of trace elements from anthropogenic sources in Europe. Atmos. Environ. **1984**, *18*, 41–50.
21. Fowler, S.W. Critical review of selected heavy metal and chlorinated hydrocarbon concentrations in the marine environment. Mar. Environ. Res. **1990**, *29*, 1–64.
22. Duce, R.A.; Unni, C.K.; Ray, B.J.; Prospero, J.M.; Merrill, J.T. Long-range atmospheric transport of soil dust from Asia to the tropical North Pacific: Temporal variability. Science **1980**, *209*, 1522–1524.

23. Prospero, J.M.; Nees, R.T. Impact of the North African drought and El Niño on mineral dust in the Barbados trade winds. Nature **1986**, *320*, 735–738.
24. Aarkrog, A. The radiological impact if the Chernobyl debris compared with than from weapons fallout. J. Environ. Radioact. **1988**, *6*, 151–162.
25. Dick, A.L. Concentrations and sources of metals in the Antarctic Peninsula aerosol. Geochim. Cosmochim. Acta **1991**, *55*, 1827–1836.
26. Boutron, C. Trace metals in remote Arctic snow: Natural or anthropogenic? Nature **1980**, *284*, 574–576.
27. Dietz, R.; Nielsen, C.O.; Hansen M.M.; Hansen C.T. Organic mercury in Greenland birds and mammals. Sci. Total Environ. **1990**, *95*, 41–51.
28. Dietz, R.; Riget, F.; Johansen P. Lead, cadmium, mercury and selenium in Greenland marine animals. Sci. Total Environ. **1996**, *186*, 67–93.
29. Donaldson, G.M.; Braune B.M.; Gaston A.J.; Noble D.G. Organochlorine and heavy metal residue in breast muscle of know-age thick-milled murres (*Uria lomvia*) from the Canadian Arctic. Arch. Environ. Contam. Toxicol. **1997**, *33*, 430–435.
30. Ahnert, A.; Borowski, C. Environmental risk assessment of anthropogenic activity in the deep sea. J. Aquat. Ecosyst. Stress Recov. **2000**, *7* (4), 299.
31. Halfar, J.; Fujita, R.M. 2007. Danger of deep-sea mining. Science **2007**, *316* (5827), 987.
32. Glasby, G.P. Lessons learned from deep-sea mining. Sci. Mag. **2000**, *289*, 551–553.
33. Nath, B.; Sharma, R. Environment and deep-sea mining: A perspective. Mar. Georesour. Geotechnol. **2000**, *18* (3), 285–294.
34. Sharma, R. Deep-sea impact experiments and their future requirements. Mar. Georesour. Geotechnol. **2005**, *23* (4), 331–338.
35. Orr, J.C.; Fabry, V.J.; Aumont, O.; Bopp, L.; Doney, S.C.; Feely, R.A.; Gnanadesikian, A.; Gruber, N.; Ishida, A.; Joos, F.; Key, R.M.; Lindsay, K.; Maier-Reimer, E.; Matear, R.; Monfray, P.; Mouchet, A.; Najjar, R.G.; Plattner, G.K.; Rodgers, K.B.; Sabine, C.L.; Sarminto, J.L.; Schlitzer, R.; Slater, R.D.; Totterdell, I.J.; Weirig, M.F.; Yamanaka, Y.; Yool, A. Anthropogenic ocean acidification over the twenty-first century and its impact on calcifying organisms. Nature **2005**, *437* (7059), 681–686.
36. Key, R.M.; Kozyr, A.; Sabine, C.L.; Lee, K.; Wanninkhof, R.; Bullister, J.; Feely, R.A.; Millero, F.; Mordy, C.; Peng, T.-H. A global ocean carbon climatology: Results from GLODAP. Global Biogeochem. Cycles **2004**, *18*, GB4031.
37. Feely, R.; Sabine, C.L.; Hernandez-Ayon, J.M.; Ianson, D.; Hales, B. Evidence for Upwelling of corrosive "acidified" seawater onto the continental shelf. Science **2008**, *10*.
38. Milkov, AV. Global estimates of hydrate-bound gas in marine sediments: how much is really out there? Earth Sci. Rev. **2004**, *66* (3–4): 183–197.
39. Skwarzec, B. Radiochemia srodowiska i ochrona radiologiczna (in Polish), Environmental radiochemistry and radiological protection, Wydawnictwo DJ sc., Gdańsk, 2002.
40. Kershaw, P.J.; McCubbin, D.M.; Leonard, K.S. Continuing contamination of the North Arctic and Arctic waters by Sellafield radionuclides. Sci. Total Environ. **1999**, *237/238*, 119–132.
41. Nakaoka, A.; Fukushima, M.; Takagi S. Environmental effect of natural radionuclides from coal-fired power plants. Health Phys. **1984**, *3*, 407–416.
42. Flues, M.; Morales, M.; Mazilli, B.P. The influence of a coal-plant operation on radionuclides in soil. J. Environ. Radioactivity **2002**, *63*, 285–294.
43. Carvalho, F.P.; Oliviera, J.M.; Lopes, I.; Batista, A. Radionuclides from post uranium mining in rivers in Portugal. J. Environ. Radioact. **2007**, *98*, 298–314.
44. Borylo, A.; Nowicki, W.; Skwarzec, B. Isotope of polonium (^{210}Po) and uranium (^{234}U and ^{238}U) in the industrialized area of Wisslinka (Northern Poland). Int. J. Environ. Anal. Chem. **2009**, *89*, 677–685.
45. Skwarzec, B.; Boryło, A.; Kosiiíska, A.; Radziejewska, S. Polonium (^{210}Po) and uranium (^{234}U and ^{238}U) in water, phosphogypsum and their bioaccumulation in plants around phosphogypsum waste heap in Wisslinka (northern Poland), Nukleonika **2010**, *55* (2), 187–193.

46. Skwarzec, B. Polonium, uranium and plutonium radionuclides in aquatic environment of Poland and southern Baltic. Baltic Coastal Zone **2009**, *13*, 127–166.
47. Skwarzec, B. Polonium, uranium and plutonium in the southern Baltic Sea. Ambio **1997**, *26* (2), 113–117.
48. Coughtrey, P.J.; Jackson, D.; Jones, C.H.; Kene, P.; Thorne, M.C. Radionuclides Distribution and Transport in Terrestrial and Aquatic Ecosystems. A Critical Review of Data; A.A. Balkema: Rotterdam, 1984.
49. Rosenberg, R. Eutrophication—The future marine coastal nuisance. Mar. Pollut. Bull. **1985**, *16*, 227–231.
50. Billen, G.; Garnier, J.; Delinge, C.; Billen C. Estimate of early-industrial inputs of nutrients to river system: implication for coastal eutrophication. Sci. Total Environ. **1999**, *243/244*, 43–52.
51. Bruneau, L. Pollution from industries in the drainage area of the Baltic. Ambio **1980**, *9*, 145–152.
52. MacLean, J.L.; White A.W. Toxic dinoflagellates bloom in Asia: A growing, concern. In *Toxic Dinoflagellates*; Anderson, D.M., White, A.W., Baden, D.G., Eds.; Elsevier: New York, 1985; 517–520.
53. Paerl, H.W. Enhancement of marine primary production by nitrogen enriched rain. Nature **1985**, *315*, 747–749.
54. Selman, M. Eutrophication: An Overview of Status, Trends, Policies, and Strategies; World Resources Institute, Washington DC, UAS, 2007.
55. Sheavly, S.B.; Register, K.M. Marine debris and plastics: Environmental concerns, sources, impacts and solutions. J. Polym. Environ. **2007**, *15* (4), 301–305.
56. Thompson, R.C. Lost at sea: Where is all the plastic? Science **2004**, *304* (5672), 843.
57. Moore, C.; Moore, S.L.; Leecaster, M.K.; Weisberg, S.B. A comparison of plastic and plankton in the North Pacific Central Gyre. Mar. Pollut. Bull. **2004**, *42* (12), 1297–1300.
58. Rios, L.M.; Moore, C.; Jones, P.R. Persistent organic pollutants carried by synthetic polymers in the ocean environment. Mar. Pollut. Bull. **2007**, *54*, 1230–1237.
59. Tanabe, S.; Watanabe, M.; Minh, T.B.; Kunisue, T.; Nakanishi, S.; Ono, H.; Tanaka, H. PCDDs, PCDFs, and coplanar PCBs in albatross from the North Pacific and Southern Oceans: Levels, patterns, and toxicological implications. Environ. Sci. Technol. **2004**, *38*, 403–413.
60. Fristrup, K.M.; Hatch, L.T.; Clark, C.W. Variation in humpback whale (*Megaptera novaeangliae*) song length in relation to low-frequency sound broadcasts. Acoust. Soc. Am. J. **2003**, *113* (6), 3411–3424.

II

COV: Comparative Overviews of Important Topics for Environmental Management

11

Rain Water: Harvesting

K.F. Andrew Lo

Introduction

Among the various alternative technologies to augment water resources, rainwater harvesting is a simple, decentralized solution and imposes insignificant impact on the environment. It is an important water source in many areas with significant rainfall but lacking any kind of conventional, centralized supply system. It is also a good option in areas where good-quality fresh surface water or groundwater is lacking. Rainwater harvesting systems have been used since ancient times and evidence of roof catchment systems dates back to early Roman times. In the Negev Desert in Israel, in Libya and Egypt, in Mexico, and in the Andes Range in South America as well as in the Arizona Desert in North America, stone dams and tanks were built to divert and store rainwater for irrigation purposes.

Advantages of Rainwater Harvesting

Rainwater harvesting systems can provide water at, or near, the point where water is needed or used. The systems can be both owner-operated and utility-operated, and owner-managed and utility-managed. Rainwater collected using existing structures (rooftops, parking lots, playgrounds, parks, ponds, and flood plains) has few negative environmental impacts compared with other water resources development technologies.[1] Rainwater is relatively clean and the quality is usually acceptable for many purposes with little or even no treatment. The physical and chemical properties of rainwater are usually superior to sources of groundwater that may have been subject to contamination.

Other advantages of rainwater harvesting include the following:

1. Rainwater harvesting can coexist with, and provide a good supplement to, other water sources and utility systems, thus relieving pressure on other water sources.
2. Rainwater harvesting provides a water supply buffer for use in times of emergency or breakdown of public water supply systems, particularly during natural disasters.

3. Rainwater harvesting can reduce storm drainage load and flooding in cities.
4. The owners who operate and manage the rainwater catchment system are more willing to exercise water conservation.
5. Rainwater harvesting technologies are flexible and can be built to meet almost any requirements.

Types of Rainwater Harvesting Systems

Collection systems can vary from simple households to large catchment systems. The categorization of rainwater harvesting systems depends on factors such as the size and nature of the catchment areas and whether the systems are in urban or rural settings.[2]

Simple Rooftop Collection Systems

The main components of a simple rooftop collection system are the cistern itself, the piping that leads to the cistern, and the appurtenances within the cistern (Figure 1). The materials and the degree of sophistication of the whole system largely depend on the initial capital investment. Some cost-effective systems involve cisterns made with ferrocement. In some cases, the harvested rainwater may be filtered or disinfected.

Large Systems for Educational Institutions, Stadiums, Airports, and Other Facilities

When the systems are larger, the overall system can become more complicated (e.g., rainwater collection from roofs and grounds of institutions, storage in underground reservoirs, and treatment and use for non-potable applications) (Figure 2).

Rooftop Collection Systems for High-Rise Buildings in Urbanized Areas

In high-rise buildings, roofs can be designed for catchment purposes and the collected roof water can be kept in separate cisterns on the roofs for non-potable uses.

Land Surface Catchments

Ground catchment techniques (Figure 3) provide more opportunity for collecting water from a larger surface area. By retaining small creek and stream flows in small storage surfaces or underground reservoirs,

FIGURE 1 A simple roof catchment system (illustrated by Chia-Ming Lin).

FIGURE 2 An indoor storage system in a monastery in China (photographed by K. F. Andrew Lo).

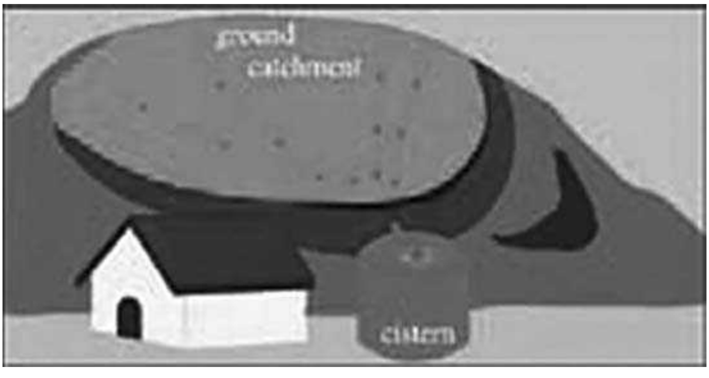


FIGURE 3 A land catchment system (illustrated by Chia-Ming Lin).

can meet water demands during dry periods. However, there is a possibility of high seepage loss to the ground. The marginal quality of the water collected is suitable for use mainly in agriculture.

Collection of Stormwater in Urbanized Catchment

The surface runoff collected in stormwater ponds/reservoirs from urban areas is subject to a wide variety of contaminants. Keeping these catchments clean is of primary importance; hence the cost of water pollution control can be considerable.

Design and Maintenance of Rainwater Harvesting Systems

Typically, a rainwater harvesting system consists of three basic elements: the collection system, the conveyance system, and the storage system.

Catchment Surface

The effective catchment area and the material used in constructing the catchment surface influence collection efficiency and water quality. Materials commonly used for roof catchment are corrugated aluminum and galvanized iron, concrete, fiberglass shingles, tiles, and slates. Mud is used primarily in rural areas. Bamboo roofs are least suitable because of possible health hazards. The catchment surface materials must be non-toxic and must not contain substances that impair water quality. Roofs with metallic paint or other coatings are not recommended because they may impart tastes or color to the collected water. Catchment surfaces and collection devices should be cleaned regularly to remove dust, leaves, and bird droppings to minimize bacterial contamination and to maintain the quality of collected water. Roofs should also be free from overhanging trees because birds and animals in the trees may defecate on the roofs.

When land surfaces are used as catchment areas, various techniques are available to increase runoff capacity: 1) clearing or altering vegetation cover; 2) increasing the land slope with artificial ground cover; and 3) reducing soil permeability by soil compaction. Specially constructed ground surfaces (concrete, paving stones, or some kind of liner) or paved runways can also be used to collect and convey rainwater to storage tanks or reservoirs. Care is required to avoid land surface damage and contamination by people and animals. If required, these surfaces should be fenced to prevent people and animal entry. Large cracks in the paved catchment because of soil movement, earthquakes, or prolonged exposure should be repaired immediately. Maintenance, typically consisting of the removal of dirt, leaves, and other accumulated materials, should take place annually before the start of the major rainfall season.

Conveyance Systems

Conveyance systems are required to transfer the rainwater collected on catchment surfaces to storage tanks. This is usually accomplished by making connections to one or more downpipes connected to collection devices. The pipes used for conveying rainwater, wherever possible, should be made of plastic, polyvinyl chloride (PVC), or other inert substance because the pH of rainwater can be acidic and may cause corrosion and mobilization of metals in metal pipes.

When it first starts to rain, dirt and debris from catchment surfaces and collection devices will be washed into the conveyance systems. Relatively clean water will only be available sometime later in the storm. The first part of each rainfall should be diverted from the storage tank. There are several possible options for selectively collecting clean water for the storage tanks. The common method is a sediment trap, which prevents debris entry into the tank. Installing a first-flush (or foul-flush) device is also useful to divert the initial batch of rainwater away from the tank.[3]

Rainwater pipes must be permanently marked in such a way that there is no risk of confusing them with drinking water pipes. Gutters and downpipes need to be periodically inspected and carefully cleaned. A good time to inspect gutters and downpipes is while it is raining, so that leaks can be easily detected.

Storage Tanks

Various types of rainwater storage facilities can be found in practice. Storage tanks should be constructed of inert material. Reinforced concrete, fiberglass, polyethylene, and stainless steel are suitable materials. Ferrocement tanks and jars made of mortar or earthen materials are commonly used. As an alternative, interconnected tanks made of pottery or polyethylene may be suitable. They are easy to clean. Bamboo-reinforced tanks are less successful because they may become infested with termites, bacteria, and fungi.

Precautions are required to prevent the entry of contaminants into storage tanks. The main sources of external contamination are pollution from debris, bird and animal droppings, and insects. A solid and secure cover is required to avoid breeding of mosquitoes, to prevent insects and rodents from entering the tank, and to keep out sunlight to prevent algae growth inside the tank.[4] A coarse inlet filter is also desirable for excluding coarse debris, dirt, leaves, and other solid materials.

All tanks need cleaning and their designs should allow for thorough scrubbing of the inner walls and floors. A sloped bottom and the provision of a pump and a drain are useful for collection and discharge of settled grit and sediment. Chlorination of the cisterns or storage tanks is necessary if the water is to be used for drinking and domestic uses. Cracks in the storage tanks can create major problems and should be repaired immediately.

The extraction system (taps/faucets, pumps) must not contaminate the stored water. Taps/faucets should be installed at least 10 cm above the base of the tank because this allows any debris entering the tank to settle on the bottom.[5] If it remains undisturbed, it will not affect the quality of the water. The handle of taps might be detachable to avoid misuse by children. Periodic maintenance should also be carried out on pumps used to lift water.

Conclusion

In the future, water scarcity in both developing and developed countries is inevitable.[6] The challenge of meeting the water demand can be largely met by appropriate understanding, study, and application of rainwater harvesting. Rainwater harvesting is about to come of age.[7] It has an appropriate image about it that meshes well with the gentler ideas of the late 20th century. Because the technique makes use of an untapped resource—precipitation that would otherwise be evaporated before it had a chance to play a useful role in feeding the human population—it looks like getting something for nothing. Making use of such a resource has certain poetry to it, particularly in a field where the resource itself can never be increased or decreased; unlike food, water cannot be grown to order, even given the right soil and the right fertilizer. But, like food, water can be harvested more efficiently. Doing so is a major priority for the 21st century.

References

1. Geiger, W.F. Integrated water management for urban areas, Proceedings of the 7th International Rainwater Catchment Systems Conference, Beijing, China, June 21–25, 1995; Liu, C.M., Ed.; Institute of Geographical Sciences and Natural Resources, Chinese Academy of Sciences: Beijing, 1995, 7-1–7–21.
2. Gould, J.; Nissen-Peterson, E. *Rainwater Catchment Systems for Domestic Supply: Design, Construction and Implementation*; Intermediate Technology Publications: London, 1999.

3. Macomber, P.S.H. *Guidelines on Rain Water Catchment Systems for Hawaii*; College of Tropical Agriculture and Human Resources, University of Hawaii at Manoa: Hawaii, 2001.
4. Cunliffe, D.A. *Guidance on the Use of Rainwater Tanks;* National Environmental Health Forum Monographs, Water Series No. 3; National Environmental Health Forum: Australia, 1998.
5. Texas Water Development Board *Texas Guide to Rainwater Harvesting*, 2nd Ed.; Texas Water Development Board: Texas, 1997.
6. Clarke, R. *Water: The International Crisis*; The MIT Press: Cambridge, 1993.
7. Pacey, A.; Cullis, A. *Rainwater Harvesting: The Collection of Rainfall and Runoff in Rural Areas*; Intermediate Technology Publications: London, 1986.

12

Water Harvesting

Gary W. Frasier

Introduction

Water is one of the critical elements to support life. Without water, life as we know it cannot exist. Without water, plants will die. Humans and animals will die without drinking water. Yet, the planet we call earth has more area covered with water than landmass. From space, the planet looks blue from the water in the seas and the clouds. Even with all this water, there are many places on the land, classified as arid and semiarid land, where there is not enough drinking water for man and animals to live or plants to grow. In many of these arid and semiarid places, sufficient water can be obtained for drinking-water supplies and growing of plants by a process called water harvesting. In technical terms, water harvesting is defined as "the process of collecting and storing precipitation (rain or snow) for beneficial use from an area that has been treated or modified to increase precipitation runoff."[1]

With the modern advent of water transfers via canals and deep wells for supplying water for homes and irrigation of crops, there is a tendency to consider water harvesting as a "new" water supply technique. In actuality, water harvesting is an ancient practice that has been dated back to the Edom Mountain areas of Southern Jordan 9000 years ago.[2] There is evidence that water harvesting systems were used 3000 to 4000 years ago in various places of what we now call the Middle East. Some of these early systems were located in areas having less than 200 mm annual precipitation.[3] Water harvesting systems are currently being used for water supply in various places around the world such as Israel, Egypt, Jordan, Mexico, Australia, and the United States.

General Description

There are two general classifications of water harvesting systems. One group is used for providing drinking-water supplies for human and animal consumption. The other group is used to provide water for the growing of plants (runoff farming). All water harvesting systems have two major components: an area where the precipitation is collected (catchment area) and some means for storing the collected water.

The water collection area can be any surface that has been modified to increase precipitation runoff. They can include rock outcroppings with diversion facilities at the lower edge to direct the runoff water to a central point; smoothed compacted soil, chemically treated soil surfaces; soil covered with membranes of rubber, wood, or metal; and building rooftops. The water storage facility can be a tank, pond, or cistern with an impermeable lining or, in the case of runoff farming, the soil around the plants.

Water Harvesting for Livestock Drinking Water

Providing animal, domestic, and wildlife drinking water supplies is one of the newer applications of water harvesting (past 20–30 years). The catchment area can be a hillslope, an area covered or treated to reduce water infiltration and increase surface water runoff, or even roofs of buildings. The catchment area can be modified via various techniques, frequently separated into 3 general categories: 1) topography modification; 2) soil modification; and 3) impermeable coverings or membranes. Topography modification consists of land smoothing and clearing and frequently yields 20%–35% runoff efficiency. Some typical soil modification treatments used are sodium salts mixed in the top 6–12 cm of soil and compacted with runoff yields of 50%–80%. Water repellents and paraffin wax applied to the soil surface will yield 60%–95% runoff. Asphalts or bitumen sprayed on the soil surface provide 50%–80% runoff. More expensive catchment treatments consist of various impermeable coverings. They consist of gravel-covered sheeting (75%–95% runoff); asphalt membranes (85%–95% runoff); and concrete, sheet metal, and artificial rubber sheeting (60%–95% runoff).[4,5]

Storage of the collected water is a key component of any water harvesting system, but it is also one of the most expensive components. Unlined pits or ponds are usually not suitable unless seepage losses can be stopped. There are many types, shapes, and sizes of wooden, metal, concrete, or reinforced plastic storage containers that can be used. Also, in most arid and semiarid areas where the precipitation events are erratic, it is necessary to provide some means of controlling evaporative losses from the stored-water facility. Evaporation control on sloping-sided pits or ponds is difficult because the water surface area changes with depth.[5]

Water Harvesting for Domestic Use

Most of the techniques used for water harvesting for livestock and wildlife drinking supplies can be used to provide water for domestic uses. One technique that is used in several places in the world today is the collection of water from the roofs of buildings.[6] A relatively small roof area can provide significant quantities of water if collected and stored. One millimeter of precipitation will yield 1 L of water. This technique is very effective in areas where there are frequent precipitation events but little or no surface water, because of highly permeable soils or groundwater that has high concentrations of undesirable minerals. A simple wooden or metal tank or concrete cistern stores the collected water.

Water Harvesting for Growing of Crops

One form of water harvesting used for growing crops is termed runoff farming. Runoff farming maximizes the effect of limited precipitation by collecting surface runoff from a large area and applying it to a smaller cropped area. The collected water can be applied directly to the crop area during the precipitation event, referred to as floodwater farming,[3] or stored and applied later by some form of irrigation system.[7] A modified version is termed microcatchment farming, where a small runoff area is situated directly upslope of the growing area. This technique has been used for tree crops such as pistachio, olive, and almond.[3]

Other Considerations

Water harvesting should not be considered as a cheap or inexpensive means of water supply. It can supply water in areas where other water sources are unattainable or unsuitable. Water harvesting is not a new means of water supply or source but is receiving renewed interest in many places in the world.

It became a water source of interest with the work of Evenari and his colleagues[8] in the early 1960s reconstructing ancient water harvesting systems in the Negev Desert of Israel. Following this work, several research organizations around the world conducted extensive studies, primarily exploring various low-cost means of treating the soil to reduce infiltration and increase precipitation runoff. Much of this effort has currently been completed. Many water harvesting systems are being used around the world, most notably in arid and semiarid, underdeveloped areas where there is plentiful labor. At the same time, there are many places where the systems have failed. One of the most common reasons for failure is the lack of adequate maintenance of the collection area and the water storage facility. When properly designed and maintained, a water harvesting system can last for decades.

References

1. Frasier, G.W. Water harvesting for rangeland water supplies: a historic perspective. In Proceedings of a Symposium on Environmental, Economic, and Legal Issues Related to Rangeland Water Development, Tempe, Arizona, November 13–15, 1997; The Center for the Study of Law, Science, and Technology, Arizona State University: Tempe, Arizona, 1998; 17–24.
2. Bruins, H.M.; Evenari, M.; Nessler, U. Rainwater-harvesting agriculture for food production in arid zones. The challenges of the African famine. Appl. Geogr. **1986**, *6*, 13–33.
3. Frasier, G.W. Water harvesting. In *Encyclopedia of Soil Science;* Marcel Dekker Inc.: New York, 2002; 1387–1389.
4. Frasier, G.W.; Cooley, K.R.; Griggs, J.R. Performance evaluation of water harvesting catchments. J. Range Manage. **1979**, *32* (6), 453–456.
5. Frasier, G.W.; Myers, L.E. *Handbook of Water Harvesting,* Agriculture Handbook Number 600; U.S. Department of Agriculture: Washington, D.C., 1983; 48 pp.
6. Frasier, G.W. Harvesting water for agricultural, wildlife, and domestic uses. J. Soil Water Conserv. **1980**, *35* (3), 125–128.
7. Scrimgeour, F.G.; Frasier, G.W. Runoff impoundment for supplemental irrigation: An economic assessment. Amer. J. Alter. Agric. **1991**, *6* (3), 139–145.
8. Evenari, M.; Shanan, L.; Tadmor, N.H. Runoff farming: In the desert. I. Experimental Layout. Agron. J. **1968**, *60,* 29–32.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

13

Groundwater: Saltwater Intrusion

Alexander
H.-D. Cheng

Introduction

Human beings have a tendency to live near coastal areas. About 23% of the world's population lives within 100 km of the coast. This figure is likely to increase to 50% in the next 25 years. In the United States, 54% of all Americans now live in 772 coastal counties adjacent to the Atlantic and Pacific Oceans, the Gulf of Mexico, and the Great Lakes. By the year 2025, nearly 75% of all Americans are expected to live in coastal counties.

Coastal aquifers serve as major sources for freshwater supply in many countries around the world, especially in arid and semiarid zones. Many coastal areas are also heavily urbanized, a fact that makes the need for freshwater even more acute. Coastal aquifers are highly sensitive to disturbances. Inappropriate management of a coastal aquifer can lead to its destruction as a source for freshwater much earlier than aquifers that are not connected to the sea. The reason is the threat of seawater intrusion into freshwater aquifers.

The origin of saltwater intrusion into freshwater aquifers can come from natural sources such as seawater and deep formation brines or from anthropogenic sources such as deicing salt, agricultural return flow, and leachate from landfills. The most frequent occurrences are found in coastal regions where overexploitation of groundwater has caused the encroachment of seawater into freshwater aquifers. Once an aquifer is invaded, a part of the salt will adsorb onto the solid surface, making it difficult to reverse the process and restore the aquifer. The slow movement of groundwater also makes the remediation time long. Salinity in water poses a health hazard for humans and livestock, damages crops, and corrodes pipes and boilers in industrial uses. Hence, the invasion of saltwater into a freshwater aquifer means the loss of that aquifer for water sources.

Saltwater Intrusion Problems around the World

Serious saltwater intrusion problems exist in many parts of the world. In the following, we shall review some of the more prominently reported cases.

United States

In the United States, the potential for saltwater intrusion was recognized as early as 1824 in New Jersey and 1854 on Long Island, New York.[1] In modern times, serious problems of saltwater intrusion into aquifers are found in the Miami, Long Island, Hawaii, and Los Angeles areas.

Seawater intrusion into aquifers is a problem throughout coastal Florida. Particularly, in the Miami area, a network of drainage canals was constructed from 1909 through the 1930s. The resulting drainage lowered water levels about 2 m in the Everglades. This drainage, combined with coastal pumping, caused seawater to advance progressively into the Biscayne aquifer.[2] The intrusion front has reached several kilometers inland.[3] The South Florida Water Management District is one of the regional water management agencies in the state that has the regulatory power to issue or deny a well construction permit, to require the installation of monitoring wells, and to impose restrictions on groundwater extraction during drought periods based on seawater intrusion considerations.

In Los Angeles County, California, severe groundwater overdraft from the early 1900s to the late 1950s caused water levels to drop below sea level, allowing saltwater to intrude into the potable aquifers, rendering coastal wells out of service. In an effort to halt the intrusion, groundwater management agencies took three major steps from the mid-1950s to mid-1960s, including the following: 1) construction of freshwater injection wells along the coast to prevent the saltwater intrusion; 2) limiting the amount of groundwater that could be pumped annually; and 3) creation of the Water Replenishment District of Southern California to purchase artificial replenishment water to make up the annual and accumulated overdrafts.[5]

For the Oahu Island of Hawaii, fresh groundwater exists as a lens floating on top of saltwater. In the predevelopment time (pre-1880), the freshwater lens in the Pearl Harbor area was estimated to be 400 m to 500 m thick.[6] In 1990, measured salinity profiles showed the freshwater lens to be only 200 m to 300 m thick.[7] Thus, the 100 years of development has caused the freshwater lens to reduce to half its thickness.

The Netherlands

The Netherlands is located on river deltas, with much of its territory below sea level, reclaimed from the sea. The low- lying lands, known as polders, are particularly susceptible to the invasion of seawater. To keep out the brackish water, year-round recharge of aquifers using reclaimed water is needed.[8]

A great concern for the Netherlands is the threat of sea level rise, as a result of the global warming trend. The rise of sea level can inundate land and push the saltwater-freshwater equilibrium point further inland. It will also increase the hydraulic pressure of seepage flow for the low-lying areas. Possible countermeasures have been investigated.[9]

Israel and Palestinian Territories

Due to the semiarid climate, Israel and Palestinian territories (Gaza Strip and West Bank) are critically dependent on groundwater for water supply. In fact, the sandstone Quaternary aquifer that underlies coastal Israel and the Gaza Strip is being “mined,” and the Mediterranean Sea intrudes along a saltwater front.[10,11] The water demands have been so much that the strategy is no longer to push the saltwater front outward to sea. The management goal is to allow the aquifer to be contaminated and to intercept the largest possible percentage of freshwater outflow to sea.

Mexico

The aquifer in northwestern Yucatan contains a freshwater lens that floats above a denser saline water wedge. It has been shown that at certain locations, the penetration of saline water can be more than 100 km inland.[12] The aquifer, which is unconfined except for a narrow band along the coast, is the sole freshwater source in northwestern Yucatan. Development of industry and agriculture and other land use changes pose a potential threat to the quantity and quality of freshwater resources in the Yucatan Peninsula.[13]

Italy

In the 1920s, Venice went through an Industrial Revolution. During this time period, water was constantly being pumped from the underlying aquifer. The lowering of the piezometric head has caused not only land subsidence but also the contamination of aquifer by saline water. Geophysical investigations have shown that the saline water may extend inshore up to 20 km from the Adriatic Sea coastline. The saltwater plume is observed from the near ground surface down to 100 m.[14,15] The combined effect of sea level rise and land subsidence can further enhance the saltwater contamination and the related soil salinization, with serious environmental and socioeconomic impacts.

Other Parts of the World

Saltwater intrusion problems are also found in the Nile Delta of Egypt,[16] Cap Bon of Tunisia,[17] China,[18] Cyprus,[19] Morocco,[20] and many other parts of the world.

Mechanisms of Saltwater Intrusion

Figure 1 gives a simplified view of seawater intrusion into an unconfined aquifer. We observe that saltwater is heavier and hence tends to stay underneath the freshwater. The freshwater, however, has a hydraulic gradient with head decreasing from inland toward the sea, thus driving a freshwater flow to the sea. This outflow momentum counteracts the density-driven seawater. Without it, seawater will continue to move inland until the entire aquifer below sea level is occupied by the heavier saltwater. However, since such a hydraulic gradient always exists due to the precipitation recharge inland, an equilibrium position of saltwater and freshwater will be established, schematically shown as the interface in Figure 1.

Ghyben-Herzberg Relation

The earliest and the simplest explanation of the saltwater intrusion mechanism was provided by Du Cummun,[21] Badon-Ghyben,[22] and Herzberg,[23] commonly known as the Ghyben-Herzberg relation. By making the following assumptions—1) a sharp interface exists, separating saltwater from freshwater; 2) the freshwater flow lines are approximately horizontal; and 3) there is no flow in the saltwater zone—a simple relation between the freshwater free surface elevation, h_f, and the interface location below the sea level, η, can be developed as (see Figure 1)

$$\eta = \frac{\rho_f}{\rho_s - \rho_f} h_f \approx 40 h_f \tag{1}$$

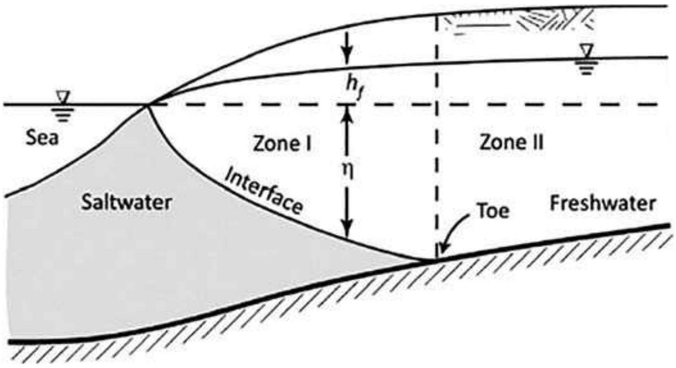


FIGURE 1 Seawater intrusion into an unconfined aquifer.

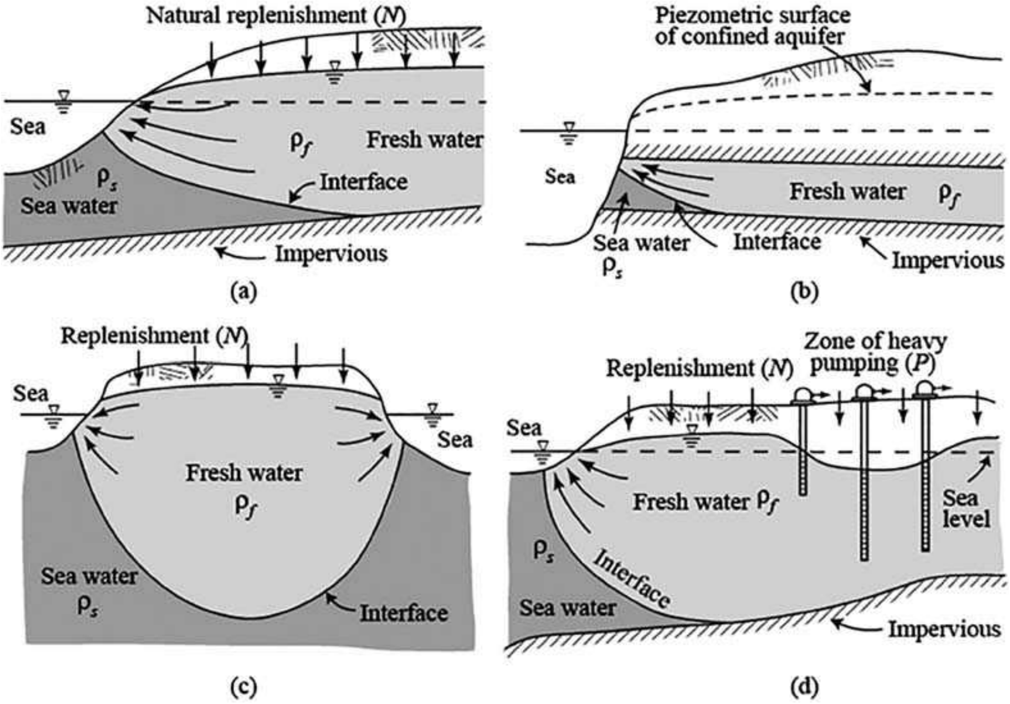


FIGURE 2 Typical vertical cross sections of seawater intrusion in coastal aquifers: (a) unconfined aquifer with replenishment; (b) confined aquifer; (c) freshwater lens on an island; (d) unconfined aquifer with pumping.

where ρ_f is freshwater density (1 g/cm^3) and ρ_s is saltwater density (approximately 1.025 g/cm^3). This is known as the

What the above relation says is that for every meter of freshwater head above sea level, the interface is pushed down 40 m below sea level. When the interface touches the bottom of an aquifer, a toe for the seawater wedge is located (Figure 1). Landward from the toe (zone 2), only freshwater exists; seaward from the toe (zone 1), saltwater and freshwater coexist.

Figure 2 presents some typical cross sections with interfaces in coastal aquifers under natural conditions (a–c) and with pumping (d). Like all figures that describe aquifers, these figures are highly distorted, with vertical scale much more magnified.

Up-Coning

This 40:1 ratio mentioned above based on the Ghyben–Herzberg relation may sound like good news for repelling saltwater by freshwater head. However, if pumping activity is increased in the coastal zone, as often is the case due to increasing population and economic activities, the reduced freshwater head will allow saltwater to move a large distance landward. Also, if a pumping well is situated above the interface in the saltwater-freshwater coexisting zone, the drawdown of a water table or piezometric head will cause the interface to rise up steeply to meet the well. This phenomenon is known as up-coning. Once the saltwater reaches the well, the well is generally considered to be lost.

For a pumping well situated at a distance d above an undisturbed interface (Figure 3), a formula that predicts the maximum allowable discharge, $Q_{\max}$, before the saltwater entering the well is given by[24,25]

$$Q_{\max} < 0.6\pi \Delta s K d^2 \approx 0.047 K d^2 \tag{2}$$

where K is the hydraulic conductivity, and $\Delta s = (\rho_s - \rho_f)/\rho_f \approx 0.25$. Take, for example, $K = 100$ m^1/day; then a well with its screened section situated 10 m above the interface can pump up to 470 m^2/day. On the other hand, if the distance d is reduced to 5 m, then the pumping rate is restricted to 117 m^3/day.

In Figure 4, we demonstrate a few saltwater intrusion scenarios with a well pumping above the interface. Figure 4a shows a confined aquifer with a well pumping at a relatively small rate. In this case, a freshwater capture zone is formed similar to a capture zone in the entirely freshwater region. Below the well, although there is an up-coning, the well is pumping freshwater.

In Figure 4b, the pumping rate is increased but not large enough to cause the interface to reach the well. At the cross section where the well is located, the (three-dimensional) capture zone touches the interface. Near the interface, there exists mixing of saltwater and freshwater due to hydrodynamic dispersion. Thus, trace salt content may be observed in the well. Figure 4c shows a similar situation as Figure 4b but for an unconfined aquifer.

Figure 4d gives a two-dimensional situation. A two dimensional pumping situation can be created either by a gallery (collector) well, that is, a perforated pipeline parallel to the coastline, or by a number of wells along the coast and close to each other, such that their influence zone overlaps. In this case, as long as the discharge per unit length of the gallery well does not exceed the freshwater outflow per unit length of coastline, the excess freshwater must flow underneath the well; hence, the interface cannot reach the well. In this way, the well can pump nearly as much as the entire freshwater outflow, not

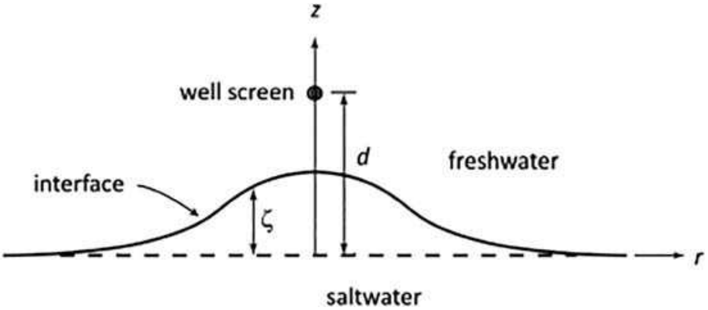


FIGURE 3 Interface up-coning below a pumping well.

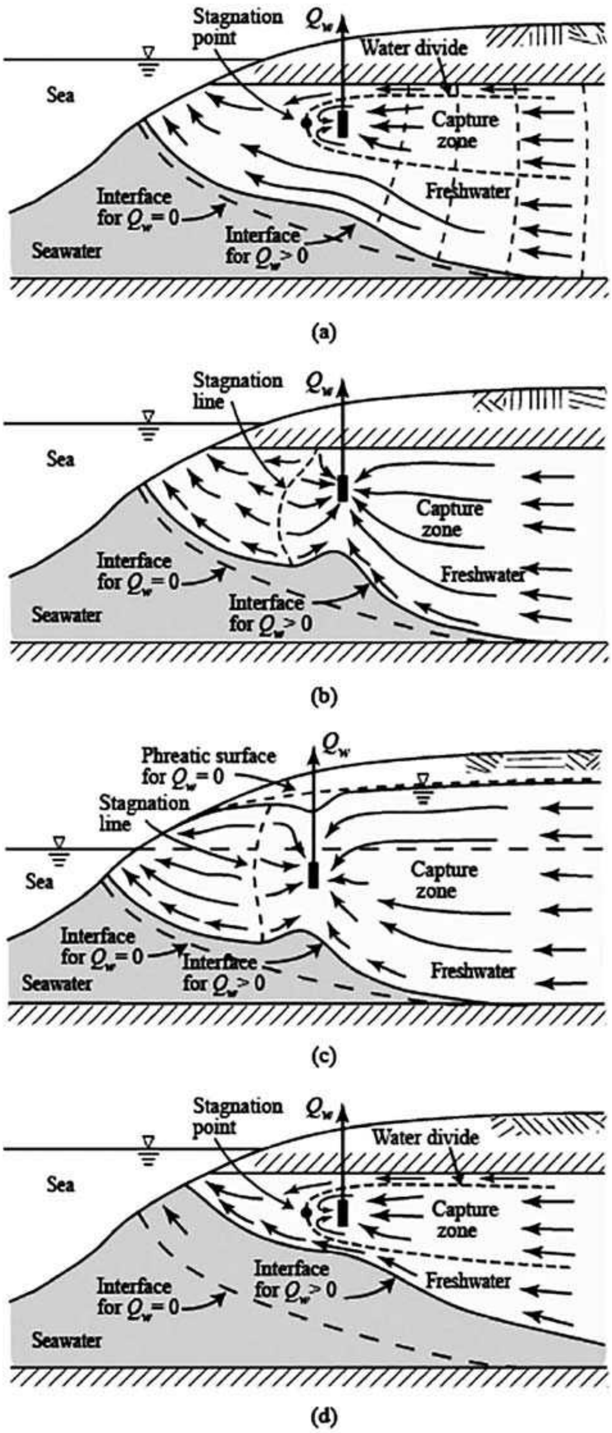


FIGURE 4 Up-coning, freshwater flow directions and capture zone in the vicinity of a well pumping above interface: (a) confined aquifer, smaller discharge; (b) confined aquifer, larger discharge; (c) unconfined aquifer; and (d) two-dimensional flow (with gallery well).
Source: Bear and Cheng.[26]

wasting any available freshwater. This type of skimming well that distributes the pumping rate over a line or over an area to avoid concentrated local up-coning beneath a point well has been practiced in Israel,[27] Palestinian territories,[28] and Pakistan.[29,30]

Transition Zone

In the above figures, saltwater and freshwater are shown to be separated by a sharp interface. This is a simplification of the actual situation, as saltwater and freshwater are miscible liquids. Actually, they constitute a single liquid phase—water—with different concentrations of total dissolved solid (TDS), such as salt. Often, the term "interface" is used for the iso-density surface that is midway between freshwater and seawater.

For the sake of simplicity, we shall continue to refer to two liquids—freshwater and seawater. The passage from the portion of the aquifer that is occupied by one liquid to the other takes the form of a transition zone, rather than a sharp interface. Depending on the extent of seawater intrusion and certain aquifer properties, this transition zone may be rather wide or narrow. The width of the transition zone is dictated by three phenomena:

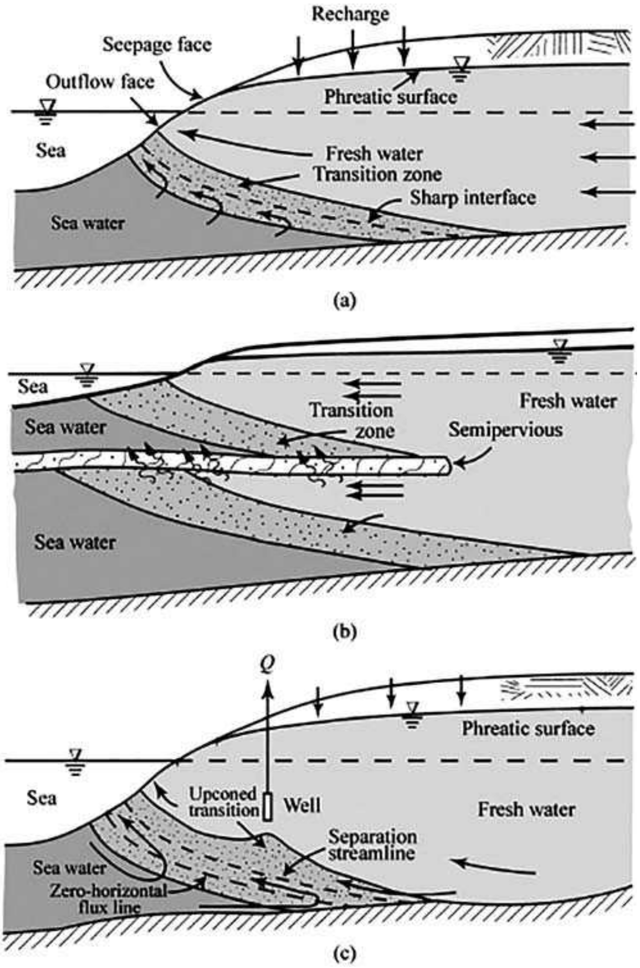


FIGURE 5 Transition zones between saltwater and freshwater regions: (a) an unconfined aquifer; (b) two aquifers separated by a semipervious layer; and (c) with a pumping well above interface.

- Hydrodynamic dispersion (dispersion and molecular diffusion)
- Advection of water—fresh and mixed—toward the sea (or, under certain conditions, landward)
- Recirculation of seawater and mixed water

Figure 5 demonstrates a few scenarios of transition zones between the saltwater and freshwater region. In Figure 5a, we observe that we can approximate a sharp interface location as the 50% relative concentration line. Figure 5b depicts a multilayer situation, where there are two transition zones. As the partition layer is semipervious, saltwater and freshwater exchange can also take place by leakage through the semipervious layer. Figure 5c shows the case with a pumping well above the interface, creating an up- coning situation.

Monitoring and Exploration of Saltwater Intrusion

Geological Investigation

A direct way of detecting saltwater intrusion is by the geological technique of drilling monitoring wells. In aquifers where a thick saltwater-freshwater transition zone exists, it is of interest to monitor the continuous change in salt concentration across the transition zone. Monitoring such information requires taking water samples at different depths. This can be achieved by the use of packers that seal sections of the well, selectively, such that each section has its own screen. Water samples taken from different sections are not mixed with each other. Obviously, such an operation is tedious, and the data obtained are limited to the well's location and to the elevation of the screened portions of the well. In order to cover a large area, a large number of wells need to be drilled, and the associated costs are usually high. The data obtained, however, are the most direct and, probably, the most accurate. Data obtained by this method are often used to calibrate data obtained by indirect methods, such as geophysical methods.

One way to reduce the effort and cost of monitoring is to utilize the Ghyben-Herzberg relation to interpret the interface location. In this case, monitoring wells need only to reach the phreatic surface, and water sampling for salt concentration is not necessary. With the Ghyben-Herzberg relation, it is possible to map the saltwater-freshwater interface using the following steps in a field investigation:

1. A network of shallow wells is deployed to observe the water table height, to give the freshwater head h_f above mean sea level.
2. The water table surface can be created using an interpolation procedure, such as kriging, expressed as $h_f(x,y)$.
3. The interface depth h(x,y) below mean sea level is represented by the same function $hf(x,y)$, except that it is amplified by a factor of 40, as indicated by Eq. 1.
4. The location of aquifer bottom $b(x,y)$ is interpolated from geological coring or map.
5. The intersection of the two surfaces h(x, y) and $b(x,y)$ is sought, which represents the toe location of the saltwater wedge.

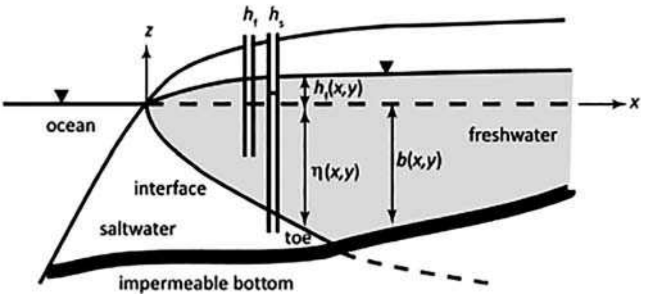


FIGURE 6 Locating saltwater toe by Ghyben–Herzberg relation.

This procedure is illustrated in Figure 6 using a vertical cross section. The above procedure is applicable for confined as well as unconfined aquifers.

The Ghyben–Herzberg relation, Eq. 1, contains the assumption that the saltwater is static. A better approximation for locating the interface that does not contain this assumption is given by[31]

$$\eta = \frac{\rho_f}{\rho_s - \rho_f} h_f - \frac{\rho_s}{\rho_s - \rho_f} h_s \tag{3}$$

The above relation states that if we can measure at the same location both the freshwater head h_f and the saltwater head h_s, as illustrated in Figure 6, then the interface location can be better determined.

Geophysical Investigation

Geophysical methods make use of physical properties, such as the velocity of seismic waves, electrical conductivity, electromagnetic permeability, thermal conductivity, etc., of the geomaterials, such as the soil, and the water occupying the void space, with the latter varying with the concentration of dissolved matter. The typical advantages of the geophysical methods are the following:

- Measurements are taken at ground surface, although some techniques also utilize existing or specially drilled boreholes.
- Sometimes, measurements can be conducted airborne by aircraft. Usually, this means less time consumed and lower cost.
- Measurements cover a large surface area, thus making the methods more suitable for field-scale investigations.

The disadvantage of the surface-based geophysical methods, however, is that they are indirect, measuring some analogous quantities, thus often requiring calibration from direct measurements. The optimal solution to this problem is to conduct an integrated survey that combines results from geophysical techniques with data obtained from direct water sampling.

Several geophysical methods that can be used for detecting the presence of saline water in geological formations are briefly discussed below, based on Stewart:[32]

- Direct current (DC) resistivity. The method is based on the increase in electrical conductivity with increasing pore water salinity. The principal advantages of the method are its simplicity and the relatively low cost of the required equipment. In the DC method, an electrical current is introduced into the ground through electrodes driven into the soil. The resulting electrical potential (voltage) is measured between the two electrodes. The measured resistance represents the integrated resistivity over the electrically heterogeneous soil. In the more advanced equipment, multiconductor cables connect a large number of electrodes along a profile. A receiver then cycles sequentially through the electrodes, producing a series of closely spaced soundings along the profile. A computer program then performs 2-D or 3-D resistivity inversions on the field data. The inversion, however, is not unique and is dependent on the assumed "model," e.g., the number of soil layers and their thickness. This technique may not be robust enough to accurately determine a continuously varying salt concentration in a transition zone. In the case of a sharp interface, however, there is a distinct change in electrical conductivity across such interface; hence, the interface location can, generally, be determined quite accurately.[33]
- Frequency-domain electromagnetic methods (FDEMs). The FDEM involves the generation of an electromagnetic field, which induces current in the soil, which in turn causes the subsurface to create a magnetic field. By measuring this secondary magnetic field, various subsurface properties and features can be deduced. The currents in the soil, termed eddy currents, are induced by time-varying magnetic fields produced by a frequency- controlled AC (alternating current) in a transmitter coil. The transmitted electromagnetic field is called the primary field. The induced

eddy currents in the soil produce a secondary field, which is usually 90° out of phase with the primary one. The ratio of the out-of-phase component of the secondary field to the in-phase component of the primary one is an indication of the terrain's conductivity. The depth of investigation by the FDEM is primarily a function of the frequency of the primary field, with lower frequencies having greater penetration.

- Airborne electromagnetic methods. The FDEM creates eddy currents through electromagnetic induction, so that, actually, no contact with the ground is required. This means that a frequency EM system can be flown by fixed-wing aircraft or by helicopters. The typical airborne system uses several receiver-transmitter coil pairs at varying frequencies. These coil pairs are placed in a "bird," which is towed behind the aircraft at elevations of 25–50 m above ground surface. The depth of investigation is determined by the transmitter's frequencies. Common frequencies range from 200 to 56,000 Hz, yielding penetration depths from tens of meters to less than one meter, respectively. The output, as an apparent resistivity map for each frequency, is produced by an inverse method. Interpretation is normally qualitative and typically needs to be calibrated with ground surface data. The airborne method is generally applied to surveys of large areas; it has been successfully applied to the detection of saltwater intrusion in freshwater aquifers and to the exploration of freshwater lenses in saltwater environments.
- Time-domain electromagnetic methods (TDEMs). The TDEM (transient) employs a transmitter that drives an AC through a square loop of insulated electrical cable laid on the ground. The current consists of equal periods of time on and time off, with base frequencies that range from 3 to 75 Hz, producing an electromagnetic field. Similar to the FDEM, the electrical current generates a primary, time-varying electromagnetic field, which in turn creates a secondary electromagnetic field. Time-domain electromagnetic method soundings can be used to detect saltwater at depths of 5 m to several hundred meters below land surface. The TDEM has several significant advantages over DC soundings, notably depths of investigation up to twice the transmitter coil dimension and the ability to sound through a conductive, near-surface unit, such as a clay confining layer. Time-domain electromagnetic method equipment, however, is more expensive and complicated to use than DC equipment, and the interpretation of TDEM data requires sophisticated interpretation software.

Other geophysical investigation techniques that have been employed for the detection of saline water in the subsurface include the ground penetrating radar, the loop-loop electromagnetic method, the very-low-frequency electromagnetic method, and various borehole geophysical methods. More details can be found in Stewart.[32]

Geochemical Investigation

As a part of exploration of freshwater contaminated by intruding saline water, it is important to identify the origin of the latter. In coastal aquifers, seawater encroachment inland is the most common reason for the increase in salinity; however, other sources or processes can contribute to groundwater salinity. Custodio[34] lists a number of salinity sources that can contaminate freshwater supply, which are not directly related to seawater encroachment. These include entrapped fossil seawater; sea-spray accumulation; evaporite rock dissolution; displacement of old saline groundwater from underlying or adjacent aquifers or aquitards through natural or man-imposed advection or by thermal convection; leaking aquitards through fault systems; and anthropogenic pollution from various sources, including sewage effluents, industrial effluents, mine water, road-deicing salts, effluents from water softening or deionization plants, and agriculture return flows.

In general, seawater has a uniform chemistry due to the long residence time of the major constituents. Its main features are[35] predominance of Cl^- and Na^+, with a molar ratio of 0.86; an excess of Cl^- over the alkali ions (Na and K); and Mg greatly in excess of Ca^{2+} (Mg/Ca = 4.5–5.2). In contrast, continental fresh groundwater is characterized by a highly variable chemical composition, although the predominant anions are HCO_3^- , SO_3^{-4} and Cl^-. If not anthropogenetically polluted, the fundamental cations are Ca^{2+} and Mg^{2+} and, to a lesser extent, the alkali ions, Na^+ and K^+. In most cases, Ca^{2+} predominates over Mg^{2+}. Seawater solutes are specifically characterized by $Mg > SO_4 + HCO_3$, whereas meteoric waters (dilute or saline), even if dominated by resolution of marine salts, reflect Na > Cl. In contrast, sedimentary basin fluids can carry significant Ca and perhaps K excess over $SO_4 + HCO_3$, due to diagenetic carbonate or silicate reactions.

Geochemists generally use the following criteria to define the signature and to distinguish the sources of salinization:[35]

- Salinity: In coastal aquifers, a time series of steadily increasing chloride concentrations can indicate the early evolution of a salinity breakthrough from seawater, due to the overexploitation of groundwater and reduction of the piezometric head.
- Cl/Br ratios: The Cl/Br ratio can be used as a reliable tracer, as both Cl and Br usually behave conservatively, except in the presence of very high amounts of organic matter. Seawater (Cl/Br weight ratio = 297) is distinguished from relics of evaporated seawater (hypersaline brine Cl/Br <297), evaporite-dissolution products (over 1000), and anthropogenic sources like sewage effluents (Cl/Br ratios up to 800) or agriculture return flows (low Cl/Br ratios).
- Na/Cl ratios: Na/Cl ratios of saltwater intrusion are usually lower than the marine values (i.e., <0.86 molar ratio). On the other hand, high (>1) Na/Cl ratios typically characterize anthropogenic sources, like domestic waste waters. Thus, low Na/Cl ratios, combined with other geochemical parameters, can foretell the arrival of saltwater intrusion, even at relatively low chloride concentrations, during early stages of salinization.
- Ca/Mg, $Ca/(HCO_3 + SO_4)$ ratios: One of the most conspicuous features of saltwater intrusion is the enrichment of Ca over its concentration in seawater. High Ca/Mg and $Ca/(HCO_3 + SO_4)$ ratios (>1) are further indicators of the intrusion of seawater.
- O and H isotopes: The stable O and H isotopes can be used to describe the mixing process between saline water and freshwater. Fresh groundwater is generally depleted in both ^{18}O and ^2H (deuterium) relative to seawater. Mixing of freshwater and seawater should result in a linear correlation. Different sources with high salinity (e.g., agriculture return flows, sewage effluents) would result in different slopes due to evaporation processes that would change the isotopic composition.
- Boron isotopes: One of the processes that modify the chemistry of seawater intrusion is the adsorption of potassium, boron, and lithium onto clay minerals in the host aquifer. These elements are relatively depleted in saline water associated with seawater intrusion. Hence, the boron isotopic composition of groundwater can be used as a tool to discern the salinization sources, in particular, to distinguish seawater from anthropogenic contamination, such as domestic waste water.

As a conclusion of this section, we observe that many of the techniques discussed above for the exploration of an environment intruded by saline water are of a qualitative nature. Hence, in island or coastal regions, an exploration program usually requires the conjunctive and integrated use of two or more complementary geophysical, geological, and geochemical methods. Using several methods can increase the confidence level of the interpretation of observed data. The collected data can be used to validate the numerical model constructed for the simulation of seawater intrusion for management purposes.

Mathematical Modeling

The use of field surveys, such as geophysical and geochemical studies, can reveal the present state of saltwater intrusion and perhaps some insight into its history. However, it cannot make predictions into the future and, particularly, cannot be used for scenario building and impact assessment based on different levels of anthropogenic activities. Mathematical models are needed for these purposes.

The Ghyben–Herzberg relation is a highly simplified model. More rigorously, the dynamic movement of groundwater flow and the solute transport of salt need to be considered. Generally speaking, there does not exist a sharp division between saltwater and freshwater zones, as implied in Figure 1. The salt concentration continuously changes from that of seawater to that of freshwater. A solute transport model including advection and dispersion is needed for the modeling. In addition, the salt at higher concentration is an active solute, because it can affect the density of water and can drive the flow. Hence a density- dependent solute transport model should be used. There are occasions, however, when the predominant change of concentration from saltwater to freshwater takes place within a narrow region called the transition zone. In that case, a simplified model using the sharp interface assumption can be attempted. Furthermore, if the aquifer modeled is of regional scale, then the flow variables are often averaged in the depth direction to reduce the three-dimensional problems to two-dimensional ones.

While these different models are discussed in more detail in Bear[36] and Bear and Cheng,[26] in the following, only the governing equations for the density-dependent solute transport model are presented.

First, rather than using a piezometric head h associated with a variable density p, we shall define a reference piezometric head h', associated with the (constant) freshwater density ρ *f*, as[37]

$$h'(x,y,z,t)=z+\frac{p}{\rho_f g} \tag{4}$$

In the above, p is pressure, z is the elevation above datum, and g is gravity acceleration. Next, we define the mass fraction (mass of dissolved salt, or TDS, per unit mass of fluid) as

$$\omega=\frac{c}{\rho} \tag{5}$$

where c is the concentration of salt in fluid (mass of dissolved salt per unit volume of fluid). The normalized salt mass fraction is defined as

$$\mathrm{C}=\frac{\omega-\omega_f}{\omega_{s-}\omega_f};\ 0\le C\le 1.0 \tag{6}$$

where ω, ω_s, ω_f, are respectively the salt mass fraction in mixed (salt- and fresh-) water, unmixed seawater, and unmixed freshwater.

Constitutive Equations

Constitutive equations express the relationship between the fluid density and viscosity with the pressure and the salt mass fraction. For fluid density, we ignore the pressure effect and present only the linearized constitutive equations here:

$$\rho=\rho_f\,(1+\beta_c C) \tag{7}$$

where β_c is a density difference factor defined as

$$\beta_c=\beta''_\omega\left(\omega_s-\omega_f\right) \tag{8}$$

and β''_ω is a coefficient expressing the effect of the change in salt mass fraction on the fluid density (at constant pressure), given by

$$\beta''_\omega=\frac{1}{-\rho_f}\frac{d\rho}{d\omega}. \tag{9}$$

The constitutive equation for the fluid dynamic viscosity can be expressed as[38]

$$\mu = \mu_0 \left(1 + 1.85\omega - 4.1\omega^2 + 44.5\omega^2\right) \tag{10}$$

in which the reference viscosity μ_o corresponds to viscosity at $\omega = 0$.

Darcy's Law

Darcy's law can be written in terms of h' and C as

$$\mathrm{q}\frac{\mathrm{K}_f}{\mu_r} = \left(\nabla h' + \beta_c C \nabla z\right) \tag{11}$$

where $\mu_r = \mu/\mu_o$ is the relative viscosity,

$$\mathrm{K}_f \frac{\rho_f \mathrm{gk}}{\mu_f} \tag{12}$$

is the hydraulic conductivity with respect to the reference density and viscosity (of freshwater), and k is the intrinsic permeability.

Mass Balance Equation for Water

The mass balance of water can be expressed in terms of h' and C as

$$\mathrm{S}_o \frac{\partial h'}{\partial t} + \phi\beta_c \frac{\rho}{\rho_f}\frac{\partial C}{\partial t} = \nabla \cdot \left[\frac{1+\beta_c C}{\mu}\mathbf{K}_f \cdot \left(\nabla h' + \beta_c C \nabla_z\right) + \phi\beta_c \mathbf{D} \cdot \nabla C\right]$$

$$+ \frac{\rho_R}{\rho_f} Q_R - \left(1 + \beta_c C\right)\mathrm{Q}_p \tag{13}$$

where S_o is the aquifer's specific storativity; ϕ is porosity; $\mathbf{D}$ is the coefficient of mechanical dispersion, a second- rank tensor; Q_R and Q_P are, respectively, rate of injection (recharge) and withdrawal (pumping) of water from the aquifer; and ρ_R is the density of the recharged water.

Mass Balance Equation for Dissolved Salt

The mass balance of TDS can be expressed as

$$\frac{\partial \phi\rho C}{\partial t} = -\nabla \cdot \left(\rho C \mathbf{q} - \phi\rho \mathbf{D}_h \cdot \nabla C\right) + \rho_R C_R Q_R - \rho C \mathrm{Q}_P \tag{14}$$

where $\mathrm{D}_h = \mathrm{D} + \mathrm{D}^*$ is the coefficient of hydrodynamic dispersion, D^* is the coefficient of molecular diffusion, and C_R is the salt mass fraction of recharged water.

Well-Posed Initial and Boundary Value Problem

Eq. 11, Darcy's law, predicts the specific discharge q, which can be used to eliminate that quantity in Eqs. 13 and 14. The equations of state given in the Constitutive Equations subsection tie the properties, such as density and viscosity, to the normalized salt mass fraction C. Hence, the two mass balance equations, Eqs. 13 and 14, contain two unknowns, h' and C. Given a domain, together with a set of well-posed initial and boundary conditions (see Bear[36] or Bear and Cheng[26] for a full description of initial and boundary conditions under various physical conditions), the system can be solved for these two unknowns.

Computer Models

With the exception of some simple geometries of saltwater intrusion, for which analytical solutions are available,[39] numerical solutions are needed for practical applications. Two of the most widely used computer codes are the SHARP[40,41] for sharp interface model and the SUTRA[42,43] for the density-dependent solute transport model; both are developed by the U.S. Geological Survey (USGS). However, like many complex engineering problems, there is no single code that can be most versatile, efficient, accurate, and stable at the same time, thus dominating the rest of the codes. Depending on the availability and reliability of input data, and the limited resources dedicated to modeling, different computer codes have been developed to offer a wide range of choices. A few of these codes are listed below. A comprehensive survey can be found in Sorek and Pinder[44] and Bear and Cheng.[26]

- SHARP (a quasi-three-dimensional, numerical finite difference method that simulates freshwater and saltwater flow separated by a sharp interface in layered coastal aquifer systems)[40,41] is an implicit finite-difference code that simulates layered aquifers with Dupuit assumption.
- MOCDENSE (a two-constituent solute transport model for groundwater with variable density)[45] and MOCDENSE3D[46] are, respectively, the two-dimensional and three-dimensional rendition of MOC (computer model of two-dimensional solute transport and dispersion in groundwater)[47] and MOC3D (three dimensional method-of-characteristics groundwater flow and transport model) [48] to allow for variable- density modeling.
- SUTRA (model for 2D or 3D saturated-unsaturated, variable-density groundwater flow, with solute or energy transport)[42,43] is a Galerkin finite-element code that solves groundwater flow and transport problems under saturated and unsaturated conditions.
- SEAWAT (computer program for simulation of three dimensional variable-density groundwater flow)[49] is a USGS code that combines MODFLOW[50] and MT3DMS[51] into a single computer program for the purpose of simulating saltwater intrusion. It is a finite- difference, Eulerian-Lagrangian code, in contrast to SUTRA, which is a finite-element code.
- CODESA-3D (coupled variable-density and saturation 3D model)[52] is0 a finite-element code, similar to SUTRA.
- FEMWATER (three-dimensional finite-element model of water flow through saturated-unsaturated media).[53] The code is available from the U.S. Environmental Protection Agency.

Combating Saltwater Intrusion

One of the most effective ways of combating saltwater intrusion is to regulate pumping activities. Generally speaking, the amount of groundwater extraction should not exceed that of natural replenishment. Optimization of pumping patterns to maximize the yield and minimize the extent of intrusion is a high-priority management issue. Recharge of natural surface water or reclaimed wastewater into aquifers can increase the freshwater outflow rate to push back saltwater wedge. A recharge near the coast can build a local freshwater mound that forms a barrier to protect the water table depression inland. Extraction of saltwater in an invaded saltwater wedge can also protect the freshwater behind, if a proper way can be found to dispose of the extracted saltwater. A similar method involving pumping simultaneously in the upper freshwater zone and the lower saltwater zone to prevent up-coning, known as double pumping, has been attempted. Using collector wells (horizontal wells) to skim the thin layer of freshwater floating on top of the saltwater wedge has been effectively used in water-poor countries such as Israel. Land reclamation has the added effect of pushing saltwater to the sea. Finally, in places where large freshwater springs flowing to the sea can be identified, physical barriers, such as solid walls or slurry curtains, can be used to intercept freshwater.

Planning and Management

Groundwater resource planning and management in coastal areas is similar to the traditional water resource planning in inland areas but with the additional complication of saltwater intrusion to eventually render portions of the coastal aquifer unusable as a source of drinking water. The recommended planning elements for managing coastal aquifers include the following:[54]

1. Collect and analyze existing data: Data like piezometric head, chlorine concentration, aquifer transmissivity, formation thickness, well location, screen depth, etc., need to be collected.
2. Develop an integrated database: The data need to be organized and integrated, typically using geographic information system software.
3. Identify saltwater intrusion problem and hypothesize intrusion scenarios, such as horizontally from the sea or vertically from the bottom brackish water.
4. Develop a groundwater simulation model using one of the above-mentioned computer codes and collected data.
5. Perform field studies for model validation: The collected existing data are likely to be insufficient, and the computer simulations are likely to be unreliable, lacking field validation. Hence, field studies are needed to fill the information gap. The geological, geophysical, and geochemical methods reviewed above can be used to conduct field studies.
6. Set planning objective: Once the extent and the time scale of the threat are understood, planning objectives, such as reducing pumping, increasing recharge, etc., should be set by consulting with the stakeholders.
7. Identify solutions and actions: One or more of the saltwater intrusion combating strategies can be recommended.
8. Evaluate management alternatives: Water resource management involves complex tradeoffs. The alternative actions need to be evaluated and selected using a multicriteria decision-making process.

References

1. Back, W.; Freeze, R.A., Eds. *Chemical Hydrogeology. Benchmark Papers in Geology*; Hutchinson Ross Publ. Co.: Stroudsburg, PA, 1983; Vol. 73.
2. Cooper, H.H.J.; Kohout, F.A.; Henry, H.R.; Glover, R.E. *Sea Water in Coastal Aquifers*; U.S. Geological Survey Water-Supply Paper 1613-C; 1964.
3. Sonenshein, R.S. Delineation and Extent of Saltwater Intrusion in the Biscayne Aquifer, Eastern Dade County, Florida, U.S. Geological Survey Water-Resources Investigation Report 96–4285; 1997.
4. Konikow, L.F.; Reilly, T.E. Seawater intrusion in the United States. In Seawater Intrusion in Coastal Aquifers— Concepts, Methods and Practices; Bear, J., et al., Eds; Kluwer, 1999; 463–506.
5. Johnson, T.A.; Whitaker, R. Saltwater intrusion in the coastal aquifers of Los Angeles County, California. In *Coastal Aquifer Management-Monitoring, Modeling, and Case Studies*; Cheng, A.H.D., Ouazar, D., Eds.; Lewis Publ., 2003; 29–48.
6. Mink, J.F. State of the groundwater resources of southern Oahu. In *Board of Water Supply, City and County of Honolulu*; 1980; 83 pp.
7. Voss, C.I.; Souza, W.R. Dynamics of a Regional Freshwater-Saltwater Transition Zone in an Anisotropic Coastal Aquifer System, U.S. Geological Survey Open-File Report 98–398; 1998; 88 pp.

8. Stakelbeek, A. Movement of brackish groundwater near a deep-well infiltration system in the Netherlands. In *Seawater Intrusion into Coastal Aquifers—Concepts, Methods and Practices;* Bear, J., et al., Eds.; Kluwer, 1999; 531–541.
9. Oude Essink, G. Impact of sea level rise in the Netherlands. In *Seawater Intrusion into Coastal Aquifers—Concepts, Methods and Practices;* Bear, J., et al., Eds.; Kluwer, 1999; 507–530.
10. Melloul, A.J.; Zeitoun, D.G. A semi-empirical approach to intrusion monitoring in Israeli coastal aquifer. In Seawater Intrusion into Coastal Aquifers—Concepts, Methods and Practices; Bear, J., et al., Eds.; Kluwer, 1999; 544–558.
11. Melloul, A.J.; Collin, M.L. Sustainable groundwater management of the stressed coastal aquifer in the Gaza region. Hydrolo. Sci. J. **2000**, *45* (1), 147–159.
12. Steinich, B.; Marin, L.E. Hydrogeological investigations in northwestern Yucatan, Mexico, using resistivity surveys. Ground Water **1996**, *34* (4), 640–646.
13. Marin, L.E.; Perry, E.C.; Essaid, H.I.; Steinich, B. Hydrogeological investigations and numerical simulation of groundwater flow in the karstic aquifer of Northwestern Yucatan, Mexico. In *Coastal Aquifer Management Monitoring, Modeling, and Case Studies;* Cheng, A.H.D., Ouazar, D., Eds.; Lewis Publishers, 2003; 257–277.
14. de Franco, R.; Biella, G.; Tosi, L.; Teatini, P.; Lozej, A.; Chiozzotto, B.; Giada, M.; Rizzetto, F.; Claude, C.; Mayer, A.; Bassan, V.; Gasparetto-Stori, G. Monitoring the saltwater intrusion by time lapse electrical resistivity tomography: The Chioggia test site (Venice Lagoon, Italy). J. Appl. Geo- phys. 2009, *69* (3–4), 117–130.
15. Viezzoli, A.; Tosi, L.; Teatini, P.; Silvestri, S. Surface water-groundwater exchange in transitional coastal environments by airborne electromagnetics: The Venice Lagoon example. Geophys. Res Lett. **2010**, *37,* L01402.
16. Sherif, M. Nile Delta aquifer in Egypt. In *Seawater Intrusion in Coastal Aquifers—Concepts, Methods and Practices;* Bear, J., et al., Eds.; Kluwer, 1999; 559–590.
17. Paniconi, C.; Khlaifi, I.; Lecca, G.; Giacomelli, A.; Tar- houni, J. Modeling and analysis of seawater intrusion in the coastal aquifer of eastern Cap-Bon, Tunisia. Transp. Porous Media **2001**, *43* (1), 3–28.
18. Wu, J.C.; Meng, F.H.; Wang, X.W.; Wang, D. The development and control of the seawater intrusion in the eastern coastal of Laizhou Bay, China. Environ. Geol. **2008**, *54* (8), 1763–1770.
19. Ergil, M.E. The salination problem of the Guzelyurt aquifer, Cyprus. Water Res. **2000**, *34* (4), 1201–1214.
20. Benkabbour, B.; Toto, E.A.; Fakir, Y. Using DC resistivity method to characterize the geometry and the salinity of the Plioquaternary consolidated coastal aquifer of the Mamora plain, Morocco. Environ. Geol. **2004**, *45* (4), 518–526.
21. Du Commun, J. On the cause of freshwater springs, fountains, etc. Am. J. Sci. Arts **1828**, *14,* 174–175.
22. Badon-Ghyben, W. Nota in verband met de voorgenomen putboring nabij Amsterdam (Notes on the Probable Results of Well Drilling near AMSTERDAM); Tijdschrift van het Koninklijk Instituut van Ingenieurs: Hague, 1888; 822.
23. Herzberg, A. Die Wasserversorgung einiger Nordseeb- der (The water supply of parts of the North Sea Coast in Germany). Z. Gasbeleucht. Wasserversorg. **1901**, *44/45,* 815–819, 842–844.
24. Bear, J.; Dagan, G. Some exact solutions of interface problems by means of hodograph method. J. Geophys. Res. **1964**, *69* (8), 1563–1572.
25. Dagan, G.; Bear, J. Solving the problem of local interface upconing in a coastal aquifer by the method of small perturbation. J. Hydraul. Res. **1968**, **6**, 15–44.
26. Bear, J.; Cheng, A.H.D. *Modeling Groundwater Flow and Contaminant Transport;* Springer, 2010; 834 pp.
27. Schmorak, S.; Mercado, A. Upconing of freshwater-saltwater interface below pumping wells, field study. Water Resour. Res. **1969**, **5** (6), 1290–1311.

28. Aliewi, A.S.; Mackay, R.; Jayyousi, A.; Nasereddin, K.; Mushtaha, A.; Yaqubi, A. Numerical simulation of the movement of saltwater under skimming and scavenger pumping in the Pleistocene aquifer of Gaza and Jericho areas, Palestine. Transp. Porous Media **2001**, *43* (1), 195–212.
29. Saeed, M.M.; Ashraf, M.; Asghar, M.N. Hydraulic and hydro-salinity behavior of skimming wells under different pumping regimes. Agric. Water Manage. **2003**, *61* (3), 163–177.
30. Saeed, M.M.; Ashraf, M. Feasible design and operational guidelines for skimming wells in the Indus basin, Pakistan. Agric. Water Manage. **2005**, *74* (3), 165–188.
31. Lusczynski, N.J. Head and flow of ground water of variable density. J. Geophys. Res. **1961**, *66*, (12), 4247–4256.
32. Stewart, M.T. Geophysical investigations. In *Seawater Intrusion into Coastal Aquifers—Concepts, Methods and Practices;* Bear, J., et al., Eds.; Kluwer, 1999; 9–50.
33. Stewart, M.T. Rapid reconnaissance mapping of freshwater lenses on small oceanic islands. In *Geotechnical and Environmental Geophysics, v. II, Soc. Explor. Geophysicists, Inv. in Geophysics, n. 5;* Ward, S., Ed.; 1990; 57–66.
34. Custodio, E. Studying, monitoring and controlling seawater intrusion in coastal aquifers. Guidelines for study, monitoring and control. *FAO Water Reports* **1997**, *11*, 7–23.
35. Jones, B.F.; Vengosh, A.; Rosenthal, E.; Yechieli, Y. *Geochemical investigations*. In *Seawater Intrusion into Coastal Aquifers—Concepts, Methods and Practices;* Bear, J., et al., Eds.; Kluwer, 1999; 51–71.
36. Bear, J. Conceptual and mathematical modeling. In *Seawater Intrusion into Coastal Aquifers—Concepts, Methods and Practices;* Bear, J., et al., Eds.; Kluwer, 1999; 127161.
37. Bear, J.; Zhou, Q. Sea water intrusion into coastal aquifers. In *The Handbook of Groundwater Engineering,* 2nd Ed.; Delleur, J.W., Ed.; CRC Press, 2006.
38. Lever, D.A.; Jackson, C.P. On the Equations for the Flow of a Concentrated Salt Solution through a Porous Medium, Harwell Report AERE-R 11765; HMSO: London, 1985.
39. Cheng, A.H.D.; Ouazar, D. Analytical solutions. In *Seawater Intrusion in Coastal Aquifers—Concepts, Methods, and Practices;* Bear, J., et al., Eds.; Kluwer, 1999; 163–191.
40. Essaid, H.I. The Computer Model SHARP, a Quasi-ThreeDimensional Finite-Difference Model to Simulate Freshwater and Saltwater Flow in Layered Coastal Aquifer Systems, U.S. Geological Survey Water-Resources Investigations Report 90–4130; 1990; 181 pp.
41. Essaid, H.I. USGS SHARP model. In Seawater Intrusion in Coastal Aquifers—Concepts, Methods, and Practices; Bear, J., et al., Eds.; Kluwer, 1999; 213–247.
42. Voss, C.I. USGS SUTRA code—History, practical use, and application in Hawaii. In *Seawater Intrusion in Coastal Aquifers—Concepts, Methods and Practices;* Bear, J., et al., Eds.; Kluwer, 1999; 249–313.
43. Voss, C.I.; Provost, A.M. SUTRA, A Model for Saturated- Unsaturated Variable-Density Ground-Water Flow with Solute or Energy Transport, *U.S. Geological Survey Water- Resources Investigations Report 02–4231;* 2002.
44. Sorek, S.; Pinder, G.F. Survey of computer codes and case histories. In *Seawater Intrusion into Coastal Aquifers— Concepts, Methods and Practices;* Bear, J., et al., Eds.; Kluwer, 1999; 399–461.
45. Sanford, W.E.; Konikow, L.F. *A Two-Constituent Solute Transport Model for Ground Water Having Variable Density,* U.S. Geological Survey Water Resources Investigation Report 85–4279; 1985; 89 pp.
46. Oude Essink, G.H.P. Modeling three-dimensional density dependent groundwater flow at the Island of Texel, the Netherlands. In *Coastal Aquifer Management—Monitoring, Modeling, and Case Studies;* Cheng, A.H.D., Ouazar, D., Eds.; Lewis Publ, 2004; 77–94.
47. Konikow, L.F.; Bredehoeft, J.D. Computer model of twodimensional solute transport and dispersion in ground water. In *Techniques of Water-Resources Investigations of the U.S. Geological Survey*; 1978; 90 pp.

48. Konikow, L.F.; Goode, D.J.; Hornberger, G.Z. *A ThreeDimensional Method-of-Characteristics Solute-Transport Model (MOC3D),* U.S. Geological Survey Water-Resources Investigations Report 96–4267; 1996; 87 pp.
49. Langevin, C.D.; Shoemaker, W.B.; Guo, W. MODFLOW-2000, the U.S. Geological Survey Modular Ground-Water Model—Documentation of the SEAWAT-2000 Version with the Variable-Density Flow Process (VDF) and the Inte- gratedMT3DMS Transport Process (IMT), U.S. Geological Survey Open-File Report 03–426; 2003; 43 pp.
50. Harbaugh, A.W. MODFLOW-2005, the U.S. Geological Survey Modular Ground-Water Model—The GroundWater Flow Process, U.S. Geological Survey Techniques and Methods 6-A16; 2005.
51. Zheng, C. MT3DMS v5.2 supplemental user's guide. In *Technical Report to the U.S. Army Engineer Research and Development Center*; Department of Geological Sciences, University of Alabama; 2006; 41 pp.
52. Gambolati, G.; Putti, M.; Paniconi, C. Three-dimensional model of coupled density-dependent flow and miscible salt transport. In *Seawater Intrusion in Coastal Aquifers— Concepts, Methods, and Practices;* Bear, J., et al., Eds.; Kluwer, 1999; 315–362.
53. Lin, H.C.; Richards, D.R.; Yeh, G.T.; Cheng, J.R.; Chang, H.P.; Jones, N.L. *FEMWATER: A Three-Dimensional Finite Element Computer Model for Simulating Density Dependent Flow and Transport,* U.S. Army Engineer Waterways Experiment Station Technical Report; 1996; 129 pp.
54. Maimone, M.; Harley, B.; Fitzgerald, R.; Moe, H.; Hos-sain, R.; Heywood, B. Coastal aquifer planning elements. In *Coastal Aquifer Management—Monitoring, Modeling, and Case Studies;* Cheng, A.H.D., Ouazar, D., Eds.; Lewis Publishers, 2003; 1–27.

14

Irrigation Systems: Water Conservation

Pai Wu, Javier
Barragan, and
Vince Bralts

Introduction

With the increasing consequence of limited water resources and the increasing need for environmental protection, drip irrigation will play an even more important role in the future. Drip irrigation systems can be used for many different types of agricultural crops, including fruit trees, vegetables, pastures, specialty crops such as sugarcane, ornamentals, golf course grasses, and high economic value crops grown in greenhouses. An understanding of drip irrigation systems, irrigation scheduling, crop response, and economic ramifications will encourage greater use of drip irrigation in future agricultural production.

Uniformity of Water Application and Design Considerations

The desired uniformity of water application and the specific crops to be grown guides the creation of drip irrigation systems. There are two types of drip irrigation uniformity: system uniformity and spatial uniformity in the field. The consistency of system distribution of water into the field describes the system uniformity. The spatial uniformity is the regularity of water distribution considering overlapping emitter flow and translocation of water in the soil. For drip irrigation systems designed for trees with large spacing, the system uniformity is equal to the water application uniformity in the field. For high-density plantings, the emitter spacing should be designed considering overlapped wetting patterns and the spatial uniformity in the field. The uniformity of a drip irrigation system depends primarily on the hydraulic design, but must also consider the manufacturer's variation, temperature effects, and potential emitter plugging. The effect of water temperature is generally negligible when using turbulent flow emitters. A combination of proper filtration and turbulent emitters can control emitter plugging. When grouping a number of emitters together as a unit, such as those designed to irrigate an individual plant's root system, the uniformity of water application with respect to the plant will improve.

Many expressions have been used to describe uniformity. The system uniformity, or emitter flow uniformity, can be expressed as the range or variation of water distribution in the field. This term was initially used for hydraulic design of drip irrigation systems given that the minimum and maximum emitter flows could be calculated and determined.[1] When more emitter flows are used or more samples are required for determining variation or spatial uniformity in the field, the Christiansen uniformity coefficient (UCC)[2] and coefficient of variation (CV), which is the ratio of standard deviation and the mean, are used. Each of the uniformity expressions are highly correlated with one other.

Hydraulic Design of Drip Irrigation Systems

Once selection of the type of drip irrigation emitter is complete, the hydraulic design can be made to achieve the expected uniformity of irrigation application.

The hydraulic design of a drip irrigation system involves designing both the submain and lateral lines. Early research in drip irrigation hydraulic design concentrated mainly on the single lateral line approach,[1,3,4] but in 1985 Bralts and Segerlind developed a method to design a submain unit. The hydraulic design is based on the energy relations in the drip tubing, the friction drop, and energy changes due to slopes in the field. Direct calculations of water pressures along a lateral line or in a submain unit are made by using an energy gradient line approach.[1] All emitter flows along a lateral line and in a submain can be determined based on their corresponding water pressures. Once the emitter flows are determined, the emitter flow variation, q_{var} is expressed by

$$q_{\mathrm{var}} = \frac{q_{\max} - q_{\min}}{q_{\max}} \tag{1}$$

where $q_{\max}$ is the maximum emitter flow and $q_{\min}$ is the minimum emitter flow. Based on these data, other uniformity parameters such as UCC and CV can also be determined. There is a strong correlation between any two of the three uniformity parameters in the hydraulic design of drip irrigation systems, thus any one of the uniformity parameters can be used as a design criterion. This correlation also justifies using the simple emitter flow variation q_{var} for hydraulic design. The emitter flow variation q_{var} is converted to the CV when it is combined with the manufacturer's variation of emitter flow.

The total emitter flow variation caused by both hydraulic and manufacturer's variation can be expressed by[5]

$$\mathrm{CV_{HM}} = \sqrt{\mathrm{CV_H^2}} + \mathrm{CV_M^2} \tag{2}$$

where $\mathrm{CV_{HM}}$ is the coefficient of variation of emitter flows caused by both hydraulic and manufacturer's variation; $\mathrm{CV_H}$ and $\mathrm{CV_M}$ are the coefficients of variation of emitter flows caused by hydraulic design and manufacturer's variation, respectively.

The design criterion for emitter flow variation q_{var} for drip irrigation design is arbitrarily set as 10.0%–20.0%, which is equivalent to a CV, from 0.033 to 0.076, or 3.0%–8.0%. Based on the research of last 30 years, the manufacturer's variation of turbulent emitters is maintained only in a range 3.0%–5.0%, expressed by CV. When this variation is combined with emitter flow variation caused by hydraulic design with a range 3.0%–8.0% in CV, the total variation determined by the equation above will be limited to a CV of less than 10.0%. This variation illustrates that the drip irrigation systems are designed to achieve high uniformity and irrigation application efficiency.

Economic return can also be the basis of design criteria for drip irrigation. A new set of design criteria for drip irrigation was developed,[6] based on achieving an expected economic return with various water resources and environmental considerations (Table 1).

TABLE 1 Design Criteria for Uniformity of Drip Irrigation System Design

Design Consideration	CV (%)	UCC (%)
Water is abundant and no environmental	30–20	75–85
pollution problems	20–10	80–90
Water is abundant but with environmental protection considerations	25–15	80–90
Limited water resources but with no environmental pollution problems	15–5	85–95
Considerations for both water conservation and environmental protection		

Drip Irrigation for Optimal Return, Water Conservation, and Environmental Protection

When the uniformity of a drip irrigation system is designed with a UCC of 70.0%, 30.0% or less in CV, the irrigation application is expressed as a straight-line distribution,[7,8] as shown in Figure 1. This figure was plotted using percent of area (PA) against a relative irrigation depth, *X*, which is the ratio of required irrigation depth to mean irrigation application. The straight-line distribution in the dimensionless plot can be specified by a minimum value, *a*, a maximum value, $(a+b)$, in the *X*-scale and a slope *b*, where *b* specifies the uniformity of water application.[9]

When a drip irrigation system is designed with fixed uniformity, it is possible to determine the sloped straight line with known value of *a* and *b*. A value (*X*) can then be selected between value *a* and $(a+b)$ and plotted (Figure 1). The triangle formed above the horizontal line (*X*) results in an irrigation deficit and yield reduction. The triangle below the horizontal line results in over-irrigation and deep seepage.

An important irrigation scheduling parameter, the relative irrigation depth, (*X*) indicates how much irrigation water is applied. The effectiveness of drip irrigation is shown not only by the high uniformity of the drip irrigation system, but also by the irrigation requirement and the strategy of irrigation scheduling. As illustrated in Figure 1, the irrigation scheduling parameter (*X*) affects the areas of overirrigation and water deficit conditions in the field and is directly related to the economic return. Practically

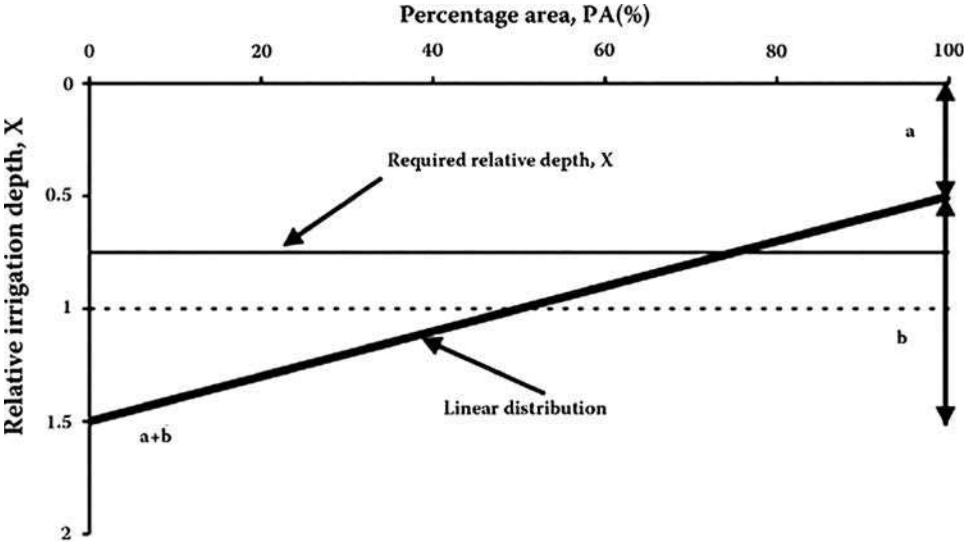


FIGURE 1 A linear water application model for drip irrigation.

speaking, the X parameter is selected in a range from a to $(a+b)$, as shown in Figure 1. Three typical irrigation schedules can be expressed by X and are as follows:

$X=a$	This schedule is a conventional irrigation schedule, which is based on the minimum emitter or minimum water application. The field is fully irrigated and whole field is over-irrigated except the point of minimum irrigation application.
$X=X_0$	For an optimal return there is a value of X for the irrigation scheduling parameter between a and (a+b).
$X=(a+b)$	This irrigation schedule is based on the maximum emitter flow or maximum irrigation application. The whole field is under deficit condition except the point of maximum water application. There is no deep percolation.

An optimal irrigation schedule for maximum economic return was determined[9] based on cost of water, price of the yield, and damage such as environmental pollution and groundwater contamination caused by over irrigation. Different irrigation strategies require different amounts of water application. Water conservation and environmental protection are realized by comparing any two of the irrigation strategies.[10]

Conclusion

Drip irrigation is an irrigation method that can distribute irrigation water uniformly and directly into the root zone of crops. It is one of the most efficient irrigation methods and can be designed and scheduled to meet the water requirement of crop and produce maximum yield in the field.

When the drip irrigation system is designed with high uniformity, the slope b of the straight line of water application function (Figure 1) can be controlled to achieve the desired variation. In this case the conventional irrigation schedule, $X=a$, optimal irrigation schedule, X_0, which is a location between a and $(a+b)$, and the irrigation schedule for environmental protection, $X=a+b$, are in close proximity. This closeness shows that the drip irrigation system can achieve optimal economic return, water conservation, and environmental protection.

References

1. Wu, I.P.; Gitlin, H.M. *Design of Drip Irrigation Lines*; HAES Technical Bulletin 96, University of Hawaii: Honolulu, HI, 1974; 29 pp.
2. Christiansen, J.E. The uniformity of application of water by sprinkler systems. Agric. Eng. **1941**, *22*, 89–92.
3. Howell, T.A.; Hiler, E.A. Trickle irrigation lateral design. Trans. ASAE **1974**, *17* (5), 902–908.
4. Bralts, V.F.; Segerlind, L.J. Finite elements analysis of drip irrigation submain unit. Trans. ASAE **1985**, *28,* 809–814.
5. Bralts, V.F.; Wu, I.P.; Gitlin, H.M. Manufacturing variation and drip irrigation uniformity. Trans. ASAE **1981**, *24* (1), 113–119.
6. Wu, I.P.; Barragan, J. Design criteria for microirrigation systems. Trans. ASAE **2000**, *43* (5), 1145–1154.
7. Seginer, I. A note on the economic significance of uniform water application. Irrig. Sci. **1978**, *1*, 19–25.
8. Wu, I.P. Linearized water application function for drip irrigation schedules. Trans. ASAE **1988**, *31* (6), 1743–1749.
9. Wu, I.P. Optimal scheduling and minimizing deep seepage in microirrigation. Trans. ASAE **1995**, *38* (5), 1385–1392.
10. Barragan, J.; Wu, I.P. Optimal scheduling of a microirrigation system under deficit irrigation. J. Agric. Eng. Res. **2001**, *80* (2), 201–208.

15

Irrigation: Erosion

David L. Bjorneberg

Introduction

Irrigation is vital to food production in the world. However, irrigation-induced soil erosion reduces productivity of irrigated land and can cause off-site water quality problems. Surface irrigation utilizes the soil to distribute water through the field. Water flowing over soil inherently detaches and transports sediment. Sprinkler and drip irrigation distribute water through fields in pipes, eliminating erosion from water distribution, but erosion can still occur if water is applied faster than it can infiltrate into the soil. This entry will briefly discuss the importance of irrigation to global food production and then discuss the important factors affecting soil erosion for surface-and sprinkler-irrigated land. Much of the information will focus on the United States, with international information included when possible.

Importance of Irrigation

Irrigated agriculture contributes a disproportionate amount to global food production. The most cited statistics indicate that irrigated cropland produces about one-third of the world's crop production on only 16% of the cropland that is irrigated.[1] In the United States, farms with all cropland irrigated account for only 8% of the total cropland and about half of the total irrigated land.[2] These farms produce 33% of the market value of crops and 12% of the market value of livestock. Over half of the crop value (55%) is produced on farms with some irrigated land, and these farms account for only 26% of the total cropland in the United States.[3] In some areas, irrigation provides essentially all of the water necessary for crop growth. In other areas, irrigation provides only a small portion of the total crop water requirement but reduces the potential for water stress during critical periods.

While irrigation is critical to global food production, applying water to soil can cause erosion. This is especially true with surface irrigation, where the soil conveys and distributes water through a field by gravity. Sprinkler irrigation and microirrigation use pipes to distribute water through the field. Surface irrigation is generally thought to cause more erosion than sprinkler irrigation; however, erosion can occur any time water flows over soil. Water can be applied with sprinkler irrigation so no runoff occurs, and therefore, no erosion will occur. However, there are situations, especially with moving irrigation systems like center pivots, where water is applied faster than it can infiltrate into the soil, resulting in ponding and, possibly, runoff.

Unique Aspects of Irrigation Erosion

The factors affecting soil erosion from irrigation are the same as rainfall. Water detaches and transports sediment in both situations. However, there are some unique differences in how the factors occur with irrigation.[4] For example, rainfall occurs relatively uniformly over an entire field, whereas irrigation is seldom applied to an entire field at the same time. Irrigation is a controlled procedure where water is applied to a specific field, or portion of a field, at a specific time. This can affect the hydrology of the erosion processes on surface- and sprinkler-irrigated fields. A center pivot, for example, is essentially a moving storm that covers only 1%–2% of the field at any given time. This results in unique runoff conditions where water can do the following: 1) flow parallel to the lateral under similar conditions as rainfall; 2) flow from wet soil onto dry soil if the lateral is moving downhill; or 3) flow onto wet soil if the lateral is moving uphill.

In surface irrigation, water flow rate decreases with distance during surface irrigation as water infiltrates. Furrow flow rates also increase with time as infiltration rate decreases (Figure 1). This creates a condition where sediment can be detached on the upper end of the field and deposited on the lower end. Trout[4] documented erosion rates on the upper end of a field that were 6 to 20 times greater than the field-average erosion rates. Figure 2 shows eroded furrows on the upper end of a field after one furrow irrigation. During rainfall, raindrops wet the soil surface and detach soil particles. As runoff begins, rills form in wet soil. In contrast, irrigation furrows are formed prior to irrigation, and water flows onto initially dry soil. Furrows with initially dry soil have greater soil erosion than furrows that were prewet immediately before furrow irrigation.[5] Irrigation water flowing in furrows is not exposed to falling raindrops that can increase sediment detachment and decrease deposition.

The quality of irrigation water can vary dramatically among water sources, or even within an irrigation tract if drainage water is reused. Conversely, electrolyte concentration of rainfall is quite consistent. Electrolyte concentration in irrigation water affects erosion for both surface and sprinkler irrigation. Furrow-irrigation erosion was greater on a silt loam when irrigation water had low electrical conductivity (EC$=$0.7 dS m^{-1}) and high sodium adsorption ratio (SAR$=$9.1) compared with low EC (0.5 dS m^{-1}) and low SAR (0.9), high EC (2.1 dS m^{-1}) and low SAR (0.5), and high EC (1.7 dS m^{-1}) and high SAR (9.3).[6] Soil erosion was also greater with low-EC water in laboratory and field rainfall simulation studies.[7,8] Lower electrolyte concentrations in water cause greater dispersion of soil particles, which tends to reduce infiltration and increase soil loss.[9]

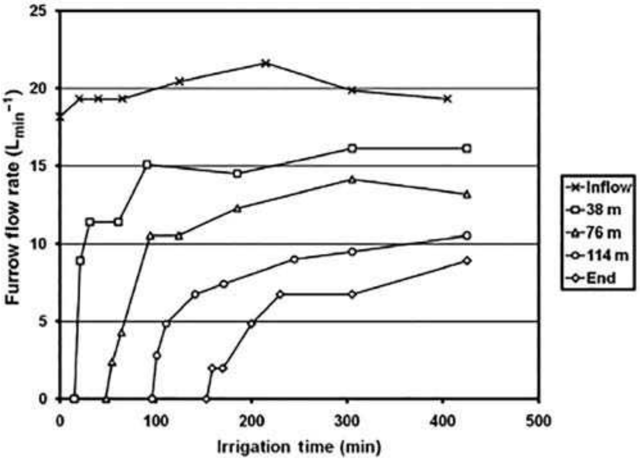


FIGURE 1 Furrow flow rate with time at five points in a 150 m–long furrow.

FIGURE 2 Eroded furrows on the upper end of a furrow-irrigated field in Idaho with approximately 1% slope.

Surface-Irrigation Erosion

Surface irrigation continues to be the most common method of irrigation in the world. The four countries with the most irrigated land are India (60.8 Mha), China (57.8 Mha), United States (22.4 Mha), and Pakistan (19.6 Mha).[10] These four countries account for 58% of the irrigated area in the world. All other countries have less than 10 Mha of irrigated land.[10] According to the country fact sheets on the Food and Agriculture Organization's Aquastat Web site,[11] surface irrigation is used on 97% of the irrigated land in India, 94% in China, 44% in the United States, and 100% in Pakistan.

Koluvek et al.[12] provided a good overview of soil erosion from irrigation in the United States. Unfortunately, this information has not been updated, and similar information is not readily available from other countries, so it is difficult to track erosion trends on irrigated lands. Some early studies documented erosion rates as great as 145 Mg ha^{-1} in 1 h[13] and 40 Mg ha^{-1} in 30 min.[14] While these rates represent extreme conditions that can occur, not typical season-long soil loss rates, these studies indicate the potential severity of the problem. One study measured annual soil losses of 1 to 141 Mg ha^{-1} from 33 fields with silt loam soils.[15] The greatest soil loss occurred on a sugar beet (*Beta vulgaris* L.) field with 4% slope. The authors noted that erosion increased sharply when field slope was greater than 1%. Close-growing crops like alfalfa *(Medicago sativa* L.) or wheat (*Triticum aestivum* L.) on fields with 1% slope had annual soil loss of less than 1 Mg ha^{-1}. A recent study in the same area documented that average soil loss from an 80,000 ha irrigated watershed decreased from 450 kg ha^{-1} in 1970 to less than 50 kg ha^{-1} in 2005.[16] This watershed was approximately 90% furrow irrigated in 1970 and 60% furrow irrigated in 2005. Another study measured daily sediment loads of 0.4 kg ha^{-1} in a watershed with no furrow irrigation compared to 19 kg ha^{-1} in a watershed with 58% of the cropland furrow irrigated.[17] Irrigation method explained 67% of the variation in soil loss measured in April and May in these nine watersheds.

The main factors affecting surface-irrigation erosion are soil type, field slope and flow rate. Soil erosion is typically not a concern where field slopes are less than 0.5% (Figure 3). However, erosion tends to increase exponentially for increasing inflow rate and field slope, with an exponent between 1 and 3 for flow rate, and between 2 and 3 for slope.[12,18,19,20] Increasing inflow rate 20% increased erosion 30% and 70% on the upper quarter of two fields.[4] Increasing inflow rate another 20% increased erosion 50% and 100%, which indicates that the exponent between erosion and flow rate was between 2 and 3.[4] Figure 4 shows soil loss from 10 furrows during a 4 h irrigation at Kimberly, Idaho, with inflow rates randomly set for each furrow.

FIGURE 3 Level furrow irrigation in Arizona. (Photo by Jeff Vanuga, USDA Natural Resources Conservation Service [NRCSAZ02037]).

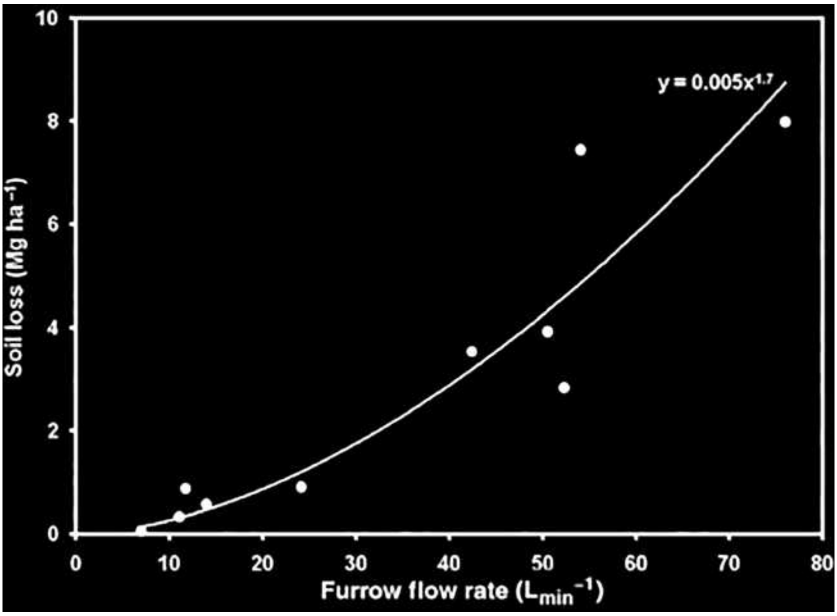


FIGURE 4 Relationship between furrow flow rate and soil loss for 10 and 30 m-long furrows with randomly set inflow rates.

Reducing field slope by grading the land is a costly practice that is not feasible in most situations compared with alternatives like installing a sprinkler-irrigation system. Reducing inflow rate is a good practice as long as the water advances down the field fast enough to uniformly irrigate the field. Slow water advance rates from low inflow rates cause overirrigation on the inflow end of the field and underirrigation on the lower end of the field due to differences in infiltration opportunity time. This results in poor distribution uniformity but little runoff. Soil loss decreases as distribution uniformity decreases.[20] An excellent practice for reducing irrigation erosion without affecting irrigation uniformity is applying small amounts of polyacrylamide (PAM) with irrigation water.[21,22] Dissolving 10 mg L^{-1} of high-molecular-weight, anionic PAM in furrow-irrigation inflow can reduce soil loss 60%–99% compared with untreated furrows. Other technologies like filter strips and sediment ponds on the lower end of the field remove sediment from the water rather than reducing erosion from occurring on the field.

Sprinkler-Irrigation Erosion

Ideally, sprinkler-irrigation systems are designed and managed to have all applied water infiltrate into the soil where it was applied. When all water infiltrates, there is no runoff or soil erosion. Solid-set (sprinklers located in the same position for the entire irrigation season) or set-move (sprinklers remaining in a location for 12 to 24 h, then moved to the next set) irrigation systems usually apply water at a low rate (e.g., 3 to 6 mm h^{-1}), so irrigation application rate does not exceed the soil infiltration rate and no soil erosion occurs. Moving irrigation systems, like center pivots, traveling guns, and lateral-move systems, often apply water faster than the infiltration rate. This occurs because the irrigation system must apply enough water as it moves across the field to meet crop water needs until the next time it irrigates that portion of the field. For example, a center pivot operating at 60 h per revolution needs to apply 20 mm per revolution to meet an 8 mm d^{-1} crop water requirement. The irrigation application rate increases with distance from the center pivot because the lateral irrigates more land as the radial distance from the pivot point increases.[23] Near the pivot point, the mean application rate could be about 4 mm h^{-1} (assuming 15 m wetting diameter). Near the end of the pivot, about 400 m, the mean application rate would be about 60 mm h^{-1}. An important fact about moving irrigation systems is that the application rate is a function of irrigation system capacity, or system flow rate. Operating the system faster decreases only the application depth, not the application rate. For example, the same center pivot operating at 48 h per revolution will apply 16 mm of water at the same application rates.

Center-pivot irrigation is the most popular type of irrigation system in the United States. According to the United States Department of Agriculture (USDA) National Agricultural Statistics Service, center-pivot irrigation was used on 47% of the irrigated land and 83% of the sprinkler-irrigated land in 2008, an increase from 25% of the irrigated land in 1988.[2] More land was irrigated by center pivots in the United States in 2008 (10.4 Mha) than all types of gravity irrigation combined (8.9 Mha). As center pivots gained popularity, researchers began to consider runoff potential, mainly to efficiently apply irrigation water. Most sprinkler-irrigation studies were not concerned with soil erosion, probably because the effects of sprinkler-irrigation erosion tend to occur within the field rather than off site.

A 1969 study evaluated center-pivot runoff from a theoretical perspective and showed the importance of modifying infiltration parameters, determined from pond infiltration tests, for the low initial application rate that occurs with moving irrigation systems.[23] Their theoretical evaluation showed that 0%–40% of the applied water could run off with typical operating conditions. A 1971 field study documented 11%–41% runoff on four center pivots operated by farmers.[24] Runoff with center-pivot irrigation became a more important issue as low-pressure sprinklers began to be used to reduce energy costs. Early types of low-pressure sprinklers applied water to a smaller area, which increased application rates and potential runoff.[25] Low-pressure sprinklers (40 and 100 Pa) averaged 69 or 70 mm of runoff compared with 8 to 10 mm of runoff for high-pressure sprinklers (170 and 345 Pa) during a 4-year field study.[26] Reducing pressure from 380 to 140 Pa increased irrigation runoff 30% for a center pivot with impact sprinklers.[27] Peak application rate at the outer end of a center pivot would be about 30 mm h^{-1} for a high-pressure impact sprinkler with 20 m wetted radius and more than 100 mm h^{-1} for a low-pressure spray sprinkler with 5 m wetted radius.[25] Figure 5 shows two sprinkler application rate curves with time and an infiltration rate curve. The volume of water applied when application rate exceeds infiltration is potential runoff. All of this water may not run off if some is ponded or stored on the soil surface.

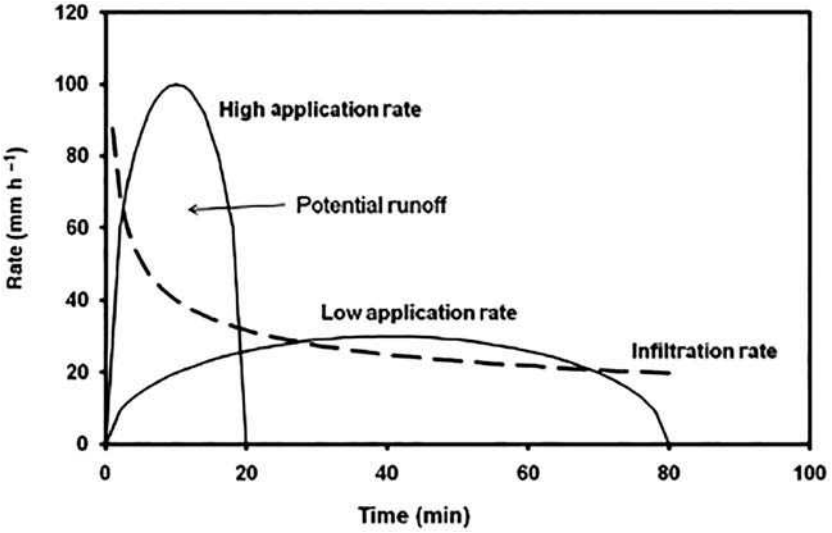


FIGURE 5 Example of soil infiltration rate and sprinkler application rates for high- and low-application-rate sprinklers. Runoff potentially occurs when sprinkler application rate exceeds the soil infiltration rate.

Many types of sprinklers are now available for center pivots. Some apply water in defined streams with a wetted diameter over 20 m with nozzle pressure of 200 Pa or less. Others distribute water evenly over the wetted area with various combinations of droplet sizes. Various sprinkler designs are the result of manufacturers trying to reduce the kinetic energy applied to the soil during irrigation, so all applied water can infiltrate. Kincaid[28] developed a model in 1996 to estimate kinetic energy per unit drop volume for common sprinkler types. Calculating area-weighted kinetic energies per unit drop volume for individual sprinklers showed that sprinklers with the smallest drop size distributions had the lowest kinetic energy. Sprinklers with smaller sized drops tend to have smaller wetted diameters because small drops cannot travel as far as large drops. Larger drops travel farther and therefore cover a greater portion of the circular wetted area for an individual sprinkler. A smaller wetted diameter also results in a higher application rate when sprinkler application patterns are overlapped like occurs on a center pivot. An alternative method for characterizing sprinkler kinetic energy is calculating the rate that energy is applied to the soil, or specific power, as a function of radial distance from the sprinkler.[29] The specific power distribution is energy per drop volume multiplied by the application rate and can be overlapped, like water application rate, to develop a composite specific power profile for a sprinkler system. A flat-plate sprinkler with small-sized drops had higher average composite specific power than two other sprinklers with larger drop sizes and larger wetted diameter.[29] Recent field research on small plots showed that soil erosion was significantly greater with the flat-plate spray sprinkler compared with the two other sprinklers with larger drop size distributions.[30] This directly contradicts previous conventional thinking that sprinklers with smaller drops caused less erosion.

The most effective way to control sprinkler-irrigation erosion is to eliminate runoff, which also increases water application efficiency. One way to control runoff is to increase water storage on the soil surface.[25] Reservoir tillage is a practice that forms small pits in the soil to store water. Each pit can hold 5 to 10 L of water.[31] This is especially important on sloping fields. Reservoir tillage reduced runoff 68% and soil erosion 92% during a 50 mm simulated irrigation on a field with 10% slope.[32] Runoff was not different when field slope was only 1%. Increasing surface residue also decreases sprinkler-irrigation runoff similar to rainfall runoff.[27] Disking corn stubble prior to planting, which left approximately 30% of the soil surface covered with crop residue, reduced runoff to 17% of applied irrigation compared with

25% runoff for moldboard plow plots in a 4 years study.[33] In addition to reducing runoff, disking also reduced soil loss about 50% compared with moldboard plowing.

Applying PAM with sprinkler irrigation can also improve infiltration, which reduces runoff and soil erosion. Field studies have shown that erosion decreased under moving sprinkler systems when 20 kg PAM ha^{-1} was applied to the soil before irrigation.[34,35] Lower PAM application rates can be effective when PAM is applied with irrigation water rather than sprayed directly on the soil surface. In laboratory studies with 2 m^2 soil boxes, applying 2 to 4 kg PAM ha^{-1} at 10 to 20 mg L^{-1} with sprinkler-irrigation water reduced soil erosion 75% compared with untreated soil, but these benefits decreased with subsequent irrigations without PAM.[36] In a similar laboratory study, applying 1 kg PAM ha^{-1} with three consecutive irrigations reduced cumulative runoff 50% compared with untreated soil, while applying 3 kg PAM ha^{-1} with one irrigation only reduced runoff by 35%.[37] Field tests in the United States showed that applying PAM with four irrigations (2 to 3 kg ha^{-1} total applied) significantly reduced soil erosion from 52 and 34 kg ha^{-1} for the control to 21 and 5 kg ha^{-1} for the PAM treatment during the 2 years of the study.[38] Soil erosion was not significantly different for a similar field study in Portugal with lower PAM application rates (0.3 kg ha^{-1}).[38]

Conclusions

Irrigation is vital to world food production, but soil erosion during irrigation threatens the long-term productivity of irrigation. Soil erosion is generally greater from surface irrigation because water flows over the soil during irrigation. Surface-irrigation management is often a tradeoff between irrigation uniformity and erosion. High flow rates can cause erosion; low flow rates can cause poor irrigation uniformity. Ideally, sprinkler irrigation should not have any runoff; however, moving irrigation systems, like center pivots, often apply water faster than it can infiltrate into the soil. Current research is attempting to quantify runoff and erosion potential for various types of center-pivot sprinklers so manufacturers can improve sprinkler designs. Irrigation can be managed to minimize erosion and maintain productivity.

References

1. Kendall, H.W.; Pimentel, D. Constraints on the expansion of the global food supply. Ambio **1994**, *23* (3), 198–205.
2. USDA National Agricultural Statistics Service. *Farm and Ranch Irrigation Survey,* available at http://www.agcensus.usda.gov (accessed December 2010).
3. Bjorneberg, D.L.; Kincaid, D.C.; Lentz, R.D.; Sojka, R.E.; Trout, T.J. Unique aspects of modeling irrigation-induced soil erosion. Int. J. Sediment Res. **1999**, *15* (2), 245–252.
4. Trout, T.J. Furrow irrigation erosion and sedimentation: Onfield distribution. Trans. ASAE **1996**, *39* (5), 1717–1723.
5. Bjorneberg, D.L.; Sojka, R.E.; Aase, J.K. Pre-wetting effect on furrow irrigation erosion: A field study. Trans. ASAE **2002**, *45* (3), 717–722.
6. Lentz, R.D.; Sojka, R.E.; Carter, D.L. Furrow irrigation water-quality effects on soil loss and infiltration. Soil Sci. Soc. Am. J. **1996**, *60* (1), 238–245.
7. Kim, K.H.; Miller, W.P. Effect of rainfall electrolyte concentration and slope on infiltration and erosion. Soil Technol. **1996**, *9* (3), 173–185.
8. Flanagan, D.C.; Norton, L.D.; Shainberg, I. Effect of water chemistry and soil amendments on a silt loam soil. Part II: Soil erosion. Trans. ASAE **1997**, *40* (6), 1555–1561.
9. Levy, G.J.; Levin, J.; Shainberg, I. Seal formation and interrill soil erosion. Soil Sci. Soc. Am. J. **1994**, *58* (1), 203–209.
10. International Commission on Irrigation and Drainage, available at http://www.icid.org/imp_data.pdf (accessed November 2010).

11. FAO Aquastat, http://www.fao.org/nr/water/aquastat/main/index.stm (accessed November 2010).
12. Koluvek, P.K.; Tanji, K.K.; Trout, T.J. Overview of soil erosion from irrigation. J. Irrig. Drain. Eng. **1993**, *119* (6), 929–946.
13. Israelson, O.W.; Clyde, G.D.; Lauritzen, C.W. *Soil erosion in small irrigation furrows.* Bull. 320. Utah Agricultural Experiment Station: Logan, UT, 1946.
14. Mech, S.J. Effect of slope and length of run on erosion under irrigation. Agric. Eng. **1949**, *30* (8), 379–383, 389.
15. Berg, R.D.; Carter, D.L. Furrow erosion and sediment losses on irrigated cropland. J. Soil Water Conserv. **1980**, *35* (6), 267–270.
16. Bjorneberg, D.L.; Westermann, D.T.; Nelson, N.O.; Kendrick, J.H. Conservation practice effectiveness in the irrigated Upper Snake River/Rock Creek watershed. J. Soil Water Conserv. **2008**, *63* (6), 487–495.
17. Ebbert, J.C.; Kim, M.H. Relation between irrigation method, sediment yields, and losses of pesticides and nitrogen. J. Environ. Qual. **1998**, *27* (2), 372–380.
18. Kemper, W.D.; Trout, T.J.; Brown, M.J.; Rosenau, R.C. Furrow erosion and water and soil management. Trans. ASAE **1985**, *28* (5), 1564–1572.
19. Mailapalli, D.R.; Raghuwanshi, N.S.; Singh, R. Sediment transport in furrow irrigation. Irrig. Sci. **2009**, *27* (6), 449–456.
20. Fernandez-Gomez, R.; Mateos, L; Giraldez, J.V. Furrow irrigation erosion and management. Irrig. Sci. **2004**, *23* (3), 123–131.
21. Lentz, R.D.; Sojka, R.E. Long-term polyacrylamide formulation effects on soil erosion, water infiltration, and yields of furrow-irrigated crops. Agron. J. **2009**, *101* (2), 305–314.
22. Lentz, R.D.; Sojka, R.E. Field results using polyacrylamide to manage furrow erosion and infiltration. Soil Sci. **1994**, *158* (4), 274–282.
23. Kincaid, D.C.; Heermann, D.F.; Kruse, E.G. Application rates and runoff in center-pivot sprinkler irrigation. Trans. ASAE **1969**, *12* (6), 790–794,797.
24. Aarstad, J.S.; Miller, D.E. Soil management to reduce runoff under center-pivot sprinkler systems. J. Soil Water Conserv. **1973**, *28* (4), 171–173.
25. Gilley, J.R. Suitability of reduced pressure center-pivots. J. Irrig. Drain. Eng. **1984**, *110* (1), 22–34.
26. DeBoer, D.W.; Beck, D.L.; Bender, A.R. A field evaluation of low, medium, and high pressure sprinklers. Trans. ASAE **1992**, *35* (4), 1185–1189.
27. Mickelson, R.H.; Schweizer, E.E. Till-plant systems for reducing runoff under low-pressure, center-pivot irrigation. J. Soil Water Conserv. **1987**, *42* (2), 107–111.
28. Kincaid, D.C. Spraydrop kinetic energy from irrigation sprinklers. Trans. ASAE **1996**, 39 (*3*), 847–853.
29. King, B.A.; Bjorneberg, D.L. Characterizing droplet kinetic energy applied by moving spray-plate center-pivot irrigation sprinklers. Trans. ASABE **2010**, 53 (*1*), 137–145.
30. King, B.A.; Bjorneberg, D.L. Evaluation of potential runoff and erosion of four center pivot irrigation sprinklers. Trans. ASABE **2011**, in press.
31. Oliveira, C.A.S.; Hanks, R.J.; Shani, U. Infiltration and runoff as affected by pitting, mulching and sprinkler irrigation. Irrig. Sci. **1987**, *8* (1), 49–64.
32. Kranz, W.L.; Eisenhauer, D.E. Sprinkler irrigation runoff and erosion control using interrow tillage techniques. Appl. Eng. Agric. **1990**, *6* (6), 739–744.
33. DeBoer, D.W.; Beck, D.L. Conservation tillage on a silt loam soil with reduced pressure sprinkler irrigation. Appl. Eng. Agric. **1991**, *7* (5), 557–562.
34. Levy, G.J.; Ben-Hur, M.; Agassi, M. The effect of polyacrylamide on runoff, erosion, and cotton yield from fields irrigated with moving sprinkler systems. Irrig. Sci. **1991**, *12* (2), 55–60.
35. Stern, R.; Van Der Merwe, A.J.; Laker, M.C.; Shainberg, I. Effect of soil surface treatments on runoff and wheat yields under irrigation. Agron. J. **1992**, *84* (1), 114–119.

36. Aase, J.K.; Bjorneberg, D.L.; Sojka, R.E. Sprinkler irrigation runoff and erosion control with polyacrylamide— Laboratory tests. Soil Sci. Soc. Am. J. **1998**, *62* (6), 1681–1687.
37. Bjorneberg, D.L.; Aase, J.K. Multiple polyacrylamide applications for controlling sprinkler irrigation runoff and erosion. Appl. Eng. Agric. **2000**, *16* (5), 501–504.
38. Bjorneberg, D.L.; Santos, F.L.; Castanheira, N.S.; Martins, O.C.; Reis, J.L.; Aase, J.K.; Sojka, R.E. Using polyacrylamide with sprinkler irrigation to improve infiltration. J. Soil Water Conserv. **2003**, *58* (5), 283–289.

16

Irrigation: River Flow Impact

Robert W. Hill and
Ivan A. Walter

Introduction

In hydrologic studies it is common engineering practice to quantify the impact upon the stream(s) from which the irrigation water is diverted. The impact upon the stream is actually of two kinds: 1) diversions that decrease the streamflow and 2) return flows that increase the streamflow. The engineering term used to describe the overall impact is "streamflow depletion" which means the net reduction in streamflow resulting from diversion to irrigation uses. Actual stream depletions are a function of many factors including the amount and timing of diversions, the type of diversion structure (well vs. ditch), crops grown, soil type, depth to groundwater, irrigation method, irrigation efficiency, properties of the alluvial aquifer, area irrigated, and evapotranspiration of precipitation, groundwater, and irrigation water.

Depletion

Depletion, in this context, is the consumptive abstraction of water from the hydrologic system as a result of irrigation. It is in addition to consumptive water use that would have occurred in the unmodified natural situation. As an example, waters of the Bear River Basin of Southern Idaho, Northern Utah, and Western Wyoming, because it is an interstate system, are administered by a federally established commission under the authority of the Bear River Compact.[1] Depletion is the basis, in the compact, for allocating Bear River water use among the three states. It is defined by a "Commission Approved Procedure" which includes consideration of land use and incorporates an equation for estimating depletion based on evapotranspiration. In a study for the commission, Hill[5] defined crop depletion as

$$\mathrm{Dpl} = \mathrm{Et} - \mathrm{Smco} - \mathrm{Pef} \tag{1}$$

where Dpl is estimated depletion for a given site or sub-basin; Et is calculated crop water use; SMco is moisture which is "carried over" from the previous non-growing season (October 1–April 30) as stored soil water in the root zone available for crop water use subsequent to May 1; and Pef is an estimate of that portion of precipitation measured at an NWS station during May–September, which could be used by crops.

The carry-over soil moisture (SMco) was estimated by assuming that 67% of adjusted precipitation from October through April could be stored in the root zone. If this exceeded 75% of the available soil water-holding capacity of the average root zone in the sub-basin, the excess was considered as lost to drainage or runoff and not available for crop use. Growing season precipitation was considered to be 80% effective in contributing to crop water use. The effectiveness factor of 80% allowed for precipitation depths throughout a sub-basin that might differ from NWS rain-gage amounts. It also included a reduction for mismatches in timing between rainfall events and irrigation scheduling.

Hydrograph Modification

Diversion of significant amounts of water from rivers and streams for irrigation and subsequent return flows alters the shape and timing of downstream hydrographs. In watersheds where mountain snowmelt provides the irrigation supply, such as in the Western United States, diversion during the spring runoff attenuates the peak flow rate while later return flows extend the flow duration into late summer and early fall.

Reservoir Storage

Storage of water in reservoirs can significantly modify the natural stream hydrograph depending on the timing and quantity of the storage right. Irrigators with junior rights may only be able to store during time periods with low irrigation demand, such as during the winter, or during peak flow periods. Reductions of stream flow during the winter time may have considerable impact on downstream instream flows. Whereas, storage during periods of peak runoff may not affect minimum in-stream flow needs, but could deposit considerable amounts of sediment in the reservoir.

Irrigation Return Flows

Irrigation return flows are comprised of surface runoff and/or subsurface drainage that becomes available for subsequent rediversion from either a surface stream or a groundwater aquifer downstream (hydrologically) of the initial use. Reusable return flow can be estimated as irrigation diversion minus crop related depletions minus additional abstractions. Additional abstractions include incidental consumptive use from water surfaces as in open drains, along with non-crop vegetation. The timing of return flow varies from nearly instantaneous (recaptured tailwater) to delays of weeks and months or perhaps longer with deep percolation subsurface drainage. In a hydrologic model study of the Bear River Basin[4] delay times between diversion and subsequent appearance of the return flow at the next downstream river gage varied from 1.5 months to as long as 6 months. The delay appeared to be related to sub-basin shape and size.

Irrigation Methods

Four general irrigation methods are used: surface, subsurface, sprinkler, and trickle (also known as low flow or drip). Surface methods include wild or controlled flooding, furrow, border-strip, and ponded water (basin, paddy, or low-head bubbler). Hand move, wheel move, and center pivot are examples of sprinkler irrigation. Trickle irrigation includes point source emitters, microspray, bubbler, and line-source drip tape (above or below ground). Whereas the efficiency of surface irrigation is dependent upon the skills and experience of the irrigator, the performance of trickle and sprinkler systems is more dependent on the design. Generally, the more control that the system design (hardware) has on the irrigation system performance, the higher the application efficiency (E_a) can be. Thus, typical wheel move sprinklers have higher E_a values than surface irrigation, but lower values than for center pivots or trickle, assuming better than average management practices for each method.

The impact on river flows can be quite different among the various irrigation methods. The nature of furrow and border surface irrigation generally produces tail water runoff, which can be immediately recaptured and reused, as well as deep percolation, which may not be available for reuse until after a period of time. Tailwater is essentially eliminated and deep percolation reduced with sprinklers (Figure 1) compared to conventional surface irrigation. Whereas, with drip methods, deep percolation can be further reduced. The reduction of deep percolation implies increased salt concentration in the root zone leachate, but, perhaps significant reduction in salt pick-up potential from geologic conditions.

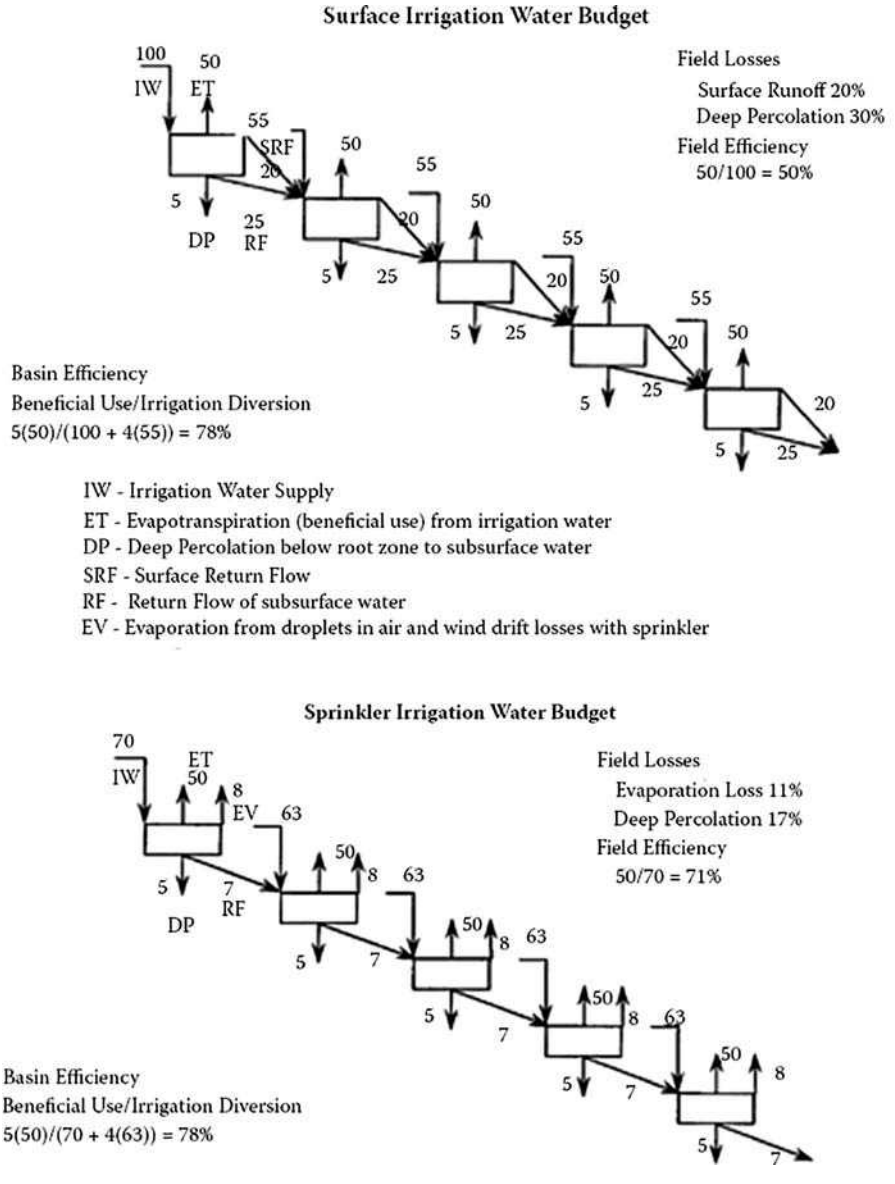


FIGURE 1 Comparison of basin efficiencies between surface and sprinkler irrigation methods with four return flow reuse cycles.

Irrigation Efficiencies

Although a full discussion of the several variations of irrigation efficiency is beyond the scope herein, two terms will be defined and discussed. More complete discussions relating to irrigation efficiencies and water requirements are given elsewhere.[6–9,13] Keller and Bliesner[9] give a particularly thorough presentation of distribution uniformity and efficiencies.

Application efficiency (E_a):

$$E_a = 100 \times \frac{\text{Volume of water stored in the root zone}\left(V_s\right)}{\text{Volume of water delivered in the farm or field}\left(V_f\right)}.$$

Distribution uniformity:

The distribution uniformity is a measure of how evenly the on-farm irrigation system distributes the water across the field. The definition of DU is:

$$\text{DU} = 100 \times \frac{\text{Average of the lowest 25\% of infiltrated water depth}}{\text{Average of all infiltrated water depths across the field}}.$$

On-farm or field application efficiencies can be affected by the distribution uniformity and vary widely for both surface and sprinkle irrigation methods. This is largely due to difference in management practices, appropriateness of design in matching the site conditions (slope, soils, and wind), and the degree of maintenance. In addition, for a given system uniformity, the higher the proportion of the field that is adequately irrigated (i.e., infiltrated water refills the soil water deficit) the lower will be the application efficiency. This is due to greater deep percolation losses in the overirrigated portions of the distribution pattern. Some values determined in recent Utah field evaluations are:

	Observed		
Method	High (%)	Low (%)	Typical (%)
	Surface irrigation		
E_a	72	24	50
Tailwater	55	5	20
Deep percolation	65	20	30
	Sprinkler irrigation		
E_a	84	52	70
Evaporation	45	8	12
Deep percolation	37	8	18

The E_a for a particular field may vary greatly during the season. Cultivation practices, microconsolidation of the soil surface and vegetation will alter surface irrigation efficiency both up and down from the seasonal average. Seasonal and diurnal variations in wind, humidity, and temperature will also affect sprinkle application efficiencies.

Basin Irrigation Efficiency

The actual irrigation efficiency realized for several successive downstream fields where capture and reuse of return flows is experienced is higher than the E_a of an individual field. This notion of "Basin Irrigation Efficiency"[12,13] is illustrated in Figure 1. This simple example comparison of surface and sprinkle methods assumes four reuse cycles. In each of the five "fields" Et is assumed to be 50 units. The surface runoff is captured for reuse on the next field. All of the irrigation-related evaporation is assumed "lost" as well as 5 units of deep percolation. After the fifth field, all surface and subsurface flows are lost. The basin efficiency for surface is 78%, which is the sameas for sprinkle. The surface irrigation basin efficiency increase is dependent upon the surface return flow reuse, which is 20 units in this example. However, the depletion is greater for sprinkler due to the extra evaporation. In a Colorado field study, Walter and Altenhofen[11] found a progressive increase in irrigation efficiencies from field (average E_a of 45%), to farm, to efficiency of ditch or sectors (average of 83%). This was due to the reuse of tailwater (10%–20% of delivery) and deep percolation (46% of delivery).

Environmental Concerns

The process of evapotranspiration, or crop water use, extracts pure water from the soil water reservoir, which leaves behind the dissolved solids (salts) contained in the applied irrigation water. The "evapo-concentration" of salts is an inevitable result of irrigation for crop production. As stated by Bishop and Peterson:[2]

…Other uses add something to the water, but irrigation basically takes some of the water away, concentrating the residual salts. Irrigation may also add substances by leaching natural salts or other materials from the soil or washing them from the surface. Irrigation return flow is a process by which the concentrated salts and other substances are conveyed from agricultural lands to the common stream or the underground water supply…

Water Quality Implications for Agriculture

Irrigated agriculture is dependent upon adequate, reasonably good quality water supplies. As the level of salt increases in an irrigation source, the quality of water for plant growth decreases. Since all irrigation waters contain a mixture of natural salt, irrigated soils will contain a similar mix to that in the applied water, but generally at a higher concentration. This necessitates applying extra irrigation water, or taking advantage of non-growing season precipitation, to leach the salts below the root zone.

Salt Loading Pick-Up

Water percolating below the root zone or leaking from canals and ditches may "pick-up" additional salts from mineral weathering or from salt-bearing geologic formations (such as the Mancos shale of Western Colorado and Eastern Utah). This salt pick-up will increase the salt load of return flows and consequently increase the salinity of receiving waters.

In the Colorado River Basin in the United States and Mexico salinity is a concern because of its adverse effects on agricultural, municipal, and industrial users.[10] The Salinity Control Act of 1974 (Public Law 93–320) created the Colorado River Basin Salinity Control Program to develop projects to reduce salt loading to the Colorado River. Salinity control projects include lining open canals and laterals (or replacing with pipe) and installing sprinklers in place of surface irrigation for the purpose of decreasing salt loading caused by canal leakage and irrigated crop deep percolation. Recently selenium in irrigation return flow has become a concern[3] and may also be reduced by salinity reduction projects.

In-Stream Flow Requirements

Diversions in some reaches in some Western United States streams are "dried up" immediately downstream of diversion structures during times of peak irrigation demand. This condition eliminates any use of the reach for fisheries and other uses which depend on in-stream flow. In some instances, negotiated agreements with senior water rights users have allowed for bypass of minimal amounts of water to sustain the fishery or habitat, and for control of tail-water runoff to reduce agricultural related chemicals in the receiving water.

References

1. Bear River Commission. *Amended Bear River Compact,* S.B. 255, 1979 Utah Legislature Session. 1979.
2. Bishop, A.A.; Peterson, H.B. (team leaders). Characteristics and Pollution Problems of Irrigation Return Flow. Final Report Project 14-12-408, Fed. Water Pollution Control Adm. U.S. Dept. of Interior (Ada, Oklahoma). Utah State University Foundation, Logan, Utah. May 1969.
3. Butler, D.L. *Effects of Piping Irrigation Laterals on Selenium and Salt Loads, Montrose Arroyo Basin, Western Colorado;* U.S. Geological Survey Water-Resources Investigations Report 01–4204 (in cooperation with the U.S. Bureau of Reclamation): Denver, Colorado, 2001.
4. Hill, R.W.; Israelsen, E.K.; Huber, A.L.; Riley, J.P. A *Hydrologic Model of the Bear River Basin*; PRWG72-1, Utah Water Resource Laboratory, Utah State University: Logan, Utah, 1970.
5. Hill, R.W.; Brockway, C.E.; Burman, R.D.; Allen, L.N.; Robison, C.W. *Duty of Water Under the Bear River Compact: Field Verification of Empirical Methods for Estimating Depletion. Final Report*; Utah Agriculture Experiment Station Research Report No. 125, Utah State University: Logan, Utah, January 1989.
6. Hoffman, G.J.; Howell, T.; Solomon, K. *Management of Farm Irrigation Systems*; The American Society of Agricultural Engineers: St. Joseph, MI, 1990.
7. Jensen, M.E. *Design and Operation of Farm Irrigation Systems;* ASAE Monograph, American Society of Agricultural Engineers: St. Joseph, MI, 1983.
8. Jensen, M.E.; Burman, R.D.; Allen, R.G.; Eds. *Evapotranspiration and Irrigation Water Requirements;* ASCE Manual No. 70, American Society of Civil Engineers: New York, 1990.
9. Keller, J.; Bliesner, R.D. *Sprinkle and Trickle Irrigation*; Van Nostrand Reinhold: New York, 1990.
10. U.S. Department of the Interior. *Quality of Water—Colorado River Basin: Bureau of Reclamation, Upper Colorado Region;* Progress Report no 19; Salt Lake City, Utah, 1999.
11. Walter, I.A.; Altenhofen, J. Irrigation efficiency studies—Northern Colorado. Proceedings USCID Water Management Seminar. Sacramento, CA Oct 5–7; USCID: Denver, CO, 1995.
12. Willardson, L.S.; Wagenet, R.J. *Basin-Wide Impacts of Irrigation Efficiency;* Proceedings of an ASCE Specialty Conference on Advances in Irrigation and Drainage, Jackson, WY, 1983.
13. Willardson, L.S.; Allen, R.G.; Frederiksen, H.D. *Elimination of Irrigation Efficiencies;* Proceedings 13th Technical Conference, USCID: Denver, CO, 1994.

17

Irrigation: Saline Water

B.A. Stewart

Introduction

As water becomes more limited, there is increasing use of saline waters for irrigation that were previously considered unsuitable. Rhoades, Kandiah, and Marshali[1] classified saline waters as shown in Table 1. Electrical conductivity is a convenient and practical method for classifying saline waters because there is a direct relationship between the salt content of the water and the conductance of an electrical current through water containing salts. Electrical conductivity values are expressed in siemens (S) at a standard temperature of 25°C.

Most waters used for irrigation have electrical conductivities less than 2 dS m^{-1}.[1] When water higher than this level is used, there can be serious negative effects on both plants and soils. As salinity in the root zone increases, the osmotic potential of the soil solution decreases and therefore reduces the availability of water to plants. At some point, the concentration of salts in the root zone can become so great that water will actually move from the plant cells to the root zone because of the osmotic effect. Salts containing ions such as boron, chloride, and sodium can also be toxic to plants when accumulated in large quantities in the leaves. The extent that plant growth is affected by saline water is dependent on the crop species. Some plants, such as barley and cotton, are much more resistant to salt than crops like beans. Rhoades, Kandiah, and Marshali[1] list the tolerance levels of a wide range of fiber, grain, and special crops; grasses and forage crops; vegetable and fruit crops; woody crops; and ornamental shrubs, trees, and ground cover. Soils are also negatively impacted by salt, particularly sodium salts. Sodium ions tend to disperse clay particles and this has deleterious effects on infiltration rate, structure, and other soil physical properties.

Irrigating with Saline Waters

Water limitations and the need to increase food and fiber production in many parts of the world have resulted in the use of water for irrigation containing increasing levels of salts. The United States, Israel, Tunisia, India, and Egypt have been particularly active in irrigating with saline waters.[1] Rhoades, Kandiah, and Marshali[1] published an extensive paper on the use of saline waters for crop production and it is a valuable guide for anyone interested in the subject. They reported that many drainage waters, including shallow ground waters underlying irrigated lands, fall in the range of 2 dS m^{-1} to 10 dS m^{-1}

TABLE 1 Classification of Saline Waters

Water Class	Electrical Conductivity (dSm–1)	Salt Concentration (mg L–1)	Type of Water
Non-saline	<0.7	<500	Drinking and irrigation
Slightly saline	0.7–2	500–1500	Irrigation
Moderately saline	2–10	1500–7000	Primary drainage and groundwater
Highly saline	10–25	7000–15,000	Secondary drainage and groundwater
Very highly saline	25–45	15,000–35,000	Very saline groundwater
Brine	>45	>45,000	Seawater

in electrical conductivity. Such waters are in ample supply in many developed irrigated lands and have good potential even though they are often discharged to better quality surface waters or to waste outlets. These waters can be successfully used in many cases with proper management. Reuse of second-generation drainage waters with electrical conductivity values of 10 dS m^{-1} to 25 dS m^{-1} is also sometimes possible but to a much lesser degree because the crops that can be grown with these waters are atypical and much less experience exists upon which to base management recommendations.

Miller and Gardiner[2] suggest that successful irrigation with saline water requires three principles. First, the soil should be maintained near field capacity to keep the salt concentration as low as possible. Second, application techniques should avoid any wetting of the foliage. Third, salts accumulating in the soil should be periodically leached. To accomplish these objectives, Miller and Gardiner[2] recommend the following general rules:

- Apply water at or below soil surface. Sprinklers should be used only if they avoid wilting the foliage (such as sprinkling before plant emergence or below-canopy to avoid salt-burn damage).
- Keep water additions almost continuous, but at or below field capacity so that most flow is unsaturated. This maintains adequate aeration.
- Enough water should be added to keep salts moving downward, thus avoiding salt buildup in the root zone.

Miller and Gardiner[2] stress that these rules are difficult to meet and are best satisfied by some form of drip irrigation. They also state that due to the need for high water levels and because of high sodium ratios that sandy soils are more adaptable to the use of saline waters than soils containing high percentages of silt and clay particles.

Rhoades, Kandiah, and Marshali[1] also list specific management practices for producing crops with salty waters. Their list includes the following guidelines:

- Selection of crops or crop varieties that will produce satisfactory yields under the existing or predicted conditions of salinity or sodicity.
- Special planting procedures that minimize or compensate for salt accumulation in the vicinity of the seed.
- Irrigation to maintain a relatively high level of soil moisture and to achieve periodic leaching of the soil.
- Use of land preparation to increase the uniformity of water distribution and infiltration, leaching and removal of salinity.
- Special treatments (such as tillage and additions of chemical amendments, organic matter and growing green manure crops) to maintain soil permeability and tilth. The crop grown, the quality of water used for irrigation, the rainfall pattern and climate, and the soil properties determine to a large degree the kind and extent of management practices needed.

Blending Low-Salt and Salty Waters

Miller and Gardiner[2] reported that countries such as Israel have developed extensive canal and reservoir systems where both low-salt and salty waters are mixed to obtain usable water. Rhoades, Kandiah, and Marshali,[1] however, state that blending or diluting excessively saline waters with good quality water supplies should only be undertaken after consideration is given to how this affects the volumes of consumable water in the combined and separate supplies. They suggest that blending or diluting drainage waters with good quality waters in order to increase water supplies or to meet discharge standards may be inappropriate under certain situations. More crop production can usually be achieved from the total water supply by keeping the water components separated. Serious consideration should be given for keeping saline drainage waters separate from the good quality water, especially when the good quality waters are used for irrigation of salt-sensitive crops. The saline waters can be used more effectively by substituting them for good quality water to irrigate certain crops grown in the rotation after seeding establishment.

Conclusion

There is ample evidence that saline waters once considered unacceptable for irrigation can be used successfully provided that they are properly managed. There is also ample evidence, however, to show that these waters can be highly damaging to the environment and to the soil resource base when improperly managed. Therefore, saline waters should be only used for irrigation after careful study and considering as many factors as possible. Then, when the waters are used for irrigation, a careful monitoring program should be implemented of both the crops produced and of the resulting soil and environmental changes.

References

1. Rhoades, J.D.; Kandiah, A.; Marshali, A.M. The use of saline water for crop production. *FAO Irrigation and Drainage Paper 48;* Food and Agriculture Organization of the United Nations: Rome, 1992.
2. Miller, R.W.; Gardiner, D.T. *Soils in Our Environment,* 8th Ed.; Prentice Hall: Upper Saddle River, NJ, 1998.

18

Irrigation: Sewage Effluent Use

B.A. Stewart

Introduction

One of the primary functions of soil is to buffer environmental change. This is the result of the biological, chemical, and physical processes that occur in soils. The soil matrix serves as an incubation chamber for decomposing organic wastes including pesticides, sewage, solid wastes, and many other wastes. Soils store, decompose, or immobilize nitrates, phosphorus, pesticides, and other substances that can become pollutants in air or water. Consequently, soil has, for centuries, been used for the application of sewage effluents. Sewage effluent provides farmers with a nutrient-enriched water supply and society with a reliable and inexpensive means of wastewater treatment and disposal. It should not, however, be assumed that irrigation is always the best solution for wastewater disposal. Disposal by irrigation should always be compared with alternative options based on environmental, social, and economic costs and benefits.

While disposal is the primary objective in many cases, the need of water for irrigation is becoming more often the driver for using sewage effluent on land. This is particularly true in areas like the Middle East where population growth is resulting in severe water shortages. The guidelines for using effluent for irrigation vary considerably among countries and other governing bodies. Cameron[1] conducted a literature review and found wide differences of guidelines for effluent irrigation projects being used throughout the world. In general, however, sustainable and environmentally sound systems can be developed in most situations provided proper management practices are followed.

Concerns of Irrigating with Sewage Effluent

In spite of the documented benefits associated with the use of sewage effluent for irrigation, there are numerous concerns. Many industrial wastewaters have been routinely dumped into municipal sewage lines. While this issue has been addressed in some jurisdictions, it has not in many others. In the United States, the Environmental Protection Agency requires that wastewaters be treated prior to disposal into municipal treatment plants or back into groundwater. Irrigating with wastewaters partially cleans water by percolation through the soil, but soluble salts and some inorganic and organic chemicals may continue to flow with the water to groundwater or surface supplies. In general, the Environmental

Protection Agency allows sewage effluents to be used for irrigation only if it does not cause: 1) extensive groundwater pollution; 2) a direct public health hazard; 3) an accumulation in the soil or water of hazardous substances that can get into the food chain; 4) an accumulation of pollutants such as odors into the atmosphere; and 5) other aesthetic losses, within the limits.[2]

Bouwer[3] has also expressed concerns about the use of sewage effluent for irrigation. He is particularly concerned with pathogens and warns that complete removal of viruses, bacteria, and protozoa and other parasites should be required before the effluent can be used to irrigate fruits/vegetables consumed raw or brought into the kitchen, or parks, playgrounds and other areas with free public access. Bouwer also stresses that long-term effects of sewage effluent irrigation on underlying groundwater should be considered in addition to the changes in nitrate and salinity. Ground water in low rainfall regions can be highly affected by percolating sewage effluent because much of the water is used by the growing crops and this greatly concentrates the chemicals in the small amounts of water that actually percolate to the groundwater. These chemicals can include disinfection byproducts, pharmaceutically active chemicals, and compounds derived from humic and fulvic acids formed by the decomposition of plant material. Bouwer claims that many of these chemicals are suspected carcinogens or toxic. Therefore, Bouwer concludes that while sewage irrigation looks good on the surface, a more extensive look reveals a potential for serious contamination of groundwater. He states that municipalities and other entities responsible for irrigation with sewage effluent should do a groundwater impact analysis to develop management protocols and be prepared for liability actions. Those who benefit are local and state institutions in water resources, environmental quality protection, public health, consultants, and operators of effluent irrigation projects.

Reuse Standards

The standards for using sewage effluent for irrigation of agricultural crops vary widely among different countries of the world. Mexico and many South American countries, e.g., use untreated wastewater for irrigation.[4] Most of these countries do not have the resources or capital to treat sewage effluents. Wastewater is utilized after little or no treatment, and health risks are minimized by crop selection. Mexico does not allow wastewater to be used to irrigate lettuce, cabbage, beets, coriander, radishes, carrots, spinach, and parsley. Acceptable crops include alfalfa, cereals, beans, chili, and green tomatoes. In contrast, Israel has very stringent water reuse requirements. Effluent water requires a high level of treatment (large soil-aquifer recharge systems with dewatering) before the water can be reused for irrigation of vegetables to be consumed raw.[5] Health guidelines for irrigation with treated wastewater developed in California indicate that effluent waters used on food crops must be disinfected, oxidized, coagulated, clarified, and filtered.[6] Total coliform counts cannot exceed a median value of 2.2/100 mL or a single sample value of 25/100 mL. Total coliforms must be monitored daily and turbidity cannot exceed 2 nephelometric turbidity units and must be monitored continuously. Less restrictive guidelines developed by Shuval et al.,[7] and adopted by most of the international agencies, suggested that effluent water reuse was relatively safe to use if it contained less than 1 helminth egg L^{-1}, and less than 1000 fecal coliforms/100 mL.

Monitoring Guidelines

Site selection is a critical and necessary step in initiating a sewage effluent irrigation system. The U.S. Environmental Protection Agency[8] published detailed information on site characterization and evaluation. Information was provided on the design of systems, site characteristics, expected quality of the effluent water after land treatment, and typical permeabilities and textural classes suitable for each land treatment process. Information was provided for designing and monitoring site characteristics for slow rate processes (sprinkler and other typical farm irrigation systems), rapid infiltration basins, and overland flow systems. Monitoring requirements will vary considerably among projects depending on the cropping patterns, soil characteristics, and specific environmental concerns. In most cases, monitoring procedures and criteria will be site specific. In all cases, however, the objectives should be to use the resources effectively, protect the land, protect the groundwater, protect the surface water, and protect the community amenity.

References

1. Cameron, D.R. *Sustainable Effluent Irrigation Phase 1: Literature Review International Perspective and Standards,* Technical Report Prepared for Irrigation Sustainability committee; Canada–Saskatchewan Agriculture Green Plan, 1996.
2. Miller, R.W.; Gardiner, D.T. *Soils in Our Environment,* 8th Ed.; Prentice-Hall Inc: Upper Saddle River, NJ, 1998.
3. Bouwer, H. *Groundwater Problems Caused by Irrigation with Sewage Effluent;* Irrigation and Water Quality Laboratory, USDA-ARS: Phoenix, AZ, 2000.
4. Strauss, M.; Blumenthal, U.J. *Human Waste Use in Agriculture and Aquaculture: Utilization Practices and Health Perspectives*; IRWCD Report No. 09/90; International Reference Centre for Waste Disposal: Deubendorf, 1990.
5. Shelef, G. The role of wastewater reuse in water resources management in Israel. Water Sci. Tech. **1990**, *23*, 2081–2089, Switzerland.
6. Ongerth, H.J.; Jopling, W.F. Water reuse in California. In *Water Renovation and Reuse;* Shuval, H.I., Ed.; Academic Press: New York, 1977.
7. Shuval, H.I.; Adin, A.; Fattal, B.; Rawitz, E.; Yekutiel, P. *Wastewater Irrigation in Developing Countries. Health Effects and Technical Solutions;* World Bank Tech. Pap.; 1986; Vol. 51, 325 pp.
8. U.S. EPA. *Process Design Manual: Land Treatment of Municipal Wastewater;* EPA 625/1–81-013; U.S. EPA Center for Environmental Research Information: Cincinnati, OH, 1981.

19

Irrigation: Soil Salinity

James D. Rhoades

Introduction

Irrigation is an ancient practice that predates recorded history. While irrigated farmland comprises only about 15% of the worlds' total farmland, it contributes about 36% of the total supply of food and fiber, and it stabilizes production against the vagaries of weather.[1] In 30 years time, irrigated agriculture is expected to have to supply 50% of the worlds' food production requirements.[1] However, over the last 20 years, irrigation growth has actually slowed to a rate that is now inadequate to keep up with the projected expanding food requirements.[1] Furthermore, irrigation has resulted in considerable salination of associated land and water. It has been estimated variably that the salinized area is as low as 20 and as high as 50% of the worlds' irrigated land.[2–4] Worldwide, about 76.6 Mha of land have become degraded by human-induced salination over the last 45–50 years.[3] It has been estimated that the world is losing at least three hectares of arable land every minute to soil salination (about 1.6 Mha per year), second only to erosion as the leading worldwide cause of soil degradation.[5–7] These data imply that the rate of salinization in developed irrigation projects now exceeds the rate of irrigation expansion.[8]

Surviving the salinity threat requires that the seriousness of the problem be recognized more widely, the processes contributing to salination of irrigated lands be understood, effective control measures be developed and implemented that will sustain the viability of irrigated agriculture, and that practical reclamation measures be implemented to rejuvenate the presently degraded lands.[9,10]

Deleterious Effects of Salts on Plants, Soils, and Waters

Salt-affected soils have reduced value for agriculture because of their content and proportions of salts, consisting mainly of sodium, magnesium, calcium, chloride, and sulfate and secondarily of potassium, bicarbonate, carbonate, nitrate, and boron. Saline soils contain excessive amounts of soluble salts for the practical and normal production of most agricultural crops. Sodic soils are those that contain excessive amounts of adsorbed sodium in proportion to calcium and magnesium, given the salinity level of the soil water. An example of a salt-affected irrigated soil is shown in Figure 1.

Soluble salts exert both general and specific effects on plants, both of which reduce crop yield.[11] Excess salinity in the seedbed hinders seedling establishment and in the crop root zone causes a general reduction in growth rate. In addition, certain salt constituents are specifically toxic to some plants. For example, boron is highly toxic to susceptible crops when present in the soil water at concentrations of

FIGURE 1 Photograph of salt-affected irrigated field.

only a few parts per million. In some woody crops sodium and chloride may accumulate in the tissue over time to toxic levels. These toxicity problems are, however, much less prevalent than is the general salinity problem.

Salts may also change soil properties that affect the suitability of the soil as a medium for plant growth.[12] The suitability of soils for cropping depends appreciably on the readiness with which they conduct water and air (permeability) and on their aggregate properties (structure), which control the friability (ease with which crumbled) of the seedbed (tilth). In contrast to saline soils, which are well aggregated and whose tillage properties and permeability to water and air are equal to or higher than those of similar nonsaline soils, sodic soils have reduced permeabilities and poor tilth. These problems are caused by the swelling and dispersion of clay minerals and by the breakdown of soil structure (slaking and crusting), which results in loss of permeability and tilth. Sodic soils are generally less extensive but more difficult to reclaim than saline soils.

Beneficial use of water in irrigation consists of transpiration and leaching for salinity control (the leaching requirement). Plant growth is directly proportional to water consumption through transpiration.[13] From the point of view of irrigated agriculture, the ultimate objective of irrigation is to increase the amount of water available to support transpiration. Salts reduce the fraction of water in a supply (or in the soil profile) that can be consumed beneficially in plant transpiration.[14] In considering the use of a saline water for irrigation and in selecting appropriate policies and practices of irrigation and drainage management, it is important to recognize that the total volume of a saline water supply cannot be consumed beneficially in crop production (i.e., transpired by the plant). A plant will not grow properly when the salt concentration in the soil water exceeds some limit specific to it under the given conditions of climate and management.[11] This is even true for halophytes.[15] Thus, the practice of blending or diluting excessively saline waters with good quality water supplies should be undertaken only after consideration is given to how it affects the volumes of consumable (usable) water in the combined and separated supplies.[14].

Causes of Salination Induced by Irrigation and Drainage

While salt-affected soils occur extensively under natural conditions, the salt problems of greatest importance to agriculture arise when previously productive soils become salinized as a result of agricultural activities (the so-called secondary salination). The extent and salt balance of salt- affected areas has been modified considerably by the redistribution of water (hence salt) through irrigation and drainage.

The development of large-scale irrigation and drainage projects, which involves diversion of rivers, construction of large reservoirs, and irrigation of large landscapes, causes large changes in the natural water and salt balances of entire geohydrologic systems. The impact of such developments can extend well beyond that of the immediate irrigated area. Excessive water diversions and applications are major causes of soil and water salination in irrigated lands. It is not unusual to find that less than 60% of the water diverted for irrigation is used in crop transpiration.[9] This implies that about 40% of the irrigation water eventually ends up as deep percolation. This drainage water contains more salt than that added with the irrigation water because of salt dissolution and mineral weathering[14] within the root zone. It often gains additional salt-load as it dissolves salts of geologic origin from the underlying substrata through which it flows in its down-gradient path. This drainage water often flows laterally to lower lying areas, eventually resulting in shallow saline groundwaters of large areas of land (waterlogging). Salination occurs in soils underlain by saline shallow groundwater through the process of "capillary rise" as groundwater (hence, salt) is driven upwards by the force of evaporation of water from the soil surface. Correspondingly, saline soils and waterlogging are closely associated problems.

Seepage from unlined or inadequately lined delivery canals occurs in many irrigation projects and is often substantial. Law, Skogerboe, and Denit[16] estimated that 20% of the total water diverted for irrigation in the United States is lost by seepage from conveyance and irrigation canals. Biswas[17] estimated that 57% of the total water diverted for irrigation in the world is lost from conveyance and distribution canals. Analogous to on-farm deep percolation resulting from irrigation, these seepage waters typically percolate through the underlying strata (often dissolving additional salts in the process), flow to lower elevation lands or waters, and add to the problems of waterlogging and salt-loading associated with on-farm irrigation there. A classic example of the rise in the water table following the development of irrigation has been documented in Pakistan and is described by Jensen, Rangeley, and Diele-man[9] and Ghassemi, Jakeman, and Nix.[2] The depth to the water table in the irrigated landscape located between three major river-tributaries rose from 20 to 30 m over a period of 80–100 years, i.e., from preirrigated time (about 1860) to the early 1960s, until it was nearly at the soil-surface. In one region, the water table rose nearly linearly from 1929 to 1950, demonstrating that deep percolation and seepage resulting from irrigation were the primary causes. Ahmad[18] concluded that about 50% of the water diverted into irrigation canals in Pakistan eventually goes to the groundwater by seepage and deep percolation.

The role of irrigated agriculture in salinizing soil systems has been well recognized for hundreds of years. It is of relatively more recent recognition that salination of water resources from agricultural activities is a major and widespread phenomenon of likely equal concern to that of soil salination. The causes of water salination are essentially the same as those of soils, only the final reservoir of the discharged salt-load is a water supply in the former case.[14] The volume of the water supply is reduced through irrigation diversions and irrigation; thus, its capacity to assimilate such received salts before reaching use-limiting levels is reduced proportionately. Only in the past 15 years has it become apparent that trace toxic constituents, such as selenium, in agricultural drainage waters can also cause serious pollution problems.[19]

Irrigation and Drainage Management to Control Soil Salinity

The key to overall salinity control is strict control that maintains a net downward movement of soil water in the root zone of irrigated fields over time while minimizing excess irrigation diversions, applications, and deep percolation.[20] The direct effect of salinity on plant growth is minimized by maintaining the soil-water content in the root zone within a narrow range at a relatively high level, while at the same time avoiding surface-ponding and oxygen deletion and minimizing deep percolation. Combined methods of pressurized, high-frequency irrigation and irrigation scheduling have been developed that permit substantially the desired control to be achieved.[21,22] These systems transfer control of water distribution and infiltration from the soil to the irrigation equipment. This results in less excess water (and hence, less salt) being applied overall to the field to meet the needs of a part of the field area having lowest intake rate, as done in the more traditional gravity irrigated systems. However, gravity irrigation systems can be designed to achieve good irrigation efficiency and salinity control, even though surface ponding still does occur. The so-called level-basin, multi-set, cablegation, surge, and tailwater-return systems are among them.[21,22] The need for irrigation and the amount required to meet evapotranspiration and leaching requirement is determined from plant stress measurements, calculations of evapotranspiration amounts, measurements of soil-water depletion, measurements of soil (or soil-water) salinity, or a combination of them.[21,22]

In addition to effective methods of irrigation scheduling and application, appropriate irrigation and salinity management also require an effective delivery system. Delivery systems have generally been designed to provide water on a regular schedule. Efficient irrigation systems require more flexible deliveries that can provide water on demand as each crop and particular field have need of it. Delivery systems can be improved by lining the canals, by containing the water within closed conduits, and by implementing techniques that increase the flexibility of delivery.

As briefly discussed earlier, irrigated agriculture is a major contributor to the salinity of many rivers and groundwaters, as well as soils. Reducing deep percolation generally lessens the salt load that is returned to rivers or groundwater and their pollution.[14] Additionally, saline drainage waters should be intercepted before being allowed to mix with water of better quality. The intercepted saline drainage water should be desalted and reused, disposed of by pond evaporation or by injection into some suitably isolated deep aquifer, or better yet it should be used for irrigation in a situation where brackish water is appropriate. Various irrigation and drainage strategies have been developed for minimizing the pollution of waters from irrigation and for using brackish waters for irrigation.[14,23] Desalination of agricultural drainage waters is not now economically feasible, but improved techniques for doing this exist and some are being implemented. However, more needs to be done in this regard.

Traditionally, the concepts of leaching requirement and salt-balance index have been used to plan and judge the appropriateness of irrigation and drainage systems, operations and practices with respect to salinity control, water use efficiency, and irrigation sustainability. However, these approaches are inadequate. The recommended method is to monitor directly the root-zone salinity levels and distributions across fields as a means to evaluate the effectiveness of salinity, irrigation, and drainage management practices, to detect problems (current and developing), to help determine the underlying causes of problems, and to determine source areas of major water and salt-load contributions to the underlying groundwater. Theory, equipment, and practical technology have been developed for these purposes.[24] More information about irrigation and drainage management to control soil and water salinity is found elsewhere.[25–27]

References

1. FAO. *World Agriculture Toward 2000:* An FAO Study; Alexandratos, N., Ed.; Bellhaven Press: London, 1988; 338 pp.
2. Ghassemi, F.; Jakeman, A.J.; Nix, H.A. *Salination of Land and Water Resources. Human Causes, Extent, Management and Case Studies;* CAB International: Wallingford, U.K., 1995; 526 pp.
3. Oldeman, L.R.; van Engelen, V.N.P.; Pulles, J.H.M. The Extent of human-induced soil degradation. In *World Map of the Status of Human-Induced Soil Degradation: An Explanatory Note;* Oldeman, L.R., Hakkeling, R.T.A., Sombroek, W.G., Eds.; International Soil Reference and Information Center (ISRIC): Wageningen, the Netherlands, 1991; 27–33.
4. Adams, W.M.; Hughes, F.M.R. Irrigation development in desert environments. In *Techniques for Desert Reclamation;* Goudie, A.S., Ed.; Wiley: New York, 1990; 135–160.
5. Buringh, P. Food production potential of the world. In *The World Food Problem: Consensus and Conflict;* Sinha, R., Ed.; Pergamon Press: Oxford, 1977; 477–485.
6. Dregne, H.; Kassas, M.; Razanov, B. A new assessment of the world status of desertification. Desertification Control Bull. **1991**, *20,* 6–18.

7. Umali, D.L. Irrigation-induced salinity. In *A Growing Problem for Development and Environment*; Technical Paper; World Bank: Washington, DC, 1993.
8. Seckler, D. *The New Era of Water Resources Management: From "Dry" to "Wet" Water Savings;* Consultative Group on International Agricultural Research: Washington, DC, 1996.
9. Jensen, M.E.; Rangeley, W.R.; Dieleman, P.J. Irrigation trends in world agriculture. In *Irrigation of Agricultural Crops;* American Society of Agronomy Monograph No. 30; ASA: Madison, WI, 1990; 31–67.
10. UNEP. *Saving Our Planet: Challenges and Hopes;* Nairobi, United Nations Environment Program: Nairobi, Kenya, 1992; 20 pp.
11. Maas, E.V. Crop salt tolerance. ASCE Manuals and Reports on Engineering No. 71. In *Agricultural Salinity Assessment and Management Manual*; Tanji, K.K., Ed.; ASCE: New York, 1990; 262–304.
12. Rhoades, J.D. Principal effects of salts on soils and plants. In *Water, Soil and Crop Management Relating to the Use of Saline Water*; Kandiah, A., Ed.; FAO (AGL) Mise. Series Publication 16/90; Food and Agriculture Organization of the United Nations: Rome, 1990; 1933.
13. Sinclair, T.R. Limits to crop yield? In *Physiology and Determination of Crop Yield*; Boone, K.J., Ed.; American Society of Agronomy: Madison, WI, 1994; 509–532.
14. Rhoades, J.D.; Kandiah, A.; Mashali, A.M. *The Use of Saline Waters for Crop Production;* FAO Irrigation and Drainage Paper 48; FAO: Rome, Italy, 1992; 133 pp.
15. Miyamoto, S.; Glenn, E.P.; Oslen, M.W. Growth, water use and salt uptake of four halophytes irrigated with highly saline water. J. Arid Environ. **1996**, *32,* 141–159.
16. Law, J.P.; Skogerboe, G.V.; Denit, J.D. The need for implementing irrigation return flow control. p. 1–17. *In Managing Irrigated Agriculture to Improve Water Quality;* Proc. Math. Conf. Manag. Irrig. Agric. Improve Water Avail., Denver, CO, May 1972; Graphics Manage. Corp.: Washington, DC, 16–18.
17. Biswas, A.K. Conservation and management of water resources. In *Techniques for Desert Reclamation*; Goudie, A.S., Ed.; Wiley: New York, 1990; 251–265.
18. Ahmad, N. Planning for Future Water Resources of Pakistan. Proceedings of Darves Bornoz Spec. Conference, National Committee of Pakistan; ICID: New Delhi, India, 1986; 279–294.
19. Letey, J.; Roberts, C.; Penberth, M.; Vasek, C. *An Agricul-turl Dilemma: Drainage Water and Toxics Disposal in the San Joaquin Valley;* Special Publication 3319; University of California: Oakland, 1986.
20. Rhoades, J.D. Soil salinity—causes and controls. In *Techniques for Desert Reclamation;* Goude, A.S., Ed.; Wiley: New York, 1990; 109–134.
21. Hoffman, G.J.; Rhoades, J.D.; Letey, J.; Sheng, F. Salinity management. In *Management of Farm Irrigation Systems*; Hoffman, G.J., Howell, T.A., Solomon, K.H., Eds.; ASCE: St. Joseph, MI, 1990; 667–715.
22. Kruse, E.G.; Willardson, L.; Ayars, J. On-farm irrigation and drainage practices. In *Agricultural Salinity Assessment and Management Manual;* Tanji, K.K., Ed.; ASCE Manuals and Reports on Engineering No. 71; ASCE: New York, 1990; 349–371.
23. Rhoades, J.D. Use of saline drainage water for irrigation. In *Agricultural Drainage*; ASA Drainage Monograph 38; Skaggs, R.W., van Schilfgaarde, J., Eds.; ASA Drainage Monograph 38; American Society of Agronomy: Madison, WI, 1999; 619–657.
24. Rhoades, J.D.; Chanduvi, F.; Lesch, S. *Soil Salinity Assessment: Methods and Interpretation of Electrical Conductivity;* FAO Irrigation and Drainage Paper 57; FAO, United Nations: Rome, Italy, 1999; 152 pp.
25. Rhoades, J.D. Use of saline and brackish waters for irrigation: implications and role in increasing food production, conserving water, sustaining irrigation and controlling soil and water degradation. Proceedings of the International Workshop on "The Use of Saline and Brackish Waters for Irrigation: Implications for the Management of Irrigation, Drainage and Crops" at the

10th Afro-Asian Conference of the International Committee on Irrigation and Drainage, Bali, Indonesia, July 23–24; Ragab, R., Pearce, G., Eds.; International Committee on Irrigation and Drainage: Bali, Indonesia, 1998; 261–304.

26. Rhoades, J.D.; Loveday, J. Salinity in irrigated agriculture. In *Irrigation of Agricultural Crops;* Stewart, B.A., Nielsen, D.R., Eds.; Agron. Monograph. No. 30; American Society of Agronomy: Madison, Wisconsin, 1990; 1089–1142.
27. Tanji, K.K. Nature and extent of agricultural salinity. In *Agricultural Salinity Assessment and Management*; Tanji, K.K., Ed.; ASCE Manuals and Reports on Engineering No. 71, ASCE: New York, 1990; 1–17.

20

Managing Water Resources and Hydrological Systems

Honglin Zhong,
Zhuoran Liang,
and Cheng Li

Introduction

Water scarcity has become a global issue as freshwater demand keeps soaring from the rapid increase in population and food demand under fast social-economic development, with additional pressure from altered water supply under climate change driven by anthropogenic emissions (Vörösmarty et al. 2000; Piao et al. 2010). Many regions in the world are already under severe water shortages and may suffer even worse water crisis under projected climate change (Konikow and Kendy 2005). Hotspots, such as the intensive cropped and heavy irrigated fertile major cropping zone of North China Plain (NCP), experience severe environmental problems from underground water over-exploitation for agricultural irrigation. This area is likely to experience more frequent deadly heatwave under projected future climate (Kang and Eltahir 2018). This will threaten both crop production and food security, risk farmer livelihoods, and constrain social-economic development in the future.

As the biggest freshwater consumption sector, agricultural production has achieved great success in food security with vast irrigation expansion since the “green revolution”. Only 20% of the cropland with irrigation produced as much as 40% of the total agricultural production in the world (Molden 2007). However, to achieve these results, irrigated crops consume 70% of global total freshwater withdrawal (Rosegrant et al. 2009) and cause rapid groundwater depletion in many major cropping zones such as North China, Northwestern India, and US high plains (Konikow and Kendy 2005). In NCP, the fast groundwater table decline has already caused severe environmental issues of dried-up rivers and lakes, seawater intrusion, land subsidence, and ground fissures (Xue et al. 2000). Public health may also be threatened when groundwater table drops to a deep level containing toxic levels of fluoride and arsenic (Currell et al. 2012).

Groundwater irrigation-intensive crops should be abandoned, and cropland fallow should be adopted for water-sustainable development (Wang et al. 2015, 2016). Shifting double cropping system (wheat–maize rotation) to single cropping system (winter fallow–spring maize rotation) and wheat fallow subsidy are highly recommended and already adopted by local farmers and government for groundwater recovery in the NCP (Feng et al. 2007; Zhong et al. 2017; Ren et al. 2018).

In this chapter, we reviewed the models, measurements, strategies, and projects for agricultural water consumption, conservation, and adaptation at the field, regional, and global levels. With a specific focus on the NCP, where experienced a significant transformation of agricultural practice and policies from previous water-intensive high-yield production-oriented to recent sustainable agricultural water management and cropland fallow with ecological compensation. With all the existing studies covering physical water saving, freshwater transfer, "virtual water" flow, and water footprint via food trade in the semi-arid NCP, farmers, scientists, stakeholders, and government from other water-deficit regions with heavy food production will get deeper insights into the water crisis and inspire local adaptations for the sustainable social-economic development around the world.

Models for Agricultural Water Consumption Estimation

Hydrological Models

Hydrological models are effective tools to simulate local water resources, hydrological processes, and crop growth at sub-basin/watershed level. Simulating cropland water balance is possible under different crop growth and water management with evapotranspiration, surface runoff, infiltration, percolation, shallow, and deep aquifer flow (Arnold et al. 1998). Major hydrological models include the Soil and Water Assessment Tool (SWAT) (Arnold et al. 1998), Variable Infiltration Capacity model (VIC) (Wood et al. 1992), and Systeme Hydrologique European (SHE) model (Abbott et al. 1986). These models provide a better understanding of the climate variability on hydrological processes and water resources for agricultural production (Sun and Li 2013; Wang et al. 2016). Advantages of the hydrological models are applied in the watershed level to optimize crop irrigation schedules for different crops (Sun and Li 2014) and identify the groundwater depletion from cropland irrigation expansion (Zhang et al. 2016). But the relatively simplified crop growth processes compared with dynamic crop models make the crop module within the macro-scale hydrological models focus more on the change of hydrological resources under given local water management conditions, rather than design the water-saving irrigation technologies or schedules (Xiong et al. 2019). In fact, the optimal crop irrigation schedule for water saving using hydrological model is obtained in tandem with field-based experiments (Yu et al. 2006).

Dynamic Crop Models

Process-based dynamic crop models have great advantages in simulating the interaction between crop growth and the sounding environment at a daily time step. Those models are widely used to determine optimal crop water management. Major crop models include Decision Support System for Agrotechnology Transfer (DSSAT) (Jones et al. 2003), Environmental Policy Integrated Climate (EPIC) (Williams et al. 1989), Agricultural Production Systems sIMulator (APSIM) (McCown et al. 1996), and AquaCrop (Steduto et al. 2009). The DSSAT model was employed to identify the long-term relationship between crop irrigation water consumption and groundwater level depletion in the Piedmont of NCP (Yang et al. 2006). For the regional agricultural water consumption simulation, a spatial parameterization by integrating the agroecological zones with site field information of all the sites was developed for the DSSAT model upscaling (Tian et al. 2012) and applied to simulate crop water consumption under different cropping systems in the NCP (Zhong et al. 2017, 2019; Tian et al. 2018).

Several limitations of the dynamic crop models need to be mentioned: (1) Crop models require heavy inputs for model calibration, validation, and simulation, but detailed information that will affect the

TABLE 1 Models for Agricultural Water Consumption Estimation

Model Types	Model Name	Description	Model References
Hydrological Models	SWAT	The SWAT model is developed to assess the impact of management on water supplies and nonpoint source pollution, and predict the impact of management on water, sediment, and agricultural chemical yields in watersheds.	Arnold et al. (1998)
	SHE/MIKE SHE	The SHE/MIKE SHE is a differential model to simulate the surface water movement from the hydrological processes of mass, energy, and momentum in a river basin.	Abbott et al. (1986)
	VIC	The VIC model simulates land-surface hydrology from infiltration variation, with a simplified approach of estimating infiltration, evaporation, and a base flow.	Wood et al. (1992)
Dynamic crop models	DSSAT	The DSSAT model is a modular structure with separate primary scientific components (soil, crop, weather, and management) and various process-based models that can be added or replaced.	Jones et al. (2003)
	EPIC	The EPIC model is developed to estimate the soil productivity as affected by soil erosion, plant growth, and related processes.	Williams et al. (1989)
	APSIM	The APSIM model focuses more on the "system" aspect of cropping, such as crop growth, rotation, fallow, and dynamic crop management to the soil processes and climate conditions.	McCown et al. (1996)
	AquaCrop	The AquaCrop model is a water-driven process model that focuses on the differential sensitivity to water stress of key plant processes and the biomass water productivity during the crop growth cycle.	Steduto et al. (2009)

total water consumption is lacking at the regional scale, especially the field water management. As most upscaling methods depend on observations from experiment sites, agricultural water consumption will be overestimated as field management levels in those sites are much higher than the farmer's practice (Tian et al. 2012). (2) Because process-based crop models mainly focus on soil water balance in the root zone during the crop growing period, hydrological processes, such as additional water sources from the surface stream, other water bodies, and underground water flow are not included. For example, the mountain-front discharge in the Piedmont of the Hebei Plain is an important groundwater recharge source but only considered in site experiment with field observations (Oort et al. 2016). Groundwater recharge from precipitation and irrigation water deep percolation are usually simplified or ignored (Yang et al. 2015) (Table 1).

Hybrid Models

To overcome the constraint of the process-based models and hydrological models in optimizing agricultural water management, there is a growing trend of integrating/coupling modeling studies. For example, there are case studies in NCP to improve the groundwater depletion simulation by coupling the crop model (DSSAT) with hydrology model (SWAT) and groundwater flow model (MODFLOW) (Hu et al. 2010). The Catchment-based Ecohydrology model was integrated with the crop model (DSSAT) to simulate the impact of agricultural water consumption on groundwater resources (Nakayama et al. 2006). In addition, there is a great potential of applying these model-based studies to identify the sustainable water management for different crops and various cropping systems under future climate projections, which will greatly support the agricultural adaptation decisions for the local farmers, stakeholders, and government.

Environmental Extended Multi-Region Input–Output Model

In addition to the water resources and direct consumption simulated by physical models, the indirect virtual water embodied in interregional food trade is also estimated by using the environmental extended multi-region input–output model (Wiedmann 2009). The virtual water or water footprint represents both the direct and indirect volumes of freshwater consumption to produce goods and services which are consumed by social and economic activities (Hubacek et al. 2009). The agricultural virtual water, which includes irrigation water as blue water, rainwater as green water, and wastewater as gray water, was introduced to identify the embodied water import and export from interregional trade and to design alternative water management strategies (Chapagain and Hoekstra 2004). A more comprehensive understanding of the direct and indirect water consumption, especially the groundwater depletion embodied food trade, will support local agricultural adaptations and reallocation/alleviation of water stress in water-deficit regions via food and product trade (Dalin et al. 2017). For example, by tracking "virtual water" embodied in the global and regional food trade, researchers showed that a great amount of food is exported from the water-deficit arid and semi-arid regions at the expense of greater groundwater depletion, and it is not reasonable for those regions to continue their role as "breadbasket" for other regions in the world (Dalin et al. 2017; Ren et al. 2018).

Water-Sustainable Agricultural Adaptations

As the biggest water consumption sector, agriculture plays a critical role in local water management. Reducing irrigation water and increasing water-use efficiency are critical to the environment and ecosystem sustainability, especially in the major cropping regions under arid and semi-arid climate.

Field Water-Saving Irrigation Technologies

Directly reducing agricultural irrigation is the most prevalent measure to save water. Field experiments have been designed and validated to identify the optimal water management for different crops in the arid and semi-arid regions. There are two major field water-saving practices: (1) water-saving irrigation and optimal irrigation schedule, and (2) plastic/straw mulching and no tillage.

Various water-saving irrigation technologies were developed to reduce the frequency and amount of irrigation. Compared with traditional flood or sprinkler irrigation, the dripping irrigation will significantly increase the water-use efficiency by the crops with little "waste" water from surface runoff and percolation (Alcon and Burton 2011; Ayars et al. 2015). More aggressive deficit irrigation with the tradeoffs of crop yield loss was also adopted in the arid regions (Geerts et al. 2010; Al-Ghobari and Dewidar 2018). Optimizing the irrigation schedule with lower frequency and a focus on the critical crop growth stages is another effective practice (Jiang et al. 2016). Experiments showed that optimal irrigation could reduce irrigation water by about 50 mm compared with local farmer's current practices in the NCP (Sun et al. 2011). Maize irrigation amount could reduce up to 50% in the Yingke irrigation district in Northwest China under the optimal irrigation schedule (Jiang et al. 2016). However, water conservation subsidies for water-efficient irrigation may consume even more water as groundwater recharge reduced while more water was extracted to meet the demand of water-intensive crops, more comprehensive social and economic policies are needed to achieve groundwater conservation instead of solely focusing on water-saving technologies (Ward and Pulido-Velazquez 2008).

Plastic mulching during the crop seedling stage is an effective way to reduce irrigation by decreasing soil evaporation and conserve soil moisture during the dry spring in Northwest China (Ran et al. 2018). Because of the negative environmental impact of plastic film, a more environment-friendly straw residual mulching and no tillage practice was developed: winter wheat straw residuals were left after harvest and no tillage was applied for soil moisture conservation before the following summer crop sowing (Tan et al. 2017). Experiments in the Luancheng Station of the NCP showed that soil evaporation could be reduced up to 40% with higher water-use efficiency of maize (Chen et al. 2007).

Cropping System Adaptations

However, no matter how much efforts were put on reducing irrigation water, the groundwater still dropped in the NCP, even when reducing the irrigation frequency to one time per crop growth cycle (deficit irrigation at the expense of limit crop yield reduction) (Sun et al. 2015). This is because the water requirements of the existing intensive wheat–maize double cropping systems are far more than local water supply. Agricultural deintensification and water-sustainable cropping systems become a feasible alternative (Zhong et al. 2017).

Many groundwater neutral (no groundwater table drop) cropping system with less water demand crops or cropland fallow were introduced to further reduce irrigation. For example, because of the dry winter and spring in NCP, winter wheat required much more irrigation water than other crops. Many alternative cropping system with partially or completely wheat fallow were introduced to replace the dominant wheat–maize cropping system in this region, including single cropping of spring maize (Pei et al. 2015), double cropping per year of early maize and late maize (Meng et al. 2017), and triple harvests in 2 years of WM-S followed by spring maize (Meng et al. 2012), and strip–relay intercropping of wheat followed by spring maize (Gao et al. 2009). A slow adaptation of wheat planting area shifting has happened at the farmer's initiative in the NCP. A "spring maize belt" was established in the driest regions of the NCP. Farmers had already abandoned winter wheat and replaced the wheat–maize double cropping system with spring maize single cropping (Feng et al. 2007; Wang et al. 2016); and, a north–south shift of winter wheat planting area occurred where more wheat was planted in the wetter part of NCP (Wang et al. 2015). Inevitably, the regional total grain production will suffer a loss when less intensive cropping systems are adopted for groundwater recovery. The current situation of groundwater over-exploitation should come to an end before there is "no water" to use when extreme drought happens and a great food production failure under projected warmer and drier climate in NCP (Oort et al. 2016).

Fortunately, policies have been made and implemented in the groundwater over-exploited NCP, where a pilot project has started to leave cropland area fallow. Ecological compensation was also estimated for winter wheat fallow in NCP (Wang et al. 2016). Investments to improve the cropland irrigation infrastructure, such as concrete canals to reduce leakage of water when transferring, are also needed to enhance the water saving and groundwater recovery.

Physical Water Transfer or Virtual Water Flow

Freshwater Diversion Projects

The uneven distribution of freshwater resource (e.g., the arid and semi-arid NCP produces 1/2 and 1/3 of the national total wheat and maize production, respectively, while south China has much more freshwater resources) and the soaring water demand from rapid urbanization and industrialization have driven many massive surface water diversion projects (Aeschbach-Hertig and Gleeson 2012; Zhao et al. 2015). The biggest one is the South-North Water Transfer (SNWT) project, with a designed water delivery capacity of 20 billion m^3 water from the western route, 9.5 billion m^3 water from the central route, and 14.8 billion m^3 water from the eastern route (Liu et al. 2013).

This SNWT project has also raised great controversy because of its huge investment and cost for maintenance/operation every year and the environmental impact to the downstream (Barnett et al. 2015). Moreover, studies also pointed out that even with the supplementary water from SNWT project, the groundwater table level would not stop declining in Shijiazhuang city in North China (Shu et al. 2012).

Adaptations, such as improving agricultural water-use efficiency and reducing crop irrigation, increasing urban water supply from rainwater harvesting and wastewater recycling, should play a more important role for the water sustainability in those regions (Barnett et al. 2015).

Virtual Water Flows via Food Trade

Compared with physical water transfer, the virtual water flow as embodied water in the interregional food and production trade is much bigger. Physical water flows via the major water transfer projects amounted to 4.5% of national water supply, whereas virtual water flows accounted for 35% in 2007 (Zhao et al. 2015). This suggests the demand-oriented water management should be highlighted rather than solely focusing on physical water saving, especially reducing the export of water-intensive products from the water-deficit regions. For example, the physically transferred freshwater was less than the embodied blue water export from Hebei Province to Beijing and Tianjin and had worsened the local groundwater depletion. Moreover, export from Hebei province with much less water-use efficiency to the higher water-use efficient Beijing and Tianjin was also not reasonable (Zhao et al. 2017). Case studies in the NCP indicated that reducing the export from low water-use efficiency and high-water-stress regions will be more effective than transfer freshwater via huge water diversion projects, despite the negative sides of huge investment, cost of maintain and migration, and environment impact to the downstream (Zhang 2009; He et al. 2010). Therefore, reducing the heavy-water-intensity products' export, increasing water-use efficiency, and giving up the role of "breadbasket" for Beijing and Tianjin should be the next step in Hebei Province (Ren et al. 2018). Appropriate subsidies to compensate the local farmers' income loss are critical to the success of cropland fallow in the NCP (Wang et al. 2016), and 87% of the local farmers would adopt the winter wheat fallow policy (Xie et al. 2017).

Conclusions

Soaring global water consumption from agriculture, urbanization, and industrialization, which is expected to grow up to 21% in 2050 (Rosegrant et al. 2009), and the unsustainable use of groundwater have caused severe environmental issues in many arid and semi-arid regions. The potential water crisis has raised great concern from the public and government. As the biggest water consumption sector, improving agricultural water-use efficiency and reducing crop irrigation water consumption is critical to global water sustainability.

Various models, agricultural water-saving adaptations, and water alleviation strategies were developed and tested at the field, regional, and global levels. Many water-saving irrigation technologies showed the potential of reducing crop irrigation and maintaining high level of food production. Initial plans and policies had already taken to slow or stop the drop in the groundwater drop in the over-exploited regions. On the other hand, instead of only focusing on improving field water-saving technologies, water-sustainable cropping systems adaptations, and supplementary freshwater transfer projects to alleviate the water deficit, it is also very important to increase the overall water-use efficiency and reduce the water footprint in all social-economic activities. "Virtual water" flow and economic cost-effective approaches should also be taken into consideration when deciding the local water diversion projects. In addition, a broader view of the water management should include social-economic water demand/consumption and local water resource. The Global Hydro-economic Model (Wada et al. 2016) can provide a feasible solution for the water management under different future climate projections and social-economic development scenarios.

Nevertheless, local water sustainability needs joint efforts from local farmers, government, and even global corporations. Future risks from projected climate change, growing population, and urbanization, shifting to more water-intensive lifestyle in many developing countries, still put global water shortage in an alarming situation, and continued efforts to improve the water management are required for sustainable development.

References

Abbott M., J. Bathurst, J. Cunge, P. O'Connell, and J. Rasmussen. 1986. "An Introduction to the European Hydrological System — Systeme Hydrologique Europeen, 'SHE', 1: History and Philosophy of a Physically-Based, Distributed Modelling System." *Journal of Hydrology* 87 (1): 45–59.

Aeschbach-Hertig Werner and Tom Gleeson. 2012. "Regional Strategies for the Accelerating Global Problem of Groundwater Depletion." *Nature Geoscience* 5: 853.

Alcon Francisco María Dolores de Miguel and Michael Burton. 2011. "Duration Analysis of Adoption of Drip Irrigation Technology in Southeastern Spain." *Technological Forecasting and Social Change* 78 (6): 991–1001.

Al-Ghobari Hussein M. and Ahmed Z. Dewidar. 2018. "Integrating Deficit Irrigation into Surface and Subsurface Drip Irrigation as a Strategy to Save Water in Arid Regions." *Agricultural Water Management* 209: 55–61.

Arnold J., R. Srinivasan, R.S. Muttiah, and J.R. Williams. 1998. "Large Area Hydrologic Modeling and Assessment Part I: Model Development." *JAWRA: Journal of the American Water Resources Association* 34 (1): 73–89.

Ayars J., A. Fulton, and B. Taylor. 2015. "Subsurface Drip Irrigation in California—Here to Stay?" *Agricultural Water Management* 157: 39–47.

Barnett Jon, Sarah Rogers, Michael Webber, Brian Finlayson, and Mark Wang. 2015. "Sustainability: Transfer Project Cannot Meet China's Water Needs." *Nature* 527 (7578): 295–97.

Chapagain A. and A. Hoekstra. 2004. *Water Footprints of Nations*. Value of Water Research Report Series BT – Water Footprints of Nations. Delft: Unesco-IHE Institute for Water Education.

Chen S., X. Zhang, D. Pei, H. Sun, and S. Chen. 2007. "Effects of Straw Mulching on Soil Temperature, Evaporation and Yield of Winter Wheat: Field Experiments on the North China Plain." *Annals of Applied Biology* 150 (3): 261–68.

Currell Matthew J., Dongmei Han, Zongyu Chen, and Ian Cartwright. 2012. "Sustainability of Groundwater Usage in Northern China: Dependence on Palaeowaters and Effects on Water Quality, Quantity and Ecosystem Health." *Hydrological Processes* 26 (26): 4050–66.

Dalin Carole, Yoshihide Wada, Thomas Kastner, and Michael J. Puma. 2017. "Groundwater Depletion Embedded in International Food Trade." *Nature* 543: 700.

Feng Zhiming, Dengwei Liu, and Yuehong Zhang. 2007. "Water Requirements and Irrigation Scheduling of Spring Maize Using GIS and CropWat Model in Beijing-Tianjin-Hebei Region." *Chinese Geographical Science* 17 (1): 56–63.

Gao Yang, Aiwang Duan, Jingsheng Sun, Fusheng Li, Zugui Liu, Hao Liu, and Zhandong Liu. 2009. "Crop Coefficient and Water-Use Efficiency of Winter Wheat/Spring Maize Strip Intercropping." *Field Crops Research* 111 (1): 65–73.

Geerts S., D. Raes, and M. Garcia. 2010. "Using AquaCrop to Derive Deficit Irrigation Schedules." *Agricultural Water Management* 98 (1): 213–16.

He Chansheng, Xiaoying He, and Li Fu. 2010. "China's South-to-North Water Transfer Project: Is It Needed?" *Geography Compass* 4 (9): 1312–23.

Hu Yukun, Juana Paul Moiwo, Yonghui Yang, Shumin Han, and Yanmin Yang. 2010. "Agricultural Water-Saving and Sustainable Groundwater Management in Shijiazhuang Irrigation District, North China Plain." *Journal of Hydrology* 393 (3): 219–32.

Hubacek Klaus, Dabo Guan, John Barrett, and Thomas Wiedmann. 2009. "Environmental Implications of Urbanization and Lifestyle Change in China: Ecological and Water Footprints." *Journal of Cleaner Production* 17 (14): 1241–48.

Jiang Yiwen, Lanhui Zhang, Baoqing Zhang, Chansheng He, Xin Jin, and Xiao Bai. 2016. "Modeling Irrigation Management for Water Conservation by DSSAT-Maize Model in Arid Northwestern China." *Agricultural Water Management* 177: 37–45.

Jones J., G. Hoogenboom, C. Porter, K. Boote, W. Batchelor, L. Hunt, P. Wilkens, U. Singh, A. Gijsman, and J. Ritchie. 2003. "The DSSAT Cropping System Model." *European Journal of Agronomy* 18 (3): 235–65.

Kang Suchul and Elfatih A. B. Eltahir. 2018. "North China Plain Threatened by Deadly Heatwaves Due to Climate Change and Irrigation." *Nature Communications* 9 (1): 2894.

Konikow Leonard and Eloise Kendy. 2005. "Groundwater Depletion: A Global Problem." *Hydrogeology Journal* 13 (1): 317–20.

Liu Junguo, Chuanfu Zang, Shiying Tian, Jianguo Liu, Hong Yang, Shaofeng Jia, Liangzhi You, Bo Liu, and Miao Zhang. 2013. "Water Conservancy Projects in China: Achievements, Challenges and Way Forward." *Global Environmental Change* 23 (3): 633–43.

McCown R., G. Hammer, J. Hargreaves, D. Holzworth, and D. Freebairn. 1996. "APSIM: A Novel Software System for Model Development, Model Testing and Simulation in Agricultural Systems Research." *Agricultural Systems* 50 (3): 255–71.

Meng Qingfeng, Hongfei Wang, Peng Yan, Junxiao Pan, Dianjun Lu, Zhenling Cui, Fusuo Zhang, and Xinping Chen. 2017. "Designing a New Cropping System for High Productivity and Sustainable Water Usage under Climate Change." *Scientific Reports* 7: 41587.

Meng Qingfeng, Qinping Sun, Xinping Chen, Zhenling Cui, Shanchao Yue, Fusuo Zhang, and Volker Römheld. 2012. "Alternative Cropping Systems for Sustainable Water and Nitrogen Use in the North China Plain." *Agriculture, Ecosystems & Environment* 146 (1): 93–102.

Molden David. 2007. *Water for Food, Water for Life: A Comprehensive Assessment of Water Management in Agriculture*. London, Earthscane, Colombo: International Water Management Institute.

Nakayama Tadanobu, Yonghui Yang, Masataka Watanabe, and Xiying Zhang. 2006. "Simulation of Groundwater Dynamics in the North China Plain by Coupled Hydrology and Agricultural Models." *Hydrological Processes* 20 (16): 3441–66.

Oort P.A.J van, G. Wang, J. Vos, H. Meinke, B. Li, J. Huang, and W. van der Werf. 2016. "Towards Groundwater Neutral Cropping Systems in the Alluvial Fans of the North China Plain." *Agricultural Water Management* 165: 131–40.

Pei Hongwei, Bridget R. Scanlon, Yanjun Shen, Robert C. Reedy, Di Long, and Changming Liu. 2015. "Impacts of Varying Agricultural Intensification on Crop Yield and Groundwater Resources: Comparison of the North China Plain and US High Plains." *Environmental Research Letters* 10 (4): 44013.

Piao Shilong, Philippe Ciais, Yao Huang, Zehao Shen, Shushi Peng, Junsheng Li, Liping Zhou, et al. 2010. "The Impacts of Climate Change on Water Resources and Agriculture in China." *Nature* 467: 43.

Ran Hui, Shaozhong Kang, Fusheng Li, Taisheng Du, Ling Tong, Sien Li, Risheng Ding, and Xiaotao Zhang. 2018. "Parameterization of the AquaCrop Model for Full and Deficit Irrigated Maize for Seed Production in Arid Northwest China." *Agricultural Water Management* 203: 438–50.

Ren Dandan, Yonghui Yang, Yanmin Yang, Keith Richards, and Xinyao Zhou. 2018. "Land-Water-Food Nexus and Indications of Crop Adjustment for Water Shortage Solution." *Science of The Total Environment* 626: 11–21.

Rosegrant Mark, Claudia Ringler, and Tingju Zhu. 2009. "Water for Agriculture: Maintaining Food Security under Growing Scarcity." *Annual Review of Environment and Resources* 34 (1): 205–22.

Shu Yunqiao, Karen G. Villholth, Karsten H. Jensen, Simon Stisen, and Yuping Lei. 2012. "Integrated Hydrological Modeling of the North China Plain: Options for Sustainable Groundwater Use in the Alluvial Plain of Mt. Taihang." *Journal of Hydrology* 464–465: 79–93.

Steduto Pasquale, Theodore C. Hsiao, Dirk Raes, and Elias Fereres. 2009. "AquaCrop-The FAO Crop Model to Simulate Yield Response to Water: I. Concepts and Underlying Principles." *Agronomy Journal* 101 (3): 426–37.

Sun Chen and L. Ren. 2013. "Assessment of Surface Water Resources and Evapotranspiration in the Haihe River Basin of China Using SWAT Model." *Hydrological Processes* 27 (8): 1200–222.

Sun Chen and L. Ren. 2014. “Assessing Crop Yield and Crop Water Productivity and Optimizing Irrigation Scheduling of Winter Wheat and Summer Maize in the Haihe Plain Using SWAT Model.” *Hydrological Processes* 28 (4): 2478–98.

Sun Hongyong, Xiying Zhang, Enli Wang, Suying Chen, and Liwei Shao. 2015. “Quantifying the Impact of Irrigation on Groundwater Reserve and Crop Production – A Case Study in the North China Plain.” *European Journal of Agronomy* 70: 48–56.

Sun Qinping, Roland Kröbel, Torsten Müller, Volker Römheld, Zhenling Cui, Fusuo Zhang, and Xinping Chen. 2011. “Optimization of Yield and Water-Use of Different Cropping Systems for Sustainable Groundwater Use in North China Plain.” *Agricultural Water Management* 98 (5): 808–14.

Tan Shuai, Quanjiu Wang, Di Xu, Jihong Zhang, and Yuyang Shan. 2017. “Evaluating Effects of Four Controlling Methods in Bare Strips on Soil Temperature, Water, and Salt Accumulation under Film-Mulched Drip Irrigation.” *Field Crops Research* 214: 350–58.

Tian Zhan, Honglin Zhong, Runhe Shi, Laixiang Sun, Gunther Fischer, and Zhuoran Liang. 2012. “Estimating Potential Yield of Wheat Production in China Based on Cross-Scale Data-Model Fusion.” *Frontiers of Earth Science* 6 (4): 388–95.

Tian Zhan, Yilong Niu, Dongli Fan, Laixiang Sun, Gunther Ficsher, Honglin Zhong, Jia Deng, and Francesco N. Tubiello. 2018. “Maintaining Rice Production While Mitigating Methane and Nitrous Oxide Emissions from Paddy Fields in China: Evaluating Tradeoffs by Using Coupled Agricultural Systems Models.” *Agricultural Systems* 159: 175–86.

Vörösmarty Charles, Pamela Green, Joseph Salisbury, and Richard B. Lammers. 2000. “Global Water Resources: Vulnerability from Climate Change and Population Growth.” *Science* 289 (5477): 284–88.

Wada Y., M. Flörke, N. Hanasaki, S. Eisner, G. Fischer, S. Tramberend, Y. Satoh, M. van Vliet, P. Yillia, C. Ringler, P. Burek, and D. Wiberg. 2016. Modeling Global Water Use for the 21st Century: Water Futures and Solutions (WFaS) Initiative and Its Approaches. *Geoscientific Model Development* 9: 175–222.

Wang Ruoyu, Laura C. Bowling, and Keith A. Cherkauer. 2016a. “Estimation of the Effects of Climate Variability on Crop Yield in the Midwest USA.” *Agricultural and Forest Meteorology* 216: 141–56.

Wang Xue, Xiubin Li, Günther Fischer, Laixiang Sun, Minghong Tan, Liangjie Xin, and Zhuoran Liang. 2015. “Impact of the Changing Area Sown to Winter Wheat on Crop Water Footprint in the North China Plain.” *Ecological Indicators* 57: 100–109.

Wang Xue, Xiubin Li, Liangjie Xin, Minghong Tan, Shengfa Li, and Renjing Wang. 2016b. “Ecological Compensation for Winter Wheat Abandonment in Groundwater Over-Exploited Areas in the North China Plain.” *Journal of Geographical Sciences* 26 (10): 1463–76.

Ward Frank A. and Manuel Pulido-Velazquez. 2008. “Water Conservation in Irrigation Can Increase Water Use.” *Proceedings of the National Academy of Sciences* 105 (47): 18215–20.

Wiedmann Thomas. 2009. “A Review of Recent Multi-Region Input–Output Models Used for Consumption-Based Emission and Resource Accounting.” *Ecological Economics* 69 (2): 211–22.

Williams J., C. Jones, J. Kiniry, and D. Spanel. 1989. “The EPIC Crop Growth Model.” *Transactions of the ASAE* 32 (2): 497–511.

Wood E., D. Lettenmaier, and V. Zartarian. 1992. “A Land-Surface Hydrology Parameterization with Subgrid Variability for General Circulation Models.” *Journal of Geophysical Research - Atmospheres* 97 (D3): 2717–28.

Xie Hualin, Lingjuan Cheng, and Tiangui Lv. 2017. “Factors Influencing Farmer Willingness to Fallow Winter Wheat and Ecological Compensation Standards in a Groundwater Funnel Area in Hengshui, Hebei Province, China”. *Sustainability* 9 (5): 839.

Xiong Lvyang, Xu Xu, Dongyang Ren, Quanzhong Huang, and Guanhua Huang. 2019. “Enhancing the Capability of Hydrological Models to Simulate the Regional Agro-Hydrological Processes in Watersheds with Shallow Groundwater: Based on the SWAT Framework.” *Journal of Hydrology* 572: 1–16.

Xue Yuqun, Jichun Wu, Shujun Ye, and Yongxiang Zhang. 2000. "Hydrogeological and Hydrogeochemical Studies for Salt Water Intrusion on the South Coast of Laizhou Bay, China." *Groundwater* 38 (1): 38–45.

Yang Xiaolin, Yuanquan Chen, Steven Pacenka, Wangsheng Gao, Min Zhang, Peng Sui, and Tammo S. Steenhuis. 2015. "Recharge and Groundwater Use in the North China Plain for Six Irrigated Crops for an Eleven Year Period." *PLOS ONE* 10 (1): e0115269.

Yang Yonghui, Masataka Watanabe, Xiying Zhang, Xiaohua Hao, and Jiqun Zhang. 2006. "Estimation of Groundwater Use by Crop Production Simulated by DSSAT-Wheat and DSSAT-Maize Models in the Piedmont Region of the North China Plain." *Hydrological Processes* 20 (13): 2787–2802.

Yu Q., S. Saseendran, L. Ma, G. Flerchinger, T. Green, and L. Ahuja. 2006. "Modeling a Wheat–Maize Double Cropping System in China Using Two Plant Growth Modules in RZWQM." *Agricultural Systems* 89 (2–3): 457–77.

Zhang Quanfa. 2009. "The South-to-North Water Transfer Project of China: Environmental Implications and Monitoring Strategy1." *JAWRA Journal of the American Water Resources Association* 45 (5): 1238–47.

Zhang Xueliang, Li Ren, and Xiangbin Kong. 2016. "Estimating Spatiotemporal Variability and Sustainability of Shallow Groundwater in a Well-Irrigated Plain of the Haihe River Basin Using SWAT Model." *Journal of Hydrology* 541: 1221–40.

Zhao Dandan, Yu Tang, Junguo Liu, and Martin R. Tillotson. 2017. "Water Footprint of Jing-Jin-Ji Urban Agglomeration in China." *Journal of Cleaner Production* 167: 919–28.

Zhao Xu, Junguo Liu, Qingying Liu, Martin R. Tillotson, Dabo Guan, and Klaus Hubacek. 2015. "Physical and Virtual Water Transfers for Regional Water Stress Alleviation in China." *Proceedings of the National Academy of Sciences* 112(4): 1031–35.

Zhong Honglin, Laixiang Sun, Günther Fischer, Zhan Tian, and Zhuoran Liang. 2019. "Optimizing Regional Cropping Systems with a Dynamic Adaptation Strategy for Water Sustainable Agriculture in the Hebei Plain." *Agricultural Systems* 173 (July): 94–106.

Zhong Honglin, Laixiang Sun, Gunther Fischer, Zhan Tian, Harrij van Velthuizen, and Zhuoran Liang. 2017. "Mission Impossible? Maintaining Regional Grain Production Level and Recovering Local Groundwater Table by Cropping System Adaptation across the North China Plain." *Agricultural Water Management* 193: 1–12.

21

Runoff Water

Zaneta Polkowska

Introduction

The composition of the stream of wet deposition changes significantly on contact with the substrate, be this natural or artificial. Following contact with the surfaces of artificial objects (e.g., roofs, roads, airport aprons and runways, steel and timber structures, communal and industrial landfill, cultivated fields) or natural surfaces (e.g., trees, shrubs, conifer needles, leaves, grass, vegetated fallow land), precipitation and atmospheric deposits (rain, dew, fog, melting snow, hoar frost, or rime), flow over them (runoff) or, trickling through the soil, replenish groundwaters (Figure 1). The extent of surface runoff depends on the properties of the soil and the degree to which it is covered by plants—the lusher the vegetation, the less runoff there is. Certain properties of the soil, such as minimal adsorption or good permeability, promote the vertical movement of water in the soil, enabling it to reach the groundwater; the intensity of surface runoff is therefore less. Conversely, if the properties of a surface favor runoff (e.g., heavy, impermeable soils, asphalt, or concrete surfaces), pollutants are transported across the land surface.[1,2]

Pollutants in runoff come from three main sources: dry deposition (from the atmosphere), wet deposition (precipitation and atmospheric deposits), and flushing from the surface of the material with which precipitation and deposits come into contact (e.g., roof coverings, the crowns and trunks of trees, cultivated fields, roads and railway lines, landfill sites).[3]

Modification of the Stream of Wet Deposition following Contact with Substances in/on the Substrate

The composition of atmospheric water changes radically on contact with the Earth's surface. The composition of runoff depends to a large extent on the type of surface (including that of the land) in which precipitation falls on. Runoff from fields may contain pesticides and nutrients. Urban areas may be the source of many different substances, such as aliphatic and aromatic hydrocarbons, Polycyclic aromatic hydrocarbons (PAHs), fatty acids, ketones, and heavy metals.

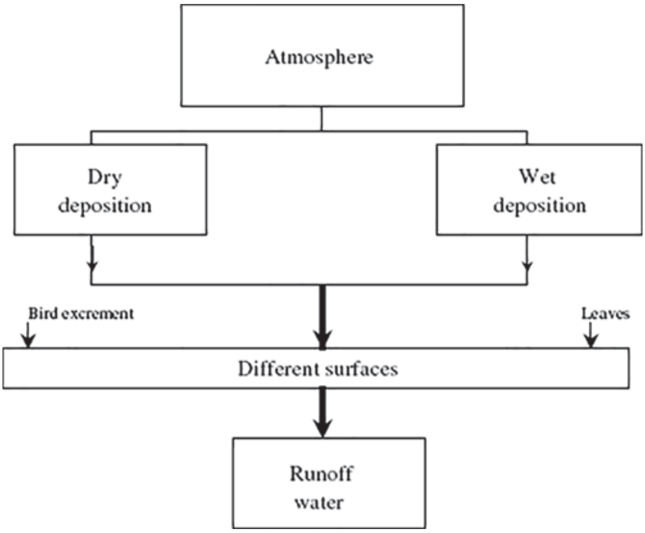


FIGURE 1 Modifications to the flow of wet deposition obstructed by different surfaces (trees, lawns, roofs, roads, motor vehicles, airport aprons and runways, steel and timber structures, communal and industrial landfill sites).

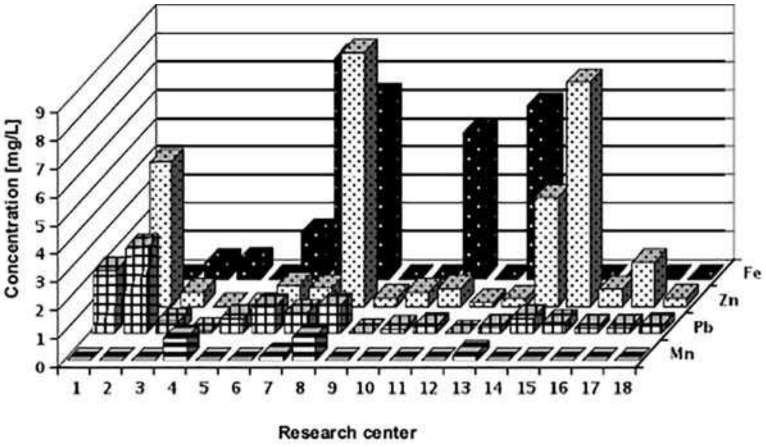
The composition of precipitation in wooded areas changes mainly under the influence of the physiological processes taking place in plants and the dissolution or flushing of atmospheric aerosols and gaseous contaminants from the surfaces of branches, bark, and leaves.[4–15]

Runoff from urban areas may also contain petroleum hydrocarbons. Road runoff contains dusts derived from the wear and tear of asphalt road surfaces and rubber tires.[16–48] The material from which a roof covering or guttering is made is a potential source of pollution taken up by water falling onto these surfaces. The compounds leached out of these materials react with the contaminants in the rainwater, changing its composition and properties. Roof runoff usually contains heavy metals and pesticides.[1,43,49–70] On the other hand, runoff from landfill sites and refuse tips are variously aggressive towards the environment, depending on the state of degradation of the refuse. On this, however, the literature supplies no data. A considerable number of papers have been published on runoff from different surfaces, although there is hardly any information on the concentrations of the constituents of runoff samples collected in Poland; what little there is refers to runoff from the crowns and trunks of trees.[5,15]

Road Runoff

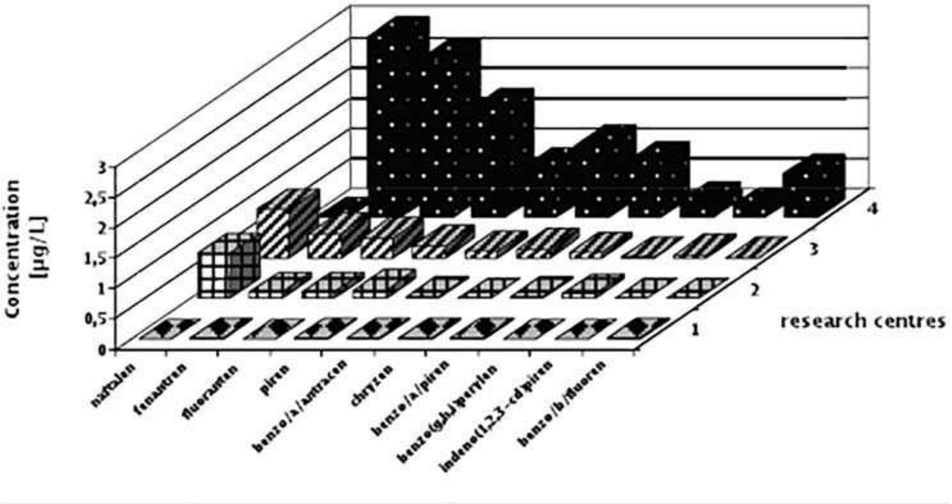
Waters running off motorways contain heavy metals in two forms: suspended solid particles and dissolved forms. Figure 2 gives the results of determinations of different heavy metal ions in samples of road runoff from Europe and the United States. Similar studies were carried out in New Zealand. The results of these studies were the basis for concluding that the content of lead ions in plants (growing near roads) depends significantly on how far from the roadside they are growing.[47]

Road runoff is to a large extent contaminated by organic compounds like petroleum constituents, fuels, fats, and greases. Studies done in Madrid have shown that there is a strict relationship between the intensity of urban road traffic and the levels of PAHs and petroleum hydrocarbons in urban road runoff.[17,19] Vehicle emissions—residues of fuel and its incomplete combustion—are the chief source of these compounds in runoff.[48] Figure 3 illustrates the results of determinations of various PAHs in runoff samples collected in Europe and the United States. The concentrations were highest in the runoff samples collected along roads with a high intensity of traffic (Spain, Poland).



1 - USA [rural area - Alabama] [34]	*10 - Denmark [motorway] [71]*
2 - Switzerland [urban area - Zürich] [61]	*11 - Switzerland [motorway] [22]*
3 - England [motorway] [35]	*12 - Switzerland [1,600 vehicles / h] [41]*
4 - England [motorway] [36]	*13 - West Germany [urban area - Hamburg] [42]*
5 - England [2,000 vehicles / h] [37]	*14 - France [urban area - Paris] [43]*
6 - Norway [motorway] [21]	*15 - USA [ca 5,000 vehicles / h] [20]*
7 - USA [ca 1,000 vehicles / h] [38]	*16 - France [2,000 vehicles / day] [44]*
8 - USA [ca 7,000 vehicles / h] [39]	*17 - France [800 vehicles / day] [29]*
9 - Belgium [urban area – Brussels] [40]	*18 - France [urban area] [30]*

FIGURE 2 Concentrations of iron, zinc, lead, and manganese ions determined in samples of runoff water in Europe and the United States (literature data).



1 - England [1,600 vehicles / h] [72]	*3 - Poland [urban area – Tri-City] [73]*
2 - Hungary [urban area - Budapest] [17]	*4 - Spain 1988 [11,500 vehicles / h] [19]*

FIGURE 3 PAH content in samples of runoff water collected in England, Hungary, Poland, and Spain.

Studies regarding the presence of volatile organochlorine compounds (VOCCs) in road runoff were carried out in Poland (Tri-City: Gdansk, Sopot, Gdynia). VOCCs were found in every single sample. The most common VOCC analytes were chloroform, dichlorobromomethane, and trichloroethene. Table 1 lists the ranges of concentrations and mean concentrations of analytes in road runoff samples.[74]

Road runoff samples from Poland have also been analyzed for the presence of PAHs. The highest levels were recorded for naphthalene and phenanthrene+anthracene, while fluoranthene, pyrene, benzo(*a*)anthracene+chrysene, benzo(*b*+*k*)fluoranthene, and benzo(*a*)pyrene were also quite frequent. The properties of asphalt road surfaces (rather smooth and impermeable) favor runoff; thus, contaminants derived mainly from road vehicles (exhaust fumes, wear and tear of asphalt, and rubber tires) are transported across the land surface. The levels of PAH analytes rise conspicuously in the winter months (November to January).[75,76]

The organochlorine pesticides usually detected in road runoff samples from Poland include heptachlor epoxide (a product of the metabolization of heptachlor, which is synthesized in plants, the soil, and homeothermic organisms) and *o,p′*-DDE (a derivative of DDT, an insecticide that used to be widely applied). The most common of the organonitrogen and organophosphorus pesticides to be detected in road runoff are propazine and bromofos (selective weed killers). Table 2 lists the mean concentrations and ranges of concentrations of pesticides determined in road runoff samples from the Tri-City area of Poland.[74,77]

The most commonly detected petroleum hydrocarbons in runoff samples from Poland were benzene and toluene, whereby the former compound was present in every sample analyzed. Weather conditions and the street layouts of towns and cities, i.e., the speed and direction of the wind and the proximity of city buildings to one another, are important factors affecting the spread of these pollutants. Table 3 lists the maximum, minimum, and mean concentrations of analytes found in samples; the number of samples in which an analyte was detected is also given.[74,76,77]

Formaldehyde was detected in all samples of road runoff from Poland (from 0.235 to 0.433 mg/L). Figure 4 compares the levels of formaldehyde in samples of rainwater and runoff from roads in the Gdansk area and from roadside sampling points in Poland.

The formaldehyde content in road runoff is high as it is a typical pollutant produced by road vehicles. It is formed as a result of the partial oxidation of fuel in the cylinder due to the chain reactions oxidizing fuel being interrupted by the “extinguishing” effect of the fairly cool cylinder walls or to a local oxygen deficit, and is emitted with the exhaust into the atmosphere, from which it may be scavenged by dry or wet deposition. Investigations have demonstrated unequivocally that formaldehyde levels increase with increasing motor vehicle traffic in both rainwater and runoff.[78,79]

The total concentration of phenol derivatives has also been determined in road runoff samples from Poland. The usual compounds in this respect were phenol and 3,4-di- methylphenol, and the highest levels were recorded for phenol (from 91 to 120 ng/L), chlorophenol (613 ng/L), and p-nitrophenol (469 ng/L). *o*-Cresol was not detected in any of the samples. This level of pollution by phenols in runoff

TABLE 1 Ranges of Concentrations and Mean Concentrations of Analytes in Road Runoff Samples

	Concentration (μg/L)		
Analyte	Minimum	Maximum	Mean
$CHCl_3$	0.01	0.53	0.59
$CHBrCl_2 + C_2HCl_3$	n.d.	0.08	0.02
$CHBr_2Cl$	n.d.	0.01	–
CCl_4	n.d.	0.01	–

Note: n.d., not determined.
Source: Polkowska et al.[74]

TABLE 2 Mean Concentrations and Ranges of Concentrations of Pesticides Determined in Road Runoff Samples

	Concentration (µg/L)		
Analyte	Minimum	Maximum	Mean
	Organo-N,-P pesticides		
Simazine	n.d.	0.48	0.27
Atrazine	n.d.	4.10	0.71
Propazine	n.d.	12.1	3.25
t-Butazine	n.d.	12.4	2.38
Bromofos	n.d.	2.86	0.27
Malathion	n.d.	0.46	0.11
Chlorfenvinfos	n.d.	0.13	0.057
Phenitrothion	n.d.	0.13	0.08
	Organochlorine pesticides		
α-HCH	n.d.	5.83	3.36
γ-HCH	n.d.	61.4	31.3
Aldrin	n.d.	21.1	2.68
Heptachlor	n.d.	57.1	8.75
epoxide		127	18.1
o,p′-DDE	n.d.	8.38	3.09
p,p′-DDD	n.d.	1308	212
o,p′-DDD	n.d.	213	29.4
p,p′-DDE	n.d.	2.41	1.38
o,p′-DDT	n.d.	2.88	1.66
p,p′-DDT	n.d.	5.70	2.63
Methoxychlor	n.d.		

Note: n.d., not determined.
Source: Polkowska et al.[77]

TABLE 3 Maximum, Minimum, and Mean Concentrations of Analytes Found in Samples

	Concentration of Analyte (µg/L)		
Compound	Minimum	Maximum	Mean
Benzene	0.86	109	48.9
Toluene	0.52	54.3	32.5
Ethylbenzene	0.12	39.8	18.6
p, *m*-xylene	<0.03	15.3	0.96
o-xylene	<0.03	27.0	0.99
Styrene	<0.03	11.3	1.18
Cumene	0.1	2.54	0.16
m-Dichlorobenzene	<0.03	12.7	4.15
Heptane	<0.03	<0.03	-
Isooctane	<0.03	14.3	3.50
Octane	<0.03	16.7	6.23
Cyclohexanone	<0.03	<0.03	-
Decane	<0.03	14.1	4.03
Dodecane	<0.03	14.2	2.50

Source: Polkowska et al.[77]

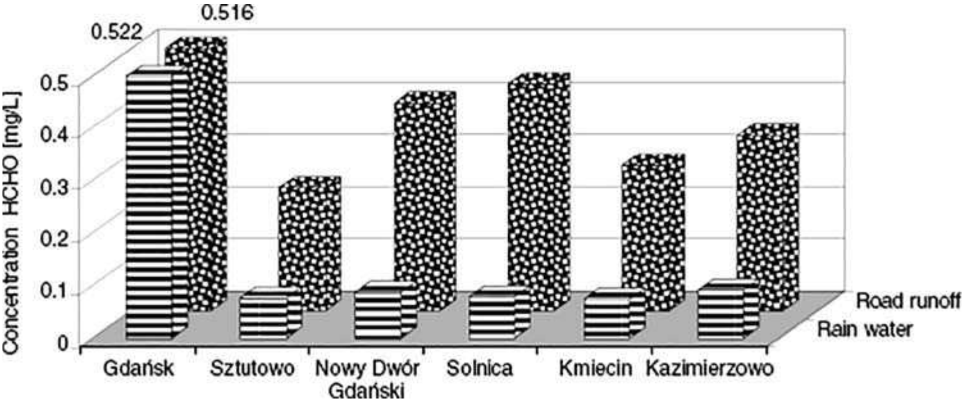


FIGURE 4 Formaldehyde (HCHO) levels determined in samples of rainwater and runoff from roads in the Gdansk area from roadside sampling points in Poland.
Source: Duncan.[78]

samples from the Tri-City area is due largely to the intensity of motor vehicle traffic and the proximity of industrial plants in the city, including the oil refinery, which discharges an average of 118 kg/yr of phenol derivatives into surface waters. The phenols present in the samples analyzed were probably formed as a result of photochemical reactions with the products of the incomplete combustion of fuel—benzene, toluene, and xylene.[80]

Roof Runoff

The most systematic investigations into the chemical composition of roof runoff have been carried out in countries where one of the sources of drinking water is water obtained from precipitation. The following factors affect the quality of water running off roofs:[81] the material from which the roof covering is made (its chemical composition, surface roughness, the area of roof covered, age, weather resistance), the physical boundary conditions (parameters) of the roof (surface area, inclination, exposure to the weather), the factors affecting precipitation (the direction and intensity of the wind, levels of pollutants in precipitation), other meteorological factors (season, type of air mass, duration of precipitation), the physicochemical properties of the pollutants (vapor pressure, partition coefficient, solubility in water, Henry's constant), and the concentration of pollutants in the atmosphere (the source and magnitude of emission, transport and spread, decomposition and degradation period, the physical form of the pollutant).

According to the available data, the pH of roof runoff is more basic than that of rainwater. In some cases, the pH may be higher, if basic soil particles have been windblown from eroded soil onto the roof surface. Roofs containing cement (concrete tiles, corrugated asbestos-cement sheet, and gravel) may cause the runoff pH to change, since they contain $CaCO_3$, which dissolves in rainwater. Table 4 presents information on the variability in runoff pH from various types of roofing materials.[54]

An in-depth study of the chemical composition of rainwater and roof runoff was carried out by German investigators. Concentrations of all the ions (except NH_4^+) were nearly twice as high in runoff waters than in rainwaters. The slight differences in the contents of ions in runoff from five types of roofing material suggest that the source of anions is more likely to be dry deposition than the roofing material.[56]

The contamination of roof runoff by heavy metals is widely described in the literature.[50,61,62,82,83] In this case, it is the roofing material that is the main factor affecting the level of these contaminants in water. Metals are the products of the corrosion of roofing and guttering materials; hence, water running off a galvanized sheet roof will contain large amounts of zinc and a certain quantity of cadmium, while roofs

TABLE 4 pH of Runoff Water from Different Types of Roofing Material

Type of Roofing Material	Median	Minimum	Maximum
Roofing felt	5.2	3.8	6.3
Cement tiles	5.4	4.6	6.5
Galvanized sheeting	6.9	6.5	7.6
Gravel roof	7.1	6.8	7.6
Concrete tiles	7.4	7.3	7.5

Source: Kennedy et al.[54]

made from copper sheeting will be a source of copper in the runoff.[64] Nonetheless, the metals present in roof runoff do not get there just as a result of leaching, since large quantities of metals have been found in runoff from roofing made from non-metallic materials, e.g., ceramic tiles, polyester, roofing felt. This does depend, however, on the location of the sampling point.[65] Heavy metal levels in roof runoff were highest in samples from the roofs of buildings situated in urban and industrial areas. Other factors affecting the amounts of heavy metals in roof runoff include the pH of rain and the angle of inclination of the roof.[64] Table 5 sets out the results of relevant research carried out in many scientific centers.

Pesticides and herbicides are frequently encountered in runoff waters. These compounds are increasingly being used in the production of roofing materials to prevent the growth on roofs of mosses and other plants. Evidence that pesticides are leached out of roofing materials was supplied by research carried out in Switzerland by a team of scientists from Eidgenössische Anstalt für Wasserversorgung, Abwasserreinigung und Gewässerschutz (EAWAG). Pesticide concentrations were lowest in runoff from ceramic tile roofs, higher in polyester roof runoff, and the highest from roofs covered with roofing felt. Particularly striking is the very high concentration of (*R*,*S*)-mecoprop in runoff from felt-covered roofs. This compound is not used on farms in Switzerland. The presence of (*R*,*S*)-mecoprop in runoff was due to the hydrolysis of Preventol® B2—a pesticide applied to root crops. It is used in large quantities as an additive to bitumen applied as a thin coating to areas on walls and roofs particularly susceptible to leaks, e.g., at the spot where the roofing felt is stuck to the wall.[51,84]

Table 6 lists the pesticides (and their concentrations) most commonly detected and determined in runoff from different types of roofing. The highest levels were recorded for heptachlor epoxide (roofing

TABLE 5 Levels of Metals in Samples of Roof Runoff with Respect to the Type of Roofing Material

	Cu	Zn	Cd	Pb
Type of Roofing Material	(μg/L)			
Ceramic tiles	1905	360	2.1	172
Polyester	6817	2076	3.1	510
Gravel	140	36	0.2	22
Gravel	1.0–56	0.05–468	0.01–0.48	0.01–2.73
Roofing felt	7.6	103.5	0.65	37.4
Cement tiles	355	53.5	0.425	38.6
Corrugated asbestos-cement sheet	10.5	22.6	0.14	24.6
Galvanized sheeting	26.4	43,500	1.23	37.6
Gravel	4.75	9,150	0.03	2.9
Ceramic tiles	71–304	10–48	0.07–0.4	13–41
Polyester	217–842	27–15	0.1–0.3	4.9–24
Gravel	18	9	0.11	2.7
Various roofing materials	14–240	582–12,357	0.2–4.5	76–2,458

Source: Quek and Forster,[50] Boller,[61] Mason et al.,[62] Zobrist et al.,[82] and Gromaire et al.[83]

TABLE 6 The Most Commonly Recorded Pesticides and Their Concentrations in Samples of Runoff Water from Roofs Depending on the Type of Roofing Material

Type of Roofing Material	Analyte (ng/L)
Roofing felt	Aldrin (6020), methoxychlor (430), heptachlor epoxide (19970), *p,p′* -DDD (520), *o, p′*-DDE (540), *o,p′* - DDD (340), *p, p′*-DDE (450), *o, p′*-DDT (190), *p, p′*-DDT (370), α-HCH (200), γ-HCH (400), propazine (1070), *tert*-butylazine (120)
Ceramic tiles	Aldrin (130), methoxychlor (430), heptachlor epoxide (430), *p,p*-DDD (4610), *o,p′*-DDE (2800), *o,p′*-DDD (1210), *p, p′* -DDE (1360), *p, p′* -DDT (1080), α-HCH (740), γ-HCH (600), propazine (1110), *tert*-butylazine (120), bromofos (2720)
Corrugated asbestos-cement sheeting	Aldrin (610), methoxychlor (130), heptachlor epoxide (14540), *p, p′* -DDD (500), *o, p′*-DDE (4640), *o, p′* - DDD (7040), *p, p′*-DDE (1040), *o, p′*-DDT (750), *p, p ′*-DDT (8460), α-HCH (460), γ-HCH (840), atrazine (1210), *tert*-butylazine (130), bromofos (310)
Metal sheet tiles	Heptachlor epoxide (1640), *o,p′* -DDE (920), *o,p′* -DDD (1890), *p,p′* -DDE (680), *o,p′* -DDT (490), *p,p′* - DDT (830), α-HCH (320), γ-HCH (500), propazine (400), bromofos (4912), atrazine (2980), simazine (620), malathion (1960), fenitrothion (630), chlorfenvinfos (180)
Galvanized sheeting	Aldrin (6020), methoxychlor (430), heptachlor epoxide (19970), *p,p*-DDD (520), *o,p′* -DDE (540), *o,p′* - DDD (340), *p,p′*-DDE (450), *o,p′*-DDT (190), *p,p′*-DDT (370), α-HCH (200), γ-HCH (400), propazine (1070), *tert*-butylazine (120)
Bituminous roofing felt	Heptachlor epoxide (1680), bromofos (50)
Teflon	Bromofos (40)

Source: Grynkiewicz et al.[80] and Polkowska.[87]

felt, galvanized sheeting, corrugated asbestos-cement sheeting), *p,p′*-DDT (corrugated asbestos-cement sheeting), *o,p′*-DDD (corrugated asbestos-cement sheeting), *p,p′*-DDD (ceramic tiles), *o,p′*- DDE (corrugated asbestos-cement sheeting, ceramic tiles), aldrin (roofing felt, galvanized sheeting), bromofos (metal sheet tiles, ceramic tiles), and atrazine (metal sheet tiles, corrugated asbestos-cement sheeting). The roofing materials from which runoff samples most often contained pesticides were corrugated asbestos-cement sheeting, ceramic tiles, and metal sheet tiles.[85–87]

Table 7 summarizes the concentrations of analytes determined in samples of roof runoff from buildings in Poland. The most frequently detected VOCC analytes were $CHCl_3$ and CQ14. The presence of VOCCs in roof runoff is probably due to the use of many of these compounds as solvents and degreasing agents.

The most frequently detected petroleum hydrocarbons were benzene and toluene. Their presence in roof runoff samples points to the possible leaching of these compounds from roofing materials like roofing felt, tar, and bituminous tiles as well as to the strong adsorption of these analytes to roof surfaces.[85]

TABLE 7 Total Concentrations of Selected Analytes Determined in Samples of Roof Runoff

Total Concentrations (μg/L)	
Organochlorine Compounds	Petroleum Hydrocarbons
Station no. 1	
0.01–0.17	0.11–35
Station no. 2 (crossroads)	
0.01–0.63	0.12–33
Station no. 3 (crossroads)	
0.03–1.07	0.61–64

TABLE 8 Levels of Cations and Anions in Samples of Runoff from Roofs Covered with Different Materials

Type of Roofing Material	Cations and Anions (meq/L)
Roofing felt	F^- (0.01), Cl^- (0.22), NO_3^- (0.08), SO_4^{2-} (0.44), Na^+ (0.14), NH_4^+ (0.21), K^+ (0.04), Mg^{2+} (0.07), Ca^{2+} (0.39)
Ceramic tiles	F^- (0.02), Cl^- (3.17), NO_3^- (1.93), SO_4^{2-} (4.59), Na^+ (1.13), NH_4^+ (0.13), K^+ (0.50), Mg^{2+} (1.30), Ca^{2+} (4.28)
Corrugated asbestos-cement sheeting	Cl^- (0.16), NO_3^- (0.12), SO_4^{2-} (1.45), Na^+ (0.20), NH_4^+ (0.14), K^+ (0.05), Mg^{2+} (0.07), Ca^{2+} (1.92)
Metal sheet tiles	F^- (0.01), Cl^- (0.20), NO_3^- (0.04), SO_4^{2-} (0.17), Na^+ (0.11), NH_4 (0.15), K^+ (0.05), Mg^{2+} (0.04), Ca^{2+} (0.27)
Galvanized sheeting	F^- (0.01), Cl^- (0.12), NO_3^- (0.09), SO_4^{2-} (0.92), Na^+ (0.11), NH_4^+ (0.10), K^+ (0.07), Mg^{2+} (0.07), Ca^{2+} (1.20)
Bituminous roofing felt	Cl^- (0.05), NO_3^- (0.02), SO_4^2 (0.32), Na^+ (0.08), NH_4^+ (0.04), K^+ (0.01), Mg^{2+} (0.06), Ca^{2+} (0.59)
Teflon	Cl^- (0.02), NO_3^- (0.03), SO_4^{2-} (0.04), Na^+ (0.04), NH_4^+ (0.04), K^+ (0.03), Mg^{2+} (0.02), Ca^{2+} (0.08)

Source: Polkowska.[86]

Roof runoff samples from Poland were also analyzed for the presence of formaldehyde. The highest level of this compound was 1.37 mg/L, while its mean concentration was 0.54 mg/L and was comparable to the mean value recorded in rainwater samples.[78,79]

Table 8 lists the most frequently detected and determined ions (and their concentrations) in samples of runoff from roofs made from different materials.

Chloride and sulfate were found in all the samples analyzed. F^- ions were detected in only 4 of 68 samples analyzed. The highest concentrations of both chlorides and sulfates were found in runoff from ceramic tiled roofs. As far as cations are concerned, Na^+, K^+, Mg^{2+}, and Ca^{2+} were determined in all samples, the highest levels being registered for Na^+ and Ca^{2+}. The highest concentration of sodium ions was recorded in runoff from roofs covered with roofing felt and ceramic tiles, whereas the highest calcium levels were found in samples from roofs covered with tiles and corrugated asbestos-cement sheeting.[85,86]

Throughfall

In forest geoecosystems, precipitation does not reach the substrate directly, but as it trickles through the crowns of trees and down their trunks, it undergoes chemical changes. The chemical composition of rainwater in forest areas changes mainly under the influence of the physiological processes of plants and the dissolution and flushing from branches, bark, and leaves of aerosols and gaseous contaminants absorbed from the air. The species composition of the trees, their age, and health can also alter the properties of rainwater, giving it new physicochemical properties.

The runoff (throughfall) from tree trunks is less interesting than that from tree crowns as it makes up at most a few percent of the rainwater. In addition, trunk runoff is much less in coniferous trees, the branches of which point downwards, taking the rainwater beyond the trunk itself.[3] The proportion of throughfall in the total deposition transported to the forest floor is less than 10%. It has been found that in pine forests, 3% of precipitation reaches the ground as trunk runoff, but 83% as crown runoff.[8] The variety and numbers of different contaminants present in crown runoff are affected primarily by the type of leaves, the density of the crown, the age of the tree, the quantity of precipitation, and the wind direction.[5]

According to the results published in the literature regarding the content of cations and anions in throughfall, the concentrations of particular ions in both crown runoff and trunk runoff are considerably higher than in precipitation falling on open terrain.[88] The elevated levels of potassium, magnesium, calcium, and manganese ions in crown runoff compared to their content in rainwater sampled in open country were due to their being flushed off leaves and needles. This was particularly noticeable in the case of spruce needles, which have a much greater surface area than the leaves/needles of other tree species.

A Japanese study has shown that a high nitrogen load in atmospheric deposition (including dry deposition) may be a factor contributing to the deterioration of pine stands. This was particularly conspicuous in urban areas, where levels of nitrate and ammonium ions in runoff were high.[8,89]

Comparative studies were carried out in Japan and Spain to illustrate the effect of the location of a woodland system on the composition and numbers of contaminants entering the environment. Samples were collected at two sites: in a forest close to a road, traffic, and urban areas (exposed), and in a wood situated far from urban areas (sheltered by mountains), where the effect of pollutants is less. The respective concentrations of nitrate and sulfate ions in trunk and crown runoff were far higher in the urban area (34.4 and 50 μeq/L) than in the mountains beyond the city (4.8 and 28.8 μeq/L) and were several times higher than the levels recorded in free precipitation.[8,12] Table 9 lists the range of concentrations and the mean concentrations of contaminants (including metals from the platinum group) in samples of runoff from tree crowns.

The data on formaldehyde concentrations in throughfall have been published in paper.[79]

Runoff from Farming Areas

Good conditions for farmland runoff containing substantial loads of nutrients exist in spring. At this time of the year, the plant cover is not yet dense enough to prevent surface runoff, and the ground, still frozen after the winter, also facilitates this. Most pesticides and fertilizers are applied directly to the soil or sprayed onto arable land, plantations, and forests; they therefore gain direct entry into the environment and are the source of much contamination (aerosols containing pesticides and nitrogen and phosphorus compounds). The inappropriate storage of pesticides and mineral fertilizers (directly on the ground, without any proper protection from the elements) contributes to the leaching of active substances, which, through the mediation of surface runoff, enter surface waters, causing their eutrophication. This leads to deleterious changes in aquatic ecosystems manifested by an increase in nutrient

TABLE 9 Range of Concentrations, Mean Concentrations and Frequency of Occurrence of Analytes in Throughfall Samples

Parameter	Concentration Range (mean) [μeq/dm³]	Parameter	Concentration Range (mean) [peq/dm³]
pH	4.2–7.4 (5.6)	Fe	0.030–0.55 (0.19)
Conductivity (μS/cm)	0.020–0.050 (0.15)	Total phenols	0.20–1.6 (0.49)
Cl^-	0.050–3.5 (0.62)	TOC	22–108 (45)
F^-	0.0010–0.11 (0.028)	Pt	$(0.21–0.6) \times 10^{-7}$ (0.4×10^{-7})
Br^-	–	Pd	$(1.1–2.6) \times 10^{-6}$ (1.7×10^{-6})
PO_4^{3-}	0.0070–1.0 (0.15)	Pb	0.035–0.053 (0.046)
SO_4^{2-}	0.098–1.0 (0.39)	Zn	0.0014–0.00409 (0.0022)
NO_3^-	0.0040–0.78 (0.20)	Cd	0.0039–0.014 (0.0068)
Li^+	0.00027–0.00040 (0.00034)	Co	0.037–0.058 (0.0508)
Na^+	0.24–1.77 (0.64)	Mn	–
NH_4^+	0.39–0.43 (0.35)	V	0.039–0.26 (0.14)
K^+	0.21–0.14 (0.13)	Cr	0.46–1.4 (1.1)
Mg^{2+}	0.080–0.13 (0.10)	Ni	0.105–0.35 (0.21)
Ca^{2+}	0.080–0.32 (0.19)	Sn	0.0067–0.0101 (0.0084)
Formaldehyde	0.060–0.51 (0.18)	Mo	0.0081–0.015 (0.012)

levels (nitrogen and phosphorus compounds), a drop in the level of dissolved oxygen, and mass blooms of algae. These contaminants can also permeate into the shallow groundwater circulation, which serves as a source of drinking water in rural areas. Studies carried out worldwide confirm that farmland runoff carries very large amounts of nitrogen and phosphorus compounds; Cl^-, K^+, Ca^{2+}, Na^+, and Mg^{2+} ions; and pesticides.[32,90–92]

Runoff from Airports

Air transport is the most modern and the fastest developing type of transport. In 1993–2000, the number of passengers traveling by air in European Union countries increased by ca. 10% each year. Forecasts for the next 20 years put the annual increase in the global number of air passengers at around 5%.

This burgeoning development of air transport also concerns Poland. To a large extent, this is the result of the rapid expansion of the so-called budget airlines onto the Polish market and the reconstruction of many regional airports. Their operation is an unavoidable interference in the natural environment. Air transport contributes to the destruction of the ozone layer, and the harmful effects of air transport are visible on the ground as well. Precipitation runoff from airports affects surface water and groundwater to a substantial extent. Along with rainwater and snowmelt, petroleum derivatives, surfactants, de-icing agents (in winter), and other contaminants get into the environment. However, it is the de-icing agents applied when air temperatures are low that pose the greatest danger of polluting waters running off airport aprons and runways. The composition of airport runoff changes during the year and depends on a number of factors, such as the weather, the quantities of de-icing agents applied, the duration of precipitation, and the intensity of traffic at the airport in question.

Airport authorities use not only the environmentally dangerous ethylene and propylene glycols for de-icing aircraft but also the environmentally rather more friendly mixtures of acetates and formates for de-icing runways, taxiing areas, and aprons.[93,94] Airports do not usually have their own sewage treatment plants, and all the effluents, carrying petroleum derivatives, surfactants, de-icing agents, and other organic and inorganic pollutants, flow with rainwater and snowmelt into the surrounding soil and surface waters. These contaminants may significantly affect the fauna and flora of rivers, causing the death of fish and aquatic vegetation; therefore, it is crucial to monitor the runoff from such areas on a continuous basis.[95–97]

Figures 5–7 show the mean concentrations of selected analytes in samples of runoff collected at airports in Poland (Gdańsk and Warsaw) and the United States.[97,98]

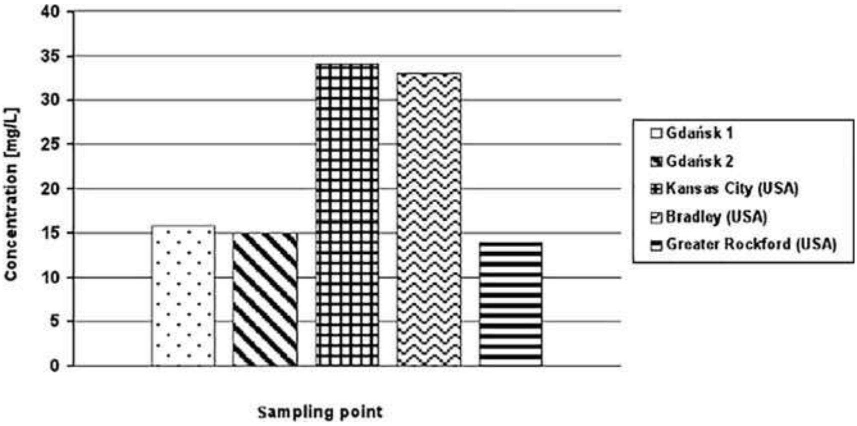


FIGURE 5 Comparison of mean concentrations of Ca in samples of runoff from airports in Gdansk and the United States.

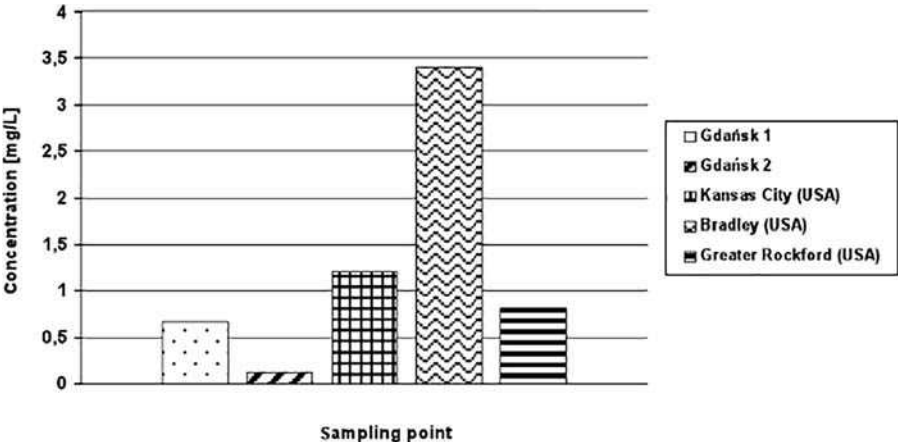


FIGURE 6 Comparison of mean concentrations of Fe in samples of runoff from airports in Gdansk and the United States.

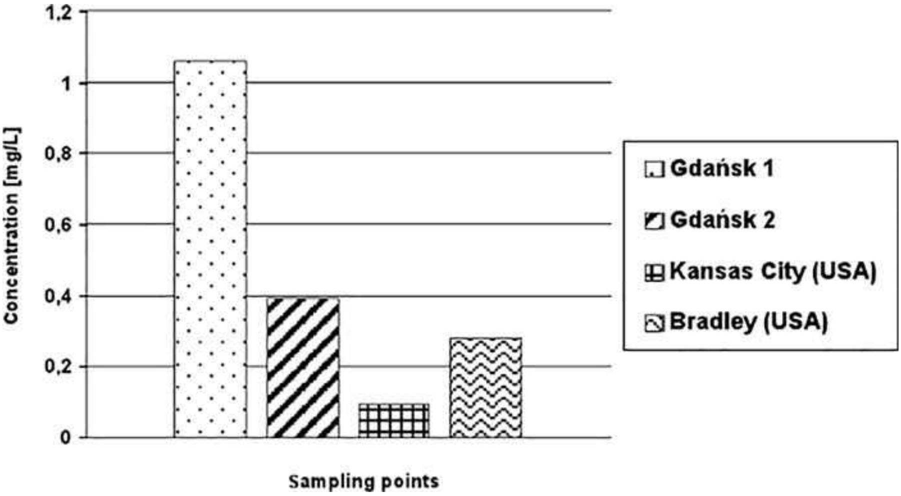


FIGURE 7 Comparison of mean concentrations of sum of phenols in samples of runoff from airports in Gdansk and the United States.

The highest levels of potassium and sodium ions were recorded in runoff from the Gdansk airport in 2009—this may have been due to the large quantities of de-icing agents (sodium and potassium formate) used during the winter. At the other airports, the mean concentrations of these analytes were very much lower. The mean concentrations of calcium ions determined in the runoff from the Gdansk airport were lower than those from American airports.

The highest level of total iron by far was found in samples from the Bradley International Airport (United States); the level of this particular contaminant at Gdansk was lower than those at U.S. airports. Comparison of the mean levels of magnesium ions shows that they have been at similar levels at all the airports, although the Mg level was lowest at Gdańsk.

The highest mean concentration of total phenols was recorded in samples taken in winter at the Gdansk airport, whereas mean Total organic carbon (TOC) levels at this time were comparable in the samples from the airports at Gdansk, Kansas City, and Bradley (United States).

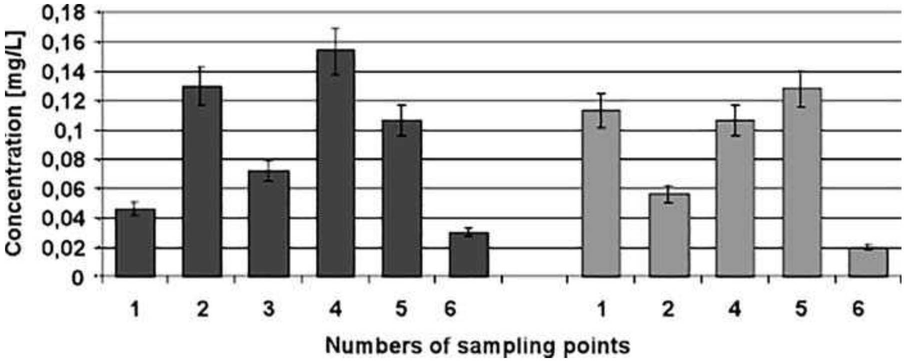


FIGURE 8 Comparison between mean concentrations of organic compounds determined in precipitation and runoff samples; 1–5, runoff samples; 6-precipitation. (March 19, 2008—black; April 8, 2008—gray).

The following organic compounds have been shown to be present in runoff from the Gdansk airport: anionic and cationic detergents, formaldehyde, phenols, PAHs, pesticides, and polychlorinated biphenyls (PCBs). Figure 8 shows the mean concentrations of cationic detergents contained in samples of precipitation and runoff.

Changes in the Composition of Samples Depending on the Moment of Sampling during Precipitation or Runoff

Literature reports indicate that atmospheric contaminants are removed during the early phases of precipitation, usually in the "first flush," and their concentrations in the precipitation drop exponentially as its volume increases. In its simplest form, this relation can be expressed by the equation

$$c = a \cdot \exp(b \cdot V) \tag{1}$$

where *c* is the concentration of contaminant in the precipitation (mol/L), *V* is the volume of precipitation collected (L), and *a* and *b* are constants.[99]

Analysis of the results of studies carried out in Germany and Switzerland shows that concentrations of contaminants (heavy metals, pesticides, cations, anions, PAHs) are the highest in the first phase of precipitation and in the first flow of runoff.[1,56,62,87] Table 10 lists the relevant data.

From these data, it can be inferred that the presence of pesticides and inorganic compounds in the first flush of precipitation and in roof runoff may be influenced by the following: the flushing of dry deposition accumulated in the period preceding rainfall, the leaching of compounds from roofing and guttering materials and the corrosion of these materials, and the reduction in concentration of constituents with increasing amounts of precipitation, as a result of the removal by raindrops of aerosol particles and gases contained in the air.[82]

When rainfall is intense, the period of contact between the rainwater and the roofing material is short and just the contaminants accumulated on its surface are washed away; hence, in this case, the type of roofing material is practically irrelevant. However, when rainfall is less intense, the type of roofing material takes on greater significance, because contaminants are adsorbed on roofs with a rough surface like roofing felt or cement tiles, which is why the first flush of rainwater and the initial runoff do not then contain larger amounts of contaminants. These are only flushed away after a long period of rain.[54]

As far as runoff from tree crowns is concerned, there is a drop in the concentration of ions in precipitation and runoff samples with increasing intensity of rainfall.[5]

TABLE 10 The Values of Some Physicochemical Parameters and the Contents of Selected Constituents in Samples of Precipitation and Runoff

		Runoff	
Parameter (mM)	Precipitation	Initial Phase	After the Initial Phase
pH	6.9	7.0–7.6	7.0–7.6
Ca^{2+}	0.06	0.4–1.3	0.1–0.5
Mg^{2+}	0.01	0.06–0.44	0.01–0.04
Na^+	0.02	0.22–0.56	0.02–0.20
K^+	0.01	0.10–0.35	0.02–0.10
NH_4^+	0.14	0.14–0.29	0.02–0.07
Alkalinity	0.08	0.7–2.8	0.3–0.6
Cl^-	0.01	0.06–0.22	0.02–0.06
SO_4^{2-}	0.04	0.1–0.2	0.01–0.1
NO_3^-	0.06	0.29–0.64	0.02–0.20
N_{total}	–	0.5–1.2	0.07–0.30

Source: Mason et al.[62]

Summary

The atmosphere is one of the main non-point sources of pollution, and its dynamics causes a diversity of the ways that it can affect the level of contamination of soils and surface water. The composition of precipitation is closely related to a degree of pollution of air. After passing through developed areas (buildings, roads, woodland, farmland, waste dumps), wet deposition enters the environment almost exclusively as runoff.

The composition of runoff water depends to a large extent on the kind of terrain on which precipitation fell. The road runoff can contain hydrocarbons of petrochemical origin, dust, and heavy metals. The level of contamination of roof runoff depends on the roof material, which can be a source of many organic and inorganic pollutants, such as pesticides and heavy metals (copper, zinc). The dump runoff is characterized by variable aggressiveness towards the environment, dependent on the degree of processing of the refuse while the edge-of-field runoff can contain pesticides and biogenic substances. A large fraction of runoff water with the pollutants present in it enters surface water and groundwater, which can be used as a source of drinking water. Although the flux of pollutants into soil and groundwater is usually small, hazardous chemicals can undergo accumulation over a period of time.

References

1. Bucheli, T.D.; Muller, S.R.; Heberle, S.; Schwarzenbach, R.P. Occurence and behavior of pesticides in rainwater, roof runoff, and artifical stormwater infiltration. Environ. Sci. Technol. **1998**, *32,* 3457–3464.
2. Wauchope, R.D.; Graney, R.L.; Cryer, S.; Eadsforth, C.; Klein, A.W.; Racke, K.D.; Pesticide runoff: methods and interpretation of field studies. Pure Appl. Chem. **1995**, *67,* 2089–2108.
3. Yates, D.J.; Hutley, L.B.; Doley, D.; Boonsaner, A. The ecological and physiological significance of occult precipitation in an Australian rainforest, 1st Int. Conf. on Fog and Fog Collection, July 19–24. Vancouver, Canada; Schemenauer, R.S., Bridgman, H., Eds.; 1998, 81–84.
4. Balestrini, R.; Galli, L.; Tagliaferi, A.; Tartari, G. Study on throughfall deposition in two North Italian forest sites (Valtellina, Lombardy). Chemosphere **1998**, *36,* 1095–1100.

5. Walna, B.; Siepak, J. Research on the variability of physicochemical parameters characterizing acid precipitation at the Jeziory Ecological Station in The Wielkopol– ski National Park (Poland). Sci. Total Environ. **1999**, *239,* 173–187.
6. Lovett, G.M.; Traynor, M.M.; Pouyat, R.V.; Carreiro, M.M.; Zhu, W.-X.; Baxter, J.W. Atmospheric deposition to oak forests along an urbanrural gradient. Environ. Sci. Technol. **2000**, *34,* 4294–4300.
7. Forti, M.C.; Moreira-Nordemann, L.M. Rainwater and throughfall chemistry in a Terra Firme Rain Forest; Central Amazonia. J. Geophys. Res. **1991**, *96,* 7415–7421.
8. Masaaki, C.; Do Hoon, K.; Hiroshi, S. Rainfall, stemflow, and throughfall chemistry at urban- and moutain-facing sites at MT. Gokurakuj, Hiroshima, Western Japan. Water Air Soil Pollut. **2003**, *146,* 93–109.
9. Hansen, K.; Draaijers, G.P.J.; Ivens, W.P.M.F.; Gundersen, P.; van Leeumen, N.F.M. Concentration variations in rain and canopy throughfall collected sequentially during individual rain events. Atmos. Environ. **1994**, *28,* 3195–3205.
10. Houle, D.; Ouimet, R.; Paquin, R.; Laflamme, J-G. Determination of sample size for estimating ion throughfall deposition under a mixed hardwood forest at the Lake Clair Watershed (Duchesnay, Quebec). Can. J. For. Res. **1999**, *29,* 1935–1934.
11. Avila, A.; Rodrigo, A. Trace metal fluxes in bulk deposition, throughfall and stemflow at two evergreen oak stands in NE Spain subject to different exposure to the industrial environment. Atmos. Environ. **2004**, *38*, 171–180.
12. Rodrigo, A.; Avila, A.; Roda, S. The chemistry of precipitation, throughfall and stemflow in two holm oak (*Querces ilex* L.) forests under a contrasted pollution environment in NE Spain. Sci. Total Environ. **2003**, *305,* 195–205.
13. Inagaki, M.; Sakai, M.; Ohnuki, Y. The effects of organic carbon on acid rain in a temperature forest in Japan. Water Air Soil Pollut. **1995**, *85,* 2345–2350.
14. Bini, C.; Bresolin, F. Soil acidification by acid rain in forest ecosystems: A case study in northern Italy. Sci. Total Environ. **1998**, *222*, 1–15.
15. Prakasa Rao, P.S.; Momin, G.A; Safai, P.D.; Pillai, A.G.; Khemani, L.T. Rain water and throughfall chemistry in the Silent Valley in South India. Atmos. Environ. **1995**, *29,* 2025–2029.
16. Lovett, G.M.; Kinsman, J. Atmospheric deposition to high- elevation systems. Atmos. Environ. **1990**, *24,* 2767–2786.
17. Kiss, G.; Varga-Puchony, Z.; Hlavy, J. Determination of polycyclic aromatic hydrocarbons in precipitation using solid-phase extraction and column liquid chromatography. J. Chromatogr. A **1996**, *725*, 261–272.
18. Garban, B.; Blanchou, H.; Motelay-Massei, A.; Chevreuil, M.; Ollivon, D. Atmospheric bulk deposition of PAHs onto France: Trends from urban to remote sites. Atmos. Environ. **2002**, *36,* 5395–5403.
19. Bomboi, M.T.; Hernandez, A. Hydrocarbons in urban runoff: Their contribution to the wastewaters. Water Res. **1991**, *25*, 557–565.
20. Stephenson, J.B.; Zhou, W.F.; Beck, B.F.; Green, T.S. Highway stormwater runoff in karst areas—Preliminary results of baseline monitoring and design of a treatment system for a sinkhole in Knoxville, Tennessee. Eng. Geol. **1999**, *52,* 51–59.
21. Lygren, E.; Gjessing, E.; Berglind L. Pollution transport from a highway. Sci. Total Environ. **1990**, *33*, 147–159.
22. Stotz, G. Investigations of the properties of the surface water run-off from federal highways in the FRG. Sci. Total Environ. **1987**, *59,* 329–337.
23. MacKenzie, M.J.; Hunter, J.V. Sources and fates of aromatic compounds in urban stormwater runoff. Environ. Sci. Technol. **1979**, *13,* 179–183.
24. Moilleron, R.; Gonzalez, A.; Chebbo, G.; Thevenot, D.R. Determination of aliphatic hydrocarbons in urban runoff samples from the "Le Marais" experimental catchment in Paris centre. Water Res. **2002**, *36,* 1275–1285.

25. Achten, Ch.; Kolb, A.; Puttmann, W. Methyl *tert-butyl* ether (MTBE) in urban and rural precipitation in Germany. Atmos. Environ. **2001**, *35*, 6337–6345.
26. Lopes, T.J.; Fallon, J.D.; Rutherford, D.W.; Hiatt M.H. Volatile organic compounds in storm water from a parking lot. J. Environ. Eng. **2000**, *126,* 1137–1143.
27. Borden, R.C.; Black, D.C; McBlief, K.V. MTBE and aromatic hydrocarbons in North Carolina stormwater. Environ. Pollut. **2002**, *118,* 141–152.
28. Thurston, K.A. Lead and petroleum hydrocarbon changes in an urban wetland receiving stormwater runoff. Ecol. Eng. **1999**, *12*, 387–399.
29. Legret, M.; Pagotto, C. Evaluation of pollutant loadings in the runoff waters from a major rural highway. Sci. Total Environ. **1999**, *235,* 143–150.
30. Pagotto, C.; Legreti, M.; Cloirec, P.L. Comparision of the hydraulic behaviour and the quality of highway runoff water according to the type of pavement. Water Res. **2000**, *34,* 4446–4454.
31. Hvitved-Jacobsen, T.; Johansen, N.B.; Yousef, Y.A. Treatment systems for urban and highway runoff in Denmark. Sci. Total Environ. **1994**, *146/147,* 499–506.
32. Tsiouris, S.E.; Mamolos, A.P.; Kalburtji, K.L.; Alifrangis, D. The quality of runoff water collected from a wheat field margin in Greece. Agric. Ecosyst. Environ. **2002**, *89,* 117125.
33. Backstrom, M.; Nilsson, U.; Hakansson, K.; Allard, B.; Karlsson, S. Specification of heavy metals in road runoff and roadside total deposition. Water Air Soil Pollut. **2003**, *147,* 343–366.
34. Matthes, S.A.; Cramer, S.D.; Covino, B.S.; Bullard, S.J.; Holcomb, G.R. Precipitation runoff from lead, outdoor and indoor atmospheric corrosion. ASTM Spec. Tech. Publ. **2002**, 1421.
35. Maltby, L.; Boxall, B.A.; Forrow, D.M.; Calow P.; Betton, C.I. The effects of motorway runoff on freshwater ecosystems: 1. Field study. Environ. Toxicol. Chem. **1995**, *14,* 1079–1092.
36. Harrison, R.M.; Wilson, S.J. The chemical composition of highway drainage waters. I. Major ions and selected trace metals. Sci. Total Environ. **1985**, *43,* 63–77.
37. Hewitt, C.N.; Rashed, M.B. Removal rates of selected pollutants in the runoff waters from a major rural highway. Water Res. **1992**, *26*, 311–319.
38. Latimer, J.S.; Hoffman E.J.; Hoffman, G.; Fasching, J.L.; Quinn, J.G. Sources of petroleum hydrocarbons in urban runoff. Water Air Soil Pollut. **1990**, *52*, 1–21.
39. Turer, D.; Maynard, J.B.; Sansalone, J.J. Heavy metal contamination in soils of urban highways: Comparison between runoff and soil concentrations at Cincinnati, Ohio. Water Air Soil Pollut. **2001**, *132,* 293–314.
40. Singh, S.P.; Tack, F.M.; Gabriels, D.; Verloo, G. Heavy metal transport from dredged sediment derived surface soils in a labolatory rainfall simulation experiment. Water Air Soil Pollut. **2000**, *118,* 73–86.
41. Mikkelsen, P.S.; Häfliger, M.; Ochs, M.; Tjell, J.C.; Jacobsen, P.; Boller M. Experimental assessment of soil and groundwater contamination from two old infiltration systems for road run–off in Switzerland. Sci. Total Environ. **1996**, *189/190,* 341–347.
42. Dannecker, W.; Au, M.; Stechmann, H. Substance load in rainwater runoff from different streets in Hamburg. Sci. Total Environ. **1996**, *93,* 385–392.
43. Gromaire-Mertz, M.C.; Garnaud, S.; Gozalez, A.; Chebbo, G. Characterisation of urban runoff pollution in Paris. Water Sci. Technol. **1999**, *39,* 1–8.
44. Legret, M.; Colandini, V.; Le Marc, C. Effect of a porous pavement with reservoir structure of the quality of runoff water and soil. Sci. Total Environ. **1996**, *189/190,* 335–340.
45. Fernández Espinosa, A.J.; Rodriguez, M.T.; Barragán de la Rosa, F.J.; Jiménez Sánchez, J.C. A chemical speciation of trace metals for fine urban particles. Atmos. Environ. **2002**, *36*, 773–780.
46. Smith, M.E.; Kaster, J.L. Effect of rural highway runoff on stream benthic macroinvertebrates. Environ. Pollut. **1983**, *32,* 157–170.
47. Ward, N.I.; Reeves, R.D.; Brooks, R.R. Lead in soil and vegetation along a New Zealand state highway with low traffic volume. Environ. Pollut. **1975**, *9*, 243–251.

48. Bomboi, M.T.; Hernandez A. Hydrocarbons in urban runoff: Their contribution to the wastewaters. Water Res. **1991**, *5*, 557–565.
49. Yaziz, M.I.; Gunting, H.; Sapari, N.; Ghazali, A.W. Variations in rainwater quality from roof catchments. Water Res. **1989**, *23,* 761–765.
50. Quek, U.; Förster, J. Trace metals in roof runoff. Water Air Soil Pollut. **1993**, *68,* 373–389.
51. Bucheli, T.D.; Müller, S.R.; Voegelin A. Bituminous roof sealing membranes as major sources of the herbicide (*R*,*S*)- mecoprop in roof runoff waters: Potential contamination of groundwater and surface water. Environ. Sci. Technol. **1998**, *32,* 3465–3471.
52. Shu, P.V.; Hirner, A. Trace compounds in urban rain and roof runoff. J. High Resol. Chromatogr. **1998**, *21,* 65–68.
53. Simmons, G.; Hope, V.; Lewis, G.; Whitmore, J.; Gao, W. Contamination of potable roof-collected rainwater in Auckland in New Zealand. Water Res. **2001**, *35*, 1518–1524.
54. Kennedy, P.; Gadd, J.; Mitchell, K. Preliminary examination of the nature of urban roof runoff in New Zealand; Report, Ministry of Transport Te Manatu Waka, New Zealand, 2001.
55. Garnaud S.; Mouchel J.M.; Chebbo G.; Thevenot D.R. Heavy metal concentrations in dry and wet deposits in Paris district: Comparison with urban runoff. Sci. Total Environ. **1999**, *235,* 235–245.
56. Foster, J. Patterns of roof runoff contamination and their potential implications on practice and regulation of treatment and local infiltration. Water Sci. Technol. **1996**, *33*, 39–48.
57. Zorbist, J.; Müller, S.R.; Bucheli, T.D. Quality of roof runoff for groundwater infiltration. Water Res. **2000**, *34,* 14551462.
58. Foster, J. The influence of location and season on the concentrations of macroions and organic trace pollutants in roof runoff. Water Sci. Technol. **1998**, *38,* 83–90.
59. Davis, A.P.; Shokouhian M.; Ni, S. Loading estimates of lead, copper, cadmium and zinc in urban runoff from specific sources. Chemosphere **2001**, *44*, 997–1009.
60. Bertling, S.; Odnevall Wallinder, I.; Leygraf, C.; Berggren, D. Environmental effects of zinc runoff from roofing ma– terials—A new multidisciplinary approach. In *Outdoor Atmospheric Corrosion;* Townstend, H.E., Ed.; ASTM STP 1421; American Society for Testing and Materials International: West Conshohocken, PA, 2002.
61. Boller, M. Tracking heavy metals reveals sustainability deficits of urban drainage systems. Water Sci. Technol. **1997**, *35*, 77–87.
62. Mason, Y.; Ammann, A.A.; Ulrich, A.; Sigg, L. Behavior of heavy metals, nutrients and major components during roof runoff infiltration. Environ. Sci. Technol. **1998**, *33*, 1588–1597.
63. Heijerick, D.G.; Janssen, C.R.; Karlen, C.; Odnevall Wall- inder, I.; Leygraf, C. Bioavailability of zinc in runoff water from roofing materials. Chemosphere **2002**, *47,* 1073–1080.
64. He, W.; Odnevall Wallinder, I.; Leygraf, C. A laboratory study of copper and zinc runoff during first flush and steady- state conditions. Corros. Sci. **2001**, *43,* 127–146.
65. Heijerick, D.G.; Janssen, C.R.; Karlen, C.; Odnevall Wallinder, I.; Leygraf, C. Runoff rates and ecotoxicity of zinc induced by atmospheric corrosion. Sci. Total Environ. **2001**, *277,* 169–180.
66. Odnevall Wallinder, I.; Korpinen, T.; Soundberg, R.; Leygraf, C. Atmospheric corrosion of naturally and pre- patinated copper roofs in Singapore and Stockholm—Run- off rates and corrosion product formation. In *Outdoor Atmospheric Corrosion;* Townstend, H.E., Ed.; ASTM STP 1421; American Society for Testing and Materials International: West Conshohocken, PA, 2002.
67. Odnevall Wallinder, I.; Verbiest, P.; Leygraf, C.; He, W. Effects of exposure direction and inclination on the runoff rates of zinc and copper roofs. Corros. Sci. **2000**, *42,* 1471–1487.
68. Davis, A.P.; Burns, M. Evaluation of lead concentration in runoff from painted structures. Water Res. **1999**, *33*, 2949–2955.
69. Horstmann, M.; McLachlan, M. Concentration of polychlorinated dibenzo-p-dioxins (PCDD) and dibenzofurans (PCDF) in urban runoff and household wastewaters. Chemosphere **1995**, *31,* 2887–2896.

70. Thomas, P.; Greene, G. Rainwater quality from different roof catchments. Water Sci. Technol. **1993**, *28,* 291–299.
71. Hvitved-Jacobsen, T.; Johansen, N. B.; Yousef, Y.A. Treatment systems for urban and highway runoff in Denmark. Sci. Total Environ. **1994**, *146/147,* 499–506.
72. Hewitt, C.N.; Rashed, M.B. An integrated budget for selected pollutants for a major rural highway. Sci. Total Environ. **1990**, *93,* 375–384.
73. Grynkiewicz, M.; Polkowska, Z.; Namieśnik, J. Determination of polycyclic aromatic hydrocarbons in bulk precipitation and runoff waters in an urban region (Poland). Atmos. Environ. **2002**, *36,* 361–369.
74. Polkowska, Z.; Zabiegala, B.; Górecki, T.; Namieśnik, J. Contamination of runoff waters from roads with high traffic intensity in the urban region of Gdansk, Poland. Pol. J. Environ. Stud. **2005**, *6,* 799–807.
75. Grynkiewicz, M.; Polkowska, Z.; Namieśnik, J. Determination of polycyclic aromatic hydrocarbons in bulk precipitation and runoff waters in an urban region (Poland). Atmos Environ. **2002**, *36,* 361–369.
76. Klimaszewska, K.; Polkowska, Z.; Namieśnik, J. The influence of mobile sources on acidity of runoff waters from roads with high traffic intensity of Gdansk. Pol. J. Environ. Stud. **2007**, *16,* 883–891.
77. Polkowska, Z.; Grynkiewicz, M.; Zabiegala, B.; Namieśnik, J. Levels of pollutants in runoff water from roads with high traffic intensity in the city of Gdansk, Poland. Pol. J. Environ. Stud. **2001**, *10,* 351–363.
78. Duncan, L.C. Chemistry of rime and snow collected at a site of central Washington Cascades. Environ. Sci. Technol. **1992**, *26*, 61–66.
79. Polkowska, Z.; Skarżyńska, K.; Górecki, T.; Namieśnik, J. Formaldehyde in various forms of atmospheric precipition and deposition from highly urbanized regions. J. Atmos. Chem. **2006**, *53*, 211–236.
80. Grynkiewicz, M.; Polkowska, Z.; Kot-Wasik, A.; Namieśnik, J. Determination of phenols in runoff. Pol. J. Environ. Stud. **2002**, *11*, 85–89.
81. Förster, J. Patterns of roof runoff contamination and their potential implications on practice and local infiltration. Water Sci. Technol. **1996**, *33*, 39–48.
82. Zobrist, J.; Muller, S.R.; Ammann, A.; Bucheli, T.D.; Mottier, V.; Ochs, M.; Schoenenberger, R.; Eugster, J.; Boller, M. Quality of roof runoff for groundwater infiltration. Water Res. **2000**, *34,* 1455–1462.
83. Gromaire, M.C.; Garnaud, S.; Saad, M.; Chebbo, G. Contribution of different sources to the pollution of wet weather flows in combined sewers. Water Res. **2001**, *25,* 521–533.
84. Bucheli, T.D.; Grüebler, F.C.; Müller, S.R.; Schwarzenbach, R.P. Simultaneous determination of neutral and acidic pesticides in natural waters at the low nanogram per liter level. Anal. Chem. **1997**, *69,* 1569–1576.
85. Polkowska, Z.; Górecki, T.; Namieśnik, J. Quality of roof runoff waters from an urban region (Gdansk, Poland). Chemosphere **2002**, *49,* 1275–1283.
86. Polkowska, Z. Examining the effect of the type of roofing on pollutant content in roof runoff waters from buildings in selected districts of the city of Gdańsk. Pol. J. Environ. Stud. **2004**, *13,* 191–201.
87. Polkowska, Z.; Górecki, T.; Namieśnik, J. Occurrence of pesticides in rain and roof runoff waters from an urban region. Urban Water **2009**, *6*, 441–448.
88. Walna, B.; Siepak, J. Research on the variability of physicochemical parameters characterising acid precipitation at the Jeziory Ecological Station in the Wielkopolski National Park (Poland). Sci. Total Environ. **1999**, *239,* 173–187.
89. Rennenberg, H.; Gessler A. Consequences of N deposition to forest ecosystems—Recent results and future research needs. Water Air Soil Pollut. **1999**, *116,* 47–64.

90. Barden, Ch.J.; Geyer, W.A.; Mankin, K.R.; Ngandu, D.; Devlin, D.L.; McVay, K. Reducing runoff contaminants with riparian buffer strips. AFTA Conf. Proc. 2003, 7–14.
91. Turtola, E.; Yli-Halla, M. Fate of phosphorus applied in slurry and mineral fertilizer: Accumulation in soil and release into surface runoff water. Nutr. Cycling Agroecosyst. **1999**, *55*, 165–174.
92. Schulz, R.; Hauschild, M.; Ebeling, M.; Nanko-Drees, J.; Wogram, J.; Liess, M. A qualitative field method for monitoring pesticides in the edge-of field runoff. Chemosphere **1998**, *36,* 3071–3082.
93. Switzenbaum, M.S.; Veltman, S.; Mericas, D.; Wagoner, B.; Schoenberg, T. Best management practices for airport deicing stormwater. Chemosphere **2001**, *43,* 1051–1062.
94. Breedveld, G.D.; Roseth, R.; Sparrevik, M.; Hartnik, T.; Hem, L.J. Persistence of the de-icing additive benzotriazole at an abandoned airport. Water Air Soil Pollut. **2003**, *3*, 91–103.
95. Revitt, D.M.; Shutes, R.B.E.; Llewellyn, N.R.; Worrall, P. Experimental reedbed systems for the treatment of airport runoff. Water Sci. Technol. **1997**, *36,* 385–390.
96. Cancilla, D.A.; Baird, J.C.; Rosa, R. Detection of aircraft deicing additives in groundwater and soil samples from fairchild air force base, a small to moderate user of deicing fluids. Environ. Contam. Toxicol. **2003**, *70,* 868–875.
97. Krzemieniowski, M.; Bialowiec, A.; Zieliński, M. The effectiveness of the storm water treatment plant at Warsaw Frederick Chopin Airport. Arch. Environ. Prot. **2006**, *32,* 25–33.
98. United States Environmental Protection Agency (Amerykańska Agencja Ochrony Środowiska), Airport Deicing Operations, EPA-821-R-00–016, Waszyngton, 2000.
99. Leister, D.L.; Baker, J.E. Atmospheric deposition of organic contaminants to the Chesapeake Bay, Atmos. Environ. **1994**, *28,* 1499–1520.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

22

Salt Marsh Resilience and Vulnerability to Sea-Level Rise and Other Environmental Impacts

Daria Nikitina

Introduction

Salt marshes are coastal wetlands within the intertidal zone, characterized by highly saline soils and anoxic conditions associated with tidal flooding and vegetated by macrophytes. Salt marshes occur globally along the low-energy depositional coasts of temporal and sub-Arctic climates both in microtidal and macrotidal regimes (Allen and Pye, 1992). Commonly, salt marshes occupy broad flat areas often referred to as the marsh platform. At the landform scale, salt marshes classify based on the physical settings that include open-coast marshes, poorly developed and vulnerable to wave action (Figure 1a); back-barrier marshes that formed on the sheltered side of barrier islands or spits (Figure 1b and f); estuarine marshes that fringe estuaries and coastal lagoons where muddy sediments accumulate (Figure 1e); embayment marshes that fringe the edge of open or restricted-entrance tidal embayment (Figure 1c and d); ria/loch-head marsh, occurring mostly in Europe (Brittany, Ireland and Scotland) (Figure 1g). Salt marshes can also develop where a gently shelving coast is combined with a high concentration of suspended sediment, such as the chenier plain (Rogers and Woodroffe, 2014).

Salt Marsh Ecosystem Characteristics

While there are several types of salt marshes, in the temporal climates, salt marshes are commonly divided into low marsh vegetated by salt-tolerant *Spartina alterniflora*, high marsh dominated by *Spartina patens*, *Distichlis spicata*, *S. alterniflora* (stunted), and *Juncus roemerianus*, and the upper marsh, the transitional zone from salt marsh to freshwater upland that is commonly occupied by *Phragmites australis*, *Iva frutescens*, *Cladium jamaicense*, *Typha* spp., and *Schoenoplectus* spp. (Adams, 1963; Tiner, 1987; Stuckey and Gould, 2000; Nikitina et al., 2003) (Figure 2).

Low biodiversity of the salt marsh ecosystem is related to frequent disturbance and stressful conditions (i.e., high salinity and hypoxia), where differences in plant community are controlled by surface

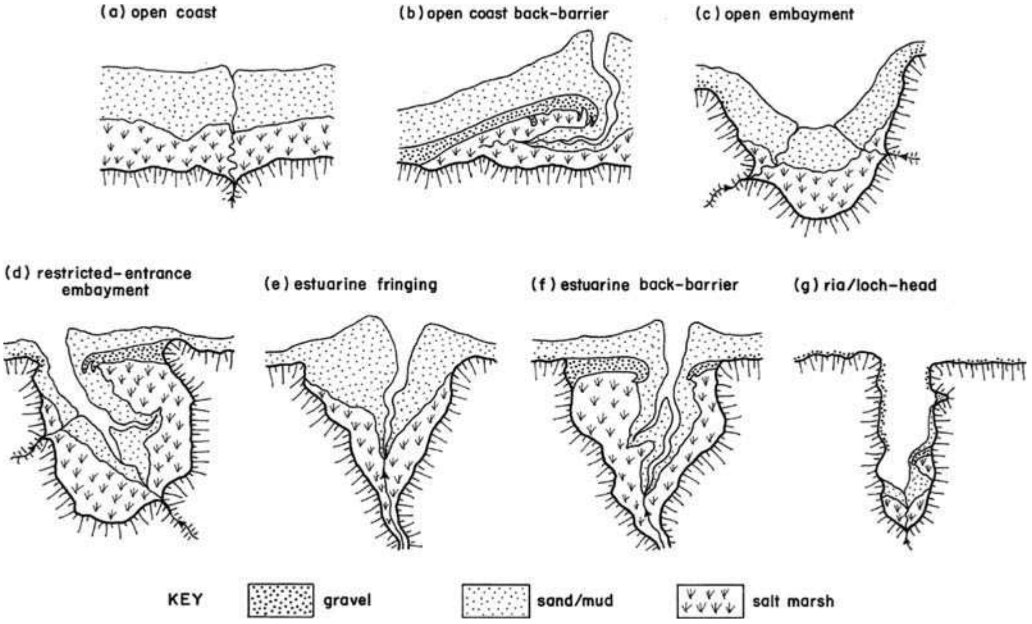


FIGURE 1 Geomorphological classification of salt marshes.
Source: Allen, 2000.

(a) Low marsh (*Sp. Alterniflora*) (b) High marsh *(Sp. Patens/ Sp. Alterniflora)*

(c) Upper marsh (*Phragmites australis)* (d) High marsh *(Distichlis spicata)*

(f) Upper marsh (*Iva frutescens/ Schoenoplectus americanus)*

FIGURE 2 Salt sub-environments and floral communities. (Photo credit: D. Nikitina.)

elevation, salinity regime, and duration and frequency of inundation (McKee and Patrick, 1988; Morris et al., 2002). Low salt marsh with monotypic stands of *S. alterniflora* occurs between mean tide level (MTL) and mean high water (MHW). Height variations within *S. alterniflora* are common, with interior marsh areas having lower vegetation and edges having taller vegetation. The tall (~1.5 m) herbaceous vegetation creates a dense habitat, both above and below ground, and provides habitat for fish, shellfish, and birds (Allen et al., 2018).

The high marsh lies between MHW and mean high high water (MHHW). Higher elevation zones can be more saline in drier climates, due to evaporative concentration of salts, or less saline in higher rainfall areas, due to frequent flushing of salts by fresh rainwater. In the area with hypersaline conditions, communities of succulent salt marsh plants, *Batis* and *Salicornia* spp. often develop (Allen et al., 2018).

The upper marsh occupies the brackish zone at elevations from MHHW to the highest astronomical tide (HAT) where tides mix with freshwater runoff. In climatic zones with warmer winter temperatures, temperate salt marshes naturally transition to mangrove or, in areas with lower precipitation, to salt flats. Arctic marshes are relatively narrow, species-poor, and uniform.

Salt Marsh Ecosystem Services

Salt marshes are economically important coastal landforms and productive ecosystems. Despite occupying <2% of the land, they provide critical ecosystem functions, such as production of organic material and nutrient cycling (Weinstein and Kreeger, 2000; Hopkinson et al., 2012; Howard et al., 2014). They act as nursery grounds for fisheries (Boesch and Turner, 1984; Tupper and Able, 2000); act as biological filters by sequestering pollutants from terrestrial runoff (Brin et al., 2010; Nelson and Zavaleta, 2012); buffer shorelines from storms by attenuating waves and currents that contribute to coastal inundation and damage (Costanza and Folke, 1997; De Groot et al., 2002); and provide tourism and recreation opportunities, education and research (Barbier et al., 2011).

In addition, coastal salt marshes contain 0.1% of the global sequestered terrestrial carbon and have been identified as a "Blue Carbon" ecosystem that plays a significant role in the Global Carbon System (Chung et al., 2011; Mcleod et al., 2011). Overall, "Blue Carbon" ecosystems are responsible for up to 50% of total ocean C burial and contain up to five times more C per unit area than terrestrial forest sediments and 20 times more than deep-sea sediments (Duarte et al., 2005, Cai, 2011, Ouyang and Lee, 2014).

C storage and accumulation capacity are high in salt marshes due to the continuous accumulation of organic C at a greater rate per unit area than any other "Blue Carbon" ecosystem (Chambers et al., 2001; Schlesinger and Lichter, 2001; Duarte et al., 2005; Murray et al., 2011, Chmura, 2013). Sources of organic C in a salt marsh include cordgrasses (primarily *Spartina* spp.), benthic algae, bacteria, allochthonous material including terrestrial and marine organic matter; however, the majority of C is being produced *in situ* by macrophytes (Ember et al., 1987; Radabaugh et al., 2018). Along the submerging coasts, rising sea level produces accommodation space for organic-rich sediments, thereby increasing the inventory of buried organic C through time, as long as salt marsh accretion keeps pace with the rate of rising sea (Kirwan and Blum, 2011; Kirwan and Mudd, 2012). After microbial decay, some portion of the buried C is sequestered over millennia in the marsh sediments (Mcleod et al., 2011, Chmura, 2013). If salt marshes are left undisturbed, then the C stored in the belowground sediments can remain for thousands of years due to anoxic conditions that prevent decomposition (Murray et al., 2011).

Salt Marsh Sustainability

However, despite protective measures, salt marshes are disappearing around the world at an alarming rate. It is estimated that at least 35% salt marshes have been lost globally (Duarte et al., 2008; Pendleton et al., 2012). In some countries, the loss is 70%–80% in the last 50 years (Hopkinson et al., 2012). Coastal marshes have been declining due to industrial development, aquaculture, agricultural reclamation, coastal erosion, and sea-level rise (French, 1997; Blackwell et al., 2004; Kirwan et al., 2010). Many salt marshes are at risk because of pollution by nutrient fluxes, with the largest increases in N flux occurring at coastlines with large areas of intertidal marshland in the temperate zones (Deegan et al., 2012). However, shoreline erosion and inundation are suggested to be the main mechanisms for current global salt marsh loss (Schwimmer, 2001; Mariotti and Fagherazzi, 2010; Theuerkauf et al., 2015).

In the United States, >96% of salt marsh loss is attributed to conversion of marsh to open water (Stedman and Dahl, 2008). In some places, this process could outpace C storage, especially if the depth of erosion at the seaward marsh edge is equal to or greater than the thickness of the marsh sediments (Theuerkauf et al., 2015). The potential future deterioration of marshes and the resulting conversion of their organic sediment to greenhouse gases (CO_2 and CH_4) represent significant impacts on the Global Carbon Budget contributing to global warming and will create positive feedbacks to sea-level rise and ocean acidification (Cai, 2011; Chmura et al., 2011).

The long-term sustainability of salt marsh ecosystem is dependent upon maintaining its elevation within intertidal zone by accumulating organic matter and trapping inorganic sediments (Stumpf, 1983; Morris et al., 2002). However, the processes controlling salt-marsh surface elevation are complex and operate in response to a number of factors including the rate of sea-level rise, sediment supply, tidal range, type of vegetation and primary production (Stevenson et al., 1986; Reed, 1995).

Despite multiple threats, salt marshes are resilient systems. Their ability to accrete vertically depends on concentration of suspended sediment, time of tidal submergence, and biomass density of the plants. Flooding is the main mechanism for sediment delivery to the marsh platform. During flooding, suspended sediments are transported with the tidal currents and partially deposited in distinctive spatial patterns across the marsh platform (Fagherazzi et al., 2012). While the hydrologic regime is largely determined by marsh position within the intertidal range, it is also controlled by geomorphological setting, mainly the proximity to tidal channel. Sedimentation rates across the salt marsh platform are found to decrease with increasing distance from tidal channels and from the seaward marsh edge (French et al., 1995; Leonard, 1997; Reed et al., 1999; Temmerman et al., 2003). Areas along the channel banks and marsh edge vegetated by dense canopy of *S. alterniflora* trap great amount of sediments (Christiansen et al., 2000). Interior areas, which are generally lower in elevation, are more susceptible to submergence and transition to open water, resulting in a disaggregated marsh landscape (Figure 3).

On the other hand, the lower portions of the marsh platform are flooded more frequently for a longer time period, so that more sediment is supplied and deposited. As sediments accrete vertically, sedimentation rates decrease with increasing platform elevation (Stoddart et al., 1989; Cahoon and Reed, 1995).

Hydrologic regime and marsh morphology can be directly modified by anthropogenic activity, including dredging, channelization, upstream modification of rivers and coastal engineering (Kennish, 2001; Turner, 2010). Indirect human modification of the environment including roads, pipelines, dams, oil and water wells and other structures not directly related to salt marsh may affect water and sediment flow, produce a barrier to plant and/or animal migration and contribute to habitat fragmentation, vegetation disturbance and marsh drowning (Allen et al., 2018). Dense vegetation cover favors sedimentation by decreasing tidal current velocities and turbulence once the water flows from the tidal channels into the marsh platform (Leonard and Luther, 1995; Yang, 1998; Christiansen et al., 2000).

FIGURE 3 Interior salt marsh platform recently converted to a salt pond, Slaughter beach salt marsh, Delaware, USA. (Photo credit: D. Nikitina.)

Perhaps the main threat to salt marshes globally is the recent acceleration in the relative sea-level rise (RSLR). Salt marshes have been developing under rising sea-level regime for the last 4000–7000 years by accreting vertically and keeping pace with rates of sea-level rise (Redfield and Rubin, 1962; Redfield, 1965; Stuiver and Daddario, 1963). During the 20th and early 21st centuries, rates of RSLR varied from 2.5 to 3.7 mm/yr (Kopp et al., 2014). For comparison, RSLR along the US Atlantic coast during the last 4000 years varied between 0.6 and 1.7 mm/yr and was mostly driven by glacio-isostatic adjustment of Earth's crust since the melting of the last ice sheet ~7000 years before present (Carlson et al., 2008; Engelhart et al., 2009). Geological data and the tide-gauge records show that RSLR became more rapid throughout the region since the Industrial Revolution identifying the US mid-Atlantic coast as a "hot spot" for sea-level rise (Kemp et al., 2011; Sallenger et al., 2012; Miller et al., 2013). Similar trends were observed elsewhere, reporting the global average rate of sea-level rise from 1880 to 2006 to be 1.7 ± 0.2 mm/yr, while the rate of global average sea-level rise measured from 1993 to 2013 was almost twice as high (3.2 ± 0.4 mm/yr), and this rate appears to be globally accelerating (Church and White, 2006; Jevrejeva et al., 2008; Nerem et al., 2010; Church et al., 2011; Rignot et al., 2011; Gehrels and Woodworth, 2012; Gregory et al., 2012).

Predicted marsh responses to RSLR is either retreat and conversion to tidal flat, tidal lagoon, or open water, if the future rates of RSLR exceed marshes' ability to maintain their elevation in equilibrium with sea level (Reed, 1995; Morris et al., 2002; French, 2006), keep pace with RSLR, or migrate landward. Analysis of geological data, modern observations and simulation studies suggest that salt marshes are resilient to rates of RSLR of ~7.1 mm/yr and likely to retreat when this value exceeds (Horton et al., 2018). Some landscape models using future projections of RSLR increase up to 20 mm/yr predict up to an 80% decrease in global tidal marsh area by 2100 (Spencer et al., 2016; Kopp et al., 2017).

By contrast, other studies suggest that, through biophysical feedback and inland marsh migration, marsh resilience to retreat is possible at RSLR rates in excess of 10 mm/yr (Rodger et al., 2012). Undisturbed vegetation on the marsh platform is a key factor in supporting equilibrium between marsh platform elevation and rapidly rising sea level (Morris et al., 2016). While salt marsh can rapidly subside and transition to open water, subsidence can be compensated by increased production of *S. alterniflora* regulating the elevation of sediment surface toward an equilibrium with mean sea level (Morris et al., 2002). In addition, dense vegetation canopy reduces peak flow velocities, turbulence and shear stress due to friction, enhancing the deposition of suspended sediment and ability of marsh to maintain its elevation within the tidal frame (Leonard and Luther, 1995; Christiansen et al., 2000; Neumeier and Ciavola, 2004; Bouma et al., 2005; Lightbody and Nepf, 2006; Möller et al., 2014).

Marshes also respond to RSLR by migrating into adjacent uplands where they are not restricted by topographic and anthropogenic barriers (Raabe and Stumpf, 2016). The rate of upslope marsh migration is determined by the slope and availability of suspended sediments. Computer simulations of marsh-upland system evolution under a variety of RSLR rates on a gently sloping upland, typical of Atlantic and Gulf coastal plains, suggest that marsh will contract in width under relatively low RSLR rates and low suspended sediment concentration (SSC) and expand under low-moderate-fast RSLR rates and high SSC (Kirwin et al., 2016). The rate of marsh landward transgression depends on the land use of the marsh-adjoining areas, where agricultural fields are more exposed to projected sea-level rise effects than grasslands or forested uplands (Fagherazzi et al., 2019). Construction of dikes, revetments and seawalls prevents salt marsh migration resulting in "coastal squeeze" that leads to marsh loss because erosion of the seaward salt marsh edge cannot be compensated by landward migration (Van der Wal and Pye, 2004).

Recent geological evidence suggests that salt marshes may be significantly eroded during storms and intense hurricane events (van de Plassche et al., 2006; Nikitina et al., 2014). Storm surges during hurricanes move large volumes of water mainly as fast sheet flow producing very high velocities, bottom shear stresses, and drag forces jeopardizing the stability of the vegetation and underlying sediments (Temmerman et al., 2005). Stripping of vegetation and platform erosion was widespread during hurricanes Katrina and Rita along the coast of Gulf of Mexico, USA (Morton and Barras 2011, Turner et al., 2006). While category 2–4 hurricanes are extreme and relatively rare events, salt marshes are often subject to erosion during the smaller storms because during a surge waves propagate over the marsh platform without impacting the edge of the marsh (Leonardi and Fagherazzi 2014, 2015).

Conclusion

Overall, salt marshes are highly dynamic coastal landforms vulnerable to many environmental stresses, mostly RSLR, storm erosion and anthropogenic impact. On the other hand, they are highly resilient coastal ecosystems. Their emergence and continued survival are driven by the rising sea level over millennia. Numerous feedbacks between tidal flooding, plant growth, organic matter production and sediment transport allow marshes to adapt to a wide range of RSLR rates in the vertical and lateral dimension. The global record of recent salt marsh decline correlates with anthropogenically induced acceleration in sea-level rise superimposed on other human activities (e.g., drainage and conversion to agriculture, coastal development, dredging). In many cases, the lateral expansion, the survival mechanism of salt marshes, is impossible due to the presence of topographic or anthropogenic barriers that limit salt marsh migration. However, the major risk of tidal marsh loss in the 21st century is the future rates of RSLR.

References

Adams, D.A., 1963. Factors influencing vascular plant zonation in North Carolina salt marsh. *Ecology*, 44: 445.

Allen, J.R.L., 2000. Morphodynamics of Holocene salt marshes: a review sketch from the Atlantic and Southern North Sea coasts of Europe. *Quaternary Science Reviews*, 19: 1155–1231.

Allen, J.R.L. and K. Pye, 1992. *Saltmarshes: Morphodynamics, Conservation, and Engineering Significance*. Cambridge University Press, Cambridge, 196 pp.

Allen, S.T., C.L. Stagg, J. Brenner, K.L. Goodin, D. Faber-Langendoen, C.A. Gabler, and K.W. Ames. 2018. Ecological resilience indicators for salt marsh ecosystems. In: Goodin, K.L. et al., *Ecological Resilience Indicators for Five Northern Gulf of Mexico Ecosystems*. NatureServe, Arlington, VA, 53 pp.

Barbier, E.B., S.D. Hacker, C. Kennedy, E.W. Koch, A.C. Stier, and B.R. Silliman, 2011. The value of estuarine and coastal ecosystem services. *Ecological Monographs*, 81(2): 169–193.

Blackwell, M.S.A, D.V. Hogana, and E. Maltbya, 2004. The short-term impact of managed realignment on soil environmental variables and hydrology. *Estuarine, Coastal and Shelf Science*, 59: 687–701.

Boesch, D.F. and R.E. Turner, 1984. Dependence of fishery species on salt marshes: the role of food and refuge. *Estuaries*, 7(4): 460–468.

Bouma, T.J., M.B. De Vries, E. Low, L. Kusters, P.M.J. Herman, I.C. Tanczos, A. Hesselink, S. Temmerman, P. Meire, and S. Van Regenmortel, 2005. Hydrodynamic measurements on a mudflat and in salt marsh vegetation: identifying general relationships for habitat characterisations. *Hydrobiologia*, 540: 259–274, doi:10.1007/s10750-004-7149-0.

Brin, L.D., I. Valiela, D. Goehringer, and B.L. Howes, 2010. Nitrogen interception and export by experimental salt marsh plots exposed to chronic nutrient addition. *MA Thesis*, Boston University.

Cahoon, D.R., and D.J. Reed, 1995. Relationships among marsh surface topography, hydroperiod, and soil accretion in a deteriorating Louisiana salt marsh. *Journal of Coastal Conservation*, 11: 357–369.

Cai, W.-J., 2011. Estuarine and coastal ocean carbon paradox: CO_2 sinks or sites of terrestrial carbon incineration? *Annual Review of Marine Science*, 3: 123–145.

Carlson, A.E., A.N. LeGrande, D.W. Oppo, R.E. Came, G.A. Schmidt, F.S. Anslow, J.M. Licciardi, and E.A. Obbink, 2008. Rapid early Holocene deglaciation of the Laurentide ice sheet. *Nature Geoscience*, 1: 620–624.

Chambers J.Q., N. Higuchi, E.S. Tribuzy, and S.E. Trumbore, 2001. Carbon sink for a century. *Nature*, 410: 429. doi.org/10.1038/35068624.

Chmura, G.L., 2013. What do we need to assess the sustainability of the tidal saltmarsh carbon sink? *Ocean & Coastal Management*, 83: 25–31. doi:10.1016/j.ocecoaman.2011.09.006.

Chmura, G.L., L. Kellman, and G.R. Guntenspergen, 2011. The greenhouse gas flux and potential global warming feedbacks of a northern macrotidal and microtidal salt marsh. *Environmental Research Letters*, 6(4): 1–6.

Christiansen, T., P.L. Wiberg, and T.G. Milligan, 2000. Flow and sediment transport on a tidal salt marsh surface. *Estuarine Coastal Shelf Science*, 50(3): 315–331, doi:10.1006/ecss.2000.0548.

Chung, I.K., J. Beardall, S. Mehta, D. Sahoo, and S. Stojkovic, 2011. Using marine macroalgae for carbon sequestration: a critical appraisal. *Journal of Applied Phycology*, 203(5): 877–886.

Church, J.A. and N.J. White, 2006. A 20th century acceleration in global sea-level rise. *Geophysical Research Letters*, 33, L01602, doi:10.1029/2005GL024826.

Church, J.A. and N.J. White, 2011. Sea-level rise from the late 19th to the early 21st century, *Surveys in Geophysics*, 32: 585–602, doi:10.1007/s10712-011-9119-1.

Costanza, R. and C. Folke, 1997. *Valuing Ecosystem Services with Efficiency, Fairness, and Sustainability as Goals, Nature's Services: Societal Dependence on Natural Ecosystems*. Island Press, Washington, DC, 49–70.

Deegan, L., D.S. Johnson, R.S. Warren, B.J. Peterson, J.W. Fleeger, S. Fagherazzi, and W.M. Wollheim, 2012. Coastal eutrophication as a driver of salt marsh loss. *Nature*, 490: 388–392.

De Groot, R.S., M.A. Wilson, and R.M. Boumans, 2002. A typology for the classification, description and valuation of ecosystem functions, goods and services. *Ecological Economics*, 41(3): 393–408.

Duarte, C.M., J.J. Middelburg, and N. Caraco, 2005. Major role of marine vegetation on the oceanic carbon cycle. *Biogeosciences*, 2: 1–8.

Duarte, C.M., W.C. Dennison, R.J.W. Orth, and T.J.B. Carruthers, 2008. The charisma of coastal ecosystems: addressing the imbalance. *Estuaries and Coasts*, 31: 233–238. doi:10.1007/s12237-008-9038-7.

Ember, L., D. Williams, and J. Morris, 1987. Processes that influence carbon isotope variations in salt marsh sediments. *Marine Ecology Progress Series* 36: 33–42.

Engelhart, S.E., B.P. Horton, B.C. Douglas, W.R. Peltier, and T.E. Tornqvist, 2009. Spatial variability of late Holocene and 20th century sea-level rise along the Atlantic coast of the United States. *Geology*, 37: 1115–1118.

Fagherazzi, S., S.C. Anisfeld, L.K. Blum, E.V. Long, R.A. Feagin, A. Fernandes, W.S. Kearney, and K. Williams, 2019. Sea level rise and the dynamics of the marsh-upland boundary. *Frontiers in Environmental Science*, 27, doi:10.3389/fenvs.2019.00025.

Fagherazzi, S., M.L. Kirwan, S.M. Mudd, Glenn R. Guntenspergen, S. Temmerman, A. Andrea D'Alpaos, J. van de Koppel, J.M. Rybczyk, E. Reyes, C. Craft, and J. Jonathan Clough, 2012. Numerical models of salt marsh evolution: ecological, geomorphic, and climatic factors. *Reviews of Geophysics*, 50, RG1002, doi:10.1029/2011RG000359.

French, J.R., 2006. Tidal marsh sedimentation and resilience to environmental change: exploratory modelling of tidal, sea-level and sediment supply forcing in predominantly allochthonous systems. *Marine Geology*, 235: 119–136.

French, J.R., T. Spencer, A.L. Murray, and N.S. Arnold, 1995. Geostatistical analysis of sediment deposition in two small tidal wetlands, Norfolk, United Kingdom. *Journal of Coastal Research*, 11: 308–321.

French, P.W., 1997. *Coastal and Estuarine Management* (Routledge Environmental Management Series). Routledge, London, 251.

Gehrels, W.R. and P.L. Woodworth, 2012. When did modern rates of sea-level rise start? *Global and Planetary Change*, 100: 263–277, doi:10.1016/j.gloplacha.2012.10.020.

Gregory, J.M., et al., 2012. Twentieth-century global-mean sea-level rise: is the whole greater than the sum of the parts? *Journal of Climate*, 26: 4476–4499, doi:10.1175/JCLI-D-12-00319.1.

Hopkinson, C.S., W-J. Cai, and X. Hu, 2012. Carbon sequestration in wetland dominated coastal systems – a global sink of rapidly diminishing magnitude. *Current Opinion in Environmental Sustainability*, 4: 186–194, doi.org/10.1016/j.cosust.2012.03.005.

Horton, B.P, I. Shennan, S. Bradley, N. Chahill, M. Kirwan, R.E. Kopp, and T.A. Shaw, 2018. Predicting marsh vulnerability to sea-level rise using Holocene relative sea-level data. *Nature Communications*, 9: 2687.

Howard, J., S. Hoyt, K. Isensee, E. Pidgeon, and M. Telszewski, 2014. *Coastal Blue Carbon: Methods for Assessing Carbon Stocks and Emissions Factors in Mangroves, Tidal Salt Marshes, and Seagrass Meadows*. Conservation International, Intergovernmental Oceanographic Commission of UNESCO, International Union for Conservation of Nature, Arlington, VI, USA.

Jevrejeva, S., J.S. Moore, A. Grinsted, and P.L. Woodworth, 2008. Recent global sea level acceleration started over 200 years ago? *Geophysical Research Letters*, 35, L08715, doi:10.1029/2008GL033611.

Kemp, A.C., B.P. Horton, J.P. Donnelly, M.E. Mann, M. Vermeer, and S. Rahmstorf, 2011. Climate related sea-level variations over the past two millennia. *Proceedings of the National Academy of Sciences of the United States of America*, 108: 11,017–11,022, doi:10.1073/pnas.1015619108.

Kennish, M.J., 2001. Coastal salt marsh systems in the U.S.: a review of anthropogenic impacts. *Journal of Coastal Research*, 17: 731–748.

Kirwan, M.L. and L.K. Blum, 2011. Enhanced decomposition offsets enhanced productivity and soil carbon accumulation in coastal wetlands responding to climate change. *Biogeosciences* 8(4): 987. doi:10.5194/bg-8-987-201.

Kirwan, M.L. and S.M. Mudd, 2012. Response of salt-marhs carbon accumulation to climate change. *Nature*, 489: 550–553, doi:10.1038/nature11440.

Kirwan, M.L., G.R. Guntenspergen, A. D'Alpaos, J.T. Morris, S.M. Mudd, and S. Temmerman, 2010. Limits on the adaptability of coastal marshes to rising sea level. *Geophysical Research Letters*, 37. doi:10.1029/2010GL045489.

Kirwan, M.L., S. Temmerman, E.E. Skeehan, G.R. Guntenspergen, and S. Fagherazzi, 2016. Overestimation of marsh vulnerability to sea level rise. *Nature Climate Change*, 6: 253–260.

Kopp, R.E., R.M. DeConto, D.A. Bader, C.C. Hay, R.M. Horton, S. Kulp, M. Oppenheimer, D. Pollard, and B.H. Strauss, 2017. Evolving understanding of Antarctic ice-sheet physics and ambiguity in probabilistic sea-level projections. *Earth's Future*, 5: 1217–1233, doi.org/10.1002/2017EF000663.

Kopp, R.E., R.M. Horton, C.M. Little, J.X. Mitrovica, M. Oppenheimer, D.J. Rasmussen, B.H. Strauss, and C. Tebaldi, 2014. Probalisitic 21st and 22nd century sea-level projections at a global network of tide-guage sites. *Earth's Future*, 2: 383–406.

Leonard, L.A., 1997. Controls on sediment transport and deposition in an incised mainland marsh basin, southeastern North Carolina. *Wetlands*, 17: 263–274, doi:10.1007/BF03161414.

Leonard, L.A. and M.E. Luther, 1995. Flow hydrodynamics in tidal marsh canopies. *Limnology and Oceanography*, 40(8), 1474–1484, doi:10.4319/lo.1995.40.8.1474.

Leonardi, N. and S. Fagherazzi, 2014. How waves shape salt marshes. *Geology*, 42(10): 887–890.

Leonardi, N. and S. Fagherazzi, 2015, Effect of local variability in erosional resistance on largescale morphodynamic response of salt marshes to wind waves and extreme events. *Geophysical Research Letters*, 42(14): 5872–5879.

Lightbody, A.F. and H.M. Nepf, 2006. Prediction of velocity profiles and longitudinal dispersion in emergent salt marsh vegetation. *Limnology and Oceanography*, 51: 218–228, doi:10.4319/lo.2006.51.1.0218.

Mariotti, G., and S. Fagherazzi, 2010. A numerical model for the long-term evolution of salt marshes and tidal flats. *Journal of Geophysical Research*, 115, F01004. doi:10.1029/2009JF001326.

McKee, K.L. and W.H. Patrick, 1988. The relationship of smooth cordgrass (*Spartinaalternifl ora*) to tidal datums: a review. *Estuaries*, 11: 143–151.

Mcleod, E., G.L. Chmura, S. Bouillon, R. Salm, M. Björk, C.M. Duarte, C.E. Lovelock, W.H. Schlesinger, and B.R. Silliman, 2011. A blueprint for blue carbon: toward an improved understanding of the role of vegetated coastal habitats in sequestering CO_2. *Frontiers in Ecology and the Environment*, 9: 552–560.

Miller, K.G., R.E. Kopp, B.P. Horton, J.V. Browning, and A.C. Kemp, 2013. A geological perspective on sea-level rise and its impacts along the U.S. mid-Atlantic coast. *Earth's Future*, doi:10.1002/2013EF000135.

Möller, I., M. Kudella , F. Rupprecht, T. Spencer, M. Paul, B.K. van Wesenbeeck, G. Wolters, K. Jensen, T.J. Bouma, M. Miranda-Lange, and S. Schimmels, 2014. Wave attenuation over coastal salt marshes under storm surge conditions. *Nature Geoscience*, 7(10): 727–731.

Morris, J.T., D.C. Barber, J.C. Callaway, R. Chambers, S.C. Hagen, C.S. Hopkinson, P. Johnson, B.J. Megonigal, S.C. Neubauer, T. Troxler, and C. Wigand, 2016. Contributions of organic and inorganic matter to sediment volume and accretion in tidal wetlands at steady state. *Earth's Future*, 4. doi:10.1002/2015EF000334.

Morris, J.T., P.V. Sundareshwar, C.T. Nietch, B. Kjerfve, and D.R. Cahoon, 2002. Responses of coastal wetlands to rising sea level. *Ecology*, 83(10): 2869–2877, doi:10.1890/0012-9658.

Morton, R.A. and J.A. Barras, 2011. Hurricane impacts on coastal wetlands: a half-century record of storm-generated features from southern Louisiana. *Journal of Coastal Research*, 27(6A): 27–43.

Murray, B.C., L. Pendleton, and S. Sifleet, 2011. *State of the Science on Coastal Blue Carbon: A Summary for Policy Makers*. Nicholas Institute for Environmental Policy Solutions Report NIR 11-06: 1–43.

Nelson, J.L. and E.S. Zavaleta, 2012, Salt marsh as a coastal filter for the oceans: changes in function with experimental increases in nitrogen loading and sea-level rise. *PloS One*, 7(8): e38558.

Nerem, R.S., D.P. Chambers, C. Choe, and G.T. Mitchum, 2010. Estimating mean sea-level change from the TOPEX and Jason Altimeter Missions, Mar. *Geodesy*, 33: 435–446, doi:10.1080/01490419.2010.491031.

Neumeier, U. and P. Ciavola, 2004. Flow resistance and associated sedimentary processes in a Spartina maritima salt marsh. *Journal of Coastal Research*, 20: 435–447, doi:10.2112/1551-5036.

Nikitina, D.L., A.C. Kemp, B.P. Horton, C.H. Vane, O. van de Plassche, and S.E. Engelhart, 2014. Storm erosion during the past 2000 years along the north shore of Delaware Bay, USA. *Geomorphology*, 208: 160–172.

Nikitina, D.L., J.E. Pizzuto, R.E. Martin, and S.P. Hippensteel, 2003. Transgressive valley-fill stratigraphy and sea-level history of the Leipsic River, Bombay Hook National Wildlife Refuge, Delaware, U.S.A. In Olson, H.C. and Leckie, R.M. (Eds.), *Micropaleontologic Proxies for Sea-Level Changes and Stratigraphic Discontinuities*. SEPM Special Publication, Tulsa, OK, 75, 51–62.

Ouyang, X. and S.Y. Lee, 2014. Updated estimates of carbon accumulation in coastal marsh sediments. *Biogeosciences*, 11: 5057–5071.

Pendleton, L., D.C. Donato, B.C. Murray, S. Crooks, W.A. Jenkins, S. Sifleet, C. Craft, J.W. Fourqurean, J.B. Kauffman, N. Marbá, P. Megonigal, E. Pidgeon, D. Herr, D. Gordon, A. Balbera, 2012. Estimating global "Blue Carbon" emissions from conversion and degradation of vegetated coastal ecosystems. *PLoS ONE* 7(9): e43542, doi:10.1371/journal.pone.0043542.

Raabe, E.A. and R.P. Stumpf, 2016. Expansion of tidal marsh in response to sea-level rise: Gulf Coast of Florida, USA. *Estuaries Coasts*, 39: 145–157.

Radabaugh, K.R., Moyer, R.P., Chappel, A.R., Powell, C.E., Bociu, I., Clark, B.C., Smoak, J.M., 2018, Coastal Blue Carbon Assessment of Mangroves, Salt Marshes, and Salt Barrens in Tampa Bay, Florida, USA. *Estuaries and Coasts*, 41: 1496, doi:10.1007/s12237-017-0362-7.

Redfield, A.C., 1965. Ontogeny of a salt marsh estuary. *Science*, 147: 50–55, doi:10.1126/science.147.3653.50.

Redfield, A.C. and M. Rubin, 1962. The age of salt marsh peat and its relation to recent changes in sea level at Barnstable, Massachusetts. *Proceedings of the National Academy of Sciences of the United States of America*, 48(10): 1728.

Reed, D.J., 1995. The response of coastal marshes to sea-level rise: survival or submergence? *Earth Surface Processes and Landforms*, 20: 39–48. doi:10.1002/esp.3290200105.

Reed, D.J., T. Spencer, A.L. Murray, J.R. French, and L. Leonard, 1999. Marsh surface sediment deposition and the role of tidal creeks: implications for created and managed coastal marshes. *Journal of Coastal Conservation*, 5: 81–90, doi:10.1007/BF02802742.

Rignot, E., I. Velicogna, M.R. van den Broeke, A. Monaghan, and J. Lenaerts, 2011. Acceleration of the contribution of the Greenland and Antarctic ice sheets to sea-level rise. *Geophysical Research Letters*, 38, L05503, doi:10.1029/2011GL046583.

Rogers, K. and C.D. Woodroffe, 2014. Tidal Flats and Salt marshes. In G. Masselink, and R. Gehrels (Eds.), *Coastal Environments and Global Change*. Wiley & Sons, Chichester, 227–248.

Rogers, K., N. Saintilan, and C. Copeland, 2012. Modelling wetland surface elevation dynamics and its application to forecasting the effects of sea-level rise on estuarine wetlands. *Ecological Modelling* 244: 148–157.

Sallenger, A.H., Jr., K.S. Doran, and P.A. Howd, 2012. Hotspot of accelerated sea-level rise on the Atlantic Coast of North America. *Nature Climate Change*, 2: 884–888, doi:10.1038/nclimate1597.

Schlesinger, W.H. and J. Lichter, 2001. Limited carbon storage in soil and litter of experimental forest plots under increased atmospheric CO_2. *Nature*, 411: 466–469.

Schwimmer, R.A., 2001. Rates and processes of Marsh shoreline erosion in Rehoboth Bay, Delaware, U.S.A. *Journal of Coastal Research*, 17(3): 672–683.

Spencer, T., M. Schürch, R.J. Nicholls, J. Hinkel, A. Vafeidis, R. Reef, L. McFadden, et al., 2016. Global coastal wetland change under sea-level rise and related stresses: the DIVA Wetland Change Model. *Global and Planetary Change*, 139: 15–30, doi:10.1016/j.gloplacha.2015.12.018.

Stedman, S. and T.E. Dahl. 2008. Status and trends of wetlands in the coastal watersheds of the Eastern United States 1998 to 2004. National Oceanic and Atmospheric Administration, National Marine Fisheries Service and U.S. Department of the Interior, Fish and Wildlife Service. 20 pp.

Stevenson, J.C., L.G. Ward, and M.S. Kearney, 1986. Vertical accretion in marshes with varying rates of sea level rise, In D.A. Wolfe (Ed.), *Estuarine Variability*. Academic Press, San Diego, CA, 241–259.

Stoddart, D.R., D.J. Reed, and J.R. French, 1989. Understanding salt marsh accretion, Scolt Head Island, Norfolk, England. *Estuaries*, 12(4): 228–236, doi:10.2307/1351902.

Stuckey, I.H. and L.L. Gould, 2000. *Coastal Plants from Cape Cod to Cape Canaveral*. University of North Carolina Press, Chapel Hill.

Stuiver, M. and J.J. Daddario, 1963. Submergence of the New Jersey coast. *Science*, 142(3594): 951.

Stumpf, R.P., 1983. The process of sedimentation on the surface of a salt marsh, Estuarine. *Coastal and Shelf Science*, 17(5): 495–508.

Temmerman, S., T.J. Bouma, G. Govers, and D. Lauwaet, 2005. Flow paths of water and sediment in a tidal marsh: relations with marsh developmental stage and tidal inundation height. *Estuaries*, 28(3): 338–352, doi:10.1007/BF02693917.

Theuerkauf, E.J., J.D. Stephens, J.T. Ridge, F.J. Fodrie, and A.B. Rodriguez, 2015. Carbon export from fringing saltmarsh shoreline erosion overwhelms carbon storage across a critical width threshold: Estuarine. *Coastal and Shelf Science*, 164: 367–378, dx.doi.org/10.1016/j.ecss.2015.08.001.

Tiner, R.W. Jr, 1987. *Field Guide to Coastal Wetland Plants of the Northern United States*. University of Massachusetts Press, Amherst.

Tupper, M. and K. Able, 2000. Movements and food habits of striped bass (Morone saxatilis) in Delaware Bay (USA) salt marshes: comparison of a restored and a reference marsh. *Marine Biology*, 137(5–6): 1049–1058.

Turner, R.E., 2010. Beneath the salt marsh canopy: loss of soil strength with increasing nutrient loads. *Estuaries Coasts* 34: 1084–1093. doi:10.1007/s12237-010-9341-y.

Turner, R.E., J.J. Baustian, E.M. Swenson, and J.S. Spicer, 2006. Wetland sedimentation from hurricanes Katrina and Rita. *Science*, 314(5798): 449–452.

van de Plassche, O., G. Erkens, F. van Vliet, J. Brandsma, K. van der Borg, and A.F. de Jong, 2006. Salt-marsh erosion associated with hurricane landfall in southern New England in the fifteenth and seventeenth centuries. *Geology*, 34(10): 829–832.

Van der Wal, D. and K. Pye, 2004. Patterns, rates and possible causes of saltmarsh erosion in the Greater Thames area (UK). *Geomorphology*, 61: 373–391.

Weinstein, M.P. and Kreeger, D.A., 2000. *Concepts and Controversies in Tidal Marsh Ecology*, Springer, Dordrecht, doi:10.1007/0-306-47534-0.

Yang, S.L., 1998. The role of Scirpus marsh in attenuation of hydrodynamics and retention of fine sediment in the Yangtze Estuary. *Estuarine Coastal Shelf Science*, 47: 227–233, doi:10.1006/ecss.1998.0348.

23

The Evolution of Water Resources Management

Francine van den Brandeler and Joyeeta Gupta

Introduction

Water resources are crucial for lives, livelihoods, biodiversity, ecosystems, and economic development. The adequate management of these water resources is therefore intrinsic to sustainable and inclusive development. Water problems can be broadly characterized in terms of quantity, quality, and the effects of extreme weather events (UNE 2019). Water quantity challenges can be absolute or relative (i.e., dependent on the ratio between water availability and demand), with 'chronic water shortages' defined as annual water supplies below 1000 m^3 per person within a certain area drop and 'absolute water scarcity' as less than 500 m^3 per person (FAO 2012). These challenges may be caused or aggravated by overabstracting water from rivers and groundwater resources. Quantitative challenges can also relate to the unequal distribution of water resources within an area and inequality of access. Qualitative water challenges relate to pollution from point sources (e.g., untreated industrial or domestic wastewater discharge) and diffuse pollution (e.g., fertilizers and pesticides from agriculture and surface runoff in urban areas) (Martinez-Santos et al. 2014; Elmqvist et al. 2013). Water-related risks linked to extreme weather events include droughts, floods, and landslides, with devastating consequences on human settlements, agriculture, and ecosystems (GRID 2018).

These challenges are affected by five driving forces that have put increased pressures on water systems around the world: (1) population dynamics, (2) urbanization, (3) technology development, (4) a constant pursuit of economic growth, and (5) climate change (UNE 2019). These, in turn, affect land use and land-use changes, such as deforestation to clear land for agriculture and urbanization; these lead to further deterioration of water quality by disturbing ecosystems and the hydrological cycle (Azzam et al. 2014; Avissar and Werth 2005). Climate variability and change affect hydrological systems by making extreme weather events more intense and more frequent (IPCC 2014).

Water resources management aims to address these challenges and underlying driving forces. Management encompasses 'the processes of decision-making, coordination, and resource deployment that occur within a given institutional setting assuming no change in rules and norms' (Hatfield-Dodds et al. 2007, 3). Although management approaches rely on scientific and technical knowledge, they are also influenced by available human and financial resources, as well as contextual factors, politics, and power relations. Indeed, "the challenge of the water crisis is first and foremost a 'crisis of governance'" (UNESCO 2006, 1), relating not only to how management systems are designed but also to how water is conceptualized, who owns water, how boundaries are delineated for management purposes, and how links to other issues are framed (Gupta et al. 2013). Water resources management approaches, in this sense, are a particular strategy of governance. They influence the institutional tools selected for allocating, regulating, and preserving water resources, the type of water (blue, black, gray, and green), infrastructure and technologies, and information systems. Water resources management approaches have evolved over time and increasingly incorporate notions of sustainability and inclusiveness, and this is reflected at global level through the adoption of a specific goal on water within the Sustainable Development Goals (SDGs) (UNGA 2015), while many other goals are directly or indirectly linked to water. Nevertheless, as water-related challenges multiply and increase in complexity, there is growing uncertainty about the most adaptive approach to manage water resources. The sections below discuss several of the main paradigms that have shaped water resources management over time, including the control of water resources to consolidate power, which started with the first ancient civilizations; water resources management as an instrument of colonization; the hydraulic paradigm; Integrated Water Resources Management (IWRM); Adaptive Management (AM); the Water–Energy–Food Nexus; and urban-centric approaches to water management.

Water Resources Management in Pre-Modern Times

Rules on managing water first developed in local customary law which later spread through the expansion of the early river basin civilizations (Gupta and Dellapenna 2009). Major rivers historically attracted communities; early states aimed at consolidating power by controlling water resources (Wittfogel 1955). Harnessing water resources was essential for the development of agriculture in Mesopotamia and Egypt, around 5700–3200 BCE, which included systems of communal canals, irrigation works, and legal frameworks to govern these (Angelakis and Zheng 2015; Kornfeld 2009). The Indus Valley civilization developed around the Indus Valley. In ancient Rome, aqueducts, fountains, and pipes on a large-scale supplied freshwater to cities and removed wastewater (Wilson 1998). Complex management systems developed first in areas where water was scarce such as the Middle East and influenced Islamic and Jewish water law and policy. The Aztecs founded Tenochtitlán (now Mexico City) in 1324 on an island surrounded by lakes within an endorheic basin. The lakes protected them from enemy invaders and provided fish and food from floating farms and through canals and irrigation infrastructures (Sosa-Rodriguez 2010). Their dependence on the lakes meant that it was essential for their survival that they preserved the hydrological balance. Exposed to frequent droughts and floods, the Aztecs also suffered from insufficient water supply, low-quality water, and inadequate sanitation. To exercise control over the basin's hydrology, they not only built drains, dams, dikes, floodgates, and aqueducts but also implemented strict rules such as the prohibition of waste disposal into water bodies. In the Netherlands, the water boards established in the 13th century to protect the lowest-lying regions from floods still exist and have been institutionalized into about 200 boards today (Lazaroms and Poos 2004). These democratically elected decentralized government authorities manage local to regional water bodies to manage water quality- and quantity-related issues and are allowed to raise resources through taxes.

The fall of civilizations has frequently been attributed to water mismanagement. The Sumerian and Indus Valley civilizations may have collapsed from soil salinization through irrigated agriculture in an arid climate (Alam et al. 2007). The Maya civilization's downfall in the early 900s may have been caused by a series of drought events that urban elites were unable to adapt to (Lucero et al. 2011). Overall, water resources management and the ability to adapt to changing conditions have been key to both the rise and fall of civilizations.

Colonial Times

During colonial times, water was central for exploration, establishing sovereignty and commercial benefits. River navigation played an important role. Portuguese colonizers in Brazil strategically used navigation to explore the newly conquered territory, map the land, control rivers, and facilitate state integration (Paganini 2008; Leite Farias 2009). In the next stage, administrating colonies required water resources management and institutions for irrigation, flood and drought prevention, and transportation (Bhattacharya 2017). After the fall of ancient civilizations, their sophisticated irrigation techniques and hydraulic works disappeared almost everywhere except in China, but these reappeared in the 19th century, as colonial powers mobilized massive corvée labor and diverted rivers to irrigate alluvial planes and deltas at great social cost (Molle et al. 2009). Colonial powers frequently cast aside indigenous water management knowledge and practices to impose theirs while possessing little understanding of local hydrology. For instance, the Spaniards transformed the basin's hydrology after they invaded Tenochtitlán. They associated the Aztec city's surrounding lakes with an unhealthy environment and a source of disease, and soon began ambitious efforts to drain the lakes, which aggravated recurrent floods (Sosa-Rodriguez 2010).

Water resources management institutions imposed as part of the colonization process often partly or largely replaced customary water regimes and asserted the colonial state's ownership over water resources (Cullet and Gupta 2009). Customary laws often continue to influence water regimes and co-exist with colonial and post-colonial institutions, leading to dual legal regimes (Nilsson and Nyanchaga 2009). In rural early America, water use was considered a community right and, as the nation consolidated, water use acquired some standing as an individual right. Conflicts between the two led to the doctrine of riparian rights, which states that water belongs to the person whose land borders a body of water (Burke 1956). With increasing competition for water resources in the 19th century, the rule of prior appropriation was introduced ('first in time, first in right'), particularly in the western states (Hanak et al. 2011).

Overall, colonialism commodified water resources while dismantling community control, causing the impoverishment of rural populations, as was the case in India where indigenous water harvesting systems were purposely destroyed (D'Souza 2006). Infrastructure and technology for drinking water provision were mainly developed in areas occupied by the colonizers as European powers were reluctant to invest in their overseas colonies unless this would be profitable, but rapidly growing cities, such as Bombay in the British Empire, experienced deterioration in urban conditions and rising tensions (Anand 2011; Nilsson and Nyanchaga 2009; Gandy 2008).

The Hydraulic Paradigm

Scientific developments, fascination with 'scientific irrigation', and ideas such as the domination of nature and the creation of new Edens in deserts cumulated into large public investments in dams and irrigation systems as of the beginning of the 20th century (Molle et al. 2009; Smythe 1905). States developed water systems as part of a 'Hydraulic mission' that aimed to increase food production, raise rural incomes, respond to growing urban water demand, strengthen state building, and legitimize governments (Hanak et al. 2011; Molle et al. 2009). In Spain, this was part of a response to the humiliating loss of colonies and the turn of colonial ambitions inward to dry and poor rural areas (Lopez-Gunn 2009). The state-led hydraulic policy served to legitimize the state, reinvent the economy, and assuage the threat of a discontented peasantry. Technological advancements, such as drilling and gasoline- and diesel-powered pumps, enhanced access to groundwater and further expanded agricultural land in dry regions such as California (Hanak et al. 2011).

The hydraulic paradigm spread worldwide, as water managers and governments strived to control the natural environment through infrastructures for reducing floods, generating hydropower and supplying cities with drinking water (Custódio 2012; Swyngedouw 1999; Lopez-Gunn 2009; Molle et al. 2009), requiring large investments and centralized management (Pahl-Wostl et al., 2007; Huitema and Meijerink, 2017). Within the United States, urban water supply systems were initially often built by small, private companies and mainly served wealthier neighborhoods (Porse 2014). However, seeking more reliable options for public health, flood control, and economic growth in urban areas, water delivery, and stormwater and sewage conveyance services were increasingly centralized and municipalized in the late 19th and early 20th centuries (Rietveld et al. 2016; Domènech 2011). As cities grew, dams and aqueducts were built to bring water from distant locations and replaced dwindling or polluted local sources (Porse 2014; Hanak et al. 2011), and the emphasis was on supply augmentation rather than demand management (Xie 2006).

The creation of national water bureaucracies in the post-World War II period and efforts to modernize customary, state, and colonial water laws soon followed to regulate many issues including large-scale engineering works (Hanak et al. 2011). The physical integration of waterworks in fragmented countries supported their integration and often became a focal point for nation-building (Huitema and Meijerink 2017). However, this process often involved the expropriation of rights from private owners to state bureaucracies in order to further national economic interests by strengthening agriculture and emerging industries. This included local and indigenous water rights under colonial and post-colonial regimes (Boelens 2009). Water management tasks were generally allocated to government agencies at provincial or national levels (Huitema and Meijerink 2017). Different tasks were the responsibility of different sectors (agriculture, domestic use, industry, environmental protection, etc.) with often little coordination between them (Xie 2006).

The Hydraulic Era gave way to different forms of water control. Large-scale, centralized infrastructure provided many societal and economic benefits but could not cover distant areas and vulnerable groups. In California, lakes dried up and rivers became wastewater canals, aquatic species declined sharply, and farmers complained about increasing salinity in their water supplies due to upstream diversions (Hanak et al. 2011). With the start of the environmental movement in the 1970s, demands for regulating pollution of water became more important. This movement spread around the world, strengthening environmental legislation and also emphasizing nature as a legitimate user of water resources (Hanak et al. 2011; Kallis and Coccossis 2002). This was institutionalized through measures such as the establishment of minimum environmental flows and maximum groundwater extractions, environmental impact assessments, and stricter regulations regarding wastewater discharges. This influenced water governance. Another feature influencing water governance was the shift from 'government to governance' and 'governance-beyond-the-state' (Huitema and Meijerink 2017; Swyngedouw 2005) and the rise of the non-state actor. This coincided with the neo-liberal demand for lean states and deregulation and led to water resources being managed by assemblages of public, private, and civil society actors with different approaches to water, shaping not only how they address problems but also how they interact with other actors (Brandeler et al. 2014). Power is transferred to both lower and higher levels. Decentralization processes have brought decision-making closer to those affected, enhancing democratic legitimacy (Dryzek 1997) and resulting in policies more likely to be supported at local levels (Kasemir et al. 2003). However, this has also led to a lack of accountability and patchy institutions that are unable to address society's needs.

Meanwhile, the recognition of global water drivers and cumulative impacts and the need for common norms have increased support for global water governance in order to address the global drivers (Gupta and Pahl-Wostl 2013; Vörösmarty et al. 2013). With the adoption of the SDGs, water governance can no longer be seen as independent of all the other goals. Some scholars see that the IWRM paradigm may have to evolve into a Nexus paradigm in order to be able to preempt the problems of tomorrow.

IWRM – Dominant Paradigm Today

The ideas behind the concept of IWRM emerged in the 1970s through practitioners, and these first appeared in official water policy circles at the global water conference in Mar del Plata in 1977, which emphasized the need for coordinating different users and authorities, participation, and legal frameworks to ensure effective water allocation (WWAP 2009). It was developed in the late 1980s and was adopted in the Dublin

Principles adopted by the International Conference on Water and the Environment in Dublin. The four principles recognized freshwater as ‘a finite and vulnerable resource’, integrated, participatory management at the ‘lowest appropriate level’, emphasis on the central role of water as well as on the economic value of water (ICWE 1992). The Global Water Partnership (GWP) defined IWRM as ‘a process that promotes the coordinated development and management of water, land, and related resources, in order to maximize the resultant economic and social welfare in an equitable manner without compromising the sustainability of vital ecosystems’ (GWP-TAC 2000). IWRM has since become the dominant paradigm for water resources management, having been adopted by international agencies and transnational actors such as the World Bank, the European Union (EU) (e.g., Water Framework Directive and EU Water Initiative), and the United Nations, most recently through the latter’s SDGs (UNGA 2015; Abers and Keck 2013; Wallington et al. 2010; Abers 2007; Molle 2009) but also by a growing number of national governments to frame their laws and policies (Gupta and Pahl-Wostl 2013; UNEP 2012).

IWRM strives to address social equity, including the water-related needs of the poor (Hooper 2005; Rahaman and Varis 2005). It promotes the integration of human and natural systems (Dzwairo et al. 2010), sectoral responses, and viewpoints and interests (Medema et al. 2008; Jønch-Clausen and Fugl 2001). While marginalized people are given a voice, IWRM also mirrors general governance shifts towards greater private sector involvement through emphasis on cost recovery, the establishment of pricing mechanisms and cost-benefit analyses, and allocation to the most beneficial uses (Huitema and Meijerink 2017; Xie 2006). Environmental sustainability is addressed through concerns about the ecological impacts of water resources management activities, including through disruption to the water cycle, pollution control, development planning, demand management, and biodiversity conservation (Medema et al. 2008; Savenije and Van der Zaag 2008; Grigg 2008). In addition, although IWRM works through national-level legislation, policies, and institutions, it emphasizes the river basin as the ideal unit for water resources management and the creation of river basin organizations (RBOs) (Butterworth et al. 2010; Hooper 2005; Abdullah and Christensen 2004). It is considered as crucial to integrate upstream and downstream concerns, to foster bottom-up planning (Molle 2009; Xie 2006). This systems approach is often seen as the key innovation of IWRM (Huitema and Meijerink 2017).

However, there are many criticisms regarding the limitations of the IWRM concept and how to implement it in practice. River basins follow hydrological boundaries and therefore frequently cut across administrative boundaries, including national borders (Butterworth et al. 2010). This can create tensions between hydrological and institutional logics and create cooperation challenges. In addition, achieving effective participation that leads to greater inclusiveness is challenging because, even with the best intentions, it requires significant resources and time commitments from participants, as well as expert knowledge to follow and intervene in discussions (Brandeler et al. 2014; Butterworth et al. 2010). Similarly, the multiple integration goals (i.e., land/water, surface/groundwater, human/nature, upstream/downstream) are unrealistic (Biswas 2008), in particular in contexts of uncertainty, insufficient data, low political will, and limited financial resources (Agyenim 2011; Molle 2008; Watson 2004; Allan 2003). Moreover, the principle of water as an economic good can lead to it being treated as a commodity, at the expense of its nonmonetary value (Rahaman and Varis 2005). IWRM is often depoliticized, even though it is fundamentally a political process, as it involves choices such as the allocation of water between competing users (Jønch-Clausen and Fugl 2001). Despite the legitimacy and mandates that they are given, RBOs are faced with the traditional powerful line agencies, such as irrigation and agriculture authorities, which create obstacles for these organizations to assert themselves as sites of decision-making (Molle 2008; Wester and Warner 2002).

A search was conducted in ScienceDirect for IWRM and other concepts discussed below (Adaptive Water Management or AWM, Water–Energy–Food Nexus, Integrated or Sustainable Urban Water Management or SUWM) in titles, abstracts, and key words for the period 1980–2018. Figure 1 illustrates the sharp rise in publications for all these concepts since the turn of the 21st century. The growing recognition of IWRM’s limitations may explain the concept’s relative loss of ground to other approaches in the past decade.

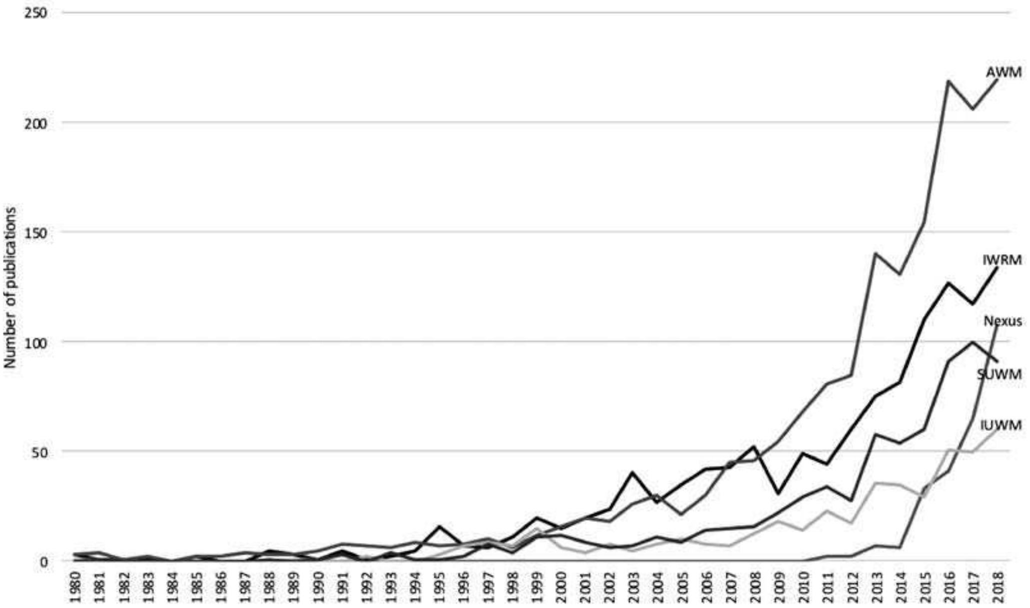


FIGURE 1 Evolution of water resources management approaches within academic literature.

Adaptive Management

AM can be traced back to the late 1970s and is based on the recognition that, although interactions between people and ecosystems can be unpredictable, it is necessary to take management action (Medema et al. 2008; Johnson 1999; Walters and Hilborn 1978; Holling 1978). It arose in a context of rising environmental challenges and public pressure to address these and aimed to support governments and policy makers through new conceptualizations (Hurlbert 2016). AM is a systematic, iterative process, centered on a learning model to not just cope with complex water-related challenges but to learn from them and improve management policies (Medema et al. 2008; Pahl-Wostl 2008). It should enable experimentation and comparison between selected policies and practices (Medema et al. 2008). As IWRM, AM therefore promotes some degree of institutional reform. However, IWRM does not clarify how implementation works within a context of uncertainty nor does it refer explicitly to adaptive capacity (Agyenim 2011; Butterworth et al. 2010). While enhancing democratic legitimacy and contextual responses, the participatory and deliberative forums of IWRM may be unequipped to deal with rapid changes (Brandeler et al. 2018).

AM is not exclusively centered around water resources and the term ‘Adaptive Water Management’ (AWM) is sometimes used to designate an explicit focus on these. As climate change and other global changes impact hydrological systems, the need for adaptive capacity and flexibility, AWM is increasingly recognized, as it allows for speeding up the learning cycle and faster assessments and responses (Pahl-Wostl, Craps, et al. 2007). AM encourages networked forms of governance and participatory processes that enable shared knowledge production and knowledge exchange between sectors and institutions (Pahl-Wostl, Craps, et al. 2007). This can foster social learning – understood as a process of collaborative learning through regular interaction between stakeholder groups – and thereby facilitate joint solutions, leading to more adaptive governance. Social learning processes can be strengthened by building trust and social capital and by providing space for out-of-the-box thinking and experiments

(Pahl-Wostl, Sendzimir, et al. 2007). In addition, more flexible and diverse management measures can reduce vulnerability to unpredictable shocks. AM is increasingly being complemented by adaptive governance, which includes the study of multiple social structures, interconnected resources and actors, and promotes resilience across different scales (Hurlbert and Gupta 2016). This includes measures such as scenario-building and strategic planning to prepare for a range of possible futures.

As with IWRM, AM has been described as difficult to implement and sustain in practice, especially in terms of attaining broad collaboration and the inclusion of stakeholders beyond experts and authorities (Huitema et al. 2009; Medema et al. 2008). The concept has multiple and ambiguous definitions that make it difficult for managers to apply (Medema et al. 2008). Adaptation can also be at odds with water management institutions, which tend to be path dependent and resistant to change due to institutional and technological 'lock-in' effects (Marlow et al. 2013). The large sunk costs and long life cycle of legacy infrastructure disincentivize innovative alternatives (Pahl-Wostl 2007). Another challenge of AM is that it assumes that current management processes are the cause of water resources problems and that these can be untangled, and it ignores the inherent political risks with making management 'experiments' (Medema et al. 2008).

Water–Energy–Food Nexus

In the context of water management, the Nexus approach refers to a set of interrelated activities – typically surrounding water, energy, and the environment – and their linkages as well as the boundary around them, providing a particular frame with which to address water-related problems (Muller 2015; Lofman et al. 2002). The origins of this approach can be traced back to the 1970s. The 1977 UN Water Conference in Mar del Plata highlighted the multiple linkages and interactions between water, food, and energy and the need for a coherent approach to multipurpose water resources development (Muller 2015). However, the Nexus approach gained little traction during the 1980s – seen as a lost decade for international water policy – and was further sidelined by the emergence of IWRM in the 1990s (Scheumann and Klaphake 2001). A search in ScienceDirect for the term 'Water–Energy–Food Nexus' in titles, abstracts, and/or keywords reveals that publications were virtually nonexistent before 2010 but have since grown exponentially (see Figure 1). The Nexus Conference in Bonn in 2011 became the first internationally recognized event on this theme, and it has since been promoted by the EU, the German Development Cooperation, and the United Nations in its SDGs (Benson et al. 2015). This sudden popularity has been described as a response to disillusionment with the IWRM paradigm (Muller 2015).

Through the Nexus approach, water is seen as a crosscutting issue rather than a sector or specific issue area (Gupta et al. 2013). Discussions on 'virtual water' or embodied or embedded water and the role of the food trade in water-scarce regions first highlighted the linkages between water and food (Allan 1998). The focus on the relationship between water and energy arose separately, in particular, regarding electrification in rural areas and groundwater overuse (Shah 2009). Similar to the IWRM approach, the Nexus approach aims for holistic and coordinated management between different aspects of water systems, and it promotes better resource use to allow societies to develop in an environmentally, socially, and economically sustainable way (Benson et al. 2015; Hoff 2011). This is also based on cooperation between actors and citizen participation (Hoff 2011). However, it differs from IWRM, which is a water-centric approach, by granting equal importance to different sectors (Benson et al. 2015). It considers these as inextricably interlinked and seeks to address externalities across them (e.g., energy intensity of desalinization), reduce trade-offs, and increase system efficiency (Olsson 2013; Hoff 2011). In addition, rather than a focus on the river basin with an overarching centralized approach for national policy, the Nexus is implemented in 'problemsheds' where boundaries are drawn around a set of interrelated activities that take place at multiple scales (Muller 2015; Benson et al. 2015; Rouillard et al. 2014).

Criticism of the Nexus approach includes its limited emphasis on integrating water resources management with broader social and economic development activities and its lack of specific emphasis on nature, ethics, power relations, public policies, and other issues (Muller 2015; Leck et al. 2015). This may be a reflection of its focus on cross-sectoral policy coherence rather than on more normative goals, but this also depoliticizes the challenges surrounding water resources management. In addition, implementation of the Nexus approach is challenging at the institutional level, as sectoral decisions often take place at different levels (i.e., national, provincial, local) and remain decoupled (Benson et al. 2015; Scott et al. 2011). Overall, there is a burgeoning literature operationalizing this concept at multiple levels of governance, each using its own approaches and methods.

Water Management in an Increasingly Urban World

The dynamics of a growing world population is accompanied by increasing rural–urban migration with 68% of the population expected to be concentrated in cities by 2050 (UN-DESA 2018). There is a short window of opportunity as many of these cities are expected to be expanded and reach maturity in the next 20–30 years (UNE 2019). The infrastructure that is developed for water is thus critical for the future. Increasing urbanization leads to the concentration of water demand and pressure on water resources in cities' rural hinterlands (ARUP 2018). The economic development often associated with urban growth further increases per capita water use (McDonald et al. 2014). This pressure on water resources can lead to the over-exploitation of aquifers, reliance on inter-basin transfers, and conflicts between users. Urbanization causes water quality degradation due to drastic interferences in ecosystems and the hydrological cycle, including from the inadequate disposal of sewage and solid waste (Azzam et al. 2014). This aggravates regional water stress by reducing water available for use, which tends to disproportionally affect marginalized communities (Varis et al. 2006). Rapid urbanization combined with inadequate urban planning also increases vulnerability to extreme weather events, as populations move into risk-prone areas (UCLG 2016). Urban dwellers are particularly exposed to climate change as many cities are located on river banks and in coastal or flood-prone areas (Varis et al. 2006). It is expected that 40% of urban land will be in high-frequency flood zones in 2030, up from 30% in 2000 (Güneralp et al. 2015).

IWRM may not be an adequate framework to address the challenges of water in and around cities. Although it aims to coordinate between all users, it does not specifically clarify how to address urban water challenges (Brandeler et al. 2018). Cities, and in particular large metropolitan areas, are characterized by complex institutional landscapes with multiple jurisdictions and differing or even conflicting interests. This can lead to fragmented decision-making; a challenge for IWRM (i.e., integrated management and participation), Adaptive Management and Governance (i.e., social learning through collaborative processes), or the Nexus approach (i.e., dominance of urban interests over those of other sectors). Moreover, there are often important power imbalances between urban and rural areas, which affect users' ability to access water resources (Swyngedouw et al. 2002). In heavily urbanized basins where water resources are relatively scarce, competition and conflict between urban and rural users may arise due to water resources allocation regimes (Molle and Mamanpoush 2012).

City-centric frameworks, such as Integrated Urban Water Management (IUWM) and SUWM, have emerged, with roots in IWRM (Bahri 2012). IUWM promotes a holistic view of the urban water cycle, involving stakeholder participation and a flexible mode of strategic planning (Closas et al. 2012; Varis et al. 2006; Brown 2005). SUWM frames water as central to sustainability in cities and a potential starting point for urban planning (Daigger 2011). It aims to combine large and centralized infrastructure with alternative, decentralized and hybrid systems, such as green infrastructure and water reuse (Closas et al. 2012; Younos 2011; Van de Meene et al. 2011). However, such shifts are often hindered in practice by investment and technological 'lock-in' (Marlow et al. 2013). Institutional inertia, caused by factors such as overcentralization, bureaucratic inefficiencies, and a lack of sustainable finance, further limits the implementation of IUWM and SUWM, in particular as urban water challenges become increasingly large and complex (Brown and Farrelly 2007; Lee 2000). Another issue with

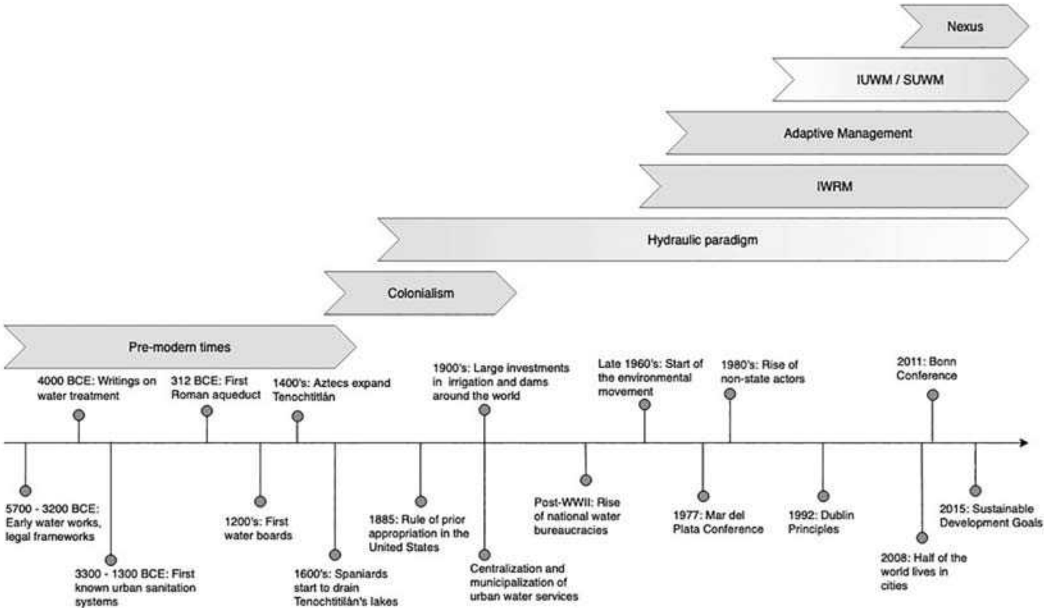


FIGURE 2 Timeline of water resources management approaches.

these frameworks is that they do not address water resources beyond urban boundaries (Brandeler et al. 2018). Ultimately, water management approaches are in constant flux (see Figure 2). As climate change and urbanization increasingly shape water-related challenges, water management approaches will need to adjust to these drivers.

Conclusion

This overview of a few of the dominant approaches to water resources management throughout history highlights society's constant struggle with harnessing the potential of water resources while mitigating its associated risks and challenges. Management approaches have had to adjust to new types of challenges, as population growth has increased water demand and competition between users and sectors and with climate change as a risk-multiplier. In the past three decades, multiple paradigms have emerged. Despite their differences, these approaches shared multiple commonalities, such as decentralization, the integration of issues and sectors, the inclusion of a larger range of actors at multiple levels, and a shift away from exclusively engineering-based solutions towards conservation, efficient water use, and green infrastructure. Around the world, water resources management actors have incorporated good governance principles within laws and policies.

However, institutional change is slow, in particular, due to path dependency and the 'lock-in' effect and as power relations often favor the status quo, leading to a lack of adherence to good governance principles in practice. To support changes in practice, it is crucial that water managers have access to reliable data and that actors across sectors and at multiple levels can effectively share knowledge and information. Decentralization has supported more inclusive and context-relevant decision-making, but effective coordination mechanisms are necessary to avoid fragmentation between actors and incoherent responses. In addition, cash-strapped agencies struggle to carry out their responsibilities, especially in the Global South. This requires not only adequate pricing mechanisms but also political will and a long-term perspective from leaders seeking fast results.

There is an urgent need to preempt possible future damage through better institutional and infrastructural design today. Although there are no panaceas to address complex water resources challenges, lessons can be learned from the advantages and limitations of past and present management approaches.

References

Abdullah, D. and B. Christensen. 2004. “Integrated River Basin Management.” *Bulletin Ingenieur* 22: 21–23.

Abers, R. and M. E. Keck. 2013. “Practical Authority: Agency and Institutional Change in Brazilian Water Politics –Excerpts.” In American Political Science Association 2013 Annual Meeting. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2302112.

Abers, R. N. 2007. “Organizing for Governance: Building Collaboration in Brazilian River Basins.” *World Development* 35 (8): 1450–63. doi:10.1016/j.worlddev.2007.04.008.

Agyenim, J. B. 2011. “Investigating Institutional Arrangements for Integrated Water Resource Management in Developing Countries: The Case of White Volta Basin, Ghana.” PhD Thesis Vrije Universiteit Amsterdam, Amsterdam.

Alam, U., P. Sahota, and P. Jeffrey. 2007. “Irrigation in the Indus Basin: A History of Unsustainability?” *Water Science and Technology: Water Supply* 7 (1): 211–18. doi:10.2166/ws.2007.024.

Allan, T. 1998. “Watersheds and Problemsheds: Explaining the Absence of Armed Conflict Over Water in the Middle East.” *Middle East Review of International Affairs* 2 (1): 49–51.

Allan, T. 2003. “IWRM/IWRAM: A New Sanctioned Discourse?” https://www.soas.ac.uk/water/publications/papers/file38393.pdf.

Anand, N. 2011. “Pressure: The Polytechnics of Water Supply in Mumbai.” *Cultural Anthropology*. doi:10.1111/j.1548-1360.2011.01111.x.

Angelakis, A. N. and X. Y. Zheng. 2015. “Evolution of Water Supply, Sanitation, Wastewater, and Stormwater Technologies Globally.” *Water (Switzerland)* 7 (2): 455–63. doi:10.3390/w7020455.

ARUP. 2018. “Cities Alive: Water for People.” Leeds, UK.

Avissar, R. and D. Werth. 2005. “Global Hydroclimatological Teleconnections Resulting from Tropical Deforestation.” *Journal of Hydrometeorology* 6: 134–45. https://journals.ametsoc.org/doi/pdf/10.1175/JHM406.1.

Azzam, R., R. Strohschön, K. Baier, L. Lu, K. Wiethoff, A. Bercht, and R. Wehrhahn. 2014. “Water Quality and Socio-Ecological Vulnerability Regarding Urban Development in Selected Case Studies of Megacity Guangzhou, China.” In *Megacities: Our Global Urban Future*, edited by F. Kraas, M. Aggarwal, and G. Mertins. Berlin: Springer.

Bahri, A. 2012. *Integrated Urban Water Management: Briefing Note*. Technical Committee Background Paper 16. Stockholm: Global Water Partnership.

Benson, D., A. K. Gain, and J. J. Rouillard. 2015. “Water Governance in a Comparative Perspective: From IWRM to a ‘nexus’ Approach?” *Water Alternatives* 8 (1). doi:10.2106/00004623-198769080-00008.

Bhattacharya, U. 2017. “From Surveys to Management: The Early Colonial State’s Intervention in Water Resources of Bengal.” *Indian Historical Review* 44 (2): 225–51. doi:10.1177/0376983617726471.

Biswas, A. K. 2008. “Integrated Water Resources Management: Is It Working?” *International Journal of Water Resources Development* 24 (1): 5–22. doi:10.1080/07900620701871718.

Boelens, R. 2009. “The Politics of Disciplining Water Rights.” *Development and Change* 40 (2): 307–31. doi:10.1111/j.1467–7660.2009.01516.x.

Brandeler, F., J. Gupta, and M. Hordijk. 2018. “Megacities and Rivers: Scalar Mismatches between Urban Water Management and River Basin Management.” *Journal of Hydrology*. doi:10.1016/j.jhydrol.2018.01.001.

Brandeler, F., M. Hordijk, K. von Schönfeld, and J. Sydenstricker-Neto. 2014. “Decentralization, Participation and Deliberation in Water Governance: A Case Study of the Implications for Guarulhos, Brazil.” *Environment and Urbanization* 26 (2): 489–504. doi:10.1177/0956247814544423.

Brown, R. R. 2005. “Impediments to Integrated Urban Stormwater Management: The Need for Institutional Reform.” *Environmental Management* 36 (3): 455–68. doi:10.1007/s00267-004-0217-4.

Brown, R. R. and M. Farrelly. 2007. “Barriers to Advancing Sustainable Urban Water Management: A Typology.” *Rainwater and Urban Design*, 229–36. http://search.informit.com.au.ezproxy.lib.monash.edu.au/documentSummary;dn=888968667496813;res=IELENG.

Burke, W. J. 1956. “The Origin, Growth and Function of the Law of Water Use.” *Wyoming Law Journal* 10 (2). http://repository.uwyo.edu/wljAvailableat:http://repository.uwyo.edu/wlj/vol10/iss2/1.

Butterworth, J., J. Warner, P. Moriarty, S. Smits, and C. Batchelor. 2010. “Finding Practical Approaches to Integrated Water Resources Management.” *Water Alternatives* 3 (31): 68–81. www.water-alternatives.org.

Closas, A., M. Schuring, and D. Rodriguez. 2012. *Integrated Urban Water Management - Lessons and Recommendations from Regional Experiences in Latin America, Central Asia, and Africa*. Washington, DC: World Bank.

Cullet, P. and J. Gupta. 2009. “India: Evolution of Water Law and Policy.” In *The Evolution of the Law and Politics of Water*, 157–73. Dordrecht: Springer. doi:10.1007/978-1-4020-9867-3_10.

Custódio, V. 2012. *Escassez de Água e Inundações Na Região Metropolitana de São Paulo*. São Paulo: Humanitas/Fapesp.

D’Souza, R. 2006. “Water in British India: The Making of a ‘Colonial Hydrology.’” *History Compass* 4 (4): 621–28.

Daigger, G. T. 2011. “Sustainable Urban Water and Resource Management.” *The Bridge* 41 (1): 13–18.

Domènech, L. 2011. “Rethinking Water Management: From Centralised to Decentralised Water Supply and Sanitation Models.” *Documents d’Anàlisi Geogràfica* 57 (2): 293–310.

Dryzek, J. 1997. *The Politics of the Earth: Environmental Discourses*. Oxford: Oxford University Press.

Dzwairo, B., F. A. O. Otieno, and G. M. Ochieng. 2010. “Making a Case for Systems Thinking Approach to Integrated Water Resources Management (IWRM).” *International Journal of Water Resources and Environmental Engineering* 1 (5): 107–13. http://www.academicjournals.org/ijwree.

Elmqvist, T., M. Fragkias, J. Goodness, B. Güneralp, P. J. Marcotullio, R. I. Mcdonald, S. Parnell, et al., eds. 2013. *Urbanization, Biodiversity and Ecosystem Services: Challenges and Opportunities. A Global Assessment.* Dordrecht: Springer.

FAO. 2012. “Coping with Water Scarcity: An Action Framework for Agriculture and Food Security.” Rome. http://www.fao.org.

Gandy, M. 2008. “Landscapes of Disaster: Water, Modernity, and Urban Fragmentation in Mumbai.” *Environment and Planning A* 40: 108–30. doi:10.1068/a3994.

GRID. 2018. *Global Report on Internal Displacement*. Geneva, Switzerland: Internal Displacement Monitoring Centre.

Grigg, N. S. 2008. “Integrated Water Resources Management: Balancing Views and Improving Practice.” *Water International* 33 (3): 1–20. doi:10.1080/02508060802272820.

Güneralp, B., İ. Güneralp, and Y. Liu. 2015. “Changing Global Patterns of Urban Exposure to Flood and Drought Hazards.” *Global Environmental Change* 31 (March): 217–25. doi:10.1016/J.GLOENVCHA.2015.01.002.

Gupta, J. and J. W. Dellapenna. 2009. “The Challenges for the Twenty-First Century: A Critical Approach.” In *The Evolution of the Law and Politics of Water,* edited by J. W. Dellapenna and J. Gupta, 391–410. Dordrecht: Springer. doi:10.1007/978-1-4020-9867-3_23.

Gupta, J. and C. Pahl-Wostl. 2013. “Global Water Governance in the Context of Global and Multilevel Governance: Its Need, Form, and Challenges.” *Ecology and Society* 18 (4): art53. doi:10.5751/ES-05952-180453.

Gupta, J., C. Pahl-Wostl, and R. Zondervan. 2013. “‘Glocal’ Water Governance: A Multi-Level Challenge in the Anthropocene.” *Current Opinion in Environmental Sustainability* 5 (6): 573–80. doi:10.1016/j.cosust.2013.09.003.

GWP-TAC. 2000. “Integrated Water Resources Management.” Stockholm. www.gwpforum.org.

Hanak, E., J. Lund, A. Dinar, B. Gray, R. Howitt, J. Mount, P. Moyle, and B. "Buzz" Thompson. 2011. *Managing California's Water from Connict to Reconciliation*. San Francisco, CA: Public Policy Institute of California. https://www.ppic.org/content/pubs/report/R_211EHR.pdf.

Hatfield-Dodds, S., R. Nelson, and D. C. Cook. 2007. "Adaptive Governance: An Introduction, and Implications for Public Policy." In ANZSEE Conference. Noosa Australia, 4–5 July 2007. https://pdfs.semanticscholar.org/8c29/116745551496ae166290acd780b2c988e21d.pdf.

Hoff, H. 2011. "Understanding the Nexus. Background Paper for the Bonn 2011 Nexus Conference: The Water, Energy and Food Security Nexus." https://mediamanager.sei.org/documents/Publications/SEI-Paper-Hoff-UnderstandingTheNexus-2011.pdf.

Holling, C. S., ed. 1978. *Adaptive Environmental Assessment and Management*. Chichester: Wiley.

Hooper, B. 2005. *Integrated River Basin Governance: Learning from International Experience*. Water Intelligence Online. London: IWA Publishing. doi:10.2166/9781780402970.

Huitema, D., and S. Meijerink. 2017. "The Politics of River Basin Organizations: Institutional Design Choices, Coalitions, and Consequences." *Ecology and Society* 22 (2). doi:10.5751/ES-09409-220242.

Huitema, D., E. Mostert, W. Egas, S. Moellenkamp, C. Pahl-Wostl, and R. Yalcin. 2009. "Adaptive Water Governance: Assessing the Institutional Prescriptions of Adaptive (Co-)Management from a Governance Perspective and Defining a Research Agenda." *Ecology and Society* 14 (1): 26.

Hurlbert, M. 2016. *Adaptive Governance of Disaster: Drought and Flood in Rural Areas*. Amsterdam: Springer.

Hurlbert, M., and J. Gupta. 2016. "Adaptive Governance, Uncertainty, and Risk: Policy Framing and Responses to Climate Change, Drought, and Flood." *Risk Analysis* 36 (2): 339–56. doi:10.1111/risa.12510.

ICWE. 1992. *The Dublin Statement on Water and Sustainable Development*. Dublin: International Conference on Water and the Environment. http://www.wmo.int/pages/prog/hwrp/documents/english/icwedece.html.

IPCC. 2014. "Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, RK Pachauri and LA Meyer (Eds.)]." Geneva, Switzerland.

Johnson, B. L. 1999. "The Role of Adaptive Management as an Operational Approach for Resource Management Agencies." *Conservation Ecology* 3 (2): 8. doi:10.5751/ES-00136-030208.

Jønch-Clausen, T., and J. Fugl. 2001. "Firming up the Conceptual Basis of Integrated Water Resources Management." *International Journal of Water Resources Development* 174: 501–10. doi:10.1080/07900620120094055.

Kallis, G., and H. Coccossis. 2002. "Water for the City: Lessons from Tendencies and Critical Issues in Five Advanced Metropolitan Areas." *Built Environment* 28 (2): 96–110.

Kasemir, B., J. Jäger, C. Jaeger, and M. Gardner, eds. 2003. *Public Participation in Sustainability Science: A Handbook*. Cambridge: Cambridge University Press.

Kornfeld, I. E. 2009. "Mesopotamia: A History of Water and Law." In *The Evolution of the Law and Politics of Water*, 21–36. Dordrecht: Springer. doi:10.1007/978-1-4020-9867-3_2.

Lazaroms, R., and D. Poos. 2004. "The Dutch Water Board Model." *Water Law* 15 (3/4): 137–40. http://www.fao.org/tempref/agl/emailconf/wfe2005/Art11_Lazaroms137-140.pdf.

Leck, H., D. Conway, M. Bradshaw, and J. Rees. 2015. "Tracing the Water-Energy-Food Nexus: Description, Theory and Practice." *Geography Compass* 9 (8): 445–60. doi:10.1111/gec3.12222.

Lee, T. R. 2000. "Urban Water Management for Better Urban Life in Latin America." *Urban Water* 2 (1): 71–78. www.elsevier.com/locate/urbwat.

Leite Farias, P. J. 2009. "Brazil: The Evolution of the Law and Politics of Water." In *The Evolution of the Law and Politics of Water*, edited by J. W. Dellapenna and J. Gupta, 69–86. Dordrecht: Springer.

Lofman, D., M. Petersen, and A. Bower. 2002. “Water, Energy and Environment Nexus: The California Experience.” *International Journal of Water Resources Development* 18 (1): 73–85. doi:10.1080/07900620220121666.

Lopez-Gunn, E. 2009. “Agua Para Todos: A New Regionalist Hydraulic Paradigm in Spain.” *Water Alternatives* 2 (3): 370–94. www.water-alternatives.org.

Lucero, L. J., J. D. Gunn, V. L. Scarborough, L. J. Lucero, J. D. Gunn, and V. L. Scarborough. 2011. “Climate Change and Classic Maya Water Management.” *Water* 3 (2): 479–94. doi:10.3390/w3020479.

Marlow, D. R., M. Moglia, S. Cook, and D. J. Beale. 2013. “Towards Sustainable Urban Water Management: A Critical Reassessment.” *Water Research* 47: 7150–61. doi:10.1016/j.watres.2013.07.046.

Martinez-Santos, P., M. M. Aldaya, and M. Ramón Llamas. 2014. *Integrated Water Resources Management in the 21st Century: Revisiting the Paradigm*. Leiden: CRC Press/Balkema.

McDonald, R. I., K. Weber, J. Padowski, M. Flörke, C. Schneider, P. A. Green, T. Gleeson, et al. 2014. “Water on an Urban Planet: Urbanization and the Reach of Urban Water Infrastructure.” *Global Environmental Change* 27 (1). doi:10.1016/j.gloenvcha.2014.04.022.

Medema, W., B. S. Mcintosh, and P. J. Jeffrey. 2008. “From Premise to Practice: A Critical Assessment of Integrated Water Resources Management and Adaptive Management Approaches in the Water Sector.” *Ecology and Society* 13 (2): 29. http://www.ecologyandsociety.org/vol13/iss2/art29/.

Meene, S. J. Van de, R. R. Brown, and M. A. Farrelly. 2011. “Towards Understanding Governance for Sustainable Urban Water Management.” *Global Environmental Change* 21 (3): 1117–27. doi:10.1016/j.gloenvcha.2011.04.003.

Molle, F. 2008. “Nirvana Concepts, Narratives and Policy Models: Insights from the Water Sector.” *Water Alternatives* 1 (1): 131–56.

Molle, F. 2009. “River-Basin Planning and Management: The Social Life of a Concept.” *Geoforum* 40: 484–94. doi:10.1016/j.geoforum.2009.03.004.

Molle, F., and A. Mamanpoush. 2012. “Scale, Governance and the Management of River Basins: A Case Study from Central Iran.” *Geoforum* 43 (2): 285–94. doi:10.1016/j.geoforum.2011.08.004.

Molle, F., P. P. Mollinga, and P. Wester. 2009. “Hydraulic Bureaucracies and the Hydraulic Mission: Flows of Water, Flows of Power.” *Water Alternatives* 2 (3): 328–49. www.water-alternatives.org.

Muller, M. 2015. “The ‘Nexus’ As a Step Back towards a More Coherent Water Resource Management Paradigm.” *Water Alternatives* 8 (1): 675–94. www.water-alternatives.org.

Nilsson, D., and E. N. Nyanchaga. 2009. “East African Water Regimes: The Case of Kenya.” In *The Evolution of the Law and Politics of Water*, edited by J. W. Dellapenna and J. Gupta, 105–20. Dordrecht: Springer. doi:10.1007/978-1-4020-9867-3_7.

Olsson, G. 2013. “Water, Energy and Food Interactions—Challenges and Opportunities.” *Frontiers of Environmental Science & Engineering* 7 (5): 787–93. doi:10.1007/s11783-013-0526-z.

Paganini, W. da S. 2008. *A Identidade de Um Rio de Contrastes o Tietê e Seus Múltiplos Usos*. ABES. https://books.google.nl/books/about/A_identidade_de_um_rio_de_contrastes.html?id=TF3JQwAACAAJ&redir_esc=y.

Pahl-Wostl, C. 2007. “Transitions towards Adaptive Management of Water Facing Climate and Global Change.” *Water Resource Management* 21: 49–62. doi:10.1007/s11269-006-9040-4.

Pahl-Wostl, C. 2008. “Requirements for Adaptive Water Management.” In *Adaptive and Integrated Water Management*, edited by C. Pahl-Wostl, P. Kabat, and J. Möltgen, 1–22. Berlin: Springer. doi:10.1007/978-3-540-75941-6_1.

Pahl-Wostl, C., M. Craps, A. Dewulf, E. Mostert, D. Tabara, and T. Taillieu. 2007. “Social Learning and Water Resources Management.” *Ecology and Society* 12 (2): 5.

Pahl-Wostl, C., J. Sendzimir, P. Jeffrey, J. Aerts, G. Berkamp, and K. Cross. 2007. “Managing Change toward Adaptive Water Management through Social Learning.” *Ecology and Society* 12 (2): 18. https://www.jstor.org/stable/26267877.

Porse, E. C. 2014. "Old Solutions and New Problems: On the Evolution of Urban Water Infrastructure and Environments." University of California, Davis. https://watershed.ucdavis.edu/shed/lund/students/PorseDissertation.pdf.

Rahaman, M. M., and O. Varis. 2005. "Integrated Water Resources Management: Evolution, Prospects and Future Challenges." *Sustainability: Science, Practice, & Policy* 1 (1). http://ejournal.nbii.org.

Rietveld, L. C., J. G. Siri, I. Chakravarty, A. M. Arsénio, R. Biswas, and A. Chatterjee. 2016. "Improving Health in Cities through Systems Approaches for Urban Water Management." *Environmental Health: A Global Access Science Source* 15 (Suppl 1). doi:10.1186/s12940-016-0107-2.

Rouillard, J. J., D. Benson, and A. K. Gain. 2014. "Evaluating IWRM Implementation Success: Are Water Policies in Bangladesh Enhancing Adaptive Capacity to Climate Change Impacts?" *International Journal of Water Resources Development* 30 (3): 515–27. doi:10.1080/07900627.2014.910756.

Savenije, H. H G, and P. Van der Zaag. 2008. "Integrated Water Resources Management: Concepts and Issues." *Physics and Chemistry of the Earth* 33: 290–97. doi:10.1016/j.pce.2008.02.003.

Scheumann, W., and A. Klaphake. 2001. "Freshwater Resources and Transboundary Rivers on the International Agenda: From UNCED to Rio+10." Bonn, Germany. https://www.ecolex.org/fr/details/literature/freshwater-resources-and-transboundary-rivers-on-the-international-agenda-from-unced-to-rio10-mon-067325/.

Scott, C. A., S. A. Pierce, M. J. Pasqualetti, A. L. Jones, B. E. Montz, and J. H. Hoover. 2011. "Policy and Institutional Dimensions of the Water–Energy Nexus." *Energy Policy* 39 (10): 6622–30. doi:10.1016/j.enpol.2011.08.013.

Shah, T. 2009. *Taming the Anarchy: Groundwater Governance in South Asia.* Colombo: Resources for the Future. https://cgspace.cgiar.org/handle/10568/36566.

Smythe, E. A. 1905. *The Conquest of Arid America*. 2nd ed. Seattle, WA: University of Washington Press.

Sosa-Rodriguez, F. S. 2010. "Impacts of Water-Management Decisions on the Survival of a City: From Ancient Tenochtitlan to Modern Mexico City." *International Journal of Water Resources Development* 26 (4): 675–87. doi:10.1080/07900627.2010.519503.

Swyngedouw, E. 1999. "Modernity and Hibridity: Nature, Regeneracionismo, and the Production of the Spanish Waterscape, 1890–1930." *Annals of the Association of American Geographers* 89: 443–65.

Swyngedouw, E. 2005. "Governance Innovation and the Citizen: The Janus Face of Governance-beyond-the-State." *Urban Studies* 42 (11). doi:10.1080/00420980500279869.

Swyngedouw, E., M. Kaïka, and E. Castro. 2002. "Urban Water: A Political-Ecology Perspective." *Built Environment* 28 (2): 124–37. https://www.jstor.org/stable/pdf/23288796.pdf?refreqid=excelsior%3A33d4175e87e690aec72bb0ffb516a88b.

UCLG. 2016. "Co-Creating the Urban Future: The Agenda of Metropolises, Cities and Territories." Barcelona.

UN-DESA. 2018. "World Urbanization Prospects – Population Division - United Nations." https://population.un.org/wup/Download/.

UNE. 2019. "Global Environment Outlook 6." Nairobi, Kenya.

UNEP. 2012. "The UN-Water Status Report on the Application of Integrated Approaches to Water Resources Management." Nairobi, Kenya.

UNESCO. 2006. *Water, a Shared Responsibility. The United Nations World Water Report 2*. Paris and New York: UNESCO and Berghahn Books.

UNGA, (United Nations General Assembly). 2015. "2015 Time for Global Action for People and Planet." New York. https://sustainabledevleopment.un.org/%0Apost2015.

Varis, O., A. K. Biswas, C. Tortajada, and J. Lundqvist. 2006. "Megacities and Water Management." *International Journal of Water Resources Development* 22 (2): 377–94. doi:10.1080/07900620600684550.

Vörösmarty, C. J., S. E. Bunn, and R. Lawford. 2013. "Global Water, the Anthropocene and the Transformation of a Science." *Current Opinion in Environmental Sustainability* 5 (6): 539–50. doi:10.1016/J.COSUST.2013.10.005.

Wallington, T. J., K. Maclean, T. Darbas, and C. J. Robinson. 2010. "Knowledge-Action Systems for Integrated Water Management: National and International Experiences, and Implications for South East Queensland." http://www.griffith.edu.au/.

Walters, C. J., and R. Hilborn. 1978. "Ecological Optimization and Adaptive Management." *Annual Review of Ecology and Systematics* 9 (1): 157–88. doi:10.1146/annurev.es.09.110178.001105.

Watson, N. 2004. "Integrated River Basin Management: A Case for Collaboration." *International Journal of River Basin Management* 2 (4): 243–57. doi:10.1080/15715124.2004.9635235.

Wester, P., and J. Warner. 2002. "River Basin Management Reconsidered." In *Hydropolitics in the Developing World: A Southern African Perspective*, edited by A. Turton and R. Henwood, 61–71. African Water Issues Research Unit, Centre for International Political Studies, University of Pretoria. https://library.wur.nl/WebQuery/wurpubs/123299.

Wilson, A. 1998. "Water Supply in Ancient Carthage." In *Carthage Papers: The Early Colony's Economy, Water Supply, a Public Bath and the Mobilization of State Olive Oil*, edited by J. T. Pena, J. J. Rossiter, A. I. Wilson, C. Wells, M. Carroll, J. Freed and D. Godden, 65–102. Portsmouth: Rhode Island.

Wittfogel, K. A. 1955. "Developmental Aspects of Hydraulic States." In *Irrigation Civilizations: A Comparative Study*. Washington, DC: Pan-American Union, Social Science Monographs.

WWAP. 2009. "Third World Water Development Report, Preface Xii."

Xie, M. 2006. "Integrated Water Resources Management (IWRM)-Introduction to Principles and Practices 1." http://pacificwater.org/userfiles/file/IWRM/Toolboxes/introductiontoiwrm/IWRM Introduction.pdf.

Younos, T. 2011. "Paradigm Shift: Holistic Approach for Water Management in Urban Environments." *Frontiers of Earth Science*. doi:10.1007/s11707-011-0209-7.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

24

Wastewater and Water Utilities

Rudolf Marloth

Introduction

Municipalities are normally responsible for providing water and disposing of wastewater, either by privately or publicly owned utilities. A water utility is responsible for the source, transmission, treatment, and distribution of potable water. Water usage is metered at the user. A wastewater treatment utility is responsible for collection, treatment, and disposal. Industrial wastewater is often metered. Residential wastewater charges are based on water usage.

Municipalities are also responsible for the collection and disposal of runoff, which consists of stormwater, misplaced or excessive irrigation, domestic car washing, etc. Runoff collects in gutters or holding ponds; then goes to storm sewers; and is dumped into a nearby river or, in coastal cities, directly into the ocean. This water carries with it a variety of pollutants: fertilizer, animal waste, oils, tire dust, etc. For municipalities that do not have combined storm and sanitary sewers, this runoff is dumped without treatment. Worse, in severe storms, combined sewer systems are likely to dump runoff combined with untreated sewage. Furthermore, runoff is not metered, so the municipality bears the cost of installing and maintaining the infrastructure. Possible approaches to reducing pollution of natural waters are (a) enactment of ordinances to make property owners responsible for reducing runoff and its pollutants, (b) diverting light flow or the first flow in a storm to a wastewater treatment plant (WWTP), and (c) installing treatment facilities to make runoff suitable for reuse. (The city of Santa Monica, California recently demonstrated its SMURRF (Santa Monica Urban Runoff Recycling Facility), the first of its kind in the nation.)

Water

The primary objective of water treatment is to make it safe for human consumption at a reasonable cost. It is possible to produce safe water that has objectionable taste, odor, or color, so a secondary goal is to make the water appealing to the consumer. Turbidity and color are qualities apparent to the naked eye.

The former is caused by particles in suspension and is measured in nephalometric turbidity units (NTU) by passing light through a sample. These particles can be removed by settling or filtering. Colloidal particles will not settle in a reasonable amount of time and must be removed by other physical processes. Dissolved substances are removed or transformed by chemical treatment. Color may be caused either by materials in solution (true color) or in suspension (apparent color).[1]

History

Early water treatment focused on what was apparent to the senses: appearance, taste, and odor. These qualities were improved by removing turbidity through filtration or precipitation. It was found much later that particles in the water harbored pathogens, which were largely removed while clarifying the water.

A slow sand filter (SSF) designed by James Simpson was commissioned in 1829, but it was some time before its full importance was realized. This simple device is essentially a tank filled with sand with water introduced at the top and removed at the bottom. Sometime after the filter is started up (a few days to a few weeks), the upper layer of sand becomes coated with a gelatinous biological layer called the schmutzedecke, made up of algae, bacteria, protozoa, and small invertebrates. This sticky layer is biologically active and converts organic matter in the water to water, carbon dioxide, and harmless salts (i.e., it is mineralized). Later, research showed the importance of biological removal and also showed that the SSF was very effective in that respect. In particular, the SSF is effective in removing *Giardia* and *Cryptosporidium* oocysts, which are nearly unaffected by chlorination.[1,2]

In the famous Broad Street pump episode of 1854, an outbreak of cholera in the Soho district of London killed more than 600 people. Dr. John Snow, who had theorized that the disease was spread by contaminated water, traced it to water from the Broad Street pump. The likely cause was from a leaking and cholera-infested cesspool located only three feet from the Broad Street well. In fact, cesspits lay under many of the houses in the district.[3] Acceptance of Snow's theory was slow, as was conversion to a sewer system that conveyed wastewater to central plants.

Recently, the Centers for Disease Control and Prevention and the National Academy of Engineering named water treatment as one of the most significant public health advancements of the 20th century.[4] This is rightly so, for in the first half of the 20th century, life expectancy in the United States increased dramatically, primarily because of water treatment, which has greatly reduced the incidence of waterborne bacterial infections such as cholera, typhoid fever, and dysentery. Even so, waterborne disease does occur in this country, the most well-known instance of which is the cryptosporidiosis outbreak of 1973 in Milwaukee. Most episodes are due to contamination of raw or treated water, inadequate treatment, and cross-contamination between sewers and water mains.

Standards and Monitoring

The Safe Drinking Water Act of 1974 and its amendments of 1986 and 1996 are the primary pieces of federal legislation protecting drinking water supplied by public water systems. Primary regulations under the act are for the protection of public health; secondary regulations are for regulations pertaining to taste, odor, and appearance.

The Surface Water Treatment Rule mandates that surface water or groundwater under the influence of surface water must be treated to remove or inactivate 99.9% of *Giardia lamblia* cysts and 99.99% of enteric viruses. These requirements are commonly stated as "log removals," where n-log removal is removal of a $1–10^{-n}$ fraction of the pollutant. Treatment processes using filtration are judged to comply by providing an adequate concentration/ contact time (Ct) product, where C is the concentration of the disinfectant in mg/L and t is the time in minutes.[5]

The Enhanced Surface Water Treatment Rule requires a 2-log removal of *Cryptosporidium*. Systems using filtration are granted credit if they meet certain turbidity criteria. The Filter Backwash Recycling

Rule requires treatment plants to recycle filter backwash water through the entire process cycle. The Disinfectants/Disinfection Byproduct Rule establishes maximum contaminant levels (MCLs) on total trihalomethanes, haloacetic acids, bromate ion, and chlorite ion.[5]

The most common way of testing the quality of drinking water is the coliform test, as specified by the U.S. Environmental Protection Agency. Coliforms are bacteria that are gram-negative, aerobic and facultative anaerobic, nonspore-forming rods, which ferment lactose with gas formation in 48 h at 35°C. When a sample tests positive for coliforms, it must be tested for fecal coliforms. Fecal or thermotolerant coliforms include all coliforms that can ferment lactose at 44.5°C. It is common to identify *Escherichia coli* uniquely with fecal coliforms.[5]

Sources and Transmission

Large-scale sources are primarily surface water and groundwater. Desalinated ocean water is not yet a major source in the United States, although Spain is a major producer and user of desalinated water. In the past 30 years, the energy required for desalination has fallen from 12 to 3 or 4 kWh/m^3 using reverse osmosis.[6] In a few areas, recycled water (treated wastewater) is used for irrigation. Water is transmitted by way of natural water courses, lined open channels, or pressure pipe. Los Angeles, for example, receives water from the Sacramento Delta via a concrete-lined open channel, from snowmelt in the Sierra via a natural watercourse that is diverted into an iron pipe aqueduct. New York reservoirs supply water to treatment plants through a series of underground tunnels. In some cases, water treatment plants are located right at the site where water is drawn from a river or lake.

Treatment

Treatment is tailored somewhat to the characteristics of the influent and the effluent limits, but in general, large particles are removed first by screening and then by settling. Next are particles that will float or settle with some assistance, such as air floatation or mixing with coagulants. Colloids and dissolved materials are removed last. The water should be nearly clear before disinfection so that contaminants may not hide in turbidity.

Standard water treatment (Figure 1) consists of coagulation, flocculation (aggregation into a wooly mass), sedimentation, filtration, and disinfection. The size of the units and major equipment are determined by the hydraulic loading. The examples below are based on a plant with an average flow of 24 mgd (million gal per day). Plants are commonly designed for a maximum daily flow that is 1.5 times the average daily flow. Plants whose main supply is groundwater will have a somewhat different process chain because the water will contain less settleable and suspended solids but probably more dissolved metals. Plants that draw from a river should have at their head a coarse screen for tree branches, etc., and a grit chamber for sand and silt, because this kind of grit will cause a serious maintenance problem with pumps.

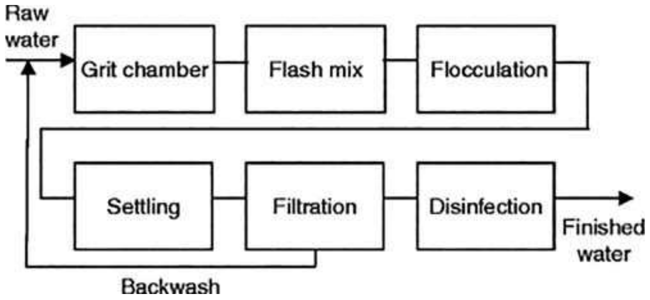


FIGURE 1 Water treatment.

The coagulation operation consists of the addition and flash mixing of chemicals meant to remove the charge on colloids and suspended solids so that they will aggregate in the flocculation step. Depending on the type of chemical used, flash mixing should occur in about one to five seconds. The addition of more than the optimum amount of chemicals can compensate for less-than-optimum mixing. A potential problem in this step is the clogging of feed lines. Various methods of flash mixing are available, the best of which are diffusion mixing by water jets and inline static mixing. Proper operation at this step requires choosing the right chemicals and quantities in response to the raw water quality and flow rate. Chemicals used in this step include aluminum sulfate (alum), polyaluminum chloride, various iron compounds, polymers, and bentonite.[5]

Flocculation is slow mixing that increases the rate of collisions between particles whose electrostatic repulsive forces have been neutralized in the coagulation step. Now the particles will stick together into sizes that will settle. Flocculation mixing is performed by mechanical mixers, baffles, or other methods. Paddle reels with a horizontal axle perform well. In a reasonable design for two parallel, 12-mgd floc tanks, each of three chambers in a tank might be about 13.5 ft^2 in cross section so that a 13-ft horizontal paddle reel would nearly fill a chamber. Perforated baffle walls separate the chambers to promote good mixing. Proper operation of this step requires continual monitoring of and adjusting for the floc size, and removing scum from the surface of the water, sludge from the bottom of the tank, and algae from the vertical surfaces (walls and baffles). Transfer of the flocculated water to the sedimentation tank must be done at low velocity to avoid breaking up the floc.

Four progressive stages of sedimentation are distinguished. Type I sediment consists of separate, destabilized particles. Type II is made of larger groups of flocculated particles. In Type III, the particles have formed a blanket that initiates hindered settling. Type IV settling is compression of the sludge blanket at the bottom of the tank.

A typical filter bed is made up of a layer of sand and a layer of coal, charcoal, or granular activated carbon. The water from the sedimentation tank is introduced at the top of the filter and moves by gravity down through the media to the underdrain. It is most desirable to have the filter backwashed once per day. Thus, the design depends upon the quality of the raw water, the required throughput, the local climate, and the skill level of the operators. The backwash water is required to be recirculated to the head of the plant.

Filter efficiency is determined by the unit filter run volume (UFRV), which is the ratio of the amount of water processed during a filter cycle to the amount that could be processed if no backwashing were necessary. The effective filtration rate, R_e, is

$$R_e = (\text{UFRV} - \text{UBWV})/T$$

where UBWV (unit backwash volume), gal/ft^2; T, filter cycle time, min; and

$$\text{UFRV} = V_f/A \quad \text{UBWV} = V_b/A$$

where V_f, volume filtered per filter cycle, gal; A, area of filter, ft^2; V_f, volume of backwash water, gal.

The design filtration rate, R_d, is the maximum filtration rate, which can be achieved only if no backwash were necessary. Then the production efficiency is

$$R_e/R_d = (\text{UFRV} - \text{UBWV})/\text{UFRBV}$$

Example 1. Find the production efficiency and the filter cycle time for a filter with UFRV equal to 7500 gal/ft^2, UBWV equal to 200 gal/ft^2, and design filtration rate equal to 5 gpm/ft^2.

Solution. ***The effective filtration rate, filter efficiency, and filter cycle time are***

$$\begin{aligned} R_e &= (5\ \text{gpm/ft}^2)\left[(7500\ \text{gal/ft}^2) - (200\ \text{gal/ft}^2)\right]/(7500\ \text{gal/ft}^2) \\ &= 4.87\ \text{gpm/ft}^2 \end{aligned}$$

$$= R_e/R_d\left[\left(7500\ \text{gal/ft}^2\right)-\left(200\ \text{gal/ft}^2\right)\right]/\left(7500\ \text{gal/ft}^2\right)$$

$$= 0.973$$

and

$$T = \left[\left(7500\ \text{gal/ft}^2\right)-\left(200\ \text{gal/ft}^2\right)\right]/\left(4.87\text{gpm/ft}^2\right)$$

$$= 1499\ \text{min}$$

Note that the filter cycle time is very close to one day (1440 min). Also, increasing the filtration rate will not necessarily increase the amount of water filtered per day. Increasing the rate increases the amount of deposition on the filter media, reducing the filter run time.

Most of the water treatment plants in this country disinfect with chlorine. Common alternatives are ozone, chloramine, and ultraviolet light. The disinfectant is sometimes added at the head of the plant to give an adequate concentration-contact time product. When chlorine is used, disinfection byproducts can be formed—most notably, tri- halomethanes. To suppress this, ammonia might be added at the end of sedimentation to form chloramines. Then the water is lightly rechlorinated in the clearwell to suppress regrowth of pollutants in the distribution system.

In small municipalities across the heartland, treated water is pumped into elevated tanks emblazoned with the name of the town. These towers serve to keep the water clean, meet surges in demand, and supply even pressure at the tap. Large cities tend to keep the water in open reservoirs, an unfortunate but perhaps necessary practice that leaves the water subject to the reintroduction of various undesirable substances.

Hydraulics

Flow through the plant is described by Bernoulli's equation, where all the components are expressed in terms of head in feet,

$$z_1 + P_1/\gamma + V_1^2/2\text{g} = z_2 + P_2/\gamma + V_2^2/2\text{g} + H_\text{L}$$

and z_i, distance of water level above datum; P_i/γ, pressure head at surface; $V_i^2/2\text{g}$, velocity head; H_L, head loss; $i = 1$, upstream; $i = 2$, downstream.

The head-loss term is made up of entrance and exit losses, pipe friction losses, and minor losses. Hydraulic calculations are best made starting at the clearwell and working upstream.[1]

It is preferable to have the water flow through the plant by gravity. The head losses through the A particle with settling velocity satisfying HRT $= D/V_\text{s}$ will reach the bottom of the tank before being swept up and out (as will particles with V_s greater than this). Then processes are approximately (a) rapid mixing, 1 ft; (b) flocculation, 2 ft; (c) sedimentation, 2 ft; (d) and filtration, 10 ft, for a total of 15 ft. If the water comes into the plant at the level of the clearwell, this head and the friction losses in the lift pipe must be provided by a pump. At the pump discharge, the water horsepower is HP_w; the pump input is the motor brake horsepower, HP_b; and the power required to drive the motor is P, where

$$HP_\text{w} = QH/C_1 \qquad HP_\text{b} = QH/C_1e_\text{P}$$

$$P = C_2HP_b/e_\text{m}$$

and Q, pump flow rate, gal/min; H, pump discharge head, ft; C_1, constant, (550 ft-lb/s-hp) (60 s/min)/(8.34 lb/gal) 3960 ft-gal/min-hp; e_p, pump efficiency; C_2, constant, 0.746 kW/hp; e_m, motor and drive efficiency.

Example 2. Size the motor, and find the input power required to provide 15 ft of head for a 24-mgd water treatment plant. Ignore plumbing losses. Take the pump efficiency to be 70% and the motor/drive efficiency to be 90%.

Solution. ***The flow rate is (24×10^6 gal/da)/[(24 h/da) (60 min/h)] = 16,667 gal/min. Thus***

$$HP_b QH/C_1 e_p = (16{,}667\ \text{gal/min})(15\ \text{ft})/\left[(3960\ \text{ft} - \text{gal/min} - \text{hp})(0.70)\right] = 90\text{hp}$$

and

$$P = C_2 HP_b / e_m = (0.76\text{kW/hp})(90\text{hp})/(0.90) = 75\text{kW}$$

Because a motor is just as efficient at 75% load as at full load, a 125-hp motor should be installed.

The important hydraulic parameters of a sedimentation tank (Figure 2) are the hydraulic retention time (HRT), the horizontal velocity of the water through the tank (V_h), and the surface overflow rate (SOR). For a tank with dimensions *L, W, D* for length, width, and depth, the volume $\forall$, is LWD; the surface area, A_s, is *LW*; and the vertical crosssectional area is $A_c = WD$. Therefore, the horizontal velocity and the SOR are

$$V_h = Q/A_c = Q/DW \text{ and } \text{SOR} = Q/A_s$$

The HRT is

$$\text{HRT} = L/V_h = L/(Q/DW) = \forall/Q$$

A particle with settling velocity satisfying HRT = D/V_s will reach the bottom of the tank before being swept up and out (as will particles with V_s greater than this). Then

$$V/Q = D/V_s$$
$$\text{or } V_s = DQ/V = Q/LW = \text{SOR}$$

That is, for a particle to settle, the settling velocity must be equal to the SOR (or greater).

The power (*P*) required for mixing or flocculation with an impeller in a tank is dependent upon the average velocity gradient (G) in the fluid, dynamic viscosity (μ) of the fluid, and volume ($\forall$) of the tank: $P = G^2 \mu \forall$ For wastewater treatment, the average velocity gradient for rapid mixing is about 500–1500/s and for flocculation is about 50–100/s.[5]

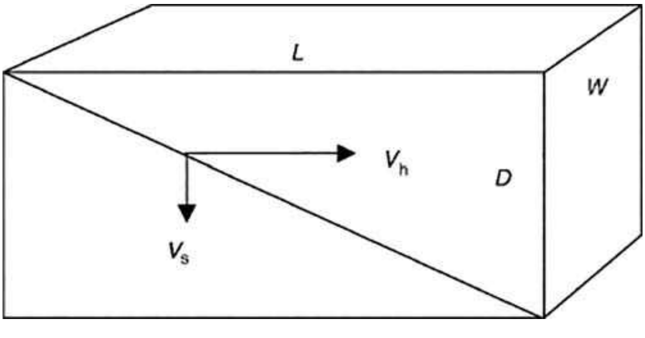


FIGURE 2 Sedimentation parameters.

Example 3. Find the power (P) required to achieve an average velocity gradient (G) of 100/s in a 1 million-gal flocculation tank whose contents are at a temperature (T) of 40°F.

Solution. ***The dynamic viscosity of water at 40°F is*** $\mu = 2.359\times10^{-5}$ ***lb-s/ft***^{2}***. Therefore,***

$$P = (100/\text{s})^2\left(2.359\times10^{-5}1\text{b}-\text{s/ft}^2\right)\left(10^6\ \text{gal}\right)$$

$$\times\left(0.746\ \text{kW/hp}\right)/\left[\left(550\ \text{ft}-1\text{b/s}-\text{hp}\right)\left(7.48\ \text{gal/ft}^3\right)\right]$$

$$= 42.8\ \text{kW}$$

Example 4. Consider a plant with an input flow rate of $Q = 24$ mgd $= 24\times10^6$ gal/da. The two sedimentation tanks are 300 ft long, 40 ft wide, and 13 ft deep. Find the HRT, the velocity of the water through the tank, and the SOR.

Solution. ***The flow rate through each tank is*** $Q = (24/2\times10^6\ \text{gal/da})/[(24\ \text{h/da})(60\ \text{min/h})] = 8333\ \text{gal/min} = (8333\ \text{gal/ min})/[(7.48\ \text{gal/ft}^3)\ (1440\ \text{min/da})] = 1114\ \text{ft}^3/\text{min}$***. Therefore, the velocity in each tank is*** $V = (1114\ \text{ft}^3/\text{min})/[(40\ \text{ft})\ (13\ \text{ft})] = 2.14\ \text{ft/min}$***, and the surface overflow rate is*** $SOR = (8333\ \text{gal/min})/[(40\ \text{ft})\ (300\ \text{ft})] = 0.694\ \text{gal/min-ft}^2$***.***

Wastewater

Some authors make the following distinction: wastewater is water that has been used for domestic, industrial, or commercial purposes, whereas sewage is more inclusive in that it can include water that has not been used, such as rain runoff. In the past 50 years, sanitation agencies have made a great effort to confine runoff to storm sewers and out of treatment plants, so the term WWTP is appropriate. These facilities are also called publicly owned treatment works (POTWs).

History

The history of wastewater treatment is a sordid one of determined ignorance and apathy. Until 1965, for example, Salt Lake City was dumping raw sewage into a 9-mi open canal that emptied into the Great Salt Lake.[7] In other parts of the world, this kind of practice continues to the present day.

Dr. John Snow in London convincingly linked cholera with the consumption of contaminated waters. This most famous episode in the history of both epidemiology and water treatment occurred in the late summer of 1854. Repeated outbreaks of cholera had occurred between 1831 and 1854 in the industrial cities of England, with little being done to prevent or contain it. The particularly sudden and violent episode in and around Broad Street in September gave Dr. Snow the opportunity to verify his belief that the cause was in contaminated water. When he persuaded the authorities to remove the handle from the Broad Street pump, the spread of the disease was halted. Later, Snow established that wastes from a single infected individual had been dumped into a leaking cesspit near the Broad Street well. After some time, people accepted the fact that fecal contamination of drinking water was a major cause of disease.[3]

With better water supplies and sewer systems, there was a sharp decrease in the incidence of waterborne diseases, even before the agents were identified. After half a century of research, the concept of waterborne disease was established. The cause was known to be microorganisms in the digestive tract and the associated health hazards had been proven. Then work proceeded on two fronts: analytical methods for the detection of fecal pollution and the development of treatment methods and facilities. Research led to publication of the first edition of *Standard Methods* in 1901,[8] and the SSF was an early and very effective method for treatment. (In fact, the newer and faster conventional rapid sand filter does not remove dissolved constituents as effectively.)

Standards and Monitoring

The Clean Water Act of 1972 and subsequent legislation placed increased emphasis on the importance of reducing the discharge of pollutants to natural waters. The minimum national standards now for secondary treatment are the "30/30" rule: a 30-day average of no more than 30 mg/L of BOD_5 (5-day biochemical oxygen demand) and TSS (total suspended solids), as well as pH to be between 6.0 and 9.0 at all times. Unfortunately, meeting these standards does not guarantee the absence of disease-causing agents—notably, *Giardia lamblia* and *Cryptosporidium parvum*. Increased sophistication of monitoring techniques is leading to better treatment techniques and stricter standards.[9]

Sources and Collection

Sanitary sewers receive some groundwater infiltration and stormwater. Otherwise, 90% or more of the intended influent is of residential or commercial origin. Industrial users are either direct dischargers dumping into a waterway or indirect dischargers dumping into a POTW. (Some industrial liquid waste may also be hauled off.)

Influent is collected in closed pipe, mostly by gravity flow. When pump lift is necessary, the flow is into short runs of pressure pipe, eventually returning to gravity flow lines. Because plants are often located next to natural waters, they are typically at the low points of terrain, which keeps pumping to a minimum.

Treatment

Several levels of increasing care are defined. Preliminary treatment is the removal of large items, sand and grit, floatables, and grease. Wastewater typically contains floatable materials—particularly fats, oils, and grease (FOG)—whereas (fresh) water does not. Removal is accomplished purely by physical processes such as screening and gravity, and is intended to protect the plant equipment. Primary treatment is the removal of suspended solids and organic matter, often by the addition of chemicals. Secondary treatment is a biological process that removes organics and suspended solids (and sometimes nitrogen and phosphorus), followed by disinfection. Tertiary treatment removes remaining suspended solids by fine filtering, and may include disinfection and nutrient removal. The standard today for wastewater is full secondary treatment, meaning that all the influent to a plant is given secondary treatment. When a plant cannot handle the flow, partial secondary treatment means that all the influent is given preliminary and primary treatment, but only part of it is given secondary treatment.

A typical process chain for secondary treatment is the activated sludge process shown in Figure 3.[9] Large items are screened out. Dense noncontaminants are removed purely by gravity in the grit chamber. Primary treatment is a chemical/physical process, whereby small particles agglomerate (flocculate)

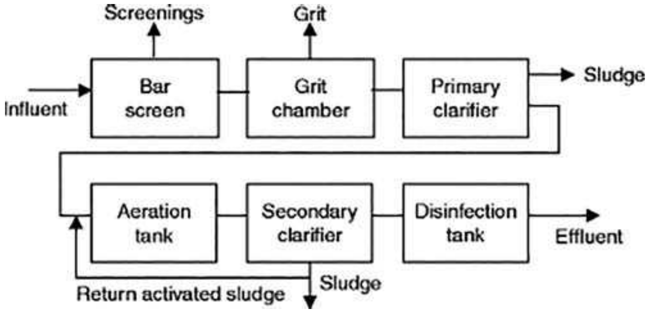


FIGURE 3 Activated sludge wastewater treatment.

and gravitate out. Secondary treatment is biological, in which microbes consume the dissolved and suspended organic matter. Disinfection kills most of the remaining contaminants.

Preliminary and primary treatment for wastewater is much like that for water. Preliminary screening of wastewater is necessary because large objects sometimes find their way into the sewer, an unfortunate example being construction debris illegally dumped into a manhole (alternatively, "maintenance hole"). (Note that a manhole cover is round for at least three reasons: (a) It will not fall through the hole, no matter how oriented; (b) it is easy to move by rolling; and (c) it need not be rotated to fit.) Primary treatment for both is nominally flash mix, flocculation, and sedimentation. Wastewater undergoes secondary treatment, in which microbes remove biological pollutants.

Both organic and inorganic particulates may be removed by settling, flotation, or filtering, depending on particle size and density. Carbon filtration is preferable to vaporizing, because the latter merely moves the substance from the water to the atmosphere.

Most reactions in waste treatment are of first order—that is, the rate of reaction is proportional to the concentration of the pollutant.

$$\mathrm{d}C/\mathrm{d}t = -kC$$

where C is concentration of pollutant (mg/L); t is time (min); k is reaction constant (mg/L-min).

Reactors are of three types: complete mix (or batch), plug flow, and dispersed flow. In a complete mix reactor, the reactor is filled; the reaction is allowed to take place; and then the reactor is emptied. The concentration of pollutant is equal throughout the tank. Complete mix reactors are approximately cubical and can be set up in sequence to provide an increasing proportion of removal. In a plug flow reactor, the flow moves through the reactor with the reaction taking place so that the concentration of pollutant is less at the outlet than at the inlet. These tanks are long in proportion to their length and width. Turbulence should be minimized to retain the form of the plug. Most reactors are of the dispersed flow type, intermediate between the other 2.[9]

Liquid and Solids Disposal

By the principle of conservation of mass, treatment does not make the contaminants disappear; it merely separates them from the water that bears them. When most of the contaminants have been removed and the water is sufficiently clean, it may be discharged to a natural waterway, such as a river, lake, or ocean. Most of the processes in Figure 3 produce residuals. These impurities—such as sediment, sludge, waste washwater, and brine—are left behind to be treated and disposed of in other ways. Large items from screening and grit from the grit chamber are sent to landfill. Sludges from primary and secondary treatment are sent to a digester, which itself produces waste. Waste washwater from the filters is recycled, but because it adds to the throughput volume, it should be kept to a minimum.

Industrial Waste

For both direct dischargers and indirect dischargers, the content of industrial wastewater is regulated. Industrial users are allowed to send the first fraction of runoff from rainstorms to the sanitary sewer, for the purpose of keeping pollutants out of the storm sewers and, ultimately, out of natural waters after which it must be diverted to storm sewers.

Because POTWs are set up to treat organic waste, much industrial waste is incompatible with public treatment systems, so is usually subject to pretreatment. The objectives of such treatment are to prevent interference with the process in the POTW, to prevent pass-through of pollutants to the receiving waters, and to make possible reuse of the effluent and sludge from the POTW.[10]

Some treatment strategies are flow equalization to prevent shock loading to the POTW; solids removal by straining or settling; removal of FOG by dissolved air floatation or centrifuging; neutralization of high-pH or low-pH solutions; and hydroxide precipitation of heavy metals. A notable exception to the last is the removal of Cr^{+6}, which will not respond to hydroxide precipitation. Instead, it is converted by chemical reduction at low pH to Cr^{+3}, which can be removed by hydroxide precipitation. Dissolved inorganics may be removed by hydroxide precipitation, ion exchange, or membrane filtering.[10] It should be noted that diluting industrial wastewater to reduce the concentration of pollutants is not acceptable, and that "dilution of pollution is not a solution, and can lead to prosecution." The most important principle is segregation: keeping the pollutants separated so that they can be treated individually.

Issues

Wastewater treatment plants are designed to treat organic wastes. Other substances (nutrients in fertilizer, pharmaceuticals, etc.), can pass through and cause problems in receiving waters (e.g., algal growth and abnormal growth in fauna). Yet others, such as heavy metals and toxic chemicals, can cause interference (also called upset)—disruption of the process in the biological reactor.

Failed equipment can cause raw or partially treated wastewater to flow into storm drains and then into natural waters. Runoff from heavy storms can flow into sanitary sewers and overwhelm treatment plants. Cities with combined sewers are especially subject to this problem. Inadequately sized treatment plants will discharge partially treated wastewater in times of heavy flow.

Everything removed from wastewater must be disposed of. Sometimes, objections are raised to the release of volatiles into the atmosphere. Although the creators believe that digested sludge is a fertilizer rich in nutrients, others are not convinced.[11]

More radically, some have questioned the wisdom of the whole process of fouling great quantities of cleaned water and then cleaning it again.[12]

Conclusion

A basic requirement of human existence is an adequate supply of clean water. Today, very few have access to clean, untreated water. Wealthy societies obtain clean water by treating it, while poor ones often rely on polluted sources and suffer from the resulting waterborne diseases.

The idea that many municipalities draw from surface waters that are used for the disposal of wastewater is sobering, if not chilling. In the United States, the Safe Drinking Water Act mandates water treatment standards, and the Clean Water Act mandates wastewater treatment standards. Observance of these standards has made the practice of having a common source and sink acceptable.

The conventional water treatment process described in this entry has been very effective in removing bacterial pathogens. Dechlorination to control disinfection byproducts created by chlorination was instituted as a result of the Safe Drinking Water Act. The most recent major issue is the resistance of *Cryptosporidium* oocysts to chlorination. Membrane filtration is an effective way to remove these and other very small suspended pathogens. New water treatment plants are likely to be based on this technique because of its effectiveness and ease of operation.[5]

The most important developments required in wastewater treatment are (a) building or expanding facilities to provide the capacity to subject all flow to complete secondary treatment, (b) repairing and maintaining the collection system to ensure that all wastewater reaches the treatment plant, and (c) finding practicable ways to dispose of the residual solids.

References

1. Reichenberger, J. Unpublished notes. Loyala Marymount University: Los Angeles, 2005.
2. Jesperson, K. *Search for Clean Water Continues.* Available at: http://www.nesc.wvu.edu/ndwc (accessed November 2006).

3. Judith Summers. *Broad Street Pump Outbreak.* Available at: http://www.ph.ucla.edu/epi/snow/broadstreetpump.html (accessed November 2006).
4. United States Environmental Protection Agency, Office of Water. *The History of Drinking Water Treatment.* Available at: http://epa.gov/safewater/consumer/pdf/hist.pdf (accessed November 2006).
5. Kawamura, S. *Integrated Operation and Design of Water Treatment Facilities,* 2nd Ed.; Wiley: New York, 2000.
6. Graber, C. *Desalination in Spain.* Available at: http://www.technologyreview.com/spain/water (accessed November 2006).
7. Salt Lake City Department of Public Utilities. *Conduits of Civilization.* Available at: http://www.ci.slc.ut.us/utili-ties/NewsEvents/news1998/news9281998-1.htm (accessed November 2006).
8. Greenberg, Arnold, et al., eds. *Standard Methods for the Examination of Water and Wastewater,* 21st Ed.; American Public Health Association (APHA), American Water Works Association (AWWA), and Water Environment Federation (WEF): Washington, D.C., 2005.
9. Tchobanoglous, G.; Burton, F.L.; Stensel, H.D. *Wastewater Engineering, Treatment and Reuse,* 4th Ed.; Metcalf and Eddy Inc., McGraw-Hill: New York, 2003.
10. Water Environment Federation Pretreatment of Industrial Wastes In *Manual of Practice FD-3;* Water Environment Federation: Alexandria, VA, 1994.
11. Voters seek to block sludge, *Los Angeles Times* 2006, January 2. L.A. Fights Kern County Sludge Ban. *Los Angeles Times* 2006, August 16.
12. Stauber, J.C.; Sheldon, R.S. *A Brief History of Slime.* Available at: http://www.prwatch.org/prwissues/1995Q3/slime.html (accessed November 2006).

25

Wastewater: Municipal

Sven Erik Jørgensen

Municipal Wastewater

The composition of domestic sewage varies surprisingly little from place to place, although to a certain extent, it reflects the economic status of the society.

A typical composition including biochemical oxygen demand (BOD_5), chemical oxygen demand (COD), suspended matter, and nutrients is shown in Table 1.

Most industrialized countries require treatment of municipal wastewater. The standards vary for the European Union, the United States, Canada, Australia, New Zealand, and Japan, but it is generally required to reduce BOD_5 to about 10 mg/L, nitrogen to about 10 mg/L, and the phosphorus concentration to 1–2 mg/L. To obtain the required reductions, a combination of the methods presented in the entry "Waste Water Treatment: Overview of Conventional Methods" can be applied. The combinations of methods applied for the treatment of municipal wastewater are presented in the next sections, and

TABLE 1 Typical Composition of Municipal Wastewater (in milligrams per liter)

Constituent	Soluble	Particulate	Total
BOD_5	100–200	50–100	150–300
COD	200–500	100–200	200–700
Ammonium-N	20–40	0	20–40
Nitrate-N	5–20	0	5–20
Organic nitrogen	0	5–20	25–60
Suspended matter	–	40–80	40–80
Carbohydrates	20–40	10–15	30–55
Amino acids	10–15	15–25	25–40
Fatty acids	0	50–80	50–80
Surfactants	10–20	5–10	15–30
Creatinine	3–5	0	3–5
Phosphorus	2–4	4–10	6–14

Source: Jørgensen[1] and Jørgensen.[2]

it will be demonstrated that it is possible by a suitable combination of methods to meet all realistic effluent standards. Municipal wastewater can, however, also be treated by ecotechnological methods, for instance, by waste stabilization ponds and by constructed or natural wetlands. These methods are presented in the entries *Pollution: Point and Non-Point Source Low Cost Treatment* (p. 2174), *Waste: Stabilization Ponds* (p. 2652), *Wetlands* (p. 2846), and *Wetlands: Treatment System Use* (p. 2885).

Combinations of Methods for the Treatment of Municipal Wastewater (Reduction of Bod$_5$)

Models are used increasingly to design and optimize wastewater treatment methods. For details about the applied models, see Jørgensen's *Handbook of Ecological Models Used in Ecosystem and Environmental Management*,[1] which has a comprehensive entry devoted to models of wastewater treatment systems.

There are a number of different possible designs of biological treatment plants.[2]

The use of mechanical-biological wastewater treatment is a classical method, or rather the combination of the two methods. Figure 1 shows the flowchart of a classical mechanical and activated sludge plan as it is generally used all over the world. The processes involve screening, a separation of grease and sand in the grit chamber, a primary settling, and an aeration step followed by sedimentation. Chlorine is often added before discharge to the receiving water, particularly when it is used for swimming and bathing. The sludge is most often digested anaerobically, which produces biogas. In the activated sludge plant, a rapid adsorption and flocculation of suspended matter take place. Organic matter is oxidized

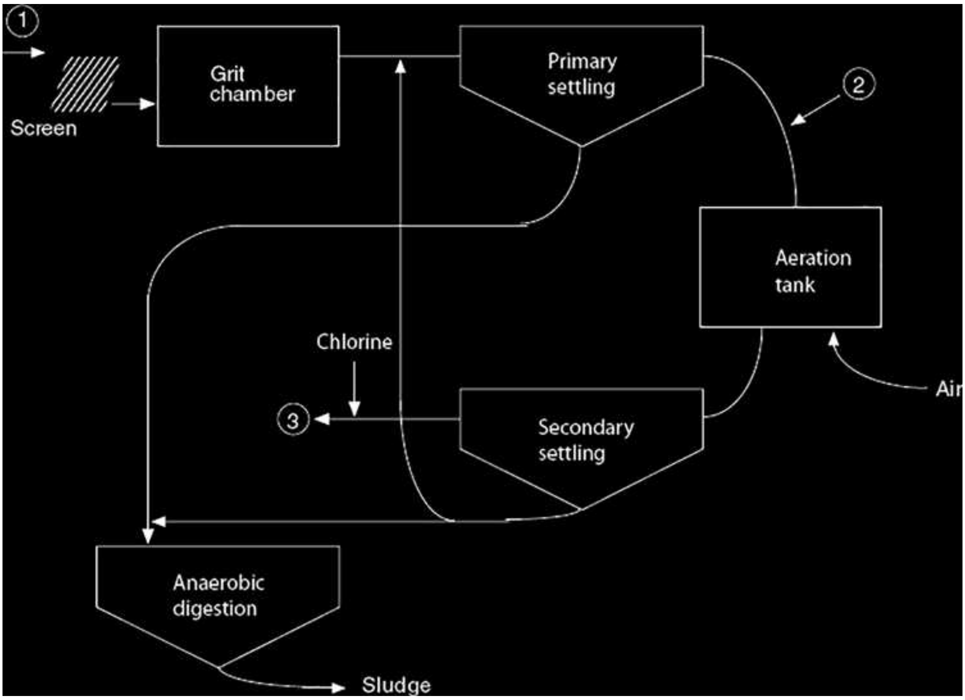


FIGURE 1 A flow diagram of a conventional activated sludge plant. It is widely used to treat municipal wastewater. Number 1 is untreated municipal wastewater that would have a BOD$_5$ of about 150–300 mg/L, total nitrogen of about 25–45 mg/L, and total phosphorus of about 6–12 mg/L. The mechanical treatment (sample at point 2) will reduce BOD$_5$ and total N by about 25%–40%, while the reduction of total phosphorus will be minor. The totally treated wastewater (point 3) will have a BOD$_5$ of about 10–15 mg/L, while the nitrogen is only reduced 35%–50% and the phosphorus concentration only about 10%–20%.

and decomposed. Sludge particles are dispersed and settled by the secondary settling. The processes are conceptualized in Figure 2.

Oxidation ditches (see Figure 3) can replace the activated sludge plant and the secondary settling. The rotor provides the aeration to oxidize the organic matter. The influent and the rotor are usually stopped during the night, when the inflow is low anyhow, to allow settling. The supernatant is withdrawn through an effluent launder. The retention time is usually 2–5 days.

A trickling filter is a bed packed with rocks, although recently, plastic media (celite pellets[3]) and bioblocks with a high surface area due to the high porosity are also applied. They require substantially less space than the stone-packed trickling filter. They have usually a specific surface of 100 m^2/m^3 or even more. It is at least 2–3 times the specific surface for a rock trickling filter. The medium is covered by a slimy microbiological film. The wastewater is passed through the bed and oxygen and organic matter diffuse into the film, where oxidation occurs. Recirculation improves the removal efficiency. A flowchart of a treatment plant combining a trickling filter and an activated sludge plant is shown in Figure 4. The alternative ecotechnological methods used for BOD_5 reduction, lagoons, and waste stabilization ponds are presented in the entry "Waste Stabilization Ponds." Various solutions to reduce BOD_5 based on natural and constructed wetlands are covered by several entries, all with wetlands as the key word in the title.

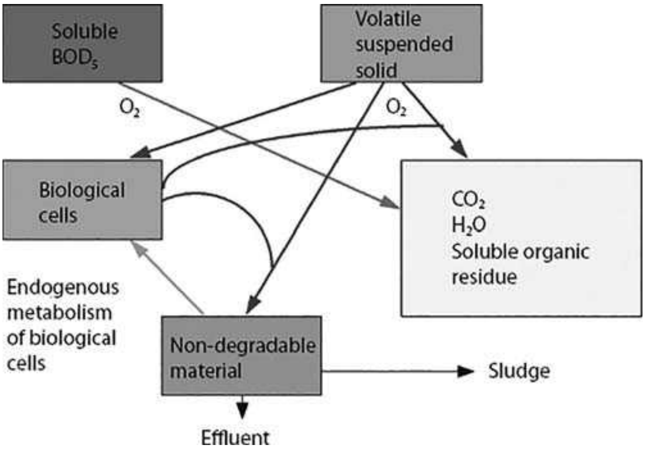


FIGURE 2 Processes characteristic for the biological treatment.

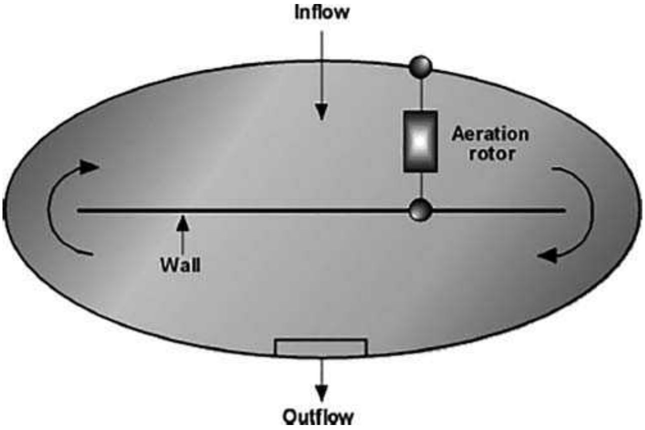


FIGURE 3 Oxidation ditch.

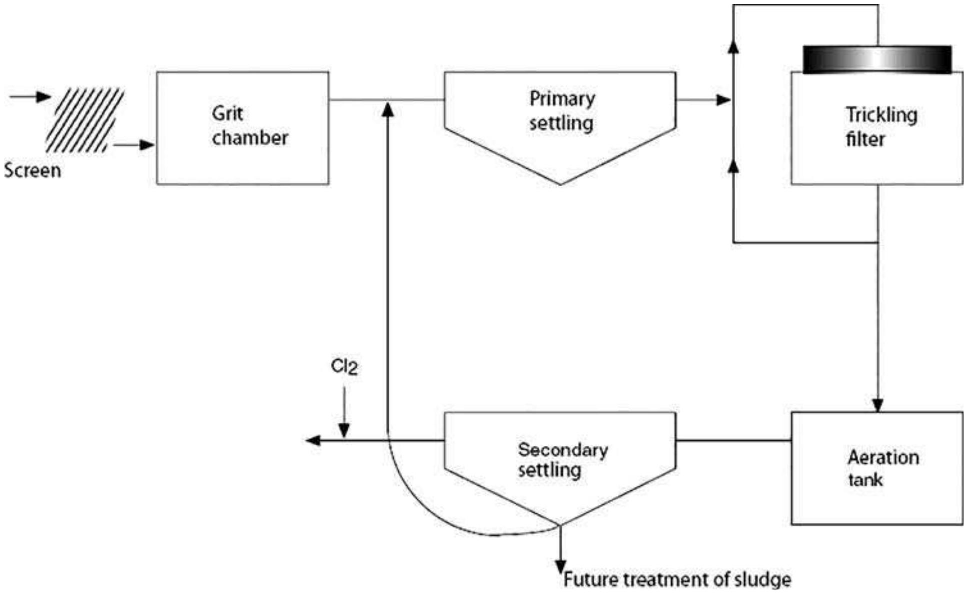


FIGURE 4 Treatment of municipal wastewater, combining trickling filter and activated sludge treatment.

Physical–chemical methods have been proposed to replace mechanical–biological treatment, for instance, the so-called AWT system, which is based on the application of a combination of mechanical treatment, precipitation, adsorption on activated carbon, and sand filtration (see Figure 5).

It has been considered that recovering proteins and grease could at least partially cover the costs of the treatment of wastewater in slaughterhouses, fish industries, starch factories, and other foodstuff industries. Figure 6 shows a method that has been applied in several, but still relatively few, cases.

If the wastewater has high concentration of grease, oil, and fat, it is possible to apply a flotation unit, offering an alternative to sedimentation. A portion of water is pressurized by 3–10 atm, and when this water is returned to the normal atmospheric pressure in the flotation unit, air bubbles are created. The air bubbles attach themselves to particles, and the air-solute mixture rises to the surface where it can be skimmed off, while the clarified water is removed from the bottom of the flotation unit. Figure 7 shows

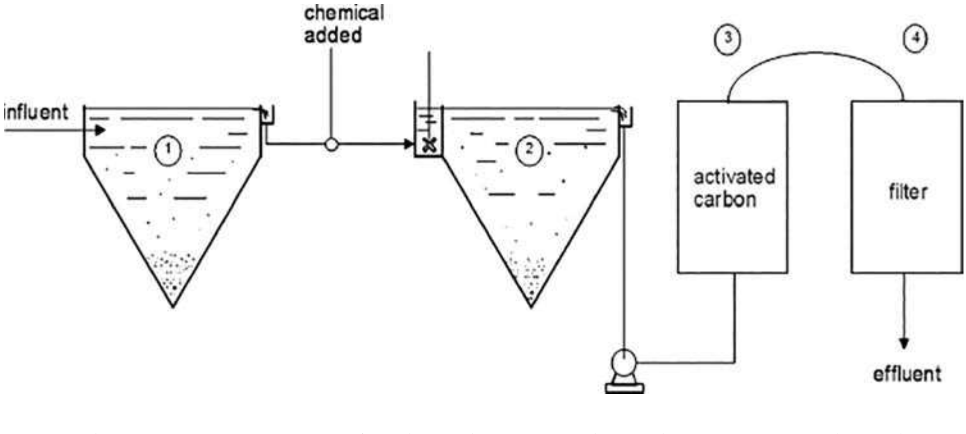


FIGURE 5 The AWT system consisting of mechanical treatment, chemical precipitation, settling, adsorption on activated carbon, and filtration. The method is not used very much. It has the advantage that it is easier to control but it is also more expensive with respect to both investment and operation.

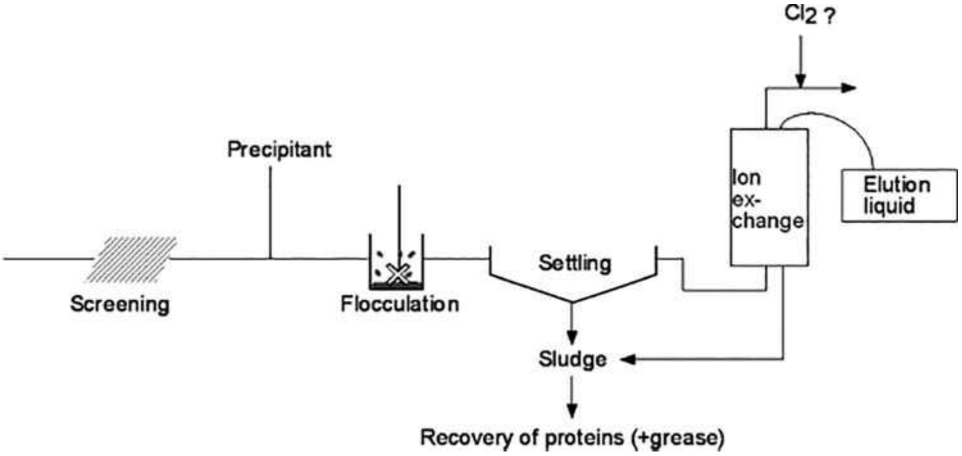


FIGURE 6 Recovery of proteins and grease from the wastewater discharge from the foodstuff industries is possible by a combination of precipitation and ion exchange. The recovery pays partially for the treatment of the wastewater by the process.

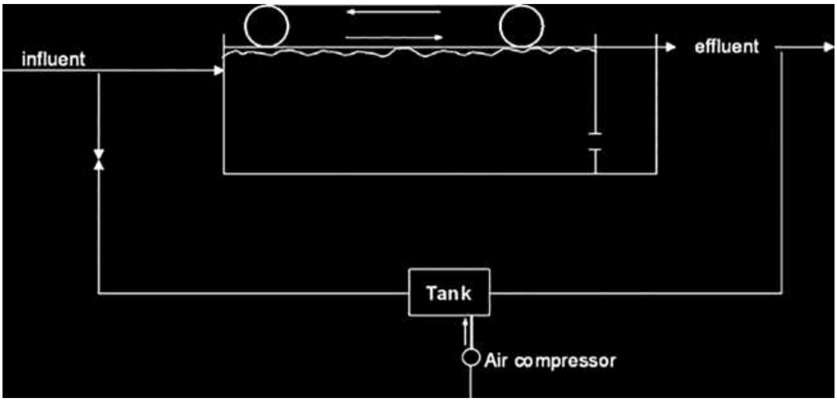


FIGURE 7 Flotation unit.

a flotation unit. Usually, the retention time in a flotation unit is 3–6 times less than for a settling unit, which means that a significant volume reduction is obtained. Flotation has therefore most frequently replaced sedimentation in slaughterhouses, fish industries, and oil industries.

Methods for the Treatment of Municipal Wastewater (Reduction of the Phosphorus Concentration)

Nutrient removal is most frequently carried out by chemical precipitation, often combined with mechanical–biological treatment. Chemical precipitation can be applied at three different points in the mechanical–biological plant as shown in Figure 8. Sometimes, both direct precipitation and posttreatment is applied, which together with sand filtration makes it possible to obtain 0.1 mg P/L or less in the effluent. In many cases where such a low phosphorus concentration is needed for particular sensitive receiving waters (mainly lakes), this double precipitation is a very attractive method, because it is

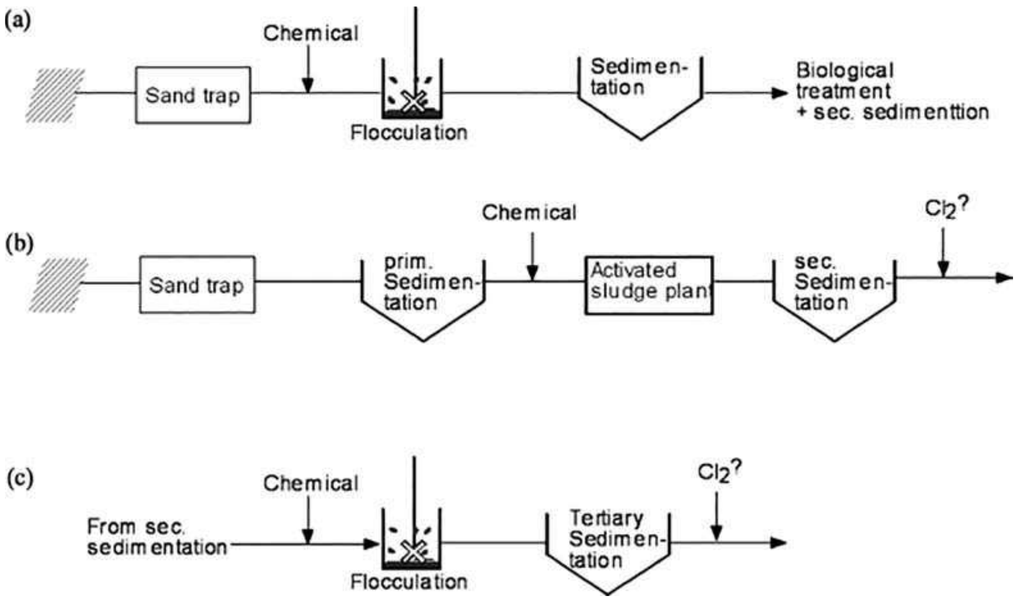


FIGURE 8 Precipitation by aluminum sulfate or other aluminum compounds, iron(III) chloride, and calcium hydroxide is able to reduce the phosphorus concentration in wastewater significantly. The precipitation is applied often in combination with mechanical biological treatment and can be carried out after the sandtrap: (a) direct precipitation, before the activated sludge plan: (b) simultaneous precipitation or after the mechanical–biological treatment: (c) posttreatment. The arrow+"Chemical" indicates the three different possibilities for addition of the precipitation chemical and "chlorine?" indicates where chlorine eventually can be added.

relatively cost moderate compared with other alternatives. Without the sand filtration and only a single precipitation, it is usually relatively easy to obtain a concentration of phosphorus in the effluent between 1.0 and 1.5 mg/L.

Aluminum sulfate, various polyaluminates, calcium hydroxide, and iron(III) chloride can be applied for the precipitation. The amount of hydrated lime or calcium hydroxide needed for the precipitation is usually 2.5–6 times higher than for the aluminum and iron compounds, because a high efficiency of the precipitation requires a pH of 10.0 or higher, which, dependent on the hardness of the water, is not possible without the indicated amount of calcium hydroxide. For a hardness of 15–30 hardness degrees, the amount is 100–480 mg calcium hydroxide per liter. It would usually give an efficiency of 90%–95% precipitation of the phosphorous compounds, which is slightly better than for most precipitations with aluminum and iron compounds. The disadvantage of precipitation with lime is the high pH, which makes adjustment of the pH after the precipitation and settling necessary. Carbon dioxide produced by incineration of the sludge or solid waste can be used for this purpose. If the sludge is incinerated, it is possible to partially recycle the calcium hydroxide and thereby reduce the costs of precipitation chemicals. Recycling 3–5 times is possible, and afterwards it can be applied as fertilizer, as it has a relatively high phosphorus concentration. A flowchart with direct precipitation and recycling of calcium hydroxide is shown in Figure 9.

If the sludge after calcium hydroxide precipitation cannot be incinerated, an adjustment of pH is needed before anaerobic digestion or aerobic sludge treatment. Heavy metals are removed more effectively by the use of calcium hydroxide than by aluminum and iron compounds. Lead, copper, and chromium are removed by a very high efficiency by all the mentioned precipitation chemicals, while only calcium hydroxide would give a high removal efficiency for cadmium and zinc.

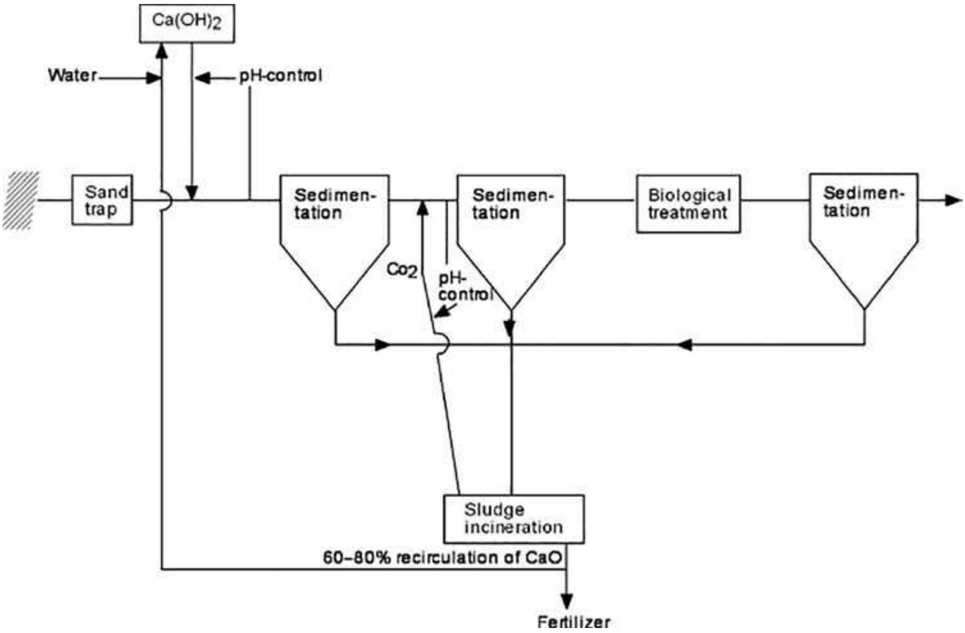


FIGURE 9 Chemical precipitation with partial recirculation of calcium hydroxide and use of carbon dioxide from incineration of the sludge for pH adjustment.
Source: Jørgensen.[1]

The amount of aluminum and iron compounds can be found in Figures 10 and 11 or by the use of the following equation based on Freundlich adsorption isotherms:

$$(C_O - C)/n = a^{*}\ C^{b}$$

where C_o is the initial concentration of phosphorus (mg P/L), C is the final concentration, n is the dose of chemical expressed as milligrams of Al per liter or milligrams of Fe per liter, and a and b are characteristic constants that can be found in Table 2.

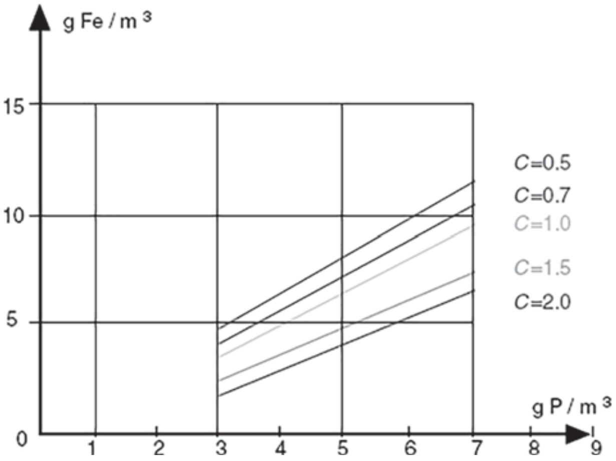


FIGURE 10 Addition of iron (III) as f(P-concentrations) at different P-concentrations in the effluent.

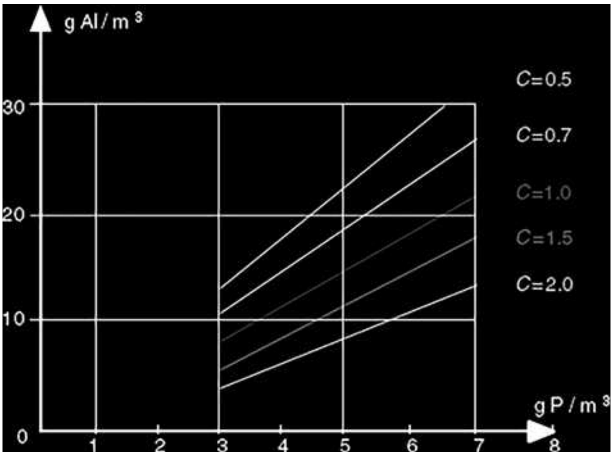


FIGURE 11 Addition of aluminum salts as f(P-concentrations) at different P-concentrations in the effluent.

TABLE 2 TThe Constants in Freundlich's Adsorption Isotherms for Aluminum Sulfate and Iron (III) Chloride

Precipitation with	a	b
Aluminum sulfate	0.63	0.2
Iron(III) c]hloride	0.26	0.4

Source: Jørgensen.[2]

Note: With good approximation, the constants for aluminum sulfate can also be applied for other aluminum compounds.

A more rapid flocculation, precipitation, and settling can be obtained by the addition of synthetic organic polymeric flocculants. They may be cationic polyelectrolytes, anionic polyelectrolytes, or non-ionic polymers. It is hardly possible to indicate which one of the polymeric flocculants would give the best result, as the ionic characteristics of municipal wastewater vary significantly. It is recommendable to test the various types of polyflocculants from case to case. The optimum design of a flocculator, before the settling of the precipitated material, should be based on a mathematical model (see Dharmappa et al.[4] and Thomas et al.[5]).

It is possible to remove phosphorus by biological treatment. Activated sludge systems with anaerobic and aerobic zones in sequence have been developed to achieve a higher phosphorus removal. The system is called EDPR—enhanced biological phosphorus removal. The shift between aerobic and anaerobic conditions activates the microorganisms to take considerably more phosphorus than under aerobic conditions, particularly if the wastewater contains relatively high concentrations of easily biodegradable organic matter. With a P/BOD_5 ratio of more than 20, a phosphorus removal of 80%–90% can be obtained.

Methods for the Treatment of Municipal Wastewater (Reduction of Nitrogen Concentration)

A combination of biological *nitrification and denitrification* can significantly reduce the nitrogen concentration in the effluent. The applied chemical processes are

$$NH_4^+ + 2O_2 \rightarrow NO_3^- + H_2O + 2H^+$$

$$4NO_3^- + 5C + H_2O \rightarrow 5HCO_3^- + 2N_2 + H^+$$

The ammonium is oxidized to nitrate and the nitrate is used to oxidize organic matter by anaerobic conditions—here indicated just as C. Thereby, the nitrate is reduced to dinitrogen, which is released to the atmosphere, where there is about 78% dinitrogen, and a minor addition of dinitrogen is therefore harmless.

Effective nitrification occurs when the sludge age is greater than the reciprocal rate of the constant of the nitrifying microorganisms.[6] The sludge age is defined as $X/\Delta X$, where X is the mass of sludge in the system and ΔX is the sludge yield per unit of time. Usually, the time unit applied is 24 hours. The relationship between nitrification efficiency in percentage and the sludge age is shown in Figure 12.

The nitrification process is a two-step biological process. Firstly, ammonia is oxidized to nitrite by *Nitrosomonas*. Secondly, nitrite is oxidized to nitrate by *Nitrobacter.* The optimum pH range for the nitrification is 6.7–7.0.[7] The oxygen concentration for the nitrification has to be at least 2 mg/L. Heavy metal ions are toxic to the nitrification at rather low concentrations. Toxic levels of about 0.2 mg/L are reported for chromium, nickel, and zinc. The nitrification process is sensitive to the temperature and an Arrhenius expression can be applied:

$$\mathrm{Kn}\left(\text{rate constant } 1/24\text{h}\right) = 0.18{*}1.128^{(\text{temperature} - 15)}$$

The denitrification takes place by a number of heterotrophic bacteria present in activated sludge. Anaerobic conditions are absolutely required. The optimum pH for this process is about 7.0. About 3 times as much BOD_5 as nitrate-N (both expressed in milligrams per liter) should be applied to ensure adequate denitrification. The process can be realized even at high concentrations of nitrate or at high salinities.[2] Addition of organic carbon—for instance, methanol—has been proposed to ensure proper denitrification, but it is of course not an attractive method to apply. Alternatives are as follows:

1. To switch between aerobic and anaerobic conditions. It is called alternative operation (for more details, see Diab et al.[8] and Halling-Sørensen and Jørgensen[9]).
2. To recycle a patty of the wastewater containing nitrate after the treatment is completed.

Both methods work properly. Other possibilities are the use of zeolite ion exchange material to obtain simultaneous nitrification and denitrification,[9] the use of electrolysis for denitrification,[10] and the use of activated carbon to enhance the denitrification.[11]

Ammonia can be removed by blowing air through the wastewater. The process is called *stripping,* and the equipment shown in Figure 13 can be used. It is of course required that the pH is sufficiently high

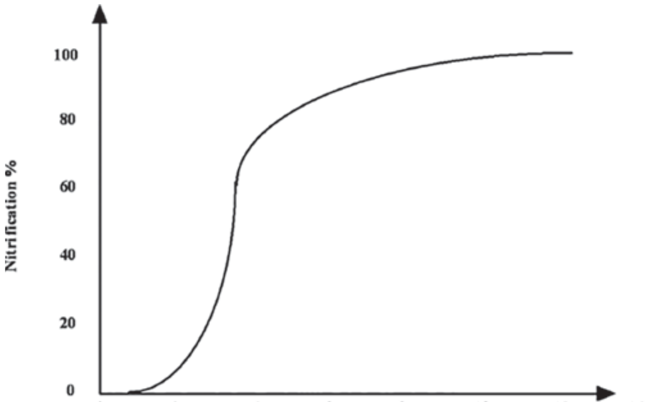


FIGURE 12 Nitrification efficiency as function of sludge age.

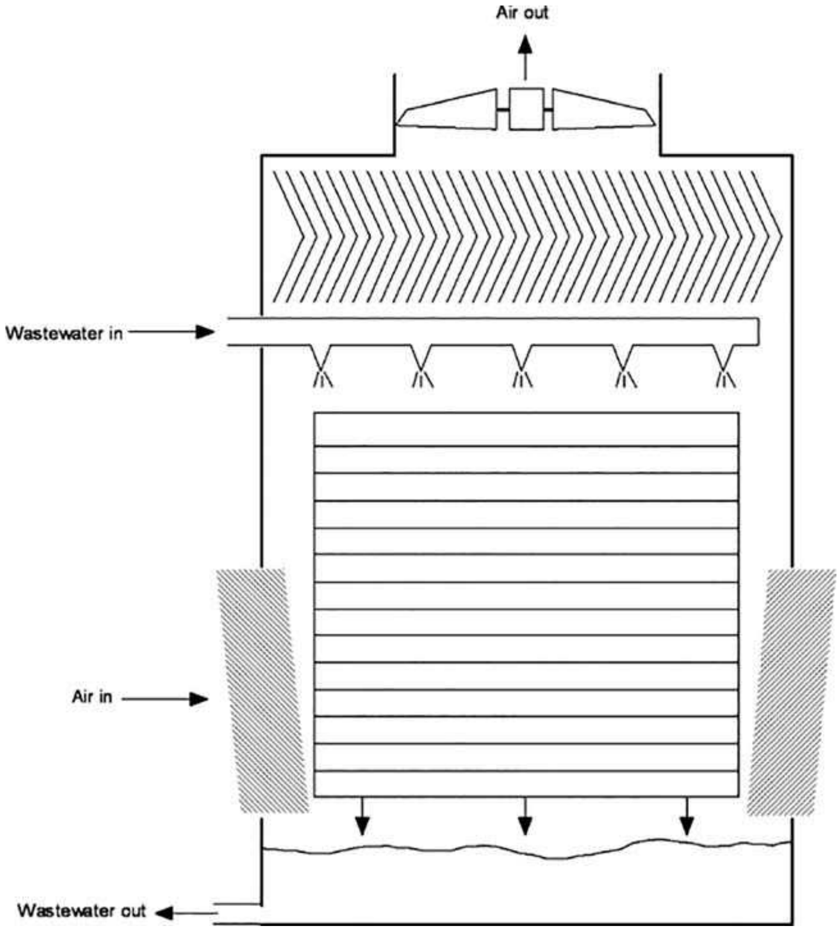


FIGURE 13 A sketch of a stripping tower is shown. The efficiency of the process is dependent on 1) pH; 2) temperature; 3) the amount of air relatively to the amount of water; and 4) the stripping tower height. Figures. 14 and 15 show how these factors influence the efficiency. As the air volume-to-water volume ratio is very high, there is a significant evaporation in the tower, which will cool the water.

enough to ensure that the gas ammonia is present. The equilibrium between ammonium and ammonia can be expressed by the following equation:

$$\mathrm{pH} = \mathrm{pK} + \log\ [\text{ammonia}]/[\text{ammonium}],$$

where pK is 9.3 for water with low concentrations of ions.

The cost of stripping is relatively low, but the process has three crucial limitations:

1. It is practically impossible to work at temperatures lower than 7°C due to the cooling effect of the evaporation.
2. Deposition of calcium carbonate can reduce the efficiency or even block the tower.
3. After the ammonia removal, pH is high and must be adjusted to 8.0 or below.

Ammonia can be removed by *chlorination* and *adsorption on activated carbon*. Chlorine can oxidize ammonia to NH_2Cl, $NHCl_2$, and NCl_3, and the activated carbon is able to adsorb the chloramines formed by these oxidation processes. By the adsorption process, it is most likely that dinitrogen and

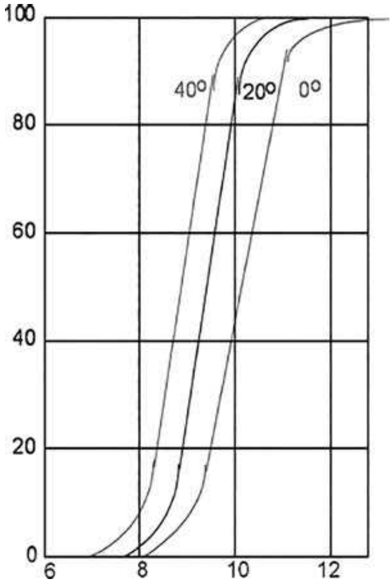


FIGURE 14 Stripping efficiency as function of temperature and pH.

chloride ions are formed. In the order of 10 parts by weight of chlorine are required for each part of ammonium-N. As the wastewater may typically contain 30 mg ammonium-N/L, 300 mg/L of chlorine is required. It would give a cost of \$0.10 per cubic meter to cover only the costs of chlorine. When the capital costs and the costs of the recovery of activated carbon are added, the total costs per cubic meter may reach about \$0.30. The method has therefore not found a wide application due to the high costs. The method has however a high removal efficiency even if the ammonium concentration is low, provided that sufficient chlorine is applied. Chlorination followed by active carbon adsorption can therefore be used as the last treatment for the removal of ammonium, where very low concentrations are required in the effluent.

Ion exchange by the use of the natural clay material clinoptilolite can be used for the removal of ammonium ions.[9] Clinoptilolite has a high selectivity for ammonium ions and about one-third to one-half of the ion exchange capacity, which is about 1.5 eq/L, and can, in many practical cases, be applied for the uptake of ammonium ions. It means that 1 L of clinoptilolite will be able to remove at least 1.5*14/3 g of ammonium-N = 7 g ammonium-N.

Ion exchange has furthermore been applied to remove nitrate from drinking water. A general anion exchanger is applied. It has an ion exchange capacity of about 2.5–3.0 eq/L, but the selectivity is not very high. It is however not necessary to remove the nitrate ions completely but to reduce the concentrations to the regional standards for nitrate in drinking water.

For further details about the ion exchange of wastewater and drinking water, see the entry denoted “Ion Exchange.”

Recycling of Municipal Wastewater

By a combination of the available treatment methods, it is possible to recycle municipal wastewater, i.e., to produce drinking water from municipal wastewater. The recycling takes about 8–10 days from toilets back to the water taps. All drinking water is in principle recycled. It takes normally nature a couple of thousands years to recycle water on average, which of course ensures a much better water quality than a recycling with a duration of 8–10 days.

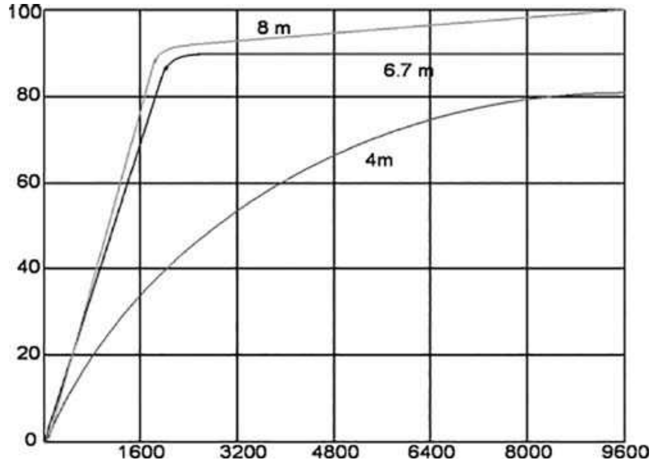


FIGURE 15 Efficiency of stripping process versus air/water ratio and tower height. Efficiency versus m^3 air/m^3 water for three different tower depths. The three tower depths are 4 m (insufficient), 6.7 m and 8 m. Tower height of 7–8 m is recommended.

Recycling of wastewater has been applied in Pretoria, South Africa, and in Windhoek, Namibia, due to an insufficient supply of natural water for production of potable water. In Pretoria, a treatment consisting of the following steps is applied: mechanical-biological treatment, aeration, lime precipitation, ammonia stripping, sand filtration, chlorination, adsorption on activated carbon, and a second chlorination. The plant in Windhoek has the same steps except ammonia stripping.

References

1. Jørgensen, S.E., Ed. *Handbook of Ecological Models Used in Ecosystem and Environmental Management;* CRC Press, 2011; 620 pp.
2. Jørgensen, S.E. *Principles of Pollution Abatement*; Elsevier: Amsterdam, 2000; 520 pp.
3. Sorial, G.A. et al. Evaluation of trickle-bed air biofilter performance for styrene removal. Water Res. **1998**, *32*, 1593–1603.
4. Dharmappa, H.B. et al. Optimal design of a flocculator. Water Res. **1993,** *27*, 513–519.
5. Thomas, D.N. et al. Flocculation modelling: A review. Water Res. *33*, 1579–1592.
6. Bernhardt, E.L. 1975. Nitrification in industrial treatment works. Second International Congress on Industrial Waste Water and WASTES, Stockholm, February 1975.
7. Groeneweg, J. et al. Ammonia oxidation by *Nitrosomonas* at ammonia concentrations near K_m: Effects of pH and temperature. Water Res. **1994**, *28*, 2561–2566.
8. Diab, S. et al. Nitrification pattern in a fluctuating anaerobic-aerobic pond environment. Water Res. **1993**, *27*, 1469–1475.
9. Halling-Sørensen, B.; Jørgensen, S.E. *The Removal of Nitrogen Compounds from Waste Water*; Elsevier: Amsterdam, 1993; 444 pp.
10. Islam, S.; Suidan, M.T. Electrolytic denitrification. Water Res. **1998**, *32*, 528–536.
11. Sison, N.F. et al. Denitrification with external carbon source utilizing adsorption and desorption capability of activated carbon. Water Res. **1996**, *30*, 217–227.

26

Water Quality and Quantity: Globalization

Kristi Denise Caravella and Jocilyn Danise Martinez

Introduction

Water is one of our most precious natural resources, covering nearly 70% of the Earth's surface, but only 2.5% of the Earth's water is fresh and suitable for consumption.[1] Thus, it is quite an understatement to say that freshwater is a scarce resource for the nearly 7 billion people living on Earth today.[2] Many nations around the world are currently ill equipped to deal with the demand for water. Nearly 1 billion people lack access to safe water.[3] The health and economic impacts of inaccessibility to water are staggering.

It is predicted that the problem will worsen with the increase in human population and the continued growth of the global economy. By 2050, the world population is expected to grow to more than 9 billion people, placing a strain on water utility systems around the world to meet the growing demand for the human consumption of water that is necessary for survival coupled with the need to utilize water as a tool of industry to maintain a competitive edge in the global marketplace.[3] Nations around the world must develop a coherent plan of action to deal with this problem. This entry discusses management issues related to water quality and quantity concerns that have occurred as a result of increased globalization. The purpose of discussing these issues is to mitigate water quality and quantity concerns that can have an adverse effect on human health and the environment on a global scale.

The results of globalization have created an international environment whereby economic integration is essential. Since World War II, the global trend toward greater interdependence among nation-states in the exchange of goods, services, and investments has accelerated. Much of this growth can be attributed to transnational corporations who have used technology to capitalize on expanding trade partnerships and exchanges. This increased international trade has led to an increase in global production, which has placed a strain on the environment. With regard to global water resources, urbanization led to heavy exploitation due to major advances in geological knowledge, well drilling, pump technology, and rural electrification. Given these circumstances, it is becoming increasingly important to put issues such as water quality and quantity in a global context.

The environment, water, and food production are closely interrelated at the international, local, and regional levels. Water is the major limiting factor for world agricultural production and implicitly tied to global food production.[4] As the human population continues to swell, food production will hinge on the ability to manage, preserve, and enhance global water resources.

Throughout the world, freshwater is used for drinking, industrial production, irrigation, transportation, recreation, waste processing, hydroelectric power, and a habitat for aquatic life.[5] Therefore, issues of contamination and local water depletion are intimately connected to the global economy. Increased global environmental trade, which increases the demand for “virtual” water, or water used to produce crops traded in international markets, can result in declines in water quality and quantity.[6] This entry discusses these important water resource topics from a global environmental management perspective.

Globalization and Water Resources

As globalization has become a more pervasive force, nations around the world have created development plans that prioritize industry over the protection of natural resources such as water.[7] This has become an increasingly notable trend in poorer nations. Many leaders in these nations see the growth of industry as the only realistic way to grow national industry in order to become a competitive force in the global economy and to increase quality of life standards for their citizens. These developing nations have been industrializing at a much faster pace than ever before, which can place an extreme burden on the environment. Some Asian countries (i.e., China) are industrializing in 20 years or less whereas the United States and Europe spent more than 100 years industrializing. However, following in the footsteps at the rate of developing nations such as Europe, Japan, and the United States can lead to a severe degradation of the environment.[7] The Millennium Ecosystem Assessment conducted by the United Nations (UN) finds that although there have been substantial gains in human well-being over the past 50 years due to industrial growth, broad ecosystems around the world are now degraded and under pressure.[8]

With rapidly increasing water demands, the competition among household, industrial, environmental, and agricultural water uses has been escalating in many regions as they are already limited by the amount and quality of available water. Human impacts on the quality and quantity of freshwater threaten economic prosperity, social stability, and the resilience of ecosystem services and natural capital.[5] According to Meyer and Turner,[9] human production and consumption is “unprecedented in its magnitude and rate and has truly reached a global scale” (p. 39).

Not only is human consumption of water an important aspect of water management, the environmental impacts of the changing hydrologic cycle can have an adverse effect on our accessibility to water. It is estimated that “less than 3 percent of all water is fresh enough to drink or to irrigate crops, and of that total, more than two-thirds is locked in glaciers and ice caps”[5] (p. 1029). In the coming century, global climate change and a growing imbalance among freshwater supply, consumption, and population will alter the water cycle dramatically.

Environmental changes that involve the “flow of materials and energy through the chain of extraction, production, consumption, and disposal of modem industrial society” or the “alteration of the land surface (i.e. wetlands, rainforests) and its biotic cover” become global either by “affecting a globally fluid system (the atmosphere, world climate, sea level) or by occurring in a localized or patchwork fashion in enough places to sum up to a globally significant total”[9] (p. 39). It is these two types of environmental change that can affect water resources at a global level.

Political Control and Global Water Management: Policies and Political Structure

With the growing demands on freshwater resources globally, there is an urgent need to link water policy research with improved water management strategies. Water resources sustainability means using water wisely and protecting complex ecosystems with future generations in mind.[10] Better monitoring, assessment, and forecasting of water resources will be needed to allocate water more efficiently among competing needs and to achieve water resource sustainability.[5] This will require coordination from a broad, ideally a global, perspective.

On an international level, the United Nations (UN) plays a critical role in water policy discussions. The UN defines public water policy as "the legislation and regulations that underpin water management"[11] (p. 7). The UN supports the goal of sustainable water management that "holistically addresses equity, economy, and the environment in a way that maintains the supply and quality of water for a variety of needs over the long-term, and ensures meaningful participation by all water resource management stakeholders"[11] (p. 7). The UN supports four dimensions of sustainability that include social sustainability (ensuring that the public has access to adequate and affordable water services), environmental sustainability (protecting against threats to the environment), economic sustainability (balancing costs), and institutional sustainability (establishing institutional frameworks that can function over a long-term period).

Much of the trouble in creating a framework that provides for international consensus, however, is that policymaking at the international level is decentralized and fragmented with multiple stakeholders. The primary stakeholders that serve as official and unofficial actors in issues of global water policy include government agencies at multiple levels, regulatory agencies, private sector firms, consultants, scientists, trade union groups, civil society, and religious groups.[12] Each of these stakeholders seeks to pursue their own individual agendas. While this statement is easily applicable to a variety of global policy issues, it is particularly well suited to global water policy issues because of the disparity in terms of how water policy is treated among nations. Essentially, public water policymaking occurs at all levels of government. The primary legislative framework is developed at the national level, whereas management and operational aspects are implemented at the local level. For example, American water policy involves a combination of state, local, and federal laws dealing with surface water and groundwater allocation and quality protection. Surface water and groundwater allocation is almost exclusively the province of state law, whereas surface water and groundwater quality protection is a mixture of state, local, and federal laws.[13] This fragmented decision-making structure leads to an increase in the number of stakeholders that influence water policymaking within and, inevitably, among nations.

There is fragmentation on water policy not only at the international level but also at the national level. Another aspect of fragmented water policy occurs when discussing specific hydrologic cycle issues and ecological water budgets. Precise data about regional hydrologic cycle issues and area-specific estimates of water supply and demand are difficult to achieve at the national level.[14] Because water policy decision making occurs at the national level in most nations, region-specific issues are the responsibility of state and local policymakers and not applicable to other regions. For example, it is often the case that hydrological knowledge and methodologies obtained for one specific geographic region cannot necessarily be adapted to other regions. This is especially problematic in large nations that have a multitude of differing geographical conditions within its political boundaries, such as the United States and the European Union. Policy fragmentation also occurs as a result in a changing conceptualization of the hydrologic cycle. Classical thinking about the hydrologic cycle has focused on the large-scale (or macro) aspects of the cycle, while largely ignoring the smaller microcycles that occur within it. These microcycles are an important factor in water resource recovery that can help to restore the natural functions of aquatic systems around the world to help achieve sustainability within watersheds. Restructuring these thoughts about water policymaking to include the smaller microcycles that occur in the classic hydrologic cycle model can be a positive step toward the successful regeneration of these ecological systems, but does little to achieve cohesion in national water policymaking.

A negative side effect of this fragmented policymaking is the lack of collaboration and equity in the context of water development and management with nations. When policy fragmentation occurs in this way, the design of policies, interventions, and programs aimed at the equitable distribution of benefits from water resources becomes impossible. It also leads to deficiency and irregularity in data collection and interpretation, which makes cohesive policymaking efforts next to impossible.[14]

Balance sheets of micro-, meso-, and macro-watersheds are another challenge in the policy environment of water policymaking. Evolving a robust and rigorous methodology for assessing the impact of watershed (or river basin) projects at the micro, meso, and macro levels is a major challenge facing academicians, policymakers, and practitioners.[15] Watershed development projects are a continuous process, rather than time-bound projects with a well-defined set of activities and clearly earmarked funds, which creates methodological problems for studying them. This leads to difficulties in creating policy that incorporates these resources. In addition, there has been disagreement on the indicators for success. Overall, studies at the watershed scale have been inconclusive with regard to positive effects on soil conservation, sustainability, or environmental regeneration.[15] Despite limitations, watershed projects have led to at least two positive outcomes, "creation of a large number of structures for soil water conservation; and … setting up of participatory institutions for management of the local resources"[15] (p. 1). Recent watershed studies include studies in India, the Indo-German Watershed Project, Latin America and the Caribbean, the Nile Basin (Africa), the Thukela river basin of South Africa, and the Pangani river basin of Tanzania.

In order for watershed projects to have significance in the development of water policy, Shah et al.[15] state the following issues must be addressed: "technology, physical treatments and economic returns; (creating) institutions that ensure equity and efficiency in resource use; and (creating) market linkages essential for enabling the project to sustain beyond the initial funding, and for spreading out to larger number of beneficiaries as well as areas" (p. 5). At the micro level, watershed projects will inherently require negotiation among territories. At the meso and macro levels, watershed projects will require extensive (often international), integrated collaboration among participatory institutions that can be very costly.[16] Finally, water policy that seeks to address issues at the watershed scale must address socioeconomic issues that affect utilization of these resources.

There is a large disparity between water policymaking when looking at developed and developing countries.[17] There is also disparity when looking at water policymaking among different developed nations. In many developed nations, issues of water quality and quantity are managed by a well-established regulatory framework that monitors the public water supply. Many developing nations do not have such a political framework in place at the national level, leaving most of the policymaking decisions for local leaders. This can lead to duplication in implementation and inaction at the local level. Ultimately, this can surmount to even greater disparity within nations, which further limits abilities to develop a cohesive national water policy agenda and regulatory framework.

Global Surface Water and Groundwater Concerns and Issues: Introduction

William Jury and Henry Vaux, at a colloquium held in 2004 to address emerging global water problems, stated that the "optimum management of global water resources presents one of the most crucial challenges of the 21st century" (p. 15,715). Of all the water resource problems, scarcity is considered to be the most serious. According to Jury and Vaux,[18] in much of the world, existing water supplies are already insufficient to meet all of the urban, industrial, agricultural, and environmental demands.

Globalization also has led to an increase in the contamination of water resources through urbanization, agricultural intensification, and land degradation. The lack of a global water quality standard has led to varying levels of water quality around the world. This is significant as water resource contamination can have global effects on human health and the environment through increased international trade and resource sharing.

Global water concerns can be divided into two main areas: water contamination and groundwater, or aquifer, depletion. The next two sections outline the major causes of these two concerns and current issues that are being explored to rectify these issues from a global perspective.

Surface Water and Groundwater Contamination

Quality surface water and groundwater in abundance is needed for human health, agriculture, industry, and recreation. Our surface water resources have a long history of being adversely affected by point sources such as industrial and municipal waste disposal, animal waste, recreational vehicles, hazardous product spills, and air deposition of particulates. Surface water is typically heavily contaminated and is not available for usage in the water supply. According to the Food and Agriculture Organization of the UN, more than 5 million people die annually from waterborne illnesses from both surface water and groundwater contamination. Roads, railways, pipelines, hydroelectric corridors, runoff from mines and mine wastes, quarries, well sites, a large variety of recreational land uses including ski resorts, boating and marinas, campgrounds, and parks, leachates, and gases (as a result of irrigation) contribute to surface water and, eventually, groundwater contamination.

Groundwater is water that is present below the surface of the Earth in underground sources (i.e., streams and aquifers). It is one of the most important natural resources in any nation and serves as a major source of water in communities around the world for households, industries, and agricultural purposes. In comparison to surface water, it is generally considered to be much less susceptible to contamination. Hence, groundwater becomes the main source for potable water in many parts of the world.

Groundwater contamination is one of the most serious problems that can have an impact on water quality. In the past 50 years, groundwater contamination has become one of the most serious environmental problems in the world because once polluted, the remediation of aquifers may be very difficult and even impossible.[19] A groundwater contaminant is any substance that impacts the quality of groundwater. Groundwater contaminants include substances found in storage tanks, septic systems, hazardous waste sites, landfills, road salts, fertilizers, pesticides, and other chemicals.

Problems with water quality have increased primarily due to contaminations associated with human and land-use activities including agricultural activities, leaking municipal sewage treatment facilities, poor septic management, uncontrolled cattle access to water bodies, effluents from industries, and runoff from urban areas.[19] This is a particularly dangerous problem when, as noted, groundwater sources constitute the main drinking water supply in a community, as is the case in most nations. Groundwater contamination also can occur as a result of contamination in surface waters. If contaminants are present in other parts of the hydrologic cycle (i.e., surface water), these contaminants can also be transferred into groundwater supplies.

Typically, groundwater contamination does not occur as a result of a lack of legislation that establishes rules and guidelines regarding groundwater contamination. In other words, much of the problem is not that anticontamination laws do not exist, but that these laws are not necessarily enforced. In China, the major culprit of worsening groundwater pollution is industrial wastes.[20] Many industrial plants discharge waste without proper treatment because environmental protection laws are not enforced. This lack of enforcement coupled with the increase in human activities, largely due to population growth and increasing industrialization, has led to an increase in groundwater contamination. In India, the primary sources of groundwater contamination include untreated wastewater (which can lead to death and disease from sewage-borne infection), large quantities of industrial effluents being discharged, and agricultural runoff filled with hazardous chemicals found in fertilizers and pesticides that taints groundwater sources.[21] These factors are largely attributed to the fact that contamination laws are not enforced. Lack of enforcement can lead not only to decreasing water quality but also to human fatalities. For example, high levels of arsenic found in the surficial aquifer Simav Plain in Kutahya, Turkey, have been linked to an increase in deaths related to gastrointestinal cancers from 1995 to 2005.[22]

Because groundwater serves as the primary source of potable water in many communities, water quality is a critical aspect of the international water crisis. Part of the problem in regulating these systems on a global level is the large amount of diversity observed in the provision of water and policy approaches to contamination among nations around the world and the different regulatory models adopted in different countries. National programs of water quality monitoring throughout the world range from limited to non-existent. Few developing countries include water quality within a meaningful national water policy context. Water supply is considered to be a national issue while contamination is often considered to be a local problem. This can be problematic as national policy frameworks develop that fail to take into consideration the differences in regions.

Many of the national programs that do exist are dysfunctional due to years of neglect, chronic underfunding, and lack of focus. In some countries, water utilities are natural monopolies whereby public or private owners of the system are able to exercise considerable control over pricing and service quality. This places users of that system in a vulnerable situation whereby they may have no other options for water access. In some developing countries, there may be more competition among water utility vendors. This gives users greater options in terms of accessing quality water.

Yet, the regulation of water systems is critical for economic and social development of a nation. Regulatory agencies have been created and established in many different countries to help police utility systems and ensure that water quality guidelines are followed. A recent study conducted by Marques[17] identified 136 water service regulatory authorities worldwide (in January 2008). These regulators are spread across 57 countries, 12 of which are in Africa, 5 in Asia, 16 in Europe (2 in Euro-Asia), 2 in Oceania, and 22 in America. About 25% of countries have regulators for the water sector. Marques[17] suggests that the trend is moving toward nations creating new agencies to provide for the regulation of their water utilities.

In terms of water quality issues, many national governments in developed countries have created a system of rules and guidelines that establish minimum criteria that utility systems must follow in the provision of drinking water to users of that system. Although the implementation of such rules can provide heavy cost burdens on those utilities, the framework provides a uniform standardization mechanism that can assure the public that the water they are consuming is safe.

However, even among developed countries, there is little consistency among the treatment of water quality issues at the national level. Some have well-developed regulatory frameworks, while others do not. The United States passed the Safe Water Drinking Act of 1974 to regulate the nation's public drinking water supply in an effort to protect public health. The Act authorizes the United States Environmental Protection Agency (USEPA) to set national health-based standards for drinking water to protect against both naturally occurring and man-made contaminants that may be found in drinking water. USEPA, individual states, and water systems then work together to make sure that these standards are met. The European Union established a Drinking Water Directive in 1983 that protects human health by establishing healthiness and purity requirements for water. The directive set quality standards for drinking water at the tap and obliges member states to regularly monitor drinking water quality.[23] The Japanese national government has not only established Environmental Water Quality Standards and water pollution control measures but also established uniform water quality regulations for public water and groundwater that operators of utility systems must implement.[24]

Surprisingly, even some larger developed countries lack an established framework. In Canada, there is no national strategy in place to address water issues and no federal leadership to conserve, protect, and regulate Canada's water. The Federal Water Policy is more than 30 years old and needs to be updated to establish more rigid criteria for water quality monitoring. It is ironic that Canada's economy is built on the myth of an abundance of freshwater, yet only 1% of Canada's freshwater is considered renewable [25] Often associated with increased regulations is a shift toward computerized water quality monitoring. Data-driven approaches to water quality can be costly and difficult to implement in developing countries.[26] Over the last several decades, there has been a large shift toward the purchase of technological upgrades that help monitor water quality and collect data in developing countries. Water quality monitoring, as developed in Western countries, is based on the premise that with enough data, a well-designed program can answer most types of water quality management issues. Traditionally, developing countries are data-poor environments. This poses a major challenge for environmental management and decision making because nations that lack water monitoring technology are not able to obtain data about the water quality in order to make appropriate surface water and groundwater contamination policy.[26] In light of the lack of enforceable regulatory systems and lack of knowledge about the actual quality of water in many nations, surface water and groundwater contamination are critical concerns for many nations around the world in a national context. Because of the diverse nature of water quality monitoring in individual nations, forming a cohesive policy regarding groundwater contamination at an international level presents considerable challenges.

Groundwater and Aquifer Depletion

By far, the most abundant and available source of freshwater is underground water supplies or well-springs known as *aquifers.* Groundwater is the primary source of water for drinking and irrigation. At least one-fourth of the world's population draws its water from underground aquifers and approximately 99% of all liquid freshwater is found in underground aquifers.[5] Serious difficulties exist in fairly allocating the world's freshwater resources between and within countries.[27]

Conflicts are escalating among new industrial, agricultural, and urban sectors. According to the Global Environment Facility, worldwide such conflicts have increased from an average of 5 per year in the 1980s to 22 in 2000.[27] In 23 countries for which data are available, the cost of conflicts related to the agricultural use of water was an estimated $55 billion between 1990 and 1997. At least 20 nations obtain more than half their water from rivers that cross national boundaries,[28] and 14 countries receive 70% or more of their surface water resources from rivers that are outside their borders.[29,30] One way that governments have sought to deal with this imbalance has been to pass laws that set limits on water allocations.

Some of the major causes of groundwater depletion include population growth and increased demand for food supply, increased trade, increased agricultural irrigation, and environmental changes including changes in rainfall, temperature, evaporation rates, soil quality, vegetation type, and water runoff. This section reviews these causes and discusses remedies where available.

In the next 30 years, the Earth's population is projected to rise by approximately one-third (or 3 billion people).[5] This growth in global population is projected to be at least 3 times greater than accessible freshwater runoff. The growth in global population and water consumption will place additional pressure on freshwater resources. People use freshwater for many purposes including irrigation for crop production, industrial and commercial activities, and residential activities. Some water that is used by people can be reused; herein lays the difference between water consumption and water use. Water consumption occurs when there is no ability to reuse the water.[27] For example, when humans use water for sanitation, it can be reused and is therefore not consumed.

To make matters worse, most of the world's population growth will occur in developing countries where water is already critically short and many of the residents are impoverished. It is estimated that, today, more than 1 billion people do not have access to safe and affordable drinking water and perhaps twice that many lack adequate sanitation services.[18] Particularly in cities and towns around the world where there is population density, there is just not large enough a recharge area to support the needs of their inhabitants on a sustainable basis.[31] For example, in Beijing, China, a highly populated city, prohibitions have been placed on farmers from using local water for irrigation. Also in the Gulf Coast countries (Bahrain, Kuwait, Oman, Qatar, Saudi Arabia, and the United Arab Emirates) where there is arid climate, these countries are experiencing a severe water shortage problem that "threatens the sustainable development and hinders the national plans for human, industrial and agricultural development" throughout the region[32] (p. 59).

While floodwater is one way to replenish global water resources, floodwater requires large-scale management and coordination of water systems including reservoirs and dams. Water for irrigation, industry, and household uses must be delivered in controlled quantities at specific times, making floodwater an impractical solution to renewing water resources without the necessary management in place.

Also, as the human population has increased, so has trade among nations and more water is used to produce these goods and services. As a result of this increase in environmental trade and urbanization, countries have tried to overproduce in an effort to maintain a competitive edge in the global marketplace. This has placed a burden on the groundwater sources in many nations that have overpumped aquifers in an effort to increase production and trade. This is most notable in the agricultural community. For example, the three main grain producers, China, United States, and India, have all felt the effects of overpumping, or pumping that exceeds the long-term average rate of replenishment. Ultimately, overpumping can lead to a depletion of aquifers and consumption of aquifer reserves, which may lead to cutbacks in grain harvests. Evidence of substantial and widespread drawdown of the piezometric surface beneath many Asian cities, as a result of heavy exploitation of aquifers, also has been accumulating since the early 1970s.

The increase in agricultural production, in both crops and livestock, is significant as much of the water that is used for production cannot be recovered or reused; in essence, it is consumed. According to the United Nations Educational, Scientific and Cultural Organization (UNESCO), world agriculture consumes 70% of the freshwater withdrawn per year.[27]

Increased crop production and irrigation negatively affects soil quality and water runoff. Irrigation causes soil erosion, which diminishes the soil quality by reducing soil depth and soil nutrients. Other problems and failures of irrigated agriculture include groundwater overdraft, water quality reduction, water logging, and salinization.[33] While these various forms of soil degradation are proven to affect water resources, on a global level, they are very hard to track.[9]

Increased livestock production directly and indirectly affects the depletion of water resources in that not only do livestock directly consume freshwater but water is also consumed in the production of grains that are fed to the livestock. According to the U.S. Department of Agriculture (USDA), increased crop and livestock production during the next five to seven decades will significantly increase the demand on all water resources, especially in the western, southern, and central United States and in many regions of the world with low rainfall.[27] Strategies such as crop rotation, replacement of trees and shrubs, soil maintenance, and the use of organic mulch promote water conservation by limiting the damage done by crop production.[27]

Appropriate water pricing is important for improved water demand and conservation of water.[27] The relatively high cost of treating and delivering water has led many world governments to subsidize water for agricultural and household use. For example, some U.S. farmers pay as little as $0.01 to $0.05 per 1000 L used in irrigation, while the public pays $0.30 to $0.80 per 1000 L treated water for personal use.[34] According to the World Bank,[35] the objectives of fair water pricing are 1) to seek revenue to pay for the operations and maintenance of water availability; 2) to improve water use efficiency; and 3) to recover the full costs of water pumping and treatment.[36]

Finally, there is a growing scientific consensus that the buildup of greenhouse gases in the atmosphere is warming the Earth.[5] The continued loss of forests and other vegetation and the accumulation of carbon dioxide, methane gas, and nitrous oxide in the atmosphere are projected to lead to global climate change or global warming. Over time, such changes may alter precipitation and temperature patterns throughout the world.[27] These temperature changes might increase soil erosion and deforestation, which could result in greater reductions in water availability worldwide. Salati and Vose[37] claim that regional forest clearance would have severe climatic consequences. Because of the high proportion of water recycled by the rainforest, deforestation would significantly lessen rainfall and increase temperatures.

Global Water Resource Policy: Toward a Potential Framework

As noted throughout this entry, there are many efforts at the national level to conserve water resources, but what is lacking is an integrated worldwide water resource management system. A well-coordinated international plan for managing the diverse and growing pressures on freshwater systems and for establishing goals and research priorities for cross-cutting water issues is required. At the most basic level, water resource management must address how to "secure water and related services, avoid degradation of water and land resources and of ecosystem integrity, and foresee changes (climate, population, diet preferences etc.)"[38] (p. 14).

This plan will have to incorporate the many national, transnational, and private organizations that already exist into global water resources conglomerations or networks. Collaboration among this myriad of agencies will inevitably require a change in the culture of water management where water agencies have been known to be proprietary in nature. In addition, a redefinition of water policy from a political, economic, cultural, and social perspective is needed.

Foster and Chilton[39] note that nations and stakeholders need to "appreciate their social and economic dependency on groundwater, and to invest in strengthening institutional provisions and building institutional capacity for its improved management" (p. 1970). This can be achieved by creating global communication strategies to fuse the dialogue between the many water management stakeholders including scientists, policymakers, water managers, and citizens.[18]

Foster and Chilton[39] also state that local field agencies might have to change their roles to aid in the development of a broader global water management system. Jury and Vaux[18] agree that many existing water management institutions were "created in times and eras when the problems of developing and managing water resources were very different from what they are today" and would therefore require substantial changes to their institutional arrangements. Currently, water management institutions tend to focus on a narrow set of interests, but with the global nature of water resources management today, these institutional arrangements will no longer be beneficial to solving world water problems.

For example, Benvenisti[40] argues that there is enough freshwater in the world to meet the existing and future needs of the world's population but that this water is poorly distributed, making the management of global water resources predominately about redistribution. Shah et al.[31] agrees that redistribution and "the co-existence of regions with undeveloped (water) resource and those with overdeveloped (water) resource" is central to the future of water management policy.

Redistribution policy management inevitably will bring into play the competing priorities of different uses and users, and since most water resources traverse political boundaries, these competing priorities often become regional conflicts between states. Policies aimed at redistribution across political boundaries must then be of "collective action despite the varied self-interest" of these regions[40] (p. 385). Benvenisti recommends an international law policy framework that borrows aspects of game theory, economics, contract legal theory, international relations theory, and international negotiations theory to encourage the cooperation needed for the redistribution of global water resources.

In addition, to move toward a global policy framework, much more precise and spatially congruent data are needed. Meyer and Turner[9] note that the reluctance of national and international institutions to fund the collection of social data is an obstacle that must be addressed. Foster and Chilton[39] note that, globally, current data used to assess the status of aquifer degradation are of questionable reliability, inadequate coverage, and poor compilation. Since data on aquifer depletion and water contamination are collected by many different agencies that might focus on different aspects of water management including engineering sciences, hydrology, climatology, and geology, institutional arrangements that can bring together these interdisciplinary concepts into a comprehensive data management system will have to be created.[18]

Finally, while there is a general consensus on some of the future challenges in a global water management system, there are still many unknown factors that may arise. Jackson et al.[5] state that "uncertainty will be the most important feature of freshwater forecasts" and that by "evaluating uncertainties, forecasters can help decision makers to anticipate the range of possible outcomes and to design flexible responses" (p. 1038). Planning for uncertainties requires "adaptive management" or the "process of designing management interventions to decrease the variance of future forecasts and recommend alternative management options" (p. 1038). Future global management of freshwater systems must, therefore, incorporate adaptive management strategies.

Conclusion

The major water concerns discussed in this entry are associated with the prevailing cultural, political, economic, scientific, and social perceptions about how humans should interact with the environment. A redefinition of water policy may lead to changes in these perceptions. On a larger scale, the global water discussion requires a redefinition of the relationships between industry and environment. A redefinition of water policy may lead to changes in human behavior that can help address some of the concerns discussed in this entry. For decades, the primary engine that has driven the growth of the global market is a continued focus of economic gain among nations. This tendency toward capitalist values has continually placed a greater value on industry at the expense of protecting the environment. Strategic thinking at the international level should emphasize sustainable development. There must be some consensus among policymakers on water policy goals. A global water policy agenda that focuses on sustainable development is necessary to meet the goals of protecting the Earth's water while meeting human water demand.

References

1. National Aeronautics and Space Administration, 2011, available at available at http://earthobservatory.nasa.gov/Features/WeighingWater/ (accessed September 7, 2011).
2. United Nations Educational, Scientific, and Cultural Organization (UNESCO), 2011, available at http://www.unesco.org/new/en/unesco/ (accessed September 7, 2011).
3. World Health Organization (WHO), 2011, available at http://www.who.int/en/ (accessed September 12, 2011).
4. Kendall, H.W.; Pimentel, D. Constraints on the expansion of the global food supply. Ambio **1994**, *23* (3), 198–205.
5. Jackson, R.B.; Carpenter, S.R.; Dahm, C.N.; McKnight, D.M.; Naiman, R.J., Postel, S.L.; Running, S.W. Water in a changing world. Ecol. Appl. **2001**, *11* (4), 1027–1045.
6. Wichelns, D. Virtual water: A helpful perspective, but not a sufficient policy criterion. Water Resour. Manage. **2010**, *24,* 2203–2219.
7. Steiner, G.; Steiner, J. *Business, Society, and Government;* McGraw-Hill: New York, 2005.
8. United Nations. Millennium Ecosystem Assessment, 2011, available at http://www.maweb.org/en/About (accessed September 10, 2001).
9. Meyer, W.B.; Turner, II, B.L. Human population growth and global land-use/cover change. Annu. Rev. Ecol. Syst. **1992**, *23,* 39–61.
10. Kinoti Mutiga, J.; Mavengano, S.T.; Zhongbo, S.; Woldai, T.; Becht, R. Water allocation as a planning tool to minimize water use conflicts in the Upper Ewaso Ng'iro North Basin, Kenya. Water Resour. Manage. **2010**, *24,* 3939–3959.
11. United Nations. United Nations Report on Guide to Responsible Business Engagement with Water Policy Report, 2010, available at http://www.unglobalcompact.org/docs/issues_doc/Environment/ceo_water_mandate/Water_Policy_Engagement_Guide_Public_Consultation_Draft.pdf (accessed September 7, 2011).
12. Ainuson, K. An advocacy coalition approach to water policy change in Ghana: A look at belief systems and policy oriented learning. J. Afr. Stud. Dev. **2009**, *1* (2), 016–027.

13. Tarlock, Dan A. Water Policy Adrift. Forum for Applied Research and Public Policy, Spring 2001, 63–70.
14. Deason, J.; Schad, T.; Sherk, G.W. Water policy in the United States: A perspective. Water Policy **2001**, *3*, 175–192.
15. Shah, A.; Devlal, R.; Joshi, H.; Desai, J.; Shenoy, R. *Benchmark Survey for Impact Assessment of Participatory Watershed Development Projects in India;* Gujarat Institute of Development Research: Ahmadabad, 2004.
16. San Martin, O. *Water Resources in Latin America and the Caribbean: Issues and Options;* Inter-American Development Bank Sustainable Development Department: Brazil, 2002.
17. Jury, W.A.; Vaux, H., Jr. The role of science in solving the world's emerging water problems. Proc. Natl. Acad. Sci. U. S. A. **2005**, *102* (44), 15715–15720.
18. Marques, R.C. *Regulation of Water and Wastewater Services*; IWA Publishing: Lisbon, 2010.
19. Shah, T.; Molden, D.; Sakthivadivel, R.; Seckler, D. *The Global Groundwater Situation: Overview of Opportunities and Challenges*; International Water Management Institute: Colombo, Sri Lanka, 2000.
20. Samake, M.; Tang, Z.; Hlaing, W.; Ndoh Mbue, I.; Kasereka, K.; Balogun, W. Groundwater vulnerability assessment in shallow aquifer in Linfen Basin, Shanxi Province, China using DRASTIC model. J. Sustainable Dev. **2011**, *4* (1), 53–71.
21. Aguilar, D. Groundwater reform in India: An equity and sustainability dilemma. Texas Int. Law J. **2011**, *46* (3), 623–653.
22. Gunduz, O.; Simsek, C.; Hasozbek, A. Arsenic pollution in the groundwater of Simav Plain, Turkey: Its impact on water quality and human health. Water, Air, Soil Pollut. **2010**, *205,* 43–62.
23. European Union Water Drinking Directive, 2011, available at http://ec.europa.eu/environment/water/water-drink/index_en.html (accessed September 7, 2011).
24. World Bank. International Trends in Water Pricing and Use, 2000, available at http://lnweb18.worldbank.org (accessed September 7, 2010).
25. The Council of Canadians, 2011, available at http://www.canadians.org/water/issues/policy/index.html (accessed September 12, 2011).
26. Ongley, E.D. Water quality management: Design, financing and sustainability considerations. Proceedings of the African Water Resources Policy Conference, Nairobi, May 26–28, 1999.
27. Pimentel, D.; Berger, B.; Filiberto, D.; Newton, M.; Wolfe, B.; Karabinakis, E.; Clark, S.; Poon, E.; Abbett, E.; Nandagopal, S. Water resources: Agricultural and environmental issues. BioScience **2004**, *54* (10), 909–918.
28. Gleick, P.H., Ed. *Water in Crisis: A Guide to the World's Fresh Water Resources*; Oxford University Press: New York, 1993.
29. Alavian, V. Shared Waters: Catalyst for Cooperation, 2003, available at http://ucowr.siu.edu/updates/pdf/V115_A2.pdf (accessed September 7, 2011).
30. Cech, T.V. *Principles of Water Resources: History, Development, Management and Policy;* John Wiley and Sons: New York, 2003.
31. Shah, T.; Molden, D.; Sakthivadivel, R.; Seckler, D. Global groundwater situation: Opportunities and challenges. Econ. Polit. Wkly. **2001**, *36* (43), 4142–4150.
32. Al-Rahsed, M.F.; Sherif, M.M. Water resources in the GCC countries: An overview. Water Resour. Manage. **2000**, *14,* 59–75.
33. Rosegrant, W.; Cai, X. Water constraints and environmental impacts of agricultural growth. Am. J. Agric. Econ. **2002**, *84* (3), 832–838.
34. Gleick, P.H. *The World's Water;* Island Press: Washington, DC, 2000.
35. World Bank. Water Resources Management in Japan. World Bank Analytical and Advisory Assistance Program, 2006, available at http://siteresources.worldbank.org (accessed September 2011).

36. Schalch, K. Providing safe drinking water to developing nations. National Public Radio, 2003, available at http://www.unesco.org/water/wwap/news/national_public_radio.shtml (accessed September 2011).
37. Salati, E.; Vose, P.B. Amazon basin: A system in equilibrium. Science **1984**, *225,* 129–138.
38. Falkenmark, M. Shift in thinking to address the 21st century hunger gap: Moving focus from blue to green water management. Water Resour. Manage. **2007**, *2*, 3–18.
39. Foster, S.S.D.; Chilton, P.J. Groundwater: The processes and global significance of aquifer degradation. Philos. Trans. R. Soc., B **2003**, *358* (1440), 1957–1972.
40. Benvenisti, E. Collective action in the utilization of shared freshwater: The challenges of international water resources. Am. J. Int. Law **1996**, *90* (3), 384–415.

Bibliography

1. Ahmad, S.; Prashar, D. Evaluating municipal water conservation policies using a dynamic simulation model. Water Resour. Manage. **2010**, *24,* 3371–3395.
2. Cobb, R.W.; Elder, C.D. *Participation in American Politics: The Dynamics of Agenda-Building,* 2nd Ed.; Allyn and Bacon: Boston, 1983.
3. Ellison, B. New thinking about water management. Public Adm. Rev. **2007**, *67* (5), 946–950.
4. EnvironmentalProtectionAgency, 2011, available at http://water.epa.gov/lawsregs/rulesregs/sdwa/index.cfm (accessed September 2011).
5. Food and Agriculture Organization of the United Nations. Control of Water Pollution from Agriculture. FAO Corporate Document Repository (accessed March 2012).
6. Gisser, M. Groundwater: Focusing on the real issue. J. Polit. Econ. **1983**, *91* (6), 1001–1027.
7. Gleick, P.H. Soft water paths. Nature **2002**, *418,* 373.
8. Hassan, R.; Scholes, R.; Ash, N. *Ecosystems and Well-Being: Current State and Trends;* Island Press: Washington, DC, 2006.
9. Ives, G.; Messerli, B. *The Himalayan Dilemma: Reconciling Development and Conservation*; Routledge: London and New York, 1989.
10. Kamel, N.H.; Sayyah, A.M.; Abdel-aal, A. Ground water in certain sites in Egypt and its treatments using a new modified ion exchange resin. J. Environ. Prot. **2011**, *2*, 435–444.
11. Leung, K.S. Public–private partnerships solve groundwater contamination issues. Am. Water Works Assoc. **2011**, *103* (1), 26–29.
12. López, E.; Schuhmacher, M.; Domingo, J. Human health risks of petroleum-contaminated groundwater. Environ. Sci. Pollut. Res. Int. **2008**, *15* (3), 278–288.
13. Shankar, B.S.; Balasubramanya, N.; Maruthesha, R. Impact of industrialization on groundwater quality—A case study of Peenya industrial area, Bangalore, India. Environ. Monit. Assess. **2008**, *142* (1–3), 263–268.
14. Wang, M. Optimal environmental management strategy and implementation for groundwater contamination prevention and restoration. Environ. Manage. **2006**, *37* (4), 553–565.
15. Water demand creates policy challenges. *Oxford Analytica Daily Brief Service,* p. 1, 2011.
16. Widdowson, M.; Mendez, III, E.; Chapelle, F. Estimating cleanup times for groundwater contamination remediation strategies. Am. Water Works Assoc. **2007**, *99* (3), 40–45.
17. Zaller, J.R. *The Nature and Origin of Mass Opinion;* Cambridge University Press: New York, 1992.

27

Water: Cost

Atif Kubursi and
Matthew Agarwala

Introduction

Is water really different from other commodities? While water is versatile and has many uses, it is generally considered as an asset, a consumer good, and a factor of production. If it is a resource, is it a renewable or a "non-renewable resource?" Is its value infinite and "thicker than blood?" Can a price for water be determined much like for any other commodity? Can water or water permits be traded? What constitutes an equitable distribution of this scarce resource? What is a fair price for water? What is the full-cost price of water that incorporates its many environmental functions?

These complex questions are provocative and can incite passionate debate, yet they are central to water resource management. In this note, we show that economics sheds useful light on how to address these questions. This is not to suggest that economics alone can answer all of them. There are complex social, environmental, moral, and strategic issues that need to be factored into such an analysis. These add complications to the general structure within which water prices will be considered but do not change in any fundamental way the conclusions reached.

Before going any further in answering some of these questions, it is useful to outline some of the specific attributes of water that distinguish it from other commodities. These attributes will also influence the way water prices are set and optimal allocations of this scarce resource may be affected.

- Water is a fugitive, reusable, and stochastically supplied resource whose production can be subject to economies of scale, which give rise to natural monopoly situations. In this respect, water has many of the characteristics of a common property resource and a quasi-public good.
- Water is typically a non-traded commodity that is rarely sold in a competitive market. There are few overt water markets where suppliers and demanders exchange water. Markets in water rights have emerged in several parts of the world; the most notable examples are in Australia, Colorado, California, and Argentina.[1–3] However, most of these markets are within national entities, are heavily regulated, and often represent simulated market solutions. There are few international examples of water trade, but it is not difficult to conceive schemes that would involve this trade.

- Water scarcity is not only about physical scarcity, but it is also complicated by economic scarcity where actual prices for water are typically only a fraction of the true scarcity price. When prices are below scarcity prices, waste and overuse are quickly observed.
- Water values generally differ from the price that would obtain in a free and competitive market. The distinction between price and value is crucial. A family's visit to a local park may have a price of $0, but the value of such a visit may be much greater. When prices and values diverge, market forces produce inefficient outcomes. Water often has social value that is above what private users are willing to pay for it.[4] The allocation of water often reflects national and social policies and priorities toward agriculture, the environment, food security, and national security that go beyond serving the interests of private farmers. Social and policy considerations apart, the diversion of actual prices from their scarcity values imposes social costs on the domestic economy as well as on neighboring countries. Water is part of the tragedy of the commons.[5]
- Water is not only a desirable commodity, but its availability is also critical for life. At subsistence levels, it has no substitute. Furthermore, that every person is entitled to a minimum quantity that is considered consistent with human dignity is now enshrined in international human rights law.[6,7]
- The secure supply of water in many regions of the world is quite low. Security of supply is defined here as the probability of its average flow availability for 9 out of 10 years. In Canada, a country rich in water, this probability is less than 30%, whereas it is less than 5% in the Middle East.[8]
- While the total water supply may be limited and few, if any, substitutes exist for it, there exist substantial possibilities for intersectoral and interregional substitutions. As well, there are a number of technologies and conservation programs that rationalize demand and raise the efficiency of its use.
- Part of the water scarcity crisis is the fact that agriculture on average uses about 70% of the total global supply.[9] It is typically the case that other needs are suppressed, but this leaves a wide room for intersectoral reallocations and shuffling water between users and uses.
- While the quantity of water is in short supply, concern for preserving its quality is perhaps more pressing. Pollution and saline intrusion of the aquifers are being increasingly recognized as critical factors in planning for the future.
- Water is the "universal solvent."[10] It performs numerous ecosystem services including the absorption, transportation, and dilution of pollutants. Water is indispensable to the environment's pollution abatement capacity.[11]

Competitive Markets and Water Prices

The scarcity of water suggests that it should have a price that reflects this scarcity. Since economics is particularly suited for dealing with how scarce resources are, or should be, allocated to various uses and users, it can shed useful light on its use and value.

Indeed, water has a value as does any scarce resource, input, or asset. There is a monetary equivalent to water, although this is not the way people typically speak about water. This value need not emerge from competitive markets. However, it has or could be constructed to have many of the characteristics that are associated with competitive prices. The price should reflect the scarcity of water and disseminate information about this relative scarcity. It should also invite, if not provoke, the correct responses (incentives) that prices in general are expected to do. Higher prices persuade consumers to economize on its use and suppliers to raise their quantities offered. In competitive markets, the market price is what buyers are willing to spend for an additional unit of the commodity and the extra cost of producing it (marginal cost). If the price is larger than the marginal cost of the last unit of water, this provides a signal that more is needed of the commodity and more should be produced.[12] Conversely, if the price is below marginal cost, additional units will not be produced as society will have to sacrifice more than what people would value the additional units.

It is generally accepted, however, that water is not bought and sold in competitive markets. This is because in the case of water, at least five of the basic properties of competitive markets are generally absent.[4] These five properties include the following:

- Free markets lead to an efficient allocation of scarce resources if these markets are characterized by competitive structures, that is, these markets include a large number of independent small sellers and a similarly large number of independent small buyers that no single supplier or buyer is significant enough to influence the price. Each and every buyer and seller in this market is a price taker.
- Competitive markets require freedom of entry and exit and that no barriers that preclude easy entrance or exit exist.
- The product must be homogeneous enough that each unit is quite similar to any other unit.
- For a free market to lead to an efficient allocation, externalities must be absent. In economics, an externality or spillover of an economic transaction is an impact on a party that is not directly involved in the transaction. An efficient allocation can emerge from a free market when social costs coincide with private costs. Water production, however, involves many "externalities."[4] In particular, extraction of water in one place reduces the amount available in another. Further, pumping water from an aquifer in one location can affect the cost of pumping elsewhere. Such externalities do not typically enter the private calculations of individual producers and drive a wedge between private and social costs.
- In a free market that efficiently allocates scarce resources, social benefits must coincide with private ones. If not, then (as in the case of cost externalities) the pursuit of private ends will not lead to socially optimal results. In the case of water, many uses have social benefits that exceed the private ones. The use of water in agriculture may result in benefits that exceed the private returns to farmers. Among these are food security, border security, environmental objectives, and national interest.

These conditions are often violated in the case of water, where water sources are relatively few; barriers to entry are real and high (high cost of infrastructure); a large gap exists between private and social costs; and benefits and water units are not homogeneous, with a large spectrum of different qualities observed. This, perhaps, is why it is often the case that water production facilities are owned by the state. In many respects, water is not a private good; it has what we alluded to above, many of the characteristics of quasi-public goods.

Although water is not bought or sold in competitive markets, it is possible to develop prices that reflect its value.[13–15] In practice, the full value of water is not revealed in competitive markets. It is possible, however, to build specific models of water allocations that simulate competitive conditions and where the optimal nature of markets is restored. The model can explicitly optimize the benefits to be obtained from water, taking into account the five points made above. It is also true that now we have a rich body of literature that informs the setting of water prices and tariffs to overcome some of the five points listed above. The prices that emerge from formal optimization models not only maximize profits for the producers (producer surplus) but also maximize the utility of the consumer (consumer surplus). Furthermore, as is shown later, a multi-objective framework can be developed to design prices that satisfy goals beyond pure economic efficiency. These constructed prices permit the optimal allocation of water to its best uses as would emerge under competitive conditions. In many respects, the constructed prices are designed to serve as guides to consumers and producers in much the same way as competitive prices.

Optimizing Model: Simulating a Competitive Water Market

The underlying economic theory of the "optimizing model" is simple and compelling.[14,16,17] (This section draws heavily on Fisher[4] and Fisher and Huber-Lee.[17]) In Figure 1, we show an individual household's demand curve for water, with the amounts of water (on the horizontal axis) that the household

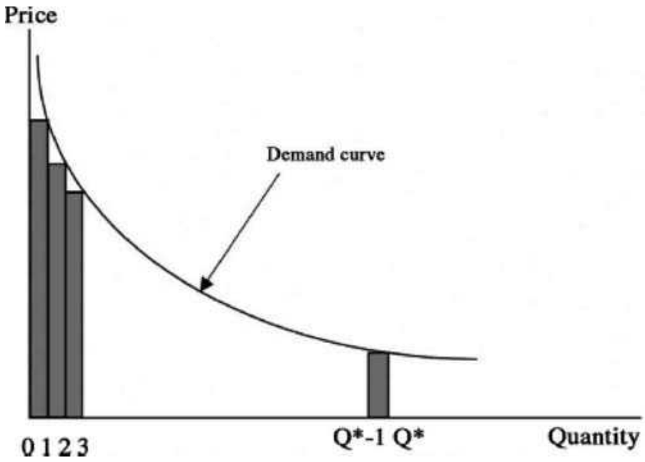


FIGURE 1 Demand curve for water.

will buy at various prices (on the vertical axis). The curve slopes down, representing the fact that the first few units of water are very valuable, while later units will be used for purposes less essential than drinking and cooking.

Now consider how much it will be worth to the household in question to have a quantity of water, *Q**, as depicted in Figure 1. We begin by asking how much the household would be willing to pay for the first small unit of water. The price that would be paid is given by a point on the curve above the interval on the horizontal axis from 0 to 1. Thus, the amount that would be paid is (approximately) the area of the leftmost vertical strip in Figure 1 (one unit of water times the price in question). Similarly, the amount that would be paid for a second unit can be approximated by the area of the second-to-left vertical strip, and so on until we reach *Q**. It is easy to see that if we make the size of the units of water smaller and smaller, the total amount that the household would be willing to pay to get *Q** approaches the area under the demand curve to the left of *Q**.

It is quite simple to reinterpret Figure 1 so that it represents not the demand curve of an individual household but the aggregate demand curve of all households in a given region. The gross (private) benefits from the water flow *Q** can thus be represented as the total area under the demand curve to the left of *Q**. These benefits are gross benefits as the cost of providing this water is not subtracted. To derive the net benefits from *Q**, we must subtract the costs of providing *Q**.

In Figure 2, the line labeled "marginal cost" shows the cost of providing an additional unit of water. That cost increases as more expensive sources of water are used. The area under the marginal-cost curve to the left of *Q** is the total cost of providing the supply, *Q**, to the households involved. Thus, the net benefit from providing *Q** to these households is the shaded area in the diagram, the area between the demand curve and the marginal-cost curve.

The amount of water that should be delivered so as to maximize the net benefits (the sum of consumer and producer surpluses) from water is *Q**, where the two curves intersect. This is the largest area between the two curves and represents the sum of the consumer surplus, defined as the difference of what households would be willing to pay and what they actually pay (the price that rules in the market at the intersection between supply and demand), and the producer surplus, defined as the amount of profits producers realize as the price of water exceeds its marginal cost. Both of these areas are defined in Figure 2.

Social value of water can exceed its private value as society may value this water beyond the benefits individuals would derive from it. A national policy to subsidize water for agriculture by say, 10 cents per cubic meter, at all quantities simply means that water to agriculture is worth 10 cents per cubic meter more to society than farmers are willing to pay for it.[17] This is represented in Figure 3. The lower

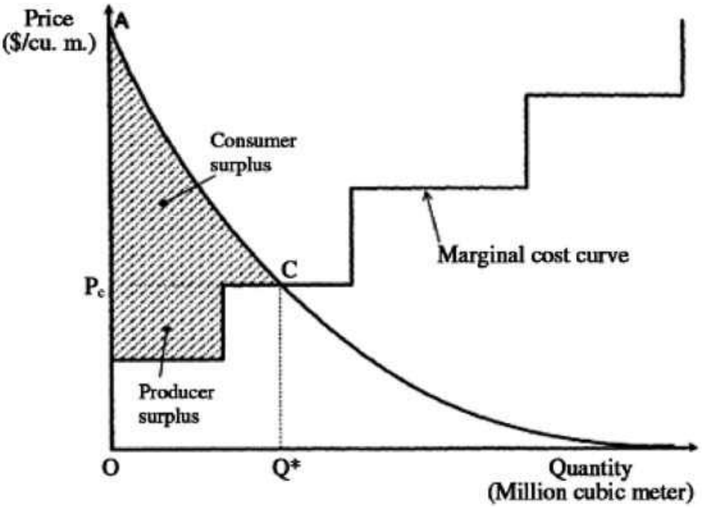


FIGURE 2 Net benefits from water.

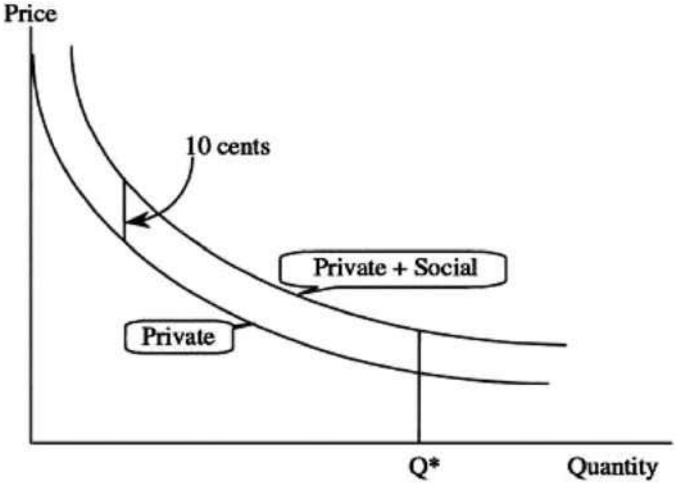


FIGURE 3 Social value of water as revealed by a subsidy.

demand curve represents the private value of water to agriculture; the upper demand curve also includes the additional public value as reflected in the policy, an additional value of 10 cents per cubic meter. As this illustrates, any consistent water policy can be represented as a change in the demand curve for water. Once such a policy has been included in the demand curves, the methods used above can be employed to measure net benefits.

Shadow Prices and Scarcity Rents

(This section too draws heavily on Fisher.[4])

As already discussed, purely private markets and the prices they generate cannot be expected to optimize net social benefits of water. Nevertheless, prices in an optimizing model can play a role very similar to that which they play in a system of competitive markets. Dr. Franklin M. Fisher[4] and his team developed such a model, called "WAS" (for "Water Allocation System") at Harvard University.[18] The model allocates water so as to maximize the net benefits (consumer surplus plus producer surplus) obtained from this allocation. The maximization of net benefits is done subject to constraints. An example of such constraints includes that at each location, the amount of water consumed cannot exceed the amount produced there plus net imports (imports minus exports) into that location.

Corresponding to the optimum quantities, there is a dual solution that defines the associated prices at the optimized level of activities. It is a general theorem that when maximization involves one or more constraints, there is a system of prices involved in the solution. These prices are called "shadow prices" or "Lagrange multipliers." Each shadow price shows the rate at which the quantity being maximized (here, net benefits from water) would increase if the associated constraint were relaxed by one unit at the optimal solution. In effect, the shadow price is the amount the maximizer would be just willing to pay (in terms of the quantity being maximized) to obtain an additional unit of the associated constrained quantity.[19]

The Role of Shadow Prices

Since water does not generally command a market price, economists have constructed models that estimate the scarcity (shadow) price of water that reflects many of its competitive equilibrium characteristics. The shadow price associated with any particular constraint shows the extent by which the net benefits from water would increase if that constraint were loosened by one unit. For example, where a pipeline is limited in capacity, the associated shadow price shows the amount by which benefits would increase per unit of pipeline capacity if that capacity were slightly increased. This is the amount that those benefiting would just be willing to pay for more capacity.

The shadow price of water at a given location is the amount by which the benefits to water users (consumers and producers) would increase were there an additional cubic meter per year available at that location. It is also the price that the buyers at that location who value additional water the most would just be willing to pay to obtain an additional cubic meter per year.

There are a number of defining characteristics of the shadow prices that need further explanation.[4] These include the following:

- Shadow prices are not necessarily the prices that water consumers are charged. That would be true in a purely private, free market system. However, in the WAS model, as in reality, the prices charged to some or all consumers can (and often will) be a matter of social or national policy. When such policy-driven prices are charged, the shadow prices of water will reflect the net benefits of additional water given the policies adopted.
- Shadow prices are outputs of the model solution, not inputs specified a priori. They depend on the policies and constraints put in by the user of the model.
- The shadow price of water in a given location does not generally equal the marginal cost of providing it there. If demand for water from a source is sufficiently high, the shadow price of that water will not be zero even if it costs nothing to produce it; benefits to water users would be increased if the capacity of the source were greater. Equivalently, buyers will be willing to pay a nonzero price for water in short supply, even though its direct marginal costs are zero.
- When demand at the source exceeds capacity, it is not costless to provide a particular user with an additional unit of water. That water can be provided only by depriving some other user of its benefits; that loss of benefits represents an opportunity cost. In other words, scarce resources have positive values and positive prices even if their direct marginal cost of production is zero. Such a positive value, the shadow price of the water in situ, is called a "scarcity rent."
- The shadow price of water used in any location equals the direct marginal cost plus the scarcity rent. For water in situ, the shadow price is the scarcity rent.

- Water will be produced at a given location only if the shadow price of water at that location exceeds the marginal cost of production. Equivalently, water will be produced only from sources whose scarcity rents are nonnegative.
- If water can be transported from location A to location B, then the shadow price of water at B can never exceed the shadow price at A by more than the cost of transporting water from A to B. If water is transported from A to B, then the shadow price at B will equal the shadow price at A plus the transportation costs. Equivalently, if water is transported from A to B, then the scarcity rent of that water should be the same in both locations.

Shadow prices generalize the role of market prices,[4] as can be seen from the following propositions:

- Where there are only private values involved, at each location, the shadow value of water is the price at which buyers of water would be just willing to buy and sellers of water just willing to sell an additional unit of water.
- Where social values do not coincide with private values, this need not hold. In particular, the shadow price of water at a given location is the price at which the user of water would be just willing to buy or sell an additional unit of water. That payment is calculated in terms of net benefits measured according to the user's own standards and values.
- Water in situ should be valued at its scarcity rent. That value is the price at which additional water is valued at any location at which it is used, less the direct marginal costs involved in transporting it there.

Structuring a Price (Tariff) of Water

Water utilities routinely design water tariffs and prices. Their choice of the price is motivated by considerations other than pure economics. They simply choose the price at a level that would help cover their operating costs plus a capital charge in addition to several other objectives, particularly environmental and sustainability objectives. Few utilities, if any, seek solely to generate sufficient revenue to cover their private costs. The generation of revenue is rarely the only purpose of a tariff, nor is it the sole consideration.[20] It has been generally recognized that water tariffs are powerful management tools with a number of complex and important functions. The tariff could create incentives for the efficient production and use of water. In this way, the tariff promotes environmental sustainability. As well, tariffs can be set to recover the full cost (including externalities, both economic and environmental) of an activity or structured to subsidize poorer customers.

Rogers, Bhatia, and Huber[21] make an interesting distinction between economic and environmental externalities. Economic externalities are more readily quantifiable and involve costs such as those imposed on downstream users by upstream diverters and polluters. Environmental externalities include the value of ecosystem degradation; the loss of biodiversity; and the health effects on human, plant, and animal populations. Each objective would lead to a different tariff design. The "best" design is the "one that strikes the most desirable balance among the different objectives."[20] However, is the "best" tariff the optimal one? Standard optimization theory tells us that net social benefits are maximized where marginal social benefit equals marginal cost. However, there is more than one cost perspective to consider when structuring an optimal tariff or price for water. Given the many different definitions for full supply cost, full economic cost, and full cost, the question remains as to which marginal cost should be equated with the marginal benefit. Full-cost pricing requires a quantitative assessment of water's intrinsic and cultural values, as well as the net of environmental externalities. The legitimacy of any such quantification is, however, severely limited by the subjectivity and unavailability of relevant data.[11,21,22] However, optimality requires quantification, and this has become a major issue in setting the optimal water tariff. The different definitions and their relevance for designing the optimal water tariff are presented in the table below.

Opportunity Cost

The opportunity cost of a commodity or service is the value of its next-best alternative use. If good X has only two possible uses, A and B, the opportunity cost of using good X for use A is the foregone benefit of employing it in use B. Similarly, the opportunity cost of using X for B is A. That is, it is the value of the foregone best alternative use. As Rogers, Bhatia, and Huber[21] argue, neglecting the value of water's opportunity costs "undervalues water, leads to failures to invest, and causes serious misallocation of the resource between users."

Full Supply Cost

This is the sum of costs directly involved in delivering water to the end user. This may include pumping, purification, storage, fixed infrastructure costs, capital investment costs, government taxes, and connection and metering costs. Full supply costs exclude all externalities and opportunity costs.[21] Capital and infrastructure costs are particularly heterogeneous vectors that vary widely depending on the types of delivery systems involved. Wells, house connections, pipes of various flow capacities, desalination projects, and recycled and treated water can lead to vastly different nominal values for full supply cost.[23]

Full Economic Cost

This is an economic concept that refers to the full supply cost plus the opportunity costs and net of economic externalities.

Full Cost

This is the full economic cost plus environmental externalities (these are more diffused and indirect and not easily quantifiable; Figure 4).

Substitutability and Environmental Considerations

Economic theory predicts that as the price of natural resource inputs rises, agents will substitute produced capital to take their place. The feasibility of the substitution of produced capital for natural resources is mixed. In some areas, the argument offers a viable solution. For example, the scarcity of whale oil, which resulted from overfishing, was made nearly irrelevant by the introduction of the electric light bulb. In telecommunications, fiber optics has been heavily substituted for copper. High energy costs promote the substitution of higher-quality insulation for fuel.[24–26] It is important to recognize and

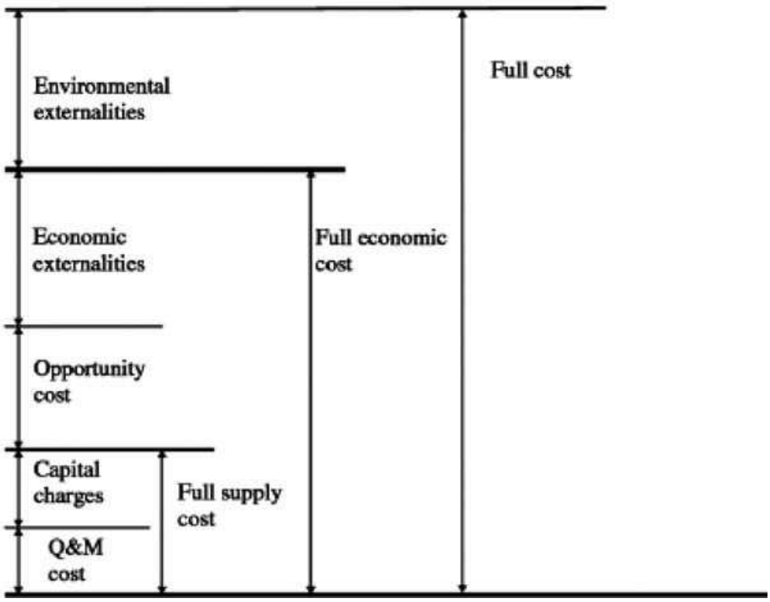


FIGURE 4 Full-cost prices.
Source: Rogers, Bhatia, and Huber.[21]

exploit the ability to substitute produced capital for natural resource inputs, but it is also important to recognize that these abilities are not universal and that they are not infinite.

This is particularly the case when it comes to water where the possibilities of substitution are more limited. As prices rise due to the introduction of full-cost pricing, human capital may be substituted for water in the form of genetically engineered seeds that yield a better harvest index. Water-scarce regions may shift production away from water-intensive activities to reflect relative scarcities (from producing wheat to producing barley) or import embodied water (virtual water, particularly agricultural products such as fruits and vegetables). Irrigation technologies that minimize evaporation and leakages will become more attractive in the face of rising water prices. Despite these possibilities, the inimitability of water and its centrality to life severely limit the potential scope and scale of substituting produced capital for water.

Given the fact that water prices are not determined in competitive markets, there exists an opportunity for economic policy makers to develop a pricing structure that incorporates broader goals. The multiple identities of water as an asset of society, a factor of production, and a consumer good are complementary and interdependent while simultaneously being rivalrous and independent. This presents a unique set of challenges for policy makers interested in setting the "right" price for water. Such a price must balance these competing identities while facilitating cooperation where they overlap. (We have abstracted in this entry from issues of intertemporal allocations of water. The models we have used are steady-state models. Treating water as an asset carries serious implications for intergenerational allocations. Suffice it to say here that sustainability requires that annual water extractions must not exceed the replenishment rates. A price that does not reflect this intertemporal conservation will misallocate resources over time.) Water as a factor of production and as an asset of society is inextricably linked by the concept of sustainable development. Similarly, the asset of society and the consumer goods are connected by the cultural and social identities associated with water use. Finally, the consumer goods and the factor of production are related by the externalities they impose upon each other. This analysis is not limited to the national level, and it is equally applicable in bilateral, multilateral, and global contexts. Given these relationships, the following principles and objectives have been identified as necessary ingredients in determining a price for water. Each of these is broadly defined and adopted in varying degrees with different weights placed on each from case to case. In Figure 5, we sketch a multiple-criteria framework for choosing prices that deal with complex objectives that balance economic, environmental, and sociocultural issues.

Economic Efficiency

Here, efficiency also incorporates multiple goals. Fees for water services must be sufficient to ensure that the utility recovers at least the full supply cost of delivery to the end user, a concept known as revenue sufficiency. Any fraction less than this fails to send the signal that water is a scarce resource; withholds information from consumers, the utility, and the regulatory agency; and prevents the proper inter-agent allocation of water (Figure 5).

It is worth noting that the standard approach to economic efficiency includes the concept of "full cost recovery," which the United Nations World Water Development Report defines as a situation in which "users pay the full cost of obtaining, collecting, treating, and distributing water, as well as collecting, treating, and disposing of wastewater."[27] In practice, as well as in theory, full cost recovery and revenue sufficiency refer only to the full supply cost.[20,21,28–30] We argue below that this is undesirable. This particular characterization is, however, of interest because it directly addresses the management of wastewater. That the importance of the economics of wastewater management has been largely ignored is demonstrated by the vast differences in progress toward the achievement of the Millennium Development Goals (MDG), and the new Sustainable Development Goals (SDGs) for water and sanitation.[27]

In stark contrast to convention, we argue that pricing at the full cost of supply does not support true economic efficiency. Supply-cost pricing brings us closer to the outcome that would exist if water were

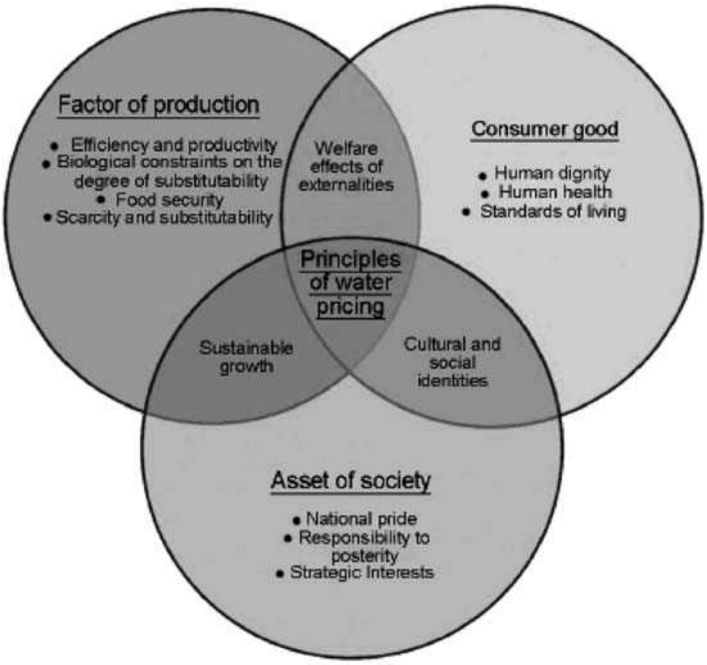


FIGURE 5 Multiple-criteria framework of prices.

traded in a perfectly competitive market. This price for water would fail to properly incorporate the value of opportunity costs, economic externalities, and environmental externalities.[21,31] Any such price prevents socially optimal allocations by promoting overuse, undervaluation, and waste. For this reason, we argue that wherever possible, full cost or full economic cost, as they are defined above, be used in the development of water tariffs.

Equity and Fairness

The principles of equity and fairness are contentious and have evoked controversy. However, given the increased recognition that a minimum amount of water is a basic human right, fairness imposes a constraint on water tariff setting. There are two dimensions to equity. On the one hand, equity has been defined as "treating equals equally" in the sense that those who purchase water that has identical costs should face identical prices.[20] On the other hand, equity and fairness are also grounded in the ideas of social justice, and water as a human right suggests that treating equals equally may satisfy only one dimension of equity. A balanced view of equity would also require treating unequals unequally. In areas of high-income inequality, poor users may be unable to afford even the minimum amount of water necessary to sustain basic human health and dignity. In such cases, some form of subsidy is required. This can take the form of internal cross-subsidization embedded in the pricing structure or of government transfers and income redistribution from other areas of the economy. The complexity of equity and fairness is hard to overstate, and policy makers are often forced to make difficult choices. The relative merits of different objectives for water subsidies are discussed in detail below.

Simplicity

Prices serve multiple objectives; one of the most important is to synthesize and convey information. Complex pricing structures confuse this information, making it difficult for producers, consumers, and regulators to respond appropriately to price signals. Simplicity and transparency also limit the potential for corruption. Insofar as simplicity facilitates the gathering and dissemination of information and economic signals, it promotes conservation, socially optimal allocations, and economic efficiency.

Sustainability

The concept of sustainability is central to each of water's roles depicted in Figure 5. Prices need to reflect this by encouraging conservation and substitution where possible.[22] A fundamental feature of the economics of sustainability is that assets are maintained over time. Water is a crucial component of natural capital. Depletion, diversion, and degradation of water assets represent an erosion of this natural capital asset, and therefore a depletion of society's wealth.

Stability and Quality

Here, stability refers to price stability, service stability, and revenue stability, all of which are interdependent. Stable prices convey stable information, avoid confusion, and promote smooth consumption. Revenue stability promotes and attracts investment while protecting service providers from shocks. Service stability reduces waste by eliminating the need to "hoard," eliminates time and travel costs of finding alternative sources of supply, promotes investment in connection to the system (which in turn increases the customer base, decreasing costs per user via economies of scale), and has the potential to raise revenue for producers as well as increase consumer surplus.

Evidence has shown that even in poor areas of developing countries, willingness to pay for improved service stability is well above current water prices. There are several reasons for this. First, the poor incur significant costs resulting from low-quality service, including travel and time costs, purification costs, reduced labor productivity due to illness and workdays lost, investment in individual water storage tanks, and purchasing expensive bottled water from private vendors.[29,32,33] A study conducted in Kathmandu, Nepal, showed that nearly 80% of respondents with a piped connection would be willing to pay as much as four times their current monthly water bill for improved water quality, stability, and more transparent billing.[32]

Water Tariffs and Pricing Strategies

The water tariff is the primary tool with which a policy maker can combine the various objectives and considerations into a socially optimal price regime. The tariff is central to the utility's ability to attract capital, creates incentives for efficiency in production and consumption, and distributes costs across consumers and time.[20]

Traditionally, decreasing block tariff (DBT) pricing strategies have been employed in water delivery systems for several reasons.[34] The DBT, as shown in Figure 6, involves prices based on quantity consumed.

The initial block, or quantity of water, costs the most, and prices decrease as consumption rises. DBTs score well on stability grounds as revenue is insulated from climatic shocks (to both supply and demand) and because the largest share of revenue comes from the initial units of water consumed. They exploit the economies of scale that arise from falling average costs as quantities supplied rise.[34] The problem, however, is that on grounds of efficiency, equity, and sustainability, this structure has less to offer than an increasing block tariff (IBT) scheme. In terms of efficiency, DBTs send exactly the wrong message to consumers. They fail to convey the fact that water is a scarce resource that must be conserved; they promote waste and fail to encourage socially optimal interagent, intersectoral, and intertemporal allocations. Furthermore, as the utility reaches capacity, the marginal cost of supply it faces rises; however, under the DBT regime, the marginal cost faced by consumers is falling. In terms of equity and fairness, the DBT fails to treat equals equally because consumers of water with identical costs of supply face

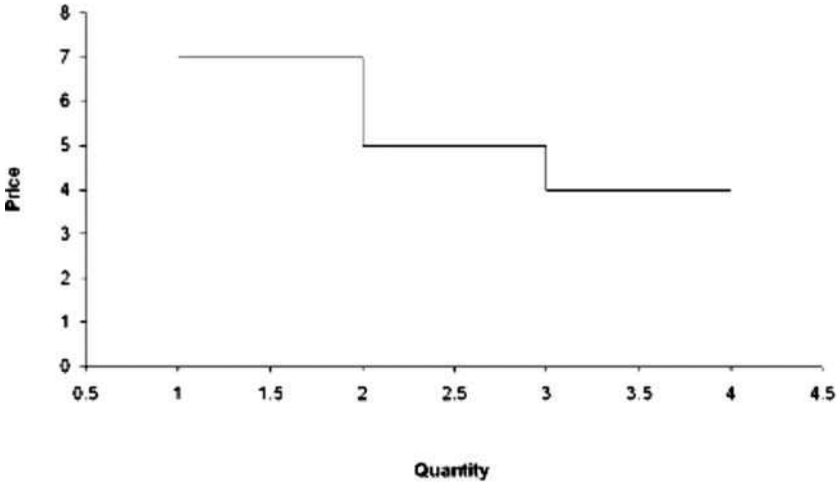


FIGURE 6 Decreasing block tariffs.

different prices. By charging higher prices for the initial units of water, the DBT also ensures that the poor who consume the least water face the highest prices (Figure 6).

In recent years, DBTs have largely been replaced by IBTs, which have become the strategy of choice in the developing world. This is because they have gained preference by donor institutions; management, engineering, and financial advisors; and water experts as being the most effective way of synthesizing the goals and principles of water supply and demand.[29] The IBT (the pricing scheme depicted in Figure 2) also features different prices based on quantities consumed; however, in this case, prices rise with quantity consumed.

In terms of economic efficiency, this is a significant improvement over the DBT. Here, tariffs are consistent with marginal-cost pricing and create an incentive structure that conveys the necessary signals to encourage socially optimal use. IBTs score well on fairness grounds because income and water demand are positively correlated and the prices faced by consumers rise with consumption. That is, IBTs are progressive, so wealthier customers that consume more water face higher prices. Furthermore, IBTs can exhibit “internal subsidization” or cross-subsidization in which large consumers pay sufficiently high prices that the first units of water consumed can be sold below cost. This feature enables the utility to provide water to the poor below the full cost of supply without violating the principle of revenue sufficiency.

The practical implementation of IBTs has yielded mixed results. If blocks are not properly structured in quantities that are relevant for the consumers in question, the benefits of IBTs are lost.[20,28,30,31,35] Specifically, the compatibility of fairness with the other principles of water pricing requires that the first block be sufficient to cover the basic requirements of human health and dignity but not more. Much attention must be paid to determining the relative block sizes and their prices, and finding the appropriate values for both requires data that are frequently unavailable.

It is possible that in some cases, low initial block prices may yield a result in which metering costs may exceed the revenue collected from low-income users. In such situations, the utility provider incurs a loss not only from the provision of water below full supply cost but also on the metering operations.[30] Alternatively, if tariffs faced by large-scale consumers such as heavy industry and agriculture are sufficiently high, these consumers may opt to disconnect from the utility and seek alternative sources of water, such as in-house direct extraction.[29,30] Even where this is illegal, government capacity to enforce such laws is frequently limited. The effect of this would be to remove the highest-paying customers, which in turn would eliminate the feasibility of internal subsidization, increase average costs, and distribute fixed-capital costs across a smaller customer base. That is, the benefits of IBTs are highly dependent upon the participation of the highest-paying customers; however, these are exactly the customers who have the greatest incentive to search for alternative sources of water.

Even where subsidies are absolutely necessary in order for the poor to meet their basic needs, it may not be ideal to provide these subsidies through the tariff structure. Chile subsidizes water consumption for the poor via transfer payments from municipal and national budgets. Many have claimed that this approach enhances equity and economic efficiency because the tariff received by the utility treats equals as equals.[30] Although we accept Yepes'[30] claim in principle, we do so with some trepidation. For the argument to hold, it must be shown that the welfare loss incurred by using the tariff structure to internally subsidize water for the poor is greater than that incurred by appropriating funds from elsewhere in the economy to fund government transfers. This reservation may be especially relevant in developing countries, where the tax base is small, government capacity is limited, and corruption is significant.

An interesting extension to the principle of IBTs involves the concept of two-tariff design. The price that end users face for water can be divided into two components: volumetric usage charges and fixed-rate charges, commonly referred to as connection fees. The two-tariff system provides an economic incentive for end users to conserve water through its volumetric component while providing an element of revenue security for the supplier through its fixed-rate component. This structure has evolved into the structure of choice in most Organization for Economic Cooperation and Development (OECD) countries.[31,34]

A two-tariff system can be carried out in the context of an IBT scheme so that a fixed cost is paid for "connection," but low levels of water can still be provided at a rate that is subsidized by larger consumers. In many areas, connection fees are sufficiently high that they exclude the poorest consumers. Since the cross-subsidization inherent in the IBT structure is beneficial only to those with connections, an argument can be made for higher subsidies on connection costs combined with lower subsidies on volumetric consumption. This approach incorporates the poorest strata of society while continuing to expand the customer base, which decreases average costs for all.

This is especially important in situations where land tenure rights are insecure. In such cases, the incentives and capacity for household investment in connection to the water utility are extremely low. It may be the case that these areas realize their greatest economic surplus by providing shared public taps that face a fixed rate rather than an IBT. There is already some evidence to support such a conclusion. Poor neighborhoods that face IBT prices often share a single connection between multiple households. These meters read high volumes of consumption, so the absolute poorest face higher prices due to the structure of the IBT.

Raising the price of water to its full cost would enhance economic efficiency and sustainability. Furthermore, Rogers, de Silva, and Bhatia[31] argue that this can be done without sacrificing equity and fairness. Situations exist in which Pareto improvements (improvements that leave at least some agents better off without making others any worse off) are possible even in cases where relative equity decreases.[29]

Conclusions

Water is not traded in competitive markets. High infrastructure costs raise barriers to entry, ruling out competitive conditions, whereas economies of scale, limited substitutes, human rights issues, and environmental externalities combine to complicate the design of a fair, efficient, and sufficient price for water.

The multiple roles that water plays in the economy, society, and the environment increase the complexity of structuring prices that capture scarcity and the multiple values of water. Economics alone cannot deal with this complex commodity, resource, and input. However, optimal allocations of this scarce resource among competing users and uses are basically an economic issue. The optimality criteria could be salvaged and applied where net social benefits are maximized within an idealized model capable of simulating competitive conditions. The shadow prices for water that emerge from the model solutions can be used to guide allocation of this scarce resource to optimize net social and environmental benefits. It is also feasible to construct a price-designing exercise where it is possible to capture the full cost of producing and delivering water and to preserve and balance multiple objectives. Many utilities do that now on a routine basis.

Water is valuable, but it is not thicker than blood, and an economic price can be set that embodies many of the competitive efficiency criteria in addition to other valuable societal and environmental objectives. Of particular importance is the sustainability consideration where the present generation must limit its water consumption to preserve quality and sufficient quantity for future generations. The emphasis must be particularly on quality, as increasingly societies face major issues of degradation of water quality and desertification.

References

1. Grafton, R.Q.; Libecap, G.D.; Edwards, E.C.; O'Brien, R.J.; Landry, C. *Water Scarcity and Water Markets: A Comparison of Institutions and Practices in the Murray Darling Basin of Australia and the Western US*; International Center for Economic Research: Washington, DC; 2010, Working Paper No. 28.
2. Newlin, B.D.; Jenkins, M.W.; Lund, J.R.; Howitt, R.E. Southern California water markets: Potential and limitations. *J. Water Resour. Plann. Manage.* 2002, *128* (1), 21–32.
3. Saliba, B.C.; Bush, D.B.; *Water Markets in Theory and Practice: Market Transfers, Water Values, and Public Policy*; Westview Press: Boulder, CO, 1987.
4. Fisher, F.M. Towards cooperation in water: The Middle East Water project. *Rev. Region Dev.* 2000, *12*, 143–165.
5. Hardin, G. The tragedy of the commons. *Science* 1968, *162*, 1243–1248.
6. United Nations General Assembly. *The Human Right to Water and Sanitation*, A/Res/64/292; United Nations General Assembly: New York, 2010; 1–3.
7. Gleick, P.H. The human right to water. *Water Policy* 1998, *1* (5), 487–503.
8. Lonergan, S.C.; Brooks, D.B. *Watershed: The Role of Fresh Water in the Israeli-Palestinian Conflict*; IDRC: Ottawa, ON, 1994.
9. Anand, P.B. *Scarcity, Entitlements and the Economics of Water in Developing Countries*; Edward Elgar Publishing Ltd.: Northampton, MA, 2007.
10. Young, R.A. Water, economics, and the nature of water policy issues. In *Determining the Economic Value of Water: Concepts and Methods*; Resources for the Future: Washington, DC, 2005; 3–16.
11. Conca, C. *Governing Water: Contentious Transnational Politics and Global Institution Building;* MIT Press: Cambridge, MA, 2006.
12. Mansfield, E.; Yohe, G. *Microeconomics: Theory and Applications*; W. W. Norton and Company: New York, 2003; Vol. 11, 764 pp.
13. Eckstein, O. *Water Resources Development: The Economics of Project Evaluation*; Harvard University Press: Cambridge, MA, 1958.
14. Maass, A.; Hufschmidt, M.M.; Dorfman, R.; Thomas, H.A.; Marglin, S.A. *Design of Water Resource Systems*; Harvard University Press: Cambridge, MA, 1962.
15. Hirshleifer, J.; DeHaven, J.C.; Milliman, J.W. *Water Supply: Economics, Technology, and Policy*; University of Chicago Press: Chicago, IL, 1960.
16. Fisher, F.M.; Huber-Lee, A. Economics, water management, and conflict resolution in the Middle East and beyond. *Environment* 2006, *48* (3), 26–41.
17. Amir, I.; Fisher, F.M. Analyzing agricultural demand for water with an optimizing model. *Agric. Syst.* 1999, *61*, 45–56.

18. Fisher, F.M.; Huber-Lee, A.T. WAS-guided cooperation in water management: Coalitions and gains. In *Game Theory and Policymaking in Natural Resources and the Environment*; Dinar, A., Albiac, J., Sánchez-Soriano, J., Eds.; Routledge: New York, 2008.
19. Dorfman, R.; Samuelson, P.; Solow, R. *Linear Programming and Economic Analysis*, The Rand Series; McGraw-Hill: Torontos, 1958.
20. Boland, J. Pricing urban water: Principles and compromises. *Water Resour. Update: Univ. Counc. Water Resour.* 1993, *92* (Summer), 7–10.
21. Rogers, P.; Bhatia, R.; Huber, A. *Water as a Social and Economic Good: How to Put the Principle into Practice*; Global Water Partnership/Swedish International Development Agency: Stockholm, 1998.
22. National Research Council. The meaning of value and use of economic valuation in the environmental policy decision-making process. In *Valuing Ecosystem Services: Toward Better Environmental Decision-Making*; The National Academies Press: Washington, DC, 2005.
23. Perks, A.R.; Kealey, T. International price of water. In *Environmental Economics and Investment Assessment*; Aravossis, K., Brebbia, C.A., Kakaras, E., Kungolos, A.G., Eds.; WIT Press: Boston, MA, 2006.
24. Krautkraemer, J. *Economics of Natural Resource Scarcity: The State of the Debate*; Resources for the Future: Washington, DC, 2005; Discussion Paper 05-14.
25. Turner, R.K.; Pearce, D.; Bateman, I. *Environmental Economics*; The Johns Hopkins University Press: Baltimore, MD, 1993.
26. Solow, R. An almost practical step toward sustainability. In *Assigning Economic Value to Natural Resources, Proceedings of the Workshop on Valuing Natural Capital for Sustainable Development*, Woods Hole, MA, July 1–3, 1993; National Academy Press: Washington, DC, 1994; 19–30.
27. World Health Organization and UNICEF. *Progress on Sanitation and Drinking Water: 2010 Update*; WHO Press: France, 2010.
28. Boland, J.J.; Whittington, D. The political economy of increasing block tariffs in developing countries. In *The Political Economy of Water Pricing Reforms*; Dinar, A., Ed.; Oxford University Press: Oxford, 2000; 215–235.
29. Whittington, D. Municipal water pricing and tariff design: A reform agenda for South Asia. *Water Policy* 2003, *5* (1), 61–76.
30. Yepes, G. *Do Cross-Subsidies Help the Poor to Benefit from Water and Wastewater Services? Lessons from Guayaquil*; UNDP-World Bank Water and Sanitation Program: Washington, DC, 1999; 1–9.
31. Rogers, P.; de Silva, R.; Bhatia, R. Water is an economic good: How to use prices to promote equity, efficiency, and sustainability. *Water Policy* 2002, *4* (1), 1–17.
32. Whittington, D.; Pattanayak, S.K.; Yang, J.C.; Kumar, K.C. Do households want improved piped water services? Evidence from Nepal. *Water Policy* 2002, *4* (6), 531–556.
33. Zerah, M.H. *Water: Unreliable Supply in Delhi*; Manohar, Centre de Sciences Humaines: New Delhi, 2000.
34. Griffin, R. Water pricing. In *Water Resource Economics: The Analysis of Scarcity, Policies, and Projects*; MIT Press: Cambridge, MA, 2006.
35. Olmstead, S.M.; Stavins, R.N. *Managing Water Demand: Price vs. Non-Price Conservation Programs, 39*; Pioneer Institute: Cambridge, MA, 2007, 1–40.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

28

Wetlands: Methane Emission

Anna Ekberg
and Tørben Røjle
Christensen

Introduction

Methane (CH_4) is a radiatively active trace gas that plays an important role for atmospheric chemistry and the energy balance of the Earth. At present, its contribution to the greenhouse effect is about 22%, which can be compared to carbon dioxide (CO_2) that contributes approximately 65% to the climate forcing of all long-lived greenhouse gases (excluding water vapor).[1] In addition to the direct warming effects, where infrared radiation is absorbed and returned to the Earth's surface, chemical and photochemical reactions with CH_4 in the troposphere and stratosphere indirectly cause greenhouse warming. The preindustrial atmospheric concentration of CH_4 was about 0.75ppmv (parts per million by volume) and since then it has more than doubled to approximately 1.73ppmv.[1] Anthropogenic sources of CH_4 include cattle (about 15% of the annual CH_4 release), rice paddies (20%), coal mining and oil production (14%), biomass burning (10%), natural gas leaks, and landfills and sewage disposal, while the major natural sources include wetlands (20%–25%) and termites (5%).[2] Because of the substantial contribution from natural wetlands, knowledge about the dynamics leading to CH_4 formation and emission is of great importance to understand how these ecosystems interact with the climate and how they would respond to climate change.

Methane Production and Consumption

Methane is produced by strictly anaerobic archaebacteria that are limited to the use of only a few simple substrates for biosynthesis and energy production.[3] If available carbon compounds are not directly supplied, they depend on other groups of micro-organisms for the initial breakdown of more complex organic structures into simpler molecules. The reduction of CO_2 with hydrogen is a common pathway to CH_4 formation, but acetate and formate are also important precursors of CH_4 in natural environments. Growth of methanogenic bacteria further requires a very low redox potential (E_h below −400mV) and the ideal environmental conditions are often met in permanently waterlogged wetlands. However, CH_4 efflux from wetlands to the atmosphere depends not only on the rate of production (methanogenesis), but also on the extent of CH_4 consumption (methanotrophy) that may occur in oxic surface layers and in the close vicinity of plant roots.[4–6] In this process, CH_4 is oxidized to CO_2, and the rate of CH_4 emission that can be measured at the surface is hence the net result of two counteracting processes—methanogenesis and methanotrophy.

Environmental Controls on Methane Emissions

Water Table Position

Water table position in relation to the surface is considered to be the most important factor controlling CH_4 flux, because it indicates the boundary between anaerobic CH_4 production and aerobic CH_4 consumption (Figure 1). Negative correlations between water table depth and rates of CH_4 emission have frequently been reported.[7–9] It has also been found that the lowering of the water table may result in increased CH_4 emissions when episodic releases of CH_4 trapped in pore water occur and the diffusivity

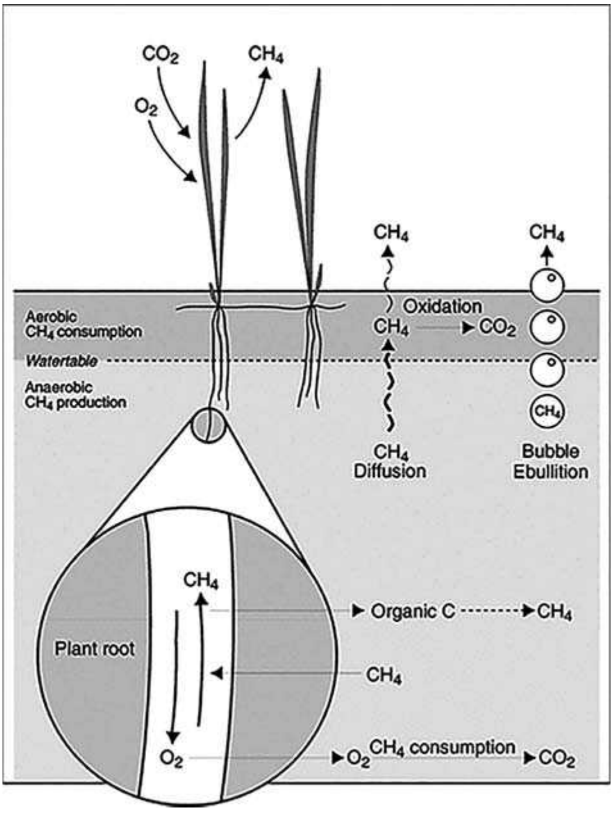


FIGURE 1 The depth of the water table is important in determining net CH_4 emissions from natural wetlands because it indicates the boundary between anaerobic CH_4 production and aerobic CH_4 consumption. Methane may be transported from the soil to the atmosphere via diffusion through the soil profile, by bubble ebullition or by plant-mediated transport. In the first case, CH_4 is subjected to methanotrophic oxidation as it passes through the oxic surface layer and this acts to decrease net CH_4 emissions. Bubbles are quite stable, meaning that little oxidation takes place, and the same is also true when CH_4 is transported through vascular plants. Vascular plants are sources of methanogenic substrate because they release labile carbon compounds into the soil via root exudation and root turnover. Leakage of O_2 from roots may lead to inhibition of methanogenesis and methanotrophic CH_4 consumption in the otherwise anoxic soil.

of gases in air-filled pore spaces increases.[10,11] Field experiments in peat-forming wetlands reveal a high spatial variability of CH_4 emissions caused by distinct microtopographical features on the wetland surface.[12] Interactions between vascular plants and mosses create differences in elevation, and therefore distance to the water table within a radius of less than a meter.[13] Scaling up of spot CH_4 emission measurements to larger areas without careful consideration of experimental plot location may therefore be misleading. Laboratory incubations have shown that there is a potential for CH_4 production both above and below the water table,[14] but the actual production in the field peaks at approximately 5–15 cm below the water table.[8] For methanotrophic CH_4 consumption to occur, a simultaneous supply of both CH_4 and oxygen is necessary. The abundance and activity of methanotrophic bacteria are therefore highest at the interface between anoxic and oxic conditions, i.e., at the depth of the water table.[5,14] Watson et al.[5] used a model to show a high capacity for CH_4 oxidation in peat, where methanotrophy accounted for 85% of the oxygen uptake potential when both oxygen and CH_4 were present in excess. In the absence of CH_4, maximum oxygen uptake potential was near the surface.

Soil Temperature

Correlations between soil temperature and CH_4 production and emission have been found at scales ranging from laboratory incubations and monolith experiments[8,11,15] to field investigations at the plot[16] and landscape scale.[9] There is a direct effect of temperature on the metabolic rate of CH_4-producing micro-organisms, but the sensitivity to temperature change is different for methanogens and methanotrophs.[11,17] In a study carried out with peat slurries under laboratory conditions, Dunfield et al.[17] found that CH_4 production in peat samples from temperate and subarctic areas was more sensitive to changes in temperature as compared to CH_4 consumption in the same samples. Increasing or decreasing temperatures also interact with other parameters with potential to control CH_4 production and emission to the atmosphere. One such interactive effect would be the relationship between temperature and water table position, where increasing temperatures lead to enhanced evaporation and plant transpiration, with lowering of the water table as a consequence.[18] It has also been suggested that the temperature dependence of CH_4 production is constrained by substrate limitation and that temperature has an effect on methanogenesis mainly via its influence on substrate availability.[19] In peat samples from a Swedish acid mire, CH_4 production was stimulated by increased temperatures only when substrate (glucose) would be the relationship between temperature and increased.[15] It has further been found that decomposition of labile material is more strongly controlled by temperature than decomposition of more recalcitrant components.[18] From the same study, it was also concluded that substrate availability lagged behind rapid temperature changes and that the thermal/hydrological history of the soil was important to determine the rate of CH_4 production.

Methane Transport: Diffusion, Bubble Ebullition, and Transport through Plants

The solubility of CH_4 in water is low, and it escapes through waterlogged soil to the atmosphere by diffusion, bubble ebullition, or by transport through vascular plants (Figure 1). The rate of CH_4 emission is largely controlled by the mode of transport because the different transportation pathways are associated with more or less extensive CH_4 consumption in the soil profile. Methane diffusing through oxic environments is subjected to methanotrophic oxidation to CO_2, and a water table depth of only a few centimeters may cause considerable reductions of CH_4 emissions to the atmosphere. Bubbles form when the rate of CH_4 production causes the sum of the partial pressures of dissolved gases to exceed the value of the hydrostatic pressure in the soil. However, the concentration of CH_4 in bubbles will be at equlibrium with dissolved pore water CH_4. At low rates of methanogenesis and with no vascular plants present, the main release of CH_4 to the atmosphere is likely to be by diffusion, but bubbles begin to form when the production rate exceeds the capacity for diffusive loss.[20] Once formed, bubbles are quite stable, meaning that little methanotrophic CH_4 oxidation occurs even as they pass through oxic environments. Studies conducted in rice paddies have shown that bubble formation tends to decrease as the plants mature because the progressively better developed rooting system acts as gas conduits and efficiently transports gases out of the soil.[21] However, a close relationship between vascular plant production and the availability of methanogenic substrate (discussed in the next section) may also speed up the rate of methanogenesis to such an extent that bubble formation is promoted. In order to supply oxygen for respiration to submerged structures, certain vascular plant species develop lacunae in stems, roots, and rhizomes.[22] Depending on the species, the ventilation of growing roots is carried out either by pressurized bulk flow or by simple diffusion that follows concentration gradients between the atmosphere and the soil. Leakage of oxygen may give rise to CH_4 oxidation and inhibition of methanogenesis in the close vicinity of the roots.[5,6] However, net CH_4 emission is generally enhanced by the presence of vascular plants with a deep rooting system because of a "chimney effect," where CH_4 transported from the soil to the atmosphere within plant lacunae is withdrawn from consumption in oxic conditions.[12,23,24]

Vascular Plants and Substrate for Methane Production

Vascular plants are sources of substrates for CH_4 production because they easily release degradable carbon compounds into the soil via root exudation and root turnover (Figure 1).[20,24,25] In peat-forming wetlands, the organic matter becomes increasingly recalcitrant with depth[26] and several studies have pointed out that plant-derived inputs of labile carbon compounds could significantly contribute to increased CH_4 production.[20,23,27] This "loading" of substrates into the soil may further be correlated to the photosynthetic rate of the vascular plants in the ecosystem,[24,28] because the amount of carbon allocated to belowground plant structures and eventually released to the soil is likely proportional to the CO_2 fixation rate.[24] The overall primary productivity in the ecosystem may therefore be of great importance for the CH_4 emission rates from a wetland area.

References

1. Lelieveld, J.; Crutzen, P.J.; Dentener, F.J. Changing concentration, lifetime and climate forcing of atmospheric methane. Tellus **1998**, *50B,* 128–150.
2. Chappelaz, J.A.; Fung, I.Y.; Thompson, A.M. The atmospheric CH_4 increase since the last glacial maximum. 1. Source estimates. Tellus **1993**, *45B,* 228–241.
3. Oremland, R.S. Biogeochemistry of methanogenic bacteria. In *Biology of Anaerobic Microorganisms*; Zehnder, A.J.B., Ed.; John Wiley: New York, 1988; 641–703.
4. King, G.M. Ecological aspects of methane oxidation, a key determinant of global methane dynamics. Adv. Microb. Ecol. **1998**, *12,* 431–468.
5. Watson, A.; Stephen, K.D.; Nedwell, D.B.; Arah, J.R.M. Oxidation of methane in peat: kinetics of CH_4 And O_2 removal and the role of plant roots. Soil. Biol. Biochem. **1997**, *29* (8), 1257–1267.
6. Popp, T.J.; Chanton, J.P.; Whiting, G.J.; Grant, N. Evaluation of methane oxidation in the rhizosphere of a Carex dominated fen in north central Alberta, Canada. Biogeochemistry **2000**, *51,* 259–281.
7. Christensen, T.R.; Friborg, T.; Sommerkorn, M.; Kaplan, J.; Illeris, L.; Soegaard, H.; Nordstroem, C.; Jonasson, S. Trace gas exchange in a high-arctic valley. 1. Variations in CO_2 and CH_4 flux between tundra vegetation. Glob. Bio- geochem. Cycles **2000**, *14* (3), 701–713.
8. Daulat, W.E.; Clymo, R.S. Effects of temperature and water table on the efflux of methane from peatland surface cores. Atmos. Environ. **1998**, *32* (19), 3207–3218.
9. Hargreaves, K.J.; Fowler, D. Quantifying the effects of water table and soil temperature on the emission of methane from peat wetland at the field scale. Atmos. Environ. **1998**, *32* (19), 3275–3282.

10. Kettunen, A.; Kaitala, V.; Alm, J.; Silvola, J.; Nykänen, H.; Martikainen, P.J. Cross-correlation analysis of the dynamics of methane emissions from a boreal peatland. Glob. Bio- geochem. Cycles **1996**, *10* (3), 457–471.
11. Moore, T.R.; Dalva, M. The influence of temperature and water table position on carbon dioxide and methane emissions from laboratory columns of peatland soils. J. Soil Sci. **1993**, *44*, 651–664.
12. Frenzel, P.; Karofeld, E. CH4 emission from a hollowridge complex in a raised bog: the role of CH_4 production and oxidation. Biogeochemistry **2000**, *51,* 91–112.
13. Waddington, J.M.; Roulet, N.T. Carbon balance of a boreal patterned peatland. Glob. Change Biol. **2000**, *6*, 87–97.
14. Moore, T.R.; Dalva, M. Methane and carbon dioxide exchange potentials of peat soils in aerobic and anaerobic laboratory incubations. Soil. Biol. Biochem. **1997**, *29* (8), 1157–1164.
15. Bergman, I.; Svensson, B.H.; Nilsson, M. Regulation of methane production in a Swedish acid mire by pH, temperature and substrate. Soil Biol. Biochem. **1998**, *30* (6), 729–741.
16. Verville, J.H.; Hobbie, S.E.; Chapin, F.S.; Hooper, D.U. Response of tundra CH_4 and CO_2 flux to manipulation of temperature and vegetation. Biogeochemistry **1998**, *41,* 215–235.
17. Dunfield, P.; Knowles, R.; Dumont, R.; Moore, T.R. Methane production and consumption in temperate and subarctic peat soils: response to temperature and pH. Soil Biol. Biochem. **1993**, *25* (3), 321–326.
18. Updegraff, K.; Bridgham, S.D.; Pastor, J.; Weishampel, P. Hysteresis in the temperature response of carbon dioxide and methane production in peat soils. Biogeochemistry **1998**, *43,* 253–272.
19. Valentine, D.W.; Holland, E.A.; Schimel, D.S. Ecosystem and physiological controls over methane production in northern wetlands. J. Geophys. Res. **1994**, *99* (D1), 1563–1571.
20. Chanton, J.P.; Whiting, G.J. Trace gas exchange in freshwater and coastal marine environments: ebullition and transport by plants. In *Biogenic Trace Gases:Measuring Emissions from Soil and Water;* Matson, P.A., Harriss, R.C., Eds.; Blackwell Science: Oxford, 1995; 98–125.
21. Holzapfel-Pschorn, A.; Conrad, R.; Seiler, W. Effects of vegetation on the emission of methane from submerged paddy soils. Plant Soil **1986**, *92*, 223–233.
22. Armstrong, W.; Justin, S.H.F.W.; Beckett, P.M.; Lythe, S. Root adaptation to soil waterlogging. Aquat. Bot. **1991**, *39,* 57–73.
23. Greenup, A.L.; Bradford, M.A.; McNamara, N.P.; Ineson, P.; Lee, J.A. The role of *Eriophorum Vaginatum* in CH_4 flux from an ombrotrophic peatland. Plant Soil **2000**, *227,* 265–272.
24. Joabsson, A.; Christensen, T.R.; Wallén, B. Vascular plant controls on methane emissions from northern peatforming wetlands. Trends Ecol. Evol. **1999**, *14,* 385–388.
25. Jones, D.L. Organic acids in the rhizosphere—a critical review. Plant Soil **1998**, *205,* 25–44.
26. Christensen, T.R.; Jonasson, S.; Callaghan, T.V.; Havström, M. On the potential CO_2 releases from tundra soils in a changing climate. Appl. Soil Ecol. **1999**, *11*, 127–134.
27. Joabsson, A.; Christensen, T.R. Methane emissions from wetlands and their relationship with vascular plants: an Arctic example. Glob. Change Biol. **2001**, *7* (8), 919–932.
28. Whiting, G.J.; Chanton, J.P. Plant-dependent CH_4 emission in a subarctic Canadian fen. Glob. Biogeochem. Cycles **1992**, *6* (3), 225–231.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

III

CSS: Case Studies of Environmental Management

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

29

Alexandria Lake Maryut: Integrated Environmental Management

Lindsay Beevers

Introduction

Lake Maryut is located on the Mediterranean coast of Egypt, in the delta of the River Nile and defines the southern boundary of the city of Alexandria.[1] The lake extends for about 80 km along the northwest coast of Alexandria and 30 km south and has a wetted surface area of 65 km^2 (Figure 1). It is a shallow water lake[2] with a water depth of approximately 1.5 m across the different basins and, unlike any of the other Nile deltaic lakes, is not directly connected to the Mediterranean Sea. The water level of the lake is kept below mean sea level, and freshwater is supplied, through irrigation canals, from the Rosetta branch of the Nile.[3] Throughout literature, the lake is referred to variously as Mariout, Mariut, or Maryut Lake. For this entry, the lake will be identified as Lake Maryut.

The lake has been in existence for more than 6000 years, as part of the Nile deltaic formation. In an extensive study of the evolution of the lake,[4] several drivers for the current form of the lake were established, namely, eustatic sea level, climate oscillations, compaction, sediment transport, and more recent anthropogenic influences such as land reclamation, irrigation, and agricultural practices. It is a mixture of these anthropogenic influences that has shaped the development of the lake in more recent years, along with its connection to the city of Alexandria.

During the Greco-Roman period, Alexandria was an active port, serving as a key navigational route from the Mediterranean, through Lake Maryut, up the Nile to Cairo. Canals linked the lake with the sea to the north and the Nile to the east. However, the siltation of the canopic mouth of the Nile in the 12th century[5,6] severed the freshwater influx to the lake and with it the navigational links. During this period, the city of Alexandria went into a phase of decline, and the lake went through several phases, which included coastal connection and influx and coastal disconnection and drying out.

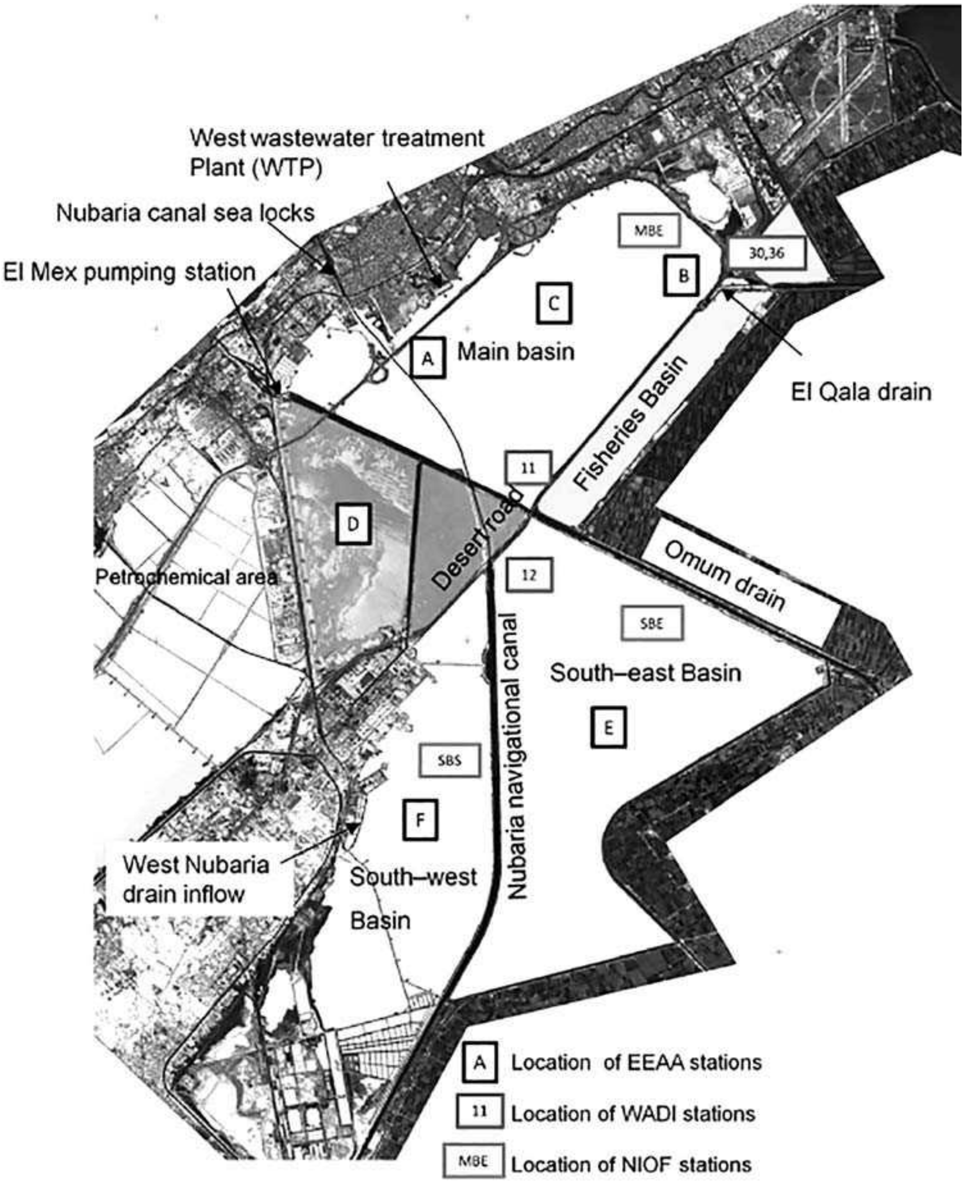


FIGURE 1 Lake Maryut: location map.

In 1892, the irrigation system of the Beheirah district was established, which restored a flow of freshwater to the lake. The lake remained disconnected from the coast, with excess water pumped to the bay through El Mex station (Figure 1), and navigation permitted through sea locks.[7] This is similar to the form the lake takes today, where the principal drainage inflows arise from agricultural irrigation channels fed by the main River Nile.

Over the years, Lake Maryut has provided Alexandria and Egypt with a source of fish, and currently, part of the lake is dedicated to aquaculture (Fisheries Basin, Figure 1). Over 7000 fishermen have rights to fish in the lake, and recent pollution levels and increased vegetation growth affect their catch.[8] *Tilapia* species are the predominant catch, representing approximately 90% of the total.[9] In addition to fisheries, the lake acts as a source of water for irrigation and for raising animals and provides areas for dwelling beside its shores.

In its current form, canals divide the lake water body into several basins. The earth embankments along the canals have several breaches that allow water to flow from the canals into the basins and vice versa, creating interaction between water bodies. There are three main inflows to the Lake Maryut: the Qala drain located to the northeast part of the lake, the Omum drain located at the east of the lake, and the Nubaria navigational canal located at the south of the lake (Figure 1). In addition to this, there are three minor inflows: one from the West Nubaria drain, one from the petrochemical industrial area, and the third from the West Wastewater Treatment Plant (WWWTP) (see Figure 1). The main outflow from the Lake Maryut is El Mex pumping station, which consists of two buildings, each housing six pumps with nominal capacities of 12.5 m^3/sec.[7] Small flows are lost to the system through the navigational locks as well as through evaporation.

As a consequence of high nutrient loadings due to the agricultural and industrial activities upstream and in its surrounds, Lake Maryut has become a highly polluted and eutrophic lake. Eutrophication of lakes is a natural process that can be accelerated by human activities (Figure 2) that introduce an excess of nutrients together with other pollutants.[10] The main sources of nutrients and pollutants to Lake Maryut are human sewage, industrial waste, farm, and urban runoff.[11,12]

Eutrophic lakes often experience an excessive growth of algae and larger aquatic plants. Such growth consumes dissolved oxygen (DO), vital for the fish and other animal life.[10] This growth and subsequent proliferation of vegetation occurs in Lake Maryut due to the high nutrient loadings that has entered the lake for a number of years. Currently, 60% of the surface area of the lake is covered by vegetation (*Phragmites australis* and *Eichornia crassipes*), which reduces the DO concentrations in the lake, especially in the main basin. Significant discharge of domestic sewage (with basic primary treatment) from the Qala drain and the East Wastewater Treatment Plant (EWWTP) enter the lake. The Omum and Nubaria Canals are less polluted than the Qala drain; however, they also contribute nutrient loadings to the lake, albeit to a lesser extent. Finally, nonpoint sources such as agricultural runoff containing pesticides and fertilizers contribute to the deterioration of the environmental quality of the lake.

The lake also suffers from other anthropogenic pressures including urbanization, unplanned settlements, and land reclamation (which has reduced the surface area of the lake gradually). An analysis of the stakeholders directly responsible for managing the lake was undertaken in 2007.[13] This showed that Alex Company for Sanitary and Drainage, the Fishing Authority, and The Ministries of Industry, Water Resources, and Environment are the main authorities responsible for the lake. The coordination

FIGURE 2 Fishing activities on the lake.

and communication regarding the management of the lake between these authorities are known to be poor, with each developing management plans in isolation.

Between 2007 and 2009, a European Union (EU)-funded SMAP III project was set up to analyze the current functioning of the lake in more detail and to develop an integrated action plan for the environmental management of the lake. This entry will cover the development of this plan, specifically focusing on the scientific models set up to support the plan development and to investigate the results of different interventions.

Current State of Lake Maryut

The current functioning of the lake has been investigated through a mixture of studies commissioned to gather specific information and ongoing monitoring surveys. Hydraulic information was gathered through a United States Agency for International Development (USAID) project in 1996,[7] hydrologic information was sourced from the local airport weather station (Nouzha airport), and a longer data series was obtained from the weather station at Port Alexandria.[14] Bathymetric information of the lake came from a survey completed in 2008 by the National Institute of Oceanography and Fisheries (NIOF).[15] The ecosystem functioning and health of the basin were constructed using a number of surveys, principally the long-term monitoring strategy undertaken annually by the Egyptian Environmental Affairs Authority (EEAA),[16] a comprehensive sampling campaign by NIOF[15] in 2008 and the EU-funded Water Demand Integration (WADI) project.[8]

Using these data, a comprehensive picture of the lake can be assembled.

Hydraulic and Hydrological Functioning

Table 1 shows the monthly average meteorological data for the lake. It is clear that, generally, precipitation in the region is very low, with the lowest values recorded between April and October. During the same period, along with significant sunlight hours recorded, evapotranspiration is highest.

Table 2 summarizes the main inflows to the lake (Figure 1) over the year. These flows are known to remain reasonably steady throughout the year, though rising slightly in the months of August to February due to the connection with the main river (Nile) and agricultural requirements. The Omum drain carries the greatest flow to the basin. Direct inputs to the lake from the WWWTP, West Nubaria drain, and the petrochemical area (Figure 1) are detailed in Table 3. These flows are considered to be

TABLE 1 Average Monthly Meteorological Data

Month	Precipitation (mm/day)[14]	Evapotranspiration	Wind Speed (m^3/sec)[15]	Solar Radiation (J/m^2/day)[20]
January	1.73	1.4	3.23	10,000,000
February	0.92	2.2	3.97	12,000,000
March	0.44	1.7	3.57	17,000,000
April	0.12	4.0	4.17	22,000,000
May	0.04	4.0	3.60	24,000,000
June	0.00	5.6	3.54	26,000,000
July	0.00	6.4	3.32	25,000,000
August	0.01	7.7	3.04	23,000,000
September	0.04	9.3	2.15	20,000,000
October	0.25	5.5	2.58	16,000,000
November	1.15	3.4	2.92	12,000,000
December	1.82	1.3	3.77	10,000,000
Average		1.4	3.32	18,083,333

TABLE 2 Monthly Flow Discharges

	Discharges (m^3/day)		
Month	Qala	Omum	Nubaria
January	529,920	3,368,640	959,040
February	466,560	2,478,640	1,270,080
March	475,200	3,248,640	462,240
April	532,224	3,318,640	666,144
May	671,328	3,680,640	691,200
June	414,720	3,453,038	243,792
July	673,920	3,136,320	285,120
August	734,400	3,412,800	838,030
September	676,500	4,275,690	492,480
October	873,500	4,813,000	559,800
November	853,500	5,353,000	559,800
December	693,500	4,103,000	559,800

Source: USAID.[7]

TABLE 3 Constant Flow Discharges

Discharges (m^3/day)		
WWWTP[17]	West Nubaria Drain	Petrochemical Area[16]
410,325	259,200	47,398

constant inflows to the lake. El Mex pumping station manages the outflow of the lake. Twelve pumps with nominal capacities of 12.5 m^3/ sec[7] maintains the water level in the lake to –2.8 m below sea level [7,17] In addition, small flows are lost from the lake through the navigational locks and of course through evaporation. Water depths vary throughout the lake, with the deepest occurring in the canals due to the navigational requirements. Depths in the basins vary between 0.5 and 1.2 m,[15] and in places, this is maintained by dredging and vegetation removal.

The flow direction in the lake is generally concentrated down the canals. Interaction of the canal water with the basin water occurs through the breached bunds; however, velocities and, hence, mixing in these sections are low. The point where the greatest interaction of canal and basin flow can be observed is at the junction between the Omum and Nubaria Canals in the main basin.

Water Quality and Ecosystems

Table 4 shows the time period available for each of the data sources, and Figure 1 shows the location of the sampling points. As a whole, the data of these sources were consistent and a similar magnitude was observed between sources. Table 5 presents the seasonal values available from the EEAA[16] data. Typical effluent concentrations from treatment plants with primary treatment can be found in literature.[18,19]

TABLE 4 Water Quality Data Sources

Source	Sampling Period	Number of Measurements	Number of Stations
EEAA[16]	2004–2008	2–3 per year	6
NIOF[15]	March–May 2008	5	10
WADI[8]	March 2007	1	31

TABLE 5 Water Quality Parameters from Selected Basins

Water Quality Parameters (EEAA 2004–2008)																
Yearly Averages																
EEAA stations	Year	DO (mg/L)	DO %	BOD	COD	NH_3-N (mg/L)	NO_2-N (mg/L)	NO_3-N (mg/L)	Total N (mgN/L)	PO_4-P (mg/L)	Temperature (°C)	Conductivity (mS/cm)	PH	TDS	TS	TSS
Main basin in front of Qala drain (B)	2004	3.8	43.7	38.0	110.0			3.6	3.6		21.5		8.0	2454.0	2479.0	25.0
	2005	2.0	22.5	30.0	98.0	12.6	0.8	1.2	14.2	5.1	19.4	3.0	7.5	1828.0		
	2006	2.6	29.2	40.0	137.0	21.0		1.8	22.8	4.5	21.0	3.6	7.2	1825.0	1852.0	27.0
	2007					20.0	0.1	1.3	21.4	4.8						
	2008	1.0			133.0	27.8		1.5	29.3		24.0		7.8	1923.0	1958.0	35.0
	Ave	2.4	31.8	36.0	119.5	20.3	0.4	1.9	18.2	4.8	21.5	3.3	7.6	2007.5	2096.3	29.0
Middle of south basin (E)	2004	7.2	83.5	2.0	11.8	0.3	0.2	4.0	4.3	0.5	23.0	6.5	7.7	5134.0	5139.3	5.3
	2005	6.7	76.9	1.0	29.3	0.3	0.2	2.2	2.7	0.8	21.2	7.5	8.1	5266.3	5349.0	6.5
	2006	7.4	82.5	4.3	34.8	0.6	0.1	1.9	2.1	0.4	22.0	8.0	7.8	4289.8	4307.5	17.8
	2007					0.1	0.3	1.0	1.2	0.7						
	2008	9.2		2.0	55.0	0.4	0.0	2.9	3.3		27.5		8.1	5141.0	5153.5	12.5
	Ave	7.6	81.0	2.3	32.7	0.3	0.2	2.4	2.7	0.6	23.4	7.3	7.9	4957.8	4987.3	10.5
Middle of west basin (F)	2004	5.4	63.8	6.0	99.0	0.8	0.2	5.5	5.9	3.8	23.5	8.1	7.5	6953.3	6965.3	12.0
	2005	7.3	84.9	2.0	49.7	0.4	0.3	2.6	2.4	0.3	21.0	8.5	8.0	5085.0	5339.0	20.0
	2006	6.7	73.9	4.3	84.5	0.3	0.1	2.4	2.1	0.8	22.2	9.4	7.8	5885.8	5907.0	21.3
	2007					0.1	0.1	0.9	1.1	1.2						
	2008	7.9		5.0	81.5	0.4	0.1	1.9	2.4		27.0		8.2	7254.0	7274.0	20.0
	Ave	6.8	74.2	4.3	78.7	0.4	0.2	2.7	2.8	1.6	23.4	8.7	7.9	6294.5	6371.3	18.3
Middle of main basin (C)	2004	4.5	52.3	21.0	97.0			1.2	1.2		24.2	3.6	8.0	2415.0	2415.0	29.0
	2005	3.3	36.3	44.0	99.0	6.8	0.1	1.0	7.9	4.2	19.0	3.1	7.9	1961.0		17.0
	2006	3.4	38.3	29.7	136.7	15.6	0.1	4.9	18.8	3.0	23.7	4.6	7.8	1935.7	1965.3	29.7
	2007					9.0	0.1	3.9	13.0	4.7						
	2008	7.5			67.0	16.5		0.2	16.7		25.0		8.0	1941.0	1973.0	32.0
	Ave	4.7	42.3	31.6	99.9	12.0	0.1	2.2	11.5	4.0	23.0	3.8	7.9	2063.2	2117.8	26.9

The data clearly show that the most polluted basin is the main basin, and the source of that pollution comes from the Qala drain, which carries with it effluent from the EWWTP. The highest. Biological oxygen demand (BOD) values and the lowest DO% values are found at station B (Figure 1) at the entrance of the Qala drain. Similarly, the highest nitrogen loading enters the main basin through the Qala drain, along with the highest phosphorus loading. Station C shows generally lower values of the N and P as the water from the drain enters the lake and becomes mixed.

In general, the proportion of total N arising from ammonia was higher than that from nitrates, particularly in the main basin. It is thought that this is due in part to the source of pollution arising from the Qala drain and the low DO values recorded. In the West Basin, the proportion of ammonia is lower than that in the main basin, which is corroborated by the findings of El Rayis.[21] Since the levels of DO are also higher in this basin, it is believed that the nitrification process is more evident here.

This poor water quality has led to a drastic reduction in biodiversity.[8] The eutrophic nature of the lake results in a significant dominance of *P. australis,* while at the outfalls from the WWTPs, significant growths of the lead tolerant species *E. crassipes* is found. Periodic cutting of the *Phragmites* is practiced to increase the nitrogen removal from the lake. This intervention is necessary to prevent the lake area being dominated by vegetation, which in turn reduces the area available for fish habitat.

It is clear that in its present state, Lake Maryut is severely polluted with reduced biodiversity. A continual deterioration in the status of the basin will lead to a significant reduction in economic and industrial development potential.

Social Considerations and Governance

There are no specific studies that survey those living by the lake. However, the Integrated Action Plan research[22] identifies three main groups, namely, fishermen, poorer communities, and a scattered population, totaling approximately 30,000. It is estimated that 62% of the population is aged between 15 and 55, which is the average employment age in Egypt. There are high unemployment rates (up to 15%–20%) within this group, leading to low annual incomes. High illiteracy, poor health services, high mortality rates, high crime levels, and a tendency to marry young are also characteristics of the demographic.

A significant proportion of the population are living in settlements and accommodation that are temporary or informal. In these scattered settlements, some common issues arise: 1) lack of, or poor infrastructure coverage, particularly water supply and sanitation/wastewater networks and paved roads; 2) inadequate local services, especially health care, education, and youth facilities; 3) poor housing conditions; 4) lack of secure land tenure in the squatter settlements; and 5) high unemployment.

An analysis of the stakeholders directly responsible for managing the lake was undertaken in 2007.[13] This showed that the Alex Company for Sanitary and Drainage, the Fishing Authority, and The Ministries of Industry, Water Resources, and Environment are the main authorities responsible for the lake. The coordination and communication regarding the management of the lake between these authorities are known to be poor. Each authority makes strategic plans for the lake in isolation, with little discussion; hence, an integrated vision of the lake did not exist.

Integrated Environmental Management

Integrated Plan Concept

The integrated action plan for Lake Maryut draws on the principles of Integrated Coastal Zone Management (ICZM) to develop integrated actions for the sustainable development of Lake Maryut. Central to the ICZM approach is participation from stakeholders in the development of a plan that integrates strategic actions to address the environmental, economic, administrative, social, and urban issues of the lake.[22] The outcomes of this process defined four specific objectives of the action plan:

- Improve the ecological and chemical status of Lake Maryut.
- Enhance levels of economic, urban, and social development in a sustainable manner.
- Increase the potential for industrial activities around the lake and their environmental management.
- Adapt the governance system of the lake to execute the plan.

Strategic actions were proposed for each of the objectives, following analysis of the current state of the lake, and where feasible, these actions were tested. This entry will focus on the process completed for the first objective in detail and will present a short summary of the actions identified for the other three objectives.

Scenario Identification

The geographic scope of the plan was taken as the general boundary of the lake (Figure 1) and the natural extension to these boundaries according to the interaction of the lake. To the north, the plan includes the navigational locks and the El Mex pumping stations (necessary connections to the sea). To the east, the limit is bounded by the main road to Alexandria from Cairo, plus the position of the EWWTP and the Qala drain. To the west, the limit was the industrial area and a 500 m perimeter to the lake, and finally, the 500 m perimeter was extended to the south of the lake as the limit to the plan.

Through a participatory process, which included four training events and six participation workshops, a series of scenarios and potential actions for the improvement of the water ecosystem were defined by key stakeholders. These can be summarized as follows:

- Redevelopment of the waterfront and potential development of islands in the lake for urban regeneration projects.
- Improve lake mixing.
- Removal or improvement of the WWTP discharges into the lake.
- Vegetation management and sediment management.
- Industrial effluent improvement.
- Installation of wetland areas for nutrient removal.

Model Development

To assess the impact of certain actions, tools were developed to predict the hydrodynamic functioning and ecosystem functioning of the lake. Given the level of information available to create these tools and the complex nature of the processes in the lake, a mixture of approaches was proposed. To investigate the hydrodynamics and mixing of Lake Maryut, a two-dimensional (2D) hydrodynamic model was developed. To investigate the ecological improvements as a result of intervention, a point model of ecosystem functioning of different basins was set up.[23]

Hydraulic Model

Using the data set out in Tables 1–3, a 2D hydraulic model of the lake was built using MIKE21 software.[24] The computational mesh is shown in Figure 3, which defines the principal flow routes through the canal. A roughness map was developed to represent vegetation growth in the basins.[25] The model was calibrated for water levels but not velocities due to lack of data. A comprehensive sensitivity study was undertaken, which showed that the main control on the lake's hydraulic system is the pumping station at El Mex.[26] Figure 4 shows the general circulation patterns predicted by the model. The interactions are mainly driven by the drain inflows and clear areas of the lake where no vegetation is found.

To investigate the impact of particular actions on the hydraulics, a number of simulations were undertaken (Figure 5). These included the land reclamation from the lake and addition of islands to

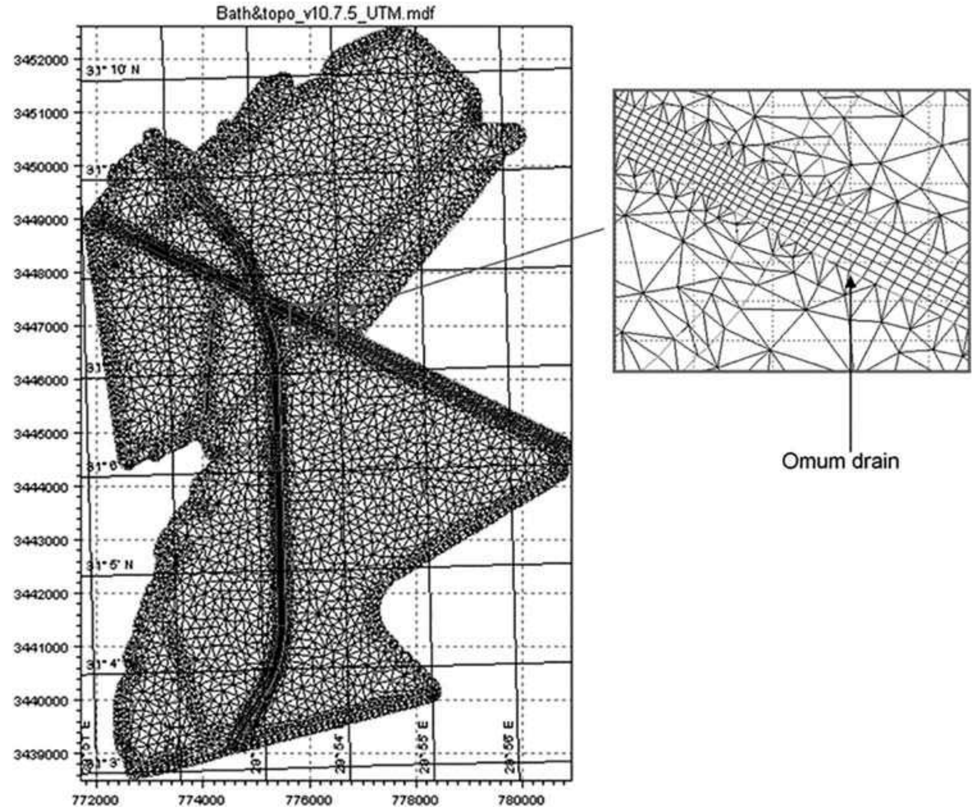


FIGURE 3 Computational mesh: Hydraulic model.
Source: Duran and Beevers.[26]

promote urban regeneration of the lakeshore, as well as the design of different entry configurations for the Qala drain to promote mixing. Figure 6 shows the magnitude of velocities predicted in the basin following these interventions. The change in velocities is highest in the main basin where most interventions are proposed; however, the magnitude of change is only in the order of a 9% increase.[26] While areas in the main basin become better mixed, areas of stagnation occur, most notably around the islands. From these results, it becomes clear that increasing mixing significantly would be difficult unless an increase to the inflow to the lake could be facilitated.

Ecological Model

Using the data set out in Tables 1–5, a dynamic 0-dimension ecological model was built for the different basins of Lake Maryut. The software PCLAKE describes the dominant ecological interaction in a shallow lake ecosystem.[23] This model was chosen for its ability to model the complex processes and allow a number of different proposed actions to be modeled, including vegetation management on a seasonal basis, dredging, wetland installation, and improvements to discharges. It should be noted that the model assumes mixing across a basin, which, for Lake Maryut, is a simplification; however, the model is detailed enough to give an indication of the potential change caused by an action.

The model was built using the seasonal data available from the EEAA monitoring for the inflows, and this was complemented by the WADI and NIOF data. Calibration was possible, comparing monitoring data available in the lake itself.

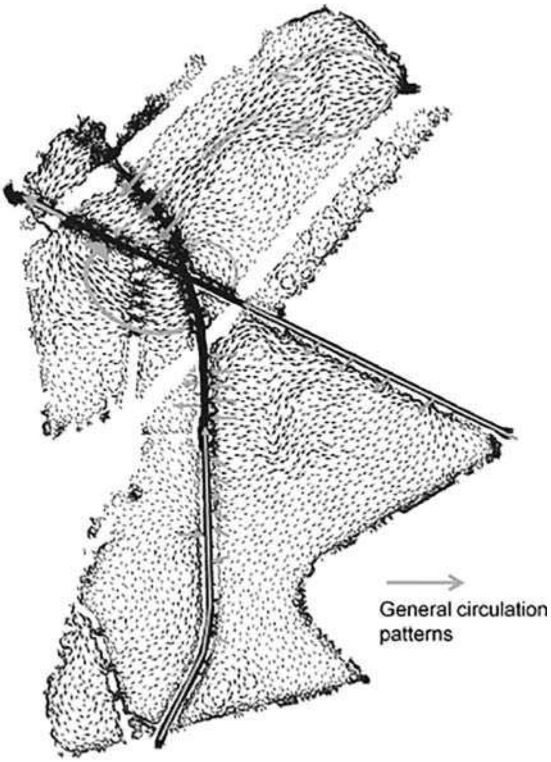


FIGURE 4 General circulation patterns: Hydraulic model.
Source: Duran and Beevers.[26]

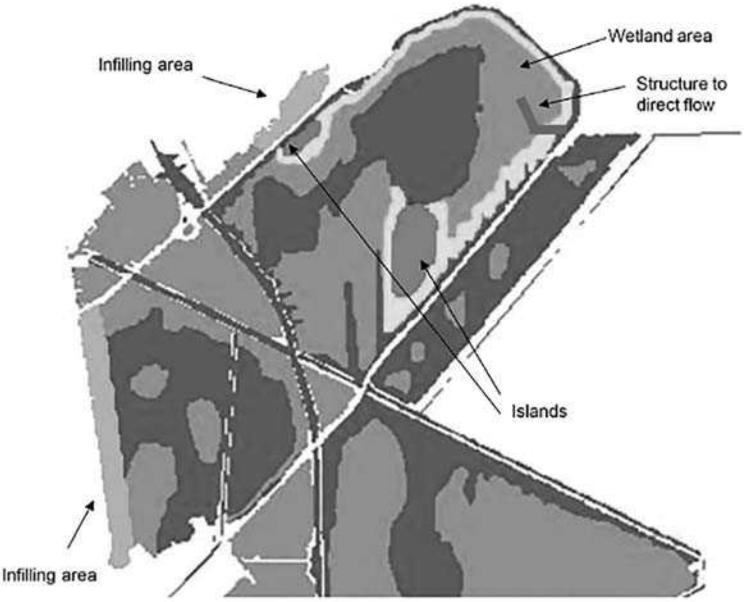


FIGURE 5 Scenario modifications: Hydraulic model.
Source: Duran and Beevers.[26]

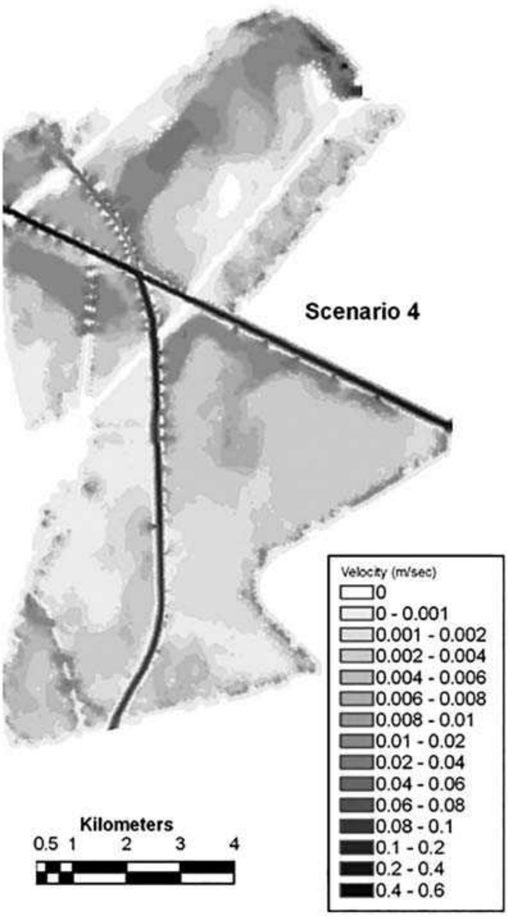


FIGURE 6 Velocity predictions: Hydraulic model.
Source: Duran and Beevers.[26]

A significant number of actions were tested in the ecological model. Available literature was used to represent the actions in the lake:

- Secondary treatment of effluent from the WWTPs.[27]
- Addition of aluminum to the WWTP process.[18]
- Installation and removal efficiencies of constructed wetlands (for the Qala drain).[28–31]
- Vegetation and sediment management.
- Table 6 shows the predicted improvement to different parameters if the following actions are taken:
- Removal of the WWWTP from the lake (outfall to the sea) after upgrade to secondary treatment.
- Upgrade of EWWTP to secondary treatment plus additional treatment using aluminum.
- Upgrade of industrial effluent through improved environmental management.
- Dredge sediments in the Qala drain and installation of a constructed wetland.
- Aquatic vegetation management.

The oxygen consumption in Lake Maryut is mainly attributed to the respiration of aquatic organisms including plankton and aerobic bacteria and the oxidation of organic matter.[32] High values of organic carbon have been reported in sediments close to the treatment plants.[21] The oxidation of these

TABLE 6 Ecological Model Results

			Improvement with Stated Actions	
Parameter	Unit	Baseline Value (average)	Increase %	Decrease %
DO	mg/L	1.3	303	
NH_4-N	mg N/L	13.4		47
NO_3-N	mg N/L	0.8	120	
Organic N	mg N/L	6.4		83
N total	mg N/L	20.5		52
P total	mg P/L	4.9		72
Chlorophyll-a	mg/m_3	93.4		85
Phyto-biomass	mg DW/L	5.9		85
Zoo biomass	g DW/m_2	0.53	39	
Detritus	mg/L	83.5		82
Secchi depth	m	0.1	377	

Source: Alvarez-Mieles and Beevers.[36]

sediments, rich in organic matter, is one of the factors that increases the DO consumption and therefore decreases the DO concentrations in the water column. Improving the treatment processes for the treatment plants has a significant impact on the model results, indicating up to a 300% improvement to DO levels in the main basin.

Ammonium is generated by heterotrophic bacteria as the primary nitrogenous end product of decomposition of organic matter. NH_4-N concentrations are usually low in oxygenated waters of oligotrophic lakes since it is utilized by plants in the nitrification process. However, at low DO concentrations, nitrification ceases, the absorptive capacity of the sediments is reduced, and an increase of the release of NH4-N from the sediments occurs. As a result, the NH_4-N concentration of shallow lakes would increase.[33] Increasing the levels of DO through improved treatment of wastewater and a reduction in pollutant loads arriving to the lake as a result of constructed wetlands has the potential to reduce overall ammonia levels in the main basin by up to 47%.

Chlorophyll-a is a primary productivity indicator and a very good estimate for monitoring and assessing the eutrophication status of lakes.[34] The supply of P and N is considered to be one of the main factors that determine the magnitude of the primary production.[35] In cases where pollution is caused by domestic wastewater with a high nutrient content, the algae production in the recipient watercourse increases considerably.[34] With the improvement to N and P and the increase in the DO predicted, a decrease in chlorophyll-a of up to 85% could be achieved for the lake.

Finally, the Secchi depth is a measure of the transparency of a water column. Secchi readings between 0.1 and 0.4 m were reported in the main basin,[32] which are directly related to the high production rate of phytoplankton. Eutrophic lakes present low values of Secchi depth due to the high concentrations of algae and detritus that increase turbidity and therefore decrease transparency. With the proposed measures and reduction in detritus, an increase in the transparency of the lake is predicted to be greater than 370%.[36]

Proposed Actions

The modeling component tested a number of actions for the improvement of the water ecosystem and showed that a significant improvement could be achieved. In addition to these, actions were proposed to address the other strategic objectives of the plan.[22] The following is a summary of such objectives:

- Prepare a Lake Maryut land and water plan.
- Promote urban regeneration around the lake shoreline and create ports and markets in the main basin.

- Promote ecotourism and encourage new ecological economic activities.
- Upgrade informal settlement areas and integrate social housing into urban development projects.
- Plan and develop fisheries activities.
- Prepare online monitoring systems along with water quantity and quality maps.
- Regulate specific limits for industrial water discharge into Lake Maryut.
- Establish a Lake Maryut Authority and set up a management unit with a clear structure and purpose.
- Establish a robust monitoring system for water quality.

The plan aims to address all four strategic objectives defined at the outset. To monitor the effect of any potential implemented action as part of the plan, a monitoring program is suggested. The implementation of such a monitoring plan would allow the incorporation of adaptive management techniques should unexpected impacts become evident.

Conclusions

An integrated action plan for Lake Maryut is proposed, using a participatory process and supported by a technical modeling study that was used to predict the impact of potential actions. It is clear from the study that significant improvement to the lake can be achieved, thus encouraging economic development and urban regeneration. However, this improvement requires substantial investment. The plan suggests that to achieve this, a financial strategy that links the urban and industrial future development potential with the environmental interventions is required to improve the lake, thereby linking potential future economic gains to the required lake renovation.

Since the completion of the proposed plan in 2009, there have been follow-up activities that move towards the adoption of a sustainable development strategy for the lake. A lake management unit is established and receives technical input from Centre for Environment and Development for the Arab Region and Europe (CEDARE) for Geographical Information System (GIS) support and NARSS (National Authority for Remote Sensing and Space Sciences of Egypt) for model development. In addition, the Ministry for Housing has proposed to use the outcomes of the study for the development of a new urban territorial plan for Alexandria that includes an integrated sustainable development vision for the Lake Maryut zone.

The study has shown that, with a coordinated and integrated approach to environmental management, the current deteriorating trend of Lake Maryut could be reversed. However, to achieve this goal, actions addressing environmental, social, and governance issues must be implemented.

Acknowledgments

The author would like to acknowledge the official partners of the ALAMIM project: MEDCITIES, Barcelona Metropolitan Area, CEDARE, Alexandria Governorate, the EEAA, NARSS, City of Marseilles, Environment and Housing Ministry of the Catalan Government, UNESCO-IHE, and the EUCC—The Coastal Union. In particular, the author is grateful to Mr. Joan Parpal, the project leader, and Ms. Roxan Duran and Ms. Gabriela Alvarez-Mieles.

References

1. Saad, M. Distribution of phosphates in Lake Maryut, a heavily polluted lake in Egypt. Water, Air, Soil Pollut. **1973**, *2*, 515–522.
2. Mesnage, V.; Picot, B. The distribution of phosphate in sediments and its relation with eutrophication of a Mediterranean coastal lagoon. Hydrobiologia **1995**, *297* (1), 29–41, DOI: 10.1007/BF00033499.

3. Hughes, R.; Hughes, S. *A Directory of African Wetlands;* IUCN: Gland, Switzerland and Cambridge, UK/UNEP: Nairobi, Kenya/WCMC: Cambridge, U.K., 1992; 820 pp., ISBN 2–88032-949–3.
4. Warne, A.; Stanley, D. Late quaternary evolutions of the Northwestern Nile Delta and adjacent coast in the Alexandria region, Egypt. J. Coastal Res. **1993**, *9* (1), 26–64.
5. Stanley, J.; Warne, A.; Schnepp, G. Geoarchaeological interpretation of the Canopic, largest of the relict Nile Delta distributaries, Egypt. J. Coastal Res. **2004**, *20* (3), 920–930, doi: 10.2112/1551–5036(2004)20[920:GI0TCL]2.0.C0;2.
6. Stanley, D.; Warne, A. Nile Delta: Recent geological evolution and human impact. Science **1993**, *260* (5108), 628–634 DOI: 10.1126/science.260.5108.628.
7. USAID. Hydrologic Studies of Lake Mariout Technical Memorandum N° 8. 1996 USAID Project N° 263–0100.
8. Mateo, M. Lake Mariut: An Ecological Assessment. Final report of the Water Demand Integration (WADI) project: INCO-C-2005—15226, 2009. Available for download at http://www.medcities.org/ocuments section (accessed April 29, 2011).
9. Bakhoum Shnoudy, A. Comparative study on length weight relationship and condition factor of the genus *Oreochromis* in polluted and non-polluted parts of Lake Mariut Egypt. Bull. Natl. Inst. Oceanogr. Fish. **1994**, *20* (1), 201–282.
10. Khan F.A.; Ansari, A.A. Eutrophication: An ecological vision. The Botanical Review. **2005**, *71* (4), 449–482. doi: 10.1663/00068101(2005)071[0449:EAEV]2.0.C0;2
11. Wahby, S.; Kinawy, S.; El-Tabbach, T.; Abdel Moneim, M. Chemical characteristics of Lake Maryût, a polluted lake south of Alexandria, Egypt. Estuarine Coastal Mar. Sci. **1978**, *7* (1), 17–28.
12. Amr, H.; El-Tawila, M.; Ramadam, M. Assessment of pollution in fish and water of main basin, Lake Maryut. J. Egypt. Public-Health Assoc. **2005**, *80* (1–2).
13. Kafafi, A. *Stakeholder analysis report,* 2007. Final report for the ALAMIM project available for download at http://www.medcities.org/ documents section (accessed April 29, 2011).
14. World Climate. Port Alexandria, Egypt Weather History and Climate Data, http://www.worldclimate.com/cgi-bin/datapl?ref=N31E029+2100+62315W (accessed April 2009).
15. National Institute of Oceanography and Fisheries (NIOF). *Lake Maryut Data Acquisition.* Final report for the ALAMIM project available for download at http://www.medcities.org/ documents section (accessed April 29, 2011).
16. Egyptian Environmental Assessment Authority (EEAA). *Monitoring Data Lake Maryut*, 2004–2008 reports.
17. CEDARE. *Alexandria Lake Mariout Integrated Management—Stocktaking Analysis report,* 2007. ALAMIM report available for download at http://www.medcities.org/ documents section (accessed April 29, 2011).
18. El-Bestawy, E.; Hussein, H. Baghdadi, H.; El-Saka, M. Comparison between biological and chemical treatment of wastewater containing nitrogen and phosphorous. J. Industrial Microbiol. Biotechnol. **2005**, *32,* 195–203.
19. Metcalf, A.; Eddy, G.; Tchobanoglous, F.; Burton, F.; Stensel, H. *Wastewater Engineering, Treatment and Reuse*, 4th Ed.; McGraw Hill Education: New York, 2003; 1329, ISBN: 0070418780.
20. Trabea, A.; Salem, I. Empirical relationship for ultraviolet solar radiation over Egypt. Egypt J. Sol. Energy **2001**, *24* (1), 123–132.
21. El Rayis, O. Impact of man's activities on a closed fishing lake, Lake Maryout in Egypt, as a case study. Mitigation Adapt. Strategies Global Change **2005**, *10,* 145–157.
22. El-Refaie, M.; Rague, X. *ALAMIM Integrated Action Plan,* 2009. ALAMIM report available for download at http://www.medcities.org/ documents section (accessed April 29, 2011).
23. Janse, J. Model studies on the eutrophication of shallow lakes and ditches, Ph.D. Thesis, Wageningen University: Wageningen, the Netherlands, 2005.
24. Warren, I.; Back, H. MIKE21: A modelling system for estuarine, coastal waters and seas. Environ. Software **1992**, *7* (4), 229–240.

25. Wang, C.; Zhu, P.; Wang, P.; Zhang, W. Effects of aquatic vegetation on flow in the Nansi Lake and its flow velocity modeling. J. Hydrodyn. **2006**, *18* (6), 640–648.
26. Duran, R.; Beevers, L. *Alexandria Lake Mariout Integrated Management: Hydrodynamic Model Report,* 2009. ALAMIM report available for download at http://www.medcities.org/ documents section.
27. Knight, R.; Rubles, R.; Kadlec, R.; Reed, S. Wetlands for wastewater treatment performance data base. In *Constructed Wetlands for Water Quality Improvement*; Moshiri, G., Ed.; CRC Press; Boca Raton, FL, 1993; 35–49.
28. Vymazal, J. Removal of nutrients in various types of constructed wetlands. Sci. Total Environ. **2007**, *380,* 48–65.
29. Moortel van de, A.; Rousseau, D.; Tack, F.; De Pauw, N. A comparative study of surface and subsurface flow constructed wetlands for treatment of combined sewer overflows: A greenhouse experiment. Ecol. Eng. **2009**, *35*, 175–183.
30. Verhoeven, J.; Meuleman, A. Wetlands for wastewater treatment: Opportunities and limitations. Ecol. Eng. **1999**, *12,* 5–12.
31. Lu, S.; Wu, F.; Lu, Y.; Xiang, C.; Zhang, P.; Jin, C. Phosphorous removal from agricultural run off by constructed wetland. Ecol. Eng. **2009**, *35*, 402–409.
32. Anwar, A.; Samaan, A. Productivity of Lake Mariut, Egypt. Part I. Physical and chemical aspects. Int. Rev. Hydrobiol. **1969**, *54* (3), 313–355.
33. Quiros, R. The relationship between nitrate and ammonia concentrations in the pelagic zone of lakes. Limnetica **2003**, *22* (1–2), 37–50.
34. Heinonen, P.; Ziglio, G.; Beken van der, A. *Hydrological and Limnological Aspects of Lake Monitoring*; Wiley: Chichester, England, 2000, ISBN 0–471–89988-7.
35. Fathi, A.; Abdelzaher, H.; Flower, R.; Ramdani, M.; Kraiem, M. Phytoplankton communities of North African wetland lakes: The Cassarina Project. Aquat. Ecol. **2001**, *35,* 303–318.
36. Alvarez-Mieles, G.; Beevers, L. *Alexandria Lake Mariout Integrated Management: Ecological Modelling for Lake Maryut,* 2009. ALAMIM report available for download at http://www.medcities.org/ documents section (accessed May 29, 2011).

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

30

Aral Sea Disaster

Guy Fipps

Introduction

The cause is attributed to a vast expansion of irrigation in the Central Asian Republics beginning in the 1950s, which greatly reduced inflows to the Sea. The diversion of water for massive irrigation development was done deliberately by Soviet Union officials, unconcerned about the consequences of their actions.

The environmental, social, and economic damage has been immense. Winds pick up dust from the dry seabed and deposit it over a large populated area. The dust likely contains pesticide and chemical residues that are blamed for the serious rise in mortality and health problems in the region. The Sea, and the now exposed dry seabed, may also be contaminated by runoff from a former Soviet military base and a biological weapons lab. The ecosystem of the Aral Sea has collapsed, and climate changes in the Aral Sea Basin have been documented. Hundreds of agreements have been signed since 1980s on programs designed to address the "Aral Sea Problem" which, to date, have not been effective at preventing the continuing shrinking of the sea.

Aral Sea Basin

The Aral Sea is located in Central Asia and lies between Uzbekistan and Kazakhstan in a vast geological depression, the Turan lowlands, in the Kyzylkum and Karakum Deserts. In the 1950s, the sea covered 66,000 km^2, contained about 1090 km^3 of water, and had a maximum depth of about 70 m. The Aral Sea supported vast fisheries and shipping industries. At that time the sea was fed by two rivers, the Amu Darya (2540 km) and the Syr Darya (2200 km), which originate in the mountain ranges of central Asia and flow through the five republics of Uzbekistan, Kazakhstan, Kyrgyzstan, Tajikistan, and Turkmenistan.

The two rivers provide most of the fresh water used in Central Asia. In the last 50 years, about 20 dams and reservoirs and 60 major irrigation schemes have been constructed. About 82% of river diversions are for agricultural use and 14% is for municipal and industrial use (Table 1).

Water demand due to population growth and industrial expansion continues to increase (Table 2). Since 1960, the population of the Central Asian republics has increased 140% and totals over 50

TABLE 1 Average Water Supply and Demand in the Aral Sea Basin

Total Water Available	km^3	%
Amu Darya Basin	84.3	64
Syr Darya Basin	47.8	36
Total	132.1	100
Water demand		
Agriculture		
Amu Darya Basin	44.8	81.6
Syr Darya Basin	34.6	
Municipal Water		
Amu Darya	3	6.5
Syr Darya	3.3	
Industry		
Amu Darya	3	8.2
Syr Darya	5	
Livestock		
Amu Darya	0.2	0.2
Syr Darya	0	
Fishery		
Amu Darya	2.6	3.5
Syr Darya	0.8	
Total	97.3	100

TABLE 2 General Statistics of the Aral Sea Basin Countries in 1995

	Kazakhstan	Uzbekistan	Turkmenistan	Kyrgyzstan	Tajikistan
Area, km^2	2,717,300	447,400	488,100	198,500	143,100
Irrigated land, km^2	23,080	41,500	12,450	10,320	6.940
Population	17,376,615	23,089,261	4,075,316	4,769,877	6,155,474
Population growth rate, %	0.62	2.08	2.5	1.5	2.6

million. Likewise, industrial production using large amounts of water has also increased. Examples include steel production which rose 200%, cement production by 170%, and electricity generation by a factor of 12.

The total inflows to the Aral Sea began decreasing rapidly in the 1960s, and by 1990 the storage volume of the sea has decreased by 600 km^3 (Table 3). As the water level fell, salinity levels have tripled, rising from about 1000 ppm to just under 3000 ppm today. By the 1980s, as the Aral Sea problem became well known in the Soviet Union, government officials proposed ambitious projects to divert water from other rivers, including ones in South Russia and Siberia, to be transported to the Aral Sea in massive canals. However, these plans died with the breakup of the Soviet Union.

The decrease in sea level has now split the Aral Sea into two separate water bodies: the Small and Large Aral Seas (Maloe More and Bol'shiye More) each separately fed by the Syr Darya and the Amu Darya, respectively. The once vast Amu Darya delta which once covered 550,000 ha has now shrunk to less than 20,000 ha.

TABLE 3 Decline of the Aral Sea during the 1980s and Total Estimated Inflows from the Amu Darya and Syr Darya Rivers

		Aral Sea	
Year	Inflows (km^3)	Volume (km^3)	Surface Area (km^2)
1911–1960	56.0	1064	66,100
1981	6.0	618	50,500
1982	0.04	583	49,300
1983	2.3	539	47,700
1984	7.9	501	46,100
1986	0.0	424	41,100
1987	9.0		
1988	23.0		41,000
1989		300	30,000

Irrigation and Cotton

For thousands of years, Central Asian farmers diverted water from the Amu Darya and Syr Darya Rivers, transforming desert into green oases and supporting great civilizations. Historically, irrigation water use was conducted at a sustainable level. The creation of the Soviet Union and the collectivization of farmlands resulted in the end for traditional agricultural practices. Beginning as early as 1918, Soviet leaders began expanding irrigated land in Central Asia for export and hard currency. Cotton was known as "white gold." The USSR became a net exporter of cotton by the 1930s, and by the 1980s, was ranked fourth in the world in cotton production.

The policy of emphasizing cotton production was accelerated in the 1950s as Central Asia's irrigated agriculture was expanded and mechanized. In 1956, the Kara Kum Canal was opened, diverting one-third of the flow in the Amu Darya to new cultivated areas in the deserts of Uzbekistan and Turkmenistan. The year 1960 represents the critical junction when the Aral Sea began to drop. Irrigated cotton production and water diversions continued to be expanded until the break-up of the Soviet Union (Table 4).

Estimates are that upwards of 80% of the workforce is employed in agriculture. The main agricultural crops in the basin are cotton (6.4 million ha), forage (1.7 million ha), rice (0.4 million ha), and tree crops (0.4 million ha).

Some Central Asian irrigation experts estimate that only 20%–25% of the water diverted from the rivers is actually used by the crops, the rest being lost in the canals that transport the water to the fields and due to inefficient irrigation practices used on-farm. It is believed that over the past decade, adequate maintenance, repair, and renovation of the irrigation infrastructure were not performed at a meaningful level, and water losses from deteriorating canals, gates, and other facilities have increased.

Most land is under furrow irrigation, with drip irrigation accounting for about 5% of the irrigated cropland (used primarily on orchard crops), and sprinkler irrigation accounts for about 3%. Even though the

TABLE 4 Cultivated Land along the Amu Darya and Syr Darya Rivers

	Before 1917	1960	1980	1992
Millions of hectares	5.2	10	15	18.3

water saving benefits of gated pipe are well known in the region, less than one-sixth of the farms use this technology. Reasons may include costs and product availability. Most farms follow the centuries' old practice of cutting earthen canals with shovels in order to divert water into the field. The volume of water available at these farm ditches is not sufficient to provide an even distribution of water over the field. As a result, water logging and soil salinity now affects about 40% of all the cultivated land in the region.

Muynak and Aralsk

Of all the villages affected by the drying of the Aral Sea, Muynak is the best known. Historically, Muynak was located on an island of the vast Aral Sea delta at the convergence of the Amu Darya River in Karakalpakstan (a semi-autonomist republic in Uzbekistan). In 1962, the island became a peninsula. By 1970 the former seaport was 10 km from the sea. The retreat of the sea accelerated and the town was 40 km from the sea by 1980, 70 km in 1995, and close to 100 km today.

Over 3000 fishermen once worked the abundant waters around Muynak which supported 22 different commercial species of fish. In 1957, Muynak fishermen harvested 26,000 tons of fish, about half of the total catch that year taken from the Aral Sea. Muynak also produced 1.1 million farmed muskrat skins which were used to produce coats and hats.

The Kazakhstan city of Aralsk, was once located on the northern edge of the Aral Sea, and like Muynak, had major fisheries and commerce industries. A major shipping and transport industry existed between these two cities. As the Aral Sea skunk, Aralsk found itself farther and farther from the shore which had retreated nearly 129 km by the 1980s. In the early 1990s, a dam was built just to the south of the mouth of the Syr Darya, to protect the northern part of the Aral Sea, letting the southern portion of the Aral Sea evaporate. Although only 10% of the water in the Syr Darya River reaches the northern part of the Aral Sea, the Little Aral has risen 3 m since the construction of the dam, and the shoreline has crept to within 16 km of the town.

Environmental Problems

The Aral Sea is an unfortunate example of an old Uzbek proverb: "at the beginning you drink water, at the end you drink poison." As the rivers flow through cultivated areas, they pick up fertilizers, pesticides, and salts from runoff, drainage water and groundwater flow. In the 1960s, it was common for about 550 kg ha^{-1} of chemicals to be applied to cotton fields in Central Asia, compared to an average of 25 kg used for other crops in the Soviet Union. Residues of these chemicals are now found on the dry seabed. Estimates are that millions of tons of dust are picked up from the seabed and distributed over the Aral Sea region.

The Sea may have been contaminated from runoff from by two former USSR military installations in the area. A chemical weapons testing facility was located on the Ust-Jurt Plateau (north shore), and was closed in the mid-1980s. Renaissance Island (Vozrozhdeniya Island), located in the central Aral Sea, was the site of the former USSR Government's Microbiological Warfare Group which produced the deadly Anthrax virus. Some scientists believe that some containers holding the virus were not properly stored or destroyed. As the Aral Sea continues to dry and water levels recede, the ever-expanding island will soon connect to the surrounding land. Scientists fear that reptiles, including snakes that have been exposed to the various viruses, will move onto the surrounding land and possibly infect the humans living around the shores of the Aral Sea.

The Aral Sea once supported a complex ecosystem, an oasis in the vast desert. Over 20 species of fish are now extinct. Karakalpakstan scientists believe that a total of about 100 species of fish and animals that once flourished in the region are now extinct, as are many unique plants.

Residents believe that there is a direct correlation between the drying of the sea and changes in climate of the Aral Sea Basin. The moderating effect of the sea has diminished and temperatures are now about 2.5°C higher in the summer and lower in the winter. Rainfall in the already arid basin has decreased by about 20 mm.

Human Tragedy

Over the last 50 years, there has been a large increase in mortality, illnesses, and poor health in the region. Some estimate that 70%–90% of the population of Karakalpakstan suffer some an environmentally induced malady. Tuberculosis is rampant. Hardest hit are women and children. Common health problems include kidney diseases, thyroid dysfunctions, anemia, bronchitis, and cancers.

Conclusion

Some accounts are that since 1984, hundreds of international agreements have been signed to address Aral Sea problems. The early agreements had the goals of first stabilizing the Sea, then slowly increasing flows to restore its ecosystem. In 1992, the Interstate Commission for Water Coordination was formed by the five central Asian republics, which also accepted, in principle, to adhere to the limits on water diversions as set during the Soviet era in 1984 and 1987. To date, however, no progress has been made on stabilizing or reversing the declining inflows. With no water reaching the Aral Sea from the Amu Darya, scientists predict that this portion of the sea (the Large Aral Sea) will disappear by 2020 (Figure 1).

FIGURE 1 This NASA photograph (STS085–503-119) was taken in August 1997 and looks toward the southeast. The Amu Darya River is visible to the right and the Syr Darya on the left. The Aral Sea is now separated into the Small Aral to the north and the Large Aral to the south. Shown are the approximate extent of the Aral Sea in 1957 before a massive expansion of irrigation diversions from the rivers.

Bibliography

1. The Aral Sea Homepage: Available at http://www.dfd.dlr.de/app/land/aralsee/
2. Requiem for a dying sea: Available at http://www.oneworld.org/patp/pap_aral.html/
3. Disappearance of the Aral Sea: Available at http://www.grida.no/aral/maps/aral.htm
4. Earth from Space: Available at http://earth.jsc.nasa.gov/categories.html

31

Chesapeake Bay

Sean M. Smith

Introduction

The Chesapeake Bay is the largest estuary within the U.S.A.[1] Fresh water flows to the Chesapeake Bay from a watershed that covers an estimated 166,709 km^2, including portions of Delaware, Maryland, New York, Pennsylvania, Virginia, Washington, DC, and West Virginia (Figure 1). The ecological productivity of the estuary has made it an important resource for Native Americans, European immigrants, and current residents in the region.[2] The population in the contributing watershed in recent years has swelled to over 15 million people, resulting in extensive direct and indirect impacts that are now a focus of a large-scale restoration effort led by the US Environmental Protection Agency.[1]

Discussion

Located along the Mid-Atlantic coast of the U.S.A. within the limits of Maryland and Virginia, the Bay is approximately 304-km long, has an estimated surface area of 11,603 km^2, a width that ranges from 5.5 to 56 km, and an average depth of approximately 6.4 m.[1] Salinity in the tidal portions of the estuary transition from "fresh" conditions (i.e., 0–5 parts per thousand (ppt) salt concentration) at the northern-most end to "marine" conditions (30–35 ppt salt concentration) at the southern boundary with the Atlantic Ocean.

Evidence indicates that the modern Chesapeake Bay began forming approximately 35 million years ago with a meteorite impact in the proximity of what is now the confluence of the Bay with the Atlantic Ocean.[2,3] The impact created a topographic depression that influenced the location and alignments of several large river valleys, including those associated with the present day Susquehanna, Rappahannock, and James Rivers. Since then, the river valleys have been periodically exposed and flooded in response to cycles of global glaciation and associated fluctuations in sea level. The most recent, the Wisconsin glaciation, began retreating approximately 18,000 years ago. The retreat resulted in a rise in sea level by almost ninety meters, drowning the river valleys and forming the current Bay.

Eleven large rivers drain the Bay's watershed, the Susquehanna River from Pennsylvania and New York providing the largest contribution with an average of 98 million m^3/day flowing into the northern end of the estuary. The rivers drain one or more of five different physiographic provinces within the watershed, including the Appalachian Plateau, Ridge and Valley, Blue Ridge, Piedmont, and Coastal Plain (Figure 1).[4] Each province's geologic composition and history creates dramatically different landscape settings from the western to eastern sides of the drainage basin. The Appalachian, Ridge and Valley, and Blue Ridge are characterized by mountainous terrain, a dominance of sandstone along ridge tops, and several carbonate valleys. The Piedmont has less relief, is dominated by metamorphic rocks, and is characterized by a surface that has been dissected by dendritic stream channel networks. Further to the east, the Coastal Plain is characterized by thick layers of unconsolidated geologic materials overlying bedrock deep beneath the surface. Waterways that flow from the Piedmont into the Coastal Plain traverse the "Fall Zone," a region that is easily distinguished by waterfalls coincident with an abrupt drop in the underlying bedrock elevations. Major ports and cities were developed along the Fall Zone, including Washington, DC, Baltimore, Maryland, and Richmond, Virginia, because of their locations at the upstream terminus of navigation from tidal waters and proximity to hydropower sources.

The Chesapeake Bay estuary is naturally dynamic and characterized by physical conditions that can be stressful to aquatic organisms. The salinity gradient broadly governs the spatial distribution of aquatic habitat types. Alterations in currents, wind, and freshwater inputs can cause salinity conditions to vary over time. The shallow depths also cause colder winter and warmer summer water temperatures compared to the open ocean. These spatial and temporal fluctuations can create physiologically challenging conditions. However, many organisms have adapted and use the abundant nutrients and physical habitat in different portions of the estuary for specific periods of their life cycles or seasons of the year. As a result, the Bay supports an estimated 3600 species of plants, fish, and animals, including 348 finfish and 173 shellfish species. Some of the most notable of these include striped bass, American shad, blueback herring, blue crab, and the American oyster.[5] The name "Chesapeake" itself was coined from the Algonquin American-Indian word "Chesepiooc" meaning "great shellfish bay."[5]

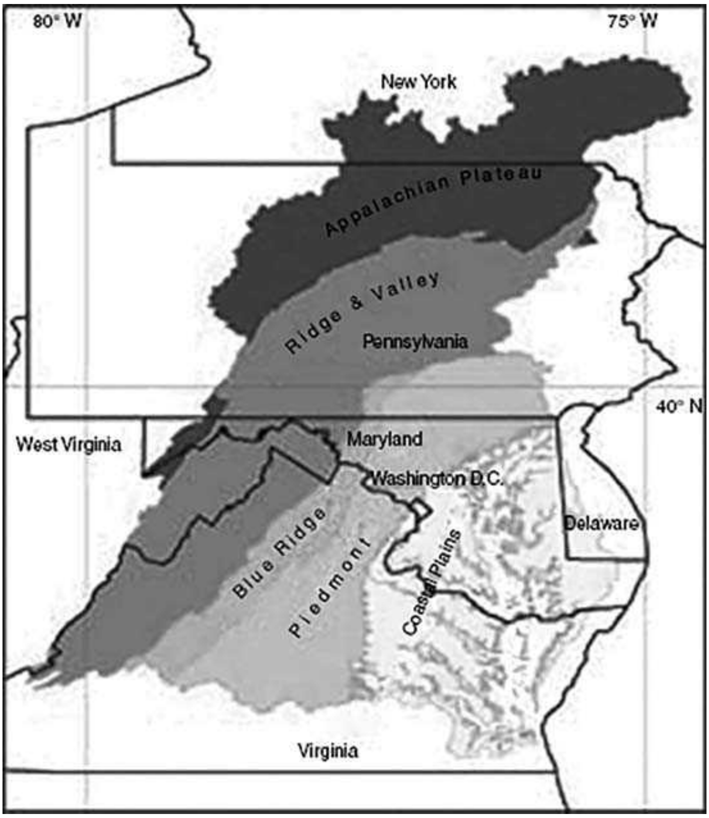


FIGURE 1 The Chesapeake Bay estuary and its watershed, including physiographic provinces and state boundaries. (Courtesy of M. Herrmann, Maryland Department of Natural Resources.)

Archaeologists estimate that Native American inhabitants first arrived in the Bay region from the south or west approximately 12,000 years ago as the ice sheets associated with the Wisconsin glaciation began to retreat and temperatures increased.[2] The first inhabitants are presumed to have been nomadic; however, archeological evidence suggests that selective food production started as early as 5000 years ago and settled towns began to be formed approximately 1300 years ago as the population density in the region increased. Recovered artifacts provide evidence of the extensive use of the Bay by the early inhabitants for travel, communication, tools, and food.

The first recorded European contact with the Chesapeake Bay region was by the Italian captain, Giovanni da Verrazano in 1524.[2] The English established one of the most well known early settlements at Jamestown, Virginia in 1607. English colonization expanded through expeditions to the north in the Bay, partly led by the famed Captain John Smith. Immigration to the region increased throughout the 1600s and much of the area was settled by the mid-1700s. The colonists made extensive use of the resources provided by the estuary, its wetlands, and tributaries. Shellfish, including oysters, blue crabs, and hard and soft clams, were harvested from shallow water areas.[2] The numerous piles of oyster shells that can be found near Coastal Plain tidal areas provide support for written claims of the extensive oyster beds that existed in the Bay when the European colonists arrived. Traps and nets were used to harvest finfish, including herring, striped bass, and shad. Migratory waterfowl, such as ducks and geese, were also plentiful food sources.

The rapid growth in the human population since European colonization of the Chesapeake Bay region dramatically increased the harvest of finfish, shellfish, waterfowl, and mammals naturally supported by the estuary. Extensive landscape alterations also caused direct and indirect physical changes to the Bay and its tributaries. The combination of overharvesting, pollution, and physical alterations has severely impacted the ecosystem and many of the species that historically flourished in the estuary.[1] Dramatic declines have been documented by the harvest records of popular commercial fisheries such as shad and striped bass. Records indicate a decline in the catch of blue crabs per unit of effort since the 1940s. The oyster harvest is currently at less than 1% of historic levels, although this reduction is partly attributed to disease. Many other species not harvested commercially have also been affected by the alterations in the Bay ecology that have accompanied European settlement and population growth.

One of the most important impacts to the Chesapeake Bay has been the increased erosion rates and downstream sedimentation caused by extensive deforestation of the watershed.[6] The influx of sediment into the tidal estuary has reduced the water depths in many embayments that once served as navigable ports.[7] Elevated suspended sediment inputs during storm events also increase turbidity in the tidal water column.[6] The resulting decrease in water clarity, which has been exacerbated by algal blooms associated with nutrient runoff pollution, reduces submerged aquatic vegetation (SAV) growth in shallow areas (i.e., depths less than 2 m). SAV coverage on the Bay bottom is estimated to have declined from approximately 80,900 hectares in 1937 to 15,400 hectares in 1984.[1] The loss has negative implications for a variety of species that use the vegetation for habitat, including blue crabs and juvenile finfish.

An extensive effort to restore the Bay has been undertaken by the US federal government in coordination with states in the watershed.[1,8] A large part of this effort has been focused on the recovery of the historic SAV distributions, as well as reversal of the abnormally low oxygen levels that now occur in the main stem of the Bay and its major tidal tributaries during summer months.[1,9] As with the water clarity problems, the low dissolved oxygen is related to excess nutrient inputs, mainly nitrogen and phosphorous, which stimulate algal production. The oxygen depletion occurs because of algal decomposition, resulting in, estuarine habitat degradation. Substantial reductions in nutrients from watershed runoff have been concluded to be necessary to achieve restoration goals related to both SAV and low dissolved oxygen.[1,9]

Conclusions

The Chesapeake Bay is a large and historically productive estuary on the Mid-Atlantic coast of the U.S.A., with extensive fisheries and wildlife resources. The Bay ecosystem has been impaired by watershed alterations and overharvesting accompanying human population growth in the region, thereby inspiring an extensive government-supported restoration effort. A large part of the restoration focuses on sediment and nutrient pollution associated with runoff from the contributing watershed.

References

1. http://chesapeakebay.net (accessed July 11, 2005).
2. Grumet, R.S. *Bay, Plain, and Piedmont: A landscape history of the Chesapeake Heartland From 1.3 Billion Years Ago to 2000;* The Chesapeake Bay Heritage Context Project; U.S. Department of the Interior: Annapolis, MD, 2000; 183 pp.
3. Powars, D.S.; Bruce, T.S. *The Effects of the Chesapeake Bay Impact Crater on the Geological Framework and Correlation of Hydrogeologic Units of the Lower York-James Peninsula, Virginia;* U.S. Geological Survey Professional Paper 1612; United States Geological Survey: Reston, VA, 1999.
4. Smith, S.; Gutuierrez, L.; Gagnon, A. *Streams of MarylandTake a Closer Look*; Maryland Department of Natural Resources: Annapolis, MD, 2003; 65 pp.
5. White, C.P. *Chesapeake Bay: Nature of the Estuary, A Field Guide;* Tidewater Publishers: Centreville, MD, 1989.
6. Langland, M.; Cronin, T. *A Summary Report of Sediment Processes in Chesapeake Bay and Watershed*; United States Geological Survey Water Resources Investigations Report 03–4123; United States Geological Survey: Reston, VA, 2003.
7. Gottschalk, L.C. Effects of soil erosion on navigation in upper Chesapeake Bay. Geogr. Rev. **1945**, 35, 219–238.
8. Ernst, H.R. *Chesapeake Bay Blues*; Rowman and Littleford Publishers, Inc.: Lanham, MD, 2003.
9. Macalaster, E.G., Barker, D.A., Kasper, M., Eds.; *Chesapeake Bay Program Technical Studies: A Synthesis*; United States Environmental Protection Agency: Washington, DC, 1983.

32

Giant Reed (*Arundo donax*): Streams and Water Resources

Gretchen C. Coffman

Introduction

Both natural and anthropogenic disturbances along rivers in Mediterranean-type climates are thought to promote the spread of invasive plant species in natural as well as altered riparian ecosystems.[1–5] A large bamboo-like member of the grass family (Poaceae), giant reed (*Arundo donax* L.), is one of the greatest invasive species threats to streams in arid and Mediterranean-type climate regions. *A. donax* forms extensive monotypic stands, successfully invading rivers in these regions. Infestations of *A. donax* are known to increase risks of flooding, compete with indigenous riparian species for scarce water resources, create unnatural fire hazards, and reduce the value of riparian habitat for most wildlife. Although many organizations are actively removing small areas of *A. donax* from streams in California, larger strategic watershed-scale removal is necessary to prevent further invasion and impacts to water resources.

Biology and Ecology

A. donax is one of the most successful weedy invaders in the highly dynamic, disturbance-defined rivers and riparian ecosystems of arid and Mediterranean-type climate regions[6,7] (Figure 1). *A. donax* is a tall, erect perennial grass species with culms 1–4 cm in diameter and 2–10 m in height, with two ranked leaves 2–6 cm wide at the base tapering to a point and up to 70 cm in length or more, and tough fibrous roots that penetrate more than 8 ft deep emanating from large horizontal creeping rhizomes. Although the seeds of this reed-like grass species are mostly sterile outside its native range,[8,9] *A. donax* colonizes readily via vegetative propagation; it is dispersed downstream when small pieces of its culm or rhizome break off during flooding and land on bare, moist substrates.[1,6,7,10,11] Fragments of the rhizome or culm as small as 2 cm^2 sprout under most soil types, depths, and soil moisture conditions.[7,10,11] Growing at an extremely rapid rate (up to 6.25 cm per day under optimal conditions), *A. donax* quickly establishes on exposed or sparsely vegetated soil and grows to more than 8 m in height after only a few months.[12,13]

FIGURE 1 *A. donax* infestation along the Santa Clara River in Ventura County, California.

Once established, *A. donax* clumps expand outward by clonal propagation (large rhizomes), crowding and displacing indigenous shrubs, herbs and grasses, and trees, especially under elevated soil moisture, nutrient, and light conditions.[8] In this manner, *A. donax* forms extensive stands, or monocultures, along floodplains and terraces of river and stream systems. *A. donax* is able to persist in ecosystems that experience seasonal drought by storing carbohydrate reserves in large rhizomes and sharing water resources through clonal integration.[14,15] When stressful conditions subside, shoots emerge.

A. donax is thought to be indigenous to freshwaters of Eurasia,[16] extending from Southeast Asia to the Mediterranean Basin, although the precise extent of its native distribution is unclear. Herbivore diversity suggests that it is Mediterranean in origin, but its native range may extend much farther. Several thousand years ago, *A. donax* was thought to have been spread around the Mediterranean Basin for use in erosion control; production of reeds for wind instruments; and construction of roofs, ceilings, fences, and baskets.[17] It has been introduced to most tropical and warm, temperate regions worldwide, including North and South America, Southern Africa, and Australia, and thrives below 350 m in elevation.[6,18] In North America, *A. donax* has become especially devastating to riparian habitats in California's Mediterranean climate region, creating significant impacts to natural river functioning and sustainability.[19] In Southern California, *A. donax* was originally planted along irrigation canals for erosion control and used as building materials and windbreaks.[20] Carried by floodwaters, *A. donax* eventually made its way to adjacent streams and rivers, and by the 1820s, patches were commonly found along floodplains of many streams, including the Los Angeles River.[20] However, it appears that *A. donax* has only recently succeeded in invading natural riparian ecosystems in Southern California (i.e., replacing native riparian vegetation).

Because of its clonal growth strategy, ability to colonize rapidly after disturbance, use of available resources, tolerance of stress, and high growth rate, *A. donax* is one of the most successful riparian weedy invaders in arid and Mediterranean-type climates.[12] Following an era of human alterations to river systems in Southern California, it was widely dispersed throughout riparian ecosystems in the floods of 1969, established in terrace and floodplain locations, and is now thriving in riparian ecosystems throughout this region.[21] Factors such as water, nutrients, light, and fire that are abundant in highly modified riparian ecosystems of arid and Mediterranean-climate regions increase the competitive ability of *A. donax*.[21] Although *A. donax* grows primarily in floodplains and terraces of low-gradient river and stream systems,[21,22] it may be found on beaches, around homes, in higher elevations, and next to hot springs where planted. *A. donax* forms huge infestations in open floodplains with high soil moisture and excess nutrients and in areas susceptible to wildfire.[21] *A. donax* successfully invades areas consisting of any soil type and once established can grow well in many soil moisture regimes.[7,21,23] Established stands recover readily after aboveground biomass is removed by wildfire, floods, frost, or mechanical means. In fact, the natural flood and wildfire regime characteristic of Mediterranean-type climates promotes growth and invasion of *A. donax*.[21]

From the time of early human settlement in arid and Mediterranean-type climates, humans have dammed, channelized, mined, diverted, and developed along rivers.[24,25] These alterations have magnified the susceptibility of streams in these regions to plant invasions by weedy species.[26,27] Human alterations associated with urbanization of watersheds in California in addition to the natural flood and wildfire processes have created ideal conditions for *A. donax* invasion. Increased water, nutrients, and light availability, as well as occurrence of fire in riparian ecosystems in Southern California, are thought to promote *A. donax* invasion.[6,21,28,29] Ever-expanding residential and agricultural development in coastal Southern California has led to increased water availability and nutrient loading of riparian ecosystems. The once vast low-lying areas of riparian forest continue to be removed to make room for agriculture, golf courses, and residential and commercial development. Consequently, open areas along floodplains formed by floods and clearing of terraces for development create an ideal location for *A. donax* to establish and invade riparian ecosystems. Furthermore, fire is more frequent in riparian corridors owing to anthropogenic ignition during the dry summer and fall months when *A. donax*-infested areas provide a large amount of dry fuel.[21,30] Because of higher post-fire growth rate and immediate growth response compared to native plant species, fire appears to contribute to the *A. donax* invasion process, especially in riparian terraces located next to fire-prone shrublands.[21]

Effects on Streams and Water Resources

Infestations of *A. donax* have created serious physical and biological problems along streams in arid and Mediterranean-type climates.[6,13,22,29] Where it grows extensively along floodplains, *A. donax* physically obstructs natural water flow, thereby increasing the risk of flooding. *A. donax* uses more water than native plant species, outcompetes native riparian species, thus reducing the value of riparian habitats for wildlife, and creating unnatural fire hazards.[21]

Flooding

Large infestations of *A. donax* within the active floodplains increase stream roughness during moderate to large flood events, forcing flood stages to higher levels and flooding adjacent property. During very high winter flows in California, *A. donax* is removed from the floodplain, floats downstream, and creates debris dams at bridges and culverts.[22] In addition, *A. donax* biomass collects in large piles along beaches after large flood events (Figure 2). Although originally planted to assist in erosion control, it now acts as an agent of erosion in California streams. The shallow rhizomes of the large, top-heavy plants growing along stream banks are undercut by high flows, causing bank erosion and instability. Economic losses due to effects of *A. donax* invasion include costs associated with repair of flood damage to property and bridges, beach cleanup, and bank stabilization repair.

Water Use

Water loss due to high evapotranspiration (ET) of *A. donax* infestations is of increasing concern in arid and Mediterranean-type climates where water resources are scarce and the plant continues to spread.

FIGURE 2 *A. donax* debris litters beaches in California after winter storms.

Using transpiration rates of rice (another C_3 species thought to have similar transpiration rates), estimates of *A. donax* water use suggest that it uses three times more water than native riparian species.[31] Other studies using a variety of methods indicate that ET of *A. donax* (1.2–7.5 m/yr) may be much higher than that of native riparian vegetation such as *Salix* spp., *Populus* spp. (1.0–3.3 m/yr), and mixed riparian communities of arid and Mediterranean-type climates (0.11–1.6 m/yr).[32–35] On the Santa Ana River alone, *A. donax* was estimated to transpire 37,500 acre-feet more water per year than native plants worth approximately $12 million at drinking water costs.[31] However, comprehensive studies are needed that compare water-use efficiency (WUE) of *A. donax* to various native species under different environmental conditions to determine exactly how much water is lost due to this invasive species. Excess water used in *A. donax* transpiration could be salvaged for groundwater recharge, drinking water supply, agricultural irrigation, and augmentation of in-stream flow for native vegetation and wildlife.[36] Furthermore, saturated soils that persist after flood events provide a lasting legacy to sustain plants over longer timescales. Using oxygen ($\delta^{18}O$) and hydrogen (δ^2H) stable isotopes analyses, Moore et al.[37] found that groundwater was not the dominant source for *A. donax* when soil moisture was sufficient for plant uptake following a flood.

Recent studies focused on water use, photosynthetic performance, and growth of *A. donax* under varying levels of soil moisture, carbon dioxide, and drought conditions. Watts and Moore[38] found that stands of *A. donax* used approximately 8.8 ± 0.9 mm of water per day during the peak of the growing season on the Lower Rio Grande in South Texas, the high end of the spectrum for plants. They also found that transpiration and leaf area index declined as water availability decreased, suggesting a sensitivity to drought and declining water tables. In another study, *A. donax* plants were subjected to a gradual drought stress for 3 weeks, after which they were returned to fully hydrated soil conditions for 1 week.[39] Overall, plant dry weight and key growth parameters were not significantly affected; however, photosynthetic performance was impaired. Physiological functions were restored after rewatering, reflecting the environmental plasticity of *A. donax*. Both atmospheric enrichment of carbon dioxide (CO_2) and increased drought are predicted to occur in regions where *A. donax* grows in arid and Mediterranean regions of the United States. Nackley et al.[40] found that increased CO_2 affected growth and water use of *A. donax*, reducing transpiration rates significantly. Reduced transpiration delayed drought responses, increased WUEs, and extended periods of assimilation but could not prevent desiccation and photosynthetic decline during extreme drought.

Wildfire

Wildfires ignited by humans at unnatural and dangerous times of the year burn rapidly through riparian corridors infested with *A. donax* and may help spread fires across watersheds and along riparian corridors.[21] Historically, dense biomass that accumulated over a period of 30–50 years or more in chaparral communities of California and shrublands in other Mediterranean-type climate regions caused fires to ignite.[41–44] Although fire was once a natural part of shrubland ecosystems in many Mediterranean-type regions, large riparian ecosystems provided natural firebreaks because native vegetation retained foliar water that resisted ignition.[13] Lightning was the primary cause of wildfires, especially during July and August under dry, low-humidity conditions and would commonly burn slowly for months.[42] Currently, however, most wildfires in these areas are anthropogenic in origin, occur much more frequently, and during strong Santa Ana wind conditions starting in September. For example, all 14 concurrent fires in October 2003 (739,597 acres burned) resulted from human activities.[41]

Invasion of annual grass species has been linked to altered fire regimes in rangelands, deserts, and wild lands of California and the Western United States.[41,47–49] However, *A. donax* may be an even bigger problem in riparian ecosystems of altered Southern California fire regimes because of its perennial growth form (the large volume of biomass produced) and rapid recovery after fire (Figure 3a and b).[21] Several accounts suggest that infestations of giant reed have increased fuel load as well as fire frequency and intensity along riparian corridors.[12,13,30,50] Thus, *A. donax* invasion appears to have created a positive feedback cycle or an invasive plant-fire regime[21] similar to those presented by other grass species.[45,46]

Biodiversity and Wildlife

A. donax has little habitat or food value for wildlife because of its dense growth structure and high content of noxious chemicals.[6,12,13] The federally endangered least Bell's vireo (*Vireo bellii pusillus*) and other riparian birds require structural diversity provided by riparian scrub and mature forest communities for breeding.[6,13,51] When naturally diverse riparian vegetation types are replaced by thick stands of *A. donax*, bird species abundance and other native wildlife have been found to decline.[6,13,45,46] Movement of medium to large mammals is most likely impaired by dense *A. donax* infestations. Herrera and Dudley[52] showed that arthropod abundance and diversity associated with native riparian vegetation was twice that associated with *A. donax* infestations. In addition, fish and aquatic invertebrates may be affected by increased stream temperature owing to lack of shading where *A. donax* has replaced mature riparian forests.[6]

Control Methods, Restoration, and Revegetation

Tens of millions of dollars ($7–$25 thousand per acre) have been spent in efforts to remove *A. donax* from riparian ecosystems in the Central Valley and coastal California. Although most attempts have been successful in removing small infestations on riparian terraces, *A. donax* continues to thrive in floodplains. An understanding of the ecological conditions that promote continued growth and invasion of *A. donax* is needed for its effective control. Management strategies for the control and removal of *A. donax* should be based on the location and size of the infestation. Priority should be given to removal of *A. donax* from riparian terrace habitats where infested areas are easily accessible and require less maintenance than along floodplains, especially infestations located adjacent to fire-prone shrubland plant communities.[53] Removal of large *A. donax* infestations on riparian terraces with high soil moisture and nutrient availability will be most difficult but is essential in removing the largest source of propagules to prevent future reinfestation. Active revegetation with native plants after *A. donax* removal

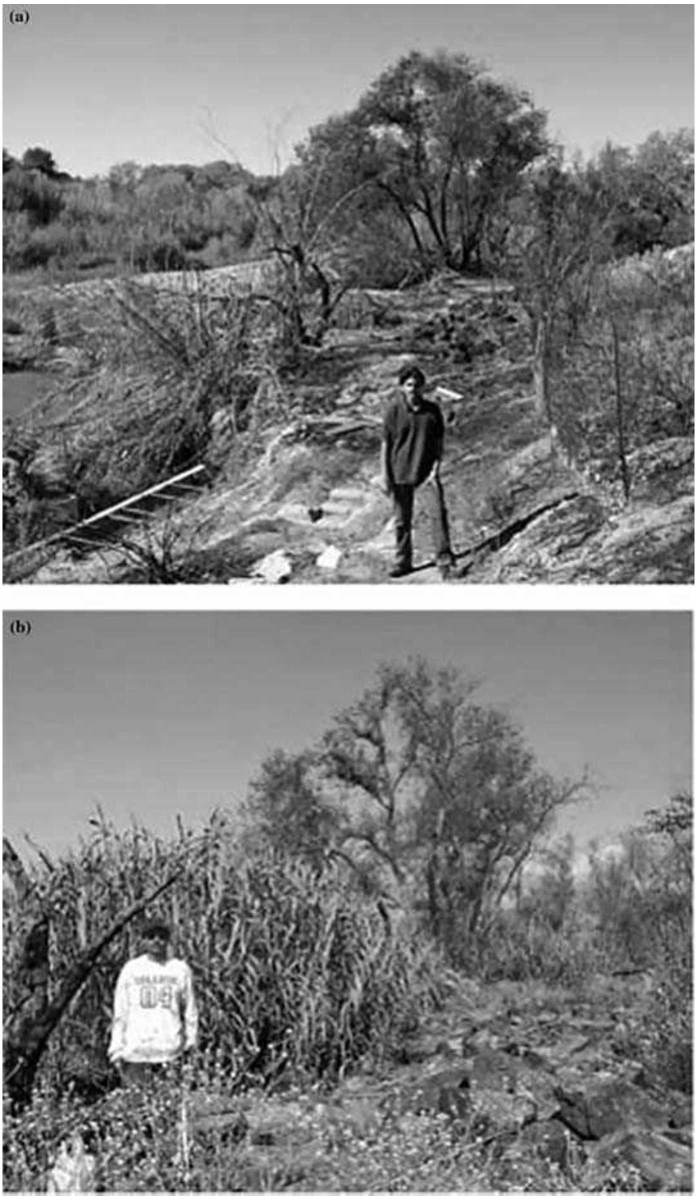


FIGURE 3 Three weeks (a) and 6 months (b) after Verdale–Simi Fire, *A. donax* invades riparian terrace along Santa Clara River in Ventura County.

is recommended to prevent reinfestation of *A. donax* or other weeds and restore functional riparian ecosystems. Unless *A. donax* is removed from floodplains on a watershed-scale working from the headwaters downstream, it is likely to recolonize removal areas after flood events. Watershed removal planning is underway in several large streams in Southern California to eradicate *A. donax* from floodplains, including the Santa Clara and Santa Ana Rivers.

Both mechanical and hand clearing techniques may be used to remove *A. donax*. Mechanical clearing methods include mulching or total excavation of all aboveground and belowground biomass. Hand clearing methods include either painting of *A. donax* stumps with herbicide after cutting or foliar applications of herbicide (glyphosate).[54] Research on biocontrol agents for *A. donax* is underway on the Santa Clara River, California and in Weslaco, Texas (Dudley, *Personal Communication*, 2006).

Conclusions

One of the biggest threats to streams and water resources in Mediterranean-type climates is invasion of *A. donax*. Forming large monocultures under ideal resource conditions along streams, *A. donax* increases flooding, promotes the spread of wildfire, outcompetes natives for water resources, and decreases wildlife value of riparian habitat. Although millions of dollars are spent every year to remove *A. donax* in California, many rivers and streams are still heavily infested. Effective removal and control strategies must be based on an ecological understanding of the invasion process and removal areas prioritized based on gaining the greatest ecological benefit for the least effort. Control of *A. donax* from watersheds in Mediterranean-type climates is an important initial step in restoration and long-term sustainability of riparian ecosystems.

References

1. Else, J.A.; Zedler, P. 1996 Annual Combined Meeting, Providence, RI, 10–14 August 1996, *Bull. Ecol. Soc. Am.* **1996**, *77*, 129.
2. Gregory, S.V.; Swanson, F.J.; McKee, W.A.; Cummins, K.W. An ecosystem perspective of riparian zones, *BioScience* **1991**, *41*, 540–551.
3. Pysek, P.; Prach, K. How important are rivers for supporting plant invasions? In *Ecology and Management of Invasive Riverside Plants*; Waal, L.C.d., Child, L.E., Wade, P.M., Brock, J.H., Eds.; John Wiley & Sons: New York, 1994; 19–26.
4. Drake, J.A.; Mooney, H.A.; diCastri, F.; Groves, R.; Kruger, F.J.; Rejmanek, M.; Williamson, M. *Biological Invasions: A Global Perspective*; John Wiley & Sons: Chichester, 1989.
5. Dudley, T. Exotic plant invasions in California riparian areas and wetlands. *Fremontia* **1998**, *26*, 24–29.
6. Bell, G.P. Ecology and management of Arundo donax, and approaches to riparian habitat restoration in Southern California. In *Plant Invasions: Studies from North America and Europe*; Brock, J.H., Wade, M., Pysek, P., Green, D., Eds.; Blackhuys Publishers: Leiden, 1997; 103–113.
7. Boose, A.B.; Holt, J.S. Environmental effects on asexual reproduction in *Arundo donax*. *Weed Res.* **1999**, *39*, 117–127.
8. Decruyenaere, J.G.; Holt, J.S. Seasonality of clonal propagation in giant reed. *Weed Sci.* **2001**, *49*, 760–767.
9. Khudamrongsawat, J.; Tayyar, R.; Holt, J.S. Genetic diversity of giant reed in the Santa Ana River, California. *Weed Sci.* **2004**, *32*, 395–405.
10. Else, J.A. San Diego State University: San Diego, 1996; 74 pp.
11. Wijte, A.H.B.M.; Mizutani, T.; Motamed, E.R.; Merryfield, M.L.; Miller, D.E.; Alexander, D.E. Temperature and endogenous factors cause seasonal patterns in rooting by stem fragments of the invasive giant reed, *Arundo donax* (Poaceae). *Int. J. Plant Sci.* **2005**, *166*, 507–517.
12. Rieger, J.P.; Kreager, D.A. *Proceedings of the California Riparian Systems Conference: Protection, Management, and Restoration for the 1990s*; Abell, D.L., Ed.; USDA Forest Service Gen. Tech. Rep. PSW-110; Berkeley, CA, 1989; 222–225.
13. Bell, G.P. *Arundo donax* Workshop Proceedings, Nov, 19, 1993; Jackson, N.E., Frandsen, P., Douthit, S., Eds.; Ontario, CA, 1994; 1–6.
14. Mann, J.J.; Barney, J.N.; Kyser, G.B.; DiTomaso, J.M. *Miscanthus x giganteus* and *Arundo donax* shoot and rhizome tolerance of extreme moisture stress. *GCB Bioenergy* **2013**, *5*, 693–700.
15. Kui, L.; Li, F.; Moore, G.; West, J. *Can the riparian invader, Arundo donax, benefit from clonal integration*. Weed Res. **2013**, *53*, 370–377.

16. Polunin, O.; Huxley, A. *Flowers of the Mediterranean;* Hogarth Press: London, 1987.
17. Perdue, R.E. *Arundo donax*—source of muscial reeds and industrial cellulose. *Econ. Botany* **1958**, *12*, 157–172.
18. Brossard, C.C.; Randall, J.M.; Hoshovsky, M.C. *Invasive Plants of California's Wildlands;* University of California Press: Berkeley, CA, 2000.
19. Rundel, P.W. Alien species in the flora and vegetation of the Santa Monica Mountains, CA: Patterns, processes, and management implications. In *2nd Interface* between *Ecology and Land Development in California*; Keeley, J.E., Baer-Keeley, M., Fotheringham, C.J., Eds.; U.S. Geological Survey Open-File Report 00-62, 2000; 145–152.
20. Robbins, W.W.; Bellue, M.K.; Ball, W.S. *Weeds of California;* California Department of Agriculture: Sacramento, 1951.
21. Coffman, G.C. *Factors Influencing Invasion of Giant Reed (Arundo donax) in Riparian Ecosystems of Mediterranean-type Climates*; University of California: Los Angeles.
22. DiTomaso, J.M. *Proceedings of the Arundo and Saltceder: The Deadly Duo Workshop*, Ontario, CA, June, 17, 1998; 1–5.
23. Singh, C.; Kumar, V.; Pacholi, R.K. Growth performance of *Arundo donax* (reed grass) under difficult site conditions of Doon Valley for erosion control. *Indian Forester* **1997**, *123,* 73–76.
24. Mount, J.F. *California Rivers and Streams: The Conflict Between Fluvial Process and Land Use*; University of California Press: Berkeley, 1995.
25. Palmer, T. *California's Threatened Environment: Restoring the Dream*; Island Press: Washington, DC, 1993.
26. Rundel, P.W. Landscape disturbance in Mediterranean-type ecosystems: An overview. In *Landscape Disturbance and Biodiversity in Mediterranean-Type Ecosystems. Ecological Studies 136*; Rundel, P.W., Montenegro, G., Jaksic, F.M., Eds.; Springer-Verlag: Berlin, 1998; 3–22.
27. Randall, J.M.; Rejmanek, M.; Hunter, J.C. Characteristics of the exotic flora of California. *Fremontia* **1998**, *26*, 3–12.
28. Wang, A. Environmental Science: Policy and Practice. In *Proceedings, Senior Research Seminar, Environmental Sciences Group Major-UGIS*; Dudley, T., Kennedy, K., Eds.; University of California: Berkeley, 1998; 720 pp.
29. Rundel, P.W. Invasive species. In *Southern California Environmental Report Card 2003*; Carlson, A.E., Winer, A.M., Eds.; UCLA Institute of the Environment: Los Angeles, CA, 2003; 4–11.
30. Scott, G. *Arundo donax* Workshop Proceedings; Jackson, N.E., Frandsen, P., Douthit, S., Eds.; Ontario, CA, 1994; 17–18.
31. Iverson, M.E. *Arundo donax* Workshop Proceedings, Ontario, CA. Nov, 1993; Jackson, N.E., Frandsen, P., Douthit, S., Eds.; 1994; 19–25.
32. Shafroth, P.B.; Cleverly, J.R.; Dudley, T.L.; Taylor, J.P.; Van Riper, C.; Weeks, E.P.; Stuart, J.N. Control of *Tamarisk* in the western United States: Implications for water salvage, wildlife use and riparian restoration. *Environ. Manage.* **2005**, *33* (3), 231–246.
33. Zimmerman, P. *Proceedings of the California Exotic Pest Plant Council*; California Exotic Pest Plant Council: Sacramento, CA, 1999.
34. Hendrickson, D.; McGaugh, S. http://desertfishes.org/cuatroc/organisms/non-native/arundo/Arundo.html, 2005; 17 pp.
35. Abichandani, S.L. *The Potential Impact of the Invasive Species Arundo donax on Water Resources along the Santa Clara River: Seasonal and Diurnal Transpiration*. University of California: Los Angeles. 2007.
36. The Nature Conservancy. Enhancing Water Supply through Invasive Plant Removal: A Literature Review of Evapotranspiration Studies on Arundo donax. file:///C:/Users/13106/Documents/Research/Arundo%20Research/TNC_Arundo_ET_Literature_Review_Feb2019.pdf, 2019; 6 pp.

37. Moore, G.; Li, F.; Kui, L.; West, J. Flood water legacy as a persistent source for riparian vegetation during prolonged drought: an isotopic study of *Arundo donax* on the Rio Grande. *Ecohydrology* **2016**, *9*, 909–917.
38. Watts, D.A.; Moore G.W. Water-use dynamics of an invasive reed, Arundo donax, from leaf to stand. *Wetlands*, **2011**, *31*, 725–734.
39. Pompeiano, A.; Remorini, D.; Vita, F., Guglielminetti, L.; Miele, S.; Morini, S. *Growth and physiological response of Arundo donax L. to controlled drought stress and recovery. Plant Biosyst.* **2017**, *151* (5), 906–914.
40. Nackley, L.L.; Vogt, K.A.; Kim, S.H. *Arundo donax* water use and photosynthetic response to drought and elevated CO_2. *Agric. Water Manage.* **2014**, *136*, 13–22.
41. Keeley, J.E.; Fotheringham, C.J. Lessons learned from the wildfires of October 2003. In *Fire, Chaparral, and Survival in Southern California*; Halsey, R.W., Ed.; Sunbelt Publications: San Diego, CA, 2005; 112–122.
42. Keeley, J.E.; Fotheringham, C.J. Historic fire regime in Southern California shrublands. *Conserv. Biol.* **2001**, *13*, 1536–1548.
43. Minnich, R.A. Fire mosaics in Southern California and northern Baja California. *Science* **1983**, *219*, 1287–1294.
44. Keeley, J.E.; Fotheringham, C.J.; Morais, M. Reexamining fire suppression impacts on brushland fire regimes. *Science* **1999**, *284*, 1829–1931.
45. D'Antonio, C.M.; Vitousek, P.M. Biological invasions by exotic grasses, the grass/fire cycle and global change. *Ann. Rev. Ecol. Syst.* **1992**, *23*, 63–87.
46. Brooks, M.L.; D'Antonio, C.M.; Richardson, D.M.; Grace, J.B.; Keeley, J.E.; DiTomaso, J.M.; Hobbs, R.J.; Pellant, M.; Pyke, D. Effects of invasive plants on fire regimes. *BioScience* **2004**, *34*, 677–688.
47. Keeley, J.E. Invasive plants and fire management in California Mediterranean-climate ecosystems. In *10th International Conference on Mediterranean Climate Ecosystems (MEDE- COS);* Arianoutsou, M., Papanastasis, V.P., Eds.; Millpress: Rhodes, Greece, 2004; (full text on cd).
48. D'Antonio, C.M. Fire, plant invasions, and global changes. In *Invasive Species in a Changing World*; Mooney, H.A., Hobbs, R.J., Eds.; Island Press: Washington, DC, 2000; 65–93.
49. Brooks, M.L. Peak fire temperatures and effects on annual plants in the Mojave Desert. *Ecol. Appl.* **2002**, *12*, 1088–1102.
50. Dukes, J.S.; Mooney, H.A. Disruption of ecosystem processes in western North America by invasive species. *Rev. Chil. Hist. Nat.* **2004**, *77*, 411–437.
51. Zembal, R. Riparian habitat and breeding birds along the Santa Margarita and Santa Ana Rivers of southern California. In *Endangered Plant Communities of Southern California*; Schoenherr, A.A., Ed.; Southern California Botanists: Claremont, CA, 1990; 98–114.
52. Herrera, A.M.; Dudley, T.L. Reduction of riparian arthropod abundance and diversity as a consequence of giant reed (*Arundo donax*) invasion. *Biol. Inv.* **2003**, *5*, 167–177.
53. Coffman, G.C.; Ambrose, R.F.; Rundel, P.W. *10th International Conference on Mediterranean Climate Ecosystems*; Arianoutsou, M., Papanastasis, V.P., Eds.; Millpress: Rhodes, Greece, 2004; (full text on cd).
54. Cornwall, C.; Dale, R.; Newhonser, M. 1999. *Arundo donax: A Landowner Handbook*. Sonoma Ecology Center and California State University, Sacramento Media Services: Long Beach, CA.

33

Inland Seas and Lakes: Central Asia Case Study

Andrey G.
Kostianoy

Introduction

Shallowing, desiccation, and degradation of freshwater and salt lakes and inland seas are among the major environmental problems at the beginning of the 21st century. All over the globe, water is being diverted for industrial, agricultural, and household uses, and many lakes are suffering from the resulting lack of inflow. There are clear indications that the growth of human population and the increasing use of natural resources, especially water, combined with climate changes, exert a considerable stress on closed or semi-enclosed seas and lakes. In many regions of the world, marine and lacustrine hydrosystems are or have been the objects of severe or fatal alterations ranging from changes in regional hydrological regimes and/or modifications of the quantity or quality of water resources, deterioration of geochemical balances (increased salinity, oxygen depletion, etc.), to mutations of ecosystems (eutrophication, decrease in biological diversity, etc.), to socioeconomic perturbations, which have been the consequences or may soon be in the near future.[1]

About 97.5% of earth's water is saltwater in oceans and seas.[1] Around 1% of that is brackish groundwater. The remaining 2.5% is freshwater. Nearly 70% of the world's freshwater is frozen in the Antarctic and Greenland ice sheets, glaciers, and permanent snow cover and ice. About 30% of all freshwater is groundwater. Lakes and rivers contain only about 0.25% of all freshwater. In 2002, the United Nations Environment Program (UNEP) announced that about 3 billion people would face severe water shortages by 2025, if the present consumption rates persist. Global and regional climate change, global and regional warming, and droughts amplify the freshwater problem. According to NASA Goddard Institute for Space Studies, the year 2018 was the fourth warmest since 1880. It was ranked behind those of 2016, 2017, and 2015. The past 5 years were the warmest years in the modern record.[2]

The goal of this chapter is to pay attention to the degradation of inland seas and lakes all over the world. They are very often called "dying" or "critical," meaning that the lake or inland sea is facing either a severe anthropogenic pressure in some form or a rapid change of its physical conditions owing to regional/global climate change.[3–6] In many cases, both mechanisms work together, and it is really very difficult to discriminate between natural and anthropogenic impact and to quantitatively assess them.

In this chapter, we address the most known examples of the degradation of the lakes all over the world, and then we focus on two Central Asia case studies—the Aral Sea and the Kara-Bogaz-Gol Bay. Conclusions summarize this chapter and briefly touch on future prospects related to the state of the Central Asia water bodies.

Degradation of Lakes in the World

A significant fall in the water level and/or increase in salinity of many large saline lakes and inland seas have taken place worldwide during the past century.[3–6] Examples include the following: in North America—Great Salt Lake (Utah), Walker and Pyramid lakes (Nevada), Lake Mead (Nevada, Arizona), Owens and Mono lakes and the Salton Sea (California), and Deadmoose Lake (Canada); in South America—Llancanelo (Argentina); in the Middle East—the Dead Sea (Israel/Jordan/West Bank) and Lake Van (Turkey); in Central Asia—the Aral and Caspian seas and Kara-Bogaz-Gol Bay (Turkmenistan), Lake Balkhash (Kazakhstan), and Lake Issyk-Kul (Kyrgyzstan); in China—Lop Nor and Qinghai Hu lakes; in Africa—Lake Chad and Lake Elmenteita (Kenya); in Japan—Lake Biwa; and in Australia—Keilambete, Eyre, Corangamite, Gnotuk, and Bullenmerri lakes.[5]

The most striking examples include (1) the Lop Nor Lake in China, which completely dried up by 1972; (2) the Aral Sea, which is following such a fate; (3) the Dead Sea, whose level dropped 23 m from 1970 to 2006; and (4) Lake Chad, which at the same time shrunk to about 5% of its size in 1963.

Northern and northwestern China has been experiencing a desiccation process since the 1950s owing to a decrease in precipitation by at least 30%.[3] As a result, Lake Lop Nor vanished completely in 1972 and became a nuclear testing site; the depth of Lake Ohlin at the head of the Yellow River dropped by more than 2 cm annually; the Qinghai Hu Lake water level decreased by an average of 10 cm/yr between 1959 and 1982 owing to a decrease in rainfall, groundwater supply, and the unsustainable use of the water for irrigation (the total drop since 1908 reached 11.7 m; the lake salinity increased from 5.6 to 12 g/L since 1950). Because of rapid population growth, the surface of Ebi Nor, the largest salt lake in northwest China's Xinjiang Uygur Autonomous Region, has shrunk to 530 km^2 in the past five decades (its surface was 1200 km^2 in the 1950s). As a result, many plant and animal species living in and around the lake have been extirpated.

The Dead Sea is a 378 m deep salty terminal lake at the border between Israel, Jordan, and the West Bank. The Dead Sea waters are among the saltiest (340 g/L) and densest (1.237 g/cm^3) lake waters/seawaters in the world.[7] Its level is determined by the balance between river runoff, precipitation, and evaporation. Since the 1960s, the hydrological regime of the Dead Sea has been strongly influenced by use of its watershed by Israel, Syria, and Jordan. Moreover, Israel and Jordan use the seawaters for mineral production at salt evaporation ponds located south of the sea, responsible for 25%–30% of the total Dead Sea evaporation. As a result, evaporation exceeded freshwater inflow to the sea. Since 1977, the length of the sea decreased from 80 to 50 km and the level has dropped by 23 m (1970–2006) at rates of 0.6–1.0 m/yr.[7] From 2000 to 2018, the sea level has dropped with a rate of 1.1 m/yr. The Dead Sea level drop has been followed by a groundwater level drop, causing brines in the underground layers near the shoreline to be flushed out by freshwater. This is the cause of the recent appearance of large sinkholes along the western shore. Plans have been developed to construct a pipeline to bring water from the Red Sea into the Dead Sea to stabilize or raise the water level. The pipeline, referred to as the "Peace Conduit" project, would bring about 450×10^6 m^3 of Red Sea water into the Dead Sea annually.[4] In May 2009 at the World Economic Forum, Jordan announced its plans to construct the "Jordan National Red Sea Development Project" (JRSP). This is a plan to convey seawater from the Red Sea near the Gulf of Aqaba to the Dead Sea. Water would be desalinated along the route to provide freshwater to Jordan, with the brine discharge sent to the Dead Sea for replenishment. In 2016, Ministry of Water and Irrigation of Jordan announced that the project will be completed by 2021.

Global warming (during the 20th century, the mean land surface temperature in Africa has increased by 0.9°C) and withdrawal and/or diversion of water from inflowing rivers are the reasons for the water level drop in several African lakes.[3–5] Lake Chad, once one of the largest in Africa, has been a source of freshwater for irrigation projects in Chad, Niger, Nigeria, and Cameroon. Since 1963, the lake has shrunk to nearly 1/20th of its original size (from 25,000 to 1,350 km^2 in 2000), owing to both climatic changes (including a 50% decline in rainfall) and high agricultural water (between 1983 and 1994, irrigation water use increased fourfold). The UNEP says that about half of the lake's decrease is attributable to human water use such as inefficient damming and irrigation methods. The other half of the shrinkage is due to regional climate change. There are considerations that the lake will shrink further and perhaps even disappear in the course of the 21st century.

Lake Victoria, shared by Kenya, Uganda, and Tanzania, with a surface area of 68,000 km^2, is the world's second largest and the largest African body of freshwater in terms of surface area.[3] It is of great socioeconomic importance for 20 million people living in the basin. Over the past few decades, Lake Victoria has been subjected to drastic ecological and water quality changes owing to pollution (including sewage discharges and agricultural runoff), sediments resulting from soil erosion in the catchment area because of deforestation and overgrazing and industrial pollution from many local industries. All these factors have resulted in the eutrophication of the lake because of the increase in nutrient supply to the lake, algal blooms, and massive fish deaths. Moreover, the introduction of exotic fish species such as the Nile perch has altered the freshwater ecosystem of the lake and driven several hundred species of native cichlids to extinction or near extinction. The water hyacinth *Eichhornia crassipes*, a native of the tropical Americas, was introduced by colonists to Rwanda and then advanced by natural means to Lake Victoria where it was first sighted in 1988. Without any natural enemies, it became an ecological plague. By forming thick mats of vegetation, it causes difficulties to transportation, fishing, hydroelectric power generation, and drinking water supply.

A number of lakes in North America have also experienced desiccation in the past century.[3,5] For instance, the Pyramid Lake in Nevada experienced a 21 m level drop since 1910, accompanied by a salinity increase from 3.8 to 5.5 g/L between 1933 and 1980.

In California, Owens Lake (once 280 km^2 large and 7–15 m deep) dried completely by 1926 due to irrigation diversions and withdrawals. The dry bed of Owens Lake has produced enormous amounts of windblown dust since the desiccation of the lake. The lake bed is probably the largest single source of PM10 dust (aerosol particles smaller than 10 microns in aerodynamic diameter) in the United States (by one estimate, 900,000–8,000,000 metric tons per year).[8] Unusually fine-grained, alkaline dust storms from the dry lake bed are a significant health hazard to residents of Owens Valley and nearby areas and impact air quality in a large region (40 km) around the lake bed.

In California, the Mono Lake's level dropped 17 m from 1919 to 1982 owing to diversions of their tributary system to the Los Angeles Aqueduct system. By 1982, the lake was reduced to 153 km^2, having lost 31% of its 1941 surface area. Before 1941, average salinity was approximately 50 g/L, but in January 1982, when the lake reached its lowest level (1942 m), the salinity had nearly doubled to 99 g/L. In 2002, it was measured at 78 g/L and is expected to stabilize at an average of 69 g/L as the lake replenishes (in August 2019, its level was at 1946 m).

Even the world's largest freshwater system, the North American Great Lakes, may be shrinking. They contain about 95% of the fresh surface water supply for the United States and 20% for the world. In 2002, the aggregate level of the five Great Lakes was at the lowest in more than 30 years. Since 1997, Lakes Huron, Michigan, and Erie have dropped over 1 m, and an additional drop of 0.5–1 m has been predicted. These changes are attributed to decrease in precipitation, enhanced evapotranspiration, and reduced ice cover because of higher temperatures and irreversible loss of water for urban and industrial uses (e.g., Chicago sends its used water taken from the lakes to the Mississippi basin after treatment, instead of back to the Great Lakes).

From the late 1990s until 2005, a multiyear drought in the Upper Colorado River Basin caused water level drop in Lake Powell. As people's demands on the river overcame the rate of recharge, the volume of the lake dropped to only 33% of capacity in 2005.[9]

In August 2010, Lake Mead reached its lowest level since 1956. The largest reservoir in the United States was straining from 12 years of persistent drought and increasing human demand after decades of population growth in the American Southwest. In August 2010, Lake Mead held just 37% of the lake's capacity (35 km^3). The lake level dropped 39 m since August 1985. Refilling the lake depends almost entirely on snowmelt from the Rocky Mountains and how much of that water is released from dams and reservoirs in the upper basin of the Colorado River (Wyoming, Colorado, Utah, and New Mexico).[10]

The Aral Sea

Before 1960, the Aral Sea (Figure 1) was a water-abundant sea lake in Central Asia that was the fourth largest in the world list of lakes after the Caspian Sea (USSR, Iran), Great Lakes (United States, Canada), and Victoria Lake (Africa).[11–18] The Aral Sea was about the size of the Netherlands and Belgium taken together. It is located in the largest deserts—Karakum and Kyzylkum. Navigation and the fishery (yearly catches of 44,000 tons) were developed here. The deltas of the Amu Darya, the major river of Central Asia, and Syr Darya bringing their waters into the Aral Sea were famous for their biodiversity, fishery, muskrat rearing, and reed production. The local population found occupation in the spheres related to the water infrastructure. This was a natural and stable period of the Aral Sea evolution that since 1960 was followed by the anthropogenic one, which continues until the present day.

The Aral Sea existed thanks to runoff from the two main rivers of Central Asia—Amudarya and Syrdarya. However, since 1960, riverine water resources have been irrationally used for increasing irrigation of agricultural lands and creation of artificial water reservoirs. As a result, the Aral Sea water balance was disrupted and irreversible alterations in the sea regime appeared that later escalated into one of the "largest ecological disasters of the 20th century." During the last 50 years, we have observed a progressive degradation of the Aral Sea and its environment. During this time period, the sea shrunk in size from 66,100 km^2 (in 1961) to 10,400 km^2 (in 2008), its volume decreased from 1066 to 110 km^3, the sea level dropped by 24 m (maximum depth of 69 m was observed in 1961), and its salinity (mineralization) rose from 10 to 116 g/L in the Western Large Aral Sea and about 210 g/L in the Eastern Large Aral

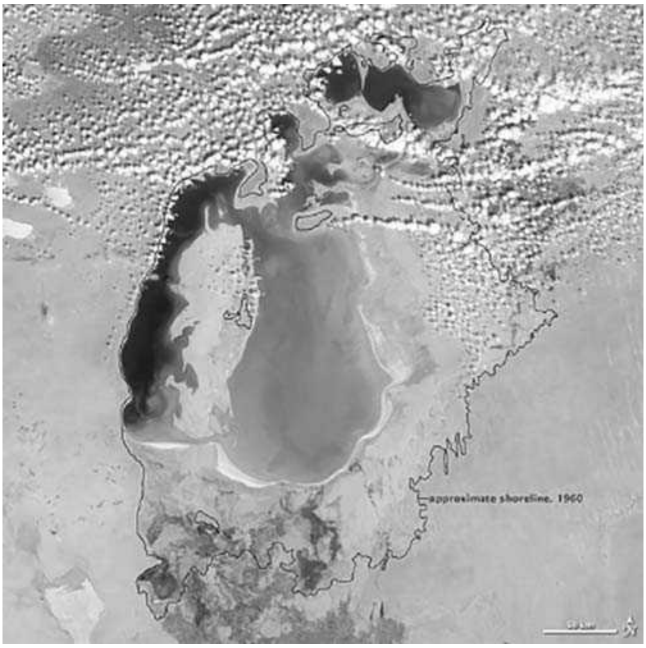


FIGURE 1 Satellite image (MODIS-Terra, 250 m resolution) of the Aral Sea on August 19, 2000, with a black line indicating the 1960 shoreline (http://earthobservatory.nasa.gov/Features/WorldOfChange/aral_sea.php, accessed on October 15, 2010).

Sea (Figure 1).[15,18,19] Today, the Aral Sea represents two isolated water bodies—a narrow and relatively deep the Western Large Aral Sea and the Small Aral Sea located in the north-eastern part of the satellite image presented in Figure 2. The former bottom of the Large Aral Sea is now a desert which is often called the Aralkum. Recent measurements (May 2019) of physical and chemical characteristics in the Western Large Aral Sea, performed by P.P. Shirshov Institute of Oceanology, Russian Academy of Sciences yearly since 2002, have recorded that the sea level has dropped by 30 m since 1961, and salinity is of 121 g/L (Personal communication by P.O. Zavialov). The ongoing desiccation, shallowing, and salinization of the Aral Sea have resulted in profound changes in its shape and its physical, chemical, and biological regime. The Aral Sea lost its economic importance, and the aftermath of its degradation represents a serious threat to the local population due to a lack of freshwater, water quality loss, salinization of soils, dust and salt storms, climate deterioration, various diseases, etc.

Until the 1960s, the deltas of the Amudarya and Syrdarya rivers, which received a lot of water and sediments, were among the most dynamic in the world, were notable for their high biodiversity and biological productivity, and resisted well against the deserts of Central Asia. As a result of dramatic man-induced reduction of river flow and a drop of the Aral Sea level, the deltas of the Amudarya and Syrdarya rivers, their hydrographic network, and landscapes have undergone severe degradation.

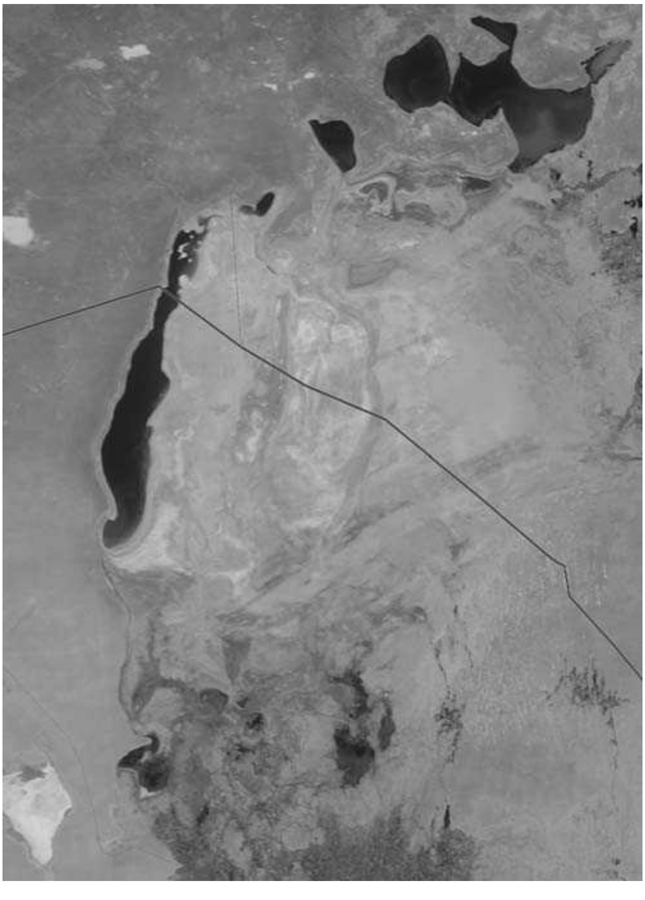


FIGURE 2 Satellite image (MODIS-Terra, 250 m resolution) of the Aral Sea on August 9, 2019 (https://wvs.earthdata.nasa.gov/?LAYERS=MODIS_Terra_CorrectedReflectance_TrueColor,Reference_Features&CRS=EPSG:4326&TIME=2019-08-09&COORDINATES=42.967532,57.612308,46.834720,61.596683&FORMAT=image/jpeg&AUTOSCALE=TRUE&RESOLUTION=250m, accessed on August 11, 2019).

The ongoing desiccation of the Aral Sea has resulted in significant changes in the distributions of dissolved gases in the residual water body. In particular, the once fully oxygenized sea developed anoxic conditions and intermittent hydrogen sulfide contamination in the bottom layers. The sulfide content depends on the density stratification and is mainly controlled by the physical regime of the sea. However, H_2S is a variable rather than a permanent feature of the present Aral Sea.[15,17]

In the first half of the 20th century in the Aral Sea, 33 species of fishes, 61 species of bottom invertebrates, 49 species of zooplankton, 306 species of phytoplankton, and 37 species of macrophytes were found. The Aral Sea accounted for 7% of the total internal waters fishery of the USSR. The main trade species were roach, sazan, and bream. From 1960 until now, the fauna of the Aral Sea evolved from mainly freshwater to hyperhalinic with the increase in water salinity. During the first decade of salinization, more than 70% of the species of fish and invertebrates vanished. By 2004, when mineralization had reached 90 ppt in the Large Aral Sea, fish fauna of the Large Aral had progressively disappeared. In 2008, zooplankton was represented by only one hypersaline species, *Artemia parthenogenetica,* introduced into the Aral Sea in 1996.[17]

It is generally accepted that the main reason for desiccation of the Aral Sea has been irrational use of Amudarya and Syrdarya waters for development of irrigation of agricultural lands and the filling of artificial water reservoirs, but regional climate change (rise of air temperature and decrease of atmospheric precipitation) also plays an important role in this process. According to estimates of the Intergovernmental Panel on Climate Change (IPCC), the trend of the mean annual air temperature in the Aral region in 1901–2005 was 1.1°C–1.7°C per 100 years, and only in 1979–2005, it was 0.3°C–0.7°C per 100 years. The results of these estimates presented in the *Fourth Assessment Report* of the IPCC have indicated that by the late 21st century, the air temperature in the Aral region depending on the emission scenarios may become 2°C–7°C higher compared to 1981–2000.[20] Our estimates of the amount of water precipitated from the atmosphere over the catchment areas of the Amudarya and Syrdarya rivers for the period 1979–2001 revealed a marked decreasing trend for the Amudarya catchment area from 7–8 to 4–5 km^3/mo on average.[21] Thus, both the effects of regional climate change significantly influenced the water balance of the Aral Sea in the past 40 years leading to its supplementary desiccation.

By the mid-1980s, the Aral crisis had been acknowledged by the whole world and became one of the most significant environment protection issues.[11–16] The Aral Sea crisis redoubled when, after disintegration of the USSR in 1990–1991, the investigations and monitoring of the sea practically ceased. Almost all hydrometeorological stations have been closed, sea expeditions have been stopped, and assessments of ongoing quick changes in the Aral Sea environment have ceased. At the same time, the interest in the Aral Sea problem has risen sharply on the international level. Many United Nations (UN) organizations (UN University, UNDP, UNESCO, UNEP, UNIDO, FAO, WMO, UN High Commissioner for Refugees, and the International Labor Organization), financial organizations (World Bank, Asian Development Bank, European Bank for Reconstruction and Development, International Monetary Fund, and Global Environment Facility), European Union Programs (TACIS, INTAS, INCO-Copernicus, OSCE, and TEMPUS), international nongovernmental organizations ("Doctors Without Borders"), regional organizations (International Fund for Saving of the Aral Sea, Interstate Coordination Water Commission, Commission on Sustainable Development, and Central Asian Economic Community), and bilateral organizations [U.S. Agency for International Development, Soros Foundation (United States), Konrad Adenauer Foundation, Friedrich Ebert Foundation, Germany Agency for Technical Cooperation (Germany), NOVIB (the Netherlands), NATO Program "Science for Peace," JAIKA, Global Infrastructure Fund Research Foundation (Japan), and others] were involved in the implementation of several hundreds of projects.[16,17]

Many of the abovementioned projects produced a lot of interesting scientific results, which traced in detail the development of the environmental crisis, but unfortunately did not result in real measures promoting salvation of the Aral Sea. Scientists, researchers, designers, and politicians could not come to a consensus on a strategy for the preservation and restoration of the Aral Sea. By the end of 2019, we can state that the main progress made towards saving of the Aral Sea occurred only in Kazakhstan with the construction of the Kokaral dam between the Small Aral Sea and Eastern Large Aral Sea in August 2005. Thus, the Small Aral Sea is now slowly reviving, while the Large Aral Sea continues to disappear progressively (Figure 2).

Kara-Bogaz-Gol Bay

Kara-Bogaz-Gol is a bay that incises deeply into the mainland and forms a highly saline bay on the eastern side of the Caspian Sea (Figure 3).[22–24] Kara-Bogaz-Gol is a Turkmen name that means "Black Trap Bay." Ancient legends say that in the bay, there was a very deep hollow that trapped the Caspian waters and even the ships that took a risk to call at this vast lagoon. This is the Caspian's largest salt-generating lagoon separated from the sea with two sand spits extending meridionally for more than 90 km (Figure 4). These sand spits form a strait 7–9 km long, 120–800 m wide, and 3–6 m deep. Due to the difference of water levels in the Caspian Sea and the bay, the waters from the sea rush at a speed of 50–100 cm/s along the strait to the bay where they evaporate completely (at a rate of 800–1000 mm/yr, on the average). Therefore, with the average annual atmospheric precipitations in this region being no more than 110 mm, Kara-Bogaz-Gol represents an enormous natural evaporation basin of seawater. Due to high evaporation, the bay is filled with brine, the salinity of which reaches 270‰–300‰ and even more. This brine is a concentrated solution of salts such as chlorides of sodium, magnesium, and potassium; magnesium sulfate; and small quantities of rare earth elements. The Kara-Bogaz-Gol Bay is the largest salt deposit where up to 20 salt minerals are found. The salt deposits of the bay accumulated a dozen billion tons of various salts that make the most valuable raw material for development of the chemical industry, agriculture, nonferrous metallurgy, medicine, and other branches of the economics.[23,24]

In the first decades of the 20th century, the level difference between the sea and the bay was about 0.5 m, and they were hydraulically linked. The level in Kara-Bogaz-Gol was close to –26.5 m relative to the world ocean. The area of the bay (in the beginning of the 20th century and after 1996) was greater than 18,300 km^2, the volume of water was 130 km^3, and the prevailing depths were 8–10 m (maximum depth was 13 m).

Starting from the 1930s, the nature of Kara-Bogaz-Gol Bay has experienced significant changes related to the sharp fall in the level of the Caspian Sea.[22] In 1939, the supply of seawater to the bay decreased to 6 km^3/yr (from 20 to 25 km^3/yr in 1929–1930) and the water balance was distorted. The maximum depth of the bay decreased from 12 m in the 1920s–1930s to 4.5 m in the early 1950s. The decrease in the water volume of Kara-Bogaz-Gol Bay, accompanied by the long-term sea-level fall and related growth in the salt concentration in the brine and changes in its chemical composition, negatively affected the conditions of sulfate mining.

In the 1960s, the difference between the levels of the sea and the bay became larger and the world's only sea waterfall about 3.7 m high (1970) was formed in the strait mouth. The maximum flow velocity in this waterfall was more than 2.5 m/s. In the 1960s to early 1970s with the water flow being 9–10 km^3, the bay area varied around 10,000–11,500 km^2, while the volume of water was 25–29 km^3 and the maximum depth of the bay decreased to 3.5 m. In the late 1970s, the Caspian water flux into the bay reduced to 5–7 km^3/yr, the water level in the lagoon dropped to –32 m, the area decreased to 10,000 km^2, and the volume of water decreased to 20–22 km^3, while the brine concentration increased to 270‰–300‰.[24] In order to reduce the Caspian Sea water use and retard the fall of the water level in the Caspian Sea that in 1977 reached its absolute minimum for the last 400 years (–29 m), in March 1980, the strait was dammed, thus stopping the inflow of seawater into the bay.

Construction of a dam led in late 1983 to complete desiccation of the bay and considerable reduction of chemical products manufactured here (sodium sulfate, etc.). By the end of 1983, the area of the bay was only 1000 km^2, its volume was 0.2 km^3, and its depth was 0.1–0.3 m. Brine salinity reached 330‰–380‰. By late August 1984, no common brine surface remained. In September 1984, tubes were installed in the dam body to direct the seawater into the bay in the amount of 1.6–1.8 km^3/yr. Such restricted supply of seawater lasted for nearly a decade and did not result in noticeable improvement of the hydrological

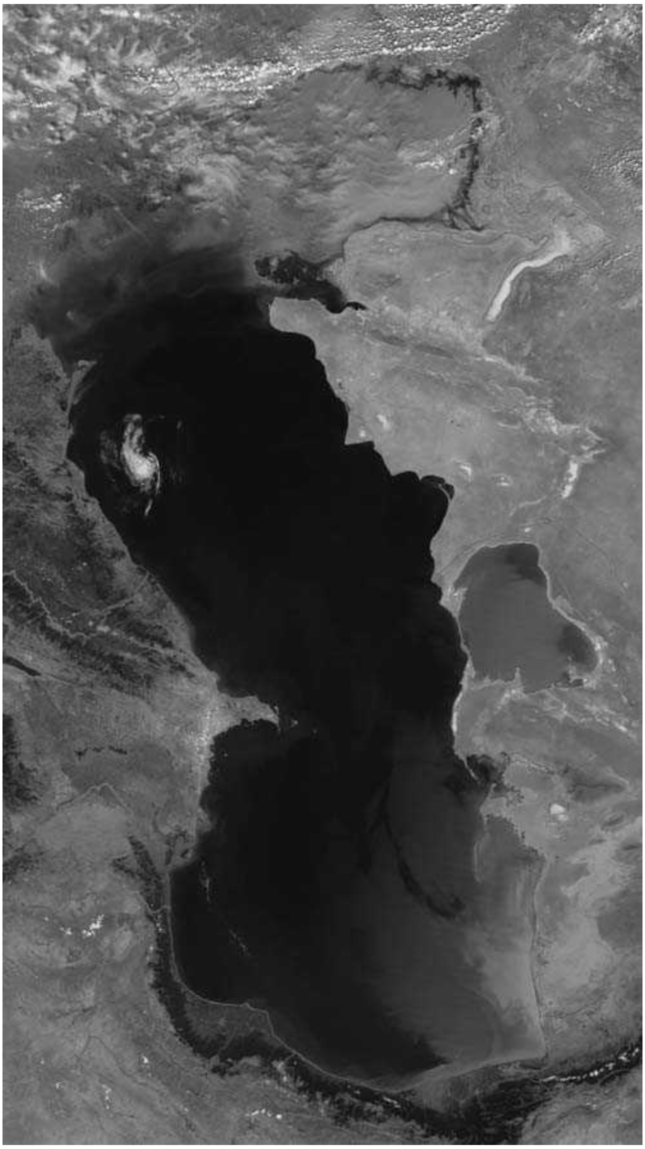


FIGURE 3 Satellite image of the Caspian Sea and Kara-Bogaz-Gol Bay obtained from MODIS-Terra on July 30, 2019 (MODIS Land Rapid Response Team, NASA GSFC, https://wvs.earthdata.nasa.gov/api/v1/snapshot?REQUEST=GetSnapshot&LAYERS=MODIS_Terra_CorrectedReflectance_TrueColor,Reference_Features&CRS=EPSG:4326&TIME=2019-07-30&BBOX=36.027841,46.596679,47.629403,55.327149&FORMAT=image/jpeg&WIDTH=1987&HEIGHT=2640&AUTOSCALE=TRUE&ts=1565516768981, accessed on August 11, 2019).

and hydrochemical situation in the bay. In June 1992, the Turkmenistan president ordered to dismantle the dam, and by 1996, the bay was filled with Caspian water until its equilibrium state (Figure 4). At the high water level in the Caspian Sea in 1993–1995, the annual water flow reached 37–52 km^3/yr and the difference between levels in the sea and bay decreased from 6.9 m in June 1992 to 0.2–0.6 m in 1996.[24] The situation became stable, the behavior of the bay returned back to its natural conditions related to the Caspian Sea, and the salt production increased.

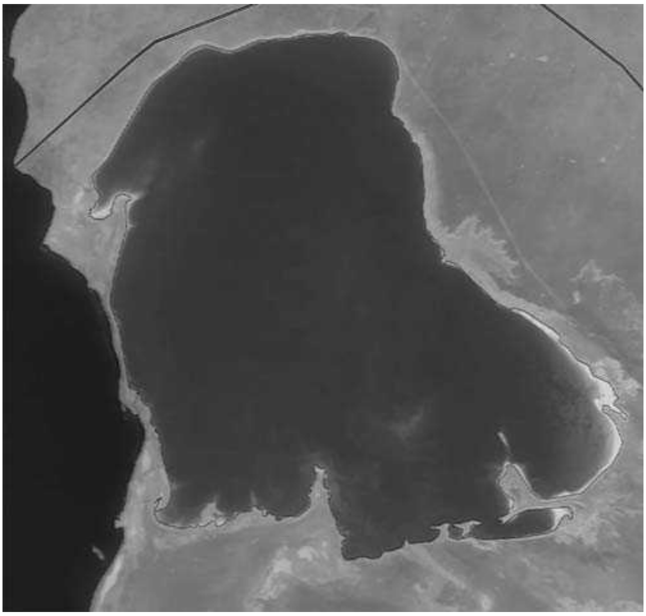


FIGURE 4 Satellite image of Kara-Bogaz-Gol Bay obtained from MODIS-Terra on August 5, 2019 (MODIS Rapid Response System, NASA GSFC, https://wvs.earthdata.nasa.gov/api/v1/snapshot?REQUEST=GetSnapshot&LAYERS=MODIS_Terra_CorrectedReflectance_TrueColor,Reference_Features&CRS=EPSG:4326&TIME=2019-08-05&BBOX=40.465395,52.395631,42.230532,54.867555&FORMAT=image/jpeg&WIDTH=1125&HEIGHT=803&AUTOSCALE=TRUE&ts=1565517161424, accessed on August 11, 2019).

Conclusions

The history of Kara-Bogaz-Gol Bay has been instructive and showed that without multidisciplinary and comprehensive research of the ecological, economic, and social aftereffects, one should not alter natural equilibriums reached over thousands of years, and human intervention into the complicated environmental processes may lead to catastrophic results.

Within a life span of only two generations, the Aral Sea as a single natural water body practically ceased to exist, and the main reason for this should be sought in man's economic activities. Since 1960, riverine water resources have been irrationally used for increasing irrigation of agricultural lands and creation of artificial water reservoirs. From the other side, regional climate change significantly influenced the water balance of the Aral Sea in the past 30 years leading to its supplementary desiccation. Thus, both effects have led the Aral Sea to one of the "largest ecological disasters of the 20th century."

Desiccation of the Aral Sea in the so-called anthropogenic period (since 1961) led not only to considerable changes in its morphometric, physical, chemical, biological, and other parameters but also to the disappearance of infrastructure in the coastal zone, including meteorological and sea-level gauge stations, ships, ports, villages, and roads. The current lack of reliable in situ measurements and time series for sea surface temperature, sea level, ice cover, and morphometric characteristics since the mid-1980s may be successfully replaced by using corresponding satellite information available directly from satellites or through the world databases.[19,21,25,26] Images from the Advanced Very High-Resolution Radiometers (AVHRR onboard National Oceanic and Atmospheric Administration (NOAA) satellites) and Moderate Resolution Imaging Spectroradiometers (MODIS onboard Terra and Aqua satellites) provide a possibility to follow the changes in the sea's coastline and observe interesting phenomena in the water, in the atmosphere, and on the dried parts of the Aral Sea.[19]

Regular measurements of the sea/lake level are practically absent in many regions such as Central Asia and Africa. Monitoring of the evolution of these and other water bodies worldwide may be accomplished by satellite altimetry from the TOPEX/Poseidon, Jason-1/-2/-3, ERS-1, ERS-2, GFO, and Envisat satellites.[26–31]

In 1986, the International Lake Environment Committee and UNEP started a project called "Survey of the State of World Lakes," aimed at collecting and analyzing environmental data on >500 lakes from 73 countries.[32,33] The results have indicated that environmental problems, common for the lakes in all continents, may be classified in the following categories:[32]

1. Lake shallowing and salinization owing to overuse of water from lakes and/or tributary rivers, resulting in a degradation of water quality and lake ecosystems.
2. Accelerated sedimentation in lakes and reservoirs resulting from anthropogenic or natural soil erosion.
3. Lake water acidification resulting from acid precipitation, which may result in the extinction of ecosystems and contamination of water with toxic agricultural and/or industrial chemicals.
4. Eutrophication owing to inflow of nitrogen and phosphorus compounds or other nutrients in the discharged water or wastewater inflows, strongly affecting biodiversity.
5. In extreme cases, a complete collapse of aquatic ecosystems and desiccation of lakes.

Increased freshwater consumption for agricultural, industrial, and urban uses and uncontrolled irrigation pose a serious threat to inland seas, lakes, rivers, and wetlands as well. In addition, shallowing and desiccation of inland seas often lead to nonhydrologic consequences, such as air pollution from dust/salt storms caused by wind erosion of exposed lake beds.

Global warming is expected to lead to large air and water temperature anomalies and make droughts and desertification more severe in the future. Climate change is also the reason for the water level drop. Shallowing or desiccation of lakes has caused local climate changes in numerous drainage basins, and if the present climatic trends persist, global climate changes could trigger almost untenable environmental effects for people and aquatic ecosystems. One of the principal tasks for future research is the delimitation of the anthropogenic and natural climate change impacts on lakes and inland seas. Degradation of many inland water bodies is a continuing global environmental problem, and its social and economic implications have attracted the growing attention of many individuals, organizations, and governments, resulting in many national and international research projects.

Acknowledgments

This research was partially supported by the Russian Science Foundation Project N 19-77-20060 "Assessing ecological variability of the Caspian Sea in the current century using satellite remote sensing data" (2019–2022).

References

1. Water, our thirsty world. *Natl. Geogr.* **2010**, *217* (special issue), N4.
2. Global Climate Change. 2018 fourth warmest year in continued warming trend, according to NASA, NOAA. Available at https://climate.nasa.gov/news/2841/2018-fourth-warmest-year-in-continued-warming-trend-according-to-nasa-noaa/ (accessed 11 August 2019).
3. Kostianoy, A.G.; Zavialov, P.O.; Lebedev, S.A. What do we know about dead, dying and endangered lakes and seas? In Dying and Dead Seas. Climatic versus Anthropic Causes; Nihoul, J.C.J., Zavialov, P.O., Micklin, Ph.P., Eds.; NATO ARW/ASI Series, Kluwer Acad. Publ.: Dordrecht, 2004; 1–48.
4. Nihoul, J.C.J.; Zavialov, P.O.; Micklin, Ph.P., Eds. *Dying and Dead Seas. Climatic versus Anthropic Causes*; NATO ARW/ASI Series, Kluwer Acad. Publ.: Dordrecht, 2004.

5. Kostianoy, A.G. Dead and dying seas. In Encyclopedia of Water Science; Taylor & Francis: London, 2007, doi:10.1081/E-EWS-120042068.
6. Lerman, A., Ed. Saline lakes and global change. *Aquat. Geochem.* **2009**, *15* (special issue), 1–5.
7. Gertman, I.; Hecht, A. The Dead Sea hydrography from 1992 to 2000. *J. Mar. Syst.* **2002**, *35*, 169–181.
8. Gill, T.E.; Gillette, D.A. Owens Lake: A natural laboratory for aridification, playa desiccation and desert dust. *Geol. Soc. Am. Abstr. Programs* **1991**, *23* (5), 462.
9. NASA Earth Observatory. April 2010 Water Level in Lake Powell. Available at http://earthobservatory.nasa.gov/IOTD/view.php?id=43933 (accessed October 2010).
10. NASA Earth Observatory. Water Level Changes in Lake Mead. Available at http://earthobservatory.nasa.gov/IOTD/view.php?id=45945 (accessed October 2010).
11. Letolle, R.; Mainguet, M. *Aral*; Springer Verlag: Paris, 1993.
12. Glantz, M.H., Ed. *Creeping Environmental Problems and Sustainable Development in the Aral Sea Basin*; Cambridge University Press: Cambridge, 1995.
13. Nihoul, J.C.J.; Kosarev, A.N.; Kostianoy, A.G.; Zonn, I.S., Eds. *The Aral Sea: Selected Bibliography*; Noosphere: Moscow, 2002.
14. Kostianoy, A.G.; Wiseman, W., Eds. The dying Aral Sea. *J. Mar. Syst.* **2004**, *47* (1–4) (special issue), 1–152.
15. Zavialov, P.O. *Physical Oceanography of the Dying Aral Sea*; Springer and Praxis Publishing: Chichester, 2005.
16. Zonn, I.S.; Glantz, M.; Kostianoy, A.G.; Kosarev, A.N. *The Aral Sea Encyclopedia*; Springer-Verlag: New York, 2009.
17. Kostianoy, A.G.; Kosarev, A.N., Eds. *The Aral Sea Environment*; The Handbook of Environmental Chemistry; Springer-Verlag: New York, 2010.
18. Kosarev, A.N.; Kostianoy, A.G. The Aral Sea under natural conditions (till 1960). In The Aral Sea Environment; Kostianoy, A.G., Kosarev, A.N., Eds.; The Handbook of Environmental Chemistry; Springer-Verlag: New York, 2010; Vol. 7, 45–63.
19. Ginzburg, A.I.; Kostianoy, A.G.; Sheremet, N.A.; Kravtsova, V.I. Satellite monitoring of the Aral Sea region. In The Aral Sea Environment; Kostianoy, A.G., Kosarev, A.N., Eds.; The Handbook of Environmental Chemistry; Springer-Verlag: New York, 2010; Vol. 7, 147–179.
20. IPCC. Climate change 2007: The physical science basis. In Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change; Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K.B., Tignor, M., Miller, H.L., Eds.; Cambridge University Press: Cambridge, 2007.
21. Nezlin, N.P.; Kostianoy, A.G.; Lebedev, S.A. Interannual variability of the discharge of Amu Darya and Syr Darya estimated from global atmospheric precipitation. *J. Mar. Syst.* 2004, 47 (1–4), 67–75.
22. Kostianoy, A.G.; Kosarev, A.N., Eds. *The Caspian Sea Environment*; The Handbook of Environmental Chemistry; Springer-Verlag: New York, 2005.
23. Kosarev, A.N.; Kostianoy, A.G. Kara-Bogaz-Gol Bay. In The Caspian Sea Environment; Kostianoy, A.G., Kosarev, A.N., Eds.; Springer-Verlag: New York, 2005; 211–221.
24. Kosarev, A.N.; Kostianoy, A.G.; Zonn, I.S. Kara-Bogaz-Gol Bay: Physical and chemical evolution. *Aquat. Geochem*. 2009, 15 (1–2), Special Issue: Saline Lakes and Global Change, 223–236, doi:10.1007/s10498-008-9054-z.
25. Nezlin, N.P.; Kostianoy, A.G.; Li, B.-L. Interannual variability and interaction of remote-sensed vegetation and atmospheric precipitation in the Aral Sea Region. *J. Arid Environ.* **2005**, *62* (4), 677–700.
26. Kouraev, A.V.; Kostianoy, A.G.; Lebedev, S.A. Recent changes of sea level and ice cover in the Aral Sea derived from satellite data (1992–2006). *J. Mar. Syst.* **2009**, *76* (3), 272–286. doi:10.1016/j.jmarsys.2008.03.016.
27. Vignudelli, S.; Kostianoy, A.G.; Cipollini, P.; Benveniste, J., Eds. *Coastal Altimetry*; Springer-Verlag: Berlin, 2011.

28. Birkett, C.; Reynolds, C.; Beckley, B.; Doorn, B. From research to operations: The USDA Global Reservoir and Lake Monitor. In Coastal Altimetry; Vignudelli, S., Kostianoy, A.G., Cipollini, P., Benveniste, J., Eds.; Springer-Verlag: Berlin, 2011; 19–50.
29. Crétaux, J.-F.; Calmant, S.; Abarca del Rio, R.; Kouraev, A.; Bergé-Nguyen, M.; Maisongrande, P. Lakes studies from satellite altimetry. In Coastal Altimetry; Vignudelli, S., Kostianoy, A.G., Cipollini, P., Benveniste, J., Eds.; Springer-Verlag: Berlin, 2011; 509–534.
30. Hydroweb. LEGOS. Available at http://ctoh.legos.obs-mip.fr/data/hydroweb/ (accessed August 11, 2019).
31. USDA. The USDA Global Reservoir and Lake Monitor. https://ipad.fas.usda.gov/cropexplorer/global_reservoir/ (accessed August 11, 2019).
32. Kira, T. Survey of the state of world lakes. In Guidelines of Lake Management. Vol.8: The World's Lakes in Crisis; Jorgensen, S.E., Matsui, S., Eds.; ILEC Foundation and UNEP: Kusatsu, 1997; 147–155.
33. World Lake Database. Available at http://wldb.ilec.or.jp/ (accessed August 11, 2019).

34

Oil Pollution: The Baltic Sea

Andrey G.
Kostianoy

Introduction

Detection of oil pollution is among the most important goals of monitoring a coastal zone. Public interest in the problem of oil pollution arises mainly during dramatic tanker and oil platform catastrophes such as those that involved the following: *Amoco Cadiz* (France, 1978), *Ixtoc I* (Gulf of Mexico, 1979–1980), *Exxon Valdez* (Alaska, 1989), *The Sea Empress* (Wales, 1996), *Erica* (France, 1999), *Prestige* (Spain, 2002), and *Deepwater Horizon* (Gulf of Mexico, 2010). However, tanker and oil platform catastrophes are only one among the many causes of oil pollution. Oil and oil product spillages at sea take place all the time, and it would be a delusion to consider tanker accidents as the main environmental danger. According to the International Tanker Owners Pollution Federation (ITOPF), over the period of 1970–2017, spillages resulting from collisions, groundings, tanker holes, and fires amounted to 52% of total leakages during tanker loading/unloading and bunkering operations.[1] In the category of 7–700 tons, some 38% of spills occurred during routine operations, most especially loading or discharging (31%). Accidents were the main cause of large spills (>700 tons), with groundings and collisions accounting for 61% of the total during the period 1970–2017.[1] Other significant causes included hull failures and fire/explosion. Discharge of wastewater containing oil products is another important source, by pollutant volume comparable to offshore oil extraction and damaged underwater pipelines. The greatest but hardest-to-estimate oil inputs come from domestic and industrial discharges, direct or via rivers, and natural hydrocarbon seeps. The long-term effects of this chronic pollution are arguably more harmful to the coastal environment than a single, large-scale accident.

Each year, ships and industries damage delicate coastal ecosystems in many parts of the world by releasing oil or pollutants into ocean, coastal waters, and rivers. Offshore environments are polluted by mineral oil mainly due to tanker accidents, illegal oil discharges by ships, and natural oil seepage. Shipping activities in European coastal seas, including oil transport and oil handled in harbors, have a number of negative impacts on the marine environment and coastal zone. Oil discharges from ships represent a significant threat to marine ecosystems. Oil spills cause the contamination of seawater, sediments, shores, and beaches, which may persist for several months and even years and represent a threat to marine resources.[1,2]

As highlighted by Oceana in its report *The Other Side of Oil Slicks*, chronic hydrocarbon contamination from washing out tanks and dumping bilge water and other oily waste represents a danger at least three times higher than that posed by the oil slicks resulting from oil tanker accidents.[3,4] For example, in the North Sea, the volume of illegal hydrocarbon dumping is estimated at 15,000–60,000 tons/year, added to which are another 10,000–20,000 tons of authorized dumping. Oil and gas platforms account for 75% of the oil pollution in the North Sea via seepage and the intentional release of oil-based drilling muds.[5] In the Mediterranean Sea, it is estimated at 400,000–1,000,000 tons/year. Of this, about 50% comes from routine ship operations and the remaining 50% comes from land-based sources via surface runoff.[5] In the Baltic Sea, this volume is estimated at 1750–5000 tons/year.[3,4] In 2004, the Finnish Environment Institute[6] estimated the total annual number of oil spills in the Baltic Sea to be 10,000 and the total amount of oil running into the sea to be as much as 10,000 tons, which is considerably more than the amount of oil pouring into the sea in accidents. In 2014, an analysis of different sources of information on oil pollution gave annual values of 20–20,000 tons for the Baltic Sea.[7] It is impressive to note that the total amount of oil spilled annually worldwide from all sources to the ocean is estimated as 1.7–8.8 million tons (the more realistic value was about 3.2 million tons) in the 1970s, 0.47–8.3 (1.3) million tons in the 1990s, and 2.6–4.8 million tons in the 2000s, which is about 0.05%–0.1% from the world oil production (4.76 billion tons in 2011).[7]

One of the most important goals of the ecological monitoring of European seas is monitoring and detection of oil pollution.[8–11] After a tanker accident or illegal oil discharge, the biggest problem is to obtain an overall view of the phenomenon, getting a clear idea of the extent of the slick and predicting the way it will move. For natural and man-made oil spills, it is necessary to operate regular and operational monitoring. Oil pollution monitoring in the Mediterranean Sea, North Sea, and Baltic Sea is normally carried out by aircrafts or ships. This is expensive and is constrained by the limited availability of resources. Aerial surveys over large areas of the seas to check for the presence of oil are limited to the daylight hours, good weather conditions, and maritime boundaries between countries.

Satellite imagery can help greatly in identifying probable spills over very large areas and in guiding aerial surveys for precise observation of specific locations. The synthetic aperture radar (SAR) instrument, which can collect data independently of weather and light conditions, is an excellent tool to monitor and detect oil on water surfaces.[7–11] This instrument offers the most effective means of monitoring oil pollution: oil slicks appear as dark patches on SAR images because of the damping effect of the oil on the backscattered signals from the radar instrument. This type of instrument is currently on board the European Space Agency's (ESA's) *Sentinel-1A and -1B* satellites, the Canadian Space Agency's *RADARSAT-2* satellite, the German Earth observation satellite *TerraSAR-X*, and other spacecraft. ASAR/SAR (advanced synthetic aperture radar) instrument is used for mapping sea ice and oil slick monitoring, measurements of ocean surface features (currents, fronts, eddies, internal waves), ship detection, oil and gas exploration, etc. Users of remotely sensed data for oil spill applications include the Coast Guard, national environmental protection agencies and departments, oil companies, shipping, insurance and fishing industries, national departments of fisheries and oceans, and other organizations.

In this entry, we will focus on oil pollution in the Baltic Sea, briefly describing main sources of oil pollution, the number of oil spills observed yearly, the results of operational satellite monitoring and numerical modeling of oil spills, and the remaining problems.

Main Sources of Oil Pollution in the Baltic Sea

Crude oil and petroleum products account for about 40% of the total exports of Russia. Russian Federation stands as one of the leading operators in the international oil business, being the largest oil exporter after Saudi Arabia. In 2000, Russia exported approximately 145 million tons of crude oil and

50 million tons of petroleum products. Since 2000, exports of petroleum and petroleum products began to grow and virtually doubled for the period from 1996 to 2005. According to the Federal Customs Service of Russia, in 2008, Russia exported 221.6 million tons, and in 2017, 253 million tons of oil.

The ports on the Baltic Sea play a huge role in the export of oil from Russia. The main oil terminals here used to be the Latvian port of Ventspils and the Port of Tallinn, Estonia. In the last 20 years, a number of new oil terminals have been built in the Baltic Sea area, resulting in increased transport of oil by ships and, consequently, an increased risk of accidents and increased risk of pollution of the marine environment. Today, in the Gulf of Finland, there are more than 20 oil terminals in Russia, Finland, and Estonia. The following are the major existing and projected oil terminals in the Gulf of Finland from Russia: Primorsk, Vysotsk, Big Port of St. Petersburg, Ust-Luga, Batareinaya, Vistino, Gorki, and Lomonosov.[12,13]

Primorsk is the largest Baltic oil terminal located on Russian territory. In 2008, 75.6 million tons of oil products were exported from Primorsk, 13.6 million tons from Vysotsk, and 14.4 million tons from the St. Petersburg oil terminal. In 2018, 53.5 million tons of oil were exported from Primorsk, maximum export possibility of the Primorsk terminal is estimated at 120 million tons, while that of Vysotsk is at 20.5 million tons. In November 2000, Lukoil has opened an oil terminal in Kaliningrad. In 2001, the company built another terminal in Kaliningrad with a declared capacity of 2.5 million tons. These terminals can overload up to 3–5 million tons of oil annually.

According to estimates of the Centre for Maritime Studies at the University of Turku (Finland), in 2007, 263 million tons of cargo were transported through the Gulf of Finland, among which the share of oil is 56%.[13] Russian ports handled 60% of goods, Finnish ports handled 23%, and Estonian ports handled 17%. The share of imports was 22%, that of export was 76%, and that of local transportation was 2%. Russian ports held 68.6% of the total turnover of petroleum products, Estonian ports held 17.2%, and Finnish ports held 14.2%.[13] The major ports were the following: Primorsk (74.2 million tons), St. Petersburg (59.5 million tons), Tallinn (35.9 million tons), Skoldvik (19.8 million tons), Vysotsk (16.5 million tons), and Helsinki (13.4 million tons). In 2007, the ports of the Gulf of Finland carried out about 53,600 ship calls, most of which were in St. Petersburg (14,651), Helsinki (11,727), and Tallinn (10,614). In 2009, vessels entered or left the Baltic Sea via Skaw 62,743 times; this is 20% more than in 2006. Approximately 21% of those ships were tankers, 46% were other cargo ships, and 4.5% were passenger ships.[14]

The growth of oil and other cargo through the terminals and the Baltic ports inevitably leads to an increase in the number of tankers and other types of vessels, which then leads to an increase in chronic sea pollution and a higher probability of ship accidents. According to statistics, shipping accounts for 45% of oil pollution in the ocean, while oil production at the shelf accounts for only 2%. In the Baltic Sea, about 2000 large ships and tankers are at sea every day; thus, shipping, including oil transport, has a major negative impact on the marine environment and coastal zone. Illegal discharges of oil and petroleum products from ships, ship accidents, collisions, and groundings represent a significant threat to the Baltic Sea.[8]

According to Global Marine Oil Pollution Information Gateway,[15] major oil spills in the region in 1970–2007 resulted from the following ship accidents: *Othello* (1970, Tralhavet Bay, Sweden, spill of 60,000 tons), *Tsesis* (1977, off Nynäshamn, Sweden, spill of 1000 tons), *Antonio Gramsci* (1979, off Ventspils, Latvia, spill of 5500 tons; another incident in 1985, off Porvoo, Finland, spill of 580 tons), *Jose Marti* (1981, off Dalarö, Sweden, spill of 1000 tons), *Globe Asimi* (1982, off Klaipeda, Lithuania, spill of 16,000 tons), *Sivona* (1984, in The Sound, Sweden, spill of 800 tons), *Volgoneft* (1990, off Karlskrona, Sweden, spill of 1000 tons), *Baltic Carrier* (2001, international waters between Denmark and Germany, spill of 2700 tons), *Fu Shan Hai* (2003, between the Danish island of Bornholm and coast of Sweden, spill of 1200 tons), *Haaga* (2003, St. Petersburg port, Russia, spill of 1300 tons), and *Golden Sky* (2007, off Ventspils, Latvia, spill of 25,000 tons).

According to HELCOM data (http://www.helcom.fi), the total number of accidents on ships in 2000–2004 is 374, of which 29 resulted in the pollution of marine waters. The number of accidents has risen since 2006, which can be linked to the 20% increase in ship traffic. Now, there are 120–140 shipping accidents yearly in the Baltic Sea area.[14] The majority of accidents are groundings and collisions. The share of groundings in the total number of accidents is higher for the Baltic Sea than for other European waters. On average, 7% of the shipping accidents in the Baltic Sea result in some kind of pollution, usually containing not more than 0.1–1 ton of oil. For the last 10 years, no major accidental oil spill has happened in the Baltic Sea.[14]

As far as oil exploitation at sea and on the coast is concerned, offshore operations have been taking place for some years in Polish waters (two jack-up rigs); Germany operated two platforms very close to the coast; in March 2004, Russia started oil production at Lukoil D-6 platform in the waters between the Kaliningrad area (Russian Federation) and Lithuania, and Latvia plans to drill for oil in the waters between them and Lithuania.[15]

Oil Spill Observations, Statistics, and Tendencies

Every ship entering the Baltic Sea must comply with the antipollution regulations of the Helsinki Convention and MARPOL (marine pollution) Convention. Even though strict controls over ships' discharges have been established by the Baltic Sea countries, illegal spills and discharges continue to happen.[14] Fortunately, the number of illegal oil spills detected by aerial surveillance has been reduced significantly over the last 30 years, from 763 spills in 1989 to 52 spills in 2017 (Table 1 and Figure 1). Also, the volume of the spills has been decreasing—most are between 1 and 0.1 m^3 today.[14,16]

A decreasing trend in the number of observed illegal oil discharges despite rapidly growing density of shipping, increased frequency of surveillance flights (Figure 1 and Table 2), and usage of satellite imagery, provided by the CleanSeaNet satellite service of the European Maritime Safety Agency (EMSA) (http://cleanseanet.emsa.europa.eu/), illustrates the positive results of the complex set of measures known as the Baltic Strategy, implemented by the Contracting Parties to the Helsinki Convention.[16]

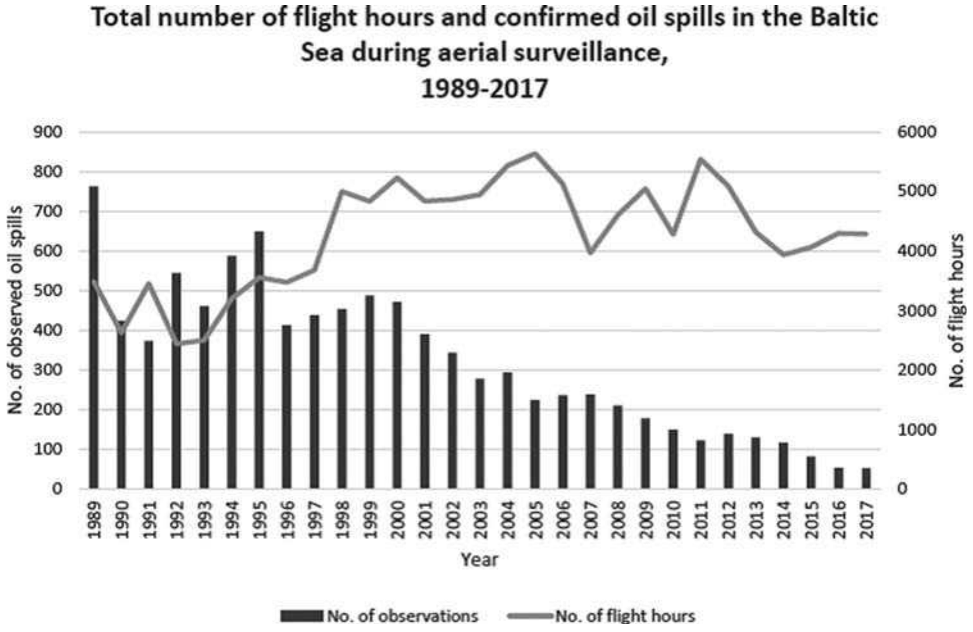


FIGURE 1 Total number of flight hours and observed oil spills in the HELCOM area during aerial surveillance in 1989–2017. (Annual Report on discharges observed during aerial surveillance in the Baltic Sea in 2017.[16])

TABLE 1 Country-Wise Data on the Number of Illegal Oil Discharges Observed in National Waters in the Baltic Sea in 1988–2017

	Denmark	Estonia	Finland	Germany	Lithuania	Latvia	Poland	Russia	Sweden	Total
1988	129			90			40	82	168	509
1989	159			139			69	184	212	763
1990	34			45		73	88		184	424
1991	46			85	8	20	14	3	197	373
1992	18	18		76	34	15	92	13	278	544
1993	17	7		43	28	6	110		250	461
1994	30	4		75			104		375	588
1995	48	3	26	55			72		445	649
1996	36		42	44			50		241	413
1997	38	3	104	34			25		234	438
1998	53	10	53	23		33	33		249	454
1999	87	33	63	72		18	18		197	488
2000	68	38	89	51		17	51		158	472
2001	93	11	107	51	0	6	24		98	390
2002	54	8	75	44		21	25		117	344
2003	37	4	40	60		14	39		84	278
2004	30	19	36	42	0	13	10		143	293
2005	28	24	32	34	0	5	5	2	94	224
2006	41	31	29	22	0	0	3		110	236
2007	43	58	29	30		2	15		61	238
2008	41	46	28	24		5	22		44	210
2009	34	20	16	15		1	27		65	178
2010	33	25	15	22	1	0	14	0	39	149
2011	18	14	16	13	0	0	5	0	56	122
2012	19	8	24	25	0	0	5	0	58	139
2013	14	8	9	7	1	0	27	0	64	130
2014	25	9	11	16	0	0	10		46	117
2015	3	12	17	12	0	0	9		29	82
2016	3	5	17	4	0	0	7		17	53
2017	4	2	8	7	0	0	9		22	52

Source: Annual Report on discharges observed during aerial surveillance in the Baltic Sea in 2017.[16]

In order to obtain the geographical distribution of oil spills, we put the spills observed in 1988–2002 on the same map (Figure 2). Analysis of Tables 1 and 2 and Figure 2 shows the following:

1. HELCOM data seem to be underestimated in figures in comparison with other estimates.[3,4,6,7]
2. Since 1993, Russia does not carry out aerial surveillance in the Southeastern Baltic Sea and in the Gulf of Finland.
3. Since 1994, Lithuania seems to have no regular and effective aerial surveillance, because no oil spills have been detected.
4. Since 2005, Latvia seems to have no effective aerial surveillance, because 0–5 oil spills yearly are not realistic figures.
5. Traces of oil spills in Figure 2 show the main ship routes in the Baltic Sea, as well as the approaches to major sea ports and oil terminals.
6. Figure 2 proves that ships are primarily responsible for the oil pollution in the Baltic Sea.

TABLE 2 Number of Aerial Surveillance Flight Hours Performed by the HELCOM Countries in 1989–2017

	Denmark	Estonia	Finland	Germany	Lithuania	Latvia	Poland	Russia	Sweden	Total
1989				142			131	1618	1600	3491
1990	292			168		400	164		1600	2624
1991	199			129	348	408	140	629	1600	3453
1992	172			267	78	127	62	32	1700	2438
1993	153	40		201	133	24	49		1900	2500
1994	253	420		290		18	179		2038	3198
1995	225	420	355	291		8	301		1953	3553
1996	275	305	400	313	65	8	345		1763	3474
1997	209	284	355	288		64	291		2189	3680
1998	325	236	649	206		577	465		2544	5002
1999	416	268	603	286		320	375		2565	4833
2000	497=	212	660	439	250	436	362		2374	5230
2001	463	161	567	466	300	412	187		2281	4837
2002	412	153	605	469		387	320		2518	4864
2003	510	201	615	446		414	228		2532	4946
2004	265	198	644	491	100	365	239		3231	5534
2005	251	178	625	549	54	384	141		3455	5638
2006	290	471	517	504	64	311	131		2842	5128
2007	271	410	529	598	41	343	380		1397	3969
2008	246	503	438	650		298	406		2063	4603
2009	240	371	351	638	66	61	561		2758	5046
2010	156	266	605	558		48	421	10	2215	4279
2011	188	315	645	648	3	18	499		3225	5541
2012	227	220	631	769		4	318		2921	5090
2013	207	327	625	470	0	19	387		2283	4317
2014	239	362	505	596	5	12	393		1823	3935
2015	254	356	490	700	0	8	259		1995	4062
2016	271	267	484	726	0	0	290		2256	4295
2017	324	268	428	891	0	0	321		2098	4331

Source: Annual report on discharges observed during aerial surveillance in the Baltic Sea in 2017.[16]

The Baltic Sea States aerial surveillance fleet today consists of more than 20 airplanes and helicopters, most of which are equipped with up-to-date remote-sensing equipment—side-looking airborne radar, SAR, infrared (IR) and ultraviolet scanner, microwave radiometer, laser fluorosensor (lidar), and forward-looking infrared (FLIR) high-resolution camera. The Baltic Sea States have to conduct aerial surveillance for detecting oil pollution and suspected ships at a minimum of twice per week over regular traffic zones including approaches to major sea ports as well as in regions with regular offshore activities. Other regions with sporadic traffic and fishing activities should be covered once per week.[16] Also, the Coordinated Extended Pollution Control Flights, which constitutes continuous surveillance of specific areas in the Baltic Sea for 24 hours or more, should be carried out twice a year.[16] Although the number of observations of illegal oil discharges shows a decreasing trend over 30 years, it should be noted that for some areas and countries, aerial surveillance is not evenly and regularly carried out and therefore there are no reliable figures for these areas.[16]

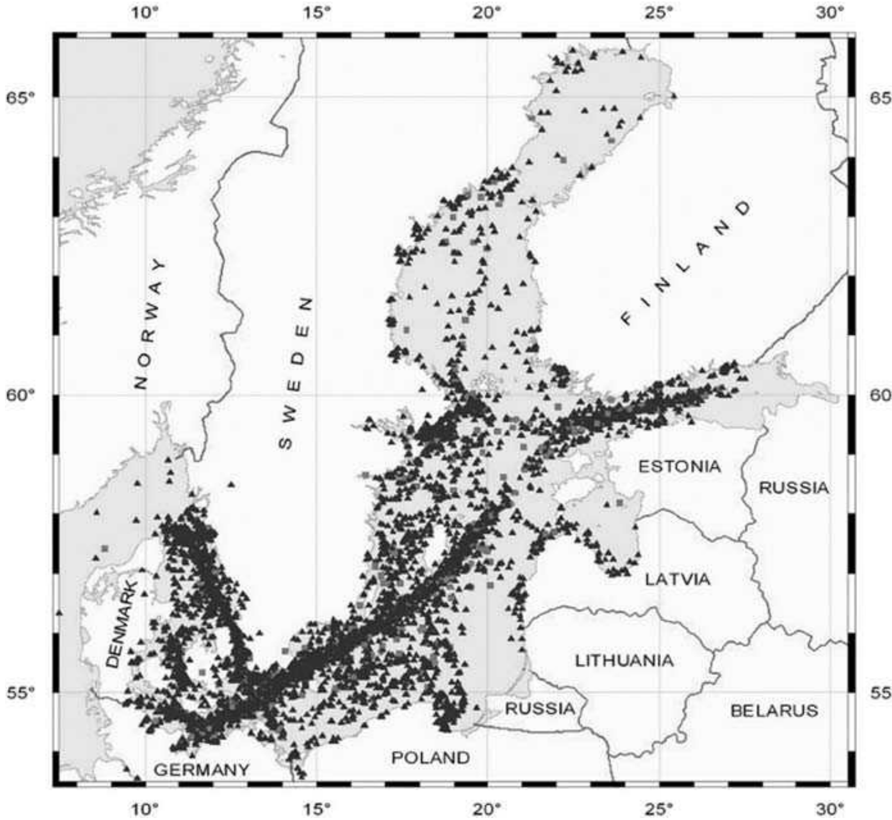


FIGURE 2 Oil spills detected in the Baltic Sea by aerial survey in 1989–2002 based on HELCOM data. (Kostianoy et al.[19])

Since 2007, aerial surveillance in the Baltic Sea was supported by the satellite remote-sensing technique for oil spill detection. CleanSeaNet is a near-real-time satellite-based oil spill and vessel-monitoring service.[8] It entered into operation on April 16, 2007.[16] The service is continually being expanded and improved and provides a range of different products to the Commission and European Union member states. The legal basis for the CleanSeaNet service is Directive 2005/35/EC on ship-source pollution and on the introduction of penalties, including criminal penalties, for pollution offenses (as amended by Directive 2009/123/EC). EMSA has been tasked to "work with the Member States in developing technical solutions and providing technical assistance in relation to the implementation of this Directive, in actions such as tracing discharges by satellite monitoring and surveillance."[16]

Operational Satellite Monitoring of Oil Pollution

A satellite-based remote-sensing system is capable of ensuring a relatively low-cost, high-standard observational system for oil pollution monitoring. SAR is the best instrument for the detection of oil slicks on the sea surface from space because slicks modify seawater viscosity and damp short waves measured by SAR. SAR images can be acquired regardless of cloud cover and light conditions. Along with 300–400 km wide swath, this is the main advantage of SAR in comparison with aerial surveillance. However, oil spill detection by SAR has a problem of distinguishing oil slicks from look-alikes, such as sea areas covered by organic films, algal bloom, sea ice, wind shadows, rain cells, and upwelling zones. Therefore, reliable automatic detection of oil spills on the basis of SAR data is not yet achieved and there is a risk of false alarms. This problem can be significantly reduced by a new approach, which consists in the combined use of all available quasi-concurrent satellite, oceanographic, and meteorological information, along with numerical modeling of oil spill transport. This operational system was specially elaborated in the beginning of 2004 for monitoring oil pollution in the vicinity of the Lukoil D-6 oil platform in the Southeastern Baltic, Russian Federation.[8,17–21]

Since 1993, regular aerial surveillance of the oil spills in the Russian sector of the Southeastern Baltic Sea and in the Gulf of Finland has stopped (Tables 1 and 2, Figure 2). In June 2004, we organized daily service for monitoring of oil spills in the Southeastern Baltic Sea based on the operational receiving and analysis of ASAR *ENVISAT* and SAR *RADARSAT-1* data as well as of other satellite IR and optical (VIS) data, meteorological information, and numerical modeling of currents required for the identification of the slick nature in the sea and forecast of the oil spill drift.[8,17–21] This work was initiated and financed by Lukoil-Kaliningradmorneft (Kaliningrad, Russia) in connection with the start of oil production from the continental shelf of Russia in March 2004. The principal differences from the existing projects and satellite services were (1) an operational monitoring regime of 24 hours/day, 7 days/week for 18 months and (2) a complex approach to the oil spill detection and forecast of their drift.

The general goals of the satellite oil pollution monitoring in the Baltic Sea were as follows:

1. Correct detection of oil spills in the vicinity of the D-6 oil platform as well as in the large area of the Southeastern Baltic Sea between 54°20′–58°N and 18°–22°E
2. Identification of sources of oil pollution
3. Forecast of the oil spill drift by different methods
4. Data systematization and archiving
5. Cooperation with authorities.

Operational monitoring of oil pollution in the sea was based on the processing and analysis of ASAR *ENVISAT* (every pass over the Southeastern Baltic Sea, frame of 400×400 km, 75 m/pixel spatial resolution) and SAR *RADARSAT-1* (300×300 km, 25 m/pixel resolution) images received from KSAT Station (Kongsberg Satellite Services, Tromsø, Norway) in operational regime (1–2 hour after the satellite's overpass). For interpretation of ASAR *ENVISAT* imagery and forecast of the oil spill drift, IR and VIS AVHRR (the Advanced Very High-Resolution Radiometer aboard the National Oceanic and Atmospheric Administration (NOAA) satellites) and Moderate Resolution Imaging Spectroradiometer (MODIS) (*Terra* and *Aqua*) images were received, processed, and analyzed, as well as the *QuikSCAT* scatterometer and the *Jason-1* altimeter data.[18–21] The total area covered by the monitoring was equal to about 60,000 km^2, which is almost one-sixth of the Baltic Sea total surface.

The satellite receiving station at the Marine Hydrophysical Institute (MHI) in Sevastopol was used for operational (24 hours/day, 7 days/week) receiving of the AVHRR NOAA data for the construction of the sea surface temperature (SST), optical characteristics of seawater, and currents maps. SST variability and intensive algae bloom (high concentration of blue–green algae on the sea surface in the summertime) allow one to highlight meso- and small-scale water dynamics in the Baltic Sea and to follow movements of currents, eddies, dipoles, jets, filaments, river plumes, and outflows from the Vistula and Curonian bays. Sequence of daily MODIS IR and VIS imagery allows reconstruction of a real field of surface currents (direction and velocity) with 0.25–1 km resolution, which is very important for a forecast of a direction and velocity of a potential pollution drift including oil spills. The combination of ASAR *ENVISAT* images with high-resolution VIS and IR MODIS images allows understanding of the observed form of the detected oil spills and prediction of their transport by currents.[18–21]

Sea wind speed fields were derived from scatterometer data from every path of the *QuikSCAT* satellite over the Baltic Sea (twice a day). These data were combined with data from coastal meteorological stations in Russia, Lithuania, Latvia, Estonia, Finland, Sweden, Denmark, Germany, and Poland and numerical weather models. Altimetry data from every track of the *Jason-1* satellite over the Baltic Sea were used for compilation of sea wave height charts, which include the results of the FNMOC (Fleet Numerical Meteorology and Oceanography Center, the United States) World War III Model. Both data were used for the analysis of the ASAR *ENVISAT* imagery and estimates of the oil spill drift direction and velocity.[18–21]

In total, 274 oil spills were detected in 230 ASAR *ENVISAT* images and 17 SAR *RADARSAT-1* images received during 18 months (June 2004 to November 2005).[8,20] One example from the oil spill gallery is shown in Figure 3 where an illegal release of oil from three ships was detected on August 25, 2005.

The interactive numerical model Seatrack Web of the Swedish Meteorological and Hydrological Institute (SMHI) was used for a forecast of the drift of satellite-detected oil spills and ecological risk assessment of the Lukoil D-6 oil platform.[18–20]

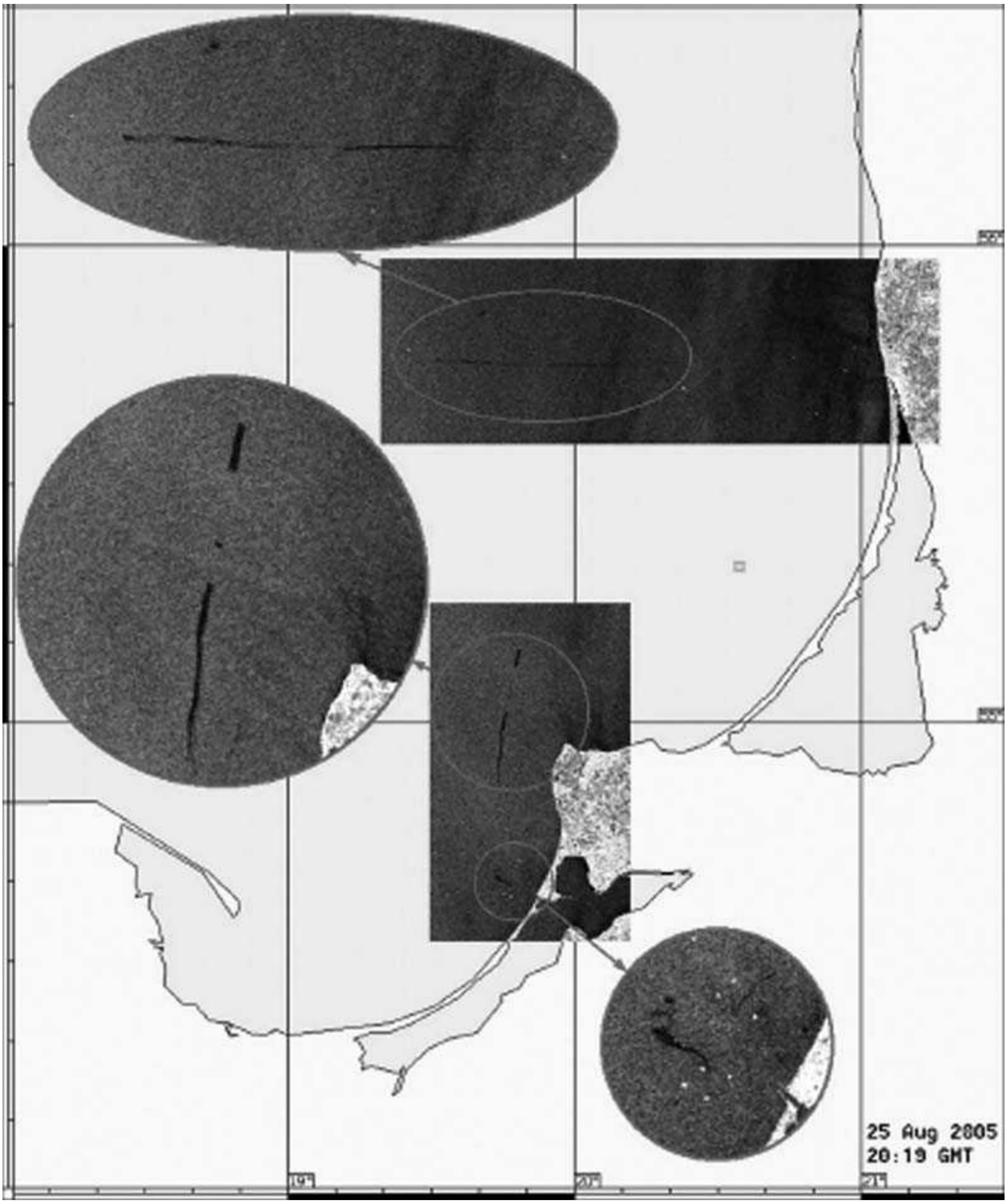


FIGURE 3 A release of oil from three ships on August 25, 2005 (ASAR *ENVISAT*). The length of the spill in front of Klaipeda is 33.6 km, surface—8.6 km^2. The length of another long spill is 22 km.

Since 2006, the satellite monitoring of the D-6 oil platform was transformed, reduced in size to about 24,000 km^2, and unfortunately, it lost its main peculiarity—a complex approach to the oil spill detection and forecast of their drift. In 2006–2009, 638 oil spills have been identified on 804 ASAR images, from which 319 spills were detected in this reduced area.[21] Combined analysis of the location and shape of the detected spills with location of the ships thanks to AIS (Automatic Identification System for ships) clearly indicates that the major source of sea pollution is shipping. In the Southeastern Baltic Sea, an area with no HELCOM statistical data, we also observe a decreasing trend in oil spill number and their total surface. In 2006, the total number of oil spills (and area of oil pollution) amounted to 114 (371.7 km^2); in 2007, 94 spills (213.7 km^2); in 2008, 67 spills (198.7 km^2); and in 2009, 44 spills (81.7 km^2).[21]

The satellite monitoring of the D-6 oil platform continues till present.[21–28] A map of all oil spills detected by the analysis of the ASAR *ENVISAT, RADARSAT-1, RADARSAT-2, Cosmo-SkyMED-1, -2, -3, -4*, and *TerraSAR-X* imagery in the given area of the Southeastern Baltic Sea from June 12, 2004, until December 31, 2015, is shown in Figure 4. A real form and dimension of oil spills are shown. A square southwestward of Klaipeda shows the location of the D-6 oil platform. Oil spills clearly revealed the main ship routes in the Baltic Sea directed to ports of Ventspils, Liepaja, Klaipeda (routes from different directions), Kaliningrad, and along Gotland Island (Figure 4). No spills originated from the D-6 oil platform were observed.

From June 12, 2004, to December 31, 2015, 1232 oil spills were detected with clear interannual tendency to reduction of oil pollution in the Southeastern Baltic Sea from 2006 to 2011 when 44 oil spills (147 km^2) were recorded. In 2012, an increase in the number of detected oil spills (86), and as a result, increase of a total area of oil pollution (470 km^2) was observed. In 2013, with a relatively small number of oil spills (52), we have observed a high total area of oil pollution (341 km^2) what is explained by huge oil

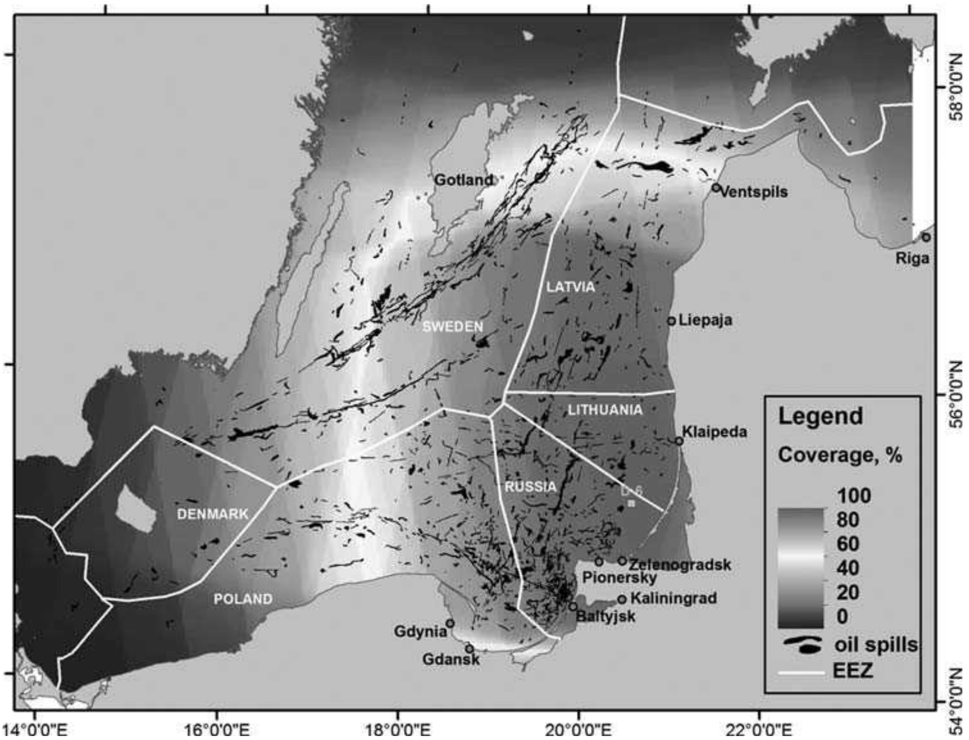


FIGURE 4 Map of all oil spills detected by the analysis of the ASAR/SAR satellite imagery from June 2004 to December 2015.

spills detected on September 13, 2013, and September 20, 2013. In 2014 and 2015, the number of oil spills has reduced to 47 and 39 with a total surface of about 188 km^2 for both years.[26]

A significant seasonal variability in oil spill detection is observed. During autumn and winter, we detect oil spills four times less than in spring and summer.[27] This huge difference is explained by limitations of the SAR method to detect oil spills when the wind is stronger than 10 m/s, which is very often during the cold season in the Baltic Sea. In addition, strong wind-wave mixing contributes to more rapid formation of emulsions ("water in oil" and "oil in water"), thus preventing formation of the oil slicks on the sea surface.

Image comparison of the number of detected oil spills in the morning (about 11:00 local time) and in the evening (about 22:00 local time) for 2006–2009 showed that the probability of finding oil pollution in the morning is about 40% higher than that in the afternoon and evening.[21] This fact indicates that the illegal discharge of oil from vessels occurs more often at night, when it is impossible for patrol aircraft or ships to record this fact by photo and video camera. This once again confirms the advantages of satellite radar imagery for monitoring of oil pollution.

Based on the number of oil spills detected in 2014 and 2015 (about 40 spills yearly on the area of about 24,000 km^2), we can estimate the total number of oil spills for the Baltic Sea (377,000 km^2) as about 630 yearly and the total surface of oil pollution as about 3000 km^2 yearly. These values are six to seven times higher than reported by HELCOM for the same years (Table 1). These values may double and even triple if we take into account the following: (1) SAR satellites pass over a specific area in the Baltic Sea every 2 days in average, (2) significant reduction of oil spill observation in autumn and winter due to unfavorable weather conditions, and (3) spatial resolution of SAR/ASAR imagery of 25–75 m/pixel.

Numerical Modeling of Oil Spill Drift

The abovementioned satellite monitoring of the Southeastern Baltic Sea was coupled with numerical modeling of oil spill transport. The interactive numerical model Seatrack Web SMHI was used for a forecast of the drift of (1) all large oil spills detected by ASAR *ENVISAT* in the Southeastern Baltic Sea and (2) virtual (simulated) oil spills from the Lukoil D-6 platform. The latter was done daily for the operational correction of the action plan for accident elimination at the D-6 and ecological risk assessment (oil pollution of the sea and the Curonian Spit).

This version of a numerical model on the Internet platform has been developed at the SMHI in close cooperation with the Danish Maritime Safety Administration, Bundesamt fur Seeschifffahrt und Hafen, and the Finnish Environment Institute.[29] The first version of Seatrack Web was introduced in 1995, and since then, Seatrack Web has been used successfully for oil spill drift forecast. It was developed to be a friendly tool for authorities responsible for oil spill response in the Baltic Sea region. Seatrack Web's main purpose is to calculate the drift and transformation of oil spills in the Gulf of Bothnia, the Gulf of Finland, the Baltic Sea Proper, the Sounds, the Kattegat, the Skagerrak, and part of the North Sea (to 3°E).[29] The program can also be used for substances other than oil, such as chemicals, algae, and floating objects. In addition to an oil drift forecast, it is possible to make a backward calculation. Then, calculation starts at the position where a substance was found and the program calculates the drift backward in time and traces the origin of the substance or an object.

Today, the Seatrack Web is hosted by SMHI and developed together by SMHI in close cooperation with the FCOO (Defence Centre for Operational Oceanography) in Denmark, the BSH (Federal Maritime and Hydrographic Agency) in Germany, and the FMI (Finnish Meteorological Institute) in Finland. Current fields are modeled with the NEMO-Nordic model (Nucleus of European Modeling of the Ocean), which is a three-dimensional circulation model covering the Gulf of Bothnia, the Gulf of Finland, the Baltic Sea, the Sounds, the Kattegat, the Skagerack, the North Sea, and the English Channel. The meteorological model used in Seatrack Web is ECMWF (the European Centre for Medium-Range Weather Forecasts), forecast for 4 days ahead and 6 days backward. The Seatrack Web provides oil spill drift with a time step of 15 min and 2 n.m. spatial resolution.

The oil spreading calculation is added to the currents, as well as oil evaporation, emulsification, sinking, stranding, and dispersion.[30]

The AIS functionality in Seatrack Web is a tool to identify the ship that causes an oil spill. The AIS function gives the ship tracks in the area where the spill was detected and backward to its probable origin. By using the ship tracks simultaneously with the oil tracks, the probability of identifying a suspected ship increases. This powerful system was recommended by HELCOM for operational use in the Baltic countries.[30]

The Seatrack Web model was very useful for ecological risk assessment related to exploitation of the Lukoil D-6 oil platform, which is installed 22.5 km from the Curonian Spit, a UNESCO World Heritage Site. Virtual (simulated) oil spills of 10 m^3 were released daily from the platform for 6 months in order to calculate the shape, direction, distance, and velocity of their drift. Then, all 180 oil tracks were accumulated on one map showing the potential impact of oil pollution in case of an accident at the platform. Statistics, based on daily forecast of the oil spill drift in July–December 2004, shows potential probability (%) of the appearance of an oil spill in any point of the area during 48 hours after an accidental release of 10 m^3 of oil (Figure 5). The probability of the oil spill drift directed to the Curonian Spit (150° sector from D-6) is equal to 67%, but only in half of these cases did oil spills reach the coast due to a coastal current.[18,20] This new technology allowed a quantitative assessment of ecological risks in the whole Baltic Sea.[31]

Later, the same methodology was applied to the risk assessment of the Nord Stream gas pipeline construction and assessment of the impact of oil pollution along the ship routes on the Baltic Sea Marine Protected Areas (BSPAs).[24,31] Figures 6 and 7 show two examples of oil spill drift modeling and a probability of the drift for specific points along main ship routes in the Gulf of Finland and southward of Gotland Island for July and August 2007. Figure 6a shows a drift of a virtual oil spill of 10 m^3 during 48 hours, which

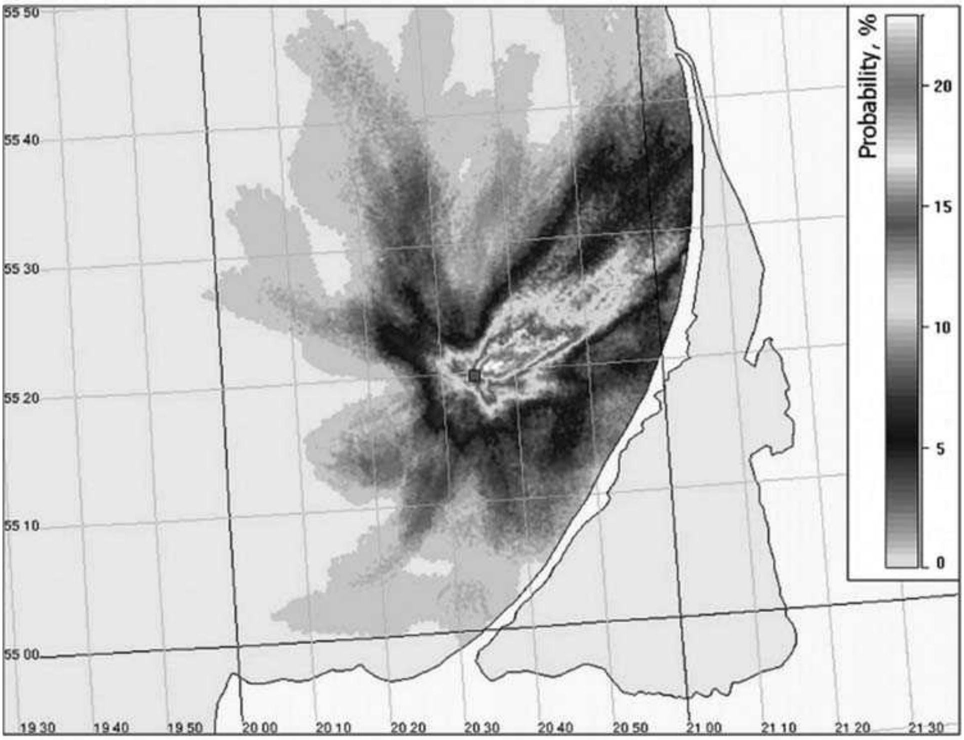


FIGURE 5 Probability of observation of potential oil pollution from the D-6 platform during the first 48 hours after an accidental release of 10 m^3 of oil (based on 6 months daily release of oil spill from the oil platform).

was released on July 23, 2007, at a specific point (red square) of the ship route passing through the Gulf of Finland. The same numerical experiment was performed daily from July 1 to August 31, 2007. Thus, based on the compilation of 62 maps of oil spill drifts, we could construct Figure 6b and calculate a probability of oil spill drift. Figure 6b shows that for this time period, there is no impact of possible oil spill drift on the surrounding BSPAs along the coasts of Finland and Estonia, which are marked by blue and rose colors. This is explained by low wind speed and weak currents observed during July and August 2007. These weather conditions differ from those observed from July 26 to August 15, 2006, when a virtual oil spill could drift 33.5 nautical miles during 2 days with a velocity up to 50 cm/s (Figure 6c). Thus, potential releases of oil spills from the ships may represent a threat to seven protected areas located along the coasts of Finland and Estonia, as well as to the islands and coasts of these Baltic countries. The same dot lies on the trajectory of the Nord Stream gas pipeline (upper line in Figure 6c), which was under construction in the Gulf of Finland since May 2010. Risk assessment for the construction of the gas pipeline was performed using the same methodology in 2006 for seven key points of the pipeline.[32]

Figure 7a shows a simultaneous release of oil from a long part of the ship route located southward of Gotland Island. Figure 7b shows a significant impact of oil pollution produced by this virtual oil release. Both BSPAs were subjected to oil pollution, but at different degrees, so that a quantitative estimation was possible with the help of the Seatrack Web model (Figure 7c and d). For instance, about 60% of the first BSPA and 95% of the second one will be polluted with the indicated probability. In this case, even the coastal zone of Gotland Island was potentially threatened by oil pollution with a clearly calculated probability.[32]

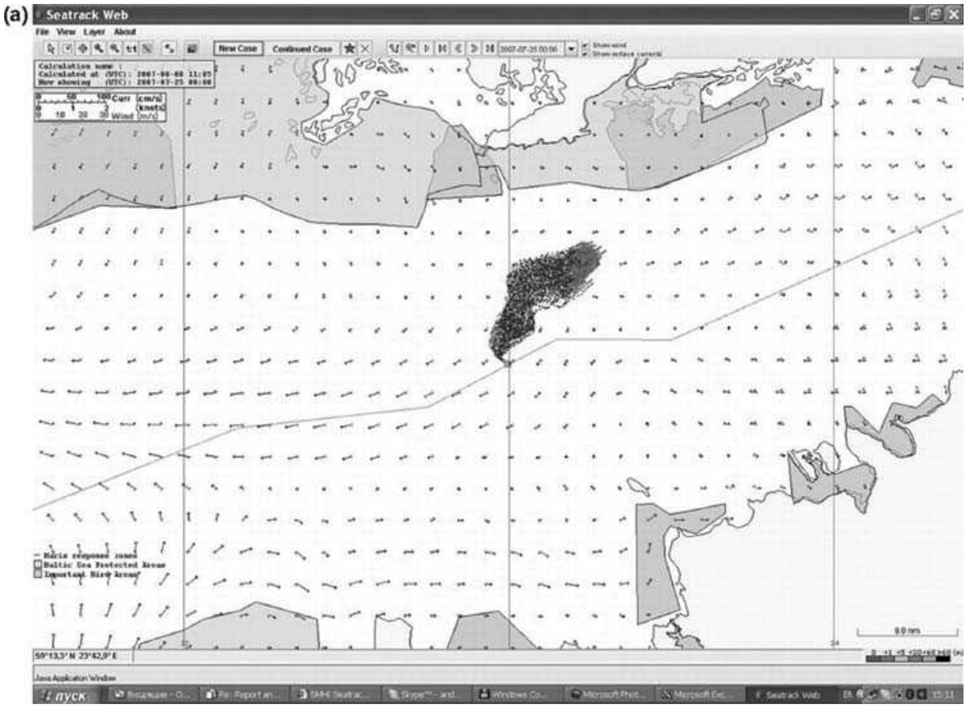


FIGURE 6 Modeling of oil spill drift in the Gulf of Finland. Panel (a) shows oil spill drift on July 23, 2007. Panel (b) shows probability (%) of oil spill drift calculated on the basis of daily modeling at this point for real wind and current conditions in July–August 2007. Panel (c) shows probability (%) of oil spill drift calculated for July 16 to August 15, 2006. BSPAs along the coastlines are shown in light grey color; important bird areas are in grey colors. The lines in panel (a) and (b) show MARIS response zones. The upper line in panel (c) shows approximate position of the Nord Stream gas pipeline, while the lower line is the MARIS response zones.

(*Continued*)

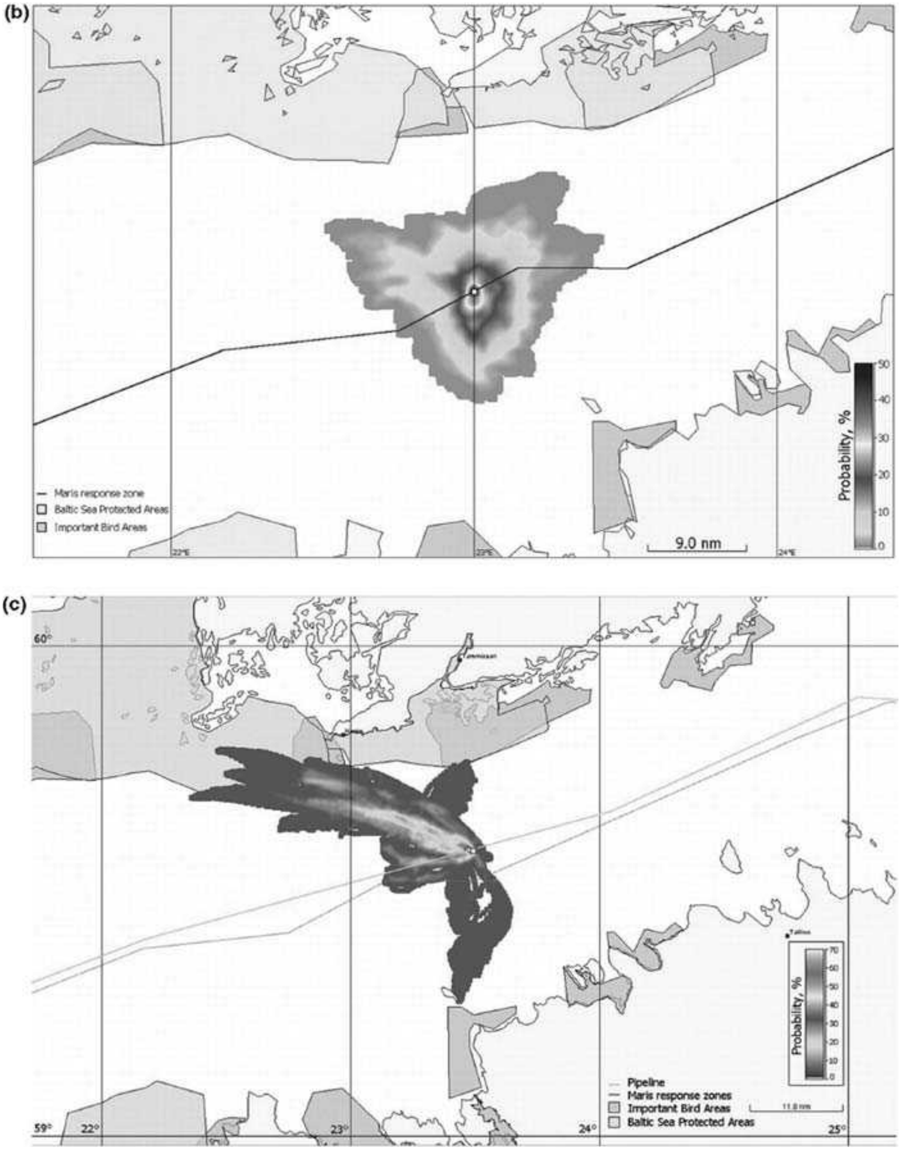


FIGURE 6 (CONTINUED) Modeling of oil spill drift in the Gulf of Finland. Panel (a) shows oil spill drift on July 23, 2007. Panel (b) shows probability (%) of oil spill drift calculated on the basis of daily modeling at this point for real wind and current conditions in July–August 2007. Panel (c) shows probability (%) of oil spill drift calculated for July 16 to August 15, 2006. BSPAs along the coastlines are shown in light grey color; important bird areas are in grey colors. The lines in panel (a) and (b) show MARIS response zones. The upper line in panel (c) shows approximate position of the Nord Stream gas pipeline, while the lower line is the MARIS response zones.

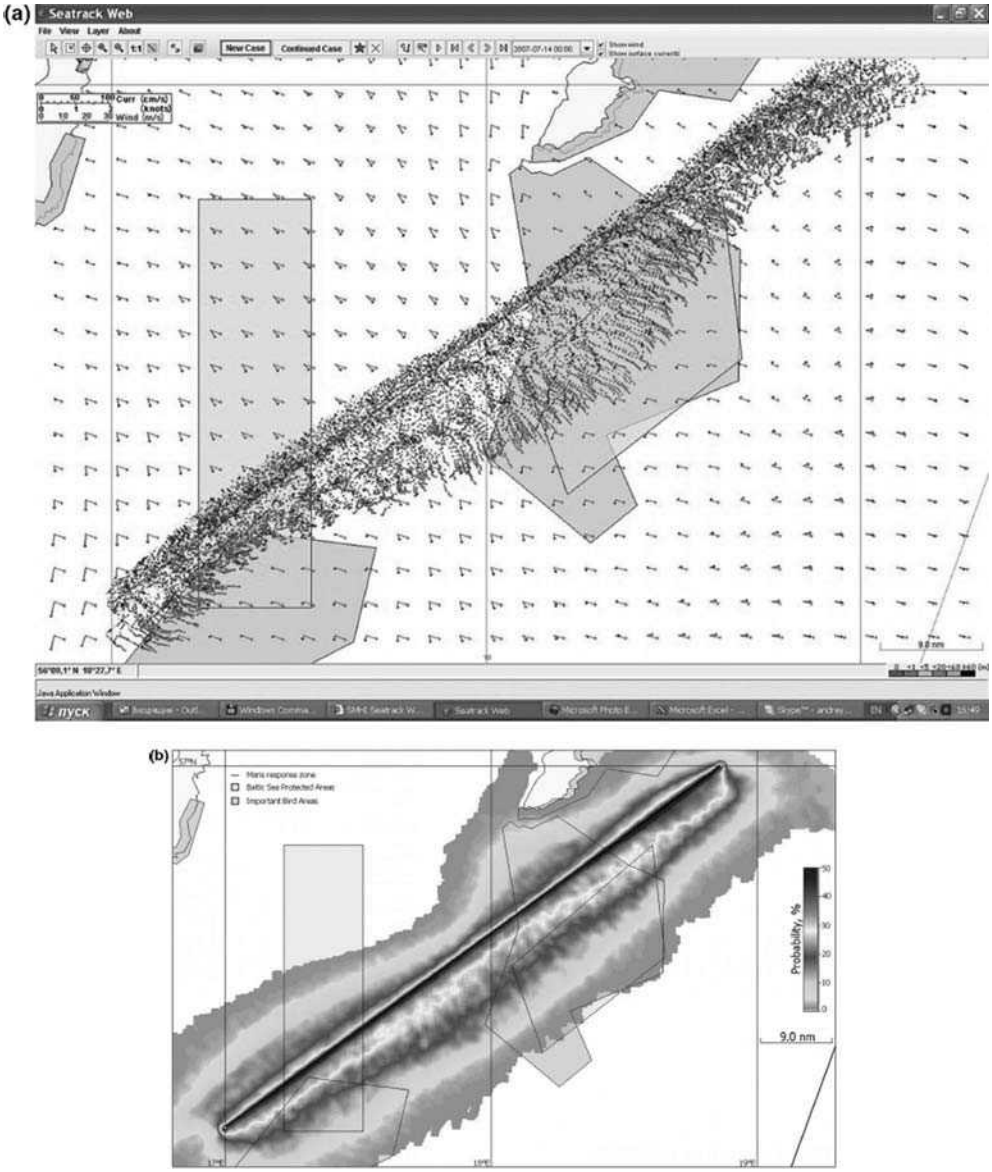


FIGURE 7 Modeling of oil spill drift released from a long part of the ship route located southward of Gotland. Panel (a) shows oil spill drift on July 12, 2007. Panel (b) shows probability (%) of oil spill drift calculated on the basis of daily modeling at this line for real wind and current conditions in July–August 2007. Panels (c) and (d) show the impact of this part of the ship route on both BSPAs.

(*Continued*)

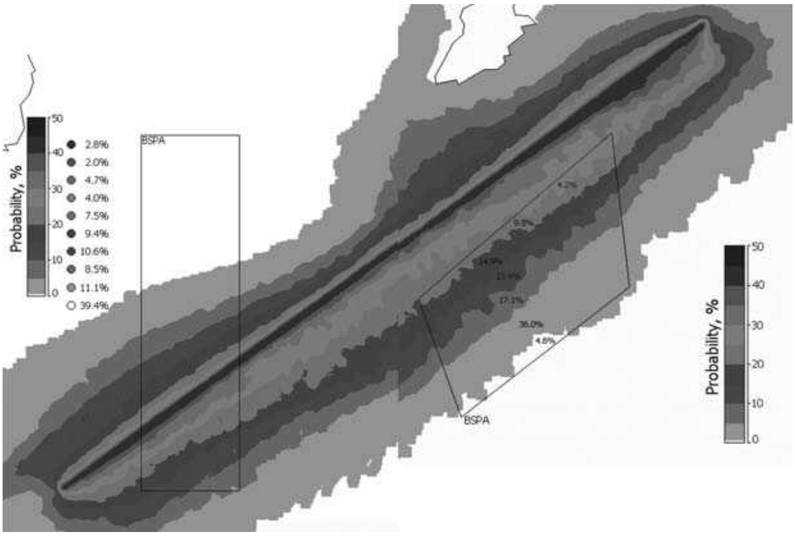


FIGURE 7 (CONTINUED) Modeling of oil spill drift released from a long part of the ship route located southward of Gotland. Panel (a) shows oil spill drift on July 12, 2007. Panel (b) shows probability (%) of oil spill drift calculated on the basis of daily modeling at this line for real wind and current conditions in July–August 2007. Panels (c) shows the impact of this part of the ship route on both BSPAs.

Problems and Solutions

The main challenge is that the real degree of oil pollution in the Baltic Sea is unknown, because the number of observed oil spills and estimates differ significantly. Partially, this is explained by known differences in aerial surveillance and satellite monitoring. Both methods have their own advantages and limitations and should complement each other. It is clear that statistics on oil spills is not complete and not comparable in different parts of the sea due to different efforts and methods applied for oil pollution monitoring (different number of oil patrol ships, aircraft, and helicopters per country and per unit of the sea area; the number of flight hours per country and unit of the sea area; use of satellites; a number of ASAR/SAR images acquired and analyzed yearly; application of complex satellite monitoring based on the multisensor and multiplatform approach along with the analysis of metocean data; local peculiarities of the water area; and numerical modeling).

For example, since 1993, Russia does not carry out aerial surveillance in the Gulf of Finland and the Southeastern Baltic Sea. The existing satellite monitoring is performed on a regular basis only in the Southeastern Baltic Sea and by a private company, Lukoil-Kaliningradmorneft. According to HELCOM data, since 1994 and 2005, respectively, Lithuania and Latvia seem to have had no regular aerial surveillance of oil pollution. Well-equipped regular aerial surveillance is very expensive, and it is clear that countries in economic recession reduce their aerial and in situ monitoring. Satellite monitoring may partially solve this problem, because satellites cover simultaneously very large areas of the Baltic Sea.

Organization of the Baltic International Satellite Monitoring Center in HELCOM could solve many problems in the operational monitoring of oil pollution in the Baltic Sea.[33] It will:

1. Ensure full and uniform coverage of the Baltic Sea area by remote-sensing control
2. Reinforce aerial surveillance and improve the oil pollution monitoring
3. Establish satellite monitoring for the countries where it is not yet applied

4. Remove duplication of satellite monitoring for the same area performed by neighboring countries
5. Significantly reduce the total cost of operational satellite monitoring for all countries
6. Provide data to all the Baltic Sea States in the same format
7. Solve the problem regarding different technologies, methods, and algorithms used for the analysis of satellite data in different countries
8. Solve the problem of the "night" oil spill pollution that is getting more and more acute
9. Stimulate exchange of data and cooperation between countries
10. Contribute to early warning in case of transboundary oil spill drift
11. Improve the ecological state of the Baltic Sea, coastal zones, and shores of the Baltic Sea States
12. Stimulate organization of analogous operational monitoring centers for the seas with a high density of shipping and/or oil/gas exploration/production industry, i.e., the North Sea, the Mediterranean Sea, the Black Sea, the Caspian Sea, the Gulf of Mexico, etc.

Wide usage of the SMHI Seatrack Web model for oil spill drift forecast is required. Originally, the model was not devoted to the ecological risk assessment, but we found it very useful for this purpose as well.[31] The ecological risk assessment for all ports, oil terminals and platforms, subsurface oil pipelines, ship routes, the Baltic Sea Protected Areas, and any part of the 8000 km long coastline of the Baltic Sea can be performed based on the methodology we elaborated in 2004 and successfully used for Lukoil D-6 oil platform and Nord Stream gas pipeline construction.[18,20,24,31,32] This will quantitatively and precisely reveal the hot spots in the marine area, islands, and coastline of the Baltic Sea that are vulnerable to the impact of the shipping oil pollution. Such a general map of the Baltic Sea with calculated probability for any point of the sea and the coastline to be polluted may serve as a guideline for the Baltic Sea States to improve their monitoring systems.

Conclusions

Oil is a major threat to the Baltic Sea ecosystems. In the last decade, maritime transportation in the Baltic Sea region has been growing steadily, reflecting the intensified trade and oil export. An increase in the number of ships also means that we can expect a larger number of illegal oil discharges. Both oil tankers and all types of ships are responsible for oil pollution of the Baltic Sea. Any discharge into the Baltic Sea of oil, or diluted mixtures containing oil in any form including crude oil, fuel oil, oil sludge, or refined products, is prohibited. This applies to oily water from the machinery spaces of any ship, as well as from ballast or cargo tanks from oil tankers. Every ship entering the Baltic Sea must comply with the antipollution regulations of the Helsinki Convention and MARPOL Convention. Even though strict controls over ships' discharges were established by the Baltic Sea countries, illegal spills and discharges still happen. The number of illegal oil spills has been reduced significantly over the last 30 years, from 763 spills in 1989 to 52 spills in 2017, and this is an evident and positive tendency, resulted from the long-term efforts of HELCOM.[16]

However, the actual total number of oil spills and their volume seem to be unknown because these values contradict significantly (10–20 times) with estimates of different organizations[3,4,6,8] and with the results of complex operational satellite monitoring performed in 2004–2017 in the Southeastern Baltic Sea.[8,21–28] Although the number of observations of illegal oil discharges shows a decreasing trend over the years, it should be kept in mind that for some areas and countries, aerial and satellite surveillance is not evenly and regularly carried out and, therefore there are no reliable figures. We have to add to the uncertainties in the oil pollution statistics considerable seasonal variability in observations of oil spills on the sea surface and predominance of the "night" discharge of oil spills from the ships used to avoid any direct visible evidence of pollution and responsibility for this fact.

So far, as the Baltic Sea ecosystem undergoes increasing human-induced impacts, especially associated with intensifying oil transport and production, further research on the link between physical, chemical, and biological parameters of the ecosystem; the complex monitoring of the Baltic Sea States;

and, especially, oil spill monitoring are of great importance. Oil spill behavior, modeling, prevention, effects, control, and cleanup techniques require supplementary information about a large number of complex physical, chemical, and biological processes and phenomena.

ASAR/SAR satellite imagery provides effective opportunities to monitor oil spills, particularly in the Baltic Sea, as well as in other European seas.[8–11] Combined with satellite remote sensing (AVHRR NOAA, MODIS-*Terra* and -*Aqua*, QuikSCAT, *Jason-1*, etc.), of the SST, sea level, chlorophyll and suspended matter concentration, meso- and small-scale dynamics, and wind and waves, this observational system represents a powerful method for long-term monitoring of the ecological state of semi-enclosed seas especially vulnerable to oil pollution. Our experience in the operational oil pollution monitoring in the Baltic Sea could be easily applied to the Caspian, Black, Mediterranean, and other European seas.

Since 2004, we have elaborated several operational satellite monitoring systems for oil and gas companies in Russia and performed complex satellite monitoring of the ecological state of coastal waters in the Baltic, Black, Caspian, and Mediterranean seas.[24,34] The accident on the BP oil platform "Deepwater Horizon" on April 20, 2010, in the Gulf of Mexico showed that the absence of such a permanent complex satellite monitoring system makes all efforts related to cleaning operations at sea and on the shore during the first weeks after the accident less effective.[35]

A large number of discharges of hydrocarbons that annually take place in European waters, the vast quantity of waste generated by the sea traffic in Europe, the lack of adequate port installations for waste management, and the toxicity of compounds thrown into the sea make solving the chronic hydrocarbon pollution problem a priority for improving the environmental quality of European seas. The growing availability of satellite and sea observation data should encourage interest, involvement, and investment into the complex operational monitoring systems from the side of the state authorities responsible for the environment, pollution control, meteorology, coastal protection, transport, fisheries, and hazard management, as well as from the side of private companies operating in the sea and coastal zone.

References

1. ITOPF (International Tanker Owners Pollution Federation Limited), Handbook 2018/2019, https://www.itopf.org/fileadmin/data/Documents/Company_Lit/ITOPF_Handbook_2018.pdf (accessed April 24, 2019).
2. European Environment Agency. *The European Environment—State and Outlook 2005*; EEA: Copenhagen, 2005.
3. Oceana. *The Other Side of Oil Slicks*; The Dumping of Hydrocarbons from Ships into the Seas and Oceans of Europe; 2003, http://www.oceana.org/north-america/publications/reports/the-other-side-of-oil-slicks (accessed May 22, 2012).
4. Oceana. *The EU Fleet and Chronic Hydrocarbon Contamination of the Oceans*, 2004, http://www.oceana.org/north-america/publications/reports/the-eu-fleet-and-chronic-hydrocarbon-contamination (accessed May 22, 2012).
5. UNESCO. *The Integrated, Strategic Design Plan for the Coastal Ocean Observations Module of the Global Ocean Observing System*, GOOS Report N 125, IOC Information Documents Series N 1183; UNESCO: Paris, 2003.
6. Finnish Environment Institute, 2004, http://www.ymparisto.fi (accessed March 15, 2005).
7. Kostianoy, A.G.; Lavrova, O.Yu. Introduction. In: *Oil Pollution in the Baltic Sea*. A.G. Kostianoy and O.Yu. Lavrova (Eds.). Springer-Verlag: Berlin, Heidelberg, New York, 2014. Vol. 27, pp. 1–13.
8. Kostianoy, A.G.; Lavrova, O.Yu. (Eds.) *Oil pollution in the Baltic Sea. The Handbook of Environmental Chemistry*. Springer-Verlag: Berlin, Heidelberg, New York, 2014, Vol. 27, 268pp.
9. Carpenter, A. (Ed.) *Oil Pollution in the North Sea*. Springer International Publishing AG: Cham, Switzerland, 2016, 312pp.

10. Carpenter, A.; Kostianoy, A.G. (Eds.) *Oil Pollution in the Mediterranean Sea: Part I – The International Context*. Springer International Publishing AG: Cham, Switzerland, 2018, 350pp.
11. Carpenter, A.; Kostianoy, A.G. (Eds.) *Oil Pollution in the Mediterranean Sea: Part II – National Case Studies*. Springer International Publishing AG: Cham, Switzerland, 2018, 291pp.
12. Hanninen, S.; Rytkonen, J. *Oil transportation and terminal development in the Gulf of Finland*; VIT publication N 547; VIT Technical Research Center of Finland, 2004, 141pp.
13. Kuronen, J.; Helminen, R.; Lehikoinen, A.; Tapaninen, U. *Maritime transportation in the Gulf of Finland in 2007 and in 2015*; Publications from the Centre for Maritime Studies, University of Turku, 2008, N A-45, 114pp.
14. HELCOM. Maritime activities in the Baltic Sea—An integrated thematic assessment on maritime activities and response to pollution at sea in the Baltic Sea Region. *Baltic Sea Environment* **2010**, *123*, 68.
15. Global Marine Oil Pollution Information Gateway, 2004, http://oils.gpa.unep.org/framework/region-2next.htm (accessed March 15, 2005).
16. Annual Report on discharges observed during aerial surveillance in the Baltic Sea in 2017. HELCOM, 2018, http://www.helcom.fi/Lists/Publications/HELCOM%20Annual%20report%20on%20discharges%20observed%20during%20aerial%20surveillance%20in%20the%20Baltic%20Sea%202017.pdf (accessed August 13, 2019).
17. Kostianoy, A.G.; Lebedev, S.A.; Litovchenko, K.Ts.; Stanichny, S.V.; Pichuzhkina, O.E. Satellite remote sensing of oil spill pollution in the southeastern Baltic Sea. *Gayana* **2004**, *68* (2), 327–332.
18. Kostianoy, A.G.; Lebedev, S.A.; Soloviev, D.M.; Pichuzhkina, O.E. *Satellite monitoring of the Southeastern Baltic Sea. Annual Report 2004*; Lukoil-Kaliningradmorneft: Kaliningrad, 2005.
19. Kostianoy, A.G.; Lebedev, S.A.; Litovchenko, K.Ts.; Stanichny, S.V.; Pichuzhkina, O.E. Oil spill monitoring in the Southeastern Baltic Sea. *Environmental Research, Engineering and Management* **2005**, *3* (33), 73–79.
20. Kostianoy, A.G.; Litovchenko, K.Ts.; Lavrova, O.Y.; Mityagina, M.I.; Bocharova, T.Y.; Lebedev, S.A.; Stanichny, S.V.; Soloviev, D.M.; Sirota, A.M.; Pichuzhkina, O.E. Operational satellite monitoring of oil spill pollution in the southeastern Baltic Sea: 18 months experience. *Environmental Research, Engineering and Management* **2006**, *43* (38), 70–77.
21. Bulycheva, E.V.; Kostianoy, A.G. Results of the satellite monitoring of oil pollution in the Southeastern Baltic Sea in 2006–2009. *Current Problems of Remote Sensing of the Earth from Space*, **2011**, *8* (2), 74–83 (in Russian).
22. Bulycheva, E.V.; Kostianoy, A.G. Results of satellite monitoring of the sea surface oil pollution in the Southeastern Baltic Sea in 2004–2013. *Current Problems in Remote Sensing of the Earth from Space*, **2014**, *11* (4), 111–126 (in Russian).
23. Lavrova, O.Yu.; Mityagina, M.I.; Kostianoy, A.G.; Semenov, A.V. Oil pollution in the southeastern Baltic Sea in 2009–2011. *Transport and Telecommunication*, **2014**, *15* (4), 322–331.
24. Kostianoy, A.G.; Bulycheva, E.V.; Semenov, A.V.; Krainyukov, A.V. Satellite monitoring systems for shipping, and offshore oil and gas industry in the Baltic Sea. *Transport and Telecommunication*, **2015**, *16* (2), 117–126.
25. Bulycheva, E.V.; Krek, A.V.; Kostianoy, A.G.; Semenov, A.V.; Joksimovich, A. Oil pollution of the Southeastern Baltic Sea by satellite remote sensing data and in-situ measurements. *Transport and Telecommunication*, **2015**, *16* (4), 296–304.
26. Bulycheva, E.V.; Krek, A.V.; Kostianoy, A.G.; Semenov, A.V.; Joksimovich, A. Oil pollution in the Southeastern Baltic Sea by satellite remote sensing data in 2004–2015. *Transport and Telecommunication*, **2016**, *17* (2), 155–163.
27. Krek, E.V.; Krek, A.V.; Kostianoy, A.G. Seasonal variability of oil pollution in the Southeastern Baltic Sea. *Current Problems in Remote Sensing of the Earth from Space*, **2018**, *15* (3), 171–182 (in Russian).

28. Krek, E.; Kostianoy, A.; Krek, A.; Semenov, A.V. Spatial distribution of oil spills at the sea surface in the Southeastern Baltic Sea according to satellite SAR data. *Transport and Telecommunication*, **2018**, *19* (4), 294–300.
29. Ambjorn, C. Seatrack Web, forecast of oil spills, a new version. *Environmental Research, Engineering and Management*, **2007**, *3* (41), 60–66.
30. SeaTrackWeb User Guide, 2019, https://stw.smhi.se/player/help/classic/?domain=helcom (accessed on August 13, 2019).
31. Kostianoy, A.G.; Ambjörn, C.; Solovyov, D.M. Seatrack web – A numerical tool for environmental risk assessment in the Baltic Sea. In: *Oil Pollution in the Baltic Sea*, A.G. Kostianoy and O.Yu. Lavrova (Eds.), Springer-Verlag: Berlin, Heidelberg, New York, 2014, Vol. 27, pp. 185–220.
32. Kostianoy, A.; Ermakov, P.; Soloviev, D. Complex satellite monitoring of the Nord Stream gas pipeline construction. *Proceedings of the US/EU Baltic 2008 International Symposium "Ocean Observations, Ecosystem-Based Management and Forecasting"*, May 27–29, 2008, Tallinn, Estonia, 2008.
33. Kostianoy, A.G. Oil pollution: Baltic Sea. In: *Encyclopedia of Environmental Management*, S.E. Jorgensen (Ed.), Taylor & Francis: New York, 2013, Vol. III, pp. 1851–1866.
34. Kostianoy, A.G.; Solovyov, D.M. Operational satellite monitoring systems for marine oil and gas industry. *Proceedings of the 2010 Taiwan Water Industry Conference*, October 28–29, 2010, Tainan, Taiwan, 2010, pp. B173–B185.
35. Lavrova, O.Yu.; Kostianoy, A.G. A catastrophic oil spill in the Gulf of Mexico in April–May 2010. *Russian Journal of Remote Sensing (Issledovanie Zemli iz Kosmosa)* **2010**, *6*, 67–72 (in Russian).

35

Status of Groundwater Arsenic Contamination in the GMB Plain

Abhijit Das, Antara Das, Meenakshi Mukherjee, Bhaskar Das, Subhas Chandra Mukherjee, Shyamapada Pati, Rathindra Nath Dutta, Quazi Quamruzzaman, Khitish Chandra Saha, Mohammad Mahmudur Rahman, Dipankar Chakraborti, and Tarit Roychowdhury

Introduction

During the past few decades with the discovery of newer arsenic sites around the world, high arsenic contamination in drinking water has emerged as a major public health problem. Asian countries, especially the Ganga–Meghna–Brahmaputra (GMB) plain of India and Bangladesh, are the worst affected among the world arsenic scenario. This chapter gives a brief update of the global arsenic contamination situation with a detailed update of arsenic contamination in the GMB plain. It appears that a good portion of all states and countries in the GMB plain, comprising an area of more than 569,749 km^2 and a population of more than 500 million, may be at risk from groundwater arsenic contamination

Background

The natural arsenic content in soil is found at an average value of 5 mg/kg in the range of 1–40 mg/kg.[1,2] Agricultural soils have been detected to contain mean As concentrations in the range of 50–60 mg/kg.[3] Large variations of arsenic content can be found in sediments of lakes, rivers, and streams.[4] In surface water and groundwater not polluted by arsenic, its concentration is 10 μg/L, and in open ocean water, its concentration is typically 1–2 μg/L.[5] Arsenic can be found in inorganic forms of As (III) and As (V). Arsenite [As (III)] is the dominant form under reducing conditions, while arsenate [As (V)] is generally

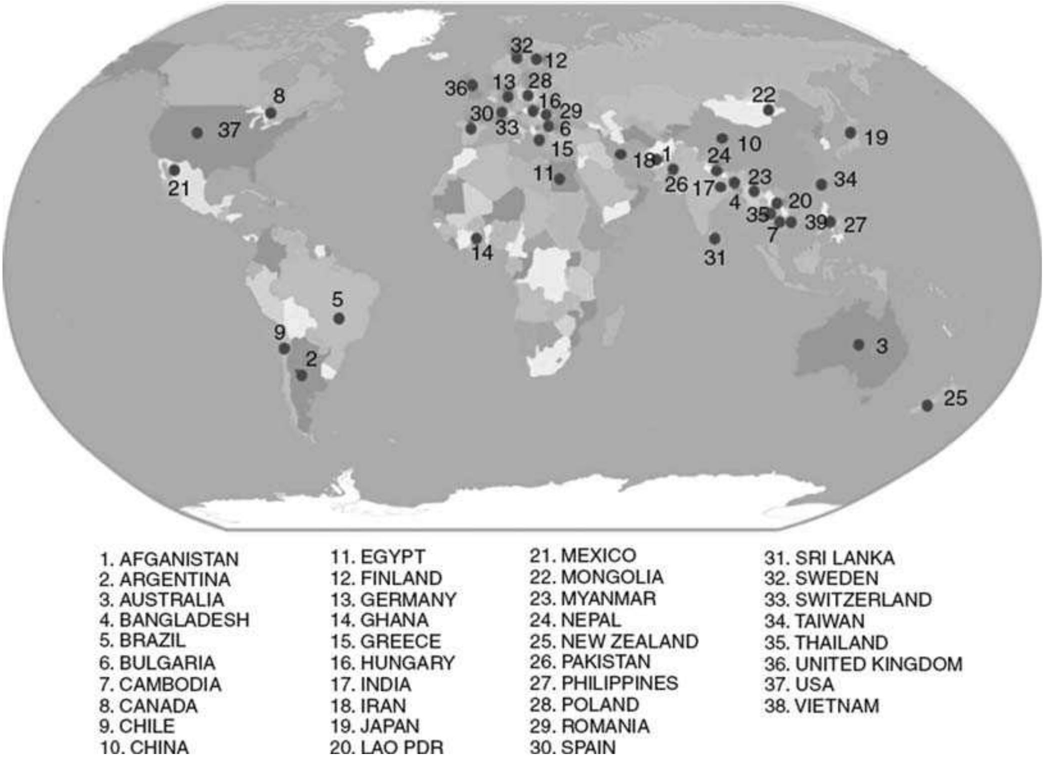


FIGURE 1 Current arsenic contamination scenario around the world.

the stable form of arsenic in oxygenated environments. Presently, there have been 42 incidents of arsenic contamination reported from 38 countries in the world. Figure 1 illustrates the update of these incidents.

However, the magnitude of arsenic contamination is considered highest in the following four Asian countries: Bangladesh, India, China, and Taiwan. Based on the survey carried out over the last 23 years, it can be said that the groundwater of some parts of all the states (Uttar Pradesh, Bihar, Jharkhand, West Bengal, Arunachal Pradesh, and Assam) in GMB plain in India and six out of seven of the northeastern states (all except Mizoram) will be arsenic affected. By the year 2014, the number of states in the pockets of GMB plain where groundwater is reported to be As polluted has increased from seven to ten; Haryana, Punjab, and Karnataka have been added to the abovementioned list of states.[6] Figure 2 shows the arsenic-affected states and countries in the GMB plain. The scenario of groundwater As contamination in the affected states in India is given in detail in Table 1.

Groundwater Arsenic Contamination in West Bengal, India

The severity of Arsenic pollution in groundwater of the districts was categorized into severely affected (>300 μg/L), mildly affected (10–50 μg/L) and unaffected (<10 μg/L).[6] The districts of Maldah, Murshidabad, Nadia, North 24 Parganas, South 24 Parganas, Bardhaman, Howrah, Hugli, and Kolkata are the nine affected districts. During the last 20 years, SOES analyzed 140,150 hand tube-well water samples for the presence of arsenic above the WHO's recommended level from the aforementioned nine districts of West Bengal.[7] From the findings, it was observed that 49.7% of these tube wells had arsenic >10 μg/L, 24.7% had arsenic >50 μg/L, and 3.4% of the analyzed tube wells had arsenic concentrations >300 μg/L (the concentration predicting overt arsenical skin lesions).[8] A maximum arsenic concentration of 3700 μg/L was found in the Ramnagar village of Gram Panchayat (GP) Ramnagar II,

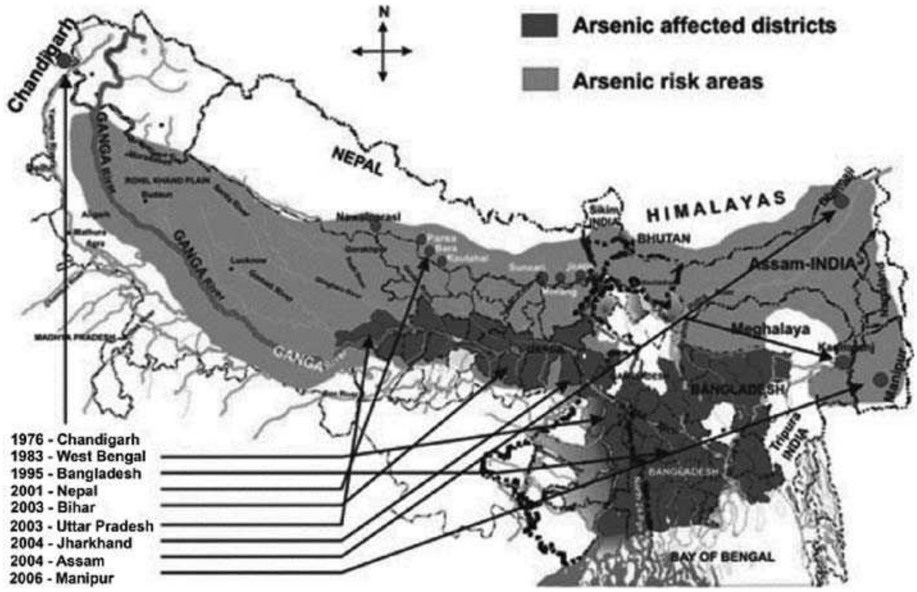


FIGURE 2 Groundwater arsenic contamination situation in India and Bangladesh in the GMB plain.

TABLE 1 Groundwater Arsenic Contamination in India[24,25]

S. No.	Name of the States	Name of the Districts	Level of As Contamination (μg/L)
1	West Bengal	Maldah, Murshidabad, Nadia, North 24 Parganas, South 24 Parganas, Bardhaman, Howrah, Hugli, and Kolkata	50–3700
2	Bihar	Darbhanga, Katihar, Khagaria, Kishanganj, Lakhisarai, Munger, Patna, Purnea, Samastipur, Saran, Vaishali, Begusarai, Bhagalpur, Bhojpur, Buxar	>50
3	Uttar Pradesh	Bahraich, Balia, Balrampur, Bareilly, Basti, Bijnor, Chandauli, Ghazipur, Gonda, Gorakhpur, Lakhimpur Kheri, Meerut, Mirzapur, Moradabad, Raebareli, Sant Kabir Nagar, Shahjahanpur, Siddharthnagar, Sant Ravidas Nagar, Unnao	>50
4	Jharkhand	Sahibgunj	>50
5	Assam	Sivasagar, Jorhat, Golaghat, Sonitpur, Lakhimpur, Dhemaji, Hailakandi, Karimganj, Cachar, Barpeta, Bongaigaon, Goalpara, Dhubri, Nalbari, Nagaon, Morigaon, Darrang, and Baksha	50–657
6	Chhattisgarh	Rajnandgaon	52–88
7	Haryana	Ambala, Bhiwani, Faridabad, Fatehabad, Hissar, Jhajjar, Jind, Karnal, Panipat, Rohtak, Sirsa, Sonepat, Yamunanagar	>50
8	Karnataka	Raichur and Yadgir	>50
9	Punjab	Mansa, Amritsar, Gurdaspur, Hoshiarpur, Kapurthala, Ropar etc.	>50
10	Manipur	Bishnupur and Thoubal	798–986
11	Tripura	North Tripura, Dhalai, West Tripura	65–444
12	Nagaland	Mokokchung and Mon	>50

Baruipur block in South 24 Parganas district. Groundwater arsenic contamination in all the districts of West Bengal is shown in Figure 3. The United Nations Children's Fund (UNICEF) collaborated with the Public Health Engineering Department (PHED), Government of West Bengal. They tested 132,262 government-installed hand tube wells and private tube wells for arsenic from eight arsenic-affected districts of West Bengal.[9] Those tube wells that have water with an arsenic concentration >10 μg/L account for 57.9% of the tested hand tube wells; similarly, 25.5% of the tested hand tube wells contained water that has arsenic concentration >50 μg/L. It is estimated that about 9.5 million people could be drinking water contaminated with an arsenic level of 10 μg/L. A recent study on Nadia, West Bengal reported that all the 17 blocks are affected with arsenic in groundwater. It has been found that about 51.4% and 17.3% of the tube wells had As >10 and 50 μg/L, respectively. The maximum observed level of arsenic in

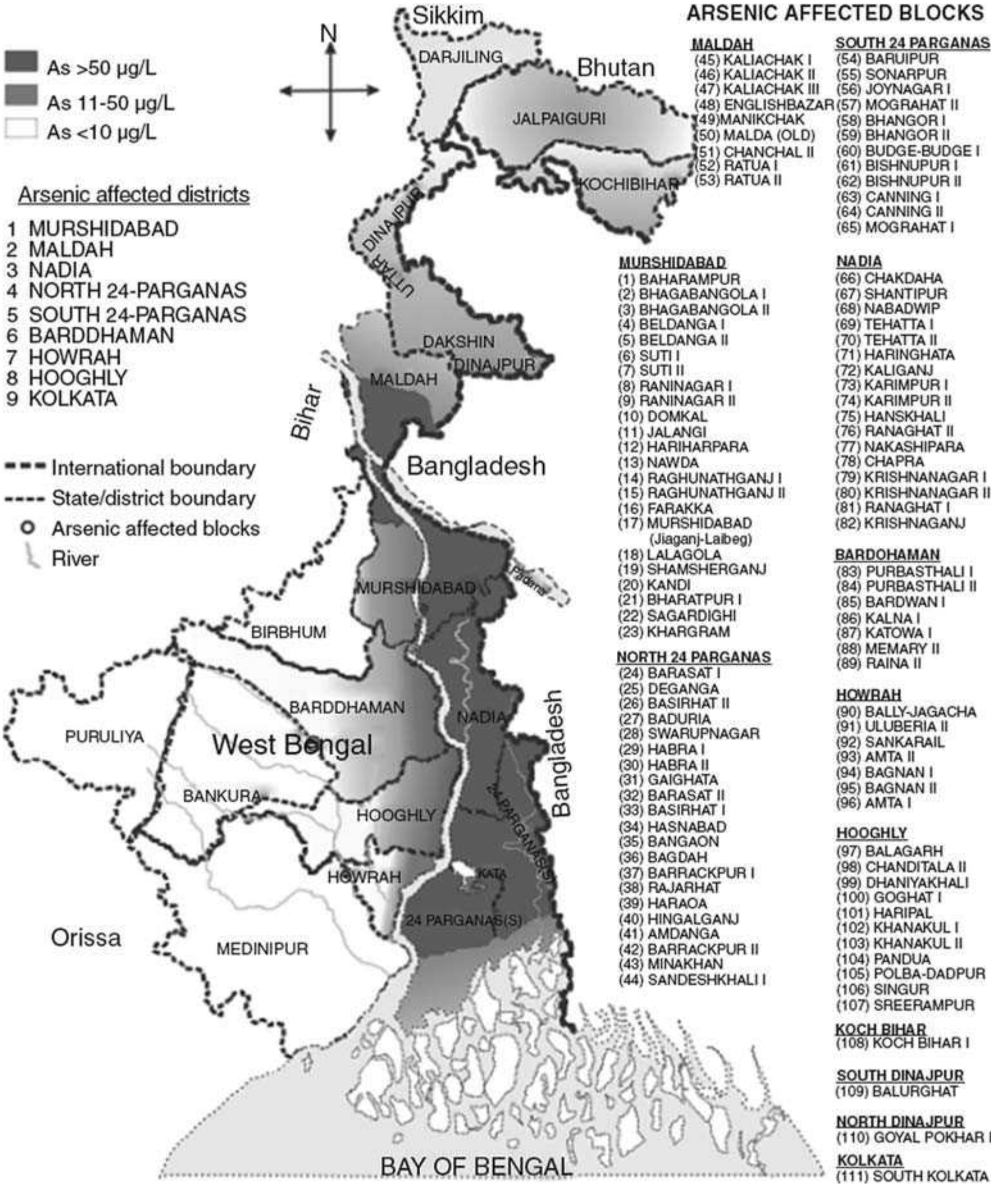


FIGURE 3 Present groundwater arsenic contamination situation in the state of West Bengal in India.

groundwater is found to be 3200 μg/L.[10] Chakdah is one of the worst affected blocks in Nadia where the shallow aquifers as well as the agricultural land show a high magnitude of arsenic content. According to a survey carried out in the nine districts, out of 135,555 samples of groundwater pumped up by tube well, As concentration of above 10 μg/L was reported in 67,306 (49.7%) samples and above 50 μg/L in 33,470 (24.7%) samples. In affected areas, the demographic survey showed that 13.85 million people were under the threat of As concentration above 10 μg/L of which 6.96 million people are exposed to a concentration above 50 μg/L.[11]

Groundwater Arsenic Contamination in Kolkata, India

The present demand of drinking water in Kolkata per day is around 1262 million liters per day (MLD) which is supplied through the pipelines laid down by Metropolitan authority. Out of that, 1096 MLD is treated surface water and the rest is groundwater. A total of 3626 samples of hand tube-well water was collected from 100 out of 141 administrative wards in Kolkata. Tube wells in 65 wards in Kolkata have arsenic concentration levels >10 μg/L, and in another 30 wards, arsenic concentration was found to be >50 μg/L. Based on water samples from 1057 wells, dug deeper than 100 m, 220 samples (20.8%) have shown arsenic concentrations >10 μg/L, 113 samples (10.7%) have arsenic concentrations >50 μg/L, and 72 samples (6.8%) have arsenic concentrations >100 μg/L. Water samples were collected from 977 wells that were 91–100 m deep. Of 977 samples, 149 (15.3%) have arsenic concentrations >10 μg/L and 13 (1.3%) have arsenic concentrations >100 μg/L. However, the tube wells that were deeper than 300 m have been detected with arsenic concentrations >50 μg/L. Maximum arsenic concentration in this region was found to be at 800 μg/L at a depth of 20 m. An analysis of water from these Kolkata Municipal Corporation (KMC) wells shows that out of 734 wells, 121 (16.5%) have arsenic concentrations >10 μg/L, 26 (3.6%) have arsenic concentrations of 50 μg/L, and only 2 (0.2%) have arsenic concentrations >300 μg/L. It has also been found that the southern part of the city is more contaminated by arsenic than the northern and central parts (Figure 4).

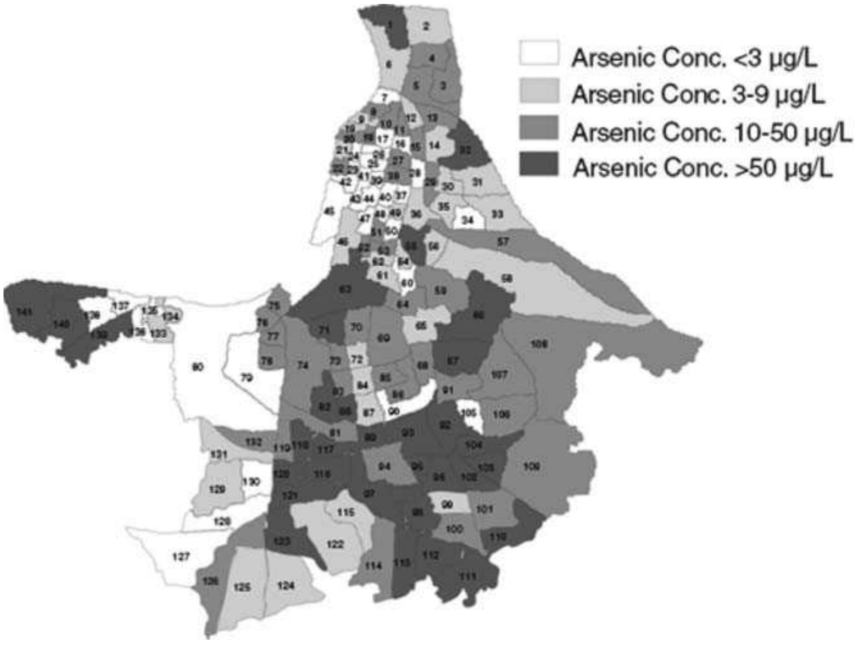


FIGURE 4 Present groundwater arsenic contamination situation in all the 141 wards of KMC, West Bengal.

At present in Kolkata, 70% of KMC area, i.e., 100 out of 144 wards, has been reported with As concentration above permissible limit in shallow unconfined aquifers with 51 wards (i.e. 35.4%) having As concentration >50 μg/L and remaining 49 wards having As concentration between 11 and 50 μg/L.[12] During the past 23 years, 14.2% of samples throughout 77 wards in KMC, exceeded the permissible limit of As in drinking water and around 5.2% samples in 37 wards showed five times higher value of the permissible limit of As in drinking water.[13]

Groundwater Arsenic Contamination in Bihar, India

Water samples from 19,961 tube wells were analyzed, and arsenic content in water samples was found to be >50 μg/L in 12 districts of Bihar. SOES's analysis of water samples from these tube wells revealed that 313 villages have water containing arsenic concentrations >10 μg/L and 240 villages have water containing arsenic concentrations >50 μg/L. Analytical results show that 32.7% of tube wells in that region have concentrations >10 μg/L, 17.75% tube wells have arsenic concentrations >50 μg/L, and 4.55% have arsenic concentrations >300 μg/L. The maximum arsenic concentration was found in the Chakani village of Brahampur GP belonging to the Barahara block of Buxar district. The level of arsenic concentration in its tube well's water is 2182 μg/L. In the two districts of Bihar, Vaishali, and Bhagalpur were reported to have arsenic contamination in groundwater where the maximum arsenic concentration was found to be 20 and 143 μg/L, respectively; 73% of the groundwater samples showed As (III)-dominant species whereas 27% showed As (V). Katihar being the most vulnerable district because of the socioeconomic and biophysical conditions needs a serious concern followed by Vaishali, Samastipur, Khagaria, and Purnia. They investigated 17 districts out of 37 and a total of 87 of 532 community blocks for groundwater arsenic contamination where 72 community blocks are reported with arsenic beyond permissible limit in drinking water.[14] On a further analysis by SOES, Jadavpur, it has been observed that on analysis of 4704 tube-well water samples from all 88 villages of Shahpur block of the district Bhojpur of Bihar, India, 21.1% of water samples is contaminated with arsenic >50 μg/L according to BIS, 2009, whereas nearly 40.3% of water samples is found containing arsenic >10 μg/L (WHO, 2004). The mean and maximum arsenic concentration is found to be 43.5 and 1805 μg/L, respectively.

Groundwater Arsenic Contamination in Uttar Pradesh, India

It was perceived in 2003 that UP is arsenic contaminated.[15] The rural population of the Allahabad district and the urban population of Shuklaganj-Kanpur of the Unnao district in the Allahabad-Kanpur track of the upper Ganga plain first reported cases of arsenic contamination in 49 districts in January 2009.[16] A recent report[11] showed that in a survey done in 2003 in 25 villages of Ballia district, Ghazipur, and Varanasi were added making a total of 20 districts affected by As where the people mostly depended on water from hand pumps which pumped the groundwater from shallow aquifers (20 m deep). Furthermore, it was reported[17] that only 3 out of 72 samples had As concentration <10 μg/L, while 95.83% had As concentration above the permissible limit. The maximum As concentration in the samples collected during pre-monsoon and monsoon was reported to be 75.6 and 74.46 μg/L, respectively. A total of 136 groundwater samples were collected and analyzed from 12 villages on the floodplain of the Rapti River. About 42% of 103 samples collected from nine villages situated on the left bank were reported to have As concentration >10 μg/L with maximum concentration of 399 μg/L while none of the samples collected from the three villages situated on the right bank had As concentration above the permissible limit.[18] On analyzing 5044 water samples from five districts of UP, SOES found 43.1% of those samples to have arsenic concentrations >10 μg/L and 27.5% to have arsenic concentrations >50 μg/L.

Groundwater Arsenic Contamination in Jharkhand, India

In this state, there are only three districts affected by arsenic concentration.[19] It was observed that the groundwater As concentration in Sahibganj district during three seasons was highest in post-monsoon ranging between 1 and 133 μg/L compared to (2–98 μg/L) in monsoon and (7–115 μg/L) in pre-monsoon with the trend of As concentration increasing with depth.[20] SOES has conducted several surveys in the nine blocks of the Sahibganj district (in the middle of Ganga plain) from December 2003 till now, where they have analyzed 3832 hand tube-well water samples for arsenic concentration. The findings suggest that 32.28% of the water samples had arsenic concentrations >10 μg/L, 13.44% contained arsenic concentrations >50 μg/L, and 2.61% had arsenic concentrations >300 μg/L. The highest arsenic concentration in water has been detected in the Hajipur Vitta village of Hajipur Porsun GP, Sahibganj block, in the Sahibganj district of Jharkhand. The amount of arsenic concentration found in the water sample from that region is 1018 μg/L.

Groundwater Arsenic Contamination in Assam, India

Dhemaji and Karimganj in Assam have been surveyed for groundwater arsenic contamination, and both have been found arsenic contaminated in January 2004. So far, water samples from 241 hand tube wells have been analyzed for arsenic contamination from these two districts, and it has been found that 42.3% of these samples have arsenic concentrations >10 μg/L and 19.1% of these samples have arsenic concentrations >50 μg/L.

A total of 27 villages have been identified from these two districts where groundwater arsenic concentrations are above 50 μg/L. Initially, As concentration >50 μg/L was found in a few samples of groundwater collected from Karimganj, Dhubri, and Dhemaji districts. Eighteen out of twenty three districts covering 76 blocks, and a population of 603 are reported to be affected by As toxicity according to a new report by UNICEF.[11] A work in 2014 reported that the mean As concentration in the groundwater of Titabor, Jorhat, Assam is 210 μg/L.[21] It was also noticed that among the six blocks of Golaghat district, Assam, the most severely affected area was Gamariguri block, where 100% of the samples had an As concentration >10 μg/L. The other blocks were also badly affected showing arsenic contamination in groundwater by 87.5%, 71.9%, 42.4%, 42.1%, and 36.4% in the blocks Podumoni, Kathalguri, Dergaon, Morangi, and Kakodonga, respectively.[22]

Groundwater Arsenic Contamination in Manipur, India

Since the preliminary study in Manipur in May 2006, SOES confirmed groundwater arsenic contamination in two districts of Manipur. They are the districts of Thoubal and Imphal. Findings suggest that 41.27% of the collected and analyzed samples have arsenic concentrations >50 μg/L and 64.72% of the samples have arsenic concentrations >10 μg/L. After the preliminary study by SOES in 2006, again 628 water samples of tube wells were analyzed in the Manipur valley which consists of four districts, where 63.3% sample holds arsenic concentration in groundwater >10 μg/L. Among the four districts, Thoubal is the most affected one where 77.6% groundwater contains arsenic >10 μg/L.[23]

Apart from Assam and Manipur, the other north-eastern states like Arunachal Pradesh, Nagaland, and Tripura are also reported with high range of arsenic contamination in groundwater (50–986 μg/L).

Groundwater Arsenic Contamination in Haryana, India

The districts on the banks of Yamuna, originating from Himalaya, and its tributaries are covered by alluvial aquifers resulting in groundwater contamination due to mobilization of As from As-containing minerals. The alluvial sediments are the main source of As in the sporadic region. Under Aquifer Mapping Project, 2013, groundwater As concentration >50 μg/L has been reported in Chandigarh.[11]

Groundwater Arsenic Contamination in Punjab, India

In 2004, 261 shallow groundwater samples were analyzed to find As concentration more than 10 μg/L in 12 districts including Ropar, Amritsar, Hoshiarpur, Gurdaspur, and Kapurthala. In July 2010, out of 105 samples collected from the district of Mansa at a depth ranging from 13 to 35 m, six locations were reported with As concentration more than 10 μg/L.[11]

Groundwater Arsenic Contamination in Bangladesh

In International Conference on Arsenic, Dhaka, held in 1998, it was reported that 66% of water from hand tube wells from Bangladesh contain arsenic concentrations >10 μg/L and 51% have arsenic concentration levels >50 μg/L. The researchers also said that the arsenic detected in the groundwater of 52 districts is >10 μg/L and that in 41 districts is >50 μg/L.[26] In surveys carried out by the British Geological Survey (BGS) along with the Department of Public Health Engineering (DPHE), it was further reported that water samples from 46% of 3534 analyzed tube wells exceeded arsenic concentrations of 10 μg/L and 27% of the samples exceeded arsenic concentration levels of 50 μg/L. This survey had taken samples from tube wells from all over Bangladesh, leaving out the Chittagong Hill tract. Based on their study, BGS-DPHE reported that 57 million and 35 million people could be drinking As-contaminated water >10 μg/L and 50 μg/L, respectively.[27] Between 2000 and 2001, half of 6500 tube wells were sampled for testing by Van Geen in the Araihazar Police Station, which is known as Thana/Upazila. It was found that half of these water samples contain arsenic with concentrations >10 μg/L and one-quarter of these samples contain arsenic concentrations >50 μg/L.[28] It can be found from this database that 1.4 million out of 4.8 million water samples, which is 30% of the water samples in Bangladesh, contain arsenic concentrations >50 μg/L.[29]

The important incidents related to groundwater arsenic contamination in the GMB plain are shown in Table 2.

Bangladesh can be divided into four existing geomorphological regions of which the hill tract and the tableland are free from arsenic contamination. The highly contaminated regions of Bangladesh fall within the deltaic region (Figure 5).

Arsenic content in water in the tableland area (detected in 17 out of 9755 analyzed hand tube wells, which is 0.2%) is >50 μg/L. In another 204 out of the total of 9755, which is 2.1%, arsenic concentration was found between 10 and 50 μg/L. In the floodplain region of Bangladesh, there are 229 thanas in total. Out of these, 19,845 water samples were collected from 158 thanas and were analyzed which suggest that 50.8% of these samples have arsenic concentrations greater than 10 μg/L and another 34.9% of the samples have arsenic concentrations >50 μg/L. Another survey by BGS-DPHE in 1999 has brought out data that indicate that 4587 out of 16,513 (27.8%) of the water samples analyzed have shown concentrations of arsenic >50 μg/L. This survey was also held in the same region. A total of 22,113 water samples from hand tube wells in the deltaic region including the coastal belt were taken for analysis. These samples were collected from 97 of the 139 thanas in that region. On analysis, it was found that 54.7% of the samples have arsenic concentrations >10 μg/L and 32.8% >50 μg/L. In 1999, BGS-DPHE also reported about 12,245 water samples from 126 thanas of this region. A total of 3344 samples (27.5%) have arsenic concentrations >50 μg/L. The study had only collected samples from Chittagong and Cox's Bazar districts. The percentage of samples that have arsenic concentrations >1000 μg/L in the floodplains is 1.14% (227 samples), which is greater than % found in the deltaic region, which is 0.33% (74 samples). Shallow tube wells varying in depth from 10 to 70 m are highly contaminated compared to deep aquifers leading to Arsenic contamination in 61 out of 65 districts of Bangladesh.[30] The maximum arsenic concentration found in the three regions are 134, 4730, and 2190 μg/L, respectively ($n = 9755$, 19,845 and 22,113, respectively). Groundwater As contamination in different areas of Bangladesh reported so far has been given in Table 3.

TABLE 2 Groundwater Arsenic Contamination in India and Bangladesh (September 2010)

Parameters	West Bengal	Bihar	UP	Jharkhand (Sahibganj)	Assam	Manipur	Chhattisgarh (Rajnandgaon)	Bangladesh
Area (km^2)	88,750	94,163	238,000	1,600	78,438	22,327	6,396	147,570
Population (in millions)	80.2	83	166	1	26.6	2.29	1.5	124.3
Arsenic-affected area (km^2)	38,861	21,271	17,919	725	8,822	2,238	6,396	
Total arsenic-affected districts (As > 50 μg/L)	12	12	5	1	3	4	1	50
Total arsenic-affected blocks/PS (As >50 μg/L)	111	36	9	3	9	–	–	197
No. of villages where groundwater arsenic >50 μg/L	3417	235	70	68	–	–	–	–
Total hand tube-well water samples analyzed	140,231	19,961	5,044	3,354	1,448	628	146	52,202
% of samples having arsenic >10 μg/L	48.1	32.70	46.1	36.1	39.0	63.3	25.34	43.0
% of samples having arsenic >50 μg/L	23.8	17.75	26.9	15.4	13.81	23.2	8.22	27.2
Total number of biological samples analyzed (hair, nail, urine)	39,624	1,833	258	367		57 (urine)		11,298
Total people screened by SOES-JU	96,000	3,012	989	522	–	–	a	18,991
People registered with arsenical skin lesions	9,356	457	154	71	–	–	80	3,762
People drinking arsenic-contaminated water >10 μg/L (in millions)	9.5	3.1	1.1	0.15				36.6
People drinking arsenic-contaminated water >50 μg/L (in millions)	4.6	1.7	0.6	0.06				22.7
Population potentially at risk from arsenic contamination >10 μg/L (in millions)	26	9	3	0.4	1.2	1		

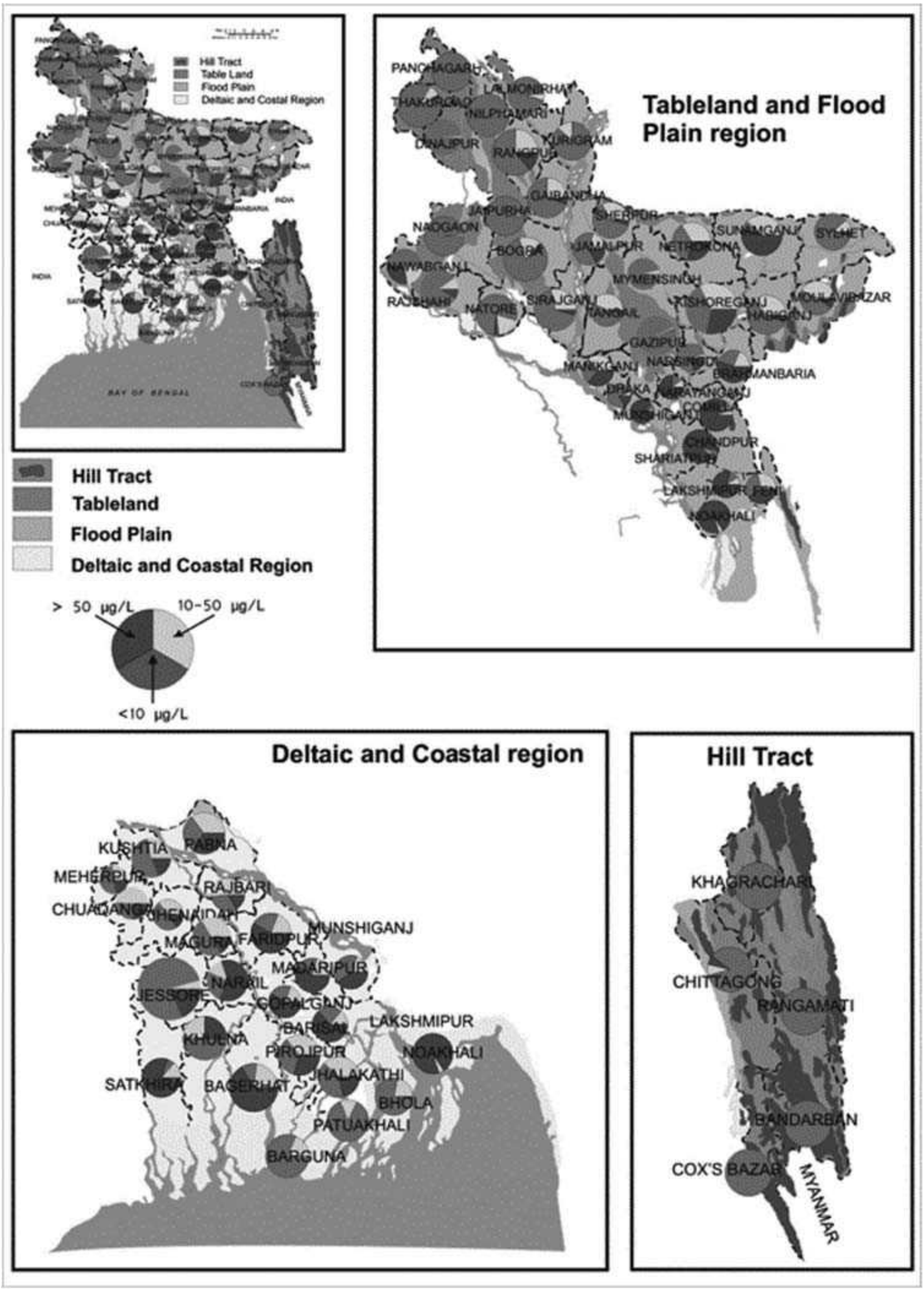


FIGURE 5 Status of As contamination in groundwater in Bangladesh.[26]

Bangladesh and West Bengal are considered the two worst arsenic-affected areas in the world. In Bangladesh, 42.1%, 27.2%, and 7.5% of the hand tube wells had As 10, 50, and 300 μg/L, respectively, whereas in West Bengal, the corresponding values were 48.1%, 23.8%, and 3.3%. SOES has detected 301—(0.6%) hand tube-well water samples with arsenic concentrations >1000 μg/L in Bangladesh, whereas 187 (0.1%) water samples had arsenic concentrations >1000 μg/L in West Bengal.

TABLE 3 Arsenic Contamination Scenario in Groundwater of Bangladesh

S. No.	Cases of As Contamination Reported	Reference
1	Groundwater of 47 districts was reported to be As contaminated on the basis of 22,003 tube-well water samples collected over 64 districts and 6000 samples which were collected over 25 km^2 area and reported to have As concentration exceeding the Bangladesh drinking water standard of 50 μg/L	[31]
2	1,00,000 tube wells in Chandpur	[32]
3	Groundwater As concentration varies from 6 to 934 μg/L with an average of 347 μg/L in Meghna river delta	[33]
4	As concentration >50 μg/L found in 27.2% of 5202 water samples collected from 64 districts	[34]
5	60% tube wells are severely affected in southern and eastern Bangladesh with a maximum concentration of 1000 μg/L	[35–37]

Conclusion

The situation of As contamination has been deteriorating ever since the first evidence of As in Asian countries such as India, Bangladesh, China, Nepal, Taiwan, and Vietnam as well as Mexico, Argentina, etc. Besides the geogenic cause, excessive withdrawal of groundwater to satisfy the increasing demand of the growing population has played an important role in groundwater As pollution. In spite of several mitigation strategies already undertaken, the global situation is even worse than before due to lack of awareness and insufficient steps taken to deal with the problem in domestic level. Since the toxicity of As is directly dependent on the nutritional status as well as economic condition of the people, efforts should be made to improve the condition at ground level.

References

1. Beyer, W.N.; Cromartie, E.J. A survey of Pb, Cu, Zn, Cd, Cr, As and Se in earthworms and soil from diverse sites. *Environmental Monitoring and Assessment* **1987**, *8*(1), 27–36.
2. Bowen, H.J.M. *Elemental Chemistry of the Elements*; Academic Press: London, 1979.
3. Sanok, W.J.; Ebel, J.G.; Manzell, K.L.; Gutenmann, W.H.; Lisk, D.J. Residues of arsenic and lead in potato soils on Long Island. *Chemosphere* **1995**, *30*(4), 803–806.
4. Welch, A.H.; Lico, M.S.; Hughes, J.L. Arsenic in groundwater of the western United States. *Ground Water*, **1988**, *26*(3), 333–347.
5. Smith, R.A.; Alexander, R.B.; Wolman, M.G. Water quality trends in the nation's rivers. *Science* **1987**, *235*, 1607–1615.
6. Bhattacharya, A.K. An analysis of arsenic contamination in the groundwater of India, Bangladesh and Nepal with a special focus on the stabilisation of Arsenic Laden Sludge from arsenic filters. *Electronic Journal of Geotechnical Engineering* **2019** *24*(1).
7. Chakraborti, D.; Das, B.; Rahman, M.M.; Chowdhury, U.K.; Biswas; Goswami, A.B.; Nayak, B.; Pal, A.; Sengupta, M.K.; Ahamed, S.; Hossain, M.A.; Basu, G.; Roychowdhury, T.; Das, D. Status of Groundwater arsenic contamination in the state of West Bengal, India: A 20-year study report. *Molecular Nutrition & Food Research* **2009**, *53*, 542–551.
8. Rahman, M.M.; Chowdhury, U.K.; Mukherjee, S.C.; Mondal, B.K.; Paul, K.; Lodh, D.; Chanda, C.R.; Basu, G.K.; Biswas, B.K.; Saha, K.C.; Roy, S.; Das, R.; Palit, S.K.; Quamruzzaman, Q.; Chakraborti, D. Chronic arsenic toxicity in Bangladesh and West Bengal-India. *Journal of Toxicology: Clinical Toxicology* **2001**, *39*(7), 683–700.
9. Nickson, R.; Sengupta, C.; Mitra, P.; Dave, S.N.; Banerjee, A.K.; Bhattacharya, A.; Basu, S.; Kakoti, S.; Moorthy, N.S.; Wasuja, M.; Kumar, M.; Mishra, D.S.; Ghosh, A.; Vaish, D.P.; Srivastava, A.K.; Tripathi, R.M.; Singh, S.N.; Prasad, R.; Bhattacharya, S.; Deverill, P. Current knowledge on arsenic in groundwater in five states of India. *Journal of Environmental Science and Health Part A* **2007**, *42*, 1707–1718.

10. Rahman, M. M.; Mondal, D.; Das, B.; Sengupta, M. K.; Ahamed, S.; Hossain, M. A.; Samal, A.C.; Saha, K.C.; Mukherjee, S.C.; Dutta, R.N.; Chakraborti, D. Status of groundwater arsenic contamination in all 17 blocks of Nadia district in the state of West Bengal, India: A 23-year study report. *Journal of Hydrology* **2014**, *518*, 363–372.
11. Bhattacharya, A. K.; Lodh, R. Arsenic contamination in the groundwater of India with a special focus on the stabilization of Arsenic-Laden Sludge from arsenic filters. *Electronic Journal of Geotechnical Engineering* **2018**, *23*(1), 575–600.
12. Malakar, A.; Islam, S.; Ali, M. A.; Ray, S. Rapid decadal evolution in the groundwater arsenic content of Kolkata, India and its correlation with the practices of her dwellers. *Environmental Monitoring and Assessment* **2016**, *188*(10), 584.
13. Chakraborti, D.; Das, B.; Rahman, M. M.; Nayak, B.; Pal, A.; Sengupta, M. K.; Ahamed, S.; Hossain, M.A.; Chowdhury, U.K.; Biswas, B.K.; Saha, K. C. Arsenic in groundwater of the Kolkata Municipal Corporation (KMC), India: Critical review and modes of mitigation. *Chemosphere* **2017**, *180*, 437–447.
14. Singh, S. K. Groundwater arsenic contamination in the Middle-Gangetic Plain, Bihar (India): the danger arrived. *International Research Journal of Environmental Science* **2015**, *4*(2), 70–76.
15. Chakraborti, D.; Mukherjee, S.C.; Pati, S.; Sengupta, M.K.; Rahman, M.M.; Chowdhury, U.K.; Lodh, D.; Chanda, C.R.; Chakraborti, A.K.; Basu, G.K. Arsenic groundwater contamination in middle Ganga Plain, Bihar, India: A future danger. *Environmental Health Perspectives* **2003**, *111*, 1194–1201.
16. Chakraborti, D.; Ghorai, S.; Das, B.; Pal, A.; Nayak, B.; Shah, B.A. Arsenic exposure through groundwater to the rural and urban population in the Allahabad-Kanpur track in the upper Ganga plain. *Journal of Environmental Monitoring*, **2009**, *11*, 1455–1459.
17. Singh, A. L.; Singh, V. K. Assessment of groundwater quality of Ballia district, Uttar Pradesh, India, with reference to arsenic contamination using multivariate statistical analysis. *Applied Water Science* **2018**, *8*(3), 95.
18. Singh, C. K.; Kumar, A.; Bindal, S. Arsenic contamination in Rapti River Basin, Terai region of India. *Journal of Geochemical Exploration* **2018**, *192*, 120–131.
19. Nayak, B.; Das, B.; Mukherjee, S.C.; Pal, A.; Ahamed, A.; Hossain, M.A.; Maity, P.; Dutta, R.N.; Dutta, S.; Chakraborti, D. Groundwater arsenic contamination in Sahibganj district of Jharkhand state, India in middle Ganga plain and adverse health effects. *Toxicology & Environmental Chemistry* **2007**, *90* (4), 673–694.
20. Alam, M. O.; Shaikh, W. A.; Chakraborty, S.; Avishek, K.; Bhattacharya, T. Groundwater arsenic contamination and potential health risk assessment of Gangetic Plains of Jharkhand, India. *Exposure and Health* **2016**, *8*(1), 125–142.
21. Sailo, L.; Mahanta, C. Arsenic mobilization in the Brahmaputra plains of Assam: groundwater and sedimentary controls. *Environmental Monitoring and Assessment* **2014**, *186*(10), 6805–6820.
22. Chetia, M.; Chatterjee, S.; Banerjee, S.; Nath, M. J.; Singh, L.; Srivastava, R. B.; Sarma, H. P. Groundwater arsenic contamination in Brahmaputra river basin: a water quality assessment in Golaghat (Assam), India. *Environmental Monitoring and Assessment* **2011**, *173*(1–4), 371–385.
23. Chakraborti, D.; Singh, E. J.; Das, B.; Shah, B. A.; Hossain, M. A.; Nayak, B.; Ahamed, S. & Singh, N. R. Groundwater arsenic contamination in Manipur, one of the seven North-Eastern Hill states of India: a future danger. *Environmental Geology* **2008**, *56*(2), 381–390.
24. Sanyal, S. K. Arsenic toxicity in soil-plant-human continuum and remedial options. In: *Dialogue-Science, Scientists, and Society.* Ed. M. Dadlani, 2018, 1–17. doi:10.29195/DSSS.01.01.0006
25. Kumar, C. P. Status and mitigation of arsenic contamination in groundwater in India. *International Journal of Earth & Environmental Sciences* **2015**, *1*(1), 1–10.

26. Chakraborti, D.; Rahman, M.M.; Das, B.; Murrill, M.; Dey, S.P.; Mukherjee, S.C.; Dhar, R.K.; Biswas, B.K.; Chowdhury, U.K.; Roy, S.; Sorif, S.; Selim, M.; Rahman, M.; Quamruzzaman, Q. Status of groundwater arseniccontamination in Bangladesh: A 14-year study report. *Water Research* **2010**, *44*(19), 5789–5802.
27. BGS-DPHE. Arsenic Contamination of Groundwater in Bangladesh, BGS Technical Report WC/00/19; British Geological Survey: Keyworth, UK, 2001.
28. Van Geen, A.; Cheng, Z.; Jia, Q.; Seddique, A.A.; Rahman, M.W.; Rahman, M.M.; Ahmed, K.M. Monitoring 51 community wells in Araihazar, Bangladesh, for up to 5 years: Implications for arsenic mitigation. *Journal of Environmental Science & Health* **2007**, *42*, 1729–1740.
29. Johnston, R.B.; Sarker, M.H. Arsenic mitigation in Bangladesh: National screening data and case studies in three upazilas. *Journal of Environmental Science & Health* **2007**, *42*, 1889–1896.
30. BGS (British Geological Survey). Executive Summary of the Main Report of Phase I, Groundwater Studies of As Contamination in Bangladesh. British Geological Survey and Mott MacDonald (UK) for the Government of Bangladesh, Ministry of Local Government, Rural Development and Cooperatives DPHE and DFID (UK), 2000. Available at http://www.engconsult.com/arsenic/article/DFID-sum.html (accessed in February 2014)
31. Van Geen, A.; Zheng, Y.; Versteeg, R.; Stute, M.; Horneman, A.; Dhar, R.; Steckler, M., Gelman, A.; Small, C.; Ahsan, H. and Graziano, J.H. Spatial variability of arsenic in 6000 tube wells in a 25 km^2 area of Bangladesh. *Water Resources Research* **2003**, *39*(5).
32. Islam, A. B. M. Arsenic pollution and the probable bioremediation (Doctoral dissertation, PhD Thesis. Kanazawa University, Japan), 2003.
33. Bibi, M. H.; Ahmed, F.; & Ishiga, H. Geochemical study of arsenic concentrations in groundwater of the Meghna River Delta, Bangladesh. *Journal of Geochemical Exploration* **2008**, *97*(2–3), 43–58.
34. Chakraborti, D.; Rahman, M. M.; Das, B.; Murrill, M.; Dey, S.; Mukherjee, S. C.; Dhar, R. K.; Biswas, B.K., Chowdhury, U.K., Roy, S.; Sorif, S. Status of groundwater arsenic contamination in Bangladesh: a 14-year study report. *Water Research* **2010**, *44*(19), 5789–5802.
35. Escobar, M. O.; Hue, N. V.; Cutler, W. G. Recent developments on arsenic: contamination and remediation. *Recent Research Developments in Bioenergetics* **2006**, *4*, 1–32.
36. Kinniburgh, D. G.; Kosmus, W. Arsenic contamination in groundwater: some analytical considerations. *Talanta* **2002**, *58*(1), 165–180.
37. Uddin, R.; Huda, N. H. Arsenic poisoning in Bangladesh. *Oman Medical Journal* **2011**, *26*(3), 207–208.

36

Yellow River

Zixi Zhu, Ynuzhang Wang, and Yifei Zhu

Introduction

Since 1949, China has made unremitting efforts to harness the Yellow River. A number of water control projects were undertaken, hydropower stations had been built, and the dykes in the lower reaches had been strengthened and heightened. The eroded land on the Loess Plateau had been harnessed and water and soil loss was restricted. Soil and water conservation improved both agricultural production conditions and the ecological environment. Since 1949, the Yellow River has never been breached, thus ensuring the safety of its people and property and promoting the development of economy and society.

Discussion

The Yellow River originates from the Yueguzonglie Basin, which has an elevation of 4500 m and is located at the northern slope of the Bayankera Mountain in the Qinghai–Tibet Plateau. From west to east, it flows through nine provinces/autonomous regions: Qinghai, Sichuan, Gansu, Ningxia, Inner Mongolia, Shanxi, Shaanxi, Henan, and Shandong, and empties into the Bohai Sea in Kenli County, Shandong Province. The trunk is 5464 km in length and the basin area is 795,000 km^2. Within the river basin, the population is 110.08 million, accounting for 8.7% of the total population in China; the cultivated land is 13.1 million ha.[1] The Yellow River basin is the cradle of Chinese nationality and the birthplace of ancient Chinese civilization. As early as 1 million years ago, the "Lantian Man" had been living in the Yellow River Basin. In this very long historical period, the basin was the center of politics, economy, and culture in China.

The terrain of the Yellow River basin is high in the west and low in the east, thus showing great disparity. The west section of the basin belongs to the Qinghai–Tibet Plateau with an elevation of 3000–5000 m. The middle section is mostly located in the Loess Plateau and reaches the east side of Taihang Mountain, with an elevation of 1000–2000 m. The east section starts from east of Taihang Mountain and ends at the Bohai Sea, lying in the Huanghuaihai Plain, with an elevation of lower than 100 m.[1] Based on the geographical position and the feature of the river, the trunk stream of the Yellow River can be divided into upper, middle, and lower reaches. The upper reaches are from the source of the river to Hekouzhen (Tuoketuo county) in Inner Mongolia, with length of 3472 km and area of 428,000 km^2. The middle reaches are from Hekouzhen, then extend to Taohuayu (Zhengzhou) of Henan province, comprising 1206 km in length and 344,000 km^2 in area. Starting from Taohuayu, the lower reaches run through nearly 800 km, occupy 23,000 km^2 of area, and end at the mouth of the river.[1] Because of the height difference, the climate in the Yellow River basin is very distinct, ranging from arid climate in the west through semiarid to semihumid climate in the east. Average annual temperature is −4°C at the source, 1–8°C in the upper reaches, 8–12°C in the middle reaches, and 12–14°C in the lower reaches. The average annual precipitation of the basin is 452 mm. The maximum is in the southeast part of the basin, reaching 800–1000 mm; whereas the minimum annual precipitation is less than 200 mm in the northwest part of the basin, including Ningxia and Inner Mongolia, which features the inland climate.[1] The Yellow River basin is rich in mineral and energy resources that are of great importance in China. Out of the 45 proved major mineral resources in China, 37 are found in the Yellow River basin.[1] The water energy in the upper-middle reaches of Yellow River, the coal in the middle reaches, and the oil and natural gas in the middle-lower reaches are all quite rich in deposits. Thus the Yellow River basin is called the "energy resources basin," playing an important role in China.[1]

The Yellow River provides the major water resource for northwest and north China, but the amount is comparatively poor. The average annual runoff in the basin is 58 billion m^3, making up only 2% of that in China, whereas the area of the Yellow River basin accounts for 8.3% of the land area of the country.[1] The distributions of runoff in different areas are different. Annual runoff in the upper, middle, and lower reaches contributes 55.6%, 40.8%, and 4.6% of the total annual runoff, respectively. It also varies with seasons. More than 60% of the annual runoff happens during the period from July to October, while less than 40% occurs from November to June.[1] The Yellow River basin is an important production area in agriculture. The main crops are wheat, cotton, oil-bearing plants, tobacco, etc. The irrigation area along the Yellow River basin is about 7.3 million ha,[2] most of which is in the Ningxia–Inner Mongolia Plain at the upper reaches, Fen-Wei basin at the middle reaches, and the irrigation area drawing water from the Yellow River at the lower reaches. Agricultural irrigation consumes 28.4 billion m^3 of water from the Yellow River annually, accounting for 92% of the overall annual water consumption of the river.[2] With high-speed development of the economy and continued population growth, water consumption has increased rapidly and the competition between water supply and demand has become more acute. As a result, zero flow in the lower reaches has resulted in recent years. From 1972 through 1998, zero flow occurred in 21years with accumulated duration of 1051 days. The worst was in 1997, when Lijin, near the river mouth, had zero flow for 226 days; the zero-flow section extended upstream even to Kaifeng of Henan Province.[2]

The Yellow River flows through the Loess Plateau, where the surface is characterized by loose soil and sparse vegetation, and the climate is dominated by dry weather and heavy storms concentrated mostly in the summer. Therefore, it is the largest area affected by water and soil loss and the strongest intensity in erosion in China. Based on the data collected by remote sensor in 1990, the area experiencing water loss and soil erosion is up to 454,000 km^2, which makes up 70.9% of the Loess Plateau's area. The area of water erosion with the annual erosion mean exceeding 8000 Mg/km^2 is 85,000 km^2, accounting for 64% of the congener area in China. The severe water degradation area with annual erosion mean exceeding 15,000 Mg/km^2 is 36,700 km^2, about 89% of the congener area.[1] Average annual amount of sediment and sand washed into the river is about 1.6 billion Mg with sand content of 35 kg/m^3. The maximum sand content is 933 kg/m^3, which was measured at Longmen on July 18, 1966.[3] In the lower reaches of the Yellow River, a large amount of sediment is left behind to raise the riverbed. Nearly 9.2 billion Mg of sediment were deposited in the lower reaches of the Yellow River from 1950 through 1998. Thus, the riverbed is 4–6 m higher than the ground outside the river on an average. At some places, this number could even reach 10 m or more.[2] For this reason, the lower reaches of the Yellow River are called the "suspended river." Because of rainstorms occurring in the middle reaches and the channel in the lower reaches being wider in the upper part and narrower in the lower part, the lower reach area frequently suffers from heavy floods. Records indicated that from 602 BC to 1949 AD, the Yellow River was breached 1590 times and changed its route 26 times, i.e., on the average, "breach twice every three years and changing its route every century."[4] Heavy floods affected a large area from Tianjin City in the north to Huaihe River in the south, crossing 250,000 km^2. On the other hand, drought is another disaster that occurs frequently in the Yellow River basin. In 582 years from 1368 to 1949, severe drought occurred in 107 years, once every 5.4 years.[5] Records show that an extraordinarily serious drought lasted 4 years, from 1875 to 1878, through the whole basin of the Yellow River.

Harnessing of the Yellow River has always been a major issue concerning China's prosperity and the people's peaceful life. Since 1949, China has made unremitting efforts to harness the Yellow River and, therefore, ensured the safety of its people and property, promoted the development of economy and society, and improved the ecological environment. On the trunk of the Yellow River, 15 key water control projects and hydropower stations have been built or are being built, providing a total water capacity of 56.6 billion m^3, a total installed capacity of 11.13 million kW, and an average annual power supply of 40.1 billion kWh.[2] By the end of 2000, 1400 km of dykes along the lower reaches had been strengthened and heightened four times. Large-scale channel improvement had been performed. Reservoirs exceeding 10,000 in number and in different sizes have been built, with a total storage capacity of 72 billion m^3. Among them, 22 of the largest reservoirs are able to hold 61.7 billion m^3 of water.[2] Other projects for irrigation and water supply had also been accomplished and the groundwater had begun to be developed. As a result, the irrigation area has increased from 0.8 million ha in 1950 to 7.3 million ha now, including 2.4 million ha that are located outside of the basin.[2] Until 2000, the area of 180,000 km^2 of eroded land on the Loess Plateau, which is one-third of the land with soil erosion, had been harnessed. To some extent, water and soil loss and desertification have been restricted. The amount of sediment and sand being washed into the Yellow River each year had decreased by about 300 million Mg.[2] Soil and water conservation improved the agricultural production condition and ecological environment. By taking these measures, the average annual grain yield has increased by more than 5 billion kg, which can provide enough food and clothing for more than 10 million people.[2] Since 1949, the Yellow River has never been breached, even when heavy flood with flow of 22,300 m^3/sec occurred in July of 1958.[3]

Conclusion

Compared with other rivers in China, the Yellow River is characterized by shortage of water, high content of sand, and serious loss of water and soil. Average annual runoff is only 58 billion m^3, taking up only 2% of that in China, and annual amount of sediment and sand washed into the river is about 1.6 billion Mg with sand content of 35 kg/m^3. The riverbed in lower reaches is 4–6 m higher than the ground outside the river on average. Floods and droughts are two major disasters in the Yellow River basin. Breach occurred frequently in its history. The Chinese government has made great efforts to harness the river. A number of water control projects have been undertaken, hydropower stations have been built, and dykes in the lower reaches had been strengthened and heightened. To a certain extent, the eroded land on the Loess Plateau had been harnessed and water and soil loss was restricted. The implementation of the great-development-of-the-west strategy of China will further prompt the harnessing and exploitation of the Yellow River along with the development of economy and society.

References

1. Yellow River Conservancy Commission. A survey of the yellow river basin (Ch. 1). *The Short-Term Program on Harnessing and Exploiting the Yellow River*; The Yellow River Conservancy Publishing House: Zhengzhou, China, 2002; 1–17.
2. Yellow River Conservancy Commission. Current situation on harnessing and exploiting the yellow river and the main problems (Ch. 2). *The Short-Term Program on Harnessing and Exploiting the Yellow River*; The Yellow River Conservancy Publishing House: Zhengzhou, China, 2002; 18–32.

3. Lanqin, Z. Section of natural geography. *300 Questions on Yellow River*; The Yellow River Conservancy Publishing House: Zhengzhou, China, 1998; 3–33.
4. Xioufeng, M. Introduction. *Floods and Droughts in Yellow River Basin*; The Yellow River Conservancy Publishing House: Zhengzhou, China, 1996; 1–9.
5. Xioufeng, M. Drought in agriculture (Ch. 10). *Floods and Droughts in Yellow River Basin*; The Yellow River Conservancy Publishing House: Zhengzhou, China, 1996; 270 pp.

IV

DIA: Diagnostic Tools: Monitoring, Ecological Modeling, Ecological Indicators, and Ecological Services

37

Groundwater: Modeling

Jesus Carrera

Introduction

A model is an entity built to reproduce some aspect of the behavior of a natural system. In the context of groundwater, aspects to be reproduced may include: groundwater flow (heads, water velocities, etc.); solute transport (concentrations, solute fluxes, etc.); reactive transport (concentrations of chemical species reacting among themselves and with the solid matrix, minerals dissolving or precipitating, etc.); multiphase flow (fractions of water, air, non-aqueous phase liquids, etc.); energy (soil temperature, surface radiation, etc.); and so forth.

Depending on the type of description of reality that one is seeking (qualitative or quantitative), models can be classified as conceptual or mathematical. A conceptual model is a qualitative description of "some aspect of the behavior of a natural system." This description is usually verbal, but may also be accompanied by figures and graphs. In the groundwater flow context, a conceptual model involves defining the origin of water (areas and processes of recharge) and the way it flows through and exits the aquifer. In contrast, a mathematical model is an abstract description (abstract in the sense that it is based on variables, equations, and the like) of "some aspect of the behavior of a natural system." However, the motivation of mathematical models is not abstraction, but rather quantification. For example, a groundwater flow mathematical model should yield the time evolution of heads and fluxes (water movements) at every point in the aquifer.

Both conceptual and mathematical models seek understanding. Some would argue that understanding is not possible without quantification. In the reverse, one cannot even think of writing equations without some sort of qualitative understanding. The methods of conceptual modeling are those of conventional hydrogeology (study geology, measure heads and hydraulic parameters, hydrochemistry, etc.). On the other hand, the methods of mathematical modeling (discretization, calibration, etc.) are more specific. Yet, it should be clear from the outset that conceptualization is the first step in modeling and that mathematical modeling helps in building firm conceptual models.

Depending on the manner in which equations are solved, models can be classified as analog, analytical, and numerical. Analog models are based on a physical simulation of a phenomenon governed by the same equation(s) as that of our natural system. For example, because of the equivalence between electrostatics and steady state flow, one may use conductive paper subject to an electrical current to solve the flow equation (a parallelism can be established between electric potential and hydraulic head). This kind of application, however, is restricted mainly to teaching. Boxes of resistances and condensators were used in the 1950s and 1960s as analog aquifer models, but they have become inefficient compared to computers. As a result, analog models are no longer used in practice.

Analytical models are based on closed-form solutions to the groundwater flow and transport equations. They are convenient in the sense that they are easy to evaluate and intuitive (visual inspection of the equation may yield an idea of the phenomenon). As a result, they are used very frequently. Examples include solutions of problems in well hydraulics, tracer movement, etc.

Numerical models are based on discretizing the partial differential equations governing flow and transport. This leads to linear systems of equations that can only be solved with the aid of computers. The advantage of numerical models lies in their generality. Analytical models are constrained to homogeneous domains and very simple geometry and boundary conditions. Numerical models, on the other hand, can handle spatially and temporally variable properties, arbitrary geometry and boundary conditions, and complex processes. The price to pay is methodological singularity. Analytical models are easy to use. Numerical models can be complex and, often, difficult.

Because of the methodological singularity mentioned above, this entry concentrates on mathematical numerical models. Analytical solutions are not discussed. In addition, conceptual modeling will be discussed as the first step in modeling, but not by itself.

What Can Be Modeled and What For?

Modeled Phenomena

The most basic phenomenon is groundwater flow (Figure 1) because of its intrinsic importance and because it is needed for subsequent processes. In essence, the flow equation expresses two things. First, groundwater moves according to Darcy's law. Second, a mass balance must be satisfied in the whole aquifer and in each of its parts. Therefore, the main output from flow models is a mass balance: classified inflows, outflows, and storage variations. The output also includes where water flows through the aquifer (water fluxes) and heads (water levels in the aquifer). In essence, input data are a thorough description of hydraulic conductivity (and/or transmissivity), storativity, recharge/discharge throughout the model domain, as well as conditions at the model boundaries. Obviously, these data are never available, and the modeler has to use a good deal of ingenuity to generate them. This is where the conceptual model becomes important.

Specific cases of flow phenomena are unsaturated and multiphase flow. In the first case, one models water flow in the vadose zone or, in general, in areas where water does not fill all the pores.[1] Therefore, besides heads and fluxes, one must work with water contents (volume of water per unit volume of aquifer), capillary pressures and suctions (difference between water and air pressure). From the input viewpoint, the main singularity of unsaturated flow is the need to specify the retention curve (water content vs. suction) and relative permeability (permeability vs. water content). The multiphase flow case is similar, but includes several fluids (phases). It is used to represent the flow of air or mixtures of liquids, singularly non-aqueous phase liquids (NAPLs), which have been the subject of much research in recent years.[2]

Conservative transport refers to the movement of inert substances dissolved in water. Solutes are affected by advection (displacement of the solute as linked to flowing water) and dispersion (dilution of contaminated water with clean water, which causes the size of the contaminated area to grow while reducing peak concentrations). The main input to a solute transport model is the output of a flow model (water fluxes). Additionally, porosity and dispersivity need to be specified (Figure 1). The output is the time evolution and spatial distribution of concentrations. While the amount of data needed for solute

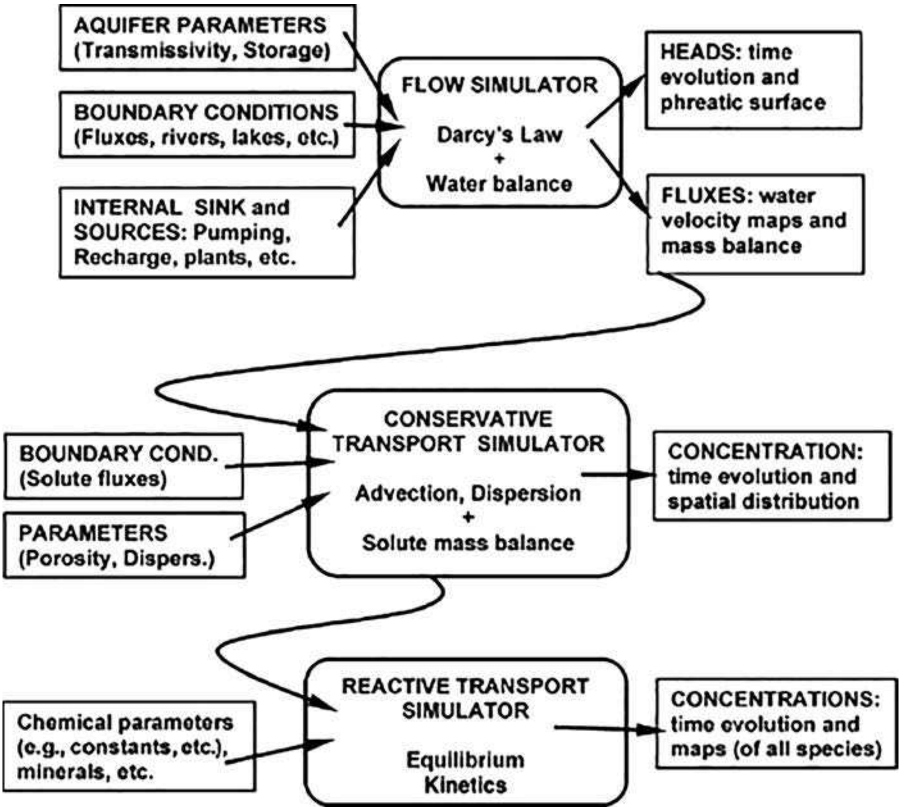


FIGURE 1 A groundwater flow model involves using a flow simulator to take aquifer parameters, boundary conditions, and internal sink and sources as inputs and obtain heads and water fluxes as output. Water fluxes are used in conservative transport models, together with porosity, diffusivity, and solute mass inflows, to yield time evolution and spatial distribution of inert tracers.

transport modeling is relatively small, it must be stressed that solute transport is extremely sensitive to variability and errors in water fluxes. A flow model may be good enough for flow results (heads and water balances) but insufficiently detailed to yield water fluxes good enough for solute transport. Therefore, modeling solute transport ends up being rather difficult.

Reactive transport refers to the movement of solutes that react among themselves and with the soil phase. Reactions can be of many kinds, ranging from sorption of a contaminant onto a solid surface to redox phenomena controlling the degradation of an organic pollutant. Input for reactive transport modeling includes not only the output of flow and conservative transport models but also the equilibrium constants of the reactions (usually available from chemistry databases) and the parameters controlling reaction kinetics. However, the most difficult input is the proper identification of relevant chemical processes. Model output includes the concentrations of all chemical species, the reaction rates, etc.

Coupled models refer to models in which different phenomena are affected reciprocally. Density dependent flow is a typical example. Variations in density affect groundwater flow (e.g., dense sea water sinks under light fresh water), which in turn affects solute transport and, hence, density distribution. Other coupled phenomena are the non-isothermal flow of water (coupling flow and energy transport) and the mechanically driven flow of water (coupling flow and mechanical deformation equations).

What Are Models Built For?

While discussing the usage of models, it is convenient to distinguish between site-specific models and generic models. The former are aimed at describing a specific aquifer while the latter emphasize processes, regardless of where they take place.

Groundwater management is the ideal use of site- specific models. Management involves deciding where to extract and/or inject water to satisfy water needs while ensuring water quality and other constraints. In this context, it is important to point out that a model is essentially a system for accounting water fluxes and stores (Figure 2) in the same way that the accounting system of a company keeps track of money fluxes and reserves. No one would imagine a well-managed company without a proper accounting system. Aquifers will not be managed accurately until they have a model running on real time. Unfortunately, at present, this is still a dream. Because of the difficulties in building and maintaining models and because of legal and practical difficulties to manage aquifers in real time, models are rarely, if ever, used in this fashion.

Instead, models are often used as decision support tools. Building an accurate model is very difficult and time consuming. As a result, one can rarely expect models to yield exact predictions. However, approximate models are much easier to build. These do not result in precise forecasts but normally allow reasonable assessments of the outcome of different management alternatives, i.e., the relative advantages and disadvantages of each alternative can be evaluated and the options ranked. This is usually all one needs for decision making.

This type of use is very frequent in aquifer rehabilitation, where one has to choose among several alternatives, including the option of doing nothing.[3] Models are also used for supporting aquifer exploration policies, i.e., for answering questions such as “how much water can be extracted?,” “where should one pump to minimize environmental impact?,” etc. In fact, a large body of literature is devoted to this kind of questions in an optimal fashion.[4]

Site-specific models are most frequently used, however, as a tool to support aquifer characterization efforts. This is somewhat ironic because a model is an essentially quantitative tool while site characterization is rather qualitative. Yet, experience dictates that modeling is the only way to consistently integrate the kind of data available in site characterization. These data are very diverse and range from geologic maps to isotope concentrations. One can use vastly different models to verbally explain all observations. Quantitative consistency is not so easy to check and requires the use of a model. Because of the difficulties in fully describing all data, this kind of model use is rarely described in the scientific literature.

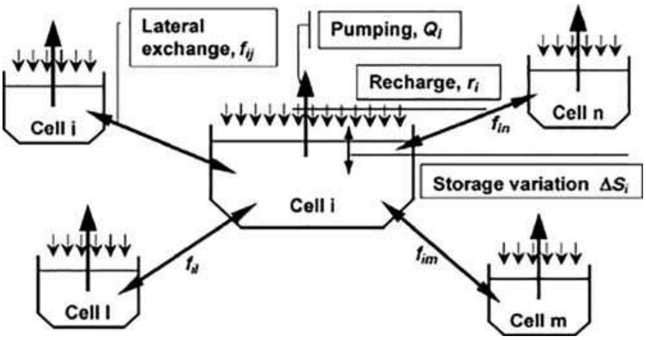


FIGURE 2 A groundwater model is the accounting system of an aquifer. It keeps track of the balance of each section (cells or compartments in the groundwater language) by evaluating exchanges with the outside (pumping Q_i, recharge, r_i, etc.) and with the adjacent sections (f_{ij}). The difference between inflow and outflow is equal to the variation in reserves (storage variation, ΔS_i,). A well-managed company needs an accounting system, and so does an aquifer.

Models can also be used in generic fashion as teaching or research tools to gain understanding on physicochemical phenomena. In these cases, they do not aim at representing a specific aquifer, but at evaluating the role of some processes under idealized conditions. A classical example of this type of use is the analysis of flow on regional basins.[5] Models are used in this fashion to explain geological processes.[6,7] Much emphasis has been placed in recent years on the evaluation of the effects of spatial variability. This involves issues such as upscaling, i.e., finding the relationship between large-scale effective parameters and small-scale measurements;[8] or analysis of hydraulic tests.[9]

How Are Models Built: The Modeling Process

The procedure to build a model is outlined in Figure 3. First, one defines a conceptual model (i.e., zones of recharge, boundaries of aquifers, etc.). Second, one discretizes the model domain into a finite element or finite difference grid. This can be entered as input data for a simulation code. Unfortunately, output data will rarely fit the observed aquifer heads and concentrations. This is what motivates calibration, i.e., the modification of model parameters to ensure that model output is indeed similar to what has been observed in reality. The model thus calibrated can be considered a "representation of the natural system" and can be used for management or simulation purposes.

The above procedure is formally described in Figure 4. This section is devoted to discussing in detail the modeling steps as previously described.[10–12] In practice, the effort behind each of these tasks may be very sensitive to the objectives of the studies and model. For the time being, we will assume that one is building a model aimed at describing reality in detail for the purpose making predictions.

Conceptualization

Modeling starts by defining which processes are important and how they are represented in the model. Definition of the relevant processes is termed "process identification" and it is needed for several reasons. First, the number of processes that may affect flow and transport is very large. For practical reasons, the modeler is forced to select those that affect the phenomenon under study, most significantly. Second, not all processes are well understood and they have to be treated in a simplified manner. In short, process identification involves simplifications, both in the choice of the processes and in the way they are implemented in the model.

Model structure identification refers to the definition of parameter variability, boundary conditions, etc. In a somewhat narrower but more systematic sense, model structure identification implies expressing the model in terms of a finite number of unknowns called model parameters. Parameters controlling the above processes are variable in space. In some cases, they also vary in time or depend on heads and/or concentrations. As discussed earlier, data are scarce so that such variability cannot be expressed accurately. Therefore, the modeler is also forced to make numerous simplifications to express the patterns of parameter variations, boundary conditions, etc. These assumptions are reflected on what is denoted as model structure.

The conceptualization step of any modeling effort is somewhat subjective and dependent on the modeler's ingenuity, experience, scientific background, and way of looking at the data. Selection of the physicochemical processes to be included in the model is only rarely the most difficult issue. The most important processes affecting the movement of water and solutes underground (advection, dispersion, sorption, etc.) are relatively well known. Ignoring a relevant process will only be caused by misjudgments and should be pointed out by reviewers, which illustrates why reviewing by others is important. Difficulties arise when trying to characterize those processes and, more specifically, the spatial variability of controlling parameters.

In spite of the large amount of data usually available, their qualitative nature prevents a detailed definition of the conceptual model. Thus, more than one description of the system may result from the conceptualization step. Selecting one conceptual model among several alternatives is sometimes performed during calibration, as discussed later.

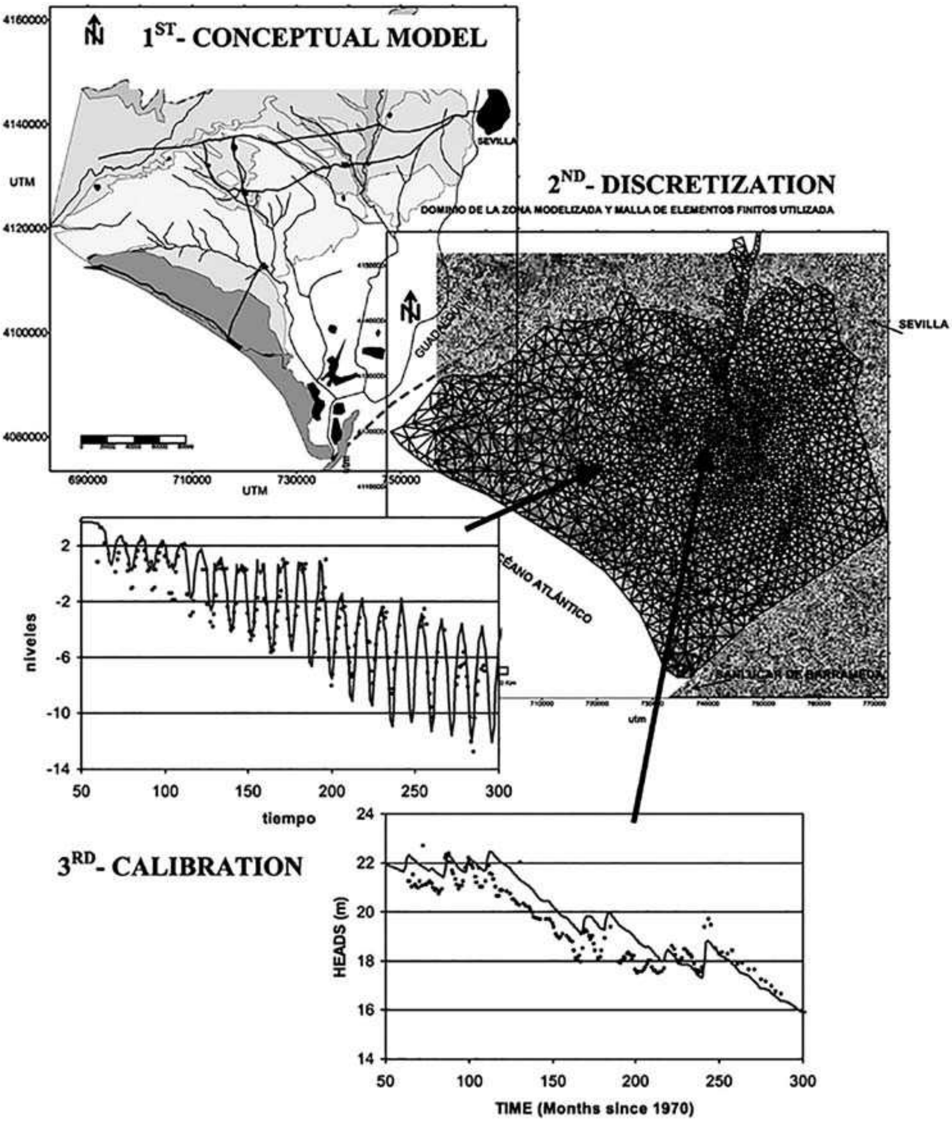


FIGURE 3 Building a model involves three basic steps: conceptualization, discretization, and calibration. Example from the Almonte-Marismas aquifer.

Discretization

Strictly speaking, discretization consists of substituting a continuum by a discrete system. However, we are extending this term here to describe the whole process of going from mathematical equations, derived from the conceptual model, to numerical expressions that can be solved by a computer. Closely related is the issue of verification, which refers to ensuring that a code accurately solves the equations that it is claimed to solve. As such, verification is a code-dependent concept. However, using a verified code is not sufficient for mathematical correctness. One should also make sure that time and space discretization is adequate for the problem being addressed. Moreover, numerical implementation of a conceptual model is not always straightforward international code comparison projects; INTRANCOIN

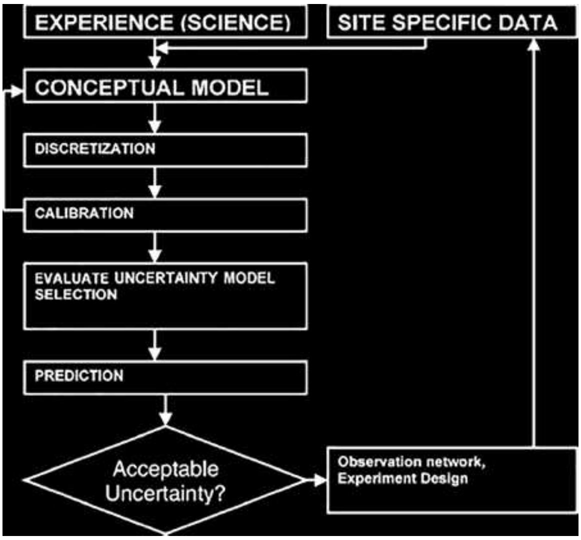


FIGURE 4 A formal description of the modeling process. Modeling starts with an understanding of the natural system (conceptual model), which is based on experience about such kind of systems (science) and on data from the site. Writing the conceptual model in a manner adequate for computer solution requires discretization. The resulting model is still dependent on many parameters that are uncertain. During calibration, these parameters are adjusted so that model outputs are close to measurements (recall Figure 3). Model predictions may be uncertain because so are the fitted parameters or because different models are consistent with observations. If uncertainty is unacceptably high, one should perform additional measurements or experiments and redo the whole process.

and HYDROCOIN have shown the need for sound conceptual models and independent checks of calculation results. Even well-posed mathematical problems lead to widely different solutions when solved by different people, because of slight variations in the solution methodology or misinterpretations in the formulation.[13] The reasons behind these differences and ways to solve them only become apparent after discussions among them.

The main concern during discretization is accuracy. In this sense, it is not conceptually difficult, although it can be complex. Accuracy is not only restricted to numerical errors (differences between numerical and exact solutions of the involved equations) but also refers to the precision with which the structure of spatial variability reproduces the natural system.

Calibration and Error Analysis

The choice of numerical values for model parameters is made during calibration, which consists of finding those values that grant a good reproduction of head and concentration data (Figure 3) and are consistent with prior independent information.

Calibration is rarely straightforward. Data come from various sources, with varying degrees of accuracy and levels of representativeness. Some parameters can be measured directly in the field, but such measurements are usually scarce and prone to error. Furthermore, since measurements are most often performed on scales and under conditions different from those required for modeling purposes, they tend to be both numerically and conceptually different from model parameters. The most dramatic example of this is dispersivity, whose representative value increases with the scale of measurement so that dispersivities derived from tracer tests cannot be used directly in a large-scale model. As a result, model parameters are calibrated by ensuring that simulated heads and concentrations are close to the corresponding field measurements.

Calibration can be tedious and time consuming because many combinations of parameters have to be evaluated, which also makes it prone to be incomplete. This, coupled to difficulties in taking into account the reliability of different pieces of information, makes it very hard to evaluate the quality of results. Therefore, it is not surprising that significant efforts have been devoted to the development of automatic calibration methods.[14–16]

Model Selection

The first step in any modeling effort involves constructing a conceptual model, describing it by means of appropriate governing equations, and translating the latter into a computer code. Model selection involves the process of choosing between alternative model forms. Methods for model selection can be classified into three broad categories. The first category is based on a comparative analysis of residuals (differences between measured and computed system responses) using objective as well as subjective criteria. The second category is denoted parameter assessment and involves evaluating whether or not computed parameters can be considered as "reasonable." The third category relies on theoretical measures of model validity known as "identification criteria." In practice, all three categories will be needed: residual analysis and parameter assessment suggest ways to modify an existing model and the resulting improvement in model performance is evaluated on the basis of identification criteria. If the modified model is judged an improvement over the previous model, the former is accepted and the latter discarded.

The most widely used tool of model identification is residual analysis. In the groundwater context, the spatial and frequency distributions of head and concentration residuals are very useful in pointing towards aspects of the model that need to be modified. For example, a long tail in the breakthrough curve not properly simulated by a single porosity model may point to a need for incorporating matrix diffusion or a similar mechanism. These modifications should, whenever possible, be guided by independent information. Qualitative data such as lithology, geological structure, geomorphology, and hydrochemistry are often useful for this purpose. A particular behavioral pattern of the residuals may be the result of varied causes that are often difficult to isolate. Spatial and/or temporal correlation among residuals may be a consequence of not only improper conceptualization, but also measurement or numerical errors. Simplifications in simulating the stresses exerted over the system are always made and they lead to correlation among residuals. Distinguishing between correlations caused by improper conceptualization and measurement errors is not an easy matter. This makes analysis of residuals a limited tool for model selection.

An expedite way of evaluating a model concept is based on assessing whether or not the parameters representing physicochemical properties can be considered "reasonable;" i.e., whether or not their values make sense and/or are consistent with those obtained elsewhere. Meaningless parameters can be a consequence of either poor conceptualization or instability. If a relevant process is ignored during conceptualization, the effect of such process may be reproduced by some other parameter. For example, the effect of sorption is to keep part of the solute attached to the solid phase, hence retarding the movement of the solute mass; in linear instantaneous sorption, this effect cannot be distinguished from standard storage in the pores. Therefore, if one needs an absent porosity (e.g., larger than one) to fit observation, one should consider the possibility of including sorption in the model. However, despite this example, parameter assessment tends to be more useful for ruling out some model concepts than for giving a hint on how to modify an inadequate model. Residual analysis is usually more helpful for this purpose.

Instability may also lead to unreasonable parameter estimates during automatic calibration, despite the validity of the conceptual model. When the number of data or their information content is low, small perturbations in the measurement or deviations in the model may lead to drastically different parameter estimates. When this happens, the model may obtain equally good fits with widely different parameter sets. Thus, one may converge to a senseless parameter set while missing other perfectly meaningful sets. This type of behavior can be easily identified by means of a thorough error analysis and corrected by fixing the values of one or several parameters.[14]

Predictions and Uncertainty

Formulation of predictions involves a conceptualization of its own. Quite often, the stresses, whose response is to be predicted, lead to significant changes in the natural system, so that the structure used for calibration is no longer valid for prediction. Changes in the hydrochemical conditions or in the flow geometry may have to be incorporated into the model. While numerical models can be used for network design or as investigation tools, most models are built in order to study the response of the medium to various scenario alternatives. Therefore, uncertainties on future natural and man-induced stresses also cause model predictions to be uncertain. Finally, even if future conditions and conceptual model are exactly known, errors in model parameters will still cause errors in the predictions. In summary, three types of prediction uncertainties can be identified: conceptual model uncertainties; stresses uncertainties; and parameters uncertainties.

The first group includes two types of problems. One is related to model selection during calibration. That is, more than one conceptual model may have been properly calibrated and data may not suffice to distinguish which one is the closest to reality. It is clear that such indetermination should be carried into the prediction stage because both models may lead to widely different results under future conditions. The second type of problems arises from improper extension of calibration to prediction conditions, i.e., from not taking into account changes in the natural system or in the scale of the problem. The only way we think about dealing with this problem consists of evaluating carefully whether or not the assumptions in which the calibration was based are still valid under future conditions. Indeed, model uncertainties can be very large.

We do not think that, strictly speaking, the second type of uncertainties, those associated with future stresses, falls in the realm of modeling. While future stresses may affect the validity of the model, they are external to it. In any case, this type of uncertainty is evaluated by carrying out simulations under a number of alternative scenarios, whose definition is an important subject in itself.

The last set of prediction uncertainties is the one associated with parameter uncertainties, which can be quantified quite well.

Conclusion

Groundwater modeling involves so many subjective decisions that it can be considered as an art. This is somewhat contrary to the widely accepted perception of models as something objective. The fact is that numerous assumptions need to be made both about the selection of relevant processes and about the manner of representing them in the computer. All these assumptions are specified in the conceptual model.

The result relies so heavily on conceptualization that models ought to be viewed as theories about the behavior of natural systems. Model predictions should rarely be viewed as firm statements about the future evolution of aquifers. Rather, they should be considered references against which actual data has to be compared. Codes do exist for modeling most processes affecting groundwater (flow, transport, reactions, thermomechanics, etc.). It is lack of understanding and lack of data what limits the actual application of those codes.

Having specified a conceptual model, the remaining steps (discretization, calibration, uncertainty analysis, prediction) are relatively objective, in the sense that systematic procedures can be followed. This explains why conceptualization is so important. It also explains why modeling is the best way of integrating widely different data. Uncertain as it is, it may represent unambiguously the overall knowledge of the aquifer.

Models represent the water balance (or solute balance, or energy balance) at the overall aquifer and at each of its parts. Therefore, they can also be viewed as accounting systems. It is argued that well-managed aquifers need real time models to help decision making, the same way that well managed companies need financial accounting systems. This is the challenge modelers must meet in the near future.

References

1. Neuman, S.P. Saturated–unsaturated seepage by finite elements. ASCE, J. Hydraulics Div. 1973, *99* (12), 2233–2250.
2. Abriola, L.M.; Pinder, G.F. A multiphase approach to the modeling of porous media contamination by organic compounds. Water Resour. Res. 1985, *21* (1), 11–26.
3. Konikow, L.F.; Bredehoeft, J.D. Groundwater models cannot be validated. Adv. Water Resour. 1992, *15* (1), 75–83.
4. Gorelick, S. A review of distributed parameter groundwater management modeling methods. Water Resour. Res. 1983, *19,* 305–319.
5. Freeze, R.A.; Witherspoon, P.A. Theoretical analysis of regional groundwater flow. II, effect of water table configuration and subsurface permeability variations. Water Resour. Res. 1966, *3,* 623–634.
6. Garven, G.; Freeze, R.A. Theoretical analysis of the role of groundwater flow in the genesis of strat-abound ore deposits. II, quantitative results. Am. J. Sci. 1984, *284,* 1125–1174.
7. Ayora, C.; Tabener, C.; Saaltink, M.W.; Carrera, J. The genesis of dedolomites: a discussion based on reactive transport modeling. J. Hydrol. 1998, *209,* 346–365.
8. Sánchez-Vila, X.; Girardi, J.; Carrera, J. A synthesis of approaches to upscaling of hydraulic conductivities. Water Resour. Res. 1995, *31* (4), 867–882.
9. Meier, P.; Carrera, J.; Sánchez-Vila, X. An evaluation of Jacob's method work for the interpretation of pumping tests in heterogeneous formations. Water Resour. Res. 1998, *34* (5), 1011–1025.
10. Carrera, J.; Mousavi, S.F.; Usunoff, E.; Sánchez-Vila, X.; Galarza, G. A discussion on validation of hydrogeological models. Reliab. Eng. Syst. Saf. 1993, *42*, 201–216.
11. Anderson, M.P.; Woessner, W.W. *Applied Groundwater Modeling;* Academic Press: San Diego, 1992.
12. Konikow, L.F.; Bredehoeft, J.D. *Computer Model of TwoDimensional Solute Transport and Dispersion in Groundwater: Reston,* VA; U.S. Geo. Surv. TWRI, Book 7, 1, 1978; 40 pp, Chap. C2.
13. NEA-SKI. *The International HYDROCOIN Project, Level 2: Model Validation;* OECD: Paris, France, 1990; 194 pp.
14. Carrera, J.; Neuman, S.P. Estimation of aquifer parameters under steady state and transient conditions: I through III. Water Resour. Res. **1986**, *22* (2), 199–242.
15. Loaiciga, H.A.; Marino, M.A. The inverse problem for confined aquifer flow: identification and estimation with extensions. Water Resour. Res. 1973, *23* (1), 92–104.
16. Hill, M.C. A Computer Program (MODFLOWP) for Estimating Parameters of a Transient, Three Dimensional, Groundwater Flow Model Using Nonlinear Regression, USGS Open-File Report, 1992; 91–484.

38

Groundwater: Numerical Method Modeling

Jesus Carrera

A Generic Numerical Method for Solving Groundwater Flow

As mentioned earlier, all methods require, first, discretizing and, second, approximating the physical phenomenon. For the generic method we are going to present here, discretization will be performed as shown in Figure 1. That is, the continuum aquifer domain will be substituted by a discrete number of cells. Furthermore, the continuum aquifer heads, $h(x,y)$, are substituted by a discrete number of model heads, h_i.

The second step, approximation, can be made in different manners. For the purpose of this section, it is sufficient to bear in mind that the flow equation is nothing but a mass balance. Therefore, we will express the mass balance in cell i as change in storage equals inflows minus outflows.

$$\Delta S_i = f_{ij} + f_{il} + f_{im} + f_{in} + g_i \tag{1}$$

where ΔS_i is the rate of change in storage during one time step (say, between time t^k and time t^{k+1}); f_{ij} is the inflow into cell i from cell j (and the same for f_{il}, f_{im}, and f_{in}); and g_i are external inflows into cell i (for example, recharge, minus pumping, minus evaporation, minus river outflow, etc.). Each of the terms in Eq. 1 is relatively easy to approximate. Storage variation can be derived from the definition of storage coefficient, (S is the change in volume of water stored per unit surface area of aquifer and per unit change in head):

$$\Delta S_i = SA_i \frac{\left(h_i^{k+1} - h_i^k\right)}{\Delta t} \tag{2}$$

where A_i is the surface area of cell i, h_i^k is the head in node i at time k and $\Delta t = t^{k+1} - t^k$ is the time step. Darcy's law gives lateral inflows

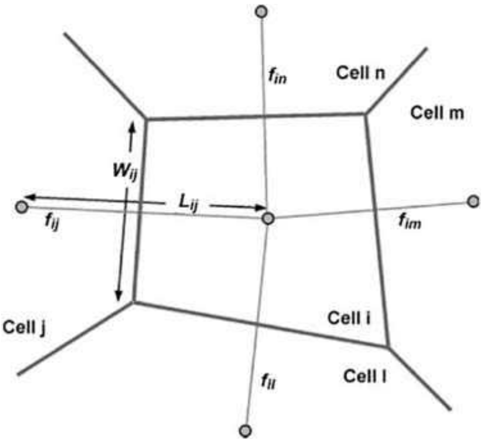


FIGURE 1 The system of equations for a generic groundwater model is obtained by establishing a water balance (inputs minus outputs equal storage variations) at each cell. Inputs and outputs include water exchanges with the outside (pumping Q_i, recharge, r_i, etc.) and with adjacent cells f_{ij}). The latter are expressed, using Darcy's law, as $f_{ij} = T_{ij} w_{ij} (h_j - h_i)/L_{ij}$.

$$f_{ij} = -Tw_{ij}\frac{h_i - h_j}{L_{ji}} = a_{ij}\left(h_i - h_j\right) \tag{3}$$

where T is transmissivity; w_{ij} is the width of the connection between nodes i and j; L_{ji} is the length of such connection; and a_{ij} is implicitly defined as $-Tw_{ij}/L_{ij.}$ The remaining inflow terms, f_{il}, f_{im}, and f_{in} are defined likewise. Changing these terms to the left-hand side of Eq. 1 and rearranging terms yield:

$$SA_i\frac{\left(h_i^{k+1} - h_i^k\right)}{\Delta t} + a_{ii}h_i + a_{ij}h_j + a_{il}h_l + a_{im}h_m + a_{in}h_n = g_i \tag{4}$$

where $a_{ii} = -a_{ij} - a_{il} - a_{im} - a_{in}$. If an equation like Eq. 4 is written for all cells from $i = 1$ through N, N being the number of cells (nodes), the resulting system of equations can be rewritten in matrix form as:

$$\mathbf{D}\frac{\left(\boldsymbol{h}^{k+1} - \boldsymbol{h}^k\right)}{\Delta \boldsymbol{t}} + \mathbf{Ah} = \mathbf{g} \tag{5}$$

where **D** is a diagonal matrix whose i-th diagonal term is precisely SA_i. This matrix is often called storage matrix. **A** is the conductance matrix, a square symmetric matrix whose components are a_{ij}. Finally, **g** is the source vector.

All the numerical methods to be outlined in subsequent sections lead to equations analogous to Eq. 5. Moreover, the meaning of the terms in such equations is always similar to that in Eq. 5. Namely, the system represents the mass balance at each of the N nodes (cells); specifically the i-th equation represents the mass balance at the i-th node. The first term, $\mathbf{D}(\mathbf{h}^{k+1} - \mathbf{h}^k)/\Delta t$, always represents storage variations. The second term, **Ah**, represents outflows from minus inflows into the i-th cell from the adjacent cells. Finally, term **g** represents external inflows minus outflows (recharge, pumping, etc.) at all i.

Equation 5 needs to be integrated in time. For this purpose, let us assume that **Ah** is evaluated at time $k + 1$ ($\mathbf{Ah}^{k+1}$). Then, Eq. 5 can be rewritten as:

$$\left(\boldsymbol{A} + \frac{\boldsymbol{D}}{\Delta t}\right)\boldsymbol{h}^{k+1} = \mathrm{g} + \frac{\boldsymbol{D}}{\Delta t}\boldsymbol{h}^k \tag{6}$$

This is simply a linear system, which can be solved using conventional methods.

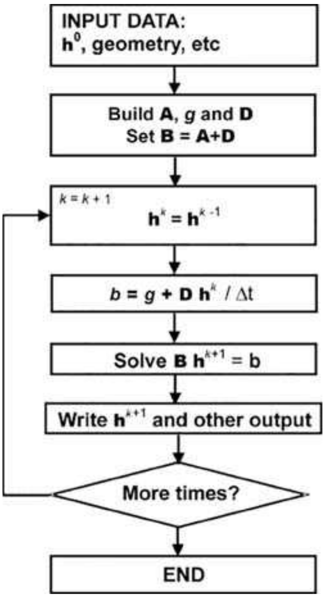


FIGURE 2 Basic steps involved in simulating groundwater flow. Heads $\mathbf{h}^{k+1}$ are computed by solving equation $\mathbf{Bh}^{k+1} = \mathbf{b}$. They may be written for later drawings. They are used as initial head for the next time increment. These steps (time loop) are repeated sequentially until the last time is reached.

$$\boldsymbol{Bh}^{k+1} = \mathbf{b} \tag{7}$$

where $\mathbf{B} = \mathbf{A} + \mathbf{D}/\Delta t$ and $\mathbf{b} = \mathbf{g} + \mathbf{D}\mathbf{h}^k/\Delta t$. This system is solved sequentially in time.

That is, most codes solve Eq. 7 using the following steps (Figure 2):

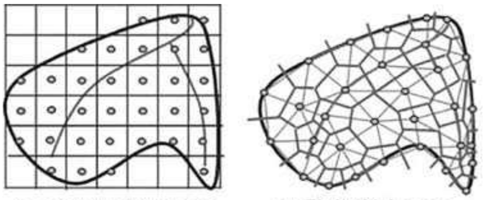
1. Input all data. Set $k = 0$
2. Compute **g**, **A** (Eq. 3); **D** (Eq. 5) and **B** (Eq. 7)
3. Set $k = k + 1$
4. Build **b** (Eq. 7)
5. Solve $\mathbf{Bh}^{k+1} = \mathbf{b}$
6. If $k = k_{\max}$ (maximum number of time steps), end. Otherwise, return to step 3

Most codes follow a structure such as this, although each method displays specific features. Some of these are outlined below.

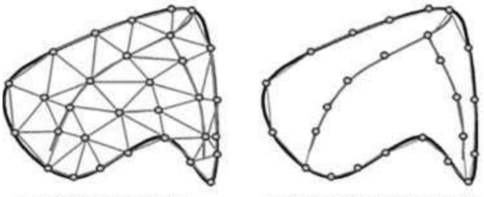
Finite Differences (FD)

As mentioned at the beginning, numerical methods differ in the way in which the domain is discretized and in the way in which the partial differential equation is transformed into a linear system of equations. In finite differences, the problem domain is discretized in a regular grid (Figure 3a), usually rectangular (equilateral triangles or hexagons are possible, but very rare). The grid may be centered at the corners (nodes are located at the vertices of the squares) or at the cells (nodes are located at the center of the squares, such as in Figure 3a).

Regarding the approximation of the partial differential equations, several alternatives are possible. The most intuitive consists of substituting all derivates by an incremental ratio. That is, the derivative between adjacent nodes *i* and *j* is approximated as:



(a) Finite Differences (b) Finite Volumes



(c) Finite Elements (d) Boundary Elements

FIGURE 3 The most widely spread methods of discretization are Finite Differences, which consists of subdividing the model domain into regular rectangles, and Finite Elements, which is based on dividing the aquifer region into elements of arbitrary shape (often triangles). Finite Volumes, also called Integrated Finite Differences, divides the region into polygons. The Boundary Element Method is very convenient, when applicable, because it only requires discretizing boundaries (both internal and external).

$$\frac{\partial h}{\partial x} = \frac{h_i - h_i}{\Delta x} \tag{8}$$

where h_i and h_j are heads at nodes i and j, respectively, and Δx is the distance between them. Approximating all derivatives by means of equations analogous to Eq. 8 leads to a system identical to Eq. 5. In fact, the finite differences method is often introduced using a mass balance approach such as the one in the section "A Generic Numerical Method for Solving Groundwater Flow," only using a regular instead of a generic grid. This is the method used in MOD- FLOW,[1] HST3D,[2] and their children.

Integrated Finite Differences (IFD)

The basic philosophy of this method is very similar to that of the generic method introduced in the generic numerical section. Basically, the domain is discretized in a number of cells centered around arbitrarily located nodes. Frequently, the cells are the Thiessen polygons of the set of nodes. This allows adapting the node density to the problem (e.g., increasing nodes density where accuracy is needed most).

Model equations can be derived using a mass balance approach, such as in the generic method section. Integrating the flow equation over each cell and applying Green's identity to transform volume integrals in boundary fluxes can also yield model equations. This type of approach is the basis of the finite volume method, which is widely used nowadays.

Finite Element Method

Finite element method (FEM) discretization consists of elements and nodes. Elements are generalized polygons (normally triangles or curvilinear quadrilaterals). Nodes are points located at the vertices and, sometimes, at the sides or the middle of the element. Unlike FD or IFD, cells around the nodes are not defined. Still, in many cases, one may write the equations in such a way that the mass balance formulation of the generic method section is still valid. However, the most singular feature of the FEM is the way the solution is interpolated, so that it becomes defined at every point. That is, head (or concentrations) is approximated as:

$$h(x) \cong \hat{h}(x) = \sum_{i} h_i N_i(x) \tag{9}$$

where h_i are nodal heads and N_i are interpolation functions. Since is not the exact solution, it would yield a residual if substituted in the flow equation. Minimizing this residual, which requires somewhat sophisticated maths, leads to a system similar to Eq. 5.

Boundary Element Method

The idea behind the boundary element method (BEM) is similar to that of the FEM. The main difference stems from the choice of interpolation functions, which are taken as the fundamental solutions of the flow equation (or whatever equation is to be solved). As a result, when the corresponding $\hat{h}$ is substituted in the flow equation, the residuals are zero. Since the equation is satisfied exactly in the model domain, one is only left with boundary conditions. In fact, as shown in Figure 3d, discretization is only required at the boundaries, where boundary heads are defined so as to satisfy approximately the boundary conditions. This method is extremely accurate, but its applicability is limited by the need of finding the fundamental solutions. This constrains the BEM to flow problems in relatively homogeneous domains.

Simulating Solute Transport

All the methods discussed above can be used for simulating solute transport. They are called Eulerian methods because they are based on a fixed (as opposed to moving) grid and all derivatives are based on a fixed coordinate system. They work fine when dispersion is dominant. Otherwise, they may lead to numerical problems (Figure 4). Two dimensionless numbers are used to anticipate numerical difficulties. Specifically, discretization must satisfy the following conditions.

Peclet number

$$Pe = \frac{v\Delta x}{D} \cong \frac{\Delta x}{\alpha} < \frac{1}{2} \tag{10}$$

Courant number

$$Co = \frac{v\Delta t}{\Delta x} < 1 \tag{11}$$

where v is the solute velocity, α is dispersivity, Δx is the distance between nodes and we have assumed that $D \cong \alpha v$. The condition on the Courant number implies that the solute at one node will not move

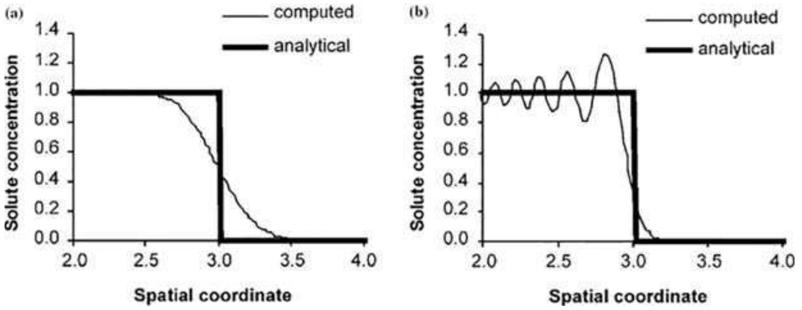


FIGURE 4 Difficulties typically associated to numerical simulation of solute transport: a) front smearing: the concentration front is more dispersed than it should; b) instability oscillations: too high and/or too low (even negative) concentrations.

beyond the following node downstream during one time step. This condition is easy to meet because usually groundwater moves slowly and also reducing Δt to satisfy Eq. 11 is not difficult. The condition on the Peclet number, Eq. 10, implies that Δx is smaller than $\alpha/2$, which may require very small elements, leading to a huge computational burden. Because of this, conventional Eulerian methods are not applicable to many groundwater problems, which have motivated the search of alternative methods.

Alternative methods can be Eulerian or Lagrangian. Among the former, the most popular is upstream weighting, introduced by Heinrich et al.[3] in the FEM, but with a huge number of papers thereafter. It consists of slightly modifying Eulerian equations so as to ensure stability. The problem is that, in doing so, it introduces numerical dispersion. As a result, the wiggles at the solute front in Figure 4 are substituted by an artificially smeared front. However, the vast majority of alternative formulations for solute transport are Lagrangian in the sense that time variations are written in terms of the material derivative, which expresses the rate of change in concentration of a particle that moves with the water. In this way, the advective term, which is the cause of problems in Eulerian formulations, disappears.

The number of Lagrangian methods is very large and many researchers have devoted much effort to find one, which is universal. The fact that so many methods have survived to date suggests that the effort has not been fully successful. Still, in practical problems, one can usually find a suitable method. Following is an outline of some of the most popular methods, with a discussion of their advantages and disadvantages and early references. The interested reader should seek further.

The most natural Lagrangian method is to write the equations on a moving grid, that is on a grid whose nodes move with water. This method is highly accurate, but expensive because the grid has to be updated every time step. Moreover, the grid can become highly deformed over time.

To avoid problems with moving grids it is frequent to work with particles. Displacing the particles with the moving water represents advection, while dispersion can be represented with a variety of methods. One such possibility is to add a random component to each particle basis displacement. This is statistically equivalent to each particle basis dispersion and is the basis of the “random walk” method.[4] The method requires careful implementation, but its main drawback is the fact that the solution is given in terms of number of particles per cell. If one is interested in spatial distributions of the solute, a huge number of particles may be needed.

The method of characteristics (MOC) overcomes the above difficulty by assigning concentrations to particles and interpolating them onto a fixed grid, where dispersion and, possibly, other transport processes are modeled. Concentrations are then interpolated back onto the particles. The method has become very popular in groundwater because of the USGS MOC Code.[5] The method is very practical, but the interpolation back-and-forth between particles and grid may introduce numerical dispersion and mass balance errors.

The modified method of characteristics[6] tries to overcome the problems associated to interpolating particle concentrations by redefining them in each time interval so that at the end of the time step they coincide with a node location. The method is very accurate, although some interpolation errors still occur when the front is abrupt. Some of these problems are overcome by the Eulerian–Lagrangian Localized Adjoint Method,[7] which looks as the most promising method.

Conclusion

Computer codes are available for simulating all phenomena affecting groundwater. In essence, they represent the balance of water (or salt, contaminants, or energy) in a manner that can be solved by the computer. This is achieved by, first, discretizing the problem and, second, rewriting it as a system of equations. This type of approach has been successful for flow problems. Solute transport, on the other hand, remains elusive. No single method is universally successful. Instead, one must seek the appropriate code in each case.

References

1. McDonald, M.G.; Harbaugh, A.W. A modular threedimensional finite difference groundwater flow model. *Techniques of Water Resources Inv.*, of the USGS, Book 6, 1988, Ch. A1.
2. Kipp, K.L. HST3D, A Computer Code for Simulation of Heat and Solute Transport in Three Dimensional Groundwater Flow Systems. USGS Water Recourses Inv. Rep. 86–4095; 1987.
3. Heinrich, J.C.; Huyakorn, P.S.; Mitchell, A.R.; Zienkie- wicz, O.C. An upwind finite element scheme for twodimensional transport equation. Int. J. Numer. Methods Eng. **1977**, *11*, 131–143.
4. Prickett, T.A.; Naymik, T.G.; Lonnquist, C.G. A "Random Walk" solute transport model for selected groundwater quality evaluations. Illinois State Water Survey Bull. **1981**, *55*, 62.
5. Konikow, L.F.; Bredehoeft, J.D. *Computer Model of TwoDimensional Solute Transport and Dispersion in Ground Water,* U.S. Geo. Surv. TWRI: Reston, Va., 1978; Book 7, 40 pp. Chap. C2.
6. Neuman, S.P. A Eulerian–lagrangian numerical scheme for the dispersion–convection equation using conjugate spacetime grids. J. Comput. Phys. **1981**, *41,* 270–294.
7. Celia, M.A.; Russell, T.F.; Herrera, I.; Ewing, R.E. An Eulerian–lagrangian localized adjoin method for the advection–diffusion equation. Adv. Water Resour. **1990**, *13,* 187–206.

39

Nitrogen (Nitrate Leaching) Index

Jorge A. Delgado

Introduction: Importance of Nitrogen

The United Nations recently reported that a changing climate will be one of the biggest challenges that humanity will be confronting this century, especially with respect to the potential impacts of climate change on food security.[1] One of the problems of climate change is the potential increase in extreme events.[2] Adapting to the occurrence of these extreme events will be critical since the potential for higher soil erosion is also significant.[3] Increased precipitation or higher amounts of irrigation increase the potential for nitrate leaching losses.[4–9] However, there is potential to use conservation management practices to reduce losses of nitrate leaching and adapt to a changing climate.[10] Humanity will need to implement the best management policies for soil and water conservation, including nutrient management, if it is to be able to adapt to a changing climate to help ensure that agricultural production is increased over the next decades and food security is achieved.[10,11]

With the projected global population increase of over two billion by 2050 and the challenges humanity faces as far as a changing climate, reduction of water resources, and more frequent occurrence of extreme events, there is a need to continue developing new management practices that help to increase nitrogen use efficiencies. This is because increases in grain and other agricultural products will require additional inputs of nitrogen, yet we do not want to increase the pressure in the metaphorical nitrogen pipes of these intensive agricultural systems, which could then potentially increase nitrogen leakages from these systems. Conservationists and nutrient managers across the world will be asking themselves questions about how to increase yields and nitrogen inputs while simultaneously minimizing the losses of reactive nitrogen to the environment, especially leaching losses of nitrate nitrogen, which can quickly move to water bodies in cases of shallow aquifers, tile systems, and surface water bodies. Increases in agricultural production will be necessary to supply grain for a larger world population, and larger quantities of grain and forage will be needed to meet the additional demand for meat and dairy products as large populations in growing economies such as China, and other regions increase in affluence.

Nitrogen management, including management of inorganic fertilizer inputs, management of animal manure and other by-products, and management of nitrogen across the environment from agricultural fields to non-agricultural areas across the watershed, will certainly be at the center of the world's emerging sustainability challenges in the coming decades. Independent of what policies are implemented, nitrogen management will be at the core of agricultural systems since most of the world's systems depend on agricultural inputs to maximize agricultural productivity.[10] As we move closer to the middle of the 21st century and we have more limited soil and water resources, we will need a global agricultural system that is both more productive and more sustainable. Best nitrogen management practices will be an essential component of efforts to increase productivity and sustainability.

Nitrogen Leaching Contributes to Environmental Problems

Although agricultural production has increased with the use of nitrogen fertilizer, nitrogen inputs also have contributed to increased losses of reactive nitrogen from agricultural systems that have contributed to global environmental impacts.[12–16] The US Environmental Protection Agency (EPA) has reported that drinking water with nitrate concentrations above 10 mg NO_3–N L^{-1} can negatively impact humans, and is thus unfit for human consumption.[17] It is important that we reduce the transport of reactive nitrogen from agricultural systems to reduce the negative impacts of reactive nitrogen on the environment and human health.[18,19] Several scientists have reported that nitrogen contributes to the hypoxic zone in the Gulf of Mexico.[20–25] There are hypoxic zones across the USA and around the world, and another example of an area impacted is Lake Erie, with its hypoxic zone and high microcystin levels.[26] Microcystin-LR concentrations above 1.0 μg/L in treated drinking water have been reported to be unsafe by the World Health Organization,[27] while the USEPA reported a much lower concentration of 0.3 μg/L to be unsafe for children less than six years old.[28] Microcystin concentrations above 0.3 μg/L can negatively impact children less than six years old.[28] Health effects from exposure could include gastroenteritis, and liver and kidney damage.[28]

It has been widely reported across the scientific literature that nitrogen losses could impact surface and groundwater quality.[29,30] Nitrate leaching is a major pathway for nitrogen loss that contributes to poorer water quality of aquifers across the world; examples of affected areas include regions of the North China Plain,[31] Europe,[15,32] South America,[33] and the USA.[29,30,34] The Midwest region of the USA has been reported to contribute large amounts of nitrate which is transported to the Gulf of Mexico, and tile systems have been reported as a major pathway for this nitrate mass that moves from agricultural fields to coastal systems.[20–23,34,25] Domestic wells that are sources of drinkable water across the USA have also been impacted by nitrate leaching as reported by Dubrovsky et al.[25] in his extensive study that found the nitrate levels were higher than the 10 mg NO_3–N L^{-1} safe limit established by the USEPA for a significant percentage of the wells.

Applying higher-than-needed N rates does not necessarily mean that all the pathways for nitrogen losses will be affected at the same magnitude, since site-specific properties affect the main mechanism for nitrogen losses. For example, in our adaptation of the metaphorical pipe concept, a higher-than-needed N rate for a coarser-textured soil or a soil system with tile drainage will have a bigger metaphorical pipe for nitrate leaching losses since coarser soils or tile drainage systems will have higher nitrate leaching potential. In other words, applying more nitrogen than needed to a coarser-textured soil or a tile system will most likely increase the pressure in the nitrate leaching pipe and the potential for NO_3–N leaching losses in this system. Another point to consider is that leaching is driven by water movement, so arid systems are less susceptible to leaching than humid systems with higher precipitation. Similarly, irrigated systems are also more susceptible to leaching, and water management practices (e.g., management of type of irrigation) can be used to reduce the leaching potential in irrigated systems. An example of a very sensitive system would be shallow-rooted crops grown in coarse-textured soils in flood-irrigated systems, which results in high potential for nitrate leaching. The good news is that we could use management practices (e.g., the inclusion of cover crops in a rotation with irrigated, shallow-rooted crops grown in coarse-textured soils) to try to reduce the pressure on the NO_3–N leaching pipe and the losses of nitrogen from the system. Randall et al., Delgado et al., and Delgado and Follett discussed in detail how best management practices could be used to reduce nitrate leaching to the environment.[5,35,36]

As previously discussed, nitrogen application rates that are higher than needed increase the flow through the metaphorical pipes to the environment, increasing the potential for nitrogen losses that will potentially impact air and water quality. However, best nitrogen management practices can be used to reduce the pressure in the pipes and minimize the leakage. Practices that will increase denitrification, contributing to conversion of NO_3–N as it is transported across water bodies to N_2 (e.g., wood chips, wetlands), will minimize the environmental impact of reactive nitrogen on the environment. Although this is a beneficial approach environmentally, there is still an economic loss due to the losses of NO_3–N from field applications, and increasing nitrogen use efficiencies at a field level while reducing NO_3–N leaching losses from the field will be a way to reduce the environmental impact and reduce field-level economic losses. Aside from this consideration, best management practices such as denitrification traps will be good practices to reduce the transport of NO_3–N to the environment.

For global sustainability, we could use many different management practices to increase nitrogen use efficiencies and yields while reducing losses of reactive nitrogen to the environment across global agricultural systems. Quick nitrogen tools that can use some general principles of soil hydrology and soil chemical and physical properties while accounting for irrigation and precipitation balances can be used to assess practices for different cropping systems on an annual basis. New, advanced tools such as the Nitrogen Index are among the new ecotechnologies that can be applied.[35–37]

New Ecotechnology: Quick Tools/Indicators

There are a variety of indexes that have been used to assess the risk of nitrate leaching.[38] Shaffer and Delgado[39] proposed a nitrate leaching index that considers the effects of management (e.g., manure, crops, fertilizer, irrigation), soils, climate, and off-site factors; that is easy to use; and that can conduct a quick but strong assessment of the risk of nitrate leaching. They discussed the different indexes that have been developed in the past to assess the risk of nitrate leaching and the advantages and disadvantages of these indexes. Delgado et al. [35,40–44] developed such an index considering hydrological and ecological factors. This new ecotechnology has been used for about the last decade by different US agencies and other national and international institutions to assess the risk of nitrate leaching for a given set of management practices under a given set of weather, soil, and off-site conditions. The tool is available for download at the website of the National Agricultural Library, and the code of the Nitrogen Index is available to download as well (https://data.nal.usda.gov/dataset/cce-nitrogen-index-tool/resource/ec68e0b2-f3e6-4adc-88f4-aee430f6603d).

In addition to the Nitrogen Index, there are other indexes that have been used to assess nitrogen use efficiencies. Shaffer and Delgado[39] and Buczko et al.[38,45] discussed the advantages and disadvantages of several nitrogen indexes approaches. Assessment of the risk of nitrate leaching potential has been conducted in New York with the New York Nitrate Leaching Index[46] and in California with the Nitrate Leaching Hazard Index.[47] However, none of these indexes conduct a quick assessment of the mass of nitrate leaching at different root depths, and the new California Nitrogen Index and new generic Nitrogen Index can be used to conduct a quick assessment of the mass of nitrate leaching at different root depths.[35,40] One of the indexes used to assess nitrate leaching is the water Leaching Index (LI) by Williams and Kissel[48] that assesses the potential for water leaching and extrapolates that risk to nitrate leaching. However, since there is no assessment of the mass of nitrate leaching or the risk of nitrate leaching, this water leaching index is at a disadvantage in assessing the risk since it does not consider N inputs and N dynamics.[39]

Delgado[49–51] studied the movement of nitrate across the soil profile to assess the nitrate leaching potential of cropping systems with different root depths and found a significant correlation between the root depths of cropping systems and nitrate leaching, with a higher nitrate leaching potential for crops with more shallow root systems. Similarly, he assessed the nitrate leaching potential across different soil textures and nitrate leaching spatially across fields and found a significant correlation between soil texture and nitrate leaching potential when fields were managed similarly, with a higher nitrate leaching potential for the areas of the field with coarser-textured soils.[52]

Sharpley et al.[53–55] and Lemunyon and Gilbert et al.[56] reported about the need to join the Phosphorus Index with a nitrogen index to conduct a multi-nutrient assessment of the risk of losses of nitrogen and phosphorus to the environment. Heathwaite et al.[57] developed a joint nitrogen and phosphorus index. OMAFRA[58] also proposed a joint index. Similarly, Delgado et al.[35] proposed to join the new Nitrogen Index to currently available P indexes. Bolster et al.[59] developed a nitrogen and a phosphorus index analysis for USDA-NRCS in Kentucky.

De Paz[15] tested the new Nitrogen Index approach joined with GIS capabilities across one of the regions of Europe most impacted by nitrate leaching.[15,32] De Paz et al.[15] found a correlation between the nitrate leaching and water leaching predicted by the Nitrogen Index, and measured nitrate leaching and water leaching. They also found that the nitrate leaching predicted by the Nitrogen Index was significantly correlated with the measured nitrate leaching across this Mediterranean region and that the areas with higher groundwater nitrate concentrations were correlated with the areas estimated to have higher nitrate leaching potential by the Nitrogen Index.

These results from De Paz et al. are another example indicating that the hole-in-the-pipe model proposed by Firestone and Davidson[60] describing the losses via emissions of greenhouse gases from the soil in the nitrogen cycle can be applied to the whole nitrogen cycle in agricultural systems, as described in the present work, and that the application of higher nitrogen rates than needed increases the pressure in the system, resulting in increased leaking from the system, and thus higher concentrations of nitrate, impacting groundwater quality.

Figueroa et al.[61] found that in intensive manure systems in Mexico that received large manure applications, the risk of nitrate leaching was higher. The tool was used to assess the effects of manure management practices on residual soil nitrate in the profile.[61] The Instituto Nacional de Investigaciones Forestales Agricolas y Pecuaris (INIFAP), a research institution in Mexico, conducted studies in Mexico and reported that the Nitrogen Index was a tool that can be used to assess the nitrate leaching risk in forage systems in this country.[42] Saynes et al.[16] were also able to use the Nitrogen Index to assess the effects of management practices in Mexico on emissions of N_2O from these fertilized systems. This is another example that this simple tool can be used to validate the hole-in-the-pipe model proposed by Firestone and Davidson[60] as adapted in the present work in international systems since the emissions of N_2O increased with higher fertilizer rates.[16]

Conclusion

The Nitrogen Index has been tested and calibrated using data from North America,[4,35,43,44,59,62] different regions in Mexico,[16,41,42,61,63,64] the Caribbean[65] region, different regions of South America[35,66,67] (Andean region and Pampas region), a Mediterranean region,[15] and the North China Plain,[35] where personnel from various national and international agencies, universities, and/or research centers have used the tool. Results of these efforts show that the Nitrogen Index is able to evaluate the effect of management on residual soil nitrate, water leaching, nitrate leaching, emissions of nitrous oxide, crop nitrogen use efficiencies, and crop nitrogen uptake. Additionally, measured groundwater nitrate concentrations have also been correlated with nitrate leaching predicted by the Nitrogen Index, validating the potential of using the tool to assess the risk of nitrate leaching to the environment.

Although it is recommended that users get familiar with the capabilities and limitations of the Nitrogen Index, the correlation between predicted and observed values for these different losses of reactive nitrogen suggests that this is a robust tool capable of being used to assess the risk of management practices on potential nitrate leaching and losses via other pathways. These results also support using the Nitrogen Index as part of an ecological engineering approach to soil and water conservation planning.[35,43] There is potential to use this tool to assess the effects of nitrogen inputs from manure, fertilizer and other organic sources, and management practices, and to rank the risk of losses of reactive nitrogen, including the nitrate leaching pathways, as very low, low, medium, high, or very high risk.

The Nitrogen Index can potentially be used as an indicator to assess the risk of nitrate leaching to the environment, helping conservation practitioners and nutrient managers implement best management practices to increase nutrient use efficiencies while reducing the nitrate leaching risk and losses of reactive nitrogen through different pathways. This tool is a new approach and is available for download, along with the code for the tool. The Nitrogen Index shows promise for being used as a quick and effective assessment tool for global sustainability.

References

1. IPCC (Intergovernmental Panel on Climate Change). IPCC Special Report on Global Warming of 1.5: Summary for Policymakers. 2018. https://www.ipcc.ch/2018/10/08/summary-for-policymakers-of-ipcc-special-report-on-global-warming-of-1-5c-approved-by-governments/.
2. Lal, R.; Delgado, J.A.; Gulliford, J.; Nielsen, D.; Rice, C.W.; Van Pelt, R.S. Adapting agriculture to drought and extreme events. *J. Soil Water Conserv.* **2012**, *67* (6), 162A–166A.
3. Nearing, M.A.; Pruski, F.F.; O'Neal, M.R. Expected climate change impacts on soil erosion rates: A review. *J. Soil Water Conserv.* **2004**, *59* (1), 43–50.
4. Marchi, E.C.S.; Zotarelli, Lincoln; Delgado, J.A.; Rowland, D.L. Use of the nitrogen index to assess nitrate leaching and water drainage from plastic-mulched horticultural cropping systems of Florida. *Int. Soil Water Conserv. Res.* **2016**, *4* (4), 237–244.
5. Randall, G.W.; Delgado, J.A.; Schepers, J.S. Nitrogen management to protect water resources. In *Nitrogen in Agricultural Systems*; Schepers, J.S., Raun, W.R., Follett, R.F., Fox, R.H., Randall, G.W., Eds.; Agronomy Monograph 49; American Society of Agronomy: Madison, WI, 2008; 911–946.
6. Randall, G.W.; Goss, M.J.; Fausey, N.R. Nitrogen and drainage management to reduce nitrate losses to subsurface drainage. In *Advances in Nitrogen Management for Water Quality*; Delgado, J.A., Follett, R.F., Eds.; Soil and Water Conservation Society: Ankeny, IA, 2010; 61–93.
7. Pierce, F.J.; Shaffer, M.J.; Halvorson, A.D. Screening procedure for estimating potentially leachable nitrate-nitrogen below the root zone. In *Managing Nitrogen for Groundwater Quality and Farm Profitability*; Follett, R.F., Keeny, D.R., Cruse, R.M., Eds.; Soil Science Society of America: Madison, WI, 1991; 259–283.
8. Van Es, H.M.; Delgado, J.A. Nitrate leaching index. In *Encyclopedia of Soil Science*; Lal, R., Ed.; Markel and Decker: New York, 2006; 1119–1121.
9. Delgado, J.A. Quantifying the loss mechanisms of nitrogen. *J. Soil Water Conserv.* **2002**, *57*, 389–398.
10. Delgado, J.A.; Groffman, P.M.; Nearing, M.A.; Goddard, T.; Reicosky, D.; Lal, R.; Kitchen, N.; Rice, C.; Towery, D.; Salon, P. Conservation practices to mitigate and adapt to climate change. *J. Soil Water Conserv.* **2011**, *66*, 118A–129A.
11. Delgado, J.A. The Nanchang Communication about the potential for implementation of conservation practices for climate change mitigation and adaptation to achieve food security in the 21st century. *Int. Soil Water Conserv. Res.* **2016**, *4* (2), 148–150.
12. Delgado, J.A.; Gagliardi, P.; Shaffer, M.J; Cover, H.; Hesketh, E.; Ascough, J.C.; Daniel, B.M. New tools to assess nitrogen management for conservation of our biosphere. In *Advances in Nitrogen Management for Water Quality*; Delgado, J.A., Follett, R.F., Eds.; SWCS: Ankeny, IA, 2010; 373–409.
13. Cowling, E.; Galloway, J.; Furiness, C.; Erisman, J.W., et al. *Optimizing nitrogen management and energy production and environmental protection: Report from the Second International Nitrogen Conference*, Bolger Center, Potomac, MD, Oct 14–18, 2001; 2002, available at http://www.initrogen.org/fileadmin/user_upload/Second_N_Conf_Report.pdf (accessed May 2010).
14. Galloway, J.N.; Aber, J.D.; Erisman, J.W.; Seitzinger, S.P.; Howarth, R.W.; Cowling, E.B.; Cosby, B.J. The nitrogen cascade. *BioScience* **2003**, *53* (4), 341–356.

15. De Paz, J.M.; Delgado, J.A.; Ramos, C.; Shaffer, M.; Bar-barick, K. Use of a new nitrogen index-GIS assessment for evaluation of nitrate leaching across a Mediterranean region. *J. Hydrol.* **2009**, *365,* 183–194.
16. Saynes, V.; Delgado, J.A.; Tebbe, C.; Etchevers, J.D., Lapidus, Otero-Arnaiz, A. Use of the new Nitrogen Index tier zero to assess the effects of nitrogen fertilizer on N2O emissions from cropping systems in Mexico. *Ecol. Eng.* **2014**, *73*, 778–785.
17. USEPA (U.S. Environmental Protection Agency). National Primary Drinking Water Regulations. n.d. https://www.epa.gov/ground-water-and-drinking-water/national-primary-drinking-water-regulations.
18. Follett, J.R.; Follett, R.F.; Herz, W.C. Environmental and human impacts of reactive nitrogen. In *Advances in Nitrogen Management for Water Quality*; Delgado, J.A., Follett, R.F., Eds.; Soil and Water Conservation Society: Ankeny, IA, 2010; 1–37.
19. Follett, R.F.; Walker, D.J. Groundwater quality concerns about nitrogen. In *Nitrogen Management and Groundwater Protection;* Follett, R.F., Ed.; Elsevier Sci. Pub.: Amsterdam, 1989; 1–22.
20. Turner, R.E.; Rabalais, N.N. Linking landscape and water quality in the Mississippi River Basin for 200 years. *BioScience* **2003**, *53* (6), 563–572.
21. Mitsch, W.J.; Day, J.W. Restoration of wetlands in the Mississippi–Ohio–Missouri (MOM) River Basin: Experience and needed research. *Ecol. Eng.* **2006**, *26,* 55–69.
22. Rabalais, N.N.; Turner, R.E.; Wiseman, W.J., Jr. Gulf of Mexico hypoxia, a.k.a. "The Dead Zone." *Annu. Rev. Ecol. Syst.* **2002**, *33*, 235–263.
23. Rabalais, N.N.; Turner, R.E.; Scavia, D. Beyond science into policy: Gulf of Mexico hypoxia and the Mississippi River. *BioScience* **2002**, *52* (2), 129–142.
24. Goolsby, D.A.; Battaglin, W.A.; Aulenbach, B.T.; Hooper, R.P. Nitrogen input to the Gulf of Mexico. *J. Environ. Qual.* **2001**, *30,* 329–336.
25. Dubrovsky, N.M.; Burow, K.R.; Clark, G.M.; Gronberg, J.A.M.; Hamilton, P.A.; Hitt, K.J.; Mueller, D.K.; Munn, M.D.; Puckett, L.J.; Nolan, B.T.; Rupert, M.G.; Short, T.M.; Spahr, N.E.; Sprague, L.A.; Wilbur, W.G. *Nutrients in the Nation's Streams and Groundwater, 1992–2004, Circular 1350*; U.S. Geological Survey: Reston, VA, 2010.
26. Smith, D.R.; Wilson, R.S.; King, K.W.; Zwonitzer, M.; McGrath, J.M.; Harmel, R.D.; Haney, R.L.; Johnson, L.T. Lake Erie, phosphorus, and microcystin: Is it really the farmer's fault?. *J. Soil Water Conserv.* **2018**, *73*, 48–57.
27. World Health Organization (WHO) (2011.) *Guidelines for Drinking-Water Quality*. 4th Ed. WHO Press: Geneva, Switzerland.
28. USEPA (United States Environmental Protection Agency). 2015 Drinking Water Health Advisories for Two Cyanobacterial Toxins. Office of Water 820F15003, June 2015, Washington, DC. https://www.epa.gov/sites/production/files/2017-06/documents/cyanotoxins-fact_sheet-2015.pdf.
29. Follett, R.F.; Delgado, J.A. Nitrogen fate and transport in agricultural systems. *J. Soil Water Conserv.* **2002**, *57* (6), 402–408.
30. Hatfield, J.L.; Follett, R.F. *Nitrogen Management in the Environment. Sources Problems, and Management*, 2nd Ed.; Elsevier Science Publ.: Amsterdam, 2008.
31. Li, X.; Hu, C.; Delgado, J.A.; Zhang, Y.; Ouyang, Z. Increased nitrogen use efficiencies as a key mitigation alternative to reduce nitrate leaching in North China Plain. *Agric. Water Manage.* **2007**, *89,* 137–147.
32. Juergens-Gschwind, S. Ground water nitrates in other developed countries (Europe)—Relations to land use patterns. In *Nitrogen Management and Groundwater Protection*; Follett, R.F., Ed., Elsevier Science Publishers: New York, 1989; 75–138.
33. Lavado, R.S.; de Paz, J.M.; Delgado J.A.; Rimski-Korsakov, H. Evaluation of best nitrogen management practices across regions of Argentina and Spain. In *Advances in Nitrogen Management for Water Quality*; Delgado, J.A., Follett, R.F., Eds.; Soil and Water Conservation Society: Ankeny, IA, **2010**; 313–342.

34. Rupert, M.G. Decadal-scale changes of nitrate in ground water of the United States, 1988–2004. *J. Environ. Qual.* **2008**, *37*, S240–S248.
35. Delgado, J.A.; Shaffer, M.; Hu, C.; Lavado, R.; Cueto Wong, J.A.; Joosse, P.; Sotomayor, D.; Colon, W.; Follett, R; Del Grosso, S.; Li, X.; Rimski-Korsakov, H. An index approach to assess nitrogen losses to the environment. *Ecol. Eng.* **2008**, *32*, 108–120.
36. Delgado, J.A.; Follett, R.F., Eds. *Advances in Nitrogen Management for Water Quality*; SWCS: Ankeny, IA, 2010.
37. Delgado, J.A.; Shaffer, M.J.; Lal, H.; McKinney, S.; Gross, C.M.; Cover, H. Assessment of nitrogen losses to the environment with a Nitrogen Trading Tool (NTT). *Comput. Electron. Agric.* **2008**, *63*, 193–206.
38. Buczko, U.; Kuchenbuch, R.O.; Lennartz, B. Assessment of the predictive quality of simple indicator approaches for nitrate leaching from agricultural fields. *J. Environ. Manage.* **2010**, *91*, 1305–1315.
39. Shaffer, M.J.; Delgado, J.A. Essentials of a national nitrate leaching index assessment tool. *J. Soil Water Conserv.* **2002**, *57*, 327–335.
40. Delgado, J.A.; Shaffer, M.; Hu, C.; Lavado, R.S.; Cueto Wong, J.; Joosse, P.; Li, X.; Rimski-Korsakov, H.; Follett, R.; Colon; W.; Sotomayor, D. A decade of change in nutrient management requires a new tool: A new nitrogen index. *J. Soil Water Conserv.* **2006**, *61*, 62A–71A.
41. Figueroa Viramontes, U., Delgado, J.A., Cueto Wong, José. Índice de Nitrógeno Ver. 4.4 Adaptado para la Producción de Forrajes en México. Manual del Usuario. (In Spanish.) Mexico Nitrogen Index User Manual. Instituto Nacional de Investigaciones Forestales, Agrícolas, y Pecuarias (INIFAP) and Secretaría de Agricultura, Ganadería, Desarrollo Rural, Pesca y Alimentación (SAGARPA), Matamoros, Coahuila, Mexico, 2011.
42. Instituto Nacional de Investigaciones Forestales Agricolas y Pecuaris (INIFAP). Reporte Anual 2009 Ciencia y Tecnología para el Campo Mexicano Instituto Nacional de Investigaciones Forestales, Agrícolas y Pecuarias. Reporte Anual 2009 México, D. F. Publicación Especial Núm. 5; ISBN 978-607-425-316-0: Mexico, 2010; 31 pp.
43. Gross, C.M.; Delgado, J.A.; Shaffer, M.J.; Gasseling, D.; Bunch, T.; Fry, R. A tiered approach to nitrogen management: A USDA perspective. In *Advances in Nitrogen Management for Water Quality*; Delgado, J.A., Follett, R.F., Eds.; Soil and Water Conservation Society: Ankeny, IA, 2010; 410–424.
44. USDA. National NRCS Nutrient Management Tools. https://www.nrcs.usda.gov/wps/portal/nrcs/detail/national/nwmc/partners/?cid=nrcs143_015091.
45. Buczko, U.; Kuchenbuch, R.O. Environmental indicators to assess the risk of diffuse nitrogen losses from agriculture. *Environ. Manage.* **2010**, *45*, 1201–1222.
46. Van Es, H.M.; Czymmek, K.J.; Ketterings, Q.M. Management effects on nitrogen leaching and guidelines for a nitrogen leaching index in New York. *J. Soil Water Conserv.* **2002**, *57*, 499–504.
47. Wu, L.; Letey, J.; French, C.; Wood, Y.; Bikie, D. Nitrate leaching hazard index developed for irrigated agriculture. *J. Soil Water Conserv.* **2005**, *60*, 90A–95A.
48. Williams, J.R.; Kissel, D.E. Water percolation: An indicator of nitrogen-leaching potential. In *Managing Nitrogen for Groundwater Quality and Farm Profitability*; Follett, R.F., Keeney, D.R., Cruse, R.M., Eds.; Soil Science Society of America: Madison, WI, 1991; 59–83.
49. Delgado, J.A. Sequential NLEAP simulations to examine effect of early and late planted winter cover crops on nitrogen dynamics. *J. Soil Water Conserv.* **1998**, *53*, 241–244.
50. Delgado, J.A. Use of simulations for evaluation of best management practices on irrigated cropping systems. In *Modeling Carbon and Nitrogen Dynamics for Soil Management*; Shaffer, M.J., Ma, L., Hansen, S., Eds.; Lewis Publishers: Boca Raton, FL, 2001; 355–381.
51. Delgado, J.A.; Shaffer, M.; Brodahl; M.K. New NLEAP for shallow and deep rooted rotations: Irrigated agriculture in the San Luis Valley of south central Colorado. *J. Soil Water Conserv.* **1998**, *53* (4), 338–340.

52. Delgado, J.A.; Bausch, W.C. Potential use of precision conservation techniques to reduce nitrate leaching in irrigated crops. *J. Soil Water Conserv.* **2005**, *60*, 379–387.
53. Sharpley, A.N.; Daniel, T.; Sims, T; Lemunyon, J.; Stevens, R.; Parry, R. *Agricultural Phosphorus and Eutrophication*, U.S. Department of Agriculture Agricultural Research Service No. ARS-149; 1999.
54. Sharpley, A.N.; Kleinman, P.; McDowell, R. Innovative management of agricultural phosphorous to protect soil and water resources. *Commun. Soil Sci. Plant Anal.* **2001**, *32* (7 & 8), 1071–1100.
55. Sharpley, A.N.; Weld, J.L.; Beegle, D.B.; Kleinman, P.J.A.; Gburuk, W.J.; Moore, P.A., Jr.; Mullins, G. Development of phosphorus indices for nutrient management planning strategies in the United States. *J. Soil Water Conserv.* **2003**, *53,* 137–151.
56. Lemunyon, J.L.; Gilbert, R.G. The concept and need for a phosphorus assessment tool. *J. Prod. Agric.* **1993**, *6,* 483–486.
57. Heathwaite, L.; Sharpley, A.; Gburek, W. A conceptual approach for integrating phosphorous and nitrogen management at watershed scales. *J. Environ. Qual.* **2000**, *29*, 158–166.
58. Ontario Ministry of Agriculture, Food and Affairs (OMAFRA). Fundamentals of Nutrient Management Training Course; Queen's Printer of Ontario: Ontario, Canada, 2005. Available at http://www.omafra.gov.on.ca/english/nm/nman/default.htm (accessed May 2011).
59. Bolster, C.H.; Horvath, T.; Lee, B.D.; Mehlhope, S.; Higgins, S.; Delogado, J.A. Development and testing of a new phosphorus index for Kentucky. *J. Soil Water Conserv.* **2014**, *69* (3), 183–196.
60. Firestone, M.K.; Davidson, E.A. Microbiological basis of NO and N2O production and consumption in soil. In *Exchange of Trace Gases between Terrestrial Eco systems and the Atmosphere;* Andreae, M.O.; Schimel, D.S., Eds. John Wiley & Sons: New York, 1989; 7–21.
61. Figueroa, V.U.; Núñez, H.G.; Delgado, J.A.; Cueto, J.A.; Flores, M.J.P. Estimación de la producción de estiércol y de la excreción de nitrógeno, fósforo y potasio por bovino lechero en la Comarca Lagunera. In *Agricultura Orgánica*, 2nd Ed.; Orona et al., Eds.; Sociedad Mexicana de la Ciencia del Suelo, Consejo Nacional de Ciencia y Tecnología, and Universidad Autónoma del Estado de Durango-Facultad de Agricultura y Zootecnia: Gomez Palacio, 2009; 128–151, 425.
62. Delgado, J.A.; Gagliardi, P.M.; Rau, E.J.; Fry, R.; Figueroa, U.; Gross, C.; Cueto-Wong, J.; Shaffer, M.; Kowalski, K.; Neer, D.; Sotomayor-Ramirez, D.; Alwang, J.; Monar, C.; Escudero, L. and Saavedra-Rivera. A.K. Nitrogen Index 4.4 User Manual. USDA-ARS-SPNR, Fort Collins, CO. 2011 (USDA-ARS-SPNR & NRCS User Manual; available at official USDA ARS SPNR. http://www.ars.usda.gov/npa/spnr/nitrogentools).
63. Figueroa-Viramontes, U.; Delgado, J.A.; Cueto-Wong, J.A.; Núnez-Hernández, G.; Reta-Sánchez, D.G.; Barbarick, K.A. A new Nitrogen Index to evaluate nitrogen losses in intensive forage systems in Mexico. *Agric. Ecosyst. Environ.* **2011**, *142*, 352–364.
64. Figueroa-Viramontes, U.; Delgado, J.A.; Sánchez-Duarte, J.I.; Ochoa-Martínez, E.; Núñez-Hernández, G. A nitrogen index for improving nutrient management within commercial Mexican dairy operations. *Int. Soil Water Conserv. Res.* **2016**, *4*, 1–5.
65. Oliveras-Berrocales, M.; Delgado, J.A.; Pérez-Alegría, L.R. The new Caribbean Nitrogen Index to assess nitrogen dynamics in vegetable production systems in southwestern Puerto Rico. *Int. Soil Water Conserv. Res.*, **2017**, *5* (1), 69–75.
66. Escudero, L.; Delgado, J.A.; Monar, C.; Valverde, F.; Barrera, V.; Alwang, J. A new nitrogen index for assessment of nitrogen management of Andean mountain cropping systems of Ecuador. *Soil Sci.* **2014**, *179*, 130–140.
67. Saavedra, A.K.; Delgado, J.A.; Botello, R.; Mamani, P.; Alwang, J. A new N Index to assess nitrogen dynamics in potato (*Solanum tuberosum* L.) production systems of Bolivia. *Agrociencia*, **2014**, *48*, 667–678.

40

Nitrogen (Nutrient) Trading Tool

Jorge A. Delgado

Importance of Nitrogen

Nitrogen is an essential nutrient with very important functions for life; it is a component of nucleic acids, the building blocks of genetic information encoded in DNA, and protein, the building blocks of biological tissues such as muscles or hair. As evidenced by its contribution to increased yields of cropping systems during the 20th century, nitrogen is also important for food security and therefore global stability and peace. Almost all global cropping systems respond to nitrogen inputs with higher yields and economic returns. When its management is combined with conservation practices, the results are very positive; for example, just recently Delgado et al. reported that in systems implementing conservation agriculture in the Andean region of South America, nitrogen inputs increased profits for farmers by 22%, and that these improved best management practices of fertilizer application and conservation agriculture could increase the profits for 200,000 small farmers.[1]

The ever-increasing global population creates additional pressure to increase agricultural intensification to achieve the needed increase in agricultural production needed to feed 2 billion additional people by 2050. The success of the green revolution has contributed to increased global production of grains and other crops such as potatoes, as well as increased yields per area during the last 100 years. [2] Similarly, the total use of nitrogen and nutrients worldwide has also been increasing as the human population and fertilizer needed to increase total global production also increase.[3] It is expected that the use of nitrogen and other nutrients will continue to increase during the 21st century; thus, management of nitrogen and other nutrients will be key to reducing the environmental impacts of adding additional nutrients to intensive agricultural systems to avoid repeating the errors of the 20th century that contributed to losses of reactive nitrogen to the environment in large quantities.

Environmental systems around the world are under pressure due to different factors such as severe droughts, extreme events, reduction of water sources, deforestation, erosion, and impacts of nutrients on aquatic and other natural systems, among others. The projected additional use of nitrogen and other nutrients at a global scale will continue to put additional pressure on these already impacted environmental systems. Nitrogen will continue to be key during the 21st century and important for food security, global stability and peace. The projected increase in the use of nutrients, the effects of extreme events and a changing climate will all put additional pressure on agricultural systems and increase the potential for losses of nutrients such as nitrogen. Adding more reactive nitrogen to the global nitrogen cycle will also increase pressure on these systems since nitrogen cycles through the environment in a phenomenon known as the nitrogen cascade,[4,5] creating a cascade of effects that accompanies this cycling.

Ribaudo et al. discussed the effects of reactive nitrogen from agricultural sources across the environment and concluded that removing nitrate from drinking water costs about 1.7 billion dollars annually in the United States.[6] Follett et al. reported on the effects of reactive nitrogen on human health and the importance of reducing losses of nitrogen to the environment.[7] In a recent paper, Temkin et al. calculated that exposure to nitrate could be connected to close to 3,000 cases of low birth weight and about 2,300–12,500 cancer cases annually in the United States, with annual economic impacts ranging from 250 million to 1.5 billion US dollars, as well as potentially 1.3–6.5 billion US dollars in lost productivity.[8] These impacts will be much higher on a global scale.[8,9] In addition to the impact of nitrate on groundwater and surface water (which will need to be removed for safety if the concentration is above 10 ppm NO_3–N),[10] surface water with hypoxic zones and algae blooms around the world could potentially develop high microcystin levels.[11] Microcystin-LR concentrations above 1.0 µg/L have been reported to be unsafe by the World Health Organization,[12] while the US Environmental Protection Agency (USEPA) reports that safe concentration limits of microcystin are 0.3 µg/L for children less than 6 years old[13] and that microcystin could contribute to gastroenteritis, and liver and kidney damage.[13] Since current projections suggest that use of nitrogen and other nutrients will continue to increase around the world to produce the additional food needed to feed the growing population, the use of conservation practices and management to reduce losses of nutrients will be of great importance.[14,15] Use of nitrogen trading tools that could assess the positive effects of best management practices will be needed so that savings (reductions in nitrogen loss) could be traded in ecosystem markets during the 21st century. Nitrogen is important for global food security, sustainability and peace; the question is how to manage it in the 21st century in the face of a changing climate, which could contribute to higher losses and make management of nitrogen more complex, and how to avoid the errors that were made in the 20th century, which contributed to a stable food supply but resulted in greater losses of nitrogen to the atmosphere, groundwater and surface water.

Need for Nitrogen Loss Measurements

As climate change continues, the occurrence of extreme events such as droughts and floods, desertification, shrinking water resources and other challenges are beginning to put additional stress on agricultural systems throughout the world. These challenges increase the pressure on these systems to maintain productivity and keep up with the growth in food demand that follows human population growth. Humans are significantly impacting the global nitrogen cycle; these impacts are reflected in, for example, increased concentrations of N_2O in the atmosphere, increased nitrate (NO_3) levels in groundwater or surface water resources and increased ammonia (NH_3) deposition in natural areas. Not only can these reactive nitrogen losses impact human health, as has been previously reported, they can also impact ecosystem health and ecosystem balances.[7–9] Over 20 years ago, the USEPA established that the safe limit for drinking water was 10 mg NO_3–N/L[10]; however, a recent paper has reported a positive correlation of NO_3–N in drinking water as low as 0.14 mg/L nitrate with a colorectal cancer risk of one in a million and greater risk at higher concentrations.[8]

The correlation of food security with nitrogen application is strong, and there is a delicate balancing act to be made: on the one hand, nitrogen applications provide food production benefits, but on the other hand, there are environmental impacts from nitrogen applications. This presents a quandary for humanity as far as how to continue to maximize productivity across agricultural systems, which requires nitrogen inputs, while minimizing nitrogen losses to the environment. Nitrogen trading tools and quantification of savings in nitrogen losses that could be traded in water and air quality markets can help provide some answers to this quandary since farmers that improve management of nitrogen could be compensated by trading these savings in air and water quality markets.[6,16,17] Users of the NTT should look at the effects of improved management in the context of the nitrogen cycle; this will help make it possible to sort out the effects of the practices on losses of reactive nitrogen via emissions of N_2O to the atmosphere (i.e., direct emissions of N_2O) and the losses of nitrogen (e.g. NO_3) through surface transport or leaching, which could impact groundwater or surface waters. How management practices interact with crop rotations depends on the previous and current crops, previous history at the site due to previous manure applications, and the site-specific factors (weather, crops, hydrology and landscape) of a given field, and this can be assessed by these types of tools to reduce nitrogen losses.[6,16,18]

Water Quality Trading Programs

Several researchers have proposed the concept of using payments as part of an ecosystem services payment system in an environmental trading/water quality trading program to reduce the negative effects of reactive nitrogen in the environment.[19–23] Hey and Hey et al.[19,21] recommended the strategic placement of wetlands across the Mississippi watershed that could be used as filters to remove nitrate as it enters these wetlands to reduce the transport of nitrogen to the Gulf of Mexico and have farmers receive environmental payments for developing wetlands for their farms to remove the nitrate. This approach is one of a series of precision conservation approaches that could be used with other conservation practices that could be set up strategically in fields and the landscape to increase the efficiency of conservation practices such as sediment traps, riparian buffers and denitrification traps to remove nitrate, phosphorus and sediment from water traveling to the Mississippi River and water bodies.[24–28] These precision conservation approaches together with precision regulation can contribute to increased adoption of conservation practices that will reduce losses of nitrogen and other nutrients to the environment. These practices can be among the key tools in the 21st-century toolbox to adapt to a changing climate and avoid some of the errors made in the 20th century.[26,28] The concept of Sustainable Precision Agriculture and Environment (SPAE)[29] is similar to the 7 R's approach to reduce the transport of nitrogen to the environment and the transport of other nutrients and sediment.[26,28] There are tools such as the Nutrient Tracking Tool and COMET-VR that can be used to assess the effect of management practices and quantify the reductions of nitrogen losses, nutrient losses and sediment transport across the environment so these reductions can potentially be traded in water quality and air quality markets.[30,31]

A review of potential trading markets by Lal et al.[32] found that precision conservation approaches have been considered by the Connecticut Department of Environmental Protection (DEP), Pennsylvania, Ohio and Oregon, and that these markets could benefit from the use of an NTT approach. The state of Pennsylvania's DEP has a nutrient trading program and additional information could be found at the program's website (https://www.dep.pa.gov/Business/Water/CleanWater/NutrientTrading/Pages/default.aspx). The nutrient trading program in Pennsylvania has developed Excel files used in calculations to assess practices such as cover crops, wetlands restoration, riparian buffers and grass buffers (https://www.dep.pa.gov/Business/Water/CleanWater/NutrientTrading/Pages/TradingResources.aspx) and their reduction of nitrogen losses.

The World Resources Institute (WRI) and the Chesapeake Bay Foundation have reported that these nutrient trading approaches in the states that are part of the Chesapeake Bay area can contribute to reduced environmental impact and could provide income for a farmer that installs a grass buffer along a stream to filter out fertilizer runoff and trades these credits with municipal wastewater treatment plants or stormwater management units that are having trouble reducing their pollution (https://www.wri.org/blog/2017/03/podcast-how-we-can-clean-chesapeake-and-save-money-process).

The concept of nutrient trading for water quality has been supported by the USEPA (https://www.epa.gov/nutrient-policy-data/collaborative-approaches-reducing-excess-nutrients#creating) with policies and regulations. Figure 1 shows nutrient trading for water quality efforts in the United States up to 2016. These nutrient trading efforts have contributed to localities and regions achieving compliance with the regulatory requirements of the Clean Water Act. The USEPA has implemented efforts to leverage

FIGURE 1 A map of water quality trading in the United States in 2016 (USEPA webpage).

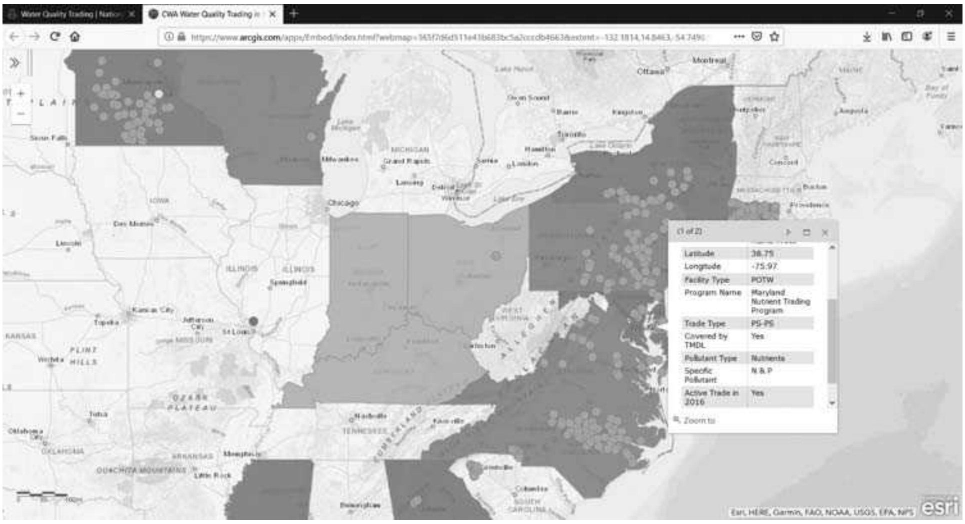
emerging technologies and facilitate broader adoption of market-based programs. Customers of the water quality trading effort under the Clean Water Act can use this option for compliance with water quality-based effluent limitations with a National Pollutant Discharge Elimination System (NPDES) permit. Ribaudo and Gottlie[33] and Ribaudo et al.[34,35] reviewed different trading markets that can be used to reduce environmental effects from agriculture and reported that tools like the NTT could be used to reduce the uncertainty and provide information that could be used for these market systems.

Payments for Environmental Services Programs

Various reviews have been conducted about worldwide water quality trading programs or other types of trading programs. Selman et al. reported in 2009 that water quality trading programs were gaining traction and that there were 57 water quality trading programs worldwide, with 26 active programs, 21 under consideration for development and 10 inactive, with the majority of the programs in the United States and others in Australia, New Zealand and Canada. Stanton et al.[36] reported that there were 288 watershed payment programs. They reported that in 2008 there were 14 active water quality trading programs and 113 payments for watershed service programs with a value of transactions of 9.3 billion dollars, protecting 50.1 billion hectares.[36] They reported that in 2008 there were programs across 24 different countries including the United States, Brazil, Canada, France, China and New Zealand.

Forest Trends has a map of current active ecosystem markets that displays current watershed projects worldwide (https://www.forest-trends.org/project-list/#s). These projects have been identified as projects where people or organizations pay for conservation practices. They identified 466 watershed projects worldwide covering over 330,000,000 hectares. The water quality trading in the United States in 2016 can be found at the USEPA website (https://www.epa.gov/nutrient-policy-data/collaborative-approaches-reducing-excess-nutrients#creating), with a large number of programs concentrated around the north-east and in the Midwest. At this webpage, the user could click on each identified point on the map and find specific information about the program (Figure 2).

The Nature Conservancy identified 391 carbon sequestration projects worldwide covering over 52,000,000 hectares. These carbon projects can be used to produce carbon offsets (reductions in CO_2 emissions) that are sold in carbon markets. These sold offsets in CO_2 can be used to offset greenhouse gas emissions elsewhere. The Environmental Defense Fund has a website describing 19 multinational, national,



(1 of 2)

Latitude	38.75
Longitude	-75.97
Facility Type	POTW
Program Name	Maryland Nutrient Trading Program
Trade Type	PS-PS
Covered by TMDL	Yes
Pollutant Type	Nutrients
Specific Pollutant	N & P
Active Trade in 2016	Yes

FIGURE 2 A map of water quality trading in the United States in 2016 (USEPA webpage). The user could click on each identified point on the map and find specific information about the program.

regional and local emissions trading systems (https://www.edf.org/worlds-carbon-markets). They reported that there are over 50 jurisdictions around the world with emission trading programs (or carbon markets).

The NTT has been reported as a potential tool that could be used to assess the effects of management practices on reductions of greenhouse gases (GHG).[6,16,30,31] Delgado et al. reported that since nitrogen fertilizers increase N_2O emissions and we could use the NTT to assess the effects of practices that reduce emissions of N_2O and that reduce NH_3 volatilization and NO_3 leaching losses that contribute to indirect emissions of N_2O, the NTT could be used to trade reductions of direct and indirect N_2O emissions.[6,16] Recent advances with tools such as the Nutrient Tracking Tool and COMET-VR also put these tools in a strong position to be used to assess the effects of management practices on carbon sequestration and reductions of direct and indirect N_2O emissions.[30,31] Both of these tools are being used by NRCS.[30,31]

The Nutrient Tracking Tool (NTT) is a tool being used by NRCS and is a free, online, user-friendly tool that quantitatively calculates the nitrogen, phosphorus and sediment losses from crop and pasture lands, and these savings could potentially be used for trade in water and air quality markets. The COMET-Farm tool is the official greenhouse gas quantification tool of USDA, and it could be used to the assess effects of management practices on reductions of GHG (https://www.nrcs.usda.gov/wps/portal/nrcs/detail/national/programs/?cid=stelprdb1261363).

The American Carbon Registry is a nonprofit carbon market in the United States that is a voluntary greenhouse gas registry (https://americancarbonregistry.org/). Among the different factors that this voluntary registry can keep track of is carbon sequestration in soils as well as potential reductions in N_2O emissions from agricultural fields. This registry lists accepted methodologies for quantifying reductions in greenhouse gas emissions resulting from changes in fertilizer management, such as improvement in nitrogen fertilizer management practices, which could contribute to reductions in N_2O emissions (https://americancarbonregistry.org/carbon-accounting/standards-methodologies/emissions-reductions-through-changes-in-fertilizer-management). Examples like this show that water and air (carbon) quality markets are connected, in that improved nitrogen management practices that increase nitrogen use efficiencies and reduce losses of nitrate, generate savings that can be traded in water quality markets, and increase efficiency, simultaneously reducing emissions of N_2O and generating savings that can be traded in air (carbon) trading markets. The 21st century will continue to bring opportunities to farmers to improve their nitrogen management practices and trade these savings in nitrate leaching losses and direct and indirect emissions of N_2O in water and air (carbon) trading markets.

Conclusion

There are new ecotechnologies such as the Nitrogen Trading Tool (NTT) that could be used to quickly assess how best management practices may reduce the losses of reactive nitrogen via several pathways, including nitrate leaching and direct and indirect N_2O emissions. The first NTT was developed by ARS in cooperation with NRCS.[6,16,37] The concept and capabilities of the original Nitrogen Trading Tool (NTT) were expanded in the Nutrient Tracking Tool (NTT), which can assess the effects of management practices on reductions of nitrogen and phosphorus losses and reductions in erosion.[18,30] Additionally, the new COMET-VR model can also assess the effects of conservation practices on reductions of greenhouse gas emissions and assess carbon sequestration in farms. There is potential to use the Nutrient Tracking Tool (NTT) and COMET-VR tools, which are both robust and have a strong scientific background, across the United States and internationally. Both the NTT and COMET-VR are currently being used by USDA as their tools to assess the effects of conservation practices on potential reductions in greenhouse gases and protection of water quality.

There is also potential to use other tools and approaches that could quantify the effects of management practices on reduction of nitrogen losses to the environment. For example, various water quality trading markets have developed coefficients and Excel files to quantify the reductions in nitrogen losses when a given conservation practice is applied (www.dep.pa.gov/Business/Water/CleanWater/NutrientTrading/Pages/default.aspx; https://www.wri.org/blog/2017/03/podcast-how-we-can-clean-chesapeake-and-save-money-process). Webpages developed by the EPA and Forest Trends are examples of online tools that track the development of active water and air quality markets where there are opportunities for tools like the NTT be to applied (https://www.epa.gov/nutrient-policy-data/collaborative-approaches-reducing-excess-nutrients#creating; https://www.forest-trends.org/project-list/#s). Independent of what tools are being used to assess the benefits of management practices in reducing losses of reactive nitrogen to the environment, application of conservation practices across watersheds will provide opportunities to reduce losses of nitrogen that can potentially impact water quality and air (carbon) quality. The use of new technologies and tools like the Nutrient Tracking Tool and COMET-VR that can assess temporal and spatial variability will allow us to assess hot-spot sources of nitrate leaching and greenhouse gas emissions across the environment and effectively apply conservation practices by considering spatial and temporal variability with precision conservation management and precision regulation.[27–31] The concept of the NTT and new ecotechnologies such as COMET-VR and the Nutrient Tracking Tool will help nutrient managers, conservation practitioners, farmers and other users quickly identify conservation practices that maximize nitrogen use efficiency while reducing nitrogen losses to the environment, providing an opportunity to trade the saved nitrogen in water and air (carbon) markets and contributing to conservation of the biosphere.[6,16,18,17,29–31]

Disclaimer

Trade and manufacturer's names are necessary to report factually on available data; however, the USDA neither guarantees nor warrants the standard of the product, and the use of the name by USDA implies no approval of the product to the exclusion of others that may also be suitable.

References

1. Delgado, J.A.; Barrera Mosquera, V.H.; Escudero López, L.O.; Cartagena Ayala, Y.E.; Alwang, J.R.; Stehouwer, R.C.; Arévalo Tenelema, J.C.; D Adamo, R.E.; Domínguez Andrade, J.M.; Valverde, F.; Alvarado Ochoa, S.P.; Conservation Agriculture increases profits in an Andean region of South America. *Agrosystems, Geosciences & Environ.* **2019** doi:10.2134/age2018.10.0050.

2. Max Roser, M.; Ritchie, H.; *Yields and Land Use in Agriculture*. University of Oxford. Global Change Data Laboratory. Published online at OurWorldInData.org. Retrieved from https://ourworldindata.org/yields-and-land-use-in-agriculture [Online Resource].
3. Smil, V.; Global population and the nitrogen cycle. *Scientific American* **1997**, *227*, 77–81.
4. Cowling, E.; Galloway, J.; Furiness, C.; Erisman, J.W. Optimizing nitrogen management in food and energy production and environmental protection. *Report* from *the* Second *International Nitrogen Conference*, October 14–18, 2001; Bolger Center: Potomac, MD, 2002.
5. Galloway, J.N.; Aber, J.D.; Erisman, J.W.; Seitzinger, S.P.; Howarth, R.W.; Cowling, E.B.; Cosby, B.J. The nitrogen cascade. *BioScience* **2003**, *53* (4), 341–356.
6. Delgado, J.A.; Shaffer, M.J.; Lal, H.; McKinney, S.; Gross, C.M.; Cover, H. Assessment of nitrogen losses to the environment with a Nitrogen Trading Tool. *Comput. Electron. Agric.* **2008**, *63,* 193–206.
7. Follett, J.R.; Follett, R.F.; Herz, W.C. Environmental and human impacts of reactive nitrogen. In *Advances in Nitrogen Management for Water Quality*; Delgado, J.A., Follett, R.F., Eds.; Soil and Water Conservation Society: Ankeny, IA, 2010; 1–37.
8. Temkin, A.; Evans, S.; Manidis, T.; Campbell, C.; Naidenko, O.V.; Exposure-based assessment and economic valuation of adverse birth outcomes and cancer risk due to nitrate in United States drinking water. *Environ. Res.* **2019**, doi:10.1016/j.envres.2019.04.009.
9. Ribaudo, M.; Delgado, J.; Hansen, L.; Livingston, M.; Mosheim, R.; Williamson, J. *Nitrogen in Agricultural Systems: Implications for Conservation Policy*; ERS: Economy Research Report: Washington, DC, 2010.
10. USEPA (U.S. Environmental Protection Agency). National Primary Drinking Water Regulations. n.d. https://www.epa.gov/ground-water-and-drinking-water/national-primary-drinking-water-regulations.
11. Smith, D.R.; Wilson, R.S.; King, K.W.; Zwonitzer, M.; McGrath, J.M.; Harmel, R.D.; Haney, R.L.; Johnson, L.T. Lake Erie, phosphorus, and microcystin: Is it really the farmer's fault? *J. Soil Water Conserv.* **2018**, *73*, 48–57.
12. World Health Organization (WHO). *Guidelines for Drinking-water Quality*. 4th Edn. WHO Press: Geneva, 2011.
13. USEPA (United States Environmental Protection Agency). Drinking Water Health Advisories for Two Cyanobacterial Toxins. 2015, Office of Water 820F15003 June 2015; Washington D.C. https://www.epa.gov/sites/production/files/2017-06/documents/cyanotoxins-fact_sheet-2015.pdf.
14. Delgado, J.A.; Groffman, P.M.; Nearing, M.A.; Goddard, T.; Reicosky, D.; Lal, R.; Kitchen, N.; Rice, C.; Towery, D.; Salon, P. Conservation practices to mitigate and adapt to climate change. *J. Soil Water Conserv.* **2011**, *66*, 118A–129A.
15. Lal, R.; Delgado, J.A.; Groffman, P.M.; Millar, N.; Dell, C.; Rotz, A.; Management to mitigate and adapt to climate change. *J. Soil Water Conserv.* **2011**, *66*, 276–285.
16. Delgado, J.A.; Gross, C.M.; Lal, H.; Cover, H.; Gagliardi, P.; McKinney, S.P.; Hesketh, E.; Shaffer, M.J. A new GIS nitrogen trading tool concept for conservation and reduction of reactive nitrogen losses to the environment. *Adv. Agron.* **2010**, *105,* 117–171.
17. Delgado, J.A., Follett, R.F., Eds. *Advances in Nitrogen Management for Water Quality*; Soil and Water Conservation Society: Ankeny, IA, 2010.
18. Saleh, A.; Gallego, O.; Osei, E.; Lal, H.; Gross, C.; McKinney, S.; Cover, H. Nutrient Tracking Tool—A user-friendly tool for calculating nutrient reductions for water quality trading. *J. Soil Water Conserv.* **2011**, *66*, 400–410.
19. Hey, D.L. Nitrogen farming: Harvesting a different crop. *Restor. Ecol.* **2002**, *10,* 1–10.
20. Greenhalch, S.; Sauer, A. *Awakening the Dead Zone: An Investment for Agriculture, Water Quality, and Climate Change*; World Resources Institute: Washington, DC, 2003.
21. Hey, D.L.; Urban, L.S.; Kostel, J.A. Nutrient farming: The business of environmental management. *Ecol. Eng.* **2005**, *24,* 279–287.

22. Ribaudo, M.O.; Heimlich, R.; Peters, M. Nitrogen sources and Gulf hypoxia: Potential for environmental credit trading. *Ecol. Econ*. **2005**, *52*, 159–168.
23. Glebe, T.W. The environmental impact of European farming: How legitimate are agri-environmental payments? *Rev. Agric. Econ*. **2006**, *29,* 87–102.
24. Berry, J.R.; Delgado, J.A.; Khosla, R.; Pierce, F.J. Precision conservation for environmental sustainability. *J. Soil Water Conserv.* **2003**, *58,* 332–339.
25. Delgado, J.A.; Berry, J.K. Advances in precision conservation. *Adv. Agron.* **2008**, *98,* 1–44.
26. Delgado, J. A.; 4 Rs are not enough. We need 7 Rs for nutrient management and conservation to increase nutrient use efficiency and reduce off-site transport of nutrients. In *Soil Specific Farming: Precision Agriculture;* Lal, R.; Stewart, B. A., Eds. Advances in Soil Science Series; CRC Press: Boca Raton, FL, 2016; 89–126.
27. Delgado, J. A., Sassenrath, G., Mueller, T., Eds. *Precision Conservation: Geospatial Techniques for Agricultural and Natural Resources Conservation*. Agronomy Monograph 59. American Society of Agronomy, Crop Science Society of America, and Soil Science Society of America: Madison, WI, 2018.
28. Sassenrath, G. F., Delgado, J.A. Precision Conservation and Precision Regulation Delgado, J. A., Sassenrath, G., Mueller, T., Eds. *Precision Conservation: Geospatial Techniques for Agricultural and Natural Resources Conservation*. Agronomy Monograph 59. American Society of Agronomy, Crop Science Society of America, and Soil Science Society of America: Madison, WI, 2018.
29. Delgado, J.A.; Short Jr.; N.M.; Roberts, D.P.; Vandenberg, B. Big Data Analysis for Sustainable Agriculture on a Geospatial Cloud Framework. *Front. Sust. Food Syst*. **2019;** doi:10.3389/fsufs.2019.00054.
30. Saleh, A., Osei, E., Precision conservation and water quality markets. In *Precision Conservation: Geospatial Techniques for Agricultural and Natural Resources Conservation*; Delgado, J. A., Sassenrath, G., Mueller, T., Eds. Agronomy Monograph 59. American Society of Agronomy, Crop Science Society of America, and Soil Science Society of America: Madison, WI, 2018; 313–340.
31. Paustian, K., Easter, M., Brown, K., Chambers, A., Eve, M., Huber, A., Marx, E., Layer, M., Sterner, M., Sutton, B., Swan, A., Toureene, C., Verlayudhan, S., Williams, S., Field- and farm-scale assessment of soil greenhouse gas mitigation using COMET-Farm. In *Precision Conservation: Geospatial Techniques for Agricultural and Natural Resources Conservation*; Delgado, J. A., Sassenrath, G., Mueller, T., Eds. Agronomy Monograph 59. American Society of Agronomy, Crop Science Society of America, and Soil Science Society of America: Madison, WI, 2018; 341–360.
32. Lal, H.; Delgado, J.A.; Gross, C.M.; Hesketh, E.; McKinney, S.P.; Cover, H.; Shaffer, M. Market-based approaches and tools for improving water and air quality. *Environ. Sci. Policy* **2009**, *12,* 1028–1039.
33. Ribaudo, M.; Gottlie, J. Point-nonpoint trading—Can it work? *J. Am. Water Resour. Assoc.* **2011**, *47,* 5–14.
34. Ribaudo, M.; Delgado, J.; Hansen, L.; Livingston, M.; Mosheim, R.; Williamson, J. *Nitrogen in Agricultural Systems: Implications for Conservation Policy*; ERS (Economy Research Report): Washington, DC, 2010.
35. Ribaudo, M.; Greene, C.; Hansen, L.; Hellerstein, D. Ecosystem services from agriculture: Steps for expanding markets. *Ecol. Econ*. **2010**, *69,* 2085–2092.
36. Stanton, T.; Echavarria, M.; Hamilton, K.; Ott, C. State of Watershed Payments: An Emerging Marketplace. Ecosystem Marketplace 2010, available at https://www.forest-trends.org/publications/state-of-watershed-payments/.
37. Gross, C.; Delgado, J.A.; McKinney, S.; Lal, H.; Cover, H.; Shaffer, M. Nitrogen Trading Tool (NTT) to facilitate water quality credit trading. *J. Soil Water Conserv.* **2008**, *63*, 44A–45A.

41

The Accounting Framework of Energy–Water Nexus in Socioeconomic Systems

Saige Wang and
Bin Chen

Introduction

Water use and energy consumption are highly interwoven in economic networks due to their interconnections in both the production and consumption of primary, intermediate, and final goods and services (Chen 2016; Chen and Lu 2015). At all stages of the production chain, water and energy resources are used and "embodied" in the products. Nexus analysis highlights the interconnections and interdependencies for all kinds of economic activities (Dai et al. 2018; Hoff 2011). The tracking of energy and water flows among regions and the quantification of their interdependencies are fundamental for coordinated management of regional resources (Liu et al. 2018).

Most of the recent research on energy–water nexus has explored the energy or water sector alone or impacts of specific technologies on sectors with energy minimization and water use options (Kenway et al. 2011; Wang and Chen 2016; Wang et al. 2017a). However, there are few works published that conceptualize the role of the nexus in reconfiguring the interactions between energy- and water-related sectors in a socioeconomic system (Chen et al. 2018; Owen et al. 2018). Some studies have explored the direct and indirect interdependencies between energy and water in socioeconomic systems (Duan and Chen 2016; Fang and Chen 2017; Liu et al. 2019). By tracking sectoral economic flows through a supply chain, input–output analysis (IOA) can be used to investigate these interdependencies, which are a common focus in nexus studies (Wang et al. 2019; Wang and Chen 2016). Based on these sectoral interactions and exchanges with other economies through the supply chain, IOA-based approaches can assess both the direct and indirect energy (or water) consumption required to produce goods and services in a region (Wang et al. 2017b; Wang and Chen 2018). Compared with conventional nexus studies, which prioritize specific interdependencies (e.g., by defining that specific processes require certain resources), the system-wide approach of IOA presents a more comprehensive picture that includes direct and indirect use profiles and hotspots. Pioneering works have used IOA case studies to focus on the interplay involved in the water–energy nexus (Cao et al. 2018; Chen and Chen 2016; Wang et al. 2019). For example, Wang et al. (2017b) proposed a modified IOA that provided a unified framework to analyze the tradeoffs between urban energy and water systems. Fang and Chen (2017) used IOA and linkage analysis to detect synergetic effects of water and energy consumption and their interactions among economic sectors. Marsh suggested various IOA techniques to address multiple dimensions of the nexus (linkage, dependency, multiplier, and scenario analyses) (Marsh 2008). Kahrl and Roland-Holst built relevant metrics to quantify the nexus from physical, monetary, and distributive perspectives (Kahrl and Roland-Holst 2008).

Ecological network analysis (ENA), a system-based approach, has its unique strength in examining the structure and function of system from a system perspective (Fath 2007, 2015). The control and dependence analysis of ENA may also provide insights into the interwoven relationships among nodes due to the direct and indirect flows, which can be used to identify the regulating pathways for the system. Also, Wang and Chen (2016) established a multiregional nexus network based on multiregional IOA and ENA to explore the structural properties and sectoral interactions within urban agglomerations. However, nexus issues in energy and water policy are rarely studied from a systems perspective, an area that urgently requires investigation. Particularly in energy development planning, the role of water stress issues is insufficiently considered as a part of nexus management (Font et al. 2017; Williams et al. 2014).

Water is directly or indirectly required in all types of power generation technologies for cooling purposes, steam generation, and infrastructure manufacturing. There is a great diversity of water intensities among different energy types. The water intensity gap among different electricity types provides us the opportunity to reduce the power generation's vulnerability to water shortages. In this study, we will propose an accounting framework to extend the energy development scenarios into nexus to study interwoven connections of energy consumption and water use in economic supply chain based on IOA and ENA.

Accounting Framework

In our accounting framework, we will first quantify the energy-related water and water-related energy from consumption and production perspectives combining a bottom-up method and a life cycle analysis. Then, energy and water flows among regions and sectors will be calculated based on IOA. The direct and indirect energy and water-related energy were then combined as embodied energy to build the nexus network model. By merging the energy-related water and direct water, we will build a regional energy–water nexus network to investigate its system properties. The nexus point is built to form the energy and water flows in the same nexus network. The tools from ENA were used to explore the energy and water control and dependence relationship among sectors under the nexus impact background. Flow analysis was used to identify critical pathways for nexus management. We investigated the sectoral nexus impact for energy system by comparing energy and water-related energy network to identify critical sectors for energy-side nexus management. The sectoral nexus impact under different scenarios was compared and analyzed based on ENA. Finally, network impact through direct and indirect pathways is used to reveal the effect of energy–water nexus on all sectors in a more explicit and predictable way and the tradeoffs among energy, water, and carbon using nexus accounting and the scenario features can be discussed.

Thus, we can capture both intensities and structures of regional nexus and how they alter the local consumption. Integrated analysis and modeling of the water–energy nexus require the simulation of many human and natural systems and their complex interactions and dynamics (Wang et al. 2018, 2019). By doing this, we are trying to reveal energy–water nexus impacts of pathways and sectors on the system and track energy and water flows among sectors. System properties, such as cycling index and system robustness, can help us to identify the risks in the energy–water nexus and leverage points to promote sustainable regional energy and water management.

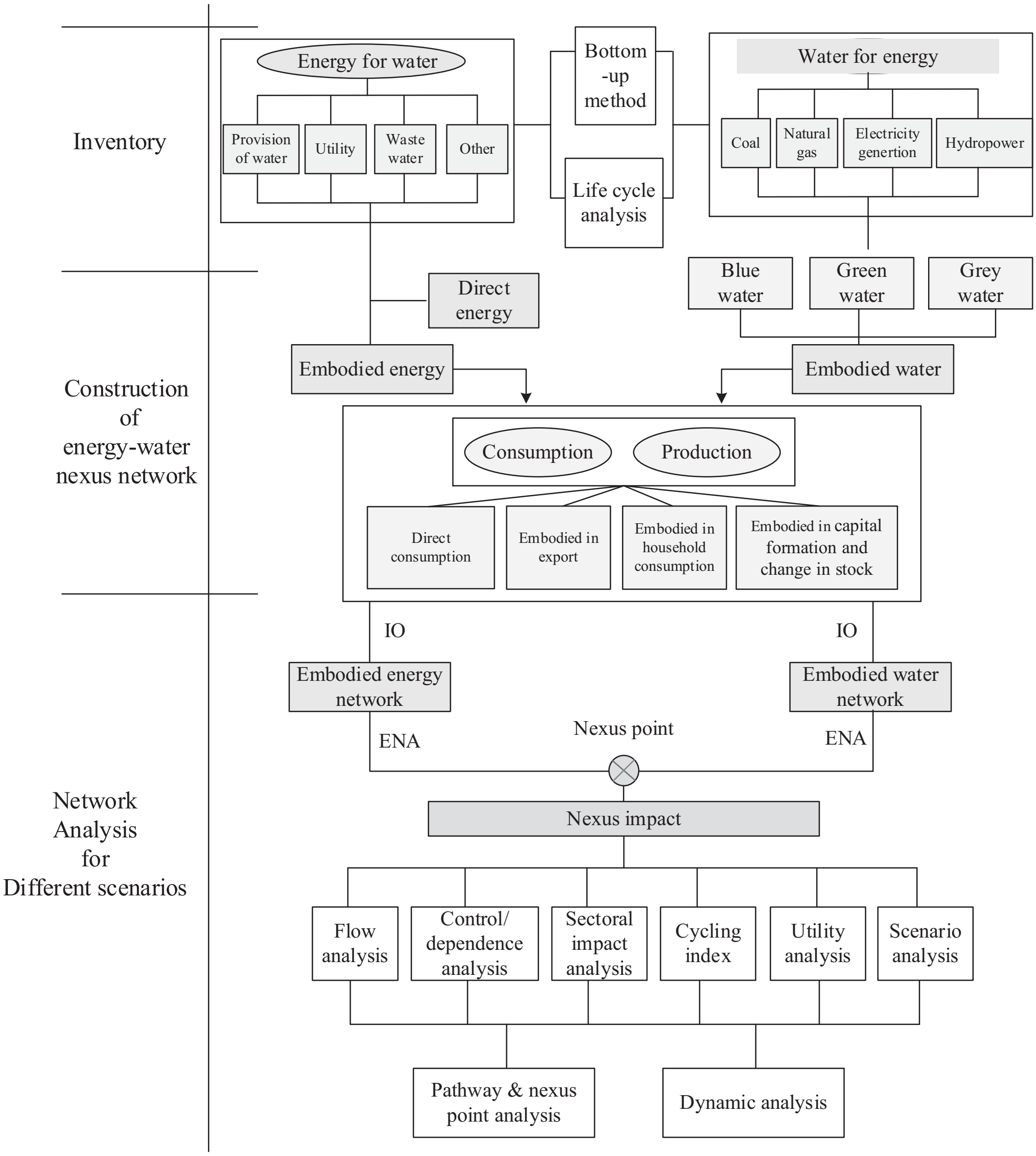


FIGURE 1 Energy–water nexus accounting framework for socioeconomic system.

References

Cao, T., Wang, S., and Chen, B. 2018. Priority in Energy-water Nexus Based on Input-output Analysis: A Case Study of China. DEStech Transactions on Environment, Energy and Earth Sciences, (iceee).

Chen, B. 2016. Energy, ecology and environment: A nexus Perspective. *Energy Ecology Environment* 1(1): 1–2.

Chen, B., and Lu, Y. 2015. Urban nexus: A new paradigm for urban studies. *Ecological Modelling* 318: 5–7.

Chen, P. C., Alvarado, V., Hsu, S. C. 2018. Water energy nexus in city and hinterlands: Multi-regional physical input-output analysis for Hong Kong and South China. *Applied Energy* 225: 986–97.

Chen, S.Q., and Chen, B. 2016. Urban energy–water nexus: A network perspective. *Applied Energy* 184: 905–14.

Dai, J., Wu, S., Han, G., Weinberg, J., Xie, X., Wu, X., … Yang, Q. 2018. Water-energy nexus: A review of methods and tools for macro-assessment. *Applied Energy*, 210: 393–408.

Duan, C.C., and Chen, B. 2016. Energy–water nexus of international energy trade of China. *Applied Energy* 194:725–34.

Fang, D.L., and Chen, B. 2017. Linkage analysis for the water–energy nexus of city. *Applied Energy* 189: 770–9.

Fath, B.D. 2007. Network Mutualism: Positive community level relations in ecosystems. *Ecological Modelling* 208: 56–67.

Fath, B.D. 2015. Quantifying economic and ecological sustainability. *Ocean Coast Management* 108: 13–9.

Font V., D., Wang, R., Hertwich, E. 2017. Nexus strength: A novel metric for assessing the global resource nexus. *Journal of Industrial Ecology* 22(6): 1473–86.

Hoff, H. 2011. Understanding the nexus. *Background Paper for the Bonn 2011 Conference: The Water, Energy and Food Security Nexus*. Stockholm Environment Institute, Stockholm.

Kahrl, F., and Roland-Holst, D. 2008. China's water–energy nexus. *Water Policy* 10(S1):51–65.

Kenway, S. J., Lant, P. A., Priestley, A., and Daniels, P. 2011. The connection between water and energy in cities: A review. *Water Science and Technology* 63(9): 1983–90.

Liu, J., Hull, V., Godfray, H. C. J., Tilman, D., Gleick, P., Hoff, H., … Li, S. 2018. Nexus approaches to global sustainable development. *Nature Sustainability*, 1(9): 466.

Liu, Y., Wang, S., Chen, B. 2019. Water–land nexus in food trade based on ecological network analysis. *Ecological Indicators* 97: 466–475.

Marsh, D.M. 2008. The water-energy nexus: A comprehensive analysis in the context of New South Wales. *Doctoral Dissertation*.

Owen, A., Scott, K., Barrett, J. 2018. Identifying critical supply chains and final products: An input-output approach to exploring the energy-water-food nexus. *Applied Energy* 210: 632–42.

Wang, S.G., Cao, T., and Chen, B. 2017a. Water–energy Nexus in China's Electric Power System. *Energy Procedia*, 105, 3972–977.

Wang, S.G., Cao, T., and Chen, B. 2017b. Urban energy–water nexus based on modified input–output analysis. *Applied Energy* 196: 208–21.

Wang, S.G., and Chen, B. 2016. Energy–water nexus of urban agglomeration based on multiregional input–output tables and ecological network analysis: A case study of the Beijing–Tianjin–Hebei region. *Applied Energy* 178: 773–83.

Wang, S. G., and Chen, B. 2018. Three-Tier carbon accounting model for cities. *Applied Energy* 229: 163–75.

Wang, S.G., Fath, B., and Chen, B. 2019. Energy–water nexus under energy mix scenarios using input–output and ecological network analyses. *Applied Energy* 233, 827–39.

Wang, S.G., Liu, Y. T., Chen, B. 2018. Multiregional input–output and ecological network analyses for regional energy–water nexus within China. *Applied Energy* 227: 353–64.

Williams, J., Bouzarovski, S., Swyngedouw, E. 2014. Politicising the nexus: Nexus technologies, urban circulation, and the coproduction of water–energy. Nexus Network Think Piece Series, Paper 001, November.

42

Water Quality: Modeling

Richard Lowrance

Introduction

Water quality models are based on some representation of hydrology and may include movement of surface water, groundwater, and mixing of water in lakes and water bodies. Based on the hydrology, water quality models then simulate some combination of sediment, nutrients, heavy metals, and xenobiotics such as pesticides. Some water quality models, especially those that deal with nutrients, may contain substantial detail related to biological processes including algal growth, nutrient transformations, and respiration. Most water quality models that portray the movement of water within a landscape or landscape components (e.g., fields, forests, streams) portray the interaction of water with soil in a variety of ways. Newer water quality models and add-ons to older water quality models are able to portray the effects of water quality parameters on the biota of lakes and streams or incorporate stream bank, riparian zone, and/or channel functions to understand the effects of these areas on chemical and sediment transport. Other water quality models are used to simulate the effects of critical inputs on the biological communities of lakes and rivers. These aquatic ecosystem models may or may not be tied to watershed models that provide simulated loading to the aquatic ecosystem under varying land use and management.

Classification of Water Quality Models

Water quality models are either built on hydrologic models, are used in conjunction with hydrologic models, or use empirical hydrologic data. Although water quality models can be physical representations of the real world such as channels and ditches built to scale, mathematical or formal models are more common.[1] Mathematical water quality models are quantitative expressions of processes or phenomena that are known to occur in the real-world. The expressions are simplifications of real-world systems through a series of equations governed by conservation of mass. Mathematical water quality models are often a combination of theoretical and empirical representations of the real-world system. Empirical models use water quality observations to provide estimates of water quality parameters through regression analysis. Process based or theoretical representations use physical, chemical, and biological causal relationships to describe the workings of a conceptual system.

Although the real world is subject to random occurrences of weather and management that drive hydrology and water quality, many models ignore the randomness of inputs and spatially distributed attributes and assume that there is a known value for all model parameters. Conversely, stochastic (or random) models use probability distributions of parameters in time or space and can provide outputs based on the distribution. Most water quality models are deterministic models in the sense that one set of inputs will provide only one set of outputs. The difference in a stochastic and deterministic model can be illustrated by how models deal with something simple like how fast water moves in a soil. A deterministic model would use one value for each soil while a stochastic model would vary the movement rate based on the range and distribution of measured water movement rates. Deterministic models are often used with a range of key parameters in order to produce a range of outputs that would better represent real world conditions. Another critical distinction among water quality models is whether they provide continuous or event-based simulations. Continuous simulation models generally provide at least some representation of groundwater/surface water interactions, while event-based models are more likely to provide only representations of hydrologic processes that take place during rainfall events.

A final distinction among models is whether they are lumped or distributed parameter models. A lumped parameter model contains little or no spatial realism and represents landscape units as homogeneous with respect to the parameters and inputs that drive the model. A distributed parameter model represents certain aspects of the landscape structure, typically by representing areas that are homogeneous with respect to soils, vegetation, and/or land use. Each of these discrete areas is modeled separately and then outputs from all the discrete areas are put together and routed through the system. Because most water quality models are tied to hydrologic models, the water quality outputs from source areas in the model are typically routed through either surface flow pathways, subsurface flow pathways, or both. Models that deal only with events are typically routed through surface flow pathways. Models that simulate continuous or daily water quality in a watershed or field generally must deal with both subsurface or groundwater routing and surface water routing.[2]

Uses of Water Quality Models

Risk Assessment of Pesticides

Knowledge of fate and transport of pesticides in the environment is essential to the assessment of risk due to dietary and drinking water exposure. The passage of the Food Quality Protection Act (FQPA) lead to a pressing need to quantitatively predict ranges and magnitudes of expected environmental pesticide concentrations in drinking water. Health-based safety standards mandated by FQPA require USEPA to consider drinking water exposures of humans to pesticides during the risk assessment process. Some state agencies and USEPA use screening models to estimate pesticide concentrations in groundwater and surface water to identify those food-use pesticides that are not expected to contribute enough exposure via drinking water to result in unacceptable levels of aggregate risk.[3] The models are used to guide regulatory agencies such as USEPA to identify where more detailed field data are needed.

Evaluation of Best Management Practices (BMPs)

Water quality improvement from extensive land uses such as agriculture and forestry depends largely on the use of BMPs. Agricultural water quality modeling attempts to adequately represent the differences among various management practices in order to compare and choose which BMPs lead to the least transport of pollutants. These models are typically structured to represent homogeneous landscape units such as fields or portions of fields in order to compare management features such as tillage, fertilizer sources, manure use, and pesticide use and predict the relative impacts on local transport of pollutants such as sediment, nitrogen, phosphorus, and pesticides. Existing models may be used to test the application of BMPs to areas for which no water quality data are available or to determine the effects of BMPs that are similar to those for which water quality effects have been quantified.

Evaluation of Sources and/or Impacts of Pollutants

Both process based and empirical models have been used successfully to examine the sources of pollutants in watersheds and the impact of pollutants or non-pollutants on aquatic ecosystems. The need to quantify the non-point source contributions for watersheds and small basins is largely driven by total maximum daily load (TMDL) assessments and implementation plans mandated by the federal Clean Water Act.[4] The TMDL assessments are done with a water quality accounting approach that typically uses water quality models to estimate non-point source pollution. The non-point and point sources of a pollutant that are causing the water quality impairment are then combined and compared to observations in the water body. If the water quality is impaired due to the direct presence of a pollutant, then the model estimates of non-point source pollution are used to design a plan for reducing non-point sources or trading point sources for non-point. If pollutants are tied indirectly to the impairment, for instance nutrient enrichment that causes low dissolved oxygen, then the behavior of the pollutant in the waterbody is modeled in order to determine the necessary pollutant load reduction.

Explanation of Large-Scale Systems Behavior

As the behavior of large-scale systems becomes more of an issue and as water quality monitoring data become more available, attempts have been made to combine monitoring and modeling to predict the transport of water-borne pollutants on large scales—river basins and continents. Regression models are used to relate measured pollutant transport in streams to spatially referenced descriptors of pollutant sources, land surface characteristics, and stream channel characteristics.[5] Although mechanisms of pollutant transport are not modeled directly, coefficients that serve as surrogates for processes are used to achieve substantial explanatory power for observed water quality data.

References

1. Tim, U.S. Emerging issues in hydrologic and water quality modeling research. In *Water Quality Modeling*; Heatwole, C., Ed.; American Society of Agricultural Engineers: St. Joseph, MI, 1995; 358–373.
2. Overton, D.E.; Meadows, M.E. *Stormwater Modeling;* Academic Press: New York, 1976.
3. USEPA, OPP. 1999. Estimating the Drinking Water Component of a Dietary Exposure Assessment.
4. *NAS.* Watershed Management for Potable Water Supply: Assessing the New York City Strategy; *National Academy of Sciences: Washington, DC, 2000; 549 pp.*
5. Smith, R.A.; Schwarz, G.E.; Alexander, R.B. Regional interpretation of water quality monitoring data. Water Resources Research **1997**, *33*, 2781–2798.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

V

ELE: Focuses on the Use of Legislation or Policy to Address Environmental Problems

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

43

Drainage: Hydrological Impacts Downstream

Mark Robinson
and D.W. Rycroft

Introduction

Land drainage is the practice of removing excess water from the land, and it is one of the most important land management tools for improving crop production in many parts of the world. Drainage systems may be broadly divided into surface drainage (comprising land grading and open ditches), shallow drainage (such as subsoiling to mechanically loosen the upper layer of soil), subsurface or groundwater drainage (buried perforated pipes or deep ditches), and the main drainage systems (commonly open channels) used to convey the drain water away.[1] Drainage will inevitably affect the pattern of water flows from the land and into the receiving watercourses. It is these downstream impacts of farmland drainage on the timing and magnitude of peak flows that are considered here, using the results of experimental studies and computer simulations, to present a coherent picture, and to answer most of the apparent anomalies and conflicts.

Hydrologic Impacts

Concern about the possible downstream effects of drainage is shown by many published papers worldwide, in North America,[2,3] Great Britain,[4,5] and continental Europe, including France,[6] Netherlands,[7] Ireland,[8,9] Fin- land,[10] and Germany.[11] The role of drainage has often been highlighted by serious flood events—for example, in the Midwest of the United States in 1993, and across Europe in 1997—which reawakened concerns that drainage could aggravate flooding downstream.

There has been a debate about the effects of drainage on streamflow for well over a century, but until recently, due to the lack of appropriate data, the debate has been largely speculation. Too often, the absence of evidence has erroneously been taken as evidence of an absence of effect. The earliest published account[12] was a report of a 4-day meeting held at the Institution of Civil Engineers in London in 1861. Many of the arguments and opinions expressed have resonance today, but due to the absence of objective measurements, the participants were unable to reach any conclusions and the meeting was inconclusive.

These conflicting opinions resulted from differences in the emphasis given to the two processes of water storage and routing. Considering the former, it may be argued that because drainage lowers the water table, the available storage capacity in the soil is enlarged and able to absorb more storm rainfall, thereby reducing peak flow rates. In contrast, according to the routing argument, the purpose of drainage is to "remove water from the land more quickly" than under natural conditions, so peak outflows must necessarily increase.

Probably more work has been carried out in Britain upon the effects of agricultural drainage upon streamflow than in any other country. Britain was the originator of modern field drainage[13] and so became the first country where concern arose about its downstream effects; it is also one of the most extensively drained countries in the world.

It is only in the last few years that it has been possible to obtain a coherent picture based on observations of field processes and supported and extended by computer modeling. This has shown that general statements that drainage "causes" or "reduces" flood risk downstream are oversimplifications of the complex processes involved, and that any consideration of the impact of drainage on streamflow must identify the point of interest, whether at the outfall from the field, along the main channel, or a combination of both at the catchment scale.

Experimental studies indicate that the provision of surface drainage will result in higher peak flows downstream. This was shown by a long-term experiment at Sandusky in Northern Ohio[14] and is a result of the reduction/ elimination of surface storage capacity, as well as the provision of more efficient faster flow routes. This has been demonstrated conclusively both by experimental studies and by computer simulations.

In contrast, there seems to be general agreement from experimental studies that subsurface drainage of waterlogged, poorly permeable clay soils reduces peak out- flows.[15–17] Since this is one of the most common situations where artificial drainage is used, it might be considered to represent the most general result of field drainage.

There are, however, instances where even on heavy soils this result may not apply. Due to their low hydraulic conductivity, most water movement in clay soils is confined to flow through macropores, such as cracks. As a result of clay shrinkage and cracking in warm, dry summers, rapid macropore flow can result in larger peak flows from the drained land than from the undrained land. The role of macropores on the seasonality of peak flows from drained land was demonstrated in detail.[18]

More permeable, drier soils may also be drained where there is an economic justification—for example, drainage of land producing high-value crops. In contrast to clay soils, relatively few scientific field studies have investigated the impact of draining lighter, more permeable soils. This may be partly due to the emphasis on draining clay soils, but also, no doubt, results from the greater practical difficulty encountered in plot definition where the soils are more permeable. Nevertheless, data from several drainage experiments on permeable soils are available. At Withernwick,[19] flow peaks were increased in the first year after drainage and there was then a reduction in the following years due to the progressive deterioration of the secondary system of subsoiling designed to improve the soil structure. Supporting evidence of increased peak flows following the drainage of more permeable soils also comes from studies at Cockle Park in northern Britain[20] and Ellingen in Central Germany.[21]

To identify factors influencing drainage response, the results of field drainage experiments under temperate northern European climates were analyzed in terms of their site characteristics.[22,23] This included topography, precipitation, drainage depth and spacing, natural (i.e., pre-drainage) soil water regime, and the soil properties. The only characteristic distinguishing sites, where drainage increased peak flows from those where they were reduced, were those relating to the soil water regime before drainage. The experimental sites all had similar land practices on the drained and the undrained land.

Drainage reduced peak flows on sites that had wetter soils and with poor natural drainage, and significant amounts of storm runoff were generated as overland flow and near-surface flow in the thin upper layers of the soil. These sites had higher topsoil clay contents and shallower depths to a poorly permeable subsoil horizon. When artificially drained, the surface saturation was largely eliminated, greatly increasing the soil water storage capacity.

In contrast, at sites with more permeable, loamy soils that were not routinely saturated before drainage, natural storm flow occurred predominantly by slower subsurface flow, and the artificial drainage pipes provided more rapid flow routes leading to increases in peak outflows.

The findings are summarized in Figure 1. This shows the topsoil texture, together with the effect of drainage on peak flows, and provides the engineer or conservationist with an initial guide to predict the effect on flows of the drainage of a site, based on knowledge of the predrainage site characteristics.

Further insights into the factors controlling the impact of drainage may be obtained by the application of modeling techniques to investigate the important interaction between soil properties and climate in determining soil water regimes. DRAINMOD[24] was applied to two of the field sites with similar climates: a heavy clay soil at Grendon and a more permeable loam at Withernwick. The model was applied to each site using actual field values of drain and soil parameters, and the simulated peak flows from drained and undrained land were compared for similar rainfall inputs. The results showed a 70% lower median peak flow after drainage of clay soil and an increase of 40% in the median peak flow from the more permeable land.[23]

The modeled fluxes and water stores confirmed that the reduction in peaks from the clay soil after drainage was achieved by a change in storm runoff generation from overland flow (caused by soil saturation) to subsurface drainflow. For the loamy soil, the model indicates that the increase in peak subsurface flow rates was due to the steeper hydraulic gradients created by the closer-spaced artificial drains.

The model also demonstrated the effect of different climatic conditions. If the loam soil site at Withernwick had double the normal rainfall (1200 mm yr^{-1} instead of 600 mm yr^{-1}), the resulting increase in ground wetness would be sufficient to generate substantial amounts of overland flow on the undrained land. Artificial drainage in this case would then reduce peak flows—exactly as happens for a clay soil (where, in contrast, the ground wetness is caused by the low soil permeability). Using the model in this way enables these effects of site characteristics to be explored in an objective manner. The overall dominant criterion—the amount and frequency of surface runoff from undrained land—can be assessed in terms of both soil properties and climatic characteristics.

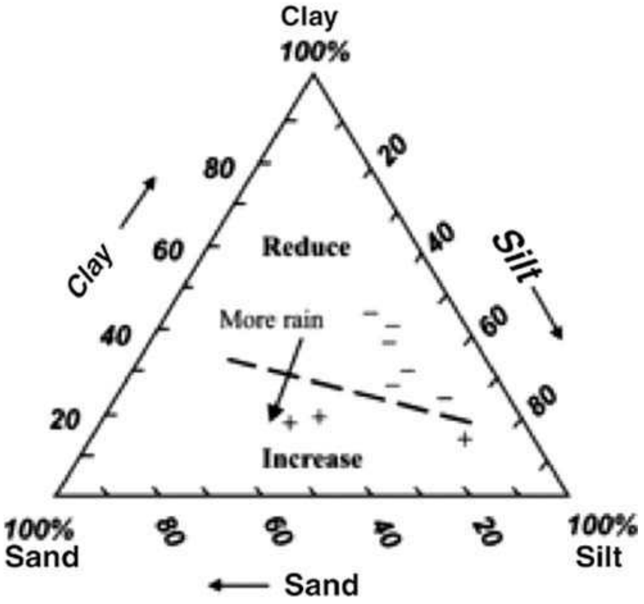


FIGURE 1 Observed impact of pipe drainage on downstream peak flows (increase/reduce), showing the importance of soil texture. Model simulations of climate changes indicate that higher rainfall and wetter ground conditions will shift the balance towards drainage schemes reducing peak flows. See text for details.

Conclusions

The effect of subsurface drainage on peak flows depends upon site wetness. If the water table is close to the surface (due to high rainfall or poor permeability), natural flows occur either over the surface or through the upper, more permeable layers of the soil. Drainage will increase soil water storage capacity and hence the amount of water that can infiltrate, thereby reducing surface runoff and peak storm flows. If the water table is deeper, due to a dry climate or due to more permeable soils, natural flows will occur through the body of the soil. In this case, artificial drainage will increase peak flows as a result of the shorter flow paths and steeper hydraulic gradients.

It must be noted that these conclusions depend upon the scale of the drainage considered. At the river catchment scale, main channel improvements will undoubtedly increase the speed of flow routing, and the timing of arrival of flows from different subcatchments will influence the peak discharge at the point of interest. The relative importance of field drainage and main drainage channels will vary with storm size: field drainage being dominant for small and medium storms, but main channel improvements becoming dominant for large events. In extreme situations where the rainfall intensity exceeds the infiltration capacity of the soil, the effects of the subsurface drains will be minimal, but the associated improved watercourses will rapidly carry away the surface runoff.

Overall, it seems likely that in large catchments, drainage schemes with substantial associated surface drainage and main channel improvements will lead to higher flow peaks downstream, even though locally, the effect of drainage may be to lower the peak flows.

References

1. Skaggs, R.W.; van Schilfgaarde, J., Eds. *Agricultural Drainage*; Agron Monograph 38; ASA, CSSA, and SSSA: Madison, WI, 1999.
2. Whiteley, H.R. Hydrologic implications of land drainage. Can. Water Res. J. **1979**, *4*, 12–19.
3. Serrano, S.E.; Whiteley, H.R.; Irwin, R.W. Effects of agricultural drainage on streamflow in the Middle Thames River, Ontario, 1949–1980. Can. J. Civil Eng. **1985**, *12,* 875–885.
4. Bailey, A.D.; Bree, T., Eds. Effect of improved land drainage on river flows. *Flood Studies Report—5 Years on*; Thomas Telford: London, 1981; 131–142.
5. Rycroft, D.W., Ed. The hydrological impact of land drainage. 4th International Drainage Workshop, Cairo; ICID- CHD, CEMAGREF; 1990; 189–197.
6. Oberlin, G. Influence du drainage et de l'assainissement rural sur l'hydrologie. CEMAGREF Bull. **1981**, *285,* 45–56.
7. Warmerdam, P.M.M., Ed. The effect of drainage improvement on the hydrological regime of a small representative catchment area in the Netherlands. *Application of Results from Representative and Experimental Basins*; UNESCO Press: Paris, 1982; 318–338.
8. Burke, W. Aspects of the hydrology of blanket peat in Ireland. Int. Assoc. Hydrol. Sci. **1975**, *105*, 171–181.
9. Wilcock, D.N. The hydrology of a peatland catchment in N Ireland following channel clearance and land drainage. In *Man's Impact on the Hydrological Cycle in the U.K.*; Hollis, G.E., Ed.; Geo Abstracts: Norwich, 1979; 93–107.
10. Seuna, P.; Kauppi, L., Eds. *Influence of Subdrainage on Water Quantity and Quality in a Cultivated Area in Finland*; Water Research Institute Publ. No. 43; Nat. Board of Waters: Helsinki, Finland 1981.
11. Harms, R.W. The effects of artificial subsurface drainage on flood discharge. In *Hydraulic Design in Water Resources Engineering: Land Drainage*; Smith, K.V.H., Rycroft, D.W., Eds.; Computational Mechanics Publication: Southampton, 1986; 189–198.
12. Denton, J. Bailey On the discharge from underdrainage and its effects on the arterial channels and outfalls of the country. Proc. Inst. Civil Eng. **1862**, *21*, 48–130.

13. Van der beken. The development of the theory and practice of land drainage in the 19th century. In *Water for the Future*; Wunderlich, W.O., Prins, J.E., Eds.; A.A. Balkema: Rotterdam 1987; 91–99.
14. Schwab, G.O.; Thiel, T.J.; Taylor, G.S.; Fouss, J.L. Tile and surface drainage of clay soils 1. Hydrologic performance with grass cover USDA. Agric. Res. Serv. Bull. **1963**, 935.
15. Robinson, M.; Beven, K.J. The effect of mole drainage on the hydrological response of a swelling clay soil. J. Hydrol. **1983**, *63*, 205–223.
16. Harris, G.L.; Goss, M.J.; Dowdell, R.J.; Howse, K.P.; Morgan, P. A study of mole drainage with simplified cultivation for autumn sown crops on a clay soil. II. Soil water regimes, water balances and nutrient loss in drain water, 1978–80. J. Agric. Sci. **1984**, *102,* 561–581.
17. Armstrong, A.C.; Garwood, E.A. Hydrological consequences of artificial drainage of grassland. Hydrol. Process. **1991**, *5*, 157–174.
18. Robinson, M.; Mulqueen, J.; Burke, W. On flows from a clay soil—Seasonal changes and the effect of mole drainage. J. Hydrol. **1987**, *91*, 339–350.
19. Robinson, M.; Ryder, E.L.; Ward, R.C. Influence on stream- flow of field drainage in a small agricultural catchment. J. Agric. Water Manage. **1985**, *10*, 145–148.
20. Armstrong, A.C., Ed. *The Hydrology and Water Quality of a Drained Clay Catchment—Cockle Park, Northumberland;* Report RD/FE/10; MAFF: London, 1983.
21. Schuch, M. Regulation of water regime of heavy soils by drainage, subsoiling and liming and water movement in this soil. In *International Institute for Land Reclamation and Improvement,* Proc. International Drainage Workshop; Wesseling, J., Ed.; Wageningen, Paper 1.14, 1978; 253–267.
22. Robinson, M., *Impact of Improved Land Drainage on River Flows*; Institute of Hydrology Report 113; Wallingford, U.K., 1990. ISBN 0–948540-24–9. http://nora.nerc.ac.uk/7349.
23. Robinson, M.; Rycroft, D.W. The impact of drainage on streamflow. *Agricultural Drainage*; Skaggs, R.W.; van Schilfgaarde, J., Eds.; Agron Monograph 38; ASA, CSSA, and SSSA: Madison, WI, 1999; Chap. 23, 767–800.
24. Skaggs, R.W. Drainage simulation models. *Agricultural Drainage*; Skaggs, R.W., van Schilfgaarde, J., Eds.; Agron Monograph 38; ASA, CSSA, and SSSA: Madison, WI, 1999; Chap. 13, 469–500.

44

Drainage: Soil Salinity Management

Glenn J. Hoffman

Introduction

Soil water must drain through the crop root zone when salinity is a hazard to prevent salts from increasing to levels detrimental to crop production. Drainage occurs whenever irrigation and rainfall provide soil water in excess of the soil's storage capacity. In humid regions, rainfall normally satisfies crop water requirements and precipitation infiltrating into the soil in excess of this requirement leaches (drains) salts present below the crop root zone. In subhumid areas, rainfall is often inadequate in amount or temporal distribution to satisfy crop needs and irrigation is implemented. For arid regions, rainfall is never abundant and the preponderance of the crop water requirement must be provided by irrigation. Regardless of the climate, if soluble salts are present, water in excess of that needed to satisfy the crop water requirement must be provided to leach excess salts. Leaching may be accomplished continuously or at intervals, depending on the degree of salinity control required. It may take decades or as little as one season, depending on the hydrogeology of the area, but without drainage, agricultural productivity cannot be sustained where salinity is a threat. For a more complete discussion on drainage design for salinity control, the reader is referred to Hoffman and Durnford.[1]

Drainage Conditions

All soils have an inherent ability to transmit soil water provided a hydraulic gradient exists. If the hydraulic gradient is positive downward, drainage occurs. Soils with compacted layers, fine texture, or layers of low hydraulic conductivity may be so restrictive to downward water movement that drainage is insufficient to remove excess salts. In some areas, the hydrogeology may be such that the hydraulic gradients are predominantly upward. This leads to water logging and salination.

Before designing a man-made drainage system, the natural drainage rate should be determined. If the natural hydraulic gradient causes soil water to drain out of the crop root zone, the capacity of the artificial system can be reduced, thereby decreasing the cost for drainage. In some situations, upward flow into the crop root zone from a shallow aquifer can significantly increase the drainage requirement.

The upward movement of groundwater leads to salination as the water evaporates at the soil surface, leaving salts behind. If upward flow is ignored, the drainage system may be inadequate. Regardless of the source, an artificial drainage system will not function unless it is below the surface of the water table.

Drainage Requirement

Saline Soils

The amount of drainage required to maintain a viable irrigated agriculture depends on the salt content of the irrigation water, soil, and groundwater; crop salt tolerance; climate; soil properties; and management. At present, the only economical means of controlling soil salinity is to ensure an adequate net downward flow of water through the crop root zone to a suitable disposal site. If drainage is inadequate, harmful amounts of salt can accumulate.

In irrigated agriculture, water is supplied to the crop from irrigation, rainfall, snow melt, and upward flow from groundwater. Water is lost through evaporation, transpiration, and drainage. The difference between water inflows and outflows is the change in soil water storage. A water balance, expressed in terms of equivalent depths (D) of water, can be written as

$$D_s = D_i + D_r + D_g - D_e - D_t - D_d \tag{1}$$

where the subscripts s, i, r, g, e, t, and d designate storage, irrigation, rainfall and snow melt, groundwater, evaporation, transpiration, and drainage, respectively. The corresponding salt balance, where S is the amount of salt and C is salt concentration, can be expressed as

$$S_s = D_iC_i + D_rC_r + D_gC_g + S_m + S_F - D_dC_d - S_p - S_c \tag{2}$$

with S_s being salt storage, S_m is the salt dissolved from minerals in the soil, S_f indicates salt added as fertilizer or amendment, S_p is precipitated salts, and S_c is the salt removed in the harvested crop.

Rarely do conditions prevail long enough for steady state to exist in the crop root zone. However, it is instructive to assume steady state to understand the relationship between drainage and salinity. If upward movement of salt, the term ($S_m+S_f-S_p-S_c$), and the change in salt storage are all essentially zero, then the salt balance Eq. (2) can be reduced to

$$D_dC_d = D_iC_i + D_rC_r \tag{3}$$

The leaching fraction, L, is the ratio of the amount of water draining below the crop root zone, D_d, and the amount applied, D_i+D_r. The ratio of the salt concentration entering and leaving the root zone can also be used to estimate L. Since C_r is essentially zero.

$$L = C_i/C_d = D_d/D_i + D_r \tag{4}$$

The concept in Eq. (4) is important because it illustrates the relationship between leaching fraction and salinity.

The minimum leaching fraction that a crop can endure without yield reduction is termed the leaching requirement, L_r. The leaching requirement is the minimum amount of drainage required to prevent excess accumulations of salt that result in loss of crop yield. Several models have been proposed to estimate the drainage (leaching) requirement. Of the four models tested,[2] the one presented in Figure 1 agrees well with measured values of the drainage requirement through the range of agricultural interest. The drainage requirement given in Figure 1 is the fraction of the volume of applied water that must pass through the crop root zone as a function of the salinity of the applied water and the salt tolerance of the crop.

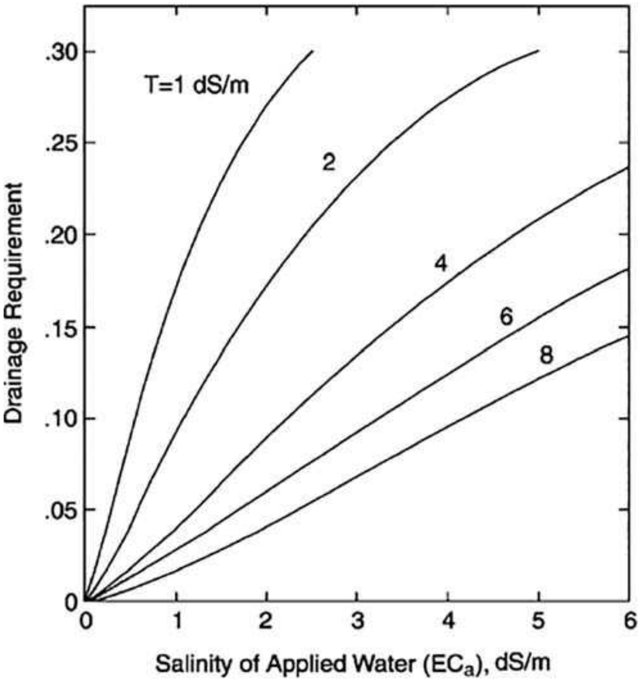


FIGURE 1 Drainage requirement as a function of the salinity of the applied water (reported as the volume weighted electrical conductivity) and the salt tolerance threshold value for the crop (T).
Source: Adapted from Hoffman and Van Genuchten.[15]

Sodic Soils

A soil is said to be sodic if an excessive concentration of sodium causes a deterioration of soil structure. The impact of excess sodium is a reduction in hydraulic conductivity and crust formation. Sodic conditions decrease the rate of drainage. Before a sodic soil can be restored to full productivity the excess sodium in the soil must be replaced with calcium or magnesium. This process frequently requires copious amounts of leaching to reclaim the soil. The design of an artificial drainage system that may be required, however, is based upon the long-term requirement for drainage as estimated in Figure 1 rather than the anticipated high drainage requirements for reclaiming a sodic soil.

Drainage System Design

There are three types of subsurface systems used to control soil salinity: relief drains, shallow wells, and interceptor drains. Relief drains, usually consisting of perforated corrugated plastic tubes buried in a regularly spaced pattern, is the most common subsurface system. Laterals for relief drains are typically placed 2.0–3.5 m deep and are spaced horizontally ten to hundreds of meters apart where salinity is a hazard. Shallow wells, called tube wells in some regions, can also be used to lower the water table by allowing pumping from shallow, unconfined aquifers. Tube wells are spaced at distances of a few hundred meters to several kilometers and may be a few meters to a hundred meters deep. Interceptor drains are used to remove excess soil water from saline seeps. Frequently, one subsurface drain, properly located at the upslope side of the seep, is sufficient. Regardless of the type of drainage system, the depth of the water table must be maintained low enough that (1), salts in the soil profile move to the water table (2), the rate of water movement by capillary flow to the soil surface because of evaporation is minimal, and (3), upflow of saline groundwater into the root zone is prevented.

Relief Drains

A relief drainage system consists of a main drain, collector drains, and field drains (laterals). The main drain is frequently a surface stream or an open drainage canal. Collectors and laterals are usually buried in a regular parallel pattern. Either open ditches or perforated pipes can serve as collectors and laterals. Open ditches are not normally installed now because they occupy land, are difficult to maintain, and are only capable of shallow drainage. Laterals are up to 300 m long and terminate in a collector drain. Both single- and double-sided entries by laterals into a collector are common.

Drain Depth

Subsurface drains are installed much deeper for salinity control in arid regions than drains for water table control in humid regions. The goal for salinity control is to place the drains deep to limit salination of the root zone by capillary upflow. Drains are placed at depths of 2.0–3.5 m in arid regions.[3] The appropriate drain depth depends upon the depth capacity of the installation machinery, the location of a shallow soil layer that impedes water movement, and anticipated benefits compared to additional costs of deeper installation.

Drain Spacing

The spacing between laterals is often estimated using simple drainage design equations. Drain spacing determinations can be based on criteria of steady-state, falling- water-table, or fluctuating-water-table conditions.[4] For large drainage projects or where more accurate values are desired, computerized drainage design models are available. An early computer model developed by Skaggs[5] has been altered by several for irrigated conditions.[6,7] Other models present drainage designs for irrigated areas based on optimization,[8] decision support systems,[9] or reuse of drainage water.[10]

Drainage Wells

Shallow or tube wells offer a viable alternative to relief drains when the aquifer has sufficient transmissivity to provide a significant yield of drain water and the vertical permeability between the crop root zone and the aquifer is adequate. Under these conditions, tube wells have the advantages of being able to lower the water table to greater depths than relief drains and also provide supplemental water for irrigation if the quality is appropriate.

Because drainage wells can be installed at convenient locations within the area to be drained and can be operated either continuously or intermittently, the management of a system of drainage wells is more versatile than relief drains. Relief drains are typically a passive drainage system relying on gravity and designed to operate continuously.

Economic comparisons between the costs of drainage wells and relief drains vary. It is generally found that relief drains have lower construction and operation costs.[11] However, Mohtadullah[12] showed tube wells were a better economic choice than relief drains for the Indus Basin.

Saline Seeps

The occurrence of saline water at the soil surface downslope from a recharge area is referred to as a saline seep. Saline seeps can occur because of the reduction of evapotranspiration that occurs when grasses or forests are converted to cropland in the upland (recharge) areas of a watershed. Dryland farming practices that include fallow periods tend to aggravate the seepage problem. Salination occurs as water infiltrating in the upper elevations of the watershed moves through salt-laden substrate on its path to a discharge site at a lower elevation. In the discharge area of the seep, crop growth is reduced or the plants killed by an intolerable level of salinity. Saline seeps can be distinguished from other saline soil conditions by their recent origin, relatively local extent, saturated soil profile, and sensitivity to precipitation and cropping systems.[13] Saline seeps occur throughout the Great Plains of North American and in Australia, India, Iran, Turkey, and Latin America.[14]

Planting crops in the recharge area that consume soil water before it percolates below the crop root zone will prevent saline seeps. Failing this, improved drainage may provide a solution. Installing an interceptor subsurface drain immediately upslope from the saline seep is frequently a successful solution. Interceptor drains to control seepage should be installed as deep as practical. If the layer restricting soil water flow is not too deep, placing the interceptor drain just above this layer is the most effective location.

References

1. Hoffman, G.J.; Durnford, D.S. Drainage design for salinity control. *Agricultural Drainage;* Agronomy Monograph No. 38; American Society of Agronomy: Madison, WI, 1999; 579–614; Chap. 17.
2. Hoffman, G.J. Drainage required to manage salinity. J. Irrig. Drain. Div., American Society of Civil Engineers, New York **1985**, 111, 199–206.
3. Ochs, W.J. *Project Drainage Issues,* Proceedings of the 3rd International Workshop on Land Drainage, Columbus, OH, Dec 7–11, 1987; E-83–E-88. 202.
4. Bouwer, H. Developing drainage design criteria. *Drainage for Agriculture;* Agronomy Monograph No. 17; American Society of Agronomy: Madison, WI, 1974; 67–79.
5. Skaggs, R.W. *A Water Management Model for Artificially Drained Soils,* Report 267; Water Resources Research Institute, North Carolina State University: Raleigh, NC, 1980.
6. Chang, A.C.; Hermsmeir, L.F.; Johnston, W.R. *Application of DRAINMOD on Irrigated Cropland,* Paper No. 81–2543; American Society of Agricultural Engineers: St. Joseph, MI, 1981.
7. Skaggs, R.W.; Chescheir, G.M. Application of drainage simulation models. *Agricultural Drainage*; Agronomy Monograph No. 38; American Society of Agronomy: Madison, WI, 1999; 529–556; Chap. 15.
8. Knapp, K.C.; Wichlens, D. Dynamic optimization models for salinity and drainage management. *Agricultural Salinity Assessment and Management*; American Society of Civil Engineers: New York, 1990; 530–548; Chap. 25.
9. Gates, T.K.; Wets, R.J.-B; Grisiner, M.E. Stochastic approximation applied to optimal irrigation and drainage planning. J. Irrig. Drain. Div., Am. Soc. Civil Eng. **1989**, *115,* 488–510.
10. El-Din El-Quosy, D.; Rijtema, P.E.; Boels, D.; Abdel- Khalik, M.; Roest, C.W.J.; Adbel-Gawad, S. Prediction of the quantity and quality of drainage water by the use of mathematical modeling. *Land Drainage in Egypt*; Drainage Research Institute: Cairo, Egypt, 1989; 207–241.
11. Zhang, W. Drainage inputs and analysis in water master plans. Proceedings of the 4th International Drainage Workshop, Cairo, Egypt, Feb 1990; 181–187.
12. Mohtadullah, K. Interdisciplinary planning, data needs and evaluation for drainage projects. Proceedings of the 4th International Drainage Workshop, Cairo, Egypt, Feb 1990; 127–140.
13. Brown, P.L.; Halvorson, A.D.; Siddoway, F.H.; Mayland, H.F.; Miller, M.R. *Saline-Seep Diagnosis, Control, and Reclamation,* USDA Conservation Research Report No. 30; 1983.
14. Halvorson, A.D. Management of dryland saline seeps. *Agricultural Salinity Assessment and Management;* American Society of Civil Engineers: New York, 1990; 372–392; Chap. 15.
15. Hoffman, G.J.; Van Genuchten, M.Th. Soil properties and efficient water use: water management for salinity control. *Limitations to Efficient Water Use in Crop Production;* American Society of Agronomy: Madison, WI, 1983; 73–85; Chap. 2C.

45

Lakes: Restoration

Anna Rabajczyk

Introduction

The 21st century is anticipated as the century of the environment. In this regard, it is crucial to remedy pollution and conserve resources, particular with respect to aqueous ecology, for fostering a healthy environment. Water reservoirs, whether natural or artificial, differ in methods of lake basin formation. The former exist as a result of physical and physicochemical processes occurring in nature; the latter appear due to human interference. Natural water reservoirs of relatively slow exchange of the liquid found therein are defined as lakes.[1,2]

Artificial reservoirs, in turn, also referred to as barrier lakes, are formed by closing river valleys with water dams or drops as damming structures. Several of these reservoirs may also be natural barrier lakes with regulated outflows.[3]

According to their functions, three main types of water reservoirs may be distinguished:[2]

- Dry reservoirs: periodical storage of water during the passage of flood peaks
- Flow-through reservoirs: maintenance of steady retention levels
- Retention reservoirs: storage of water at times of surplus in order to utilize it at other times (subtypes include flood control, navigation, power engineering, water equalization, municipal, industrial, agricultural, or rock waste control reservoirs)

Another classification, which takes into account land configuration, includes mountain, submountain, and lowland reservoirs. The first of these are usually the deepest, with short and high dams. The lowland ones occupy the largest areas and are less deep.[2]

Such considerable diversification of both formation methods and localities causes lakes to be characterized by varying parameters of lake basin morphometry, water transfer speed and routes, water quality therein, etc. Lake shape and dimensions condition lake susceptibility to degradation. Wind speed is higher over large lakes than over small ones, which results in stronger waving and better water oxygenation. Waters in lakes whose elongation follows the most frequent wind directions are also better mixed than in those with transverse orientation.[4]

Low values of the mean-to-maximum depth ratio suggest that large depths are found in a small part of the lake; hence, they may not be of considerable importance for lake functioning. The occurrence of thermal stratification and the size of the littoral zone that uses the biogenes found in the water are both dependent on lake depth.

Each lake, together with the surrounding land and watercourses, forms a catchment characterized by parameters such as topographic features, plant cover, basal complex, and climate conditions, which determine the volume of water resources in a given area. A catchment shaped by low atmospheric precipitation and lowland topographic features provides less water than one with high precipitation and mountainous topographic features, facilitating surface wash. The direction of precipitation waters, accompanied by a variety of compounds, is determined by the basal complex that conditions the river network density and, thus, inflow and outflow volumes. The vegetation found in a particular area stores the water and then gives it off to the atmosphere via the process of transpiration.[2]

At moderate latitudes, water exchange in reservoirs occurs twice yearly: in spring and in autumn. During spring circulation, winter-specific temperature distribution changes when bottom temperatures reach 4°C and upper layers of the reservoir become cooler. As a result of more intensive solar radiation and air warming, surface water temperatures rise. As they reach 4°C, the temperatures of the vertical section equalize because water temperature at the reservoir bottom is 4°C.[5]

As a consequence of wind-induced circulation currents, the entire water mass is mixed (spring homothermy). Further warming of surface waters causes near-surface waters to become lighter than the near-bottom waters whose temperature reaches 4°C (summer stagnation), and an upper warm layer called the epilimnion is formed. The lower layer of cold water is called the hypolimnion. In the metalimnion, i.e., the lake middle or transition water layer, a sharp temperature drop occurs at that time.[6]

In autumn, surface waters cool until they reach a state in which, upon reaching the temperature of 4°C, they attain density identical to that of near-bottom waters. In a way analogous to spring, waters are mixed by wind-induced motion (autumn homothermy). Further cooling of surface waters makes them obtain temperature that approximates 0°C, as well as lesser density than that of near-bottom waters. Subsequently, on the surface, the ice cover is formed (winter stagnation).[5]

Thermal stratification is related to oxygen content stratification in reservoir waters, thus affecting organic life growth. Because the epilimnion is a layer of intensive wind-induced water mixing as well as easy solar radiation access, it displays better oxygenation than deeper layers. As a result of intensive photosynthesis, in summer, water may become supersaturated with oxygen. Spring and autumn circulations cause oxygen to be distributed over the entire water mass.[5,7]

Temperature also conditions other processes that occur in the lake, including the rates of all chemical, biological, and physical reactions.[5] Higher temperature, for example, intensifies the processes of phosphorus release from bottom sediments (Figure 1).

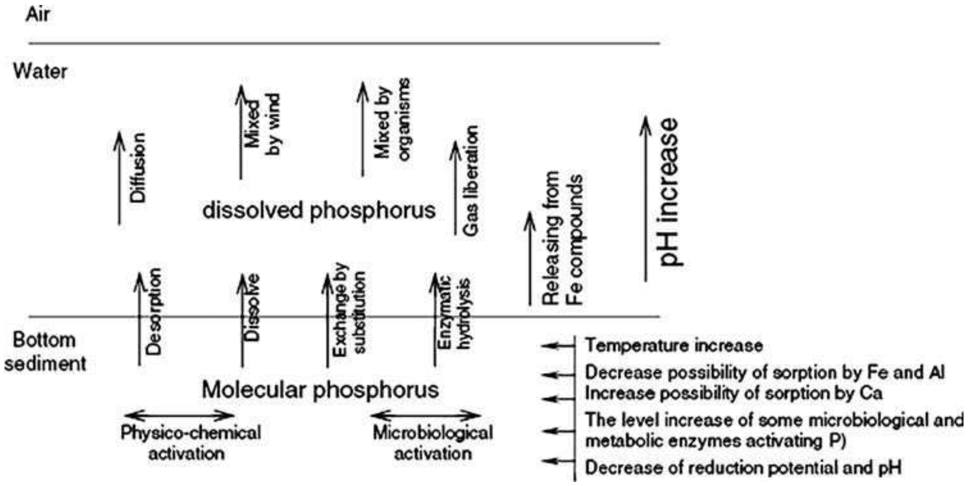


FIGURE 1 Phosphorus release from bottom sediments affected by temperature.
Source: Bajtkiewicz-Grabowska and Zdanowski.[8]

Water temperature increase in the reservoir decreases water viscosity and density. This is particularly significant for plankton organisms as it accelerates the sedimentation speed and changes locomotive conditions. Consequently, the rates of chemical and biochemical reactions increase, and solubility of most substances is enhanced as well, while gas solubility is reduced. The intensification of the processes occurring in the reservoir results in increased use of oxygen. Photosynthetic processes are accelerated, which may lead to cyclical supersaturation of upper water layers. A sudden water temperature rise by 10°C may lead to thermal shock or even death of organisms.[9]

Temperature increase may cause spontaneous discharge of bottom sediments and intensify gas liberation to the extent that the material collected at the bottom is loosened and discharged into water.

Access to solar radiation conditions the growth of autotrophs in the water, as well as that of the organisms situated behind them in the food chain, because the fluctuations in producer numbers affect level 1 (and further) consumer numbers. The propagation of solar radiation in the lake is affected by lake size and shape, i.e., morphometry. For instance, it determines thermal stratification of lake waters as well as chemical, physical, and biological processes that occur in reservoirs.[10]

Natural water reservoirs are efficiently operating ecosystems, capable of maintaining internal homeostasis despite the operation of adverse factors. Yet, human proximity and in particular the effects of human existence have far-reaching consequences (frequently negative) for the environment. Human activity takes forms against which a lake cannot defend naturally. Pollutants rich in biogenic elements from a variety of sources (including municipal waste, agricultural fertilizers, degraded woodland and communication routes) are discharged into lakes and rivers, causing considerable increase in the fertility of lake waters and rivers that serve as lake outflows. In raw municipal waste, for instance, nitrogen concentrations are contained in the 25–50 mgdm^{-3} range, while total phosphorus concentrations are in the range 4.0–12 mgdm^{-3}.[3] According to Vollenweider,[11] limiting concentrations of phosphorus and nitrogen compounds in waters, in excess of which mass algal growth may occur, amount to 0.01 mg Pdm^{-3} and 0.3 mg Ndm^{-3}, respectively.

For the assessment of external factors affecting a particular lake, an assessment of the catchment in terms of yearly delivery of biogenic (N and P) compounds into lake waters must be performed. The calculations are based on the formulas:[3,11]

$$L = I/P_j \left(g \cdot yr^{-1} \cdot m^{-2} \right) \tag{1}$$

$$I = I_{pr} + I_{pl} + I_{pz} + I_o + I_k + I_w + I_l \left(kg \cdot yr^{-1} \right) \tag{2}$$

where P_j is lake area (m^2), I is total load, I_{pr} is grassland load, I_{pl} is woodland load, I_{pz} is built-up area load, I_o is atmospheric precipitation load, I_k is bathers load, I_w is anglers load, and I_l is linear source load.

The values of external nitrogen and phosphorus loading in a particular lake are then compared with Vollenweider's criteria,[11] which define permissible and hazardous nitrogen and phosphorus loadings to a lake:

$$L_{d(\mathrm{N})} = 15 \left(25\ z^{0.6} \right)$$

$$L_{\mathrm{n(N)}} = 15 \left(50\ z^{0.6} \right)$$

$$L_{\mathrm{d(P)}} = 25\ z^{0.6}$$

$$L_{\mathrm{n(P)}} = 50\ z^{0.6}$$

where z is mean lake depth, $L_{\mathrm{d(N)}}$ is permissible reservoir N loading ($\mathrm{mg \cdot yr^{-1} \cdot m^{-2}}$), $L_{\mathrm{d(P)}}$ is permissible reservoir P loading ($\mathrm{mg \cdot yr^{-1} \cdot m^{-2}}$), $L_{\mathrm{n(N)}}$ is hazardous reservoir N loading ($\mathrm{mg \cdot yr^{-1} \cdot m^{-2}}$), $L_{\mathrm{n(P)}}$ is hazardous reservoir P loading ($mg \cdot yr^{-1} \cdot m^{-2}$), 25 is the limiting loading rate for oligo- and mesotrophic reservoirs, and 50 is the limiting loading rate for meso- and eutrophic reservoirs.

External loading of a reservoir with biogenic compounds per area unit is a major indicator specifying the reservoir's trophy. Trophy increase is based primarily on increased concentrations of biogenic compounds such as nitrogen and phosphorus, and this phenomenon is referred to as eutrophication. It is worth noting that the process of eutrophication is a most natural phenomenon. Organic substances are delivered to the lake starting from the moment of its formation and deposited at the bottom in the form of bottom sediments. Yet, due to intensified pollutant inflow, the phenomenon of eutrophication begins to intensify until it becomes a major threat to lakes.

Recultivation Methods

The growing eutrophication rate of water reservoirs and the increasing number of degraded reservoirs create the demand for effective prevention methods. In the 1960s, attempts at recultivation of degraded lakes were made. However, adequate know-how resulting from experience in this field was lacking, particularly because each water reservoir operates under different internal and external conditions. The current know-how and further developments in other fields of science, not only in natural sciences, have made it possible to explore theoretical issues related to water reservoir recultivation on the basis of the practical experience obtained.

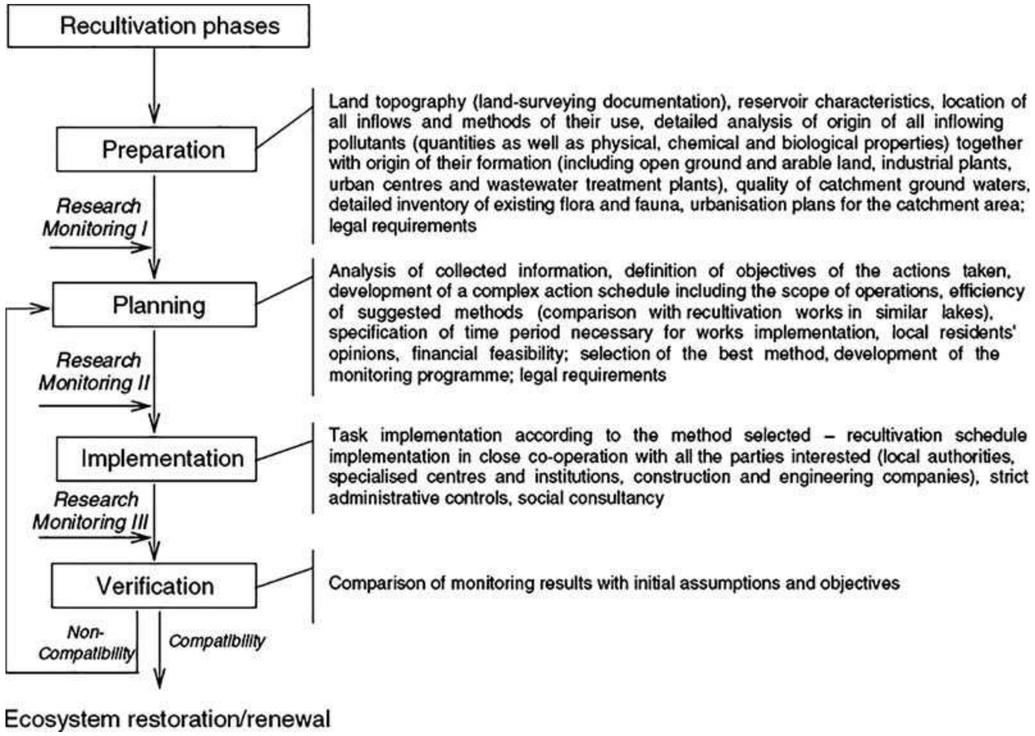
As water ecosystems, lakes are located in land depressions; hence, they provide natural receiving water for the pollutants that come from the catchment area. Due to the functions that they perform in human environment and economy, it is crucial to take appropriate steps in order to improve the quality of degraded waters or protect those in better condition. To this end, scientific, organizational, and technological actions are undertaken jointly, known as recultivation. This process consists in elimination of possible chemical contamination of waters and sediments, improvement of oxygen and nutrient balance, maintenance of flora and fauna at a level appropriate for a given water ecosystem, and provision of engineering elements of flow regulation in the form of desired depths, erosion control, and speeds.[12–15]

The idea of lake recultivation is to restore the previous functions of lakes (e.g., water storage, recreation, household, and agriculture), as well as physical, chemical, and biological features approximating as closely as possible the natural ones. The selection of an appropriate method is determined by the diversity of individual lakes, differences in the ways and scopes of pollution, together with their location in a catchment. The type of the recultivation method used depends on the reservoir size, the nature of the fauna and flora resident therein, connection with watercourses, and proximity to clean water reservoirs in the vicinity of the reservoir under recultivation.[12,14]

Recultivation of water reservoirs is a four-phase process, including the phases of preparation, planning, implementation, and verification monitoring (Figure 2). The duration of actions aimed at restoration of utility values to a particular area, from the initial phase to the final stage of implementation, is typically measured in years.

The first stage is the collection of information as well as the development of a detailed inventory of the immediate catchment. The following may also be helpful: soil maps; maps of localities with erosion-susceptible land and soils, utility or arable lands adjacent to rivers and reservoirs, and critical terrain (with its specific problems); maps of areas where surface waters supply groundwaters; and urbanization plans for the building in catchment areas and industrial development trends.

In view of the above, an individualized approach to each reservoir and planning a complete ecosystem restoration process are necessary. While taking any recultivation actions, it should be remembered that physical, chemical, external, and internal environmental factors may cause profound changes, destabilize the water ecosystem, and push it into one of the two alternative stable states. It must also



Land topography (land-surveying documentation), reservoir characteristics, location of all inflows and methods of their use, detailed analysis of origin of all inflowing pollutants (quantities as well as physical, chemical and biological properties) together with origin of their formation (including open ground and arable land, industrial plants, urban centres and wastewater treatment plants), quality of catchment ground waters, detailed inventory of existing flora and fauna, urbanisation plans for the catchment area; legal requirements

Analysis of collected information, definition of objectives of the actions taken, development of a complex action schedule including the scope of operations, efficiency of suggested methods (comparison with recultivation works in similar lakes), specification of time period necessary for works implementation, local residents' opinions, financial feasibility; selection of the best method, development of the monitoring programme; legal requirements

Task implementation according to the method selected – recultivation schedule implementation in close co-operation with all the parties interested (local authorities, specialised centres and institutions, construction and engineering companies), strict administrative controls, social consultancy

Comparison of monitoring results with initial assumptions and objectives

FIGURE 2 Recultivation phases.

be borne in mind that biological elements and processes conditioned thereby, which occur in a water ecosystem, stabilize and perpetuate its status quo.

A significant element of recultivation is the specification of a monitoring program, i.e., catchment control at critical points as well as in terms of registration of environmental quality indicators important for the recultivation process. This program assumes the establishment of observation sites, equipped with control and measurement apparatus. After a set monitoring time, all data from field observation sites are covered in a technical study that enables developing a representation of a catchment demonstrating the origin of possible pollutants and the problems to be tackled. Following from an extensive data base, including input from various sources, and complete with monitoring program results, it becomes possible to enter the definition phase for particular elements and actions necessary to achieve recultivation effect for a particular reservoir.

Nevertheless, it must be stressed that Research Monitoring I usually differs from the subsequent phases of monitoring, which results from task implementation at individual stages. During the inventorying process, maximum quantities of data concerning pollutant sources and migration routes, as well as the quantity and quality of substances delivered, ought to be collected. In contrast, the objective of Monitoring II is to obtain information on whether the selected recultivation method is well adjusted to the lake's individual characteristics.

To improve the situation in a degraded lake, it is not enough to restore its condition to that from before the disturbance: much feedback in the lake perpetuates the postchange status, and lakes tend to demonstrate resilience, i.e., immunity to recultivation actions. Very often, the condition for the application of any recultivation method is prior reduction of phosphorus concentration in the water to the 0.050–0.100 mg $PO_4 \cdot dm^{-3}$ level and phosphorus removal/ inactivation in sediments (internal import inhibition).[12–14]

TABLE 1 Most Commonly Used Lake Recultivation Methods and Techniques with Their Characteristics

Method/Technique	Characteristics	Model Application	Literature
Dredging	Bottom sediment removal from the entire lake, deepest waters and inflow (in flow-through lakes); complete sediment removal to reach mother rock guarantees radical improvement in water quality; necessity of thorough physico-chemical tests to determine the chemical composition of the sediments, their thickness and distribution in the reservoir (specification which sediment layer and in which lake parts ought to be removed); a sensitive issue is to find sites for storage and management of extracted sediments treated as hazardous	Mission Lake, Kansas; Lake Elkhorn, Columbia; Saluda Lake, South Carolina; Chain Lake, British Columbia	[16–19]
Macrophyte reintroduction	Bentonite is a good macrophyte substratum, which enables macrophyte layer reconstruction as well as restoration of destroyed trophy web structures	Alderfen, Barton, Belaugh and Cockshoot Broads, Norfolk;	[20]
Flushing, rewashing	Provision of clean, nutrient-deficient water, removal of strongly eutrophicated hypolimnion waters and replacing them with well-oxygenated waters from outside the lake; requires a source of clean water in the vicinity of the lake; lake water exchange ought to be performed a couple of times a year	Lake Veluwe, the Netherlands; Willow Lake, Arizona	[21,22]
Macrophyte removal	Comprises artificial lowering of lake water level, lake icing in winter, in-freezing of plant stalks into ice, raising the river level and plant uprooting, collection of floating plant remains and removal from the lake; the quantities of phosphorus thus removed are much lower than the loads introduced to the reservoir; effective in the case of lakes which are strongly polluted with biogenes, accumulated mostly in littoral vegetation	Lakes: Mary, Ida and Bass, Minnesota; Boreal Lake, Ontario	[23,24]
Hypolimnion water removal	Procedure possible in flow-through reservoirs; consists in liquid removal from the profundal layer via a hose; a disadvantage of this method is the pollution of the watercourse discharged from the lake; during near-bottom water removal, the sediments which remain in chemical balance with them become depleted; the process may cause reservoir pollution with H_2S, accumulated in the sediments and temperature increase near the bottom, which leads to hypolimnion oxygen depletion	Bled, Yugoslavia; Lake Kortowo, Olsztyn	[14,25]
Sparging, oxygenation (aeration)	Introduction of a duct into the water through which air is pumped, without disturbance (near-bottom liquid temperature is low: slow rate of organic matter decomposition) or with disturbance of natural lake stratification (entire water mass mixing and water oxygenation: near-bottom layer temperature increase, intensification of chemical and biological processes, including enhanced mineralization and internal supply); enhances hypolimnion oxygen conditions as well as enforcing lake water circulation	Canyon Lake, Teksas; Lake Commabbio, Lombardy; Kieleckie Bay, Kielce	[26–28]

(*Continued*)

TABLE 1 (*Continued*) Most Commonly Used Lake Recultivation Methods and Techniques with Their Characteristics

Method/Technique	Characteristics	Model Application	Literature
Artificial destratification	Destruction or prevention of the lake's thermal stratification, stimulating mass algal growth in the top (warmed and lighted) water layer; very often linked to sparging	Canyon Lake, California; Lake Catherine, Arkansas; Lake Starodworskie, Olsztyn	[29–31]
Nutrient deposition and deactivation in water and sediments	Possible in small reservoirs; does not guarantee permanent phosphorus removal from water; enduring results depend on coagulant properties ($FeCl_3$, $Al_2(SO_4)_3$, $FeSO_4$); optimum water pH 6–8 and high Redox potential level; phosphorus inactivation in sediments is possible due to application of a mixture of bentonite clay with lanthanum, which becomes a phosphorus bonding element upon adsorption on bentonite	Kielecki Bay, Kielce; Jessie Lake, Alberta; Lake Głęboczek, Tuchola; Lake Sønderby, Sønderby	[28,32–34]
Sediment isolation (capping)	Possible in small reservoirs where water transfer is low and does not cause mobility of the introduced layer; used in order to counter chemical exchange in the sediments-water system; physical isolation: sand and foil; chemical isolation: $+Al^{+3}$ diatomite, bentonite or other clayey minerals	Great Lakes, North America; Onondaga Lake, Central New York; Venice Lagoon, Porto Marghera	[35–37]
Precipitation (chemical methods)	Use of herbicides, including algaecides (e.g. barley straw, tree leaves) to fight algae; short-lasting procedure results due to appearance of other algae in place of the removed blooms; causes introduction of large amount of hazardous substances into water	Shoecraft Lake, Washington; Eight lakes located in the eastern and western portions of southern Michigan; Joliet Junior College Lake (JJC Lake), Illinois	[38–40]
Biomanipulation	Change in living conditions of organisms or quantitative ratios in a given ecosystem by means of several food chain dependencies: limiting the populations or complete elimination of individual groups of organisms (e.g. increase in the amount of zooplankton and introduction of selected fish species in order to reduce algal population, introduction of silver carp to the lake to limit the growth of phytoplankton; introduction of white amur to eliminate excess of macrophysical vegetation; introduction of predators such as pike, pike perch or perch in order to limit plankton-eating fish populations); use of biomanipulation requires thorough analysis of dependencies occurring in a given ecosystem	Lake Terra Nova, Loosdrecht Lakes ; Lake Eymir, Ankara	[41–43]
Fish catch	Regular catch of fast-growing fish (important is good cooperation with the lake management) enables reducing lake fertility (so-called trophy) and systemic enhancement of water quality; effective in the case of lakes which are strongly polluted with biogenes, accumulated mostly in the ichtiofauna	Lake Fure, Copenhagen; Lake Ringsjön, Skåne County	[44,45]
Seston removal/ catch	Pumping out water at sites richest in seston, its subsequent filtration in appropriate apparatus; in barrier lakes, zooplankton removal may be accomplished by water inflowing through bottom culverts when zooplankton density is highest at the bottom	Lake Jussi, Pikkjärv	[46]

Source: Cooke,[12] Klapper,[13] Sengupta and Dalwani,[14] and Gupta et al.[15]

As a rule, each catchment and its environment are distinct, but from a practical viewpoint, it is possible to distinguish several actions, techniques, and methods enabling renewal or protection of the status quo of the lake at issue (Table 1).

The most popular methods that provide significant effects in biogene removal are those based on sparging and dredging. Reservoir water sparging enables reduction of excessive algal and waterweed growth. Aeration is performed with the use of either mechanical apparatus or compressed air. Water sparging by means of mechanical apparatus may be performed by water spraying or direct penetration of air into the reservoir.[26,27]

Spraying occurs when water passes along a perforated hose through which it is catapulted into the air in the form of fine droplets. This effect can also be achieved by directing water to purpose-made diffusers. The most important element of this technique is the generation of adequately sized water droplets. The finer the water droplets that penetrate into the atmosphere, the larger the area of gas exchange between the droplet and the air, thanks to which better aeration effects are obtained—more oxygen is transferred into the water.[12–14]

The apparatus used for direct transfer of oxygen into a reservoir includes, for example, aerators with horizontal and vertical rotation axes, including surface, turbine- driven subsurface, and combination aerators. Aerators may also be classified according to paddle-wheel shapes: some may be straight, whereas others have single- or multisided curvatures.

The simplest is the surface aerator. Its operation relies on a turbine or rotor pumping the water upwards and then downwards, thus causing turbulence and splashing on water surface. Oxygen is delivered through water droplets in contact with atmospheric air. This type of aerator is used in ponds, water reservoirs, and rivers.

Combination aerators, which display features of both surface and turbine-driven aerators, are characterized by high capacity for oxygen transfer and provide good sparging results in deep reservoirs. In this type of devices, the surface rotor and the turbine are usually powered by the same aggregator, which causes high power demand, high noise levels, and possible icing during operation.[13,14]

The other types of aeration devices have vertical axes, which enables operation in both open and closed setups. In open aerators, the liquid is centrifugally catapulted, which causes the occurrence of the suction force that sucks the air into the so-called hopper and inter-paddle-wheel space. At the outlet of this space, a mixture of air and liquid appears and is then splashed on the surface of the aerating chamber. This construction provides for the mixing of the liquid and the air in the entire volume. In closed aerators, the liquid is propelled onto paddle-wheels by means of a pump. The air is thus sucked from the atmosphere by means of suction resulting from the flowing liquid at set sites in the inter-paddle-wheel channel.[47]

The compressed air aerator consists of a blower or a fan, sources of compressed air, and distribution ducts and sparging devices. In order to obtain the largest possible contact area between the two phases, a variety of porous elements are used for air dispersion. They are made of ceramic or artificial materials, and are pipe or plate shaped.[47]

In the process of selection of lake aeration systems and methods, the following factors ought to be taken into consideration:

- Oxygen concentration in the water and at the bottom of water reservoirs, which depends on water reservoir eutrophication degree, oxygen deficit volume in bottom sediments, with allowances being made for sediment resuspension, temperature, spring and autumn circulation intensity, and possible waving effects.
- Nominal result of aerator operation, which depends on total nominal hydraulic efficiency, reservoir depth at installation site, reservoir area, air volume fed into the aerator, and aerator construction type, such as the following:
  - Pump and surface dispersion
  - Pump and reversal-to-bottom with surface dispersion (with a thin water jacket)

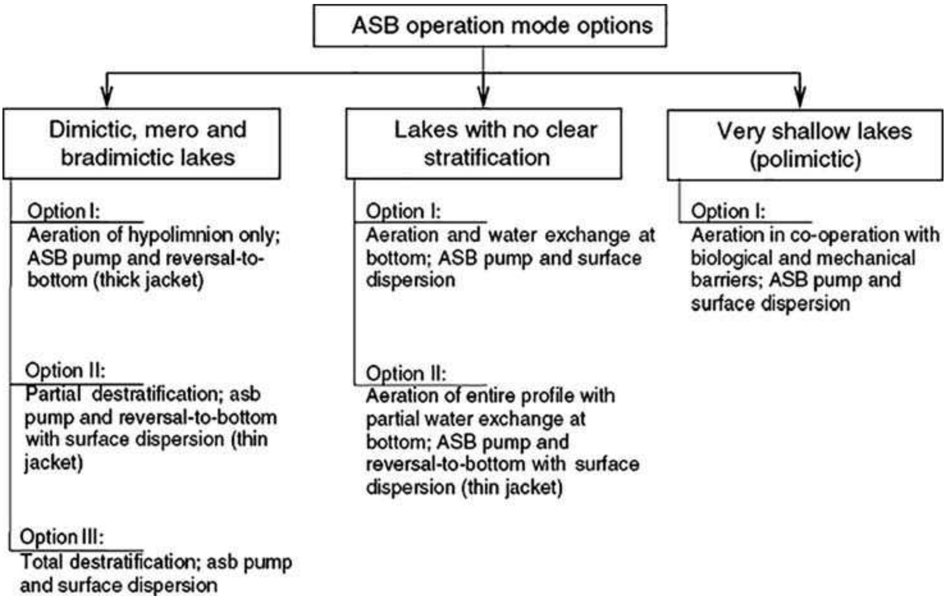


FIGURE 3 ASB operation mode options depending on lake type.

- Pump and reversal-to-bottom (with a thick water jacket)
- Pump with deep suction—in all the above types[48–50]

Overall change to oxygen and dynamic conditions at aerator installation site, i.e., genuine efficiency, depends on lake type and operation mode selected, i.e., aerator construction, as exemplified by the air-stream bottom aerator (ASB) (Figure 3).

During the method selection process, it is also necessary to answer the question of what immediate effect is expected from aeration since the procedure of aeration alone can bring positive results only in few cases.

An equally popular method of lake recultivation is dredging. During sediment removal, it is necessary to construct single or multiple thick-structure islands. Scarps, or island shores, should be equipped with bottom baffle piers that will constitute initial base for the benthos. At the early stage of island construction, the circumference wall should not be closed completely. Through the gap, waters overlying the sediment will be discharged. The gap ought to be secured with sedimentation barriers as well as a set of directional ASBs operating in the countercurrent to the outlet from the closed area. The applied cofferdam will only be temporary. At the end of island filling, the outlet gap ought to be closed with a cofferdam. Upon completion of sediment dredging, the whole ought to be covered with stable material.[48–50]

Upon completion of engineering work, red osier (for instance) needs to be planted. This type of islands can be used for recreation as well as natural purposes. Appropriate, separated parts of the islands can become nesting areas for many bird species. Sediments remaining at the bottom, whose thickness should not exceed 20 cm, may be immobilized by means of a baffle pier grid.

In the course of recultivation procedures, supply points ought to be perpetually monitored. If biogenes are found to be introduced via these routes, it is necessary to place biological barriers, possibly active (with aeration) at inflow inlets.[51,52]

The negative influence of surface wash can also be limited by barriers of appropriate plantings parallel to the shoreline. When forming the islands, it would be advisable to dredge a couple of local deepest waters (ca. 5–6 m). At those sites, an aerator is placed to disable the occurrence of oxygen deficits at windless periods. Artificial sparging enhancement ought to be sustained until balance is obtained between food and consumers, as well as between producers and reducers. Hence, in order to reduce power costs, in near-shore island zones and non-developed shore segments, installation of artificial reef as initial base for benthos is recommended. The ichtiopopulation status and composition also need to be considered.[48–50]

Another method aimed at restoration of good water quality is chemical bonding and inactivation of nitrogen and phosphorus. This system, used in Scandinavia, has been adapted to Polish conditions and tested at several sites, e.g., Lake Dlugie in Olsztyn. The phosphorus inactivation method is recommended for those lakes in which, despite cutoff of external sources of biogenic compounds, high fertility is sustained through deposition of these compounds from bottom sediments. Used primarily in shallow lakes, this method relies on removal of excess phosphorus from the pelagic zone and trapping it in bottom sediments. It also increases sorption capacity of degraded lake bottom sediments. Strongly eutrophicated lakes are characterized by high biogenic compound content, which causes blooms (increased phytoplankton population). At this degree of degradation, the lake no longer fulfills its recreational or economic functions. Chemical deposition of phosphorus by means of aluminum (PAX) and ferric (PIX) coagulants reduces the quantity of biogenic (nutrient) compounds, thus limiting algal growth intensity, which results in improvement of water quality and transparency.[53,54]

In the case of shallow reservoir recultivation, it is necessary to mitigate the deoxygenated layer of benthic water, the so-called oxicline. The oxicline generally occurs in the summer, appearing as a result of oxygen depletion in the course of chemical reactions. These reactions occur at the contact zone between bottom sediment and water, depending, among others, on ambient temperature: the warmer the water, the faster the reactions proceed. Under the circumstances, more and more frequently, actions are taken to place aerators equipped with a phosphorus inactivation system in reservoirs. This method has also been used in the Kielce Lake, where a pulverizing aerator was installed in May 2008. It is powered with wind energy and has a built- in installation dosage system for iron sulfate, which bonds phosphorus.[28]

Most efforts to alleviate the detrimental and undesirable effects of eutrophication on aquatic systems address the problem of P reduction in the inflows. Despite many cases of success in the reduction of nutrient loads of lakes in recent years, the expected corresponding reduction of phytoplankton abundances has often been delayed by many years. Sometimes, lakes have been reported to be resistant in their response to loading reductions. Recycling of P from the sediments becomes more important if the P in inflows is reduced.[55]

Phosphate inactivation in bottom sediments enables the elimination of internal transport in the reservoir (release of biogenes accumulated in bottom sediments into the pelagic zone), which is the cause of self-maintenance of high trophy levels—a feature of overfertile lakes, conditioning the occurrence of blue-green algae blooms. This means that as a result of the application of this method, the cause is removed, rather than the effects of excess of biogenes in the water. This testifies to greater efficiency of the above technique.[56]

The only alternative for application of the phosphorus inactivation method in bottom sediments is lake dredging. This is a technique that is equally effective, but very expensive, long-lasting, and accompanied by the problem of storage and further treatment of the removed sediments. It must be stressed that the storage site must not be located in the lake catchment area so that effluent waters together with their biogene load cannot find their way back to the reservoir. Similarly, treatment of the sediments is connected with the need for a recipient, transportation, etc. The phosphorus inactivation method in bottom sediments is far less costly, is incomparably faster at application, and provides enduring results of reduced phosphorus concentration in the water.[56,57]

The oldest method of phosphate inactivation in sediments, with simultaneous mineralization of organic matter, is the Riplox method, which yields good results. The required labor intensity and technical complexity make it extremely unpopular. Other techniques are its modifications, the Prote method being a case in point. The similarity lies in the fact that in both cases, air and the flocculant are simultaneously supplied to bottom sediments. A major difference is prior induction of intense sediment resuspension and flocculant application to a very carefully delineated sediment layer. The responsibility for the success of lake recultivation with this method lies with an effective device, providing precision dosage of flocculants to lake bottom sediments.[56,58]

To that end, several purpose-made mixes, containing a variety of substances as well as enzyme complexes, selected bacteria, biological activators, complete with extended-surface mineral carriers and stabilizers, have been developed. A preparation introduced into a water reservoir strongly activates the detritus food chain and increases the participation of autotrophic bacteria in matter circulation.[56,58]

The effects are so enduring that the following results are observed over 2 years:[56,58]

- Acceleration of organic matter decomposition in the sediments: quantitative reduction of accumulated bottom sediments and deceleration of natural reservoir shallowing
- Elimination of anaerobic zones in the hypolimnion and acceleration of sediment surface oxidation
- Enhanced elimination of biogenic compounds from matter circulation
- Blue-green algae bloom intensity control
- Boosting the growth of submerged water vegetation.
- Increased fish biomass[56,58]

Achieving enduring improvement of reservoir condition takes time and efficient implementation of the stages aimed at reduction of external supply, internal supply (elimination of anaerobic zones), and the quantity of biogenes in circulation (trophy reduction).

The experience in terms of external supply reduction has shown that the main condition that determines the efficiency of this technique is possibly the fastest reduction of external biogene supply. If this requires improvement of wastewater management in the reservoir catchment, the issue may seem problematic and distant in time due to the need for a wastewater treatment plant to be modernized or constructed. In contrast, the application of the Trigger-1 and Trigger-2 biopreparations in already-existing household, municipal, or industrial treatment plants enables limitation of pollutant inflow into the lake. Because the elimination of 1 kg of phosphorus and nitrogen in the process of wastewater treatment, even at plants with inadequate technical equipment, is far easier and cheaper than in the water environment, it is not advisable to wait for the construction to be completed, instead using all the methods to enable the reduction of external supply.[59]

In reservoirs where storage of large internal loads at stagnation periods has already occurred, the occurrence of anaerobic zones above the bottom is observed in the water. This triggers the mechanism of internal supply, i.e., releasing the previously accumulated biogenes from bottom sediments. This leads to accelerated degradation of the reservoir, although initially, only increased reservoir productivity is noted (primary and secondary productivity). Natural processes that occur in the reservoir are accelerated several times and result in rapid expansion of anaerobic zones. The methods applied to check this phenomenon, such as aeration, are expensive and carry a high risk of failure. A less costly method that does not require complex technical procedures is the application of a biopreparation, such as Trig- ger-3max, which, without changing water stratification, eliminates anaerobic zones in the hypolimnion, reducing the internal supply and eliminating anaerobic zones.

Recently, many biopreparations containing saprophytic microorganisms are commercially available and commonly offered for use in ponds, lakes, and reservoirs. These preparations usually consist of selected bacterial strains immobilized on a mineral carrier. Sometimes, the preparations are also enriched by bacterial enzymes such as biocatalyzators. Some preparations are even enriched by nutrients as growth stimulants at the beginning of growth after addition into the water. However, in most cases, parallel aeration or sediment oxidation by adding another electron acceptor such as nitrate will be necessary to support the growth and activity of added microbes. Otherwise, the method hardly brings any effects. With regard to many bacterial extracellular exudates that have been reported to inhibit the growth of cyanobacteria, a direct effect on cyanobacterial development may be possible as well.

However, no scientific proof of such an effect in the case of application of these commercially available biopreparation is available.[56,60]

All biogenes that occur in a water reservoir ecosystem are in circulation, the major stages of which are transformations of the dissolved phase into the solid phase and back into the dissolved phase. The main mechanisms of this change are the initial placement of elements in the biomass (solid and molecular phase), and their subsequent release from mortified organisms and fecal matter (dissolved phase). Phosphorus is the quickest to be released, and consequently, it circulates in the epilimnion, passing many times from the dissolved phase to the solid before falling onto the sediments. The application of a biopreparation will thus enable a reduction of the quantity of biogenes in circulation, thus limiting lake trophy.[48]

More and more frequently, attempts are made to develop and apply sustainable methods of water reservoir qualitative enhancement, based on "bio" structures. A case in point is the bio-hydro structure (biofilter filling), as well as active biological and mechanical filters, which may be used in open water reservoirs, thanks to which the processes of mineralization and biomass growth are induced.[48,61]

Components of biofilter filling are manufactured as standard panels. They are five-layer grids, made of wide strips, with single net-mesh dimensions. Each subsequent layer is shifted in relation to the previous layer in two directions, by half a net-mesh module.[48]

When this packet is placed vertically, the vertical strip plane is perpendicular to the panel plane. The horizontal strip plane is differently inclined to the panel plane. This geometric arrangement enables directing the stream upward or downward, depending on flow direction. Consequently, this structure may be used as, for instance, a sedimentation element, a lift component, or a resuspension barrier and initial base for the complex of plant (phytobenthos) and animal (zoobenthos) organisms, which occupy the water reservoir bottom.[61]

Strip surface is rough (knurled) and made of chemically neutral plastic, which enables occupancy by a variety of sedentary species. Despite the fact that the extended surface of such filters amounts to 122 $m^2 \cdot m^{-3}$, the filter displays a low degree of flow suppression and relatively large vacant internal space. Due to this, upon placement in the pelagic zone, it very soon becomes the habitat of a very wide range of periphyton. By being an ecological niche in its own right, it also acts as an attractant to many sedentary species as well as fish fry. Above the pelagic zone, autonomously, on the top elements of the panels, a wide variety of hydrophilous plants settle.[48]

In the 1980s, at Biotechnika Company, the occupancy extent of these filters was studied in Lakes Karczemne and Klasztorne. On average, one standard panel came to be occupied by biomass in the amount of ca. 70 kg. If such an occupied filter is removed from the reservoir at the right moment, then, together with the biomass, the built-in biogenic compound load will also be removed. The certainty of the effect achieved, in contrast to chemical methods for example, consists in the fact that the application of chemical deposition does not guarantee that, under certain conditions (pH, redox potential), the sorption capacity of metals in the coagulants applied will not be destroyed. As data demonstrate, this is a very frequent occurrence and then, apart from the still unsolved problem of phosphorus removal, the additional "foreign" load in the form of metal compounds is found in the water. Then, the phenomenon of system adjustment to new conditions, which can moreover take a completely unknown direction, arises.[48]

Sparging, in connection with bio-hydro barriers, also offers several options, including passive biological barriers as well as active biological and mechanical filters. If a lake receives a strongly polluted watercourse, the inlet of the watercourse may be separated from the remaining area with a barrier (consisting of several layers to make it most advantageous) made of bio-hydro structures placed jointly, streamwise—first downwards and then upwards. Mechanically, this setup will function as a sedimentation barrier, which will, due to autonomous occupancy, after a short period of time, also and perhaps primarily become a biological barrier.

On account of its structure, despite initial passivity, as occupancy progresses, it becomes an attractant for an increasing variety of species, in most cases beneficial to further development of reservoir biocenosis and further recultivation procedures. During the observation of this type of barrier placed in Lake Rybnickie, it was noted that it had become a "refuge," initiating changes in both animal and plant populations. A double-positive effect was thus achieved: the inflowing pollutant load that degraded the lake has been set to work towards beneficial renewal of the reservoir.

In Lake Rybnickie, the positive result of barrier operation was visible as soon as after 1 year of working. Positive results of water quality tests restored the recreational value of the reservoir, and after many years, the State Sanitary and Epidemiological Board ultimately permitted bathing.[48]

Conclusions

Fertility or abundance of a water reservoir with biogenic mineral substances gradually increases in time, and may be either a natural process or one resulting from human activity. Natural eutrophication proceeds slowly, whereas anthropogenic eutrophication is characterized by a rapid change rate. At the initial stage, an increase of primary and secondary productivity is observed, but as eutrophication progresses, the efficiency of secondary productivity decreases and algal biomass gradually accumulates. The reservoir basin gradually fills with sediments and the water volume decreases, which, even at no biogenic inflow, simultaneously accelerates eutrophication. Further succession leads to the reservoir being turned into land.

In order to preserve a lake, human intervention becomes necessary. All recultivation strategies require sensible planning, time, and, frequently, considerable financial outlay. Their results are not always certain, but the current know-how concerning the subject enables appropriate selection and implementation of an effective recultivation process in a particular lake. The type of the method and technique used depends on the size of the reservoir, the nature of fauna and flora resident therein, the character of the catchment, and the type of adjacent land development. Effectiveness and efficiency of the work performed depend on accurate assessment of the reservoir status and adoption of an appropriate action plan.

While analyzing the efficiency of various procedures carried out in the world's lakes, it must be stated that before embarking on recultivation, it is necessary to

- Define the current trophic status of the lake as well as causes and sources of its degradation basing on physicochemical and biological tests of lake waters, inflows and outflows, complete with an analysis of sediment chemical composition, considering exchange processes between sediments and water.
- Specify the feasibility of the water reservoir recultivation in view of its natural, recreational, and economic value.
- Design and perform protective procedures consisting of the elimination, or at least limitation, of individual pollution sources.
- Develop the most advantageous concept of a recultivation method that ought to make allowances for the trophic status as well as morphometric and hydrological conditions of the reservoir, together with possible uses of margin land, its development, ownership status, access to the shore and a power supply, duration of the process, and financial outlay required.
- Ensure that recultivation will be performed by an expert team under the supervision of an experienced limnologist at all times.
- Document the course of recultivation and its effects through the results of monitoring carried out at particular stages.
- Deem recultivation unfeasible in view of its failing to offer an opportunity for water quality enhancement, if it is found that excess loading of the lake is impossible to eliminate.
- Have a close connection with appropriate regulations, enabling the protection of a lake subject to recultivation procedures from further degradation processes.

Acknowledgments

This study has been financed in part by the State Committee for Scientific Research (KBN) (Grant No. N N305 306635).

References

1. USEPA. *The Lake and Reservoir Restoration and Guidance Manual,* 2nd Ed. (OWRS, EPA 440/4–90–006); United States Environmental Protection Agency, Office of Water: Washington, DC, 1990.
2. Bajkiewicz-Grabowska, E.; Mikulski Z. *Hydrology*; PWN: Warszawa, 2006.
3. Leonard, J.; Crouzet, P. *Lakes and Reservoirs in the EEA Area*; European Environment Agency: Copenhagen, 1998.
4. Viekman, B.E.; Flood, R.D.; Wimbush, M.; Faghri, M.; Asako, Y.; Van Leer, J.C. Sedimentary furrows and organized flow structure: A study in Lake Superior. Limnol. Oceanogr. **1992**, *37* (4), 797–812.
5. French, R.H.; McCutcheon, S.C.; Martin, J.L. Environmental hydraulics (Ch. 5). In *Hydraulic Design Handbook*; Mays, L., Ed.; McGraw-Hill Professional: New York, 1999;–5.33.
6. Wetzel, R.G. *Limnology*, 2nd Ed.; Saunders College Publishing: Fort Worth, 1983.
7. Martin, K.L.; McCuthceon, S.C. *Hydrodynamics and Transport for Water Quality Modeling*; CRC Press: Boca Raton, FL, 1999.
8. Grabowska, E.; Zdanowski, B. Phosphorus retention in lake section of Struga Siedmiu Jezior. Limnol. Rev. **2006**, *6,* 5–12.
9. Tipton, M.J.; Brooks, C.J. The dangers of sudden immersion in cold water (Ch. 3). In *Survival at Sea for Mariners, Aviators and Search and Rescue Personnel (RTO-AG- HFM-152);* RTO/NATO, NATO Research and Technology Organization: Neuilly-sur-Seine Cedex, France, 2008,–3.10.
10. Lenntech, Water Treatment Solutions. Water ecology FAQ frequently asked questions, http://www.lenntech.com/wa-ter-ecology-faq.htm (accessed May 2011).
11. Vollenweider, R. Advances in defining critical levels for phosphorus in lake eutrophication. Mem. Ist. Ital. Idrobiol. **1976**, *33,* 53–83.
12. Cooke, G.D. *Restoration and Management of Lakes and Reservoirs*, 3rd Ed.; CRC Press, Taylor and Francis Group: Boca Raton, 2005.
13. Klapper, H. Technologies for lake restoration. Papers from Bolsena Conference (2002). Residence time in lakes: Science, management, education. J. Limnol. **2003**, *62* (suppl. 1), 73–90.
14. Sengupta, M.; Dalwani, R., Eds. Hypolimnic Withdrawal for Lake Conservation, Proceedings of Taal2007: The 12th World Lake Conference, Jaipur, India, 2008; 812–818.
15. Gupta, A.K.; Shrivastva, N.G.; Shrama, A. Pollution Technologies for Conservation of Lakes; Proceedings of Taal2007: The 12th World Lake Conference, Jaipur, India; Sengupta, M.; Dalwani, R., Eds., 2008; 894–905.
16. Lake Dredging. Illinois EPA, Northeastern Illinois Planning Commission, Chicago, Illinois, 1998.
17. Helmke, D. Dredging: Kansas Takes First Steps to Reclaim Reservoirs, The Kansas Lifeline, Seneca, Kansas, 2010, 96–99.
18. Maxwell, A.O. Upstate's Saluda Lake Revived through Grassroots Conservation Effort. USDA-NRCS, http:// www.sc.nrcs.usda.gov/news/saluda_lake.html (accessed May 2011).
19. Murphy, T.P.; Macdonald, R.H.; Lawrence, G.A.; Mawhinney, M. Chain lake restoration by dredging and hypolimnetic withdrawal. In *Aquatic Restoration in Canada*; Backhuys Publishers: Leiden, the Netherlands, 1999; 195–211.
20. Kelly, A. Appendix 5. History of lake restoration. In Lake Restoration Strategy for The Broads, Broads Authority, The Broads—A Member of the National Part Family, Norwich Norfolk, 2008.
21. Hosper, H.; Meyer, A.-L. Control of phosphorus loading and flushing as restoration methods for Lake Veluwe, the Netherlands. Aquat. Ecol. **1986**, *20* (1–2), 183–194.
22. Mankiewicz, P.S.; Mankiewicz, J.A. *Ecological Engineering and Restoration Study. Flushing Meadows Lakes and Watershed*. The Gaia Institute: City Island Avenue, Bronx, NY, 2002.

23. Cross, T.K.; McInerny, M.C.; Davis, R.A. Macrophyte Removal to Enhance Bluegill, Largemouth Bass and Northern Pike Populations. Minnesota Department of Natural Resources, St. Paul, Investigation Report 415, 1992.
24. Rabasco, R. Trophic effects of macrophyte removal on fish populations in a boreal lake. A thesis submitted to the Faculty of Graduate Studies of the degree of Master of Natural Resources Management, Natural Resources Institute, The University of Manitoba, Canada, 2000.
25. Dunalska, J. Influence of limited water flow in a pipeline on the nutrients budget in a lake restored by hypolimnetic withdrawal method. Pol. J. Environ. Stud. **2002**, *11* (6), 631–637.
26. Anderson, M.A.; Paez, C.; Men, S. Sediment nutrient flux and oxygen demand study for Canyon Lake with assessment of in-lake alternatives. Final report: Canyon Lake Nutrient Flux and In-Lake Alternatives, Dept. of Environmental Sciences, Univ. of California, 2007.
27. Riccardi, N.; Mangoni, M. Chemical consequences of oxygenation in a shallow eutrophic lake studied with meso- cosms. J. Aquat. Ecosyst. Health **1996**, *5* (1), 63–71.
28. Rabajczyk, A.; Jóźwiak, M. Littoral zone flora versus quality of Kielecki Bay waters. Ecol. Chem. Eng. A, **2008**, *15* (12), 1359–1368.
29. Fast, A.W. Lake Aeration System for Canyon Lake, California, Lake Elsinore and San Jacinto Watersheds Authority (LESJWA), Riverside, CA, 2002.
30. Kothandaraman, V.; Roseboom, D.; Evans, R.L. Pilot Lake Restoration Investigations—Aeration and Destratification in Lake Catherine. Illinois State Water Survey, Urbana, Illinois, 1979.
31. Lossow, K.; Gawrońska, H.; Jaszczult, R. Attempts to use wind energy for artificial destratification of Lake Starod- worskie. Pol. J. Environ. Stud. **1998**, *7* (4), 221–227.
32. Itasca Soil and Water Conservation District, Jessie Lake Nutrient TMDL for Jessie Lake. Wenck Associates, Inc., Minnesota, 2011.
33. Łopata, M.; Gawrońska, H. Effectiveness of the polymictic Lake Głęboczek in Tuchola restoration by the phosphorus inactivation method. Pol. J. Nat. Sci. **2006**, *21* (2), 859–870.
34. Reitzel, K.; Hansen, J.; Andersen, F.Ø.; Hansen, K.S.; Jensen, H.S. Lake restoration by dosing aluminum relative to mobile phosphorus in the sediment. Environ. Sci. Technol. **2005**, *39* (22), 4234–4140.
35. Palermo, M.R.; Maynord, S.; Miller, J.; Reible, D.D. Guidance for In-Situ Subaqueous Capping of Contaminated Sediments, Assessment and Remediation of Contaminated Sediments (ARCS) Program, Great Lakes National Program Office, US EPA 905-B96-004, 1998.
36. Draft Onondaga Lake Capping and Dredge Area and Depth Initial Design Submittal. Parsons, Anchor QEA, 2009; P:\ Honeywell-SYR\444576 2008 Capping\09 Reports\9.3 December 2009_ Capping and Dredge Area and Depth IDS\ Appendices\Appendix A -RA Delineation\Attachment A- 1.doc (accessed May 2011).
37. Bona, F.; Cecconi, G.; Affiotti, A. An integrated approach to assess the benthic quality after sediment capping in Venice Lagoon, Aquat. Ecosyst. Health Manage **2000**, *3*, 379–386.
38. Wagner, K.J. The Practical Guide to Lake Management in Massachusetts. Executive Office of Environmental Affairs, Commonwealth of Massachusetts, Boston, 160 pp, 2004.
39. Madsen, J.D.; Getsinger, K.D.; Owens, Ch.S. Whole Lake Fluridone Treatments For Selective Control of Eurasian Wa- termilfoil: II. Impacts on Submersed Plant Communities, Lake and Reservoir Management, **2002**, *18* (3), 191–200.
40. Mitchell, J. STS CONSULTANTS, LTD. Joliet Junior College Lake. Diagnostic/Feasibility Study and Restoration Plan. Phase 1 B Report, 2007.
41. Van de Haterd, R.J.W.; Ter Heerdt, G.N.J. Potential for the development of submerged macrophytes in eutrophicated shallow peaty after restoration measures. Hydrobiologia **2007**, *584,* 277–290.
42. Bontes, B.M.; Per, R.; Ibelings, B.W.; Boschker, H.T.S.; Middelburg, J.J.; Van Donk, E. The effects of biomanipulation on the biogeochemistry, carbon isotopic composition and pelagic food web relations of a shallow lake. Biogeosciences **2006**, *3*, 69–83.

43. Beklioglu, M.; Ince, O.; Tuzun, I. Restoration of the eutrophic Lake Eymir, Turkey, by biomanipulation after a major external nutrient control I. Hydrobiologia **2003**, *489,* 93–105.
44. Frederiksborg County, Final report: Restoration of Lake Fure—A nutrient-rich lake near Copenhagen. LIFE02NAT/ DK/8589, Hillerod, 2006.
45. Hamrin, S.F. Planning and execution of the fish reduction in Lake Ringsjön. Hydrobiologia, **1999**, *404,* 59–63.
46. Terasmaa, J. Seston Fluxes and Sedimentation Dynamics in Small Estonian Lakes. Tallinn University, Dissertations on Natural Sciences 11, Tallinn, 2005.
47. Wilson, D.E.; Beutel, M. Final Report: Review of the Feasibility of Oxygen Addition or Accelerated Upwelling in Hood Canal. Washington, Brown and Caldwell, 2005.
48. Sadecka, Z.; Waś, J. Non-invasive methods of reservoir restoration—Perspective. In *Wastewater and Sludge Treatment Utilization;* Sadecka, Z.; Myszograj, S., Eds.; University of Zielona Góra Press: Poland, 2008; 247–260.
49. Company WBWW-BIOPAX Sp. z o.o, http://wbww- biopax.pl/ (accessed May 2011).
50. Company Biopax.pl, http://www.biopax.pl/cm.php?id=16 (accessed May 2011).
51. Allied Biological Inc. New York, available at http://www.alliedbiological.com/lake_managementandrestoration_ specialtyservices.htm (accessed May 2011).
52. Environmental Consulting and Technology, Inc., Carpenter Lake Restoration Project, Final Report, Corporate Of- fices—Gainesville, FL, 2008.
53. Brzozowska, R.; Gawrońska, H. Effect of the applied restoration techniques on the content of organic matter in the sediment of Lake Dlugie. Limnol. Rev. **2006**, *6,* 39–46.
54. Brzozowska, R.; Gawrońska, H. The influence of a longterm artificial aeration on the nitrogen compounds exchange between bottom sediments and water in Lake Długie. Oceanol. Hydrobiol. Stud., Int. J. Oceanogr. Hydrobiol. **2009**, 37 (1), 113–119.
55. Van Donk, E.; Hessen, D.O.; Verschoor, A.M.; Gulati, R.D. Re-oligotrophication by phosphorus reduction and effects on seston quality in lakes. Limnology **2008**, *38,* 189–202.
56. Drábková, M. Methods for control of the cyanobacterial blooms development in lakes. Dissertation thesis, RECE- TOX—Research Centre for Environmental Chemistry and Ecotoxicology, Brno, Czech Republic, 2007.
57. Salgot, M. Eutrophication in Lake Maryut: Diagnosis and actions proposal. ANNEX 1, Alexandria Lake Maryiut Integrated Management (ALAMIM), Alexandria, Egypt, 2009.
58. Quaak, M.; Does, J.; Boers, P.; Vlugt, J. A new technique to reduce internal phosphorus loading by in-lake phosphate fixation in shallow lakes. Hydrobiologia **1993**, *253* (1–3), 337–344.
59. VWP Individual Permit No. 08–0572, Part I—Special Conditions, DEQ, 2008, http://stauntonriverwatch.com/VWPP%20Final%20Part%20I%20special%20conditions.pdf (accessed May 2011).
60. Duval, R.J.; Anderson, L.J.W. Laboratory and greenhouse studies of microbial products used to biologically control algae. J. Aquat. Plant Manage. **2001**, *39,* 95–98.
61. Gulati, R.D.; Pires, L.M.D.; Van Donk, E. Lake restoration studies: Failures, bottlenecks and prospects of new ecotechnological measures. Limnology **2008**, *38*, 233–247.

46

Wastewater Use in Agriculture: Policy Issues

Dennis Wichelns

Introduction

Treated and untreated wastewaters have been used for many years to irrigate farms, landscapes, and golf courses in both developed and developing countries. Public officials in arid countries have promoted wastewater use primarily to extend their limited national water supplies. Officials in other countries have promoted the benefits of using treated wastewater in agriculture or landscape maintenance as an alternative to discharging the wastewater into a stream or ocean. Their rationale is based partly on cost and partly on environmental implications. Treating wastewater is expensive, and a portion of the cost can be recovered by using the wastewater in a productive endeavor. From an environmental perspective, applying the treated wastewater to a farm field provides another layer of quality improvement before the water is finally returned to a stream or ocean.

Many farmers in developing countries use untreated wastewater for irrigation, often because it is the only source of water available. Many small-scale farmers obtain irrigation water from streams or ditches that are polluted with effluent from a nearby city, industry, or housing development. The farmers likely would prefer higher-quality water, but in most cases, they have no alternative source for irrigation. There can be agronomic value in the nutrients in the untreated wastewater (the nitrogen, phosphorus, and potassium), but there are also pathogens and chemicals that threaten the health of farmers, food vendors, and consumers. Irrigating with untreated wastewater is risky business in developing countries, yet it generates household income for families with limited livelihood alternatives. Most farmers using untreated wastewater for irrigation likely would vote to continue using the wastewater, even if they understood all the risks, in the absence of an alternative, higher-quality water supply.

In a sense, farmers using untreated wastewater provide a public service by removing effluent from polluted streams and applying it to soils, thus reducing the pollutant load in downstream locations. Soils filter some of the undesirable constituents, and plants consume some of the nutrients. However, wastewater irrigation also generates risk for farm communities and consumers of farm products. Polluted canals and ditches and wastewater-irrigated fields create hazards in which children and other residents are exposed to harmful pathogens and chemicals. Consumers of farm produce also are at risk of illness when they handle and ingest contaminated vegetables, particularly when the food is eaten raw or prepared with inadequate care toward reducing contamination risk.

The goal of this entry in the encyclopedia is to describe policy issues pertaining to wastewater use in agriculture. The issues in developed countries are somewhat straightforward and mature. Public agencies have largely determined appropriate water quality criteria and implemented treatment protocols to support wastewater use in irrigation. Future issues will include refining those standards and protocols and continually evaluating the costs and benefits of alternative levels of wastewater treatment and use in agriculture and other activities. There will also be discussions of who should pay for wastewater treatment and who should have priority in receiving limited supplies. These issues involve costs, returns, and the division of economic rents, but they generally do not involve decisions that can support or destroy livelihood opportunities, whether intentionally or as the unintended consequences of seemingly beneficial policy choices.

Policy issues are more challenging in developing countries, where most of the wastewater used for irrigation is untreated and much of the use is informal and unintentional. Farmers, communities, and consumers are at risk from harmful constituents in the untreated wastewater, yet each group also obtains critically important benefits. Farmers generate financial returns that enhance their livelihoods and improve the economic status of farm communities. Consumers gain nutritional value by having affordable access to locally grown fresh vegetables. The public, more generally, benefits also when farmers divert effluent for use in irrigation, rather than allowing it to continue flowing downstream.

There are no easy—and affordable—policy choices regarding the use of untreated wastewater in developing countries. Public funding for treating all wastewater will not be available in many countries within the foreseeable future. Lacking the treatment alternative, public agencies must identify measures that will reduce the risks of using untreated wastewater, while still maintaining a notable portion of the benefits that accrue to farmers, consumers, and the larger community. In this review of policy issues pertaining to wastewater use in irrigation, we focus largely on issues involving the informal use of untreated wastewater in developing countries. The issues in such countries are more challenging, and the potential rewards of identifying and implementing successful risk reduction measures are substantial.

Wastewater Is a Resource in Waterscarce Settings

With increasing demands on water resources in arid and humid regions, the role of wastewater as a resource in municipal, industrial, and agricultural applications will continue to gain importance in many countries. Large numbers of farmers irrigate with wastewater in developed and developing countries, yet we do not have good statistics on the areal extent or the value of output generated through wastewater irrigation. This is due partly to the difficulty of identifying and measuring farm areas irrigated fully or partially with treated or untreated wastewater. In most countries, wastewater is used to irrigate vegetables, fruits, cereals, cotton, and fodder crops (Table 1). Fruits and vegetables are often consumed without cooking and thus pose a higher risk of illness due to pathogens or chemicals in wastewater.

TABLE 1 Examples of Wastewater Reuse in Selected Countries

Region and Country	Examples of Use
Africa	
Tunisia	Citrus, fodder
Morocco	Vegetables, fodder
Americas	
Argentina	Vegetables, fodder
Chile	Vegetables, grapes
Mexico	Vegetables, cotton, fodder, parks, and greenbelts
Peru	Vegetables, cotton, fodder
United States	Vegetables, cereals, fodder, greenbelts, golf courses
Middle East	
Israel	Cotton
Kuwait	Vegetables, orchards, fodder
Saudi Arabia	Cereals, fodder, greenbelts
Asia	
India	Vegetables, cereals

Source: Gupta and Gangopadhyay.[1]

Wastewater Use in Developed Countries

Farmers and landscape managers have been using treated wastewater for many years in Australia, Israel, Spain, the United Arab Emirates, the United States, and other developed countries.[2–5] Treated wastewater is used directly in some applications and indirectly in others, often following storage in an aquifer or discharge to a drain or stream from which farmers obtain irrigation water at some distance from a water treatment facility.[6,7] Over the years, scientists and practitioners have improved their understanding of the impacts of using treated wastewater for irrigation, and public officials have implemented policies and regulations consistent with current knowledge and experience.[8–11]

Persistent water scarcity in the Middle East and North Africa has led many governments to advance the state of knowledge regarding irrigation with wastewater and desalinized water. Wastewater is a key irrigation resource in Israel, Jordan, Kuwait, Iran, Oman, and Saudi Arabia.[12,13] In Israel, where 73% of treated municipal sewage is recycled[10] the government has supported the construction of pipelines to deliver wastewater to agricultural areas where many farmers use drip irrigation systems to minimize potential harm from contact with the wastewater.[14] Treated effluent also is stored in aquifers before recovery and delivery to farmers.[10] In Jordan, 95% of the treated wastewater volume is used each year, primarily for irrigation in the Jordan Valley.[15] In Saudi Arabia and the United Arab Emirates, many farmers and municipal landscape agencies irrigate with secondary and tertiary treated wastewater.[16–18] Tunisia, Egypt, and Morocco also have notable programs in which farmers are provided with treated wastewater for use in irrigation.[13]

Treated wastewater is used extensively for irrigation also in Australia and in the arid southwest and the humid southeast of the United States. In Australia, as in Israel, treated municipal effluent is infiltrated into aquifers, and farmers later retrieve the water for irrigating horticultural crops.[19] Most of the irrigation with wastewater in the United States is found in California and Florida, where strict water quality standards are enforced to protect consumer health and to sustain positive market acceptance of crops produced in these heavily agricultural states.[13]

Wastewater Use in Developing Countries

Perhaps the greater challenge regarding wastewater use in agriculture in the future pertains to developing countries in which there is inadequate capacity to collect, treat, and distribute wastewater in ways that minimize potential harm to public health.[20] Millions of farmers in developing countries rely on untreated, partially treated, or haphazardly blended wastewater as a source of irrigation supply. They produce grains, vegetables, and fruits with the low-quality water, often for sale at attractive prices in urban markets.[21,22] The farmers and their family members who have contact with the untreated and partially treated wastewater are at risk of health damage due to biological and pathogenic constituents.[23] The consumers of leafy vegetables also are at risk if they do not take adequate precautions in preparing the vegetables prior to consumption.[24,25] In many cases, some consumers are unaware that vegetables have been produced using wastewater, and they have no experience of becoming ill after eating raw vegetables. Hence, they are not inclined to implement the necessary precautions.[26]

Worldwide, an estimated 800 million farmers are engaged in urban agriculture. One-fourth of these farmers produce crops for sale in local markets on small plots of land of which they might or might not have ownership, using low-quality irrigation water that often contains untreated wastewater.[27] Market demand for irrigated vegetables in tropical areas is strong, and those produced locally are particularly important in warm regions where modern transportation and storage facilities are largely unavailable. Farmers in urban and periurban areas can bring their highly perishable vegetables to market quickly, thus generating household income and providing an important source of vitamins for consumers.[24]

Many farmers in sub-Saharan Africa and south Asia rely on untreated wastewater for a portion of their irrigation supply. Some farmers utilize wastewater taken directly from tanker trucks or sewage lines, while others divert irrigation water from streams and ditches polluted with untreated municipal wastewater.[20,28] In many areas, farmers irrigate crops and sell them in local markets (Figure 1), with few precautions taken to minimize risks in production or consumption.[23]

More than 60% of the vegetables consumed in many West African cities are produced locally in urban and periurban areas.[29] In Kumasi, the second largest city in Ghana, urban farmers produce 90% of the

FIGURE 1 Selling wastewater-irrigated leafy vegetables in local markets generates revenue for small-scale farmers, while creating a risk of contamination for consumers.
Source: Sanjini de Silva, IWMI, Hyderabad.

lettuce and spring onions consumed in the city. With irrigation, urban farmers can harvest 9 to 11 crops per year on their small plots (0.1 hectare), thus generating annual incomes of $400 to $800 per year.[24] Farmers place substantial value on the availability and reliability of their irrigation supply, even though it contains wastewater from the city.

Farmers along the Musi River, downstream of Hyderabad, India, irrigate rice and fodder grass with untreated wastewater discharged into the largely dry riverbed from many locations in the city. Farmers rely on the wastewater as their only source of irrigation water.[30] While reporting negative health effects due to wastewater exposure, farmers also report that the availability of wastewater in the Musi River enhances their livelihoods.[31,32] Farmers irrigating with wastewater report greater occurrence of fever, headaches, and skin and stomach problems than farmers who do not use wastewater.[31] Although irrigation enhances incomes, substantial costs are imposed on households by illnesses that might be caused by exposure to wastewater. It is in such settings that effective policies are needed to enhance farm family health and to protect consumers from the potentially negative impacts of consuming wastewater-irrigated vegetables.

Farmers in periurban areas of Calcutta, India, irrigate about 5000 hectares of rice using untreated wastewater from the city's sewage collection system. Many farmers use wastewater for aquaculture as well, thus enabling them to generate a steady annual income and provide employment opportunities for local workers.[1] The sustainability of the system is uncertain, as much of the wastewater contains heavy metals that accumulate in soils and can be taken up by vegetables.

Untreated wastewater is used also by farmers in periurban areas of Addis Ababa, Ethiopia, where only 3% of the residents are served by a centralized sewer system.[33] The farmers appreciate the nutrient content of the wastewater in a region where the price of imported fertilizer has been increasing by as much as 20% per year. Controlled experiments on urine-irrigated plots in Ethiopia, and in trials on farm fields, have demonstrated the agronomic value of the nitrogen and phosphorus contained in untreated wastewater.[33]

Policies Are Needed to Ensure Wise Use

From a policy perspective, the use of wastewater in agriculture provides both opportunities and challenges that require public intervention. In one sense, wastewater is an effluent requiring treatment or disposal, subject to regulations that protect public health. In the absence of regulations, private generators of wastewater would have little incentive to reduce volume or to manage the flow of wastewater beyond their property line. Because wastewater generation is a negative externality in most settings, regulations and incentives are needed to minimize the potential harm from wastewater in the environment.

Wastewater management has public-good characteristics in that once it is provided, many members of society benefit. At the same time, it is difficult to exclude individuals from enjoying the benefits of a cleaner, healthier environment once the decision has been made to collect and treat all wastewater in a community. The non-rival nature of the benefits and the difficulty of exclusion provide the basis for managing wastewater treatment within the public sector.

The public-goods perspective is appropriate when viewing wastewater as an effluent requiring treatment or disposal. However, when viewing wastewater as a resource, there are notable private benefits for which individuals will be willing to invest time, effort, and funding to enhance their opportunities. The private-goods perspective pertains to both treated and untreated wastewater. Several water agencies in Australia, Israel, and the United States sell treated wastewater (directly or through an aquifer recharge program) to farmers and golf course owners who obtain private benefits through irrigation.[34,35] Often there is a price differential between treated wastewater and fresh water, thus providing a financial incentive for irrigators to select the treated wastewater.[36,37]

Farmers in developing countries also obtain private benefits, but the distribution of wastewater among them is much less formal, and the wastewater generally has not been treated. An estimated 80% of the sewage generated in developing countries is discharged untreated into the environment, and half the population is exposed to polluted water sources.[38,39] Many farmers acquire untreated wastewater when they divert irrigation water from a stream or ditch that carries effluent from a nearby city or from households in an urban, periurban, or rural area. Water diversions and the use of wastewater in such settings generate private benefits for the farmers. The public gains also as the farmers remove the low-quality water from streams and ditches. However, the primary motivation for farmers is to boost their productivity and increase their net returns in agriculture. By doing so, they risk the health of their families through exposure to untreated wastewater, and they create situations in which consumers also are at risk of eating harmful produce. Public policies are needed to reduce these risks and to optimize the management of wastewater from the public's perspective.

Policy Issues in Developed Countries

Current policy issues regarding wastewater use for irrigation are best described separately for developed and developing countries. In developed countries, the primary policy issues involve economics and finance. The public generally has already determined that protective water quality guidelines must be followed when using wastewater for irrigation. The technology for treating wastewater is well understood, and advances in technology that lower the costs and increase the benefits will be forthcoming in response to market demands and in accordance with government-sponsored research programs. The question will not be whether wastewater will be treated but who will pay for treatment and how much of the cost will be passed along to the users. In some areas, public officials also will seek to determine the economically optimal level of wastewater treatment, as a function of its intended uses.[40]

Key policy questions will also involve the appropriate levels of government involvement in wastewater treatment and reuse programs. Where the private, farm-level benefits of wastewater use are notable, farmers should be financially motivated to invest in production methods and develop market outlets that support irrigation with wastewater. Farmers should also be able and willing to pay for treated wastewater delivered by a public agency or water user association. Using contingent valuation methods in a survey of Greek farmers, Bakopoulou et al.[41] determined that 58% of the participants would be willing to pay half the fresh water price to purchase treated wastewater for irrigation. Small-scale farmers on the island of Crete expressed a greater willingness to pay for treated wastewater after attending a session in which they learned of the private and social benefits of wastewater irrigation.[42]

Farmers should also have incentive to invest in communal facilities for collecting, treating, and delivering wastewater to farms, as the price and availability of wastewater improve over time, relative to the price and availability of fresh water. Public agencies can hasten a farm-level switch from fresh water to wastewater with water pricing and investment incentives and also by informing farmers and consumers of the safety and benefits of irrigating with treated wastewater.

Policy Issues in Developing Countries

The primary policy questions in developing countries are more challenging, in part, because financial resources are limited and there are many competing demands on public funds. In addition, much of the wastewater irrigation takes place in decentralized, informal settings in which individual farmers gain access to wastewater very simply by diverting polluted water from a stream or ditch. Property rights to the water are not defined, and there is no communal agency or water user association that coordinates irrigation activities. Millions of individual farmers will be very reluctant to stop diverting polluted water for use in irrigation, given that their livelihoods currently depend on the sale of irrigated farm produce.

Public officials in developing countries must address the following question: How do we minimize the risks to farmers and consumers, while not destroying or severely diminishing the livelihoods of those farmers who currently irrigate with wastewater? This is not an easy question to answer. Public officials will be mindful of the benefits that farmers provide by diverting and using polluted water for irrigation. If not for that activity, larger volumes of wastewater would continue flowing downstream in

TABLE 2 Average Nutrient Availability in Human Excreta per Person, per Year

	Amount of Nutrient Available (kgs)			
	In Urine	In Feces		Amount of Nutrient Required to Produce
Nutrient	(500 liters)	(50 liters)	Sum	250 kg of Cereal (kgs)
Nitrogen	4.0	0.5	4.5	5.6
Phosphorus	0.4	0.2	0.6	0.7
Potassium	0.9	0.3	1.5	1.2

many watercourses, creating greater risk for downstream residents and causing environmental harm over a larger area. Farmers who irrigate with wastewater generate one set of risks for their families and consumers while reducing another set of risks to residents downstream.

Public officials also will note that some of the pollutants in municipal wastewater (such as nitrogen, phosphorous, and potassium) are plant nutrients. In areas where inorganic fertilizer is costly and difficult for small-scale farmers to obtain, those nutrients have substantial agronomic value. The nutrients contained in the average amount of urine and feces generated by one person each year are sufficient to produce enough grain to sustain that person. Urine is a particularly good source of nitrogen, phosphorus, and potassium (Table 2). Efforts to collect and transport urine directly to farms, rather than discharging it to waterways, would allow farmers to manage their nutrient applications with greater care and accuracy.

In summary, farmers generate both private and public benefits when they divert polluted water from streams and ditches to irrigate crops in urban areas. The nutrient content of wastewater is sufficient to increase crop yields, provided that farmers do not apply excessive amounts of some nutrients, such as nitrogen.[1,43] Public officials in developing countries must determine how to sustain these beneficial aspects of wastewater irrigation and the livelihoods of farm families, while minimizing risks to those same families and the consumers of their produce.

Policy Options Include Treatment and Non-Treatment Alternatives

The policy options available to public officials for reducing the risks associated with wastewater irrigation in developing countries, while sustaining livelihood benefits, might be placed into four categories:

1. Improve and extend centralized wastewater treatment
2. Improve and extend decentralized wastewater treatment
3. Regulate (with enforcement) the use of untreated wastewater in agriculture
4. Complement existing wastewater use patterns with risk reduction interventions to protect farm families, communities, and consumers

The first category is likely the most costly and the least likely to be implemented along a reasonable timeline. There might be affordable opportunities in some settings within developing countries, in which new, large-scale wastewater treatment plants can be constructed to improve the quality of water available for agriculture. Yet it seems that if such opportunities were affordable, if they compared favorably with alternative public investments, and if an affordable source of finance were available, then such efforts would already be underway. It is difficult to imagine that the pace of investments in large, centralized wastewater treatment plants will be sufficient to improve water quality for many of the farmers who currently use wastewater for irrigation in developing countries.

The second category includes options that should be more affordable than building large, centralized wastewater treatment plants. The goal within this category is to identify opportunities for enhancing irrigation water quality at an appropriate scale and within a meaningful distance from the point of wastewater use. Small-scale wastewater treatment plants might be designed with the expressed purpose of making higher-quality water available for irrigation. The construction costs and operating criteria for such plants might be different—and less expensive—than those pertaining to centralized wastewater treatment plants that discharge water intended for uses outside agriculture.[44] For example, it is important to remove solids, salts, and pathogens from water intended for use in irrigation, but farmers can accommodate higher nutrient levels than wastewater users in municipal and industrial settings.

The third option likely will be challenging in many developing-country settings, given the decentralized, informal nature of wastewater use and the strong dependency of farm households on wastewater. Regulations will be politically unpopular, and enforcement will be difficult to achieve. In Syria, for example, the government disallows the irrigation of vegetables with wastewater, but compliance with the restriction is not complete. Syrian officials resort to destroying vegetable crops irrigated with wastewater when they find such situations. As a result, less than 7% of the area irrigated with wastewater near the city of Aleppo is in vegetable production.[27] The opportunity costs involved in planting and cultivating crops, only to have them destroyed by the government, can be substantial for farm households with limited sources of income.

The financial burden of treating wastewater in developing countries and the challenge of regulating wastewater use by farmers will remain substantial for the foreseeable future. Hence, many farmers will continue using wastewater, and their workers and families will remain at risk of infection while applying irrigation water. Consumers will remain susceptible to sickness caused by handling and consuming the irrigated produce. Given this near-term outlook, public agencies in developing countries should seek opportunities to reduce the risks of infection and sickness by intervening at selected stages of the process, which includes wastewater generation, capture, irrigation, crop production, harvest and handling, and food preparation and consumption. Thus, we focus on the fourth category of policy options—reducing risk to farm households, communities, and consumers.

Policy Interventions Should Focus on Reducing Risk

Conventional wastewater treatment might be viewed as the ultimate risk reduction measure when considering the use of wastewater in irrigation.[45] Establishing and enforcing water quality standards, in conjunction with a wastewater treatment program, can be effective in removing potentially harmful constituents. However, the cost of treating wastewater and enforcing water quality standards will exceed affordability in many developing countries. Recognizing this challenge, the World Health Organization (WHO) recommends shifting the policy focus from reliance on wastewater treatment and water quality standards to establishing health-based targets that might be achieved by implementing a range of risk-reducing interventions.[11,45]

The WHO[46] describes three sets of health protection measures pertaining to the three groups most susceptible to health impacts of wastewater irrigation: 1) farmers and their families; 2) consumers of farm products; and 3) agricultural communities. We consider each group in turn.

Farmers and Their Families

When delivering irrigation water or working in fields irrigated with wastewater, farmers, family members, and other farm workers can be exposed to microbial pathogens, including viruses, bacteria, helminths (nematodes and tapeworms), and protozoa.[47] Wastewater also can contain endocrine-disrupting chemicals, pharmaceutically active compounds, and residuals of personal care products.[48–51] Exposure to wastewater can result in skin irritation and diseases related to pathogens in human waste products. The WHO[46] recommends the following protective measures to be considered when designing public policies and intervention strategies:

1. Treating wastewater
2. Supporting the use of personal protective equipment.
3. Providing access to safe drinking water and sanitation on farms
4. Promoting good health and hygiene practices
5. Providing chemotherapy and immunization
6. Controlling disease vectors and intermediate hosts
7. Reducing contact with disease vectors

One or more of these measures would be helpful in breaking or disrupting the pathway of contamination from wastewater to farm family members and farm workers. However, success will be determined by how effectively the benefits of these measures are communicated to farmers and how aggressively farm workers adopt them. The farm-level cost of any measure also will be a key determinant of its successful adoption.

Agricultural Communities

In a sense, many residents of agricultural communities are susceptible to the same types of risks as farmers and their families, particularly if they utilize water in irrigation canals or ditches or if they have access to farm fields. In many irrigated areas, community residents use water from irrigation canals or ditches for cleaning clothes, washing livestock, and watering kitchen gardens (Figure 2).[52] Young children often swim or play in irrigation ditches, while some residents rely on irrigation canals as a source of household drinking water.[53] The lack of knowledge regarding the potential health risks in many rural and periurban settings and the scarcity of fresh water supplies create situations in which many residents are at substantial risk. The WHO[46] recommends the following measures to protect members of agricultural communities:

1. Treating wastewater
2. Restricting access to irrigated fields and canals and ditches

FIGURE 2 Community members are at risk of contamination from wastewater carried in local stream, as they withdraw water for cooking, cleaning, and other purposes.
Source: Ben Keraita, IWMI, Ghana.

3. Providing safe recreational water, particularly for adolescents
4. Providing safe drinking water and sanitation facilities to communities
5. Promoting good health and hygiene practices
6. Providing chemotherapy and immunization
7. Controlling disease vectors and intermediate hosts
8. Reducing contact with disease vectors

Several of these measures are similar to those recommended to protect farm families and farm workers, given the similarity in exposure opportunities on farms and in the larger community. Many of the challenges involved in implementing the measures and encouraging sustainable adoption also would be similar.

Consumers of Farm Products

In many settings, in the absence of policy intervention, consumers might be the least informed group regarding the potential health risks due to wastewater irrigation. They might be unaware that farmers using wastewater have produced some of the fruits and vegetables for sale in local markets. They might also be unaware that some of the farm produce carries harmful pathogens and chemicals or that cooking the produce might reduce the likelihood of damage from infectious pathogens. Given these considerations, the WHO[46] recommends the following measures to reduce the risk to consumers:

1. Treating wastewater
2. Restricting the crops that are irrigated with wastewater
3. Promoting irrigation practices that minimize contamination of plants
4. Implementing withholding periods that allow pathogens to die between the last irrigation and harvest
5. Promoting hygienic practices at food markets and during food preparation
6. Promoting good health and hygiene practices
7. Promoting produce washing, disinfection, and cooking
8. Providing chemotherapy and immunization

Although enforcement will be difficult, public agencies might consider disallowing wastewater irrigation of vegetables and other crops that consumers often eat without cooking. Leafy vegetables, such as lettuce and spinach, are particularly prone to accumulating pathogens on edible portions of the plant when wastewater is applied directly over the plants and when irrigators splash contaminated soil particles on the leaves (Figure 3).[54] Modifying the spouts of watering cans will reduce contamination by reducing the splashing of soil particles (Figure 4).[54] Drip irrigation on the soil surface or below ground will minimize contamination,[55] but many poor farmers will not have the funds to invest in such systems (Figure 5).

Withholding periods between the date of last irrigation and harvest are sensible approaches as well, but monitoring and enforcement might be problematic in areas where wastewater irrigation is prevalent. Some farmers report that irrigating lettuce on the morning of the day of harvest freshens the crop and enhances its appearance in local markets.[54] Encouraging farmers to change such practices will be challenging, particularly given the perishable nature of leafy vegetable crops. Farmers generally want to obtain the highest price possible and to sell their produce quickly, before its appearance and quality begin to fade.

Public efforts to improve hygienic practices and food preparation at homes and in the marketplace also will be challenging. In areas where small-scale farmers sell produce to small-scale vendors who resell the produce in a restaurant or fast-food outlet, individuals have little incentive to assume the extra cost of enhanced food treatment. This situation in which information is limited and asymmetric can be described also an externality involving producers and consumers. The benefits of a cleaner, safer food

FIGURE 3 Leafy vegetables are susceptible to contamination by pathogens in wastewater used for irrigation, particularly when watered with sprinklers or cans, as in this photograph from Ghana.
Source: Ben Keraita, IWMI.

FIGURE 4 Using a watering can spout with many small holes reduces the splashing of soil particles onto leaves of lettuce and other leafy vegetables.
Source: Ben Keraita, IWMI.

FIGURE 5 Drip irrigation minimizes contact of irrigation water with plant foliage, thus reducing the risk of contamination by pathogens in wastewater.
Source: Frank Rijsberman, IWMI.

supply accrue to consumers and communities, rather than to the farmers and food shop owners who will incur higher costs if they implement improved production, washing, and handling practices. Public policy is needed to ensure that farmers and vendors internalize the external costs of their activities.

Motivating Safe Practices along the Wastewater Exposure Pathway

From a policy perspective, it is helpful to view the use of wastewater in irrigation as an activity along a pathway on which many individuals and communities are exposed to potential contamination. The pathway begins with wastewater generation in households, companies, and industries. Farmers, fishers, and members of urban and rural communities who utilize water in streams or drains carrying wastewater are exposed to contamination as they conduct their activities. Farmers and laborers applying irrigation water or working in farm fields are exposed, as are community members who spend time in fields irrigated with wastewater. The exposure pathway continues from the farm to food processors, vendors, and consumers.

Along the wastewater exposure pathway, farmers and food vendors generate value as they produce and market crops that are desired by consumers. That value provides motivation for reducing risk, while maintaining irrigation opportunities. The value also can provide a partial source of finance for implementing measures to reduce risk. Both farmers and food vendors have a financial stake in sustaining their activities, and they should be able and willing to pay some portion of the cost of achieving acceptable risk levels. Given this perspective, cost-sharing programs in which farmers pay a portion of the cost of adopting safer irrigation methods might be helpful in promoting rapid adoption. Food vendors also might be offered cost-sharing programs for adopting measures that will improve hygiene and preparation of irrigated produce.

Local officials might also consider ways to motivate farmers and food vendors to participate actively in the study and enhancement of crop production and food preparation methods in areas where wastewater is used for irrigation. At present, it is likely that many farmers and vendors prefer to operate with minimal visibility, to reduce the probability that a public agency or consumer organization might endeavor to end the practice of wastewater irrigation. The public will gain value by encouraging farmers and vendors to explain their practices, maintain production records, gather information regarding water quality and food safety, and participate in training programs on irrigation and food preparation. Public officials might offer farmers greater security of land tenure or access to affordable credit, in exchange for active participation. Food vendors also might respond positively to the offer of affordable credit or the opportunity to obtain a license that enables them to sell food products in a desirable location for several years.

Consumers also can be expected to contribute some portion of the cost of reducing risk along the wastewater exposure pathway. Consumer willingness to pay market prices for fresh produce reflects the value generated by farmers and food vendors in making the produce available. Fresh fruits and vegetables enhance dietary intake in important ways, provided the produce is free of pathogens and chemicals. The net value consumers actually obtain from wastewater-irrigated produce will be diminished if they become ill due to interaction with pathogens or chemicals. Hence, they should be willing to pay a premium for produce that is certified free from potentially harmful constituents.

Local governments might consider implementing food safety assurance programs, in which trained professionals monitor and certify the production of farm produce in areas where wastewater is used for irrigation. The programs might be started on a voluntary basis, with the goal of eventually including all farmers using wastewater to produce vegetables for sale in local markets. Funds for the program could be generated through an assessment on the sale of farm products to food vendors, or at the point of sale from vendors to consumers. In either case, consumers would pay a portion of the cost of assuring food safety.

In summary, a key question for policy makers is how to enhance the values generated along the wastewater exposure pathway, while also encouraging farmers and food vendors to use a portion of those values to improve food safety and reduce farm-level and community risks. Poor farmers might initially be resistant to changing irrigation practices in ways that reduce contamination of their produce, but they should be encouraged to consider their financial situation with and without access to wastewater. To the extent that wastewater irrigation enhances farm-level net income, there should be some farm-level willingness and ability to invest in safe irrigation practices. Public officials might consider implementing cost-sharing programs to encourage farmers to invest in drip irrigation systems or other watering methods that reduce the risks associated with wastewater irrigation.

The same perspective applies with regard to food vendors. To the extent that their sales are enhanced through access to crops irrigated with wastewater, food vendors should be able and willing to invest some portion of their net returns in efforts to improve food safety. They might be encouraged to purchase vegetable washing devices and to clean all produce carefully before it is sold to consumers. A successful campaign of improving food safety might enhance food sales, as consumers learn of the campaign and as they suffer fewer health effects from consuming vegetables produced by local farmers.

Public officials also might consider implementing farm produce certification programs in which farmers who agree to improve their cultural practices in ways that reduce contamination are given special recognition. For example, one can envision a “consumer-safe” labeling program in which vegetables produced using risk-minimizing irrigation methods are given a label denoting that status. If implemented along with an innovative marketing campaign, consumers might be willing to pay a premium price for consumer-safe produce, thus rewarding farmers for investing in safe irrigation practices.

A similar program could be implemented for food vendors who purchase consumer-safe produce from farmers and then prepare the food for sale using safe handling and preparation practices. One can envision a consumer- safe label affixed to the storefronts of participating food vendors. Over time, if the program is successful, market forces might guide both farmers and food vendors to join the consumer-safe program, as the demand for non-labeled produce and food diminishes.

Examples of Public Policies

Helpful examples of public policies regarding wastewater use in irrigation are found in the Middle East and North Africa and other regions where farmers have been using treated and untreated wastewater for many years. In some countries, such as Egypt, the volume of municipal wastewater exceeds the treatment capacity, and large volumes of untreated wastewater enter agricultural drains.[56] The government attempts to manage the blending of treated and untreated wastewater with agricultural drainage water and the use of blended water by farmers, but success is limited by the scale of the problem and the strong demand for supplemental water supplies in the Nile Delta. Irrigation with treated wastewater will increase over time, with the expansion of wastewater treatment capacity. Egypt has developed a code of practice for using treated wastewater in agriculture (Table 3).

Several countries in the region, including Algeria, Cyprus, and Tunisia, do not allow the irrigation of vegetables with treated wastewater. Cyprus also disallows the irrigation of ornamental plants destined for sale in international markets.[57] Wastewater policies are well developed in Cyprus and Tunisia, where the governments actively support and regulate wastewater treatment and reuse. In Cyprus, the government pays for large portions of the cost of water treatment plants in cities and villages, while also paying for the distribution of wastewater to farmers (Table 4). Tunisia requires that industries comply with wastewater discharge standards designed to support reuse on farms, golf courses, and landscapes, and also for aquifer recharge (Table 4). Saudi Arabia plans to use all of its treated wastewater, primarily in agriculture. The city of Muscat in Oman has installed an extensive drip irrigation system for irrigating landscapes with treated municipal wastewater.[60]

In other regions, Italy has established water quality criteria regarding the use of treated wastewater on vegetables and grazing crops, while several autonomous provinces in Spain have developed legal prescriptions or recommendations regarding wastewater reuse.[61] Wastewater accounts for an estimated 41% of the irrigation water used on Spanish golf courses.[62]

The government of Botswana encourages greater reuse of wastewater in irrigation and mining, in part, by ending its policy of providing fresh water supplies at subsidized prices.[63] Botswana also is considering how to account for wastewater volumes within its national water accounting framework.[64]

Public officials in countries with little experience in regulating the use of wastewater in irrigation can gain value by reviewing the examples presented here and by considering ways to engage producers and consumers in active discussion of wastewater issues. As in many regulatory settings, the prospect of new rules and procedures regarding wastewater irrigation and food preparation will be viewed initially as a cost-increasing outcome that will harm the financial performance of individual farmers and food vendors. Hence, the rational strategy from an individual's perspective involves a combination of maintaining a low profile and quietly lobbying against the adoption of any new programs. Yet, in aggregate, net social welfare is decreased if the sum of damages from using wastewater in irrigation exceeds the sum of the benefits.

TABLE 3 The Cairo East Bank Effluent Re-use Project

Goal

Examine the potential for reusing in agriculture the treated wastewater from three sewage treatment plants that serve Cairo communities on the east bank of the Nile River.

Key Issues

1. What is the best way to distribute and allocate the treated wastewater?
2. What are the safest and most efficient ways to supply and use the wastewater?

Outcomes

1. A draft code of practice, which addresses irrigation requirements, legislative needs, water quality standards, health issues, and monitoring programs.
2. An institutional framework to ensure the design of safe and efficient water reuse projects.

Source: Angelakis et al.[58]

TABLE 4 Examples of Wastewater Policies in the Middle East

Cyprus
The government pays for 75% of the cost of water treatment plants in villages. In cities, the government pays the full costs of construction and operation of tertiary treatment plants. In addition, the government pays the cost of distributing wastewater to farmers.
Egypt
The Ministry of Agriculture supports the restricted reuse of treated wastewater for irrigation of nonfood crops, including trees for timber and green belts designed to stabilize sand dunes. Large volumes of untreated wastewater flow into agricultural drains, where farmers withdraw water for irrigation.
Jordan
Wastewater treatment standards were introduced in 1982, and they have been modified in recent years. Vegetables may not be irrigated directly with treated wastewater.
Kuwait
Treated wastewater may not be used for landscape irrigation. It may be used only to irrigate "safe crops."
Saudi Arabia
The Kingdom plans to use all of its treated wastewater, primarily in agriculture. Other uses include landscape irrigation and aquifer recharge. Guidelines requiring secondary and tertiary treatment have been developed to support the unrestricted use of wastewater for irrigation.
Syria
Wastewater irrigation is restricted to fodder crops, industrial crops, and fruit trees, but enforcement is not complete.
Tunisia
Industries must comply with discharge standards to support intensive development of wastewater reuse on farms, golf courses, and landscapes, and also for aquifer recharge.

Source: Bakir.[59]

Perhaps the key to starting policy discussions is to demonstrate the potential gains in aggregate net benefits. Farmers, food vendors, and consumers can gain value together as they work with public officials to develop safe practices in crop production and food preparation. Individual farmers and food vendors will not be disadvantaged if everyone agrees to adopt safe practices and if consumers are willing to pay higher prices in return for safety assurances. Details regarding policy parameters and effective monitoring and enforcement programs can be developed over time, once all parties appreciate the potential gains in net benefits made possible through the safe and efficient use of wastewater in agriculture and the preparation of healthful food products.

Conclusions

Policy issues regarding the use of wastewater in irrigation are quite different in developed and developing countries. In developed countries, most municipal and industrial wastewater is treated, and thus, most of the wastewater used in agriculture is treated. Protective guidelines regarding the quality of wastewater used for irrigation have been in place for many years. Policy issues in developed countries pertain largely to financial and economic considerations regarding the improvement and expansion of wastewater treatment facilities. Public officials and water management agencies motivate greater use of wastewater by providing financial incentives and increasing public awareness of the safety and benefits of using treated wastewater on farms, golf courses, and urban landscapes.

Policy issues in developing countries also include financial and economic questions regarding investments in wastewater treatment. However, in many countries, the pace of such investments will not be sufficient to meet demand. Much of the wastewater generated in cities and rural areas will remain untreated for many years. As a result, farmers will continue to use untreated wastewater for irrigation, and its use will be largely unintentional and informal. Public officials must therefore implement risk reduction programs that protect farm families, communities, food vendors, and consumers from the potentially harmful effects of exposure to the pathogens and chemicals in untreated wastewater.

Policy options in developing countries will reflect a range of interventions along the pathway that includes wastewater generation, irrigation water capture and use, crop production and harvest, food preparation, and consumption. Public officials can implement risk-reducing guidelines and programs at each stage along the wastewater exposure pathway. For example, public officials can support improvements in wastewater treatment at the point of generation, when funds for such improvements are available. Officials also can call for changes in household and industrial production practices that would reduce the loads of harmful constituents in wastewater, thus reducing concentrations of those constituents in the irrigation water diverted from streams and ditches by farmers.

At the farm level, public agencies can provide technical assistance regarding water diversion and irrigation methods that would reduce potential exposure of farm workers to harmful pathogens and chemicals. Technical assistance regarding irrigation methods that reduce contamination of leafy vegetables and other produce consumed without cooking is essential for reducing risks to food vendors and consumers. Although difficult to enforce, regulations that establish a minimum time period between the dates of last irrigation and harvest would be helpful in reducing the risk of contamination from agricultural products.

Public officials in developing countries might also consider implementing certification programs for consumer-safe farm produce, particularly in markets where local farmers sell their irrigated vegetables. Public agencies can begin such programs, with support from farmers and food vendors, but eventually, market forces must arise to sustain them. Consumers must find value in certified produce, and they must be willing to pay a small premium that compensates farmers and vendors for their costs in providing the safer produce. Educational and marketing campaigns can be helpful in boosting demand for safe produce among consumers.

The policy issues we describe in this entry pertain largely to near-term strategies for minimizing the risk of negative health effects, while also enabling farmers to gain the potential benefits of using untreated and partially treated wastewater in agriculture. This approach is appropriate for countries that presently cannot afford to build, operate, and maintain a full complement of modern wastewater treatment facilities. Over time, as the demand for water in agriculture and other uses continues to increase, public officials in all countries should endeavor to provide wastewater treatment that matches end uses, including the irrigation of crops, landscapes, and golf courses. In developing countries, it will be necessary also to ensure that small-scale farmers retain access to a reliable source of irrigation water when the untreated and commingled wastewater they once relied on becomes unavailable, with the expansion of wastewater treatment programs.

Acknowledgments

I appreciate very much the helpful comments and suggestions of an anonymous reviewer.

References

1. Gupta, R.; Gangopadhyay, S.G. Peri-urban agriculture and aquaculture. Econ. Political Wkly. **2006**, *41* (18), 1757–1760.
2. Qian, Y.L.; Mecham, B. Long-term effects of recycled wastewater irrigation on soil chemical properties on golf course fairways. Agron. J. **2005**, *97* (3), 717–721.
3. Barker-Reid, F.; Harper, G.A.; Hamilton, A.J. Affluent effluent: Growing vegetables with wastewater in Melbourne, Australia—A wealthy but bone-dry city. Irrig. Drain. Syst. **2010**, *24* (1–2), 79–94.
4. Pedrero, F.; Kalavrouziotis, I.; Alarcón, J.J.; Koukoula- kis, P.; Asano, T. Use of treated municipal wastewater in irrigated agriculture—Review of practices in Spain and Greece. Agric. Water Manage. **2010**, *97* (9), 1233–1241.
5. Qadir, M.; Bahri, A.; Sato, T.; Al-Karadsheh, E. Wastewater production, treatment, and irrigation in Middle East and North Africa. Irrig. Drain. Syst. **2010**, *24* (1–2), 37–51.

6. Sheng, Z. An aquifer storage and recovery system with reclaimed wastewater to preserve native groundwater resources in El Paso, Texas. J. Environ. Manage. **2005**, *75* (4), 367–377.
7. Jamwal, P.; Mittal, A.K. Reuse of treated sewage in Delhi city: Microbial evaluation of STPs and reuse options. Resour., Conserv. Recycl. **2010**, *54* (4), 211–221.
8. Bixio, D.; Thoeye, C.; De Koning, J.; Joksimovic, D.; Savic, D.; Wintgens, T.; Melin, T. Wastewater reuse in Europe. Desalination **2006**, *187* (1–3), 89–101.
9. Salgot, M.; Huertas, E.; Weber, S.; Dott, W.; Hollender, J. Wastewater reuse and risk: Definition of key objectives. Desalination **2006**, *187* (1–3), 29–40.
10. Tal, A. Seeking sustainability: Israel's evolving water management strategy. Science **2006**, *213*, 1081–1084.
11. World Health Organization (WHO). *Guidelines for the Safe Use of Wastewater, Excreta and Greywater.* Volume 1: Policy and regulatory aspects; World Health Organization: Geneva, 2006a.
12. Akber, A.; Mukhopadhyay, A.; Al-Senafy, M.; Al-Haddad, A.; Al-Awadi, E.; Al-Qallaf, H. Feasibility of long-term irrigation as a treatment method for municipal wastewater using natural soil in Kuwait. Agric. Water Manage. **2008**, *95* (3), 233–242.
13. Hamilton, A.J.; Stagnitti, F.; Xiong, X.; Kreidl, S.L.; Benke, K.K.; Maher, P. Wastewater irrigation: The state of play. Vadose Zone J. **2007**, *6* (4), 823–840.
14. Oron, G. Water resources management and wastewater reuse for agriculture in Israel. In *Wastewater Reclamation and Reuse*; Asano, T., Ed.; CBC Press: Boca Raton, FL, 1998; 757–778.
15. van der Bruggen, B. The global water recycling situation. Sustainable Sci. Eng. **2010**, *2*, 41–62.
16. Hussain, G.; Al-Saati, A.J. Wastewater quality and its reuse in agriculture in Saudi Arabia. Desalination **1999**, *123* (2–3), 241–251.
17. Al-Rashed, M.F.; Sherif, M.M. Water resources in the GCC countries: an overview. Water Resour. Manage. **2000**, *14* (1), 59–75.
18. Al-Katheeri, E.S. Towards the establishment of water management in Abu Dhabi Emirate. Water Resour. Manage. **2008**, *22* (2), 205–215.
19. Dillon, P.; Pavelic, P.; Toze, S.; Rinck-Pfeiffer, S.; Martin, R.; Knapton, A.; Pidsley, D. Role of aquifer storage in water reuse. Desalination **2006**, *188* (1–3), 123–134.
20. Jiménez, B.; Drechsel, P.; Koné, D.; Bahri, A.; Raschid-Sally, L.; Qadir, M. Wastewater, sludge and excreta use in developing countries: An overview. In *Wastewater Irrigation and Health: Assessing and Mitigating Risk in Low-Income Countries;* Drechsel, P., Scott, C.A., Raschid-Sally, L., Redwood, M., Bahri, A., Eds.; Earthscan: London, 2010; 3–27.
21. Bradford, A.; Brook, R.; Hunshal, C.S. Wastewater irrigation in Hubli-Dharwad, India: Implications for health and livelihoods. Environ. Urbanization **2003**, *15* (2), 157–170.
22. Ensink, J.H.J.; Mahmood, T.; van der Hoek, W.; Raschid- Sally, L.; Amerasinghe, F.P. A nationwide assessment of wastewater use in Pakistan: An obscure activity or a vitally important one? Water Policy **2004**, *6* (3), 197–206.
23. Bos, R.; Carr, R.; Keraita, B. Assessing and mitigating wastewater-related health risks in low-income countries: An introduction. In *Wastewater Irrigation and Health: Assessing and Mitigating Risk in Low-Income Countries*; Drechsel, P. Scott, C.A., Raschid-Sally, L., Redwood, M., Bahri, A., Eds.; Earthscan: London, 2010; 29–47.
24. Keraita, B.; Drechsel, P.; Amoah, P. Influence of urban wastewater on stream water quality and agriculture in and around Kumasi, Ghana. Environ. Urbanization. **2003**, *15* (2), 171–178.
25. Amoah, P.; Drechsel, P.; Abaidoo, R.C.; Henseler, M. Irrigated urban vegetable production in Ghana: Microbiological contamination in farms and markets and associated consumer risk groups. J. Water Health **2007**, *5* (3), 455–466.
26. Tiongco, M.M.; Narrod, C.A.; Bidwell, K. Risk analysis integrating livelihood and economic impacts of wastewater irrigation on health. In *Wastewater Irrigation and Health: Assessing and Mitigating Risk in Low-Income Countries*; Drechsel, P., Scott, C.A., Raschid-Sally, L., Redwood, M., Bahri, A., Eds.; Earthscan: London, 2010; 127–145.

27. Qadir, M.; Wichelns, D.; Raschid-Sally, L.; McCornick, P.G.; Drechsel, P.; Bahri, A.; Minhas, P.S. The challenges of wastewater irrigation in developing countries. Agric. Water Manage. **2010**, *97* (4), 561–568.
28. Rutkowski, T.; Raschid-Sally, L.; Buechler, S. Wastewater irrigation in the developing world—Two case studies from the Kathmandu Valley in Nepal. Agric. Water Manage. **2007**, *88* (1–3), 83–91.
29. Drechsel, P.; Graefe, S.; Sonou, M.; Cofie, O.O. *Informal Irrigation in Urban West Africa: An Overview*, IWMI Research Report 102; International Water Management Institute: Colombo, Sri Lanka, 2006.
30. Van Rooijen, D.J.; Turral, H.; Biggs, T.W. Sponge city: Water balance of mega-city water use and wastewater use in Hyderabad, India. Irrig. Drain. **2005**, *54* (Suppl. 1), S81–S91.
31. Srinivasan, J.T.; Reddy, V.R. Impact of irrigation water quality on human health: A case study in India. Ecol. Econ. **2009**, *68* (11), 2800–2807.
32. Ensink, J.H.J.; Scott, C.A.; Brooker, S.; Cairncross, S. Sewage disposal in the Musi-River, India: Water quality remediation through irrigation infrastructure. Irrig. Drain. Syst. **2010**, *24* (1–3), 65–77.
33. Meinzinger, F.; Oldenburg, M.; Otterpohl, R. No waste, but a resource: Alternative approaches to urban sanitation in Ethiopia. Desalination **2009**, *248* (1–3), 322–329.
34. Mills, R.A.; Karajeh, F.; Hultquist, R.H. California's task force evaluation of issues confronting water reuse. Water Sci. Technol. **2004**, *50* (2), 301–308.
35. van Roon, M. Water localization and reclamation: Steps towards low impact urban design and development. J. Environ. Manage. **2007**, *83* (4), 437–447.
36. Hurlimann, A.; McKay, J. Urban Australians using recycled water for domestic non-potable use—An evaluation of the attributes price, saltiness, colour and odour using conjoint analysis. J. Environ. Manage. **2007**, *83* (1), 93–104.
37. Hurlimann, A.C. Water supply in regional Victoria Australia: A review of the water cartage industry and willingness to pay for recycled water. Resour. Conserv. Recycl. **2009**, *53* (5), 262–268.
38. United Nations Educational, Scientific, and Cultural Organization (UNESCO). Water for People, Water for Life - UN World Water Development Report. UNESCO Publishing: Paris, 2003; 36 pages.
39. Drechsel, P.; Evans, A.E.V. Wastewater use in irrigated agriculture. Irrig. Drain. Syst. **2010**, *24* (1–2), 1–3.
40. Fine, P.; Halperin, R.; Hadas E. Economic considerations for wastewater upgrading alternatives: An Israeli test case. J. Environ. Manage. **2006**, *78* (2), 163–169.
41. Bakopoulou, S.; Polyzos, S.; Kungolos, A. Investigation of farmers' willingness to pay for using recycled water for irrigation in Thessaly region, Greece. Desalination **2010**, *250* (1), 329–334.
42. Tsagarakis, K.P.; Georgantzis, N. The role of information on farmers' willingness to use recycled water for irrigation. Water Sci. Technol.: Water Supply **2003**, *3* (4), 105–113.
43. Toze, S. Reuse of effluent water—Benefits and risks. Agric. Water Manage. **2006**, *80* (1–3), 147–159.
44. van Lier, J.B.; Huibers, F.P. From unplanned to planned agricultural use: Making an asset out of wastewater. Irrig. Drain. Syst. **2010**, *24* (1–2), 143–152.
45. Keraita, B.; Drechsel, P.; Konradsen, F. Up and down the sanitation ladder: Harmonizing the treatment and multiple- barrier perspectives on risk reduction in wastewater irrigated agriculture. Irrig. Drain. Syst. **2010**, *24* (1–2), 24–35.
46. World Health Organization (WHO). *Guidelines for the Safe Use of Wastewater, Excreta and Greywater.* Volume 2: Wastewater use in agriculture; World Health Organization: Geneva, 2006.
47. Toze, S. Water reuse and health risks—Real vs. perceived. Desalination **2006**, *187* (1–3), 41–51.
48. Ternes, T.A.; Bonerz, M.; Herrmann, N.; Teiser, B.; Andersen, H.R. Irrigation of wastewater in Braunschweig, Germany: An option to remove pharmaceuticals and musk fragrances. Chemosphere **2007**, *66* (5), 894–904.

49. Lapen, D.R.; Topp, E.; Metcalfe, C.D.; Li, H.; Edwards, M.; Gottschall, N.; Bolton, P.; Curnoe, W.; Payne, M.; Beck, A. Pharmaceutical and personal care products in tile drainage following land application of municipal biosolids. Sci. Total Environ. **2008**, *399* (1–3), 50–65.
50. Siemens, J.; Huschek, G.; Siebe, C.; Kaupenjohann, M. Concentrations and mobility of human pharmaceuticals in the world's largest wastewater irrigation system, Mexico City- Mezquital Valley. Water Res. **2008,** *42* (8–9), 2124–2134.
51. Topp, E.; Monteiro, S.C.; Beck, A.; Coelho, B.B.; Boxall, A.B.A.; Duenk, P.W.; Kleywegt, S.; Lapen, D.R.; Payne, M.; Sabourin, L.; Li, H.; Metcalfe, C.D. Pharmaceutical and personal care products in tile drainage following land application of municipal biosolids. Sci. Total Environ. **2008**, *396* (1), 52–59.
52. Meinzen-Dick, R.; van der Hoek, W. Multiple uses of water in irrigated areas. Irrig. Drain. Syst. **2001**, *15* (2), 93–98.
53. Senzanje, A.; Boelee, E.; Rusere, S. Multiple use of water and water productivity of communal small dams in the Limpopo Basin, Zimbabwe. Irrig. Drain. Syst. **2008**, *22* (3–4), 225–237.
54. Keraita, B.; Konradsen, F.; Drechsel, P. Farm-based measures for reducing microbiological health risks for consumers from informal wastewater-irrigated agriculture. In *Wastewater Irrigation and Health: Assessing and Mitigating Risk in Low-Income Countries;* Drechsel, P., Scott, C.A., Raschid-Sally, L., Redwood, M., Bahri, A., Eds.; Earthscan: London, 2010; 188–207.
55. Capra, A.; Scicolone, B. Recycling of poor quality urban wastewater by drip irrigation systems. J. Cleaner Prod. **2007**, *15* (16), 1529–1534.
56. Abdel-Dayem, S.; Abdel-Gawad, S.; Fahmy, H. Drainage in Egypt: A story of determination, continuity, and success. Irrig. Drain. **2007**, *56* (Suppl. 1), S101-S111.
57. Lawrence, P.; Adham, S.; Barrott, L. Ensuring water reuse projects succeed—Institutional and technical issues for treated wastewater reuse. Desalination **2002**, *152* (1–3), 291–298.
58. Angelakis, A.N.; Marecos do Monte, M.H.F.; Bontoux, L.; Asano, T. The status of wastewater reuse practice in the Mediterranean Basin: Need for guidelines. Water Res. **1999**, *33* (10), 2201–2217.
59. Bakir, H.A. Sustainable wastewater management for small communities in the Middle East and North Africa. J. Environ. Manage. **2001**, *61* (4), 319–328.
60. Bazza, M. Wastewater recycling and reuse in the Near East Region: Experience and issues. Water Sci. Technol.: Water Supply **2003**, *3* (4), 33–50.
61. Angelakis, A.N.; Bontoux, L.; Lazarova, V. Challenges and prospectives for water recycling and reuse in EU countries. Water Sci. Technol.: Water Supply **2003**, *3* (4), 59–68.
62. Rodriguez Diaz, J.A.; Knox, J.W.; Weatherhead, E.K. Competing demands for irrigation water: Golf and agriculture in Spain. Irrig. Drain. **2007**, *56* (5), 541–549.
63. Swatuk, L.A.; Rahm, D. Integrating policy, disintegrating practice: Water resources management in Botswana. Phys. Chem. Earth **2004**, *29* (15–18), 1357–1364.
64. Arntzen, J.W.; Setlhogile, T. Mainstreaming wastewater through water accounting: The example of Botswana. Phys. Chem. Earth **2007**, *32* (15–18), 1221–1230.

47

Water: Total Maximum Daily Load

Robin Kundis Craig

Introduction

This entry explains what a "total maximum daily load" (TMDL) is and how TMDLs help regulators to achieve the ultimate goals of the Federal Water Pollution Control Act (FWPCA), more commonly known (since the 1977 amendments) as the Clean Water Act.[1] A TMDL is both a regulatory and an informational tool that helps regulators to "restore and maintain the chemical, physical, and biological integrity of the Nation's waters"[2] by ensuring that each water body in the United States meets its water quality standards. As such, the Act's TMDL provisions also force regulators to take a broader look at water body integrity, and hence, the TMDL requirements could prompt state water quality managers in particular to move water quality regulation beyond individual source requirements to a more systemic and watershed-based pollution control program.

Congress first enacted the FWPCA in 1948.[1] However, early versions of the FWPCA left water quality regulation almost entirely to the states, providing almost nothing that qualifies as a true regulatory program, although the 1965 Water Quality Act unsuccessfully tried to prompt states to adopt water quality standards programs.[1] In 1972, however, Congress significantly amended the Act in order to better protect water quality, significantly expanding the federal role in water quality regulation and creating mandatory regulatory requirements.[1] Specifically, Congress relied upon three main mechanisms to improve the quality and integrity of the Nation's waters: two federal permit programs that regulate the discharges of pollutants from point sources and state nonpoint source control programs that are supposed to manage more diffuse sources of water pollution, such as agricultural runoff.[1]

The Act's two permit programs and the state nonpoint source control programs are source-focused mechanisms for improving water quality—that is, they control water pollution by regulating or managing

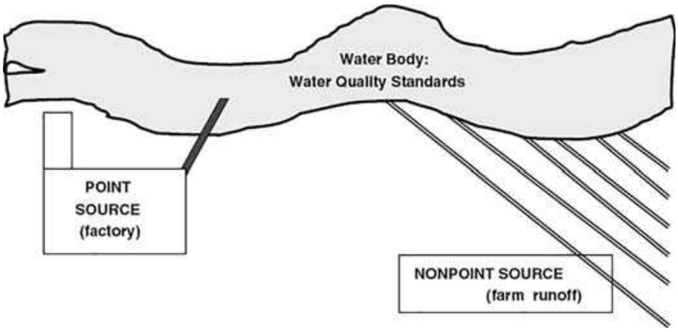


FIGURE 1 Clean Water Act overview.

the sources of that pollution.[3] Point sources—that is, "discernible, confined, and discrete conveyances," like pipes[4]—receive one of two kinds of permits. National Pollutant Discharge Elimination System (NPDES) permits dictate limits on how much of what kinds of pollutants the source can discharge into the water, predominantly based on technology-based effluent limitations.[1] Section 404 "dredge and fill" permits limit both how much dredged or fill material a source can discharge into waters and the broader impacts of construction projects in waters.[1] Similarly, when states manage nonpoint sources, they generally dictate how those nonpoint sources should manage water pollution, such as through best management practices (BMPs).[1]

However, neither the permit programs nor the state nonpoint source management programs dictate specific goals for water quality regulation—that is, nothing in the permit limitations or in BMPs specifies the ultimate water quality goals that their regulatory mechanisms are intended to achieve. Moreover, a single set of national water quality goals would not suffice, because water bodies naturally differ in their water quality characteristics. A headwater stream in the Rocky Mountains is not the Mississippi River, and Lake Erie is not Crater Lake.

Instead, the Clean Water Act uses water quality standards to determine the individual water quality goals for specific water bodies[1] (see Figure 1). The regulatory issue then becomes the following: how do regulators connect the Clean Water Act's two regulatory programs and the state nonpoint source management programs to these water quality standards? Although other mechanisms also exist,[1] the TMDL is the ultimate calculation that the Clean Water Act prescribes to connect water quality standards—the water quality goals for particular water bodies—and the Act's source-based regulatory mechanisms.

A TMDL is, literally, the "total maximum daily load"—that is, the maximum amount of a given pollutant that can be added to a particular water body on a daily basis while still having that water body meets its water quality standards. Individual states and the federal EPA use TMDLs when the national technology-based effluent limitations in NPDES permits, the water quality requirements in Section 404 dredge and fill permits, and the standard state nonpoint source control requirements are not stringent enough to allow a particular water body to achieve its water quality standards. Proper employment of a TMDL can require adjustments to the discharge permits, adjustments to the nonpoint source control requirements, or both.

Regulation of Point Sources

Statutory Trigger: A "Discharge of a Pollutant"

Total maximum daily loads cannot be understood without understanding the Clean Water Act's overall regulatory structure. As a starting point, the Act's regulatory provisions prohibit any "discharge of a pollutant" except as in compliance with the Act.[5] The Act defines "discharge of a pollutant" to be

"(A) any addition of any pollutant to navigable waters from any point source, (B) any addition of any pollutant to the waters of the contiguous zone or the ocean from any point source other than a vessel or other floating craft." [4] The Act does not define "addition," but case law has defined that term to mean any human-controlled contribution of pollutants to regulated waters.[1] The Act defines "pollutant" broadly to mean "dredged spoil, solid waste, incinerator residue, sewage, garbage, sewage sludge, munitions, chemical wastes, biological materials, radioactive materials, heat, wrecked or discarded equipment, rock, sand, cellar dirt and industrial, municipal, and agricultural waste discharged into water."[4] Case law has made clear that the Act does not apply to small, isolated waters.[6] By regulation, the "navigable waters," the "contiguous zone," and the "oceans" are defined so as to include almost everything else,[4] but in 2006, the U.S. Supreme Court issued an indecisive opinion regarding the test that the agencies and the courts should use to determine whether other smaller waters should qualify as "navigable waters" under the Act.[7] Finally, a "point source" is "any discernible, confined and discrete conveyance," like a pipe.[4]

NPDES Permit Program

In order to comply with the Clean Water Act, point sources that discharge pollutants must get a permit. The Act's most generally applicable permit program is the NPDES permit program, which applies to any "discharges of a pollutant" except discharges of dredged or fill material.[8,9] At the federal level, the EPA has the primary authority to implement the NPDES permit program,[8] although it has delegated much of its permitting authority to the individual states.[10] Even where states issue the NPDES permit, however, the EPA retains supervisory authority and can override state authority to issue individual permits.[8,11]

When a point source gets an NPDES permit, most of the requirements governing its discharge will be based on national effluent limitations[1] (see Figure 2). Effluent limitations are "end of the pipe," numerical limitations on the concentrations of pollutants that a discharger can discharge.[4] The EPA generally sets effluent limitations on an industry-wide basis.[5] Moreover, the effluent limitations are technology based. For example, the EPA currently sets most effluent limitations for most industries on the basis of the "best available technology economically achievable" for each category of industry.[5] In contrast, effluent limitations for publicly owned treatment works (or sewage treatment plants) are based on secondary treatment of the sewage. Where the EPA has not issued national technology-based effluent limitations, states write equivalent limits into individual permits.

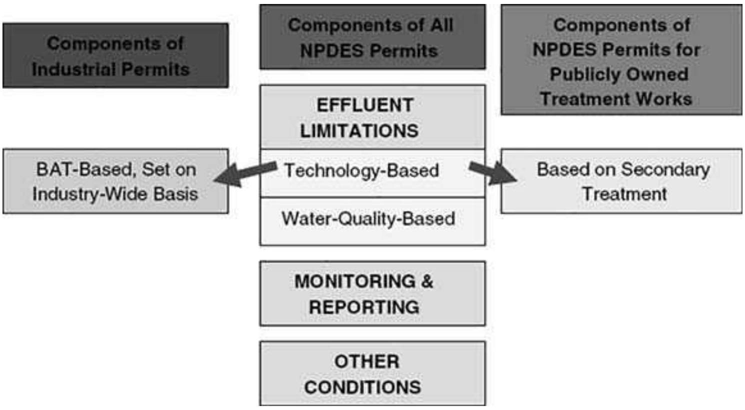


FIGURE 2 Components of an NPDES Permit.
Source: Adopted from U.S. EPA.[32]

Section 404 Dredge and Fill Permit Program

The Clean Water Act's second permit program, the Section 404 dredge and fill permit program, applies only to discharges of dredged or fill material.[9] At the federal level, the U.S. Army Corps of Engineers received the primary authority to implement this permit program, although the EPA retains oversight authority and can veto Section 404 permits.[9] States can also acquire authority to issue some Section 404 permits,[9] but to date, only two states—Michigan and New Jersey—have chosen to do so.[12] As a result, the Army Corps issues most Section 404 permits.

Under the Army Corps' regulations, "the term discharge of dredged material means any addition of dredged material into, including redeposit of dredged material other than incidental fallback within, the waters of the United States"[13]; "dredged material" is "material that is excavated or dredged from waters of the United States."[13] In turn, "[t]he term discharge of fill material means the addition of fill material into waters of the United States,"[13] where "fill material" is "material placed in waters of the United States where the material has the effect of: (1) Replacing any portion of a water of the United States with dry land; or (2) Changing the bottom elevation of any portion of a water of the United States."[13] Thus, Section 404 permits most often regulate construction in or filling of waters of the United States, especially wetlands.

In issuing an individual Section 404 permit (general permits are available for several kinds of smaller activities with limited impact on the environment), the Army Corps reviews the proposed activity against two sets of permitting criteria: the EPA's Section 404(b) Guidelines[9,14] and the Corps' public interest review requirements.[15] Both sets of criteria provide the Army Corps and, in its oversight capacity, the EPA with means to deny individual Section 404 permits to activities that are too damaging to the environment. However, the more common result of the agencies' review of projects in light of the Guidelines and the public interest criteria is that the permit applicant will have to modify the project's design, engage in mitigation, or both in order to receive the Section 404 permit.

Role of Water Quality Standards in Point Source Permitting, before the TMDL Process

The states received the primary authority to set water quality standards.[16] As noted, water quality standards establish the ultimate goals of water pollution regulation for individual water bodies. Specifically, water quality standards "consist of the *designated uses* of the navigable waters involved and the *water quality criteria* for such waters based upon such uses" (emphasis added).[16]

As far as designated uses are concerned, Congress directed states to consider waters' uses for public water supply, fish and wildlife support, recreation, agriculture, industry, and navigation.[16] However, states are free to designate other uses, as well (see Figure 3).

Water quality criteria, in turn, can be either numeric or narrative. State water quality criteria for toxic pollutants, for instance, have often been narrative in form—typically, "no toxic pollutants in toxic amounts." However, the EPA is strongly encouraging states to move to numeric water quality criteria for all pollutants of concern, and states setting numeric water quality criteria often borrow the EPA's water quality criteria guidelines.[17]

Whether numeric or narrative, however, the water quality criteria are supposed to specify the water quality, along multiple parameters (temperature, salinity, turbidity, toxicity, oxygen content, and so forth), that is necessary to achieve the waterway's designated uses.[16] For example, waters designated for salmon populations generally require cold, clear water, whereas waters designated for industrial use need not be so pristine.

In choosing the designated uses of particular water bodies and then establishing the associated water quality criteria necessary to achieve and support those uses, states particularize the meaning of the Clean Water Act's general goal of "chemical, physical, and biological integrity"[2] for specific water bodies. In theory, a state could create different water quality standards for every single water body in the

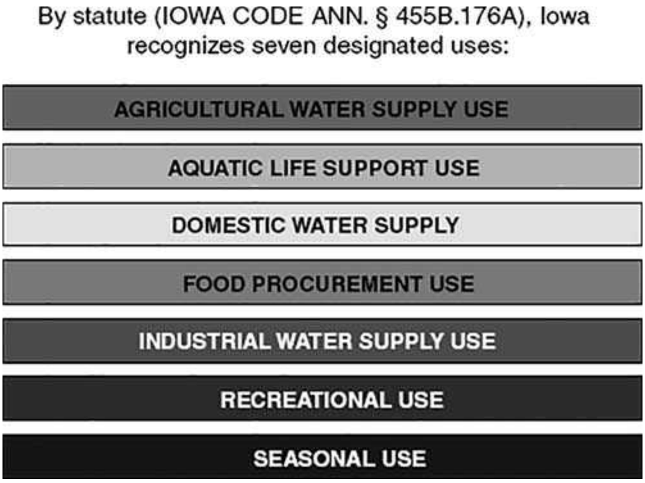


FIGURE 3 An example of Designated Uses: IOWA.

state. In practice, however, states tend to create categories of water bodies and then set water quality standards on that categorical basis.

Although the states received the primary authority to set water quality standards, the EPA retains the authority to set those standards for states that refuse to do so.[16] For example, in 2010, the EPA proposed numeric water quality criteria for nutrients (nitrogen and phosphorus) for the state of Florida.[18] The EPA also has the authority to disapprove state water quality standards that are inconsistent with the Act's requirements.[16]

Besides setting the water quality goals for individual water bodies, water quality standards also directly influence the two point source permitting programs. For example, in NPDES permitting, if the industry-wide technology-based effluent limitations are insufficient to achieve the water quality standards for a particular water body, the state or the EPA should instead use water quality–based effluent limitations in the permits of point sources discharging into that water body[19] (see Figure 2). The Army Corps, in turn, must ensure that projects requiring a Section 404 dredge and fill permit will not violate state water quality standards, and states can ensure that the project will comply with state water quality standards through their Section 401 certification authority.[20–22]

State Management of Nonpoint Sources

State Nonpoint Source Management Plans

The federal permit programs apply only to point sources—that is, "discernible, confined, and discrete conveyances"[4]—of water pollution. In contrast, a nonpoint source of water pollution is any source that is not a point source. Common nonpoint sources are uncontrolled and contaminated runoff or snowmelt or deposition onto water of air pollution.[23]

The Clean Water Act leaves the states in charge of managing nonpoint source water pollution. Initially, states were supposed to address nonpoint source pollution through their Section 208 area-wide waste treatment management programs.[24] These plans emphasized the construction of treatment works.[24] However, they were also supposed to do the following: 1) "identify, if appropriate, agriculturally and silviculturally related nonpoint sources of pollution" and establish controls for them; 2) identify and control "mine related sources of pollution"; 3) identify and control "construction activity related sources of pollution"; and 4) identify and control salt water intrusion problems.[24]

Dissatisfied with states' progress in controlling nonpoint source pollution under Section 208, Congress in 1987 added Section 319 to the Clean Water Act, which encouraged states to enact more specific and detailed nonpoint source management programs.[25] Section 319 nonpoint source management plans are supposed to include BMPs for all relevant nonpoint sources in the state, enforcement and assistance programs, and annual milestones against which to judge the state's progress in reducing nonpoint source pollution.[25]

Role of Water Quality Standards in Nonpoint Source Management before the TMDL Process

The ultimate goal of nonpoint source management, as in the regulatory provisions of the Clean Water Act generally, is to achieve the "chemical, physical, and biological integrity" of the Nation's waters,[2] which, as noted, is defined as a practical matter through the states' water quality standards. Thus, state nonpoint source management programs are often keyed to state water quality standards. Idaho, for example, explicitly requires that the designated BMPs for nonpoint sources ensure compliance with state water quality standards.[26]

As a more elaborate example, Washington explicitly ties its forest practice rules, which prescribe BMPs for forest activities that cause nonpoint source pollution, to its water quality standards, requiring the forest practice rules to be protective of water quality standards in waters affected by forestry-related nonpoint source pollution.[27] Once that connection is made, however, a forestry activity that complies with the forest practice rules is deemed to be in compliance with the state's water quality requirements.[27]

Role of TMDLs in Achieving Water Quality Standards

Introduction

As noted, Congress intended the regulatory mechanisms in the Clean Water Act "to restore and maintain the chemical, physical, and biological integrity of the Nation's waters."[2] The state-set water quality standards define such integrity for particular water bodies, and hence, water quality standards are the measure of whether the Act's goals have been achieved for an individual water body.

The Act's TMDL process provides a means for adjusting both permit requirements and the nonpoint source management requirements for the particular sources that contribute pollution to a water body that cannot achieve its water quality standards through the application of "normal" source-based requirements. The TMDL thus provides the information necessary to "ratchet down" the Clean Water Act's regulatory requirements to meet the water quality needs of specific water bodies.

Section 303 Impaired Waters List

The first step in the Clean Water Act's TMDL process is the state's creation of a Section 303(d) impaired waters list.[17] Specifically, the Act requires each state to identify water bodies or segments within its boundaries for which the effluent limitations "are not stringent enough to implement any water quality standard applicable to such waters."[17] States vary considerably in how many impaired waters they have.

Impaired waters are also known as "water-quality-limited water bodies." It is important to remember, however, that the designations of impairment depend on the following: 1) the exact water quality standards that each state has decided to apply to a particular water body; and 2) how the state decides to segment its water bodies in general. Given these particularities of state law, the number of impaired water bodies is actually a poor basis for attempting to compare actual water quality across state lines. For example, the fact that Pennsylvania has thousands of impaired water bodies may mean that most waters in Pennsylvania have poor water quality—but it could also mean that Pennsylvania has chosen very stringent water quality standards, applied water quality standards to very small segments of rivers and streams, and/or monitors water quality quite comprehensively. On the other hand, the fact that Alaska has identified less than 100 impaired waters might mean that water bodies in Alaska are generally pristine—but it could also mean that Alaska applies its water quality standards to entire rivers and lakes (creating fewer regulatory "water bodies"), has less stringent water quality standards than Pennsylvania, and/or monitors water quality less comprehensively than Pennsylvania.

Once a state identifies water bodies within its borders that do not meet their applicable water quality standards, it must rank those impaired water bodies for the purpose of establishing TMDLs for them.[16] The priority ranking must "tak[e] into account the severity of the pollution and the uses to be made of such waters." [16] States have considerable discretion in ranking their impaired water bodies. At the same time, the state must also "identify those waters or parts thereof within its boundaries for which controls on thermal discharges … are not stringent enough to assure protection and propagation of a balanced indigenous population of shellfish, fish, and wildlife."[16] States submit their lists of impaired waters to the EPA biennially,[28] and these lists—known as the Section 303(d) lists—are subject to the EPA's approval.[16]

Setting the TMDL and State Water Planning

Once a state has identified water-quality-impaired water bodies within its borders, establishing a TMDL is mandatory, and the EPA must establish the required TMDLs if the state fails to do so.[16] In establishing a TMDL, the state estimates, "in accordance with the priority ranking, the total maximum daily load" of all pollutants contributing to the violation of the relevant water quality standards that the EPA has determined can be subject to such a numeric calculation.[16] The state must establish the TMDL "at a level necessary to implement the applicable water quality standards with seasonal variations and a margin of safety which takes into account any lack of knowledge concerning the relationship between effluent limitations and water quality."[16] Thus, the TMDL should be somewhat conservative, allowing for both scientific uncertainty and natural variations in water quality.

Under the EPA's regulations, "TMDLs may be established using a pollutant-by-pollutant or biomonitoring approach. In many cases both techniques may be needed. Site-specific information should be used wherever possible."[28] Moreover, the state must establish TMDLs for "all pollutants preventing or expected to prevent attainment of water quality standards."[28]

Similarly, for thermally impaired waters, the state must estimate "the total maximum daily thermal load required to assure protection and propagation of a balanced, indigenous population of shellfish, fish, and wildlife."[16] Statutory factors for setting a thermal TMDL include "normal water temperatures, flow rates, seasonal variations, existing sources of heat input, and the dissipative capacity of the identified waters or parts thereof."[15] Moreover, like other TMDLs, thermal TMDLs are set conservatively, because they must include a margin of safety to account for scientific uncertainty.[16]

Once a state has established its TMDLs, it submits them to the EPA. The EPA must approve or disapprove the TMDLs within 30 days of submission.[16] If the EPA disapproves of the TMDLs, it has 30 days to establish its own TMDLs.[16]

For each TMDL established, the state must allocate the total daily load of the relevant pollutant among the sources that contribute that pollutant to the impaired water. The EPA recognizes that the TMDL's total pollutant load must be allocated among three sources (see Figure 4): background or "natural" sources of the pollutant; nonpoint sources of the pollutant (the "load allocation," or LA); and point sources of the pollutant (the "waste load allocation," or WLA).[29]

These allocations then provide the basis for strengthening the regulation of the relevant sources so that the water body can achieve its water quality standards. For example, the WLA provides the total amount of the relevant pollutant that the point sources can discharge. As a result, it becomes the basis for calculating water quality–based effluent limitations that should then be incorporated into the point sources' NPDES permits.[29] Moreover, the Clean Water Act limits the states' abilities to amend these

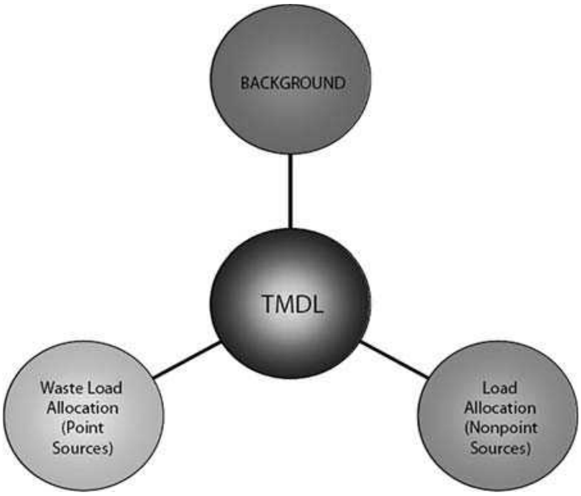


FIGURE 4 Allocation of the TMDL.

TMDL-based effluent limitations so long as the water body remains impaired: such standards "may be revised only if (i) the cumulative effect of all such revised effluent limitations based on such total maximum daily load or waste load allocation will assure the attainment of such water quality standard, or (ii) the designated use which is not being attained is removed. …"[16] Removing designated uses is difficult, however, because the Act incorporates an antidegradation policy.[16,30]

The LA, in turn, allows for the possibility that the state will also amend its nonpoint source management requirements—especially for waters impaired entirely or predominantly as a result of nonpoint source pollution. As the EPA recognizes in its regulations, "[i]f Best Management Practices (BMPs) or other nonpoint source pollution controls make more stringent load allocations practicable, then wasteload allocations can be made less stringent. Thus, the TMDL process provides for nonpoint source control tradeoffs." [29]

If the EPA approves the state's TMDLs, or if the EPA establishes TMDLs for the state, the state also incorporates those TMDLs into its continuing planning process for water quality.[16] The state's continuing planning process includes the following: 1) effluent limitations and schedules of compliance, including effluent limitations stringent enough to meet water quality standards; 2) the applicable Section 208 area-wide waste management plans; 3) the TMDLs; 4) procedures for revision; 5) authority for intergovernmental cooperation; 6) mechanisms to implement new and revised water quality standards, including schedules of compliance; 7) controls for the disposal of residual wastes from wastewater treatment; and 8) an inventory and priority ranking of treatment works.[16]

Two Examples: Applying a TMDL

The potential impacts of a TMDL, and the potential tradeoffs between point source regulation and nonpoint source regulation, are best illustrated by example (see Table 1). Suppose that, in its Section 303 water quality standards, State A has designated both the Blue River and the Green River as cold-water rivers to support native trout populations. To support this designated use, State A establishes water quality criteria for sediment. Moreover, it includes standard effluent limitations for sediment in the NPDES permits that it issues to point sources. State A's nonpoint source management program includes extensive BMPs for farming, which has been a traditional source of water pollution in the state. However, forestry is a relatively new industry in State A, and the state has not amended its nonpoint source management program to account for water pollution from those activities.

TABLE 1 Sources of Sediment in the Blue River and Green River in State A

Source	Blue River	Green River
Background/natural	5 kg/day from cliff erosion	10 kg/day from upstream erosion
Point sources	500 kg/day from 4 confined animal feeding operations, each of which has an NPDES permit	None
Nonpoint sources	10 kg/day from runoff over a small farm	600 kg/day from 3 large timber operations along the river

During its most recent biennial Section 303(d) review of impaired waters, State A discovers that both the Blue River and the Green River are violating their water quality standards for sediment, causing harm to the native trout populations. Through the TMDL process, State A estimates that 100 kg of sediment can be added to each river each day without violating the sediment water quality standard for trout, leaving an ample margin of safety and allowing for seasonal variations, such as increased sediment runoff during the rainy season.

State A then identifies the sources contributing sediment to each river (Table 1).

Given these sources, State A should take a different approach to implementing the sediment TMDL on the Blue River than it does to implementing the sediment TMDL on the Green River. To ensure that the Blue River will meet its water quality standards, State A will need to adjust the point sources' *WLA*. As a result, to ensure compliance with the Blue River's water quality standards, State A will need to amend each of the point sources' NPDES permits to incorporate more stringent water quality–based effluent limitations for sediment.

In contrast, to ensure that the Green River will meet its water quality standards, State A will have to address the nonpoint sources' *LA*. Most obviously, State A should amend its nonpoint source management plan to impose nonpoint source control requirements, such as BMPs, on the timber companies. Such BMPs may, for example, require the use of buffer zones, require selective logging instead of clear-cutting, or limit the amount of road building and land clearing allowed.

Conclusion

Congress designed the TMDL process to ensure that every "water of the United States" would eventually enjoy chemical, physical, and biological integrity, as defined by the states in water quality standards. While TMDLs have been part of the Act since 1972, states ignored the TMDL process for many years.[31] Instead, citizen litigation drove the establishment and implementation of TMDLs, and many states are still working their way through court orders that require them to set TMDLs.[31]

Given that litigation context, TMDL implementation focused initially on adjusting the effluent limitations in point sources' Clean Water Act permits. However, as it becomes clear that nonpoint source pollution is the most significant remaining source of water quality impairment, TMDLs are increasingly becoming the mechanism that encourages states to adequately control those nonpoint sources of water pollution, as well.

References

1. Craig, R.K. *The Clean Water Act and the Constitution,* 2nd Ed.; Environmental Law Institute Press: Washington, DC, 2009.
2. Federal Water Pollution Control Act, § 101, 33 U.S.C. § 1251 (2006).
3. Craig, R.K. Climate Change, Regulatory Fragmentation, and Water Triage. University of Colorado Law Review **2008**, *79*(3), 825–927.
4. Federal Water Pollution Control Act, § 502, 33 U.S.C. § 1362; 2006.
5. Federal Water Pollution Control Act, § 301, 33 U.S.C. § 1311; 2006.

6. Solid Waste Agency of Northern Cook County v. U.S. Army Corps of Engineers, 531 U.S. 159; 2001.
7. Rapanos v. United States, 547 U.S. 715 (2006).
8. Federal Water Pollution Control Act, § 402, 33 U.S.C. § 1342; 2006.
9. Federal Water Pollution Control Act, § 404, 33 U.S.C. § 1344; 2006.
10. Office of Water, U.S. Environmental Protection Agency. *National Pollutant Discharge Elimination System (NPDES),* available at http://cfpub.epa.gov/npdes/statestats.cfm. (Accessed May 11, 2012).
11. Champion International Corp. v. U.S. Environmental Protection Agency, 850 F.2d 182 (4th Cir. 1988).
12. Office of Water, U.S. Environmental Protection Agency. *State or Tribal Assumption of the Section 404 Permit Program,* available at http://water.epa.gov/type/wetlands/outreach/fact23.cfm (accessed May 2012).
13. 33 C.F.R. § 323.2; 2009.
14. 40 C.F.R. §§ 230.1–230.80; 2009.
15. 33 C.F.R. §§ 320.1, 320.4; 2009.
16. Federal Water Pollution Control Act, § 303, 33 U.S.C. § 1313; 2006.
17. Federal Water Pollution Control Act, § 304, 33 U.S.C. § 1314; 2006.
18. U.S. Environmental Protection Agency. *Water Quality Standards for the State of Florida's Lakes and Flowing Rivers,* 75 Fed. Reg. 4174; (Jan. 26, 2010).
19. Federal Water Pollution Control Act, § 302, 33 U.S.C. § 1312; 2006.
20. Federal Water Pollution Control Act, § 401, 33 U.S.C. § 1341; 2006.
21. PUD. No. 1 of Jefferson County v. Washington Department of Ecology, 511 U.S. 700; 1994.
22. S.D. Warren Co. v. Maine Board of Environmental Protection, 547 U.S. 370; 2006.
23. Office of Water, U.S. Environmental Protection Agency. *What Is Nonpoint Source Pollution?,* available at http://water.epa.gov/polwaste/nps/whatis.cfm (accessed May 2012).
24. Federal Water Pollution Control Act, § 208, 33 U.S.C. § 1288; 2006.
25. Federal Water Pollution Control Act, § 319, 33 U.S.C. § 1329; 2006.
26. Idaho Code § 39–3620; 2009.
27. Revised Code of Washington § 90.48.420; 2009.
28. 40 C.F.R. § 130.7; 2009.
29. 40 C.F.R. § 130.2; 2009.
30. 40 C.F.R. § 131.12; 2009.
31. Houck, O.A. *The Clean Water Act TMDL Program: Law, Policy, and Implementation;* Environmental Law Institute Press: Washington, DC, 1999.
32. U.S. EPA, *NPDES Permit Document: Components,* available at http://www.epa.gov/waterscience/standards/academy/supp/permit/page6.htm.

48

Watershed Management: Remote Sensing and GIS

A.V. Shanwal
and S.P. Singh

Introduction

The need for natural resource conservation must be considered in any agricultural development plans involving conversion of new land use to increase production. An action plan to minimize natural resources degradation must be based on the principles of sustainability so that soil can be handed over to the future generation under better conditions than received from the previous generation. Thus, management practices must be ecologically sound, economically feasible, and socially and politically acceptable.

Natural resources degradation can be contained by adopting the watershed as a hydrological unit for development and management. This approach is multidisciplinary, broad based, and intensive vis-a-vis the simple “Seed and Fertilizer Approach.” Despite the investment, it is an economic approach in the long run.[1]

The degradation of improperly used watersheds happens because of increase in soil erosion, decline in soil productivity, reduction in livestock-carrying capacity, decline in forest cover and perturbation in ecological equilibrium, and reduction in biodiversity. The problem requires a hightech solution with back-up policy decisions.

Remote sensing (RS) and geographic information system (GIS) are important technologies for addressing challenges in sustainable management of natural resources. Remotely sensed data (e.g., aerial photographs and satellite imagery) can be used to obtain information on soils, land use, vegetation, slope gradient, runoff, erosion, etc. However, recent developments in multispectral scanner, radar system, and a multitude of quantitative techniques for analyzing and processing such data provide opportunities for data acquisition through RS and array of techniques for data analyses.

GIS is an important tool for tracking spatial data. GIS draws the composite map by superimposing the data and image files obtained from traditional methods and satellite imageries, allows us to develop, analyze, and display spatially explicit information, and gives us the ability to deal with the larger spatial scales in process such as soil erosion and drainage (500 acres) and regional landscape (several million acres).

Inventorying and Monitoring

Despite efforts at inventorying and monitoring, conventional techniques provide only sketchy information on resources in a watershed, their location, and spatial and dynamic distribution. Of inventorying and monitoring, the latter is probably the major new thrust for resource managers and scientists.

Improved management practices are not thought of spontaneously prior to testing and implementation. Such practices have evolved over time through experimentation, experience, and trial and error. A case in point is the Bunga watershed in Ambala, Haryana, India.[10] As was expected, numerous components did not work properly. Neither the rate of runoff nor the outflow from the reservoir was in harmony in the initial stage. High rate of siltation of Sukhana Lake near Chandigarh, Haryana, India, due to severe soil erosion in the catchment area is another example of dire need for inventorying of data prior to starting any project.[11] Because of high siltation, the ponded area of Sukhana Lake decreased by about 30ha between 1980 and 2000. During this period, the courses of Sukhana and Kansal streams feeding the lake also changed considerably. The siltation of Sukhana Lake is the result of poor vegetation cover and heavy runoff in the catchment within the Shiwalik Hills.[11]

It is thus important that RS and GIS specialist and the resource managers must work together to determine the specific information needed (e.g., the nature of earth surface cover and their characteristics; the location, size, terrain, and other characteristics of the watershed area involved). Relevant information needed includes the format (e.g., maps, tables, scales, etc.); time frame for both collecting and processing the data; level of accuracy and reliability; costs of obtaining and interpreting/processing the data.

Research and Development

Using watershed, landscape or ecosystem approaches have broad support as a means in achieving sustainable use of natural resources and integrating objectives on a practical scale. Yet, addressing watershed management issues on large scales requires enhanced technical capabilities and modern set of tools and facilities.

Landsat-1 was launched in 1972 and the Landsat Thematic Mapper (TM) type of data has been in existence only since 1984. Thus, a tremendous progress has been made over short time in developing effective methods of processing and analyzing such data. The last decade of 20th century witnessed rapid advances and significant increase in the operational use of remotely sensed data.[3]

Landsat-7 launched in 1998 carries Enhanced Thematic Mappers (ETM) and pointable sensor with improved spatial resolution and signal-to-noise ratio (Table 1). The Earth Observing System consists of a morning (AM) and an afternoon (PM) components and carries five separate sensors including MODIS, a 36-narrow band imager with 1-km spatial resolution.[2] These sensors are significantly improved in their capability to map watershed information classes. These advanced sensors and computational capability have new requirements for satellite data processing algorithms. Added to the image analysis system, there is also a need for RS algorithms for estimating biophysical parameters necessary to drive and validate the watershed process models.[5]

ETM: enhanced thematic mapper; HRVIR: high resolution visible and middle infrared; LISS: linear imaging self scanner system; SAR: synthetic aperture radar; VIRS: visible infrared scanner; TMI: TRMM microwave imager; CERES: clouds earth's radiation energy system; AVHRR: advanced very high resolution radiometer; OCTS: ocean color and temperature scanner; AVNIR: advanced visible and near infrared radiometer; ASTER: advanced space borne thermal emission and reflectance radiometer; MODIS: moderate resolution imaging spectrometer-nadir; PAN: panchromatic; MS: multi spectral; IFOV: instantaneous field of view.[5]

During the 1980s, advances in computer hardware, particularly speed and data storage, catalyzed the development of software for handling spatial data. One of the most significant products of this period of rapid technological change was the development of GIS (Table 2). It has made a tremendous impact in identifying strategies of watershed development by manipulating and analyzing individual "layers"

TABLE 1 Recent RS Satellites with Advanced Sensors

		Sensor		
Satellite	Country	Name	Spectrum	IFOV
Landsat-7	U.S.A.	ETM⁺	0.45 µm ~ 12.50 µm (8 bands)	30 m
SPOT-4	France	HRVIR	0.50 µm ~ 1.75 µm (5 bands)	10 m
IRS-1D	India	PAN	0.50 ~ 0.75	5.8 m
		LISS-III	0.45 ~ 0.86 1.55 ~ 1.70 (5 bands)	24.0 m
		WIFS	0.67 ~ 0.86	188 m
RADARSAT	Canada	SAR	5.3 GHz	25 m × 28 m
TRMN	Japan	VIRS	0.63 ~ 10.7 (5 bands)	2km
		TMI	10.65 GHz ~ 85.5 GHz	5.6 km × 3.8 km
		CERES	0.3 µm ~ 50 µm (3 bands)	25 km
		LIS	0.7774 nm	5 km
NOAA-M	U.S.A.	AVHRR/3	0.58 µm ~ 12.40 µm (6 bands)	0.5 km
CRSS	U.S.A.	PAN	0.45 µm ~ 0.90 µm	0.82 m
		MSS	0.45 µm ~ 0.90 µm (4 bands)	3.20 m
ADEOS	Japan	OCTS	0.402 µm ~ 12.5 µm (12 bands)	700 m
		AVNIR	0.40 µm ~ 0.92 µm (5 bands)	8 m
EOS-AM	U.S.A.	ASTER	0.52 µm ~ 11.3 µm (3 bands)	15 m
EOS-PM	U.S.A.	MODIS-N	0.659 µm ~ 14.24 µm (5 bands)	250 m
EOSAT	U.S.A.	PAN	0.45 µm ~ 0.90 µm	1 m
		MS	0.45 µm ~ 0.90 µm (4 bands)	4 m

Note: ETM: enhanced thematic mapper; HRVIR: high resolution visible and middle infrared; LISS: linear imaging self scanner system; SAR: synthetic aperture radar; VIRS: visible infrared scanner; TMI: TRMM microwave imager; CERES: clouds earth's radiation energy system; AVHRR: advanced very high resolution radiometer; OCTS: ocean color and temperature scanner; AVNIR: advanced visible and near infrared radiometer; ASTER: advanced space borne thermal emission and reflectance radiometer; MODIS: moderate resolution imaging spectrometer-nadir; PAN: panchromatic; MS multi spectral; IFOV: instantaneous field of view.
Source: JSRS Remote Sensing Note.[5]

of spatial data and providing tools for analyzing and modeling the interrelationship among layers. The GIS also provides a means of displaying complex watershed information in a comprehensible manner.

The GIS also provides a means of predicting the outcomes of alternative courses of action, from both spatial and temporal perspective and in a timely and cost-effective manner. However, it does not preclude the need for monitoring the ground truth as a guide to development of future management practices.

The last decade of the 20th century witnessed rapid advances and a significant increase in the operational use of RS and GIS in watershed management.[8] Much of this increase in operational use is due to the continued integration of RS, GIS, GPS (global positioning system), and Crop Model (CM) techniques.[4] These four technologies form a powerful, interrelated combination (Figure 1). In watershed management, there has been a continuous increase in the use of RS data to provide input to new GIS database, upgrade existing database, and monitor land use changes.

TABLE 2 Different Types of GIS

		CPU		Data Model			Applications	
GIS	Origin	PC	WS	Vector	Raster	Analysis	DTM	Network
ARC=INFO	ESRI	ˇ	ˇ	ˇ	ˇ	ˇ	ˇ	ˇ
MGE	Intergraph		ˇ	ˇ	ˇ	ˇ	ˇ	ˇ
Geo=SQL	Generation 5 Technology	ˇ	ˇ	ˇ	ˇ	ˇ	ˇ	ˇ
GFIS	IBM	ˇ	ˇ	ˇ	ˇ	ˇ	ˇ	ˇ
IDRISI	Clark University, U.S.A.	ˇ	ˇ	ˇ	ˇ	ˇ	ˇ	ˇ
GRASS	GRASS Information Center	ˇ	ˇ	ˇ	ˇ	ˇ	ˇ	ˇ
ERDAS	ISRI	ˇ	ˇ	ˇ	ˇ	ˇ	ˇ	ˇ
GRAMM^{++}	IIT Bombay, India	ˇ	ˇ	ˇ	ˇ	ˇ	ˇ	ˇ

Source: Maguire[6] and Morehouse.[7]

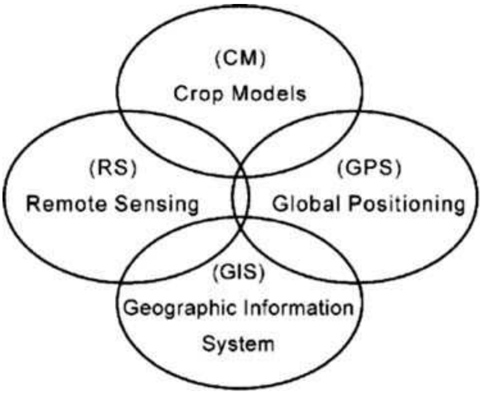


FIGURE 1 Relationship between RS, GIS, GPS, and CM.

In addition to providing input to GIS database, the RS data enables significant improvements in classification accuracies. Further, GPS capabilities provide effective cartographic control to the GIS database, and will enable field plots to be located efficiently and accurately in the data set.

The integration of GIS and CMs has proven very successful for alternative land use planning in the watershed. The important ones are DSSAT (Decision Support System for Agro-technology Transfer), AEGIS/WIN (Agricultural and Environmental Information System for Windows), IDSS (Intelligent Decision Support System), and DSSE (Decision Support System Engine). These simulation models are capable of predicting the potential yields of crops (rice, cassava, potato, sugarcane, sunflower, maize, wheat, barley, millet, sorghum, soybean, peanut, drybeans, tomato, and chickpea) in different physiographic units of a particular watershed. In addition, several GIS-based simulation models have been developed for natural resources management and are used in watershed management.[9] The important ones are SNAP (Scheduling and Network Analysis Program), LUCAS (Land Use Analysis System), and LANDSIM and LANDIS (Land Information System). All these models involve database (nonspatial), GIS, model base (simulation analysis), and a GUI (Graphic User Interface).

Challenges and Opportunities Ahead

Rapid advances in RS and GIS technologies can bring about quantum leap in watershed management. Yet the challenge lies in assuring that RS and GIS technologies continue to serve the practitioners and users, but not vice versa.

However, GIS and RS are both a panacea and a Pandora's box. These are panacea because of the promise to meet the challenges of resource inventory and monitoring, and planning and policy analysis. These are Pandora's box because of the numerous pitfalls of using the tools wrongly, capturing the data poorly, miscommunicating information, conveying incorrect results, and overselling the capabilities.

The dilemma can be addressed by a careful integration of RS, GIS, and other simulation models in watershed management and with due consideration of the following:

- Designing the classification system as totally exhaustive, mutually exclusive, and hierarchical
- Determining the temporal and spatial scale of the watershed by accommodating GIS and RS at multiple scales without compromising flexibility and quality of a project
- Identifying the appropriate data sources (video, aerial photography, satellite imagery, airborne scanner, etc.), for different land cover
- Assessing and reporting the accuracy of the data needs
- Limiting the scope of the project according to the budget and schedule
- Standardizing the formats needed for the exchange of information across projects and eliminating duplication

Future changes may include the following:

1. Developing effective techniques of integration and analysis of data from various sources such as AVHRR and Landsat TM or SPOT data or Landsat TM plus satellite radar data.
2. Converting research into operational applications in watershed management.
3. Developing effective expert system to assist the analyst.
4. Educating and training the user community in principles and theory of these technologies so that they can use these powerful tools wisely, appropriately, and effectively.

Anticipated developments and opportunities in these technologies comprise the following:

- Economic availability of satellite optical sensor data with improved spatial, spectral, radiometric, and temporal resolutions
- Availability of operational multifrequency, multipolarization synthetic aperture radar data from satellite altitudes
- Improvement in computer storage and processing capabilities
- Better understanding and use of combined data from optical, microwave, and other remote sensors
- Integration of RS, GIS, GPS, and CM technologies
- Increased use of expert systems for data analysis

These developments will improve the quality and characteristics of the data and analytical capabilities in the area of watershed management. A combination of knowledgeable resource managers and practical farmers with the technological tools and data available to them holds great promise for identifying effective watershed management technologies.

Conclusions

A variety of information about the characteristics and condition of the area is needed for judicious management of watershed. Aerial photography has been used since the 1950s for obtaining information on soils, land use, vegetation, slope, runoff, erosion, etc. The advent of space research through satellite, multispectral scanner, and radar data, and quantitative analytical techniques of processing such data has increased the array of data, analysis procedure, and results that can be obtained using RS capabilities.

During the 1980s, advances in computer hardware for processing and data storage catalyzed the development of software for handling spatial and image digital data. These technologies played an important role in the development of GIS for natural resource management, especially in preparing composite map superimposing the data and image files.

During the 1990s, significant development in GIS technology and integration of RS, GIS, GPS, and Crop Simulation Model techniques have created additional complexity and opportunities of using various data sources and analysis techniques for obtaining the information needed by the resource managers. Yet, there are some critical issues that need to be addressed for integration of RS, GIS, and other technologies to obtain information needed for sustainable natural resource management through watershed development.

References

1. Dhruva Narayana, V.V.; Sastry, G.; Patnaik, U.S. *Watershed Management*; Indian Council of Agricultural Research: New Delhi, 1990.
2. Hall, G.F. Adaptation of NASA remote sensing technology for regional-level analysis of forested ecosystem. In *Remote Sensing and GIS in Ecosystem and Management*; Sample, V.A., Ed.; Island Press: Washington, D.C., 1994; 287–314.
3. Hoffer, M.R. Challenges in development and applying remote sensing to ecosystem management. In *Remote Sensing and GIS in Ecosystem Management*; Sample, V.A., Ed.; Island Press: Washington, D.C., 1994; 25–40.
4. Hoogenboom, G.; Wilkins, P.W.; Tsuji, G.Y. *Discussion Support System for Agrotechnology Transfer*; University of Hawaii: Honolulu, Hawaii, 1999; Vol. 4.
5. JSRS Remote Sensing Note. Japan Association on Remote Sensing, University of Tokyo: Japan, 1998.
6. Maguire, D.J. The Raster GIS design model—a profile of ERDAS. Comput. Geosci. **1992**, *18* (4), 463–470.
7. Morehouse, S. The ARC/INFO geographic information system. Comput. Geosci. **1992**, *18* (4), 435–441.
8. Murai, S. *GIS Work Book;* Japan Association of Surveyors: Tokyo, Japan, 1996.
9. Sample, V.A. *Remote Sensing and GIS in Ecosystem and Management;* Island Press: Washington, D.C., 1994.
10. Singh, G.; Mittal, S.P.; Dhruva Narayana, V.V.; Agnihotri, R.C. *Watershed Management Plan (Vil. Bunga);* Central Soil and Water Conservation Research and Training Institute: Dehra Dun, 1983.
11. Sohal, H.S. Soil and water conservation measures for sediment control in Sukhana Lake, Chandigarh. In *Fifty Years of Research and Sustainable Resource Management in Shiwalik;* Mittal, S.P., Aggarwal, R.K., Samra, J.S., Eds.; Central Soil and Water Conservation Research and Training Institute Research Center: Chandigarh, India, 2000; 259–266.

49

Wetlands: Conservation Policy

Clayton Rubec

Introduction

Meeting the challenge of conserving wetlands requires comprehensive national policies to provide a foundation for domestic action and a framework for international and national cooperation. Such policy is valuable as countries seek to address the management and habitat requirements for wildlife and other natural resources, such as soil and water, as well as human needs. Implementation of a national wetland policy is a key feature of the wise use principles of the Convention on Wetlands of International Importance (the Ramsar convention). However, such policy remains an elusive goal for many of the 145 nations that today (May 2005) are contracting parties to this global environmental treaty.

Responding to recommendations by the Convention, in 1998–1999 the author led a team of writers in preparing guidelines for developing and implementing national wetland policies.[1] These Policy Guidelines complement the convention's guidance on wetland legislation.[2] The following entry provides highlights of the guidelines and reports on the status of wetland policy development around the world, effective mid-1999.

Relationship between Policy and Wise Use

The wise use principles are a hallmark of the Ramsar convention. "Wise use" applies not only to sites listed as wetlands of international importance (as of May 2005 covering over 125 million hectares at 1429 sites), but also to all wetlands in the territory of contracting parties. These *principles* help contracting parties improve institutional and organizational arrangements, address legislative and policy needs, increase knowledge and awareness of wetland values, inventory and monitor the status of wetlands, identify program priorities, and develop action plans for specific sites as components of a national wetland policy.

The formulation of national wetland policies sometimes involves a lengthy and complex process. Political, jurisdictional, institutional legal, and financial constraints affect policy formulation in addition to social and economic factors that continue to contribute to wetland loss while the policy process is under way.

What Are Wetland Policies, and Are They Needed?

A policy can simply be a document. However, making is also a process involving consensus building, encapsulation of ideas and commitments, implementation, accountability, and review complemented by legislation, strategies, and operational programs. It is a mechanism for an administration to capture the public will or mandate on an issue, and refine it with its own vision.

A national wetland policy is nationwide in scope but it may be developed at several levels of government. In Australia and Canada, for example, both the federal government and state/provincial governments have developed wetland conservation policies. This reflects the federal nature of these two nations, wherein constitutional authority for natural resources management (including wetlands) is divided between the levels of government.

Wetlands are seldom explicitly covered in other natural resource management policies such as for water, soil, forest, land, or agriculture at a national level. Development of a "stand-alone" wetland policy and/or strategy can be an important step in recognizing and solving wetland problems. A wetland policy recognizes wetlands as ecosystems requiring different approaches not masked under other sectoral management objectives. Articulation of goals and objectives for these ecosystems identifies clear responsibilities of the government and a public expectation that the government will deliver these commitments.

Wetland policy objectives need to focus on a variety of themes as they become the image of the policy. However, practical implementation of the Policy may result in only one or two of these objectives receiving the greatest public attention. For example, Canada's announcement of its federal wetland policy in 1992 contained seven objectives but "no net loss of wetland functions" has proven to be its catch phrase.

Guidelines

The Guidelines review the key steps and issues that may arise in both developing and implementing a National Wetland Policy. These include over 20 detailed sections defining the purpose of such an initiative, organizing a suitable process, deciding how to present the content of the policy document, and developing strategies for implementation and monitoring. The text is complemented by seven wetland policy essays: 1) *defining stakeholders',* 2) *consultations;* 3) *wetland policies within a federal state;* 4) *sectoral policies and legislation;* 5) *compliance strategies;* 6) *role of nongovernment organizations;* and the 7) *development and coordination process.* The guidelines provide a reference against which all nations can review their wetland action plans and strategies at the national level.

A team of contributors with governmental or nongovernmental work experience and expertise in wetland policy development prepared the guidelines. The team included writers from Ramsar national authorities in Australia, Canada, Trinidad and Tobago, Uganda, and the United States of America. Contributors from Bird Life International, University of Massachusetts, IUCN Environmental Law Centre, and Wetlands International were also involved.

Implementation Strategies

A National Wetland Policy includes specific implementation strategies that demonstrate the priorities of the government, but also fosters the cooperation and involvement of other interests. Linkages between these strategies and national water, soil, biodiversity, and sustainable development policy initiatives are explored in the guidelines.

An analysis of the strategies used in selected National Wetland Policies is summarized in Table 1. These include the policies/action plans of Australia, Cambodia, Canada, Columbia, Costa Rica, Finland, France, Jamaica, Malaysia, Peru, Trinidad and Tobago, and Uganda. These initiatives have many common strategic approaches including: a) ensuring public awareness and education; b) developing

TABLE 1 Implementation Strategies in Selected National Wetland Policies

Country	1	2	3	4	5	6	7	8	9	10	11	12
Australia	X	X	X	X	X	X						
Cambodia		X	X		X	X		X		X		X
Canada	X		X	X	X	X	X					
Colombia		X	X		X	X		X	X		X	X
Costa Rica	X		X		X			X	X	X		X
Finland	X	X		X	X		X				X	X
France		X	X		X						X	X
Jamaica		X					X	X	X			
Malaysia		X	X		X	X	X	X	X	X		X
Peru			X		X	X		X	X			X
Trinidad and Tobago	X		X	X	X		X	X				
Uganda		X	X		X	X		X	X		X	X

Policy strategies:

1. Management of national wetland networks
2. Integration with other policies such as water, soil and forests
3. Public awareness and education
4. Partnerships
5. Science, monitoring, assessment, and research
6. International commitments
7. Managing special sites
8. Administration and institutions, capacity building
9. Enforcement, regulation, and legislation
10. Financial mechanisms
11. Restoration of degraded sites
12. Sustainable use and conservation

cooperation and partnerships between levels of government from national to local; c) developing and supporting legislation and interrelated land and water use policies and programs; d) implementing wetland site management responsibilities; e) developing a sound basis for the policy through scientific research and expertise; f) developing institutional and financial capacity for policy implementation; and g) meeting international commitments. These strategies have been drafted to evoke a clear vision and acceptance across the nation.

Global Review of the Status of Wetland Policies and Strategies

Significant progress is evident globally in the development of National Wetland Policies since the Ramsar Convention focused attention on this issue in 1987.[1,3,4] Meetings of the contracting parties every three years allow regular review of the status of wetland policies. The Contracting Parties last met in November 2002 when there were 123 Contracting Parties while now (May 2005) there are 145 Contracting Parties.

Policies

As of November 2002, 60 (56%) of the 107 Ramsar Contracting Parties that submitted national reports to the Eighth Meeting of the Contracting Parties to the Convention on wetlands indicated that they were engaged in development or implementation of a National Wetland Policy. Between 1987 and 2002, the number of nations with a National Wetland Policy officially adopted grew from 0 to 41. An additional 19 nations indicated that a National Wetland Policy was in draft or under consideration. However, 47 (44%)

TABLE 2 Evolution of Ramsar Convention on National Wetland Policies/Strategies[a]

Status of National Wetland Policies or Strategies	1987 Regina COP3	1990 Montreux COP4	1993 Kushiro COP5	1996 Brisbane COP6	1999 San Jose COP7	2002 Valencia COP8
National wetland policies Policy/strategy adopted	0	0	3	6	12	41
Policy/strategy in draft or under consideration	2	5	12	21	30	19
No policy/strategy activity reported	15	40	36	65	72	47
Number of national reports tabled at this COP	17	45	51	92	98	107

[a] As of May 2005, the Ramsar Convention had 145 Contracting Parties and will next report on this issue in November 2005.

TABLE 3 Summary of National Wetland Policies and Strategies by Ramsar Regions (2002)

Ramsar Region	National Wetland Policy or Strategy Adopted by Government	National Wetland Policy or Strategy in Preparation or under Consideration
Africa	Congo Rep., Côte d'Ivoire, Senegal, South Africa, Uganda	Algeria, Botswana, Burkina Faso, Chad, Comoros, Dem. Rep. Congo, Egypt, The Gambia, Ghana, Guinea, Kenya, Malawi, Mali, Morocco, Namibia, Niger, Togo, Zambia
Asia	Indonesia, Japan, Thailand, Vietnam	Bangladesh, Cambodia, P.R. China, Georgia, India, Rep. Korea, Malaysia, Mongolia, Philippines, Russia,[a] Turkey
Europe	Belgium, Bulgaria, Denmark, Estonia, Finland, France, Greece, Iceland, Malta, Monaco, Netherlands, Norway, Romania, Sweden, Switzerland, United Kingdom	Austria, Belarus, Croatia, Czech Rep., Germany, Hungary, Ireland, Italy, Latvia, Lithuania, Poland, Portugal, Russia,[a] Slovak Rep., Slovenia, Ukraine, F.R. Yugoslavia
Neotropics (Central and	Colombia, Costa Rica, Jamaica,	Argentina, Bahamas, Chile, Ecuador, Guatemala,
South America and Caribbean)	Peru, Trinidad and Tobago	Honduras, Nicaragua, Panama, Paraguay, Venezuela
North America	Canada, United States	Mexico
Oceania (Australia and Pacific)	Australia, New Zealand	—

[a] Russia straddles Asia and Europe.

of the Ramsar Contracting Parties did not yet report any actions being taken in support of National Wetland Policy development. A number of nations, particularly those with a commonwealth or federal make-up, reported wetland policies and strategies at the sub-national level also. National Wetland Policies have also been called a "National Wetland Strategy" or "Action Plan."

Table 2 summarizes the status of the development and adoption of National Wetland Policies from 1987 through November 2002. This table was developed by reviewing the reports and conference papers that summarize Convention activities every three years by each country. In 1987, only two nations indicated that they were involved in developing any sort of national wetland policy/strategy. By 2002, this number grew to at least 60 nations. Table 3 summarizes the countries by Ramsar's regions (Africa, Asia, Europe, Neotropics, North America, and Oceania) that have either a *national wetland policy* or *national wetland strategy* adopted, being drafted or considered up to May 2005.

Conclusions

Development and implementation of a national wetland policy in the 145 countries that have acceded to the Ramsar Convention is proceeding throughout the World. The Convention's Wise Use Principles and Guidelines on National Wetland Policy,[1] complemented by guidelines for wetland legislation,[2] are effective tools in fostering the completion and use of National Wetland Policies and Strategies as important cornerstones of this Convention. At least 56% of the Ramsar Convention's Contracting Parties now are implementing or developing national wetland policies/strategies.

One of the most interesting aspects of the Ramsar Convention is its capacity to foster sharing of experience. Interchanges by wetland policy experts are now occurring internationally that involve short-term invited visits or sabbaticals, informal exchange of documents, confidential advice, and review of draft policy. Consultation workshops, working with non-government groups, meeting with senior government officials, exploring funding mechanisms, and assistance in drafting of text have also been involved. Experience gained in one nation's development of these policies can be shared and local expertise enhanced, filling a need among the Ramsar family.

References

1. Rubec, C.D.A.; Pritchard, D.; Mafabi, P.; Nathai-Gyan, N.; Phillips, B.; Mahy, M.; Lynch-Stewart, P.; Chew, R.; Cintron, G.; Larson, J.; Ramakrishna, S. *Guidelines for Developing and Implementing National Wetland Policies;* Report to COP7 Ramsar Convention. Handbook No. 2; Ramsar Bureau: Gland, Switzerland, 2000.
2. Shine, C.; Glowka, L. *Reviewing Laws and Insti.tutions to Promote the Conservation and Wise Use of Wetlands*; IUCN Environmental Law Centre Report to COP7 Ramsar Convention. Handbook No. 3; Ramsar Bureau: Gland, Switzerland, 2000.
3. Rubec, C.D.A. Status of national wetland policy development in ramsar nations. Proceedings, Sixth Meeting of the Conference of the Contracting Parties. Convention on Wetlands, Brisbane, Australia, 1996; Vol. 10/12A: 22–29.
4. Ramsar convention secretariat, National Wetland Policies in Ramsar Contracting Parties. Information Report to Eighth Meeting of the Contracting Parties to the Convention on Wetlands, Gland, Switzerland, 2002.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

VI

ENT: Environmental Management Using Environmental Technologies

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

50

Irrigation Systems: Subsurface Drip Design

Carl R. Camp, Jr.
and Freddie L.
Lamm

Introduction

Subsurface drip irrigation (SDI) is generally defined as the application of water below the soil surface through emitters, with discharge rates in the same range as drip irrigation.[1] While this definition is not specific regarding depth below the soil surface, most SDI laterals are installed at a depth sufficient to prevent interference with surface traffic or tillage implements and to provide a useful life of several years as opposed to annual replacement of surface or near-surface drip laterals.

Development of drip irrigation accelerated with the availability of plastics following World War II, primarily in Great Britain, Israel, and the United States. SDI was part of drip irrigation development in the United States beginning about 1959, especially in Hawaii and California. While early drip irrigation products were relatively crude by modern standards, SDI devices were being installed in both experimental and commercial farms by the 1970s. As drip irrigation products improved during the 1970s and early 1980s, surface drip irrigation grew at a faster rate than SDI, probably because of emitter plugging problems and root intrusion. However, interest in SDI increased during the early 1980s, increased rapidly during the last half of the 1980s, and continues today, especially in areas with declining water supplies, with environmental issues related to irrigation, and where wastewater is used for irrigation. Initially, SDI was used primarily for sugarcane, vegetables, tree crops, and pineapple in Hawaii and California. Later, SDI use was expanded to other geographic areas and to agronomic and vines crops, including cotton, corn, and grapes. Currently in the United States, the major uses of SDI are for cotton, processing tomato, corn, and onions using various installation depths and other design aspects.[2]

SDI has the advantage of multiple-year life, reduced interference with cultural practices, dry plant foliage, and a dry soil surface. Multiple-year life allows amortization of the entire system cost over several years, often more than ten. There are a few systems in the United States that have achieved 20 years are more in longevity.[3] If all system components are installed below tillage depth, surface cultural practices can be accomplished with minimal concern for system damage. Dry soil surfaces can reduce weed growth in arid climates and may reduce evaporation losses of applied water. Because the plant canopy is not irrigated, the foliage remains dry, which may reduce the incidence of disease. SDI is also very adaptable to irregularly shaped fields and low-capacity water supplies that may provide design limitations with other irrigation systems.

The major disadvantages of SDI include system cost, difficulty in locating and repairing system leaks and plugged emitters, and poor soil water manipulation near the soil surface. Most system components are installed below the soil surface and are neither easy to locate nor directly observable. In a properly designed and managed SDI system, the soil surface should seldom be wet. Consequently, seed germination, especially for small seeds, can be very difficult.

SDI systems offer considerable flexibility, both in design and operation. For example, SDI systems can apply small, frequent water applications, often multiple times each day, to very specific sites within the soil profile and plant root zone. Fertilizers, pesticides, and other chemical amendments can be applied via the irrigation system directly into the active root zone, often at a modest increase in equipment cost. In many cases, the operational cost may be less than that for applying these chemicals via conventional surface equipment.

System Design

Site, Water Supply, and Crop

Design of subsurface drip systems is similar to that of surface drip systems, especially with regard to hydraulic characteristics.[3,4] Specific crop and soil characteristics are used in the design process to select emitter spacing and flow rate, lateral depth and spacing, and the required system capacity. Emitter properties and lateral location are influenced by soil properties such as texture, soil compaction, and soil layering because these affect the rate of water movement through the soil profile and the subsequent wetting pattern for each emitter.

The water supply capacity directly affects the design of an SDI system. The size of the irrigated field or zone is often determined by the water supply capacity. For example, in some humid areas, high-capacity wells are not available but multiple low-capacity wells can be distributed throughout a farm. Fortunately, the design of SDI systems can be economically adjusted to correspond to the field size and shape, to the available water supply capacity, and to other factors. Water supply quality should be tested by an approved laboratory before proceeding with system design. This information is needed for the proper design and management of the water filtration and treatment system. Some water supplies require frequent or intermittent injection of acids and/or chlorine. Other saline and/or sodic water supplies may require treatment or special management. As water supplies become more limited, treated wastewater is becoming an increasingly important alternative water supply that can be applied through SDI systems. Camp[5] listed several reports that emphasized water supplies (saline, deficit, and wastewater) for SDI systems.

The SDI system is usually designed to satisfy peak crop water requirements, which vary with specific site, soil, and crop conditions. When properly designed and managed, SDI is one of the most efficient irrigation methods, providing typical application efficiencies exceeding 90%. In comparison with other methods of irrigation, reported yields with SDI were equal to or greater than those with other irrigation methods. Generally, water requirements with SDI are similar or slightly lower than those with other irrigation methods. In some cases, water savings of up to 40% have been reported.[5] However, unless more specific information is available, it is usually best to use standard net water requirements for the location when designing SDI systems.

Lateral Type, Spacing, and Depth

SDI lateral depth for various cropping systems is normally optimized for prevailing site conditions and soil characteristics.[3,5] Where systems are used for multiple years and tillage is a consideration, lateral depths vary from 0.20 to 0.70 m. Where tillage is not a consideration (e.g., turfgrass, alfalfa), depth is sometimes less (0.10–0.40 m). Lateral spacing also varies considerably (0.25–5.0 m), with narrow spacing used primarily for turfgrass and wide spacing used for vegetable, tree, or vine crops. In uniformly spaced row crops, the lateral is usually located under either alternate or every third midrow area (furrow). For crops with alternating row spacing patterns, the lateral is located about 0.8 m from each row, usually in the narrow spacing of the pattern.

The lateral should be installed deep enough to prevent damage by tillage or injection equipment but shallow enough to supply water to the crop root zone without wetting the soil surface. Generally, laterals in SDI systems are placed at depths of 0.1–0.5 m, at shallower depths in coarse-textured soils, and at slightly deeper depths on finer-textured soils. The selection of emitter spacing and flow rate is influenced by crop rooting patterns, lateral depth, and soil characteristics. It is also desirable to select an emitter spacing that provides overlapping subsurface wetted zones along the lateral for most row crops. For wider spaced crops such as trees and vines, emitters are normally located near each plant and may have wider spacings that do not provide overlapping patterns. Lateral spacing is determined primarily by the soil, crop, and cultural practice and should be narrow enough to provide a uniform supply of water to all plants.

Special Requirements

Site topography must be considered in system design and selection of components as with any irrigation system, but SDI is suitable for most sites, ranging from flat to hilly. For sites with considerable elevation change, especially along the lateral, pressure-compensating emitters should be used.

Two special design requirements for SDI systems, which are significantly different from those for surface drip systems, are the needs for flushing manifolds and air entry valves. Flushing manifolds are needed to allow frequent flushing of particulate matter that may accumulate in laterals. Air relief valves are needed to prevent aspiration of soil particles into emitter openings when the system is depressurized. These valves must be located in sufficient number and at the higher elevations for each lateral or zone to prevent negative pressures within the laterals.

Emitter plugging caused by root intrusion is a major problem with some SDI systems but can be minimized by chemicals, emitter design, and irrigation management. Chemical controls include the use of herbicides, either slow-release compounds embedded into emitters and filters or periodic injection of other chemical solutions (concentrated and/or diluted) into the irrigation supply. Periodic injection of acid and chlorine for general system maintenance can also modify the soil solution immediately adjacent to emitters and reduce root intrusion. In some cases, emitters plugged by roots may be cleared via injection of higher concentrations of chemicals, such as acids and chlorine.

Emitter design may also affect root intrusion. Smaller orifices tend to have less root intrusion but are more susceptible to plugging by particulate matter. Some emitters are constructed with physical barriers to root intrusion. Root intrusion appears to be more severe when emitters are located along dripline seams, which can be an area of preferential root growth. However, root intrusion problems appear to be greater for emitters, driplines, and porous tubes that are not chemically treated.

Irrigation management can also be used to influence root intrusion by controlling the environment immediately adjacent to the emitter. High-frequency pulsing that frequently saturates the soil immediately surrounding the emitter can discourage root growth in that area for some crops but not others. Conversely, deficit irrigation sometimes practiced to increase quality or maturity, or to control vegetative growth, can increase root intrusion in lower rainfall areas because of high root concentrations in the soil zone near emitters.

System Components

Pumps, Filtration, and Pressure Regulation

Pump requirements for SDI are similar to those for other drip irrigation systems, meaning water must be supplied at a relatively low pressure (170–275 kPa) and flow rate in comparison to other irrigation methods. Because of the flushing requirement for SDI systems, a flow velocity of about 0.3 m/s must be achievable, either by reducing the zone size while using the same pumping rate or by increasing the pumping rate without changing the zone size.

Water filtration is more critical for SDI systems than for surface drip systems because the consequences of emitter clogging are more severe and more costly. Generally, the better the water quality, the less complex the filtration system required. Surface and recycled or wastewater supplies require the most elaborate filtration systems. However, good filtration is the key to good system performance and long life and should be a major emphasis in system design. Filtration systems range from simple screen filters for relatively clean water to more elaborate and complex disc and sand media filters for poorer quality water.

The pressure regulation requirement in SDI systems is similar to that in surface drip systems. When non-pressure-compensating emitters are used on relatively flat areas, pressure is typically regulated within the system supply lines (main and/or submain) using pressure-regulating valves. When pressure-compensating emitters are used, typically on more hilly terrain, the pressure within the system supply lines is controlled at a higher, but more variable, pressure that is within the recommended input pressure range for the emitters used. Water pressure should be monitored on a regular basis at the pump or supply port and at various locations throughout the SDI system, especially at the both ends of laterals.

Laterals and Emitters

Many types of driplines have been used successfully for SDI and most have emitters installed as an integral part of the dripline. This is accomplished by one of three methods: (1) molded indentions created during the fusing of dripline seams, (2) prefabricated emitters welded inside the dripline, or (3) circular prefabricated in-line emitters installed during extrusion. Regardless of the emitter used, dripline wall thickness and expected longevity must be considered along with other design factors in selecting the lateral depth. Flexible, thin-walled driplines typically are installed at shallow depths and normally have a shorter expected life. Thicker-walled, flexible driplines have been used successfully for several years provided they are installed deep enough to avoid tillage, cultivating, and harvesting machinery but shallow enough to prevent excessive deformation or permanent collapse of the dripline by machinery or soil weight. Rigid tubing with thicker walls can be installed at deeper depths without deformation and is often used on perennial crops or annual crops for longer time periods (>10 years). Some driplines are impregnated with bactericides or other chemicals to reduce the formation of sludge or other material that could plug emitters.

Chemical Injection

Subsurface drip systems offer the potential for precise management of water, nutrients, and pesticides if the system is properly designed and managed. The marginal cost to add chemical injection equipment is generally competitive with other, more conventional application methods. Water and fertilizers can be applied in a variety of modes, varying from multiple continuous or pulsed applications each day to one application in several days. Choice of application frequency depends upon several factors, including soil characteristics, crop requirements, water supply, system design, and management strategies. If labeled for the purpose, some systemic pesticides and soil fumigants can be safely injected via SDI systems. Use of the SDI system for chemical applications has the potential to minimize exposure to workers and the environment, reduce the cost of pesticide rinse water disposal, and improve precision of application to the desired target (root pests). Injection of other chemicals, such as acids and chlorine, is often required to clean and maintain emitters in optimum condition. However, a high level of management with system automation and feedback control is required to minimize chemical movement to the groundwater when chemicals are used.

Air Entry and Flushing

Air entry valves must be installed at higher elevations in SDI systems to prevent the emitter from ingesting soil particles that could plug emitters when the system is depressurized. Typically, air entry valves are located in water supply lines near the head works or control station and in both the supply and flushing manifolds. In some cases, such as turf or pasture, air entry valves may be installed below the soil surface and enclosed within a protective box. Flushing valves installed on the flushing manifold are required to control periodic system flushing.

Operation and Maintenance

Operation

SDI systems can be operated in several modes, varying from manual to fully automated. Overall, SDI systems are probably more easily automated than many other types of irrigation. One reason is that most are controlled from a central point using electrical or pneumatic valves and controllers that vary from a simple clock system to microprocessor systems, which are capable of receiving external inputs to initiate and/or terminate irrigation events.

Irrigation scheduling is as important for SDI systems as for any other type of irrigation. Choosing to initiate an irrigation event and how much water to apply during each event depends on crop, soil, and irrigation system type and design. Factors that affect those decisions include soil water storage volume, sensitivity of the crop to water stress, irrigation application rate, weather conditions, and water supply capacity. Camp[5] discussed several irrigation scheduling methods that have been used successfully with SDI. However, the important point is that a science-based scheduling method can conserve the water supply and increase profit.

If seed germination and seedling establishment and growth are critical, especially in arid climates when initial soil water content is not adequate, either sprinkler or surface irrigation is often used for germination. However, the need for two systems increases cost and decreases economic return. If subsurface drip is used for germination, an excessive amount of irrigation is often required to wet the seed zone for germination, which could result in excessive leaching and off-site environmental effects as well as increased cost. Surface wetting can also occur when the emitter flow rate exceeds the hydraulic conductivity of the soil surrounding the emitter, but wetted areas are often not uniform.

Because salts tend to accumulate above the lateral, high salt concentrations may occur between the lateral and soil surface in arid areas where rainfall is not available to leach the salts downward. Salts may also be moved under the row when laterals are placed under the furrow.[6] Supplemental sprinkler irrigation may be required in some areas to control salinity if precipitation is inadequate for leaching during several consecutive years.

Maintenance

Often, SDI systems must have a long life (>10 years) to be economical for lower value crops. Thus, appropriate management strategies are required to prevent emitter clogging and protect other system components to ensure proper system operation. Locating and repairing/replacing failed components is much more difficult and more expensive with SDI systems than with surface systems because most system components are buried, difficult to locate, and cannot be directly observed by managers. Consequently, operational parameters such as flow rate and pressure must be measured frequently and used as indicators of system performance. Good system performance requires constant attention to maintain good water quality, proper filtration, and periodic system flushing to remove particulate matter that could plug emitters. Periodic evaluation of SDI system performance in relation to design performance can identify problems before they become serious and significantly affect crop yield and quality.

Conclusion

Although there is a general consensus that use of SDI is increasing on a worldwide basis, this growth is difficult to document. In the 10-year period (2003–2013), SDI in the United States increased by 89% from 164,017 to 310,361 ha according to USDA-NASS irrigation surveys.[2] Use of SDI should continue to increase in the future, depending primarily upon the economic and water conservation benefits in comparison to other irrigation methods. As water supplies become more limited, the high application efficiency and water-conserving features of SDI should increase its application. Also, SDI offers potential advantages such as reduced odors and exposure to pathogens when using recycled domestic and animal wastewater. The SDI technology offers the capability to precisely place water, nutrients, and other chemicals in the plant root zone at the time and frequency needed for optimum crop production. With proper design, installation, and management, SDI systems can provide excellent irrigation efficiency and reliable performance with a system life of 10–20 years.

References

1. ASABE. *Soil and Water Terminology*. 49th Ed.; ASAE: St. Joseph, MI, 2001; 970–990.
2. Lamm, F. R. 2016. Cotton, tomato, corn, and onion production with subsurface drip irrigation – A review. *Trans. ASABE 59* (1), 263–278.
3. Lamm, F.R. and C.R. Camp. 2007. Subsurface drip irrigation. In *Microirrigation for Crop Production - Design, Operation and Management*. F.R. Lamm, J.E. Ayars, and F.S. Nakayama (Eds.), Elsevier Publications: Amsterdam; 473–551.
4. ASAE. *Design and Installation of Microirrigation Systems*. 49th Ed.; ASAE: St. Joseph, MI, 2001; 903–907.
5. Camp, C.R. Subsurface drip irrigation: A review. *Trans. ASAE* **1998**, *41* (5), 1353–1367.
6. Ayars, J.E.; Phene, C.J.; Schoneman, R.A.; Meso, B.; Dale, F.; Penland, J. Impact of bed location on the operation of subsurface drip irrigation systems. In Microirrigation for a Changing World, *Proceedings of the Fifth International Microirrigation Congress*, Orlando, FL, April 2–6, 1995; Lamm, F.R., Ed.; ASAE: St. Joseph, MI, 1995; 141–146.

51

Recent Approaches to Robust Water Resources Management under Hydroclimatic Uncertainty

J. Pablo Ortiz-Partida, Mahesh L. Makey, and Alejandra Virgen-Urcelay

Introduction

Managing water resources under hydroclimatic uncertainty is a primary focus with respect to water resources management (Brown et al. 2015; Poff et al. 2015). There is a scientific consensus that climate change is going to impact current water systems in a variety of ways. For example, expected impacts on California's water resources include increase in evaporation rates and water demands (Hayhoe et al. 2004); reduction in crop yield as a consequence of higher temperatures (Pathak et al. 2018); less precipitation falling as snow and more as rain translating into wetter wet seasons and drier dry seasons (Mallakpour et al. 2018), which may force operators to lower reservoir levels to allow for space and reduce the risk of flood events; increase in water temperatures affecting aquatic and riparian ecosystem (Poff et al. 2012); and sea-level rise, which will increase seawater intrusion into coastal aquifers (Ferguson and Gleeson 2012). Meanwhile, there is, however, uncertainty on their magnitude and extent, which raise the question of how we can develop management strategies that account for the range of possible future alternatives.

Water systems modeling has been crucial to answer such questions because of the advancement of computational power. Water systems modeling was often deterministically developed, considering some variables such as temperatures, precipitation, streamflow, or water demands as known parameters. Despite recognizing their hydrologic variability, the only available data are often treated as a statistical description of hydrologic variables resulting into unreliable forecasts. Deterministic models treat stochastic parameters as known quantities reducing – to some extent – the complexity of the model (Chen et al. 2018). Depending on the application of the model, such assumptions can be sustained. Nevertheless, the simplifications cannot retain all the essential characteristics of the original data and may lead to unsatisfactory results given the complex behavior of a system (Puente et al. 2018). Climate change is bringing extreme hydroclimatic events of low probabilities but high impacts that – as noted – deterministic approaches do not capture well (Farmer and Vogel 2016; Philbrick and Kitanidis 1999). Thus, there is a need for integrating the seemingly random behavior of precipitation, temperature, streamflow, and other sometimes-unknown variables (e.g., soil properties, water quality) for developing water management strategies suitable for a variety of future socioeconomic and climatic alternatives.

This chapter describes recent approaches from the scientific literature that incorporate hydroclimatic uncertainty to develop robust water management strategies. This chapter is divided into three sections. The first section describes the development of robust reservoir operations, a main technical area in water resources. The second section describes an approach to achieve water systems sustainability that goes beyond common performance criteria (e.g., economic, reliability) by integrating stakeholder-defined performance metrics within the modeling framework. The last section outlines how to move forward and incorporate uncertainty into the decision-making process, a crucial component pertaining to real-life activities.

Robust Reservoirs Management

Reservoirs around the world provide storage to supply water for urban, agriculture, and industry sectors; flood protection; hydropower; and support recreational activities. However, in some cases, these benefits came at the cost of environmental and social degradation (WCD 2000). Leaving aside the controversy of building dams, it is important to recognize that most of the current reservoirs will persist despite their contribution to the degradation of river ecosystems. Given the uncertainty and variability of hydroclimatic conditions under climate change, droughts and floods may worsen, turning reservoir management vital to prevent or reduce expected drought and flood events (Cristina and Tullos 2017).

Modifying reservoir operations offers an opportunity for mitigating hydrologic responses to climate change as current operation rules tend to be static and based on historical inflows and outflows observations. Given the new recurrence and magnitude of hydroclimatic events, some reservoir operation rules are no longer suitable for managing drought periods and floods or for reaching the full potential of a reservoir (Howard 1999; Moy et al. 1986). New reservoir operations that incorporate projected hydroclimatic variability may be an effective strategy for reducing the impacts of changes in water supplies and demands (Vonk et al. 2014). Reservoir operations developed with new available tools and information contribute to increased resilience of water management systems and ecosystem restoration.

Alternative approaches to improve long-term reservoir operations have utilized multi-objective genetic programming (Ashofteh et al. 2015), machine learning (Herman and Steinschneider 2018), and stochastic modeling (Ermoliev et al. 2019) to incorporate climate change projections and a broader set of observations and forecasts. Such approaches link hydroclimatic observations and predictions with water resources management decisions to improve tradeoffs among human and environmental water supply and flood management objectives.

Another recently proposed framework presents a two-stage stochastic optimization model that maximizes regional economic benefits as a function of reservoir deliveries to water users (Ortiz-Partida et al. 2019). The first-stage decisions allocate water based on an expected deficit for the whole system, while in the second stage, the now known deficit is allocated in a way that maximizes economic benefits. This framework was applied to the single reservoir system of Luis L. Leon dam in the Rio Conchos, the main tributary to the Rio Grande of North America. The model result is a set of robust monthly reservoir releases that can cover most of the water demands – including the environment – and reduce the frequency and magnitude of flood events under a wide range of hydroclimatic conditions.

Results from these different approaches suggest that robust operations could improve long-term planning by making the water system more reliable and resilient, and less vulnerable to extreme hydroclimatic events. However, these technically complex approaches tend to have bigger scale implications and usually lack stakeholder involvement, making them difficult to implement.

Integrated Framework

Attaining water sustainability needs to go beyond historically used criteria (e.g., economic performance) and integrate socioeconomic, environmental, and climate components. As such, decision makers are seeking for robust decisions with satisfactory performance across a large range of plausible futures. To make plans that are flexible and enhance long-term decision-making, various model-based frameworks have been proposed (Kwakkel et al. 2016). One of the frameworks that distinguishes from others is decision scaling (DS). DS integrates stakeholder-defined performance metrics that allow for inclusion and empowerment to foster collaboration across historically conflicting perspectives (Brown et al. 2011, 2012). In addition, DS attempts to find alternative solutions for multi-objective water resources systems under a variety of nonstationary hydroclimatic conditions and modeling uncertainties. Steinschneider et al. (2015) advanced previous applications of DS to identify long-term planning alternatives with the use of Monte Carlo simulations and by quantifying uncertainty in each stage of the modeling chain. Such a novel approach efficiently selects future climate realizations to assess the effects of hydrologic modeling uncertainty. Its limitation lies on a computational burden and required high-performance computing facilities to handle large multi-reservoir systems.

DS has been further developed by Poff et al. (2015) into the eco-engineering decision-scaling (EEDS) framework that explicitly and quantitatively explores tradeoffs across management actions under hydroclimatic uncertainty and incorporates ecological and engineering performance metrics. EEDS has five main steps that include (1) developing a clear definition of performance criteria, unacceptable thresholds, and management options; (2) building a systems model that relates hydroclimatic variables with the defined performance criteria; (3) performing a vulnerability analysis under a variety of hydroclimatic and social alternatives to discard unacceptable options and identify robust management strategies; (4) evaluating management options on the basis of their ability to satisfy both engineering and ecosystem performance under uncertain future conditions; and lastly, (5) selecting an alternative or developing new performance criteria based on political and institutional feasibility through converting the framework into an iterative process. As a case study around Coralville Dam, Poff et al. (2015) identified the most robust management strategy that would satisfy economic and ecological goals (Figure 1).

In truly achieving robustness and avoiding undesirable consequences, Herman et al. (2015) argue that (1) alternative decisions should be searched and not prespecified, (2) uncertainty should be discovered with sensitivity analysis and not assumed, and (3) stakeholders should carefully select their satisfactory measure of robustness based on their problem-specific performance requirements.

Moving Forward

Robust water management strategies need to go beyond water data and incorporate uncertainty in socioeconomic, environmental, and climate variables to be capable of assessing a broader range of future alternatives. Moreover, other information related to global climate change, population growth, land-use change, globalization, or even political stability makes long-term water resources planning difficult. Projected information of these variables is pertinent to long-term sustainable water resources strategies. A lack of information regarding interactions among these variables often misleads outputs from models entailing uncertainties to be adopted by water managers, planners, and decision makers. A first step towards creating robust water management strategies is to improve data collection and monitoring, which can be costly, suggesting that benefits and tradeoffs between investments in monitoring and development should be considered properly (Wada et al. 2016).

Adequate performance of long-term water resources management highly depends on a modeling chain that links technical, socioeconomic, environmental, and climate components. This is because each stage of the modeling chain is vulnerable to uncertainty. Such uncertainties include nonstationarity in future climate and potential shifts in large-scale synoptic circulation (Sheridan and Lee 2010) as

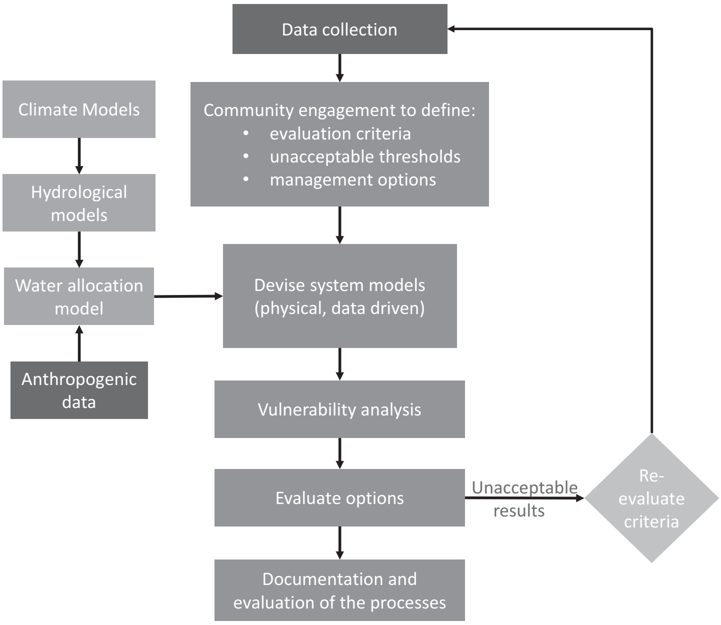


FIGURE 1 Eco-engineering decision scaling (EEDS). (Adapted from Poff et al. (2015).)

well as nonstationarity in water demands (Kang and Lansey 2012). As observed and experienced, it is difficult to propose long-term robust approaches without better prediction of climatic shifts and water demand variability (Korteling et al. 2013; Matrosov et al. 2013; Prudhomme et al. 2010; Turner et al. 2014). In fact, these frameworks attempt to bring the pieces together for decision-making rather than accounting each component individually.

To meet diverse water demands, Groves et al. (2019) recently applied "robust decision making (RDM)" to (a) identify water management strategies to reduce vulnerabilities in the Colorado River Basin and (b) develop robust investment strategies for the Green Climate Fund that reduce greenhouse gas emissions. This work suggests that RDM addresses long-term policy challenges associated with multifaceted, nonlinear, natural, and socioeconomic systems for defining policy pathways in a beneficial way. Existing water system models still need to address issues related to current human impact modeling and associated indicators and their limitations in representing regional water management. In this way, one can integrate land use and climate interaction to integrate human behavior into a large-scale modeling and to fill the gap among human–water management information. Advancing water system models coupled with human–water systems over a wide range of temporal and spatial scales will be beneficial for water science community and for understanding the climate and Earth system science communities.

Conclusions and Future Needs

Current water management strategies and infrastructure were not designed for some of the hydroclimatic changes that we have observed in the last decades. Robust water resources management strategies provide an opportunity to cope and adapt to the changing climate. This chapter presents different approaches to incorporate hydroclimatic uncertainty for developing robust water resources management strategies. Technically complex approaches tend to have bigger scale implications and usually lack stakeholder involvement, which leads to simplified social and environmental components. Long-term planning with stakeholder involvement tends to simplify climate processes. Ultimately, the social, economic, environmental, and climate components must be considered for developing sustainable and resilient water management systems. Despite potential model and data limitations, one important step is to account more realistically for the nature of socioeconomic and environmental components to identify associated uncertainties and data gaps. Such considerations may lead to accurately assess future sustainability of water use under the current envisaged population growth, ecosystems degradation, and climate change.

References

Ashofteh, Parisa-Sadat, Omid Bozorg Haddad, and Hugo A. Loáiciga. 2015. "Evaluation of Climatic-Change Impacts on Multiobjective Reservoir Operation with Multiobjective Genetic Programming." *Journal of Water Resources Planning and Management* 141 (11): 04015030. doi:10.1061/(ASCE)WR.1943-5452.0000540.

Brown, Casey, Yonas Ghile, Mikaela Laverty, and Ke Li. 2012. "Decision Scaling: Linking Bottom-up Vulnerability Analysis with Climate Projections in the Water Sector." *Water Resources Research* 48 (9). doi:10.1029/2011WR011212.

Brown, Casey M. Jay R. Lund, Ximing Cai, Patrick M. Reed, Edith A. Zagona, Avi Ostfeld, Jim Hall, Gregory W. Characklis, Winston Yu, and Levi Brekke. 2015. "The Future of Water Resources Systems Analysis: Toward a Scientific Framework for Sustainable Water Management." *Water Resources Research* 51: 6110–24. doi:10.1002/2015WR017114.

Brown, Casey, William Werick, Wendy Leger, and David Fay. 2011. "A Decision-Analytic Approach to Managing Climate Risks: Application to the Upper Great Lakes1." *Journal of the American Water Resources Association (JAWRA)* 47 (3): 524–34. doi:10.1111/j.1752-1688.2011.00552.x.

Chen, Cheng, Chuanxiong Kang, and Jinwen Wang. 2018. "Stochastic Linear Programming for Reservoir Operation with Constraints on Reliability and Vulnerability." *Water* 10 (2): 175.

Cristina, M. Mateus, and Desiree Tullos. 2017. "Reliability, Sensitivity, and Vulnerability of Reservoir Operations under Climate Change." *Journal of Water Resources Planning and Management* 143 (4): 04016085. doi:10.1061/(ASCE)WR.1943-5452.0000742.

Ermoliev, Yu., T. Ermolieva, T. Kahil, M. Obersteiner, V. Gorbachuk, and P. Knopov. 2019. "Stochastic Optimization Models for Risk-Based Reservoir Management." *Cybernetics and Systems Analysis* 55 (1): 55–64. doi:10.1007/s10559-019-00112-z.

Farmer, William H., and Richard M. Vogel. 2016. "On the Deterministic and Stochastic Use of Hydrologic Models." *Water Resources Research* 52 (7): 5619–5633. doi:10.1002/2016WR019129.

Ferguson, Grant, and Tom Gleeson. 2012. "Vulnerability of Coastal Aquifers to Groundwater Use and Climate Change." *Nature Climate Change* 2 (5): 342.

Groves, David G., Edmundo Molina-Perez, Evan Bloom, and Jordan R. Fischbach. 2019. "Robust Decision Making (RDM): Application to Water Planning and Climate Policy." In Decision Making under Deep Uncertainty: From Theory to Practice, edited by Vincent A. W. J. Marchau, Warren E. Walker, Pieter J. T. M. Bloemen, and Steven W. Popper, 135–63. Springer: Cham.

Hayhoe, Katharine, Daniel Cayan, Christopher B. Field, Peter C. Frumhoff, Edwin P. Maurer, Norman L. Miller, Susanne C. Moser, et al. 2004. "Emissions Pathways, Climate Change, and Impacts on California." *Proceedings of the National Academy of Sciences of the United States of America* 101 (34): 12422. doi:10.1073/pnas.0404500101.

Herman, Jonathan D., Patrick M. Reed, Zeff B. Harrison, and Gregory W. Characklis. 2015. "How Should Robustness Be Defined for Water Systems Planning under Change?" *Journal of Water Resources Planning and Management* 141 (10): 04015012. doi:10.1061/(ASCE)WR.1943-5452.0000509.

Herman, Jonathan D., and S. Steinschneider. 2018. "Atmospheric Rivers and Reservoir Control: Combining Short-Term Forecasts with Clustering of Weather Regimes to Inform Conjunctive Management in California." In, 2018:H21G-02. https://ui.adsabs.harvard.edu/abs/2018AGUFM.H21G..02H.

Howard, Charles D.D. 1999. “Death to Rule Curves.” *29th Annual Water Resources Planning and Management Conference, Proceedings*, 1–5. doi:10.1061/40430(1999)232.

Kang, Doosun, and Kevin Lansey. 2012. “Multiperiod Planning of Water Supply Infrastructure Based on Scenario Analysis.” *Journal of Water Resources Planning and Management* 140 (1): 40–54.

Korteling, Brett, Suraje Dessai, and Zoran Kapelan. 2013. “Using Information-Gap Decision Theory for Water Resources Planning under Severe Uncertainty.” *Water Resources Management* 27 (4): 1149–72.

Kwakkel, Jan H., Marjolijn Haasnoot, and Warren E. Walker. 2016. “Comparing Robust Decision-Making and Dynamic Adaptive Policy Pathways for Model-Based Decision Support under Deep Uncertainty.” *Environmental Modelling & Software* 86: 168–83.

Mallakpour, Iman, Mojtaba Sadegh, and Amir AghaKouchak. 2018. “A New Normal for Streamflow in California in a Warming Climate: Wetter Wet Seasons and Drier Dry Seasons.” *Journal of Hydrology* 567 (December): 203–11. doi:10.1016/j.jhydrol.2018.10.023.

Matrosov, Evgenii S., Ashley M. Woods, and Julien J. Harou. 2013. “Robust Decision Making and Info-Gap Decision Theory for Water Resource System Planning.” *Journal of Hydrology* 494: 43–58.

Moy, Wai-See, Jared L. Cohon, and Charles S. ReVelle. 1986. “A Programming Model for Analysis of the Reliability, Resilience, and Vulnerability of a Water Supply Reservoir.” *Water Resources Research* 22 (4): 489–98.

Ortiz-Partida, J. Pablo, Taher Kahil, Tatiana Ermolieva, Yuri Ermoliev, Belize Lane, Samuel Sandoval-Solis, and Yoshihide Wada. 2019. “A Two-Stage Stochastic Optimization for Robust Operation of Multipurpose Reservoirs.” *Water Resources Management* 33 (11): 3815–3830. doi:10.1007/s11269-019-02337-1.

Pathak, B. Tapan, L. Mahesh Maskey, A. Jeffery Dahlberg, Faith Kearns, M. Khaled Bali, and Daniele Zaccaria. 2018. “Climate Change Trends and Impacts on California Agriculture: A Detailed Review.” *Agronomy* 8 (3). doi:10.3390/agronomy8030025.

Poff, Boris, Karen A. Koestner, Daniel G. Neary, and David Merritt. 2012. “Threats to Western United States Riparian Ecosystems: A Bibliography.” *Gen. Tech. Rep.* RMRS-GTR-269. Fort Collins, CO: US Department of Agriculture, Forest Service, Rocky Mountain Research Station. 78 p. 269.

Philbrick, C. Russ, and Peter K. Kitanidis. 1999. “Limitations of Deterministic Optimization Applied to Reservoir Operations.” *Journal of Water Resources Planning and Management* 125 (3): 135–142. doi:10.1061/(ASCE)0733-9496(1999)125:3(135).

Poff, N. LeRoy, Casey M. Brown, Theodore E. Grantham, John H. Matthews, Margaret A. Palmer, Caitlin M. Spence, Robert L. Wilby, et al. 2015. “Sustainable Water Management under Future Uncertainty with Eco-Engineering Decision Scaling.” *Nature Climate Change* 6: 25. doi:10.1038/nclimate2765; https://www.nature.com/articles/nclimate2765#supplementary-information.

Prudhomme, Christel, Robert L. Wilby, S. Crooks, Alison L. Kay, and Nick S. Reynard. 2010. “Scenario-Neutral Approach to Climate Change Impact Studies: Application to Flood Risk.” *Journal of Hydrology* 390 (3–4): 198–209.

Puente, Carlos E., Mahesh L. Maskey, and Bellie Sivakumar. 2018. “Studying the Complexity of Rainfall Within California Via a Fractal Geometric Method.” In *Advances in Nonlinear Geosciences*, Anastasios A. Tsonis (Ed.) 519–42. Springer: Cham.

Sheridan, Scott C., and Cameron C. Lee. 2010. “Synoptic Climatology and the General Circulation Model.” *Progress in Physical Geography* 34(1): 101–9.

Steinschneider, Scott, Rachel McCrary, Sungwook Wi, Kevin Mulligan, Linda O. Mearns, and Casey Brown. 2015. “Expanded Decision-Scaling Framework to Select Robust Long-Term Water-System Plans under Hydroclimatic Uncertainties.” *Journal of Water Resources Planning and Management* 141(11): 04015023.

Turner, Sean W.D., David Marlow, Marie Ekström, Bruce G. Rhodes, Udaya Kularathna, and Paul J. Jeffrey. 2014. “Linking Climate Projections to Performance: A Yield-based Decision Scaling Assessment of a Large Urban Water Resources System.” *Water Resources Research* 50(4): 3553–67.

Vonk, E., YuePing Xu, Martijn J. Booij, Xujie Zhang, and Dionysius C.M. Augustijn. 2014. “Adapting Multireservoir Operation to Shifting Patterns of Water Supply and Demand.” *Water Resources Management* 28(3): 625–43.

Wada, Y., M. Flörke, N. Hanasaki, S. Eisner, G. Fischer, S. Tramberend, Y. Satoh, et al. 2016. “Modeling Global Water Use for the 21st Century: The Water Futures and Solutions (WFaS) Initiative and Its Approaches.” *Geoscientific Model Development* 9 (1): 175–222. doi:10.5194/gmd-9-175-2016.

WCD. 2000. *Dams and Development: A New Framework for Decision-Making: The Report of the World Commission on Dams*. World Commission on Dams: London.

52

Rivers: Restoration

Anna Rabajczyk

Introduction

River ecosystems play a significant role in the human environment and economy. They are characterized by a remarkable variety of plant and animal organisms, while simultaneously acting as ecological corridors and exerting considerable landscaping and recreational impact. The riverbed, riparian zones, and the valley support vegetation, whereas the mass of the flowing water, the bottom soil substratum, riverbanks, and riparian land support a rich generic and onthogenic variety of fauna. The occurrence of these organisms is connected with microhabitats formed owing to the river's morphological diversity, resulting from natural processes occurring over the centuries.[1,2] The naturalness of the river course, including its various curvatures, their radiuses and central angles; the diversification of flow volume, directions, and velocity; the fluctuations of the river water level; and the riverbed bottom type all affect the appearance and maintenance of the morphological diversity. All these factors are subject to change in time and along the river course under the influence of human activity, but also of natural phenomena and processes. If natural processes predominate, living organisms adapt to them and the entire natural system remains stable, due to which rivers can create favorable conditions for the development of various organisms and river ecosystems display very high biological diversity.[2,3]

However, over the centuries, humans have been using rivers, regulating and straightening their courses, building water drops, clearing riverside carrs, and draining floodplains in order to enhance river navigability, and expand farming acreage and area allocated to industrial and housing infrastructure. In many cases, this has led to changing natural watercourses into homogeneous channels deprived of natural value, whose function has become limited to water removal. Simultaneously, the rapid progress of land development, together with farming and industrial advances, has become the source of many pollutants, considerably diversified in terms of quantity and quality, and frequently hazardous to the functioning and life of plants, animals, and humans.[1–5]

Widespread environmental degradation has forced the authorities to take appropriate steps aimed at the introduction of principles of sustainable development to provide social welfare, understood as a possibility to satisfy fundamental needs not only by the present-day but also by future generations, with a focus on eco-safety. Appropriate policies and legislative regulations are being introduced to enable the determination of water quality, and guidelines are being developed for actions to be taken to improve it.[5,6]

One of the objectives in water resources management is to restore and maintain good water quality. First and foremost, it is necessary to specify the subject of river maintenance. To specify the types and

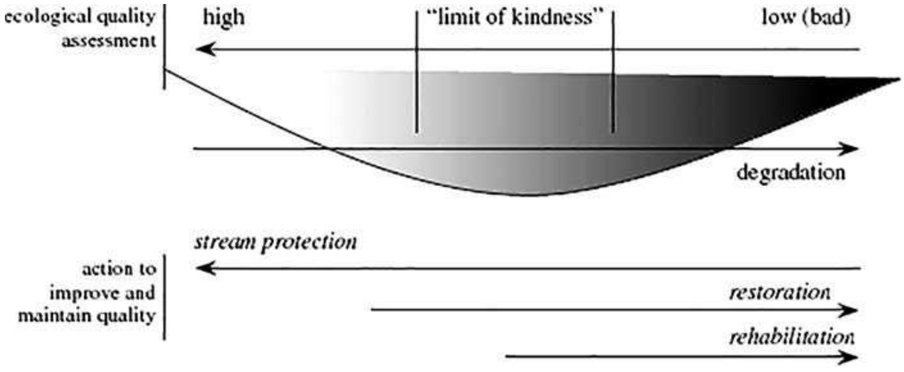


FIGURE 1 Types and scopes of actions geared at river protection and maintenance in terms of quality.
Source: Based on Phillips et al.[7]

scopes of actions geared at river protection and maintenance, it may prove helpful to use a diagram that shows river and stream quality in terms of the so-called limit of kindness (Figure 1).[7]

In the case of very good, good, and moderate quality of water ecosystems, with focus on very good and good quality, only stream protection is justified since actions need to be taken to enable both maintenance and protection of high-quality natural river ecosystems. Another type of action is the recovery of conditions, or restoration, used in the case of rivers characterized by moderate quality, whose "limit of kindness" suggests the hazard of rapid deterioration of the ecosystem's conditions. It is vital to take actions geared at the restoration of natural (primary) characteristics and values from a period taken as a reference point, including, but not limited to

- Natural water quality range
- Natural sedimentation and flow regime (including seasonal fluctuations in a year and multiyear seasons, according to flood patterns)
- Natural riverbed geometry and its stability, determined according to stability ratios, from that period
- Natural riparian vegetation
- Restoration of indigenous aquatic flora and fauna[2,5–9]

Rehabilitation, in turn, pertains to cases of more serious degradation of natural ecosystems and/or when complete restoration of natural features is impossible (e.g., due to technical or economic reasons). Rehabilitation comprises enhancement and improvement of major environmental features of the watercourse so that valuable characteristics thereof are created to define ecological quality.

Another method is to implement remediation, which does not lead to the reconstruction of the primary ecosystem but to the construction of a completely new, different one (Figure 2).[7] This pertains primarily to strongly urbanized (urban) areas, in particular to situations when

- Natural watercourses became degraded in the past and are currently being "brought back to life."
- Ditches were used as storm drains and, partly, as sewage sludges in the past, and are now, upon the introduction of new, closed systems of municipal sewers, often treated as natural.[8–11]

It ought to be stressed that, as a result of adequate actions, it is possible to create new ecosystems, of high biological functionality, yet depleted in terms of biodiversity

River recultivation can only be effective when approached holistically—as an element of comprehensive protection and maintenance of water ecosystems. Consequently, as part of the recultivation process, the planned and implemented actions ought to

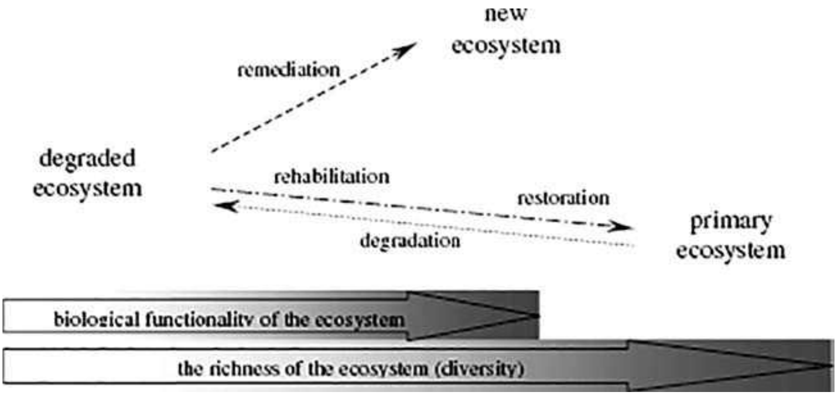


FIGURE 2 Actions aimed at restoration of relevant features of a water ecosystem.
Source: Based on Phillips et al.

- Be integrated, i.e., a particular watercourse (a river segment or the entire river subject to recultivation), together with its appurtenant ecosystem, ought to form one whole, immune to possible environmental impact (seasonal high water, depleted oxygen content, etc.), clearly, up to a limit.
- Reflect the catchment's capacity for natural self-renewal.
- Take into account the physical structure of a given watercourse that fosters the development of desirable flora and fauna (morphological changes, riverbed straightening, or cutoffs from ecosystems may promote degradation).
- Take into account the influence of catchment area users on a particular water body (this is related to the impact of industry, urban areas, or commercial centers).
- Take into account all the reasons for the existing situation, in particular the immediate causes of degradation (including toxic waste discharge, changes to land use, expanded hard road surface in proportion to green areas, and water-logged area elimination).
- Be based on strict guidelines and their methodological implementation, and on adequate monitoring of actions taken to improve the situation.[8–17]

Methods Used in River Recultivation

Adequate implementation of individual actions requires clearly specified methods, techniques, and technical procedures. However, first of all, a conceptual framework ought to be established to specify the principles of operation in terms of

- An assessment of the current status, consistent with legal requirements as well as local/domestic conditions and limitations resulting from both data availability and historical and cultural contexts.
- Principles of operation for the restoration of those features of the ecosystem that determine its capacity for the maintenance of continuity, as well as the major characteristics that constitute the value of ecological quality.
- Following the establishment of the conceptual framework and set of principles, detailed standards are to be developed that will enable the methodological, technical, and technological implementation of appropriate solutions. That is why specialists in a variety of scientific disciplines (such as ecology, water environment biology, hydrology, hydraulics, geomorphology, construction engineering, spatial planning, communications, and sociology) ought to participate in the works.[11,18–26]

Each catchment and its environment are, as a rule, unique, and each case requires an individualized approach, hence the first step is to collect information on the catchment structure and development from a number of sources. This should be complemented with results of research monitoring over a specified period (such as several months or a hydrological year) as well as the catchment area located before the final part of the segment subject to recultivation.[27,28] Only by basing on a wide range of data collected from several sources is it possible to pass on to the phase of specification of individual elements and procedures necessary to obtain recultivation results in a particular watercourse.

The development of criteria for an assessment of a particular watercourse's quality enables standard assessment of current river quality, followed by verification of the actions implemented as part of the recultivation process (Figure 3). These should take into account not only the valid worldwide, European, and domestic regulations, but also the uniqueness resulting from locality, together with environmental and economic functions.[29]

In view of the above, preparatory recultivation works cover watercourse assessment in accordance with adopted criteria, on the basis of which it is possible to describe the status quo and pinpoint shortcomings in relation to an upgraded status. A partial function assessment is based on the lowest rating

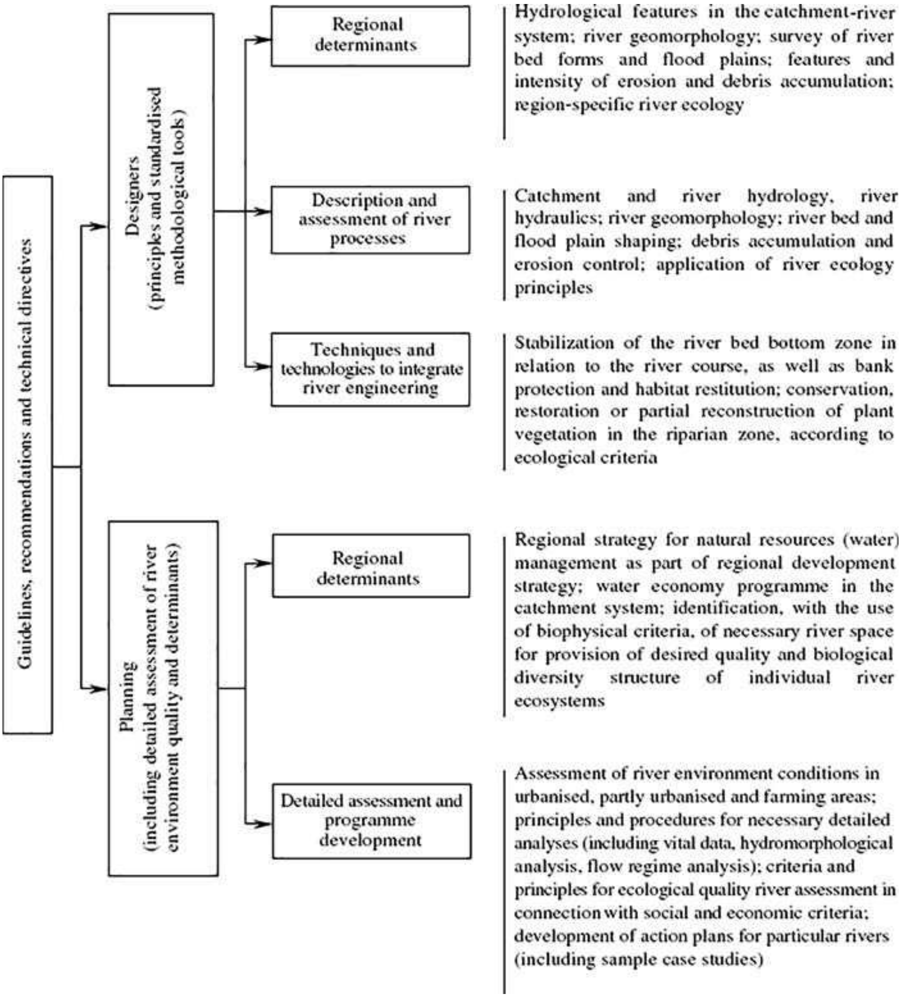


FIGURE 3 Guidelines and recommendations for planning and development of river recultivation programs.

for the criteria appurtenant with that function; the segment rating covers the function's lowest partial rating, with allowances being made for the weakest-link principle. The content of this kind of a document enables fast identification of hotspots whose adjustment will allow for value enhancement of a particular watercourse segment. On this basis, appropriate procedures and diverse actions are selected.

As part of complex works toward river ecosystem renewal, actions are taken to influence various elements of the environment (Figure 4). This course of recultivation allows the objective of river renewal to be achieved. In most cases, single-tasking actions do not yield expected results, or the results take much longer to become visible than they would in the case of a wider spectrum of actions.[29]

Multitasking actions ensure high diversity in the methods and techniques that may be, and are, used in the process of river recultivation. Yet, according to recultivation assumptions, the performed operations ought to enable the intensification of watercourses' natural capacity for cleansing. Consequently, possible works include the following[30–36]

- Construction of aeration-enhancing drops and step falls
- River course regulation
- Elimination of isolated still water pools where oxygen depletion is easy to occur

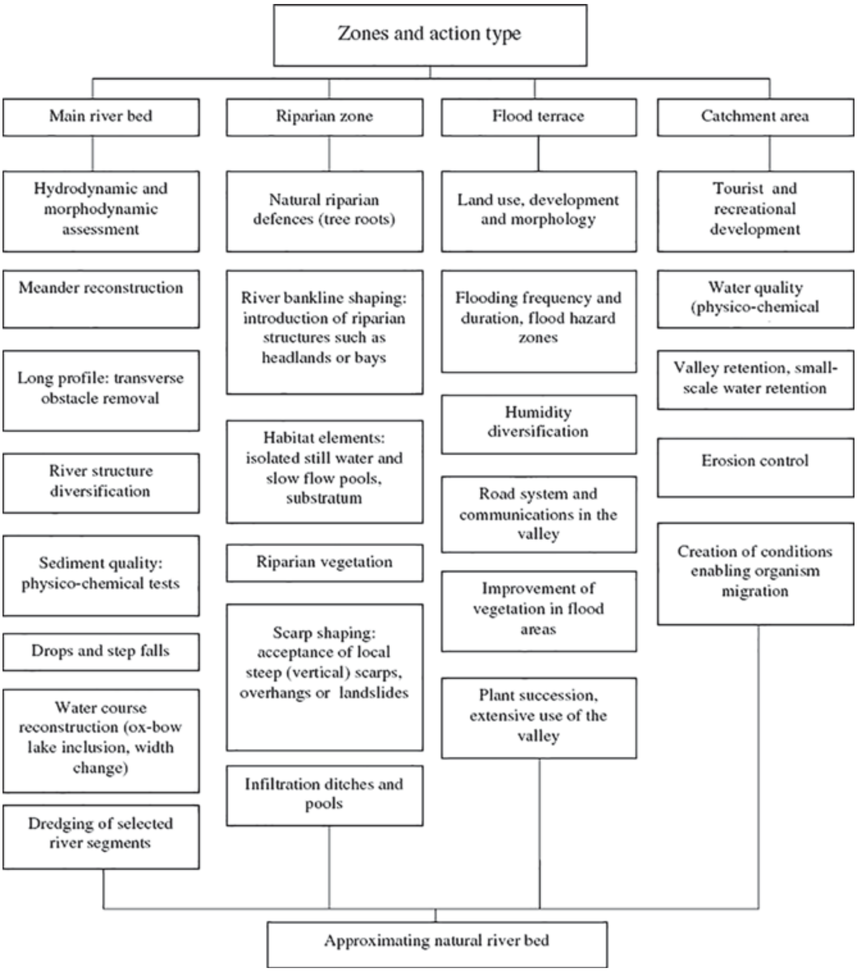


FIGURE 4 Example of multitasking for river recultivation.
Source: Bojarski et al.[30]

- Dredging sediments in selected river segments
- Formation of riparian vegetation strips
- Removal of excess vegetation from the riparian zone
- Formation of infiltration pools and ditches along river- banks to enable both surface runoff limitation and precipitation water pretreatment
- Establishment of small-scale water retention facilities
- Precipitation water management in urban areas
- Groundwater remediation

Not all cases require a full range of actions; it is not always and not everywhere that implementation of all assumptions will be possible. That is why the recultivation process ought to be implemented in stages. First and foremost, the most vital actions need to be taken, i.e., those that will yield optimum natural results, and whose implementation is acceptable to other water users. Of particular importance is, however, the development of strategic action plans that ought to outline both further actions aimed at bringing waters close to nature and result monitoring of the operations performed, complete with reactions adequate to the results obtained.

A complex approach to the problem of river recultivation is renaturization, understood as restoring the river, previously regulated or reshaped in another way, into a state close to the natural (existing before the regulation or found in nature). At times, the notion of river revitalization is used to describe these actions. Yet, it covers a narrower scope of actions than renaturization and is conceived as restoring ecological functions to the river, without technical operations to restore the natural characteristics of the riverbed, riverbank, or valley.[33,34]

One of the major objectives of renaturization is to restore diversified riverbed morphology to regulated rivers, which may be obtained, e.g., through shaping a more curvilinear river course, cross-sections varying in shape and dimensions, or the introduction of diversified vegetation. Most frequently, this is a long-lasting process, comprising a variety of technical undertakings, such as

- Anti-erosion drop replacement with slides enabling communication of water organisms as well as elimination of water-tight defenses.
- Works, which—upon completion—do not produce ready-made elements of renaturized water but foster a natural process that can restore naturalness to waters (e.g., vegetation planting).
- Maintenance works consisting in minor corrections to natural transformations when these are not progressing toward a state close to the natural; replacement of fallen trees and bushes.
- Water maintenance during the renaturization process (maintenance and conservation procedures, water quality protection and enhancement).
- Discontinuation of selected actions concerning water management and leaving them to natural impact only.[33–38]
- A river functionality analysis suggests that a significant role in river naturalness restoration is played by vegetation. As part of renaturization operations, the following are implemented: riparian plant structure, tree planting in riparian zones and flood areas, introduction of vegetation into the existing riparian technical structures (the so-called revitalization of existing structures or replacement with plant defenses).[36]

Conducted recultivation procedures, both technical and biological, constitute a long-lasting and difficult-to-design process whose course requires expert supervision, including result monitoring. Hence, riparian vegetation that improves riverbank stability and carries high natural and landscape importance ought to be distributed in stories comprising trees, bushes, perennial grasses, and gramineous plants already found on the site, which meet the necessary requirements. New replacements and seedings of species from nearby habitats are also needed. Tall vegetation ought to be rarefied so that gramineous plants receive enough light. The interface of the tree-covered area and the river ought to be as long as possible, taking tree thinning into account.[27–29]

It must be stated that, when selecting flora and fauna elements, local species are recommended, as their functioning is conditioned by local hydraulic and hydrological determinants. In the spatial distribution of vegetation, they may fulfill natural functions (e.g., those related to ornithology). The introduction of elements from beyond a given ecological system may cause unwanted disturbance in the environment. Moreover, the plants selected ought not to display antagonisms.[2,5,36]

The elongation of river bankline fosters the dovetailing of the riverbed with the adjacent land strip. Over newly formed water habitats, vegetation plantings accelerate the watercourse's integration with the landscape. This is of particular significance in urban areas. In addition, slides situated in the bottom to break up the watercourse gradient, together with islands and boulders to disturb the flow, serve to diversify the current. Hence, they enable the enrichment of the streambed structure and the formation of various minor biotopes for the animal world.[36–38]

The objective of technical actions is to eliminate obstacles that prevent or limit the course of natural fluvial processes. These works are meant to initiate a return to conditions approximating the natural, and affect both abiotic and biotic conditions.[36]

Upon removal of obstacles to shaping morphological conditions in accordance with natural laws, the second stage of renaturization follows. This is a spontaneous transformation implemented by nature as, due to succession and natural transformation, the environment comes to approximate the natural. Most frequently, this spontaneous process comprises riverbed transformations and shaping hydromorphological conditions, as well as vegetation and animal organism development. It is long lasting and may occasionally take from 10 to 15 years, or even up to 30 years.

Recultivation through renaturization ought to make allowances for the multifunctionality of rivers and valleys.[37,39] The chances for implementation of renaturization procedures in rivers increase when these actions, apart from enhancement of the natural state of the rivers, bring other benefits, including economic, municipal, recreational, etc. A case in point may be the improvement of flood control by restoration of the valley retention function. The municipal aims of renaturization may be, e.g., the enhanced self-cleansing capacity of the waters, as well as upgrading the protection status of particular infrastructure. The objective of renaturization in urban areas may be, in turn, the enhancement of a city's (and its surroundings') landscape value, thanks to the restored naturalness of existing water bodies and through the establishment of new ones. These actions are desirable in many cities; however, at the same time, they encounter much difficulty posed by urbanization.[36,40–43]

The most effective, from the natural viewpoint, are renaturization actions conducted in riverbeds and riparian zones in order to increase the diversity of riverbed morphology. This diversity is among the most significant natural features of rivers. It results from water and debris movement as well as the closely related phenomena of erosion and sedimentation. Increased morphodynamic activity may be achieved through elimination or limitation of factors that restrict the process: changes to the river course, riparian defenses and the long profile, or inducing the removal of obstacles disturbing the river debris movement regime.[36,37]

The recreation of the natural shape of the river course (altered as a result of river regulation) is rarely possible as ox-bow lakes are usually buried or permanently cut off from the regulated river. Due to these reasons, as part of renaturization procedures, reclamation usually comprises only fragments of ox-bow lakes, or provision of organisms with opportunities for mobility between ox-bow lakes and the main riverbed. Another option is that a regulated (straightened) river receives a curvilinear shape, resembling those found in natural rivers.[36,39–45]

However, even this kind of local recovery of the river course is very often impossible and the restoration of naturalness is limited to a reconstruction of structures characteristic of natural rivers, i.e., bays, headlands, islands, and local barriers that affect velocity distribution, leading to enhancement of riverbed morphology. This pertains, in particular, to large regulated rivers in which renaturization operations may only cover segments of the riverbank, elimination or replacement of riparian defenses, clearing lateral river arms, renewal of intergroyne space, or partial reclamation of floodplains.[10,14,30]

A significant element in restoration of good ecological quality to rivers is to increase the diversity of crosssection forms. Under natural conditions, this diversification results from the complex hydraulic structure of a water stream as it runs along a curving route. If the restoration of the curvilinear river system is deemed impossible, in order to diversify bottom topography, facilities that steer the water stream (deflectors, digger logs, or appropriately shaped drops) are introduced. Diversification of cross-section width and placement of barriers in the riverbed to enable changing the flow velocity, thus enhancing the diversity of the riverbed morphology, are also practicable.[35–40]

The diversified river banklines thus obtained enable the expansion of the interface between the substratum and water, causing the appearance of a variety of biotopes: a system of fissures in the watercourse gravel substratum, the water area, the interface between water and land, the water level fluctuation zone, the riparian area, and other humid habitats.

The research conducted suggests that, as a result of riverbed renaturization, in the cases of minor lowland rivers with sandy bottoms, where riparian impact significantly affects flow resistance, flow resistance increases. This changes several parameters characterizing water flow and debris drag transport conditions. If riverbed dimensions and flow regime do not change, the increase in scarp coarseness will cause increased falls of the water plane, riverbed filling, and stream impact on the bottom. At constant flow intensity, the drag rate will increase and cause deep riverbed erosion. In view of the above, in river renaturization projects, it is necessary to check not only the riverbed capacity but also the debris drag conditions.[33–40]

To maintain and enhance the ecological quality of rivers, regardless of renaturization procedures, other actions also ought to be taken. A case in point may be adequately conducted river maintenance and conservation works.

These works ought to take into account major requirements of natural protection. Special attention must be paid to the methods used in technical solutions for river development and use. The principles of environmentally friendly river regulation (natural river regulation) ought to be implemented, basing on the preservation of riverbed morphological diversity, through restricted riverbed transformation. This results in the limitation of adverse effects of planned actions on the natural environment. Natural regulation solutions are based on the following principles[36,38,40–49]:

- The river route ought to remain unchanged, and crosssections ought to preserve their diverse forms and dimensions.
- River rerouting ought to be seen as an exception, which may be justified only by very important arguments and lack of alternative solutions.
- River bed structures, i.e., islands, slower flow areas, river outwash, silts, and other riverbed diversifications, should not be eliminated.
- Riverbanks ought to be reinforced only in those segments where the occurring erosion may cause major hazards.

The application of environmentally friendly river regulation principles enables obtaining a compromise solution that will ensure the protection of the most valuable natural resources as well as the fulfillment of major economic goals.

Yet, if riverbed regulation is needed, the solutions should be based on the following principles[36,38,40–49]:

- Natural riverbed protection: rechanneling is done through initializing the meander pool by imposing a triangular section.
- Smoothing natural riverbed meanders by means of linear structures with cross-beams: dams are replaced with river walls that only reinforce concave banks and are pitched against the existing banks (apart from cases when this solution cannot be used).
- Meandering, including riverbank walls or short structures resembling groyne heads (only for riparian protection, low impact on riverbed flows and the riverbed itself).
- Transverse structures that stabilize the riverbed bottom: made as gentle-sloped erratic riffles (1:10 is the optimum gradient, tested in practice), thus preserving the biological passability of the

watercourse; if possible, the location of the riffle is selected so that it can also act as a cataract and not only as bottom gradient reducer and stabilizer.
- Stone backfill with appropriate scarp inclination from 1:2.5, the recommended ratio being 1:4 and lower (wherever possible): more capacious watercourse bed, accessible to people and animals; regulative structures are more durable.
- Preservation of the riverbed's and the river bottom's morphological form: construction of stone backfill from the meander pool bottom (elimination of the necessity to disturb the natural gravel bottom shield during the formation of partitions from these alluviums; the structures are made without partitions).
- Stone backfill durability and visual appeal enhancement: boulder placement and turfing initialization by soil filling of interboulder space and grass seeding according to pertinent formulas.
- Application of defenses made of stone netting baskets: only for protection of building structures situated in the vicinity of the watercourse bank.
- River bar or island leveling or correction—riverbank reinforcement.

As part of the recultivation process, it is necessary to take action in the field of precipitation water management, which is important in view of the current change in climate, and increasing freshwater use and degradation. The construction of retention and infiltration basins enables establishment of the so-called small-scale water retention facilities at the sites where precipitation occurs.

Enhancement of individual forms of retention may be obtained through a variety of methods. Briefly speaking, technical, planning, and agrotechnical procedures may be distinguished. The group of technical methods includes the majority of hydrotechnology and melioration works, aimed at limitation of surface water runoff, e.g.,[5,37,39]

- Surface water retention through the construction of small water reservoirs; lake damming; and construction of damming structures in watercourses, ditches, and channels.
- Regulation of water runoff from drainage systems and ditch networks.
- Increase in underground water reservoir supply through the construction of infiltration ponds and wells.
- Application of adequate methods for rainwater removal from sealed surfaces (roofs, squares, and streets), enabling water infiltration in adjacent nonsealed areas.
- Limitation of rapid surface water runoff through renaturization of small watercourses and reconstruction of floodplains where possible due to economic (farming) reasons.

A significant role in water management, including water quality protection and enhancement, is played by relevant spatial arrangement in rural areas, particularly in terms of formation of such systems in which precipitation and meltwater runoff does not occur rapidly. The methods used in this respect are defined as planning. These include, but are not limited to the following[26–29,30–39]:

- Development of an appropriate arrangement of arable fields, green areas, and woodland in the catchment.
- Peat land, swamp, and wetland protection and reconstruction.
- Formation of protective vegetation strips (bushes and trees: increased catchment woodiness), reconstruction of as many as possible ecological sites, including aquatic gardens.
- Reconstruction of floodplains in river valleys.
- Establishment of protective areas with appropriate facilities.
- The application of appropriate agrotechnical methods, including observance of the Code of Good Agricultural Practice, may contribute to improvement and reconstruction of natural conditions in rivers. The basic actions in this respect, which form the so-called group of agrotechnical methods, include but are not limited to the following[45,48–50]:
- Soil retention increase through soil structure enhancement and soil humus content increase (correct tillage, agromelioration procedures, fertilizing, and liming).

- Surface runoff limitation through erosion control and aftercrops planting.
- Evapotranspiration decrease through appropriate plant selection and limitation of soil surface evaporation.

The suggested methods for catchment retention capacity enhancement do not introduce any significant change to the natural water regime, but only necessary corrections aimed at improving the water balance structure without disturbing the ecosystem's biological balance. Also in terms of precipitation water management, complex solutions are recommended, which comprise urban area drainage, including roofing, downspouts, and underground pipes, thus providing for rapid and effective rainwater removal to a retention and infiltration box system.[27–29,32,50–54] There are several solutions available in the field of rational rainwater removal:

1. Infiltration: used in permeable and poorly permeable soils, with the maintenance of the minimum 1-m distance from the bottom of the box, which is wrapped with non-woven geotextile to the groundwater level; the water infiltrates to the soil; the most popular solution in rainwater management.
2. Retention: used in firm grounds and at high groundwater levels; the water is stored for subsequent runoff to storm drains or waterways; the box system is wrapped with a geomembrane, which prevents water infiltration into the soil.
3. Storage: an alternative solution for open surface reservoirs; precipitation water is stored in boxes wrapped with geomembrane, and may be used later, e.g., as fire water supply.

The variety of methods and techniques that may be used in the recultivation process enables their adaptation to individual needs and objectives (Table 1).

The works in the field of watercourse restoration are subject to numerous technical, legal, administrative, economic, and natural constraints. Under the circumstances, a compromise is needed between approximating nature and achieving economic results. If a river's naturalness has been lost, for instance, owing to the building of hydrotechnical facilities, it cannot be fully restored without the demolition of these facilities. Consequently, that would mean the community's return to conditions that, in the past, made living difficult (floods or swamping) to the extent that led to decisions being taken to improve the situation. While undertaking recultivation procedures, it is necessary to bear in mind that the scope will be limited in proportion to what might be implemented if there is no need to meet the economic requirements. That means that the recovery of natural values in the scope in which they are found in rivers untampered with by humans will not be achieved.

Serious limitations for river recultivation may also result from lack of space for appropriate actions. "Human nature" conflicts most frequently arise in situations when morphological diversification of the riverbed is intended, including addition of a curve to a straightened river; riverbed diversification through the introduction of changeable bottom widths or scarp gradients; or the introduction of islands, headlands, bays, breaches, precipices, and beach segments. The reason is that these require the taking of land adjacent to the river and disturbing the riverbanks, whose shaping is perceived differently by land owners and by ecologists. There is little chance of reaching a compromise on these issues, so the best solution would be to purchase the strip of riparian land where appropriate actions, such as river bankline diversification or allowing for riverbank erosion, may be taken.

Recultivation actions may also be constrained by technical reasons. The restoration of natural elements to rivers, including cutoff meanders, ox-bow lakes, or connections between the river, land depressions, and marshes, will only become possible if, after the regulation, these structures have not been completely demolished and the land has not been developed. Relatively frequent are cases when the existing, necessary riverbed facilities as well as the ways in which they are used (e.g., the river running across a built-up area, a walled riverbed, a link between the river and the water economics system) render transformation of rivers into more natural ones completely unfeasible or feasible only to a small extent.

TABLE 1 Sample Procedures Geared at River Restoration with Their Characteristics

Method/ Technique	Model Application	Characteristics	Literature
Dredging	Buffalo River	Primarily relies on natural sedimentation processes after dredging to achieve dredge residual performance standards established during detailed design; remediation area was approximately 164 acres, including 138 acres in the Buffalo River and 26 acres in the City Ship Canal	[55]
	Eel River	Excess sand and all peat excavated during site restoration work were placed in the former sand borrow pits around the perimeter of the bogs; insecticide-affected peat was buried beneath a variable layer of sand; once filled—sand borrow pits into the surrounding upland forest; a layer of top soil was imported and placed as a final cap over the former borrow pits, revegetated with native plantings, and monitored over time to ensure growth and survival	[56]
Capping	Buffalo River	Introduced for remediation of the end of the City Ship Canal, beyond the limits of the authorized navigation channel, to isolate underlying sediment contaminants and provide a clean sediment surface; capping depths and cap materials were designed to optimize and enhance habitat restoration plans while providing adequate protection against damage from root penetration; cap placement was performed by either of the following methods: 1) extending a navigational channel to the downstream end; 2) through hydraulic means; or 3) in dry conditions using earth moving equipment, by temporary sheeting and dewatering the proposed cap area and using the adjoining upland areas for material handling	[55]
Renaturization	Izara River	Along a 14-km-long segment, the riverbank structures were removed; there appeared gravel river bars, shoals, main riverbed forks, and stagnant waters; in selected segments, modern fish ladders were built and riparian communities originated	[57]
	Mistelbach Stream	Concrete gutters were removed from stream bottom and banks; so was the irregular stream bank shaping; islands and widenings were installed; stones were introduced into the streambed bottom; slides (ramps) were built and planted with vegetation	[58]
	Enz River	Irregular widenings were formed (at two sites, forking the river by means of islands); the bottom shape was diversified, the scarp inclination of 1:4–1:5 was obtained; the watercourse gradient was increased to 2.7%; dead willow fascine was used; structures were erected with the use of reed beds; plantings of tree shoots were inserted into the substratum, and seedings were made (vegetation was mostly obtained from floodplains and nearby areas)	[59,60]
	Kwacza River	Riparian alders were felled; six bypasses were made for the existing riverbed; in the riverbed, fallen trees were retained, providing valuable natural microhabitats; in the current, artificial structures in the forms of stone piles were placed; wooden and stone deflectors were made, as well as wooden log drops and in-stream digger logs made of tree trunks (deflectors); riverbank defenses were made of tree trunks; the connection between the major riverbed and the ox-bow lake was reconstructed; bays and stagnant water pools were formed; several hundreds of tons of gravel and stones were used to increase bottom material graining so that its composition would approximate the natural salmonid spawning sites	[61]
	Skjern River	Elimination of hydrotechnical structures, restoration of the river's meandering course, and purchase of farming land adjacent to the river	[62]

(*Continued*)

TABLE 1 (*Continued*) Sample Procedures Geared at River Restoration with Their Characteristics

Method/ Technique	Model Application	Characteristics	Literature
Rehabilitation	River Frome	Dig a borrow pit to create a supply of clay; use clay to raise the riverbed level (specific depth to be agreed on site); win gravels from embankments (old river dredgings or borrow pit); return gravel to the river creating riffles and gravel features; spoil created from bank reprofiling used to infill the borrow pit (less material available—in the original pit volume, a small pond or wetland was created), giving varied habitats	[63]
	River Rother	Restore the loop to these dimensions tempering them with uncovered in-channel features (basic river archaeology using remnant channel features as a guide in favor of theoretical section design); fix the channel features at the upstream and downstream end of the system to prevent channel adjustment between the canalized section and the restored naturalized section (headward recession, etc.); restore the effect of the lock gates and system to prevent any passage of flow down the canal and divert all flows through the loop	[64]
Restoration	River Avon	Physical features and geomorphological processes of the watercourse habitat were re-established at six river restoration demonstration sites; approximately 7 km of river and banks was enhanced and woody debris introduced; an estimated 0.36 ha of new spawning area for Atlantic salmon, bullhead, and lamprey was created (Atlantic salmon are already spawning on the new gravel)	[65]
Restoration of native terrestrial plants	Niagara River	Restore extirpated species to gorge; close unauthorized trails; upgrade existing trails with water bars, fencing, signs; initiate trail steward program; selectively remove aggressive nonnative species from the vicinity of gorge and gorge rim; use flow restriction or diffusion devices on outfalls; eliminate perched outfalls; restore water flow to Devil's Hole; reduce acidification; restore original character of cove; exclude birds (removal of nests, egg oiling); reduce soil acidification; reduce direct trampling or pulling of plants by gulls; reduce input of alien seeds in material brought in by gulls	[66]
Remediation	Sumida River	Waterfront amenity improvement coexisted with flood prevention; gently sloping levees and high standard levees were constructed as substitutes for linear levees to increase durability; river edge terraces (promenades) were improved and levees were forested to improve water amenities; improvement of the sewage system and activities of citizens' groups	[67]
Revitalization	River Brent	Restored two sections of this river by remeandering the straightened channel along its original route, creating a backwater channel, and naturalizing the river's banks; linked up with the earlier restoration work to enhance the entire park	[68]
	River Thames	Activities covered by the tributaries were carrying pollutants, e.g., creation of a natural channel in place of a buried culvert; creation of sustainable instream, marginal, wetland, and floodplain habitats; potential for improved biodiversity in the area and increased flood storage capacity; reconnected the two stretches of seminatural channel up- and downstream of Brookmill Park	[69,70]

Conclusion

The recultivation of flowing surface waters may be based on tested case studies from other regions of the country or abroad. Yet, it should be remembered that each environment, while similar in terms of the problems encountered, is unique. The adaptation of solutions must then be carried out with due care.

More and more often, an ideal model of a watercourse is taken as a starting point during the recultivation planning process. The model corresponds to natural conditions of the watercourse, unchanged by humans. It is only on that basis that its current state, and deficits, or degradation rating, are assessed. Subsequently, taking into account socioeconomic boundary conditions, such as the degree of valley occupancy, navigation, fire safety, and environmental protection requirements, development goals are set, i.e., the character of the watercourse after recultivation is determined. The actions taken are based on a variety of methods and techniques, individually adapted to the local situation and the specified objectives. They include technical, engineering, biological, and hydrological procedures. Nonetheless, the scope of the works conducted must demonstrate multitasking and be capable of complex inclusion of selected environmental elements, at the same time stimulating their reconstruction process.

It must be stressed that recultivation procedures do not usually yield direct economic benefits, which constitutes another obstacle. The improvement of a river's natural condition or enhancement of landscape values in themselves are not sufficient arguments in favor of beginning the recultivation.

Such actions, understandable and convincing from the point of view of the natural environment and its needs, tend to breed substantial controversy if local economic expectations are considered. It is then vital to recognize hydrological and hydraulic conditions, and describe the phenomenon of water supply to the land subject to recultivation procedures, together with recognition of farming water needs in the area. Social and sociological surveys that would answer the questions related to the intentions of the farmers themselves, the prospects for local agriculture, possibility of land sale, etc., are also needed.

The chief obstacle is indeed found in the mentality of local communities who expect the proprietor of flowing waters to fulfill the protective obligations. These include deepening and narrowing the riverbed and the streambed so that they no longer flood; reinforcing riverbanks and securing private facilities built nearby; and creating or establishing conditions for recreational river use. Yet, the responsibility to protect the environment as our shared asset remains a shared obligation.

Acknowledgments

The present study has been financed in part by the State Committee for Scientific Research (KBN), grant no. N N305 306635.

References

1. Petts, G.; Calow, P. *River Restoration*; Blackwell Sciences: Oxford, U.K., 1996.
2. Rao, R.J. Biological resources of the Ganga River, India. Hydrobiologia **2001**, *458,* 159–168.
3. U.S. Army Corps of Engineers. *Environmental Design Handbook*; Upper Mississippi River System Environmental Management Program: Rock Island, IL, 2006.
4. Theiling, C.H. Defining ecosystem restoration potential using a multiple reference condition approach; Upper Mississippi River System, USA, Theses and Dissertations, University of Iowa, 2010.
5. Zalewski, M.; Harper, D.M.; Robarts, R.D., Eds. Proceedings of the International UNESCO Workshop "Aquatic Habitats in Integrated Urban Water Management" within the framework of the International Hydrological Programme (IHP) and the Man and the Biosphere (MAB) Programme, Ecohydrology & Hydrolobiology. **2005**, 5(4), pp. 93.
6. European Parliament and the Council. Directive 2000/60/ CE of the European Parliament and of the Council of 23 October 2000 establishing a Framework for Community Action in the Field of Water Policy; Official Journal of the European Communities, L 327: Luxembourg, 2000.
7. Phillips, N.; Bennett, J.; Moulton, D. Queensland Environmental Protection Agency, *Principles and Tools for Protecting Australian Rivers,* Land & Water Australia, Canberra 2001.
8. Jasperse, P. Policy networks and the success of lowland stream rehabilitation projects in the Netherlands. In *Rehabilitation of Rivers. Principles and Implementation*; Waal, L.C., Large, A.R.G., Wade, P.M., Eds.; John Wiley & Sons; Chichester, U.K., 1998, 14–29.

9. Redondo, M.D. Social impact assessment for river restoration: A more sustainable perspective, Thesis presented in part-fulfillment of the degree of Master of Science in accordance with the regulations of the University of East Anglia, Norwich, 2003.
10. Schanze, J.; Olfert, A.; Tourbier, J.T.; Gersdorf, I.; Schwager, T. Existing Urban River Rehabilitation Schemes (Work package 2), Final Report, Urban River Basin Enhancement Methods, 2004.
11. Gore, J.A.; Shields, F.D. Can large rivers be restored? BioScience **1995**, *45* (3), 142–152.
12. Brierley, G.; Fryirs, K.; Outhet, D.; Massey, C. Application of the River Styles framework as a basis for river management in New South Wales, Australia. Appl. Geogr. **2002**, 22, 91–122.
13. Wagner, I.; Marsalek, J.; Breil, P. *Aquatic Habitats in Sustainable Urban Water Management, Science, Policy and Practice;* Urban Water Series—UNESCO-IHP, Paris, France; Taylor & Francis: Leiden, the Netherlands, 2008
14. Stalzer, W. Institutional Challenges in the Danube River Basin: Inputs from the International Commission for the Protection of the Danube River, MRC International Conference, 2–3 April 2010; Hua Hin, Thailand.
15. Reeve, D.E. Predicting morphological changes in rivers, estuaries and coasts, Executive Summary, Report number T05–07-03, Members of the FLOOD*site* Consortium, University of Plymouth, 2009.
16. Nakamura, K.; Tockner, K.; Amano, K. River and wetland restoration. Lessons from Japan. BioScience **2006**, *56* (5), 419–429.
17. River Restoration Project, Texas State University, San Marcos, http://www.aquarena.txstate.edu/Wetlands/Restoration-grants.html (accessed January 2011).
18. ARRN's Roundtable Meeting, How to Develop Technology and Guideline for River Restoration through Networks, ARRN Asian River Restoration Network; Foundation for Riverfront Improvement and Restoration and CTI Engineering Co., Ltd.: Seoul, Korea, 2011.
19. Groundwork Thames Gateway London South, Feasibility Study, Chinbrook Meadows River Naturalisation Scheme; London Borough of Lewisham: London, 2001.
20. Wheaton, J.M. *Review of River Restoration Motives and Objectives;* unpublished review, Southampton, U.K., 2005.
21. Herricks, E.E. Water quality issues in river channel restoration. In *River Channel Restoration: Guiding Principles for Sustainable Projects;* Brookes, A., Shields, F.D., Eds.; John Wiley & Sons: Chichester, U.K., 1996; 179–201.
22. River Restoration Centre. *River Restoration Manual of Techniques,* 1999, http://www.therrc.co.uk/rrc_manual.php (accessed May 2012).
23. U.S. Army Corps of Engineers. *Upper Mississippi River System Environmental Design Handbook;* U.S. Army Corps of Engineers: Rock Island, IL, 2006.
24. Bolton, S.; Shellberg, J. *Executive Summary: Ecological Issues in Floodplains and Riparian Corridors*, White paper prepared for Washington Department of Fish and Wildlife, Washington Department of Ecology and Washington Department of Transportation, Seattle, Washington; University of Washington, Center for Streamside Studies, Seattle, Washington, 2001.
25. Rhoads, B.L.; Herricks, E.E. Naturalization of headwater streams in Illinois: Challenges and possibilities. In *River Channel Restoration: Guiding Principles for Sustainable Projects;* Brookes, A., Shields, F.D., Eds.; John Wiley & Sons: Chichester, U.K., 1996; 331–369.
26. Sear, D.A. The sediment system and channel stability. In *River Channel Restoration: Guiding Principles for Sustainable Projects;* Brookes, A., Shields, F.D., Eds.; John Wiley & Sons: Chichester, U.K., 1996; 149–179.
27. De Waal, L.; Wade, P.M.; Large, A., Eds.; *Rehabilitation of Rivers. Principles and Implementation*; John Wiley & Sons: London, 1998.
28. Jormola, J., Ed. *Urban Rivers, Flood Risk Management and River Restoration,* Gumiero, B.; Rinaldi, M.; Fokkens, B., (Eds), 4th ECRR Conference on River Restoration, ECRR - European Centre for River Restoration, CIRF - Centro Italiano per la Riqualificazione Fluviale, Venice, San Servolo Island, 2008; Chapter 12, pp. 889–948.

29. Wisconsin DNR. *Wisconsin Water Quality Report to Congress,* WDNR Pub WT-924–2010, 2010.
30. Bojarski, A.; Jeleński, J.; Jelonek, M.; Litewka, T.; Wyżga, B.; Zalewski, J. *Good Practice in Maintaining Mountain Rivers and Streams*; Ministry of the Environment, Department of Water Resources: Warsaw, Poland, 2005 (in Polish).
31. Hulse, D.; Gregory, S. Integrating resilience into floodplain restoration. Urban Ecosyst. **2004**, *7,* 295–314.
32. Jones, M.; Bradley, W.; Dunsford, D. An integrated approach to flood defence design. River Restor. News Newsl. River Restor. Centre **2002**, *12,* 5.
33. Kondolf, G.M.; Downs, P.W. Catchment approach to planning channel restoration. In *River Channel Restoration: Guiding Principles for Sustainable Projects*; Brookes, A., Shields, F.D., Eds.; John Wiley & Sons: Chichester, U.K., 1996; 129–149.
34. Pedroli, B.; de Blust, G.; van Looy, K.; van Rooij, S. Setting targets in strategies for river restoration. Landsc. Ecol. **2002**, *17* (Suppl. 1), 5–18.
35. Darby, S.; Sear, D. *River Restoration, Managing the Uncertainty in Restoring Physical Habitat;* John Wiley & Sons, Ltd., 2008.
36. Doratli, N.; Hoskara, S.O.; Fasli, M. An analytical methodology for revitalization strategies in historic urban quarters: A case study of the Walled City of Nicosia, North Cyprus. Cities **2004**, *21* (4), 329–348. doi: 10.1016/ j.cities.2004.04.009.
37. Żelazko, J. Rivers and valleys restoration, infrastructure and ecology of the countryside, Polish Academy of Sciences, Cracow Branch, Technical Committee on Rural Infrastructure, 2006, No. 01/04/2006, 11–31.
38. Davis, J.R. Revitalization of a northcentral Texas river, as indicated by benthic macroinvertebrate communities. Hydrobiologia **1997**, *345,* 95–117.
39. Osugi, T.; Tate, S-I.; Takemura, K.; Watanabe, W.; Ogura, N.; Kikkawa, J. Ecological research for the restoration and management of rivers and reservoirs in Japan. Landsc. Ecol. Eng. **2007**, 3, 159–170.
40. Ryder, D.; Boulton, A.; De Deckker, P. Setting goals and measuring success: Linking patterns and processes in stream restoration. Hydrobiologia **2005**, 552, 147–158.
41. Williams, C.A.; Cooper, D.J. Mechanisms of riparian cottonwood decline along regulated rivers. Ecosystems **2005**, *8,* 382–395.
42. Wang, W.; Tang, X.Q.; Huang, S.L.; Zhang, S.H.; Lin, C.; Liu, D.W.; Che, H.J.; Yang, Q.; Scholz, M. Ecological restoration of polluted plain rivers within the Haihe river basin in China. Water Air Soil Pollut. **2010**, *211,* 341–357.
43. Zhang, L. Ecological water requirement based on ecological protection and restoration targets in the lower reaches of the Heihe River, northern China. Front. For. China **2009**, *4* (3), 263–270. doi: 10.1007/s11461-009-0052-0.
44. Buijs, A.E. Public support for river restoration. A mixed- method study into local residents' support for and framing of river management and ecological restoration in the Dutch floodplains. J. Environ. Manage. **2009**, *90,* 2680–2689.
45. Bar-Or, Y. Restoration of the rivers in Israel's coastal plain. Water Air Soil Pollut. **2000**, *123,* 311–321.
46. Willis, K.G.; Garrot, G.D. Angling and recreation values of low-flow alleviation in rivers. J. Environ. Manage. **1999**, *57* (2), 71–83.
47. Lüderitz, V.; Jüpner, R.; Müller, S.; Feld, C.K. Renaturalization of streams and rivers—The special importance of integrated ecological methods in measurement of success. An example from Saxony-Anhalt (Germany). Limnology **2004**, *34,* 249–263.
48. Nowak, M.C. *Burnt River Subbasin Plan, DRAFT.* Cat Tracks Wildlife Consulting. Prepared for Northwest Power and Conservation Council, 2004.
49. Arlettaz, R.; Lugon, A.; Sierro, A.; Werner, P.; Kéry, M.; Oggier, P.-A. River bed restoration boosts habitat mosaics and the demography of two rare non-aquatic vertebrates. Biol. Conserv. **2011**, *144,* 2126–2132.

50. Junker, B.; Buchecker, M. Aesthetic preferences versus ecological objectives in river restorations, Landsc. Urban Plan. **2008**, *85,* 141–154.
51. Ivannikov, F.A.; Prokofieva, T.V. Artificial soil-like bodies of a river valley and their evolution in an urban environment (based on the example of the Moscow river valley). Moscow Univ. Soil Sci. Bull. **2010**, *65* (4), 145–150.
52. Pollard, P.; Devlin, M.; Holloway, D. Managing a complex river catchment: A case study on the River Almond. Sci. Total. Environ. **2001**, *265,* 343–357.
53. Rohde, S.; Kienast, F.; Bürgi, M. Assessing the restoration success of river widenings: A landscape approach. Environ. Manage. **2004**, *34* (4), 574–589.
54. Connelly, N.A.; Knuth, B.A.; Kay, D.L. PROFILE, public support for ecosystem restoration in the Hudson River Valley, USA. Environ. Manage. **2002**, *29* (4), 467–476. doi: 10.1007/s00267-001-0033-Z.
55. Jun, K.S.; Kim, J.S. The Four Major Rivers Restoration Project: Impacts on river flows. KSCE J. Civil Eng. **2011**, *15* (2), 217–224. doi: 10.1007/s12205-011-0002-x.
56. Feasibility Study by Buffalo River, New York, Draft Final. ENVIRON International Corporation, MACTEC Engineering and Consulting, Inc., LimnoTech, 2010.
57. Eer River Headwaters Restoration Project, Phase 1—Cranberry Bog and River Restoration, Massachusetts Department of Environmental Protection, Bureau of Resource Protection—Wetlands and Waterways, Plymouth, 2008.
58. How to reconcile the interests of man and nature? The Regional Water Management Board in Krakow, available at http://www.krakow.rzgw.gov.pl/ (accessed May 19, 2011) (in Polish).
59. Chardon, W.J.; Koopmans, G.F., Eds. Critical evaluation of options for reducing phosphorus loss from agriculture. IPW 4—Proceedings of the 4th Int. Phosphorus Workshop, Wageningen, the Netherlands, 2004.
60. Larson, P. Restoration of river corridors: German experiences. In: *River Restoration;* Petts, G.E., Calow, P., Ed.; Blackwell: Cambridge, U.K., 1996; 124–143.
61. Meixner, H.; Schnauder, I.; Bölscher, J.; Iordache, V., Eds. Hydraulic, sedimentological and ecological problems of multifunctional riparian forest management RIPFOR, guidelines for end-users. Berl. Geogr. Abh. **2006**, 66, 1–168.
62. Obolewski, K.; Glińska-Lewczuk, K.; Osadowski, Z.; Kobus, S. Water quality monitoring of agricultural River catchments In the example of the Kwacza River Valley (Middle Slupia River basin). Polish Journal of Environmental Studies, **2008**, *17* (2c), 39–46.
63. Alwan, A.A.; Appiah-Kubi, G.; Majland-Kristensen, P. *The Possible Impact of the Restoration of River Skjern;* Environmental Studies; University of Aarhus: Aarhus, Denmark, 2001.
64. River Frome Rehabilitation Plan, Environment Agency, Newsletter No. 2, 2010.
65. Bramley, M. Development of Engineering in River Restoration. River Restor. News Newsl. River Restor. Centre **2002**, *12*, 2–3.
66. STREAM—River Avon cSAC demonstrating strategic restoration and management LIFE05 NAT/UK/000143, STREAM Newsletter no. 7, 2007.
67. Niagara Power Project (Ferc No. 2216), Feasibility Study for the Restoration of Native Terrestrial Plants in the Vicinity of the Niagara Gorge, New York Power Authority, 2008.
68. Wada, A. Sumida River, Japan. Japan River Restoration Network JRNN, Panel on Urban Water/River Front Redevelopment, World Bank Tokyo Office, May 30th, 2011.
69. Environment Agency. Bringing Your Rivers Back to Life. *A Strategy for Restoring Rivers in North London*; Environment Agency, Rio House, Waterside Drive, Aztec West, Almondsbury, Bristol, 2006.
70. Environment Agency. The London Rivers Action Plan, A tool to help restore rivers for people and nature, available at http://www.therrc.co.uk/lrap.php (accessed May 2011).
71. Environment Agency. *River Restoration. A Stepping Stone to Urban Regeneration Highlighting the Opportunities in London*; Environment Agency: Bristol, U.K., 2002

53

Waste: Stabilization Ponds

Sven Erik Jørgensen

Application of Waste Stabilization Pond Systems

A more detailed review of waste stabilization ponds (WSPs), their kinetics, and design can be found in Kayombo et al.[1]

WSPs are large, shallow basins in which raw sewage is treated entirely by natural processes involving both algae and bacteria.[2] They are used for sewage treatment in temperate and tropical climates and represent one of the most cost-effective, reliable, and easily operated methods for treating domestic and industrial wastewater. WSPs are very effective in the removal of fecal coliform bacteria. Sunlight energy is the only requirement for its operation. Further, it requires minimum supervision for daily operation, by simply cleaning the outlets and inlet works. The temperature and duration of sunlight in tropical countries offer an excellent opportunity for high efficiency and satisfactory performance for this type of water-cleaning system. They are well suited for low-income tropical countries where conventional wastewater treatment cannot be achieved due to the lack of a reliable energy source. Further, the advantage of these systems, in terms of removal of pathogens, is one of the most important reasons for its use.

Types of WSPS and Their Specific Uses

WSP systems comprise a single string of anaerobic, facultative, and maturation ponds in series, or several such series in parallel. In essence, anaerobic and facultative ponds are designed for removal of biochemical oxygen demand (BOD), and maturation ponds are designed for pathogen removal, although some BOD removal also occurs in maturation ponds and some pathogen removal occurs in anaerobic and facultative ponds.[3,4] In most cases, only anaerobic and facultative ponds will be needed for BOD removal when the effluent is to be used for restricted crop irrigation and fishpond fertilization, as well as when weak sewage is to be treated prior to its discharge to surface waters. Maturation ponds are only required when the effluent is to be used for unrestricted irrigation, thereby having to comply with the World Health Organization guideline of >1000 fecal coliform bacteria/100 mL. The WSP does not require mechanical mixing, needing only sunlight to supply most of its oxygenation. Its performance may be measured in terms of its removal of BOD and fecal coliform bacteria.

Processes in WSPS

Anaerobic Ponds

Anaerobic ponds are commonly 2–5 m deep and receive wastewater with high organic loads (i.e., usually greater than 100 g/m^2 24h, equivalent to more than 3000 kg/ha/ day for a depth of 3 m). They normally do not contain dissolved oxygen or algae. In anaerobic ponds, BOD removal is achieved by sedimentation of solids and subsequent anaerobic digestion in the resulting sludge. The process of anaerobic digestion is more intense at temperatures above 15°C. The anaerobic bacteria are usually sensitive to pH <6.2. Thus, acidic wastewater must be neutralized prior to its treatment in anaerobic ponds. A properly designed anaerobic pond will achieve about a 40% removal of BOD at 10°C and more than 60% at 20°C. A retention time of 1.0–1.5 days is commonly used.

Facultative Ponds

Facultative ponds (1–2 m deep) are of two types: Primary facultative ponds that receive raw wastewater, and secondary facultative ponds that receive particle-free wastewater (usually from anaerobic ponds, septic tanks, primary facultative ponds, and shallow sewerage systems). The process of oxidation of organic matter by aerobic bacteria is usually dominant in primary facultative ponds or secondary facultative ponds. The processes in anaerobic and secondary facultative ponds occur simultaneously in primary facultative ponds, as shown in Figure 1. It is estimated that about 30% of the influent BOD leaves the primary facultative pond in the form of methane.[2] A high proportion of the BOD that does not leave the pond as methane ends up in algae. This process requires more time, more land area, and possibly 2–3 weeks water retention time, rather than 2–3 days in the anaerobic pond. In the secondary facultative

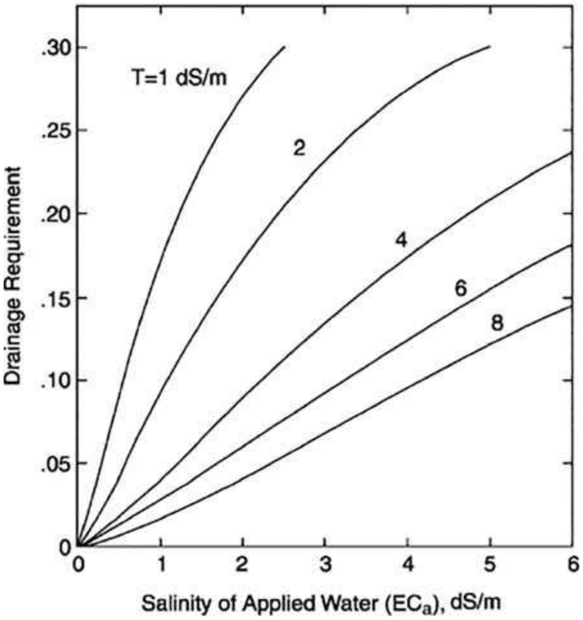


FIGURE 1 Pathways of BOD removal in primary facultative ponds.
Source: After Marais.[2]

pond (and the upper layers of primary facultative ponds), sewage BOD is converted into "algal BOD" and has implications for effluent quality requirements. About 70%–90% of the BOD of the final effluent from a series of well-designed WSPs is related to the algae they contain.

In secondary facultative ponds that receive particle-free sewage (anaerobic effluent), the remaining non-settleable BOD is oxidized by heterotrophic bacteria (*Pseudomonas, Flavobacterium, Achromobacter,* and *Alcaligenes* spp). The oxygen required for oxidation of BOD is obtained from photosynthetic activity of the micro-algae that grow naturally and profusely in facultative ponds. Facultative ponds are designed for BOD removal on the basis of a relatively low surface loading (100–400 kg BOD/ha/day), in order to allow for the development of a healthy algal population, since the oxygen for BOD removal by the pond bacteria is generated primarily via algal photosynthesis. The facultative pond relies on naturally growing algae. The facultative ponds are usually dark green in color because of the algae they contain. Motile algae (*Chlamydomonas* and *Euglena*) tend to predominate the turbid water in facultative ponds, compared to non-motile algae (*Chlorella*).

The algal concentration in the pond depends on nutrient loading, temperature, and sunlight, but is usually in the range of 500–2000 μg chlorophyll-*a* per liter.[3,4] Because of the photosynthetic activities of pond algae, there is a diurnal variation in the dissolved oxygen concentration. The dissolved oxygen concentration in the water gradually rises after sunrise, in response to photosynthetic activity, to a maximum level in the mid-afternoon, after which it falls to a minimum during the night, when photosynthesis ceases and respiratory activities consume oxygen. At peak algal activity, carbonate and bicarbonate ions react to provide more carbon dioxide for the algae, leaving an excess of hydroxyl ions. As a result, the pH of the water can rise to above 9, which will kill fecal coliform. Good water mixing, which is usually facilitated within the upper water layer by strong wind, ensures a uniform distribution of BOD, dissolved oxygen, bacteria, and algae, thereby leading to a better degree of waste stabilization.

Kinetics of the Processes in Facultative Ponds

The growth of the mixed culture was studied at concentrations ranging between 200 and 800 mg COD/L, in a series of batch static reactors. From laboratory data, the specific growth rate (μ) was determined, using the modified Gompertz model.[1,5] There are several growth models in the literature used to evaluate growth parameters, first-order reaction, logistic models, etc. The maximum specific growth rate ($\mu_{\max}$) and half saturation coefficients (K_s) were calculated using the Monod kinetic equation, which is the generally applied expression:

$$\mu = (\mu \text{max})^{*} \, \text{S}/(\text{S} + (\text{K}_\text{s}))$$

S in this equation is the substrate concentration usually expressed as mg/L BOD5 or COD.

The maximum observed growth rate ($\mu_{\max}$) for heterotrophic bacteria was 3.8 L/24 hr, with a K_s of 200 mg/L when COD was applied as unit for *S*.

The $\mu_{\max}$ for algal biomass, based on volatile suspended solids, was 2.7 L/24 hr, with a K_s of 110 mg COD/L. The $\mu_{\max}$ of algae, based on chlorophyll-*a*, was 3.5 L/24 hr with a K_s of 50 mg COD/L. The indicated constants were all found on the basis of literature values that were found from data. The regression coefficients for the constants were all 0.96–0.99 according to the literature.

The observed specific substrate removal by heterotrophic bacteria varied between the concentrations of substrate used, with an average value of 0.82 (mg COD/mg biomass). Thus, the determined Monod kinetic parameters are useful for defining the operation of secondary facultative ponds. The specific substrate utilization rate in the bioreactors was directly proportional to the specific growth rate.

Although the Monod equation was developed for the pure culture of bacteria growing in a single organic substrate, the results may be used correctly also for a mixed culture of algae and heterotrophic bacteria growing in a heterogeneous-mixed organic carbon. The K_s values obtained are the characteristic of the wastewaters containing a complex mixture of organic carbon. The Monod constants for algae depend on the method used in determining the biomass as indicated above.

Knowledge of the effects of the pH on the growth rate of algae and heterotrophic bacteria, and its subsequent impacts on the degradation of organic matter, is useful for better operation and design of secondary facultative ponds. The kinetics of microbial growth and product formations are influenced by diurnal variation in the pH in the pond resulting from diurnal variations in the carbon dioxide.[6] Algae require large quantities of dissolved carbon dioxide during photosynthesis, causing a depletion of carbon dioxide (CO_2) and leading to a shift in the carbonate-bicarbonate ($CO3^{2-}$–$HCO3^{-}$) equilibrium, resulting in an increase in pH due to the formation of hydroxyl (OH^{-}) ions.[7] The effects of pH on the growth rate of heterotrophic bacteria and algae in secondary facultative ponds were investigated, using batch growth at a pH value between 5 and 11. The optimum pH was found to be between 7.0 and 8.0. The results indicate that, for a pH value higher than 8, the chlorophyll-*a* content decreases. However, the specific growth rate of heterotrophic bacteria and algae was high with a pH value between pH 6.5 and 8. At a pH value above 9, the specific growth rates of both biomasses decrease.

Oxygen tension in WSPs is an operational parameter with a great deal of daily and hourly variation. The rate of oxygen production is a function of the algal concentration. Because algal growth is both light and temperature dependent, the rate of oxygen production (photosynthetic) follows the same pattern. Temperature also is a parameter exhibiting marked seasonal and daily variation in WSPs. It influences photosynthesis, the growth of microorganisms, and the bio-decomposition of organic carbon in the system. The fluctuation in pH influences the kinetics of microbial growth, species competition, and product fomations in the pond.[6] Microbial species can grow within a specific pH range, which typically extends over 3 to 4 pH units, with an optimum growth rate near the midpoint of the range. Values of pH up to 11 are not uncommon in WSPs, with the highest levels being reached in the late afternoon. These results lead to the conclusion that the most useful rate that may be used to govern the processes from the WSP should be those determined from the combined influence of the various forcing functions.

Maturation Ponds

The maturation ponds, usually 1–1.5 m deep, receive the effluent from the facultative ponds. Their primary function is to remove excreted pathogens. Although maturation ponds achieve only a small degree of BOD removal, their contribution to nutrient removal can be significant. Maturation ponds usually show less vertical biological and physicochemical stratification, and are well oxygenated throughout the day. The algal population in maturation ponds is much more diverse than in the facultative ponds. The algal diversity generally increases from pond to pond along the series.[3] Although fecal bacteria are partially removed in the facultative ponds, the size and numbers of the maturation ponds especially determine the numbers of fecal bacteria in the final effluent. There is some removal of solids-associated bacteria in anaerobic ponds, principally by sedimentation. The principal mechanisms for fecal bacterial removal in facultative and maturation ponds are now known to be

1. Time and temperature
2. High pH (>9)
3. High light intensity, combined with high dissolved oxygen concentration

Time and temperature are the two principal parameters used in designing maturation ponds. Fecal bacterial die-off in ponds increases with both time and temperature. High pH values (above 9) occur in ponds, due to rapid photosynthesis by pond algae, which consumes CO2 faster than it can be replaced by bacterial respiration. As a result, carbonate and bicarbonate ions dissociate, as follows:

$$2\ HCO_3^- \rightarrow CO_3^{2-} + H_2O + CO_2$$

$$CO_3^{2-} + H_2O \rightarrow 2\ OH^- + CO_2$$

The resulting CO_2 is fixed by the algae, and the hydroxyl ions accumulate, often raising the pH to values even above 10. Fecal bacteria (with the notable exception of *Vibrio cholerae*) die very quickly at pH values higher than 9.[8] The role of high light intensity and high dissolved oxygen concentration has recently been elucidated. Light of wavelengths between 425 and 700 nm can damage fecal bacteria by being absorbed by the humic substances ubiquitous in wastewater. They remain in an excited state sufficiently long to damage the cell. Light-mediated die-off is completely dependent on the presence of oxygen and is enhanced at high pH values. Thus, the sun plays a threefold role in directly promoting fecal bacterial removal in WSP and in increasing the pond temperature, and more indirectly by providing the energy for rapid algal photosynthesis. This not only raises the pond pH value above 9, but also results in high dissolved oxygen concentrations, which are necessary for its third role, namely, promoting photo-oxidative damage.

Nutrient Removal in WSPs

In anaerobic ponds, organic nitrogen is decomposed to ammonia. Thus, the effluent from anaerobic ponds usually has higher concentrations of ammonia than what is found in raw sewage. In facultative and maturation ponds, ammonia is incorporated into algal biomass. At high pH values, ammonia leaves the pond through volatilization. There is little evidence for nitrification (and hence denitrification, unless the wastewater has a high nitrate content).[3] This is due to the fact that the population of the nitrifying bacteria is low because of the lack of physical attachment sites in the aerobic zone. Total nitrogen and ammonia removal from WSP can reach 80% and 95%, respectively.[3] Phosphorus removal in WSP is associated with its uptake by algal biomass, precipitation, and sedimentation. Mara[3] suggested that the best way to remove much of the phosphorus in the wastewater by WSP is to increase the number of maturation ponds. However, both nitrogen and phosphorus must be removed in order to prevent eutrophication in receiving water bodies. The common practice in the design of the WSP is not based on nutrient removal; rather, it is based on BOD and fecal coliform removal.

Design of WSPs

Design Parameters

There are four important design parameters for WSP, including temperature, net evaporation, flow, and BOD. The climate also is important in as much as the processes responsible for BOD5 and fecal bacterial removal are temperature dependent. Further, algal photosynthesis depends on solar insulation, which is a function of latitude and cloud cover. Cloud cover periods are seldom a problem because the solar insulation during the day in tropical and subtropical regions generally greatly exceeds the saturation light intensity of the algae in the ponds. The design temperature usually is the mean air temperature in the coolest month (or quarter). The pond water is usually 2–3°C warmer than the air temperature in the cool season, with the reverse also being true in the warm season.

Because the bacteria responsible for treatment are mesophilic, high temperatures are not a problem. However, low temperatures can be, since they slow down the treatment process. In the case of the methanogenic bacteria (crucial to anaerobic digestion), methane production virtually ceases below temperatures of 15°C. Thus, in areas where the pond temperature remains below 15°C for more than a couple of months of the year, careful consideration should be given to deciding whether or not anaerobic units are needed. Net evaporation (evaporation minus rainfall) must be taken into account during the design of facultative and maturation ponds, but not for anaerobic ponds.[9] Anaerobic ponds generally have a scum layer, which effectively prevents significant evaporation of the water.

Total nitrogen and free ammonia (NH_3, rather than $NH_4^+ + NH_3$) are important in the design of wastewater-fed fish ponds. Typical concentrations of total nitrogen in raw domestic wastewater are 20–70 mg N/L, and total ammonia ($NH_4^+ + NH3$) concentrations are 15–40 mg N/L. Fecal coliform numbers are important if the pond effluent is to be used for unrestricted crop irrigation or for fishpond fertilization. Grab samples of the wastewater may be used to measure the fecal coliform concentration if wastewater exists.

Estimation of Water Flows and BOD Concentrations

The mean water flows should be carefully estimated, since they have direct effects on the size of the ponds and the construction costs. A suitable design is 85% of the in-house water consumption. The BOD may be measured if wastewater exists, based on 24 hr flow weighted data. Alternatively, the BOD may be estimated from the value 30 to 70 g/(capita/24 hr), with rich communities producing more BOD than poor communities. In medium-sized towns, a value of 50 g/(capita/24 hr) is more suit-able.[10,11] A typical design figure for an urban area in a developing country would be 40–50 g BOD_5/(capita/24 hr).[9] Although it is dangerous to generalize, in view of the wide variations that can be expected with differing social customs, religion, etc., a BOD_5 per capita contribution of 40 g/day with a wastewater contribution of about 100 L/(capita/24 hr) is probably a reasonable initial estimate where there is a household water supply; however, flows also could be considerably less. The usual range of fecal coliform in domestic wastewater is 10^7–10^8 fecal coliform/100 mL, with a suitable design value being 5×10^7/100 mL.

Design of Anaerobic Ponds

The anaerobic ponds are designed on the basis of a recommended volumetric loading between 100 and 400 g BOD_5/m^3/24 hr, in order to maintain anaerobic conditions. Once the organic loading is selected, the volume of the pond is then determined and the hydraulic retention time is calculated. A retention time of less than 1 day should not be used for anaerobic ponds. If it occurs, however, a retention time of 1 day should be used, and the volume of the pond should be recalculated accordingly.

Design of Facultative Ponds

Facultative ponds can be designed on the basis of kinetic or empirical models. The rate at which the organic matter is oxidized by bacteria is a fundamental parameter in the rational design of biological wastewater treatment systems. It has been found that BOD removal often approximates first-order kinetics; that is, the rate of BOD removal (rate of oxidation of organic matter) at any time is proportional to the quantity of BOD (organic matter) present in the system at that time. The simple approach to the rational design of facultative ponds assumes that they are completely mixed reactors in which BOD_5 removal follows first-order kinetics[12] with the kinetic constant, k_1 at 20°C to be equal to 0.3/day.[13] For the temperature dependence, the Arrhenius expression is used, with the constant between 1.01 and 1.09—an average value of 1.05 is recommended to be applied generally. A minimum retention time value of 5 days should be adopted for temperatures below 20°C, and 4 days for temperature above 20°C. This is to minimize hydraulic short-circuiting and to give algae sufficient time to multiply (i.e., to prevent algal washout).

Design of Maturation Ponds for Fecal Coliform Removal

The method of Marais[2] is generally used to design a pond series for fecal coliform removal. This assumes that fecal coliform removal can be reasonably well represented by a first-order kinetic model in a completely mixed reactor. The numbers of fecal coliform/100 mL in the effluent and influent are used as unit in the applied first-order equation. k_T, the first-order rate constant for fecal coliform removal, is 2.6 L/24 hr, and it is highly dependent on the temperature. If an Arrhenius expression is used to calculate the influence of the temperature, an Arrhenius constant as high as 1.19 should be applied. Thus, k_T changes by 19% for every change in temperature of 1°C. The first-order equation represents the removal of fecal coliform in a series of ponds as a whole reasonably well. Maturation ponds require careful design to ensure that the fecal coliform removal is sufficient. Marais,[2] and Mara,[13] recommend a value of 3 days as the minimum retention time of the maturation ponds, although at temperatures below 20°C, values of 4–5 days are preferable. The BOD loading on the first maturation pond must be checked and must not be higher than that on the preceding facultative pond; in fact, it is preferable that it be significantly lower. The maximum BOD loading in the first maturation pond should normally be 75% of that on the preceding facultative pond. It is not necessary to check the BOD loadings on subsequent maturation ponds, as the non-algal BOD contribution to the load is very low. The loading on the first maturation pond is calculated on the assumption that 70% of unfiltered BOD has been removed in the preceding anaerobic and facultative ponds (or 80% for temperatures above 20°C). Mara and Pearson[11] and Mare[14] also suggested 90% cumulative removal in anaerobic and facultative ponds, and then 25% in each maturation pond, for temperatures above 25°C (80% and 20%, respectively, for temperatures below 20°C), when the BOD is based on filtered BOD values.

Helminth Egg Removal

Helminth eggs are normally removed by sedimentation, with the process occurring in the anaerobic or primary facultative ponds. If the final effluent is to be used for restricted irrigation, it is necessary to ensure that it contains no more than one egg per liter.

Water Quality and WSPs

By the right design of WSPs, it is possible to obtain close to 90% removal of BOD_5 (i.e., a BOD_5 of about 10–20 mg/L), a phosphorus removal of about 20%–35%, and a total nitrogen removal of at least 25%–50%. Usually, nitrate is removed very effectively. The organic nitrogen is easily oxidized to ammonia but ammonia may be removed by a relatively low efficiency due to insufficient oxygen concentrations in the ponds. By aeration (even rather moderate aeration), for instance, in the last facultative pond, it is possible to increase the nitrogen removal to almost 85% by oxidation of ammonium to nitrate, which, as indicated above, is removed usually with high efficiency. By careful design of the maturation ponds, the removal of coliform bacteria will be satisfactory.

Kayombo et al.[1] present several models that are able to give a fully acceptable design of WSPs. The same reference gives several equations for an empirical design, but as the model can consider interactions of several factors of importance for the design, it is recommendable to apply models for the design of WSPs whenever it is possible.

WSPs are, with very few exceptions, only applied in the tropical region, because the treatment results would not be sufficient during the winter months in the temperate zone, where the industrialized countries are using mostly high-technological treatment methods.

References

1. Kayombo, S.; Mbwette, T.S.A.; Katima, J.H.Y.; Ladegaard, N.; Jørgensen, S.E. Waste Stabilization Ponds and Constructed Wetlands. Design Manual Published by Danida and UNEP, 2004.
2. Marais, G.V.R. Dynamic behavior of oxidation ponds. Proceedings, 2nd International Symposium for Wastewater Lagoon, Missouri Basin Engineering Health Council and Federal Water Quality Administration, University of Kansas, Lawrence, 1970; 15–46.
3. Mara, D.D. Waste stabilization ponds. Problems and controversies. Water Qual. Int. **1987**, *1*, 20–22.
4. Mara, D.D. *Sewage Treatment in Hot Climates*; John Willey & Sons: New York, 1976; Vol. 120, 149.
5. Kayombo, S.; Mbwette, T.S.A.; Mayo, A.W.; Katima, J.H.Y.; Jorgensen, S.E. Modelling diurnal variation of dissolved oxygen in waste stabilization ponds. J. Ecol. Model. **1999**, *127*, 21–31.

6. Fritz, J.J.; Middleton, A.C.; Meredith, D.D. Dynamic process modelling of wastewater stabilization ponds. J. Water Pollut. Control Fed. **1979**, *51* (11), 2724–2742.
7. Neel, J.K.; McDermott, J.H.; Monday, C.A. Experimental lagoon of raw sewage at Fayette, Missouri. J. Water Pollut. Control Fed. **1961**, *33* (6), 603–641.
8. Pearson, H.W.; Mara, D.D.; Mills, S.W.; Smallman, D.J. Factors determining algal populations in waste stabilization ponds and the influence of algae on the performance. Water Sci. Technol. **1987**, *19* (12), 131–140.
9. Arthur, J.P. Notes on the Design and Operation of Waste Stabilization Ponds in Warm Climates of Developing Countries. Technical Paper No. 7, World Bank, Washington, DC, 1983.
10. Mara, D.D.; Pearson, H. Artificial freshwater environment: Waste stabilization ponds. Biotechnology **1986**, 8–16.
11. Mara, D.D.; Pearson, H. *Waste Stabilization Pond. Design Manual for Mediterranean Europe*; WHO Regional Office for Europe: Copenhagen, 1987.
12. Marais, G.v.R.; Shaw, V.A. A rational theory for the design of sewage stabilization ponds in tropical and sub-tropical areas. Trans. S. Afr. Inst. Civil Eng. **1961**, 3, 205.
13. Mara, D.D. *Design Manual for Waste Stabilization Ponds in India*. Ministry of Environment and Forests, National River Conservation Directorate, Lagoon Technology International: Leeds, U.K., 1997.

54

Wastewater Treatment Wetlands: Use in Arctic Regions 5-Year Update

Colin N. Yates,
Brent Wootton, and
Stephen D. Murphy

Introduction

In the last 50 years, conventional and alternative wastewater treatment systems (e.g., lagoons and tundra wetlands) have been used in remote Arctic communities in Canada for wastewater treatment. Current knowledge of the performance of these systems is limited, as little research has been conducted and regulatory monitoring has been poorly documented or not observed at all. Also, in the past, the rational design process of treatment systems in Arctic communities has not acknowledged cultural and socioeconomic aspects, which are important for the long-term management and performance of the treatment facilities in Canadian Arctic communities.

Wastewater treatment in extreme cold climate regions of world, such as the Canadian Arctic, is a comparatively recent exercise in the history of wastewater treatment. The term extreme cold climate wastewater treatment is used in this entry as cold climate wastewater has been coined for treatment in more temperate locations where mean annual ambient air temperatures are above 0°C. In comparison, the regions discussed in this entry all have mean annual ambient air temperatures well below –5°C, and often below –10°C.

Most communities in the far north regions of Canada, as well as other localities globally such as Greenland, have primarily relied on dilution of small volumes of wastewater by the receiving environment. However, with rapid population growth occurring in Arctic regions because of resource extraction and exploration, risk to the receiving environment has become a concern. In communities where wastewater treatment technologies have been employed, performance evaluations of the systems are now required to determine whether they can cope with the growing demand on them. This entry includes an overview of wastewater treatment in the Canadian Arctic with a specific emphasis on the use of tundra wetlands to treat municipal wastewater from the region's remote communities.

The corresponding objectives of this entry are as follows: (1) to provide a detailed description of current wastewater treatment practice in the Canadian Arctic with special consideration on the use of wetland systems; (2) to convey the current knowledge of wastewater treatment performance using natural wetlands in the Canadian Arctic; (3) to explore the potential for the application of ecotechnologies, namely, constructed or engineered wetlands in remote Arctic communities; and (4) to discuss a case study on a natural treatment wetland (Paulatuk, Northwest Territories, Canada) as an example to illustrate performance as well as describe other challenges associated with remote extreme cold climate wastewater treatment.

State of Current Knowledge and Practice

Current knowledge of performance of Canadian Arctic wastewater treatment systems is largely restricted to governmental reports, consultant reports, and other sources of gray literature. Only a few peer-reviewed documents exist to contribute to our current understanding of performance and are confined to the performance of lagoons.

In the Canadian Arctic, wastewater treatment facilities such as lagoons and wetlands are largely designed and managed using southern engineering standards, adopting design models to reflect Arctic temperature.[1–3] Since the 1970s, our knowledge of wastewater treatment in remote Canadian Arctic communities has grown very little despite a half-century of operation. Much of our understanding has been developed from site-specific consultant and government reports[4,5] and only a few peer-reviewed entries, as well as conference proceedings.[6,7]

The entries that have been listed above primarily address the performance of lagoon treatment systems. Only Doku and Heinke discuss the potential for greater use of natural and constructed wetlands (CWs) to treat wastewater in Northern Canada in detail.[8] The work of Dubuc et al. is one of very few papers to investigate long-term performance of treatment wetlands in Northern Canada, with a study on a hydro-construction camp along the 55th parallel in the province of Quebec.[9] To date, no long-term monitoring of treatment wetlands has occurred in the Canadian Arctic, except for research conducted by the authors in several wetland systems between the years 2008 and 2010. Unfortunately, in the past 10 years, additional long-term studies have not continued. Nor has there been any extensive discussion or study of mechanistic functions of tundra wetlands to treat wastewater in peer-reviewed literature except for some sporadic example works conducted since 2013.[10,11] Still, the closest approximation for mechanistic functionality in Arctic treatment wetlands is drawn from cold temperate climate regions of southern Canada, Scandinavia, and northern United States; examples from extensively studied locations being from Minot Wetland in North Dakota and Houghton Lake wetland in Michigan.[12,13] Only an entry by Kadlec and Johnson addresses some mechanistic function in a Canadian Arctic treatment wetland but does not provide significant background data.[1] Furthermore, much of the current knowledge on plant and microbial influence on wastewater treatment in the Arctic derives from smaller-scale fertilizations and carbon cycling studies in different Arctic environments.[14,15]

The authors would also like to note the lack of attention that has been given to the planning practice of wastewater treatment in the Canadian Arctic. In both peer-reviewed and gray literature, evidence of planning practice for wastewater treatment has been minimal. Only Ritter and Johnson directly touch upon the issue of planning and wastewater management in remote Arctic communities in North America.[16,17] More recently, more discussion has been made on public health risks in the Canadian Arctic with respect to wastewater.[18,19] The remainder of current thought on the subject relies on contributions from indirect sources on waste management and contamination in the Arctic.[20,21] Unpublished work by author Yates discusses these issues in great detail and identifies several primary factors that have compounded the problem of sanitation in Arctic communities, including the climate, their remote localities, and in some cases physiographic features. A few experts in sanitation and wastewater treatment have proposed and implemented conventional techniques in the Canadian Arctic with varying degrees of success in terms of (1) performance of the technology and (2) acceptance or understanding by the community.[6,23,24] Despite this, there remains a great deal of uncertainty with regard to which approaches are most suitable for Arctic communities. This uncertainty is because there is limited knowledge of how Arctic environments respond to increased loads of nutrients, other pollutants, and water, and how conventional systems respond to Arctic conditions.[7] Therefore, appropriate loading rates and predictions of expected performance are still largely speculative. Socioeconomic issues are also an ever-present concern, especially in the Canadian Arctic. This is related to a lack of resources and trained personnel as well as other factors.[16] Furthermore, there is distrust caused by a lack of communication and discussion between aboriginal groups in the Canadian Arctic and government agencies over treatment approaches and, more recently, concerns over compliancy to existing and proposed new regulatory standards.[25]

As previously described, there is a paucity of seasonal and long-term performance data of tundra treatment wetlands. Therefore, unsurprisingly, we know even less with respect to treatment mechanisms in tundra treatment wetlands. Attention still needs to be given to testing alternative technologies for wastewater treatment, such as CWs or engineered wetlands in these remote communities as well as emerging technologies that may provide low cost, effective, and socioculturally acceptable. The following discussion presents some evidence on the capability of tundra wetlands' ability to treat municipal wastewater.

A Discussion of Performance

Cold climate treatment wetlands have been identified as a significant area of interest for those studying treatment wetlands in the past three decades.[26] Vymazal also identified the important role that natural wetlands historically have played in our understanding of wetland function for wastewater treatment.[26] However, because of the growing knowledge of the importance of wetland function and values early in the adoption of wetlands to treat wastewater, their use has largely ceased except in controlled conditions[29,30] and in a few other locations around the world.[22] The tundra wetlands in the Canadian Arctic are among those still used to treat wastewater.

During a period over about 5 years (2008–2013), significant effort and funding were made available to characterize wastewater treatment wetlands in the Canadian Arctic for performance and to make recommendations on how to best manage these extreme cold climate systems. This work was largely conducted by CAWT (Fleming College) and Dalhousie. In the following discussion, a case study is used to illustrate the performance of Arctic tundra wetland wastewater treatment the authors conducted during that time. The authors also draw on some preliminary results collected in other tundra treatment wetlands that they have studied throughout the Canadian Arctic (Figure 1) and provide discussion on the potential for CWs.

Case Study of Paulatuk Treatment Wetland

This case study documents unpublished data collected by the authors in 2009 in a remote community of the Canadian Arctic (see summary of water quality data from Paulatuk and other selected wetland sites in Table 1). The Hamlet of Paulatuk is located in the Northwest Territories, Canada (69°N 124°W). The system is composed of a facultative lake and wetland serving approximately 294 residents.[31] Since this time, the population has decreased to 265 residents.[32] Wastewater from households and a small number of businesses is trucked to the facultative lake daily. The community is only accessible by aircraft, as few physical transportation corridors exist between the majority of communities in the Canadian Arctic (Figure 2).

Background Information

In 2007, it was estimated that approximately 11,200 m^3 of wastewater was being discharged into the lake (~31 m^3/day). The lake is estimated to have a volume of 103,000 m^3.[33] Basic estimates of effluent flow rate from a preferential flow channel as measured by Yates and Wootton showed a rate of 1.2 m^3/day.[34]

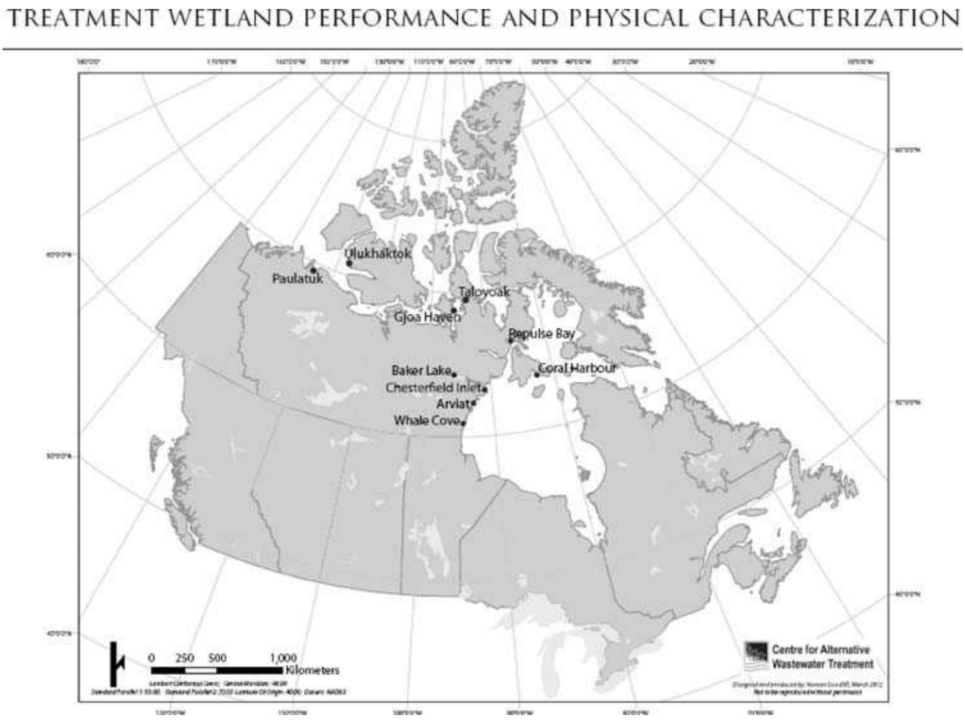


FIGURE 1 Map of Arctic Canada and its regions. Locations of each of the communities where treatment wetlands were studied are located in Nunavut and Northwest Territories. (Created by Noreen Goodliff.)

TABLE 1 Influent and Effluent Concentrations from Selected Tundra Treatment Wetlands in Arctic Canada

	Taloyoak			Paulatuk			Gjoa Haven		
	Influent	Effluent	% Removal	Influent	Effluent %	% Removal	Influent	Effluent	% Removal
Ammonia (NH_3N)	4.58	0.127	97	3.19	0.01	99	102	5.75	94
Total phosphorus	3.86	0.324	91	2.42	0.04	98	12.3	2.08	94
$cBOD_5$	12	3	75	40	2	95	138	3	83
Total suspended solids	–	–	–	35	3	91	22	3	97
E. coli	1.30E+03	2.40E+01	98	2.85E+03	1.00E+00	99	2.42E+03	3.30E+01	86

Note: All values are expressed as milligrams per liter (mg/L) except for *E. coli,* which was recorded as colony-forming units (cfu)/100 mL.

The wetland ranged from 40 to 80 m in width, extending approximately 350 m from the facultative lake to the Arctic Ocean (Figures 3–8). The wetland was characterized as wet-sedge tundra, dominated by *Carex* and *Poa* spp. In drier upland areas along the wetland boundaries, *Salix* spp. were observed to be dominant. The highest daily maximum temperature for the area is 15°C for July; mean annual temperature is approximately −9.2°C.[35] Paulatuk's wetland hydraulic loading rate into the facultative lake was estimated at 31 m^3/day, assuming that the amount of water continuously discharged from the facultative lake is equal to the amount as it receives a Carbonaceous Biochemical Oxygen Demand ($cBOD_5$) areal loading rate of 0.9 kg/ha/day. However, the flow rate into the wetland is much less because of evaporation and some loss into groundwater. Precipitation and runoff from neighboring hillsides may add to the flow through the wetland but only a minimal amount as this region only receives 84 mm of precipitation

FIGURE 2 Aerial view of Paulatuk facultative lake and treatment wetland. The wetland is located on the upper right of the photo. Effluents flow to the right and a distance of approximately 350 m to the Arctic Ocean.
Source: Photo courtesy of Aboriginal Affairs and Northern Development Canada North Mackenzie District.

FIGURE 3 View of facultative lake flowing into the beginning of the wetland system.

FIGURE 4 View of productive vegetation in wetland.

FIGURE 5 View of wetland showing discrete boundaries.

FIGURE 6 Upslope view of wetland.

FIGURE 7 Downslope view of wetland.

FIGURE 8 View of wetland discharging into the Arctic Ocean.

from June to October.[35] Background water quality concentrations for $cBOD_5$, NH_3–N, and *Escherichia coli* from a nearby reference wetland were measured to be 2 mg/L, 0.01 mg/L, and 9 colony-forming units (cfu)/100 mL, respectively.

Methods

The authors conducted an extensive characterization analysis on the treatment wetland, collecting surface and subsurface water samples throughout the wetland. A onetime sampling of the site was conducted over a week in early September 2009 with a sample size of $n = 41$. Spatial interpolative analysis was conducted to illustrate concentration changes throughout the wetland. Measurements for $cBOD_5$, NH_3–N, and *E. coli* were taken. Samples were shipped within 24 hours to an ISO 17025 accredited laboratory and processed using *Standard Methods for the Examination of Water and Wastewater.* Tension spline was used to conduct interpolation analyses mapping of water quality parameters.

Results

The interpolation analysis showed that treatment for all parameters was found to be occurring in the first 50–100 m of the wetland. After 150 m, flows of wastewater were difficult to detect, as wastewater appeared to be flowing evenly at low velocities across much of the wetland. With a basic understanding of the rate of flow and loading of the wetlands and interpolation of concentration of specific wastewater parameters, it is possible to discuss the performance of the system. Treatment of wastewater was observed to occur primarily in the upper portions of the wetlands, with concentrations quickly dissipating to background levels (Figure 9).

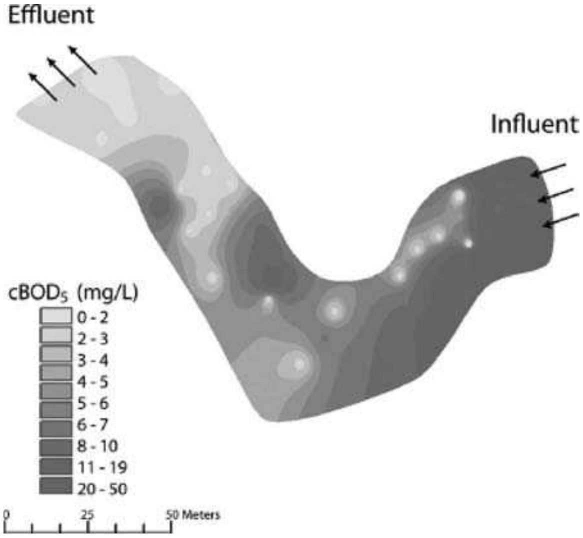


FIGURE 9 Concentration gradients of $cBOD_5$ in the Paulatuk treatment wetland.

Expected effluent concentrations for the identified active portions of the wetlands were calculated using a first-order kinetic model at 10°C. We calculated expected effluent concentrations using P-*k*-C* in order to determine expected effluent values for $cBOD_5$. The van't Hoff–Arrhenius equation as described in Crites and Tchobanoglous was adopted[31]:

$$\frac{\mathrm{d}(\ln k)}{dT} = \frac{\mathrm{E}}{RT^2} \tag{1}$$

The P-*k*-C* model is described in Campbell and Ogden as follows[32]:

$$\mathrm{As} = \frac{Q(\ln \mathrm{Co} - \ln \mathrm{Ce})}{k_\mathrm{t} * d * n} \tag{2}$$

The k_t value for the P-*k*-C* model was determined by using a k_{10} value of 1.0; the θ-factor used was 1.14. A high θ-factor was deemed appropriate for extreme temperature cases as determined for a Minnesota horizontal subsurface flow wetland with a temperature range from 1°C to 17°C, as outlined in Kadlec and Wallace.[30]

Expected $cBOD_5$ effluent values calculated for the Paulatuk wetland were 3 mg/L. In the Paulatuk wetland, $cBOD_5$ levels were actually observed to decrease rapidly (Table 1) within the first 50–100 m. This was likely due to the low influent levels from pretreatment facilities, allowing the top end of the wetland to assimilate or treat remaining organic matter. The influent demand was observed to be 40 mg/L. Effluent demands were observed to be 2 mg/L. However, a level of 2 mg/L was observed to consistently appear less than halfway down the wetland during the study. Effluent oxygen demands were observed to be the same as background conditions measured in the nearby reference sites.

Nutrient parameters, specifically ammonia (NH_3–N), also decreased rapidly down each of the wetlands; concentrations quickly dissipated within the first 50 m of the wetland, again likely due to the low influent concentrations (Figure 10).

Observed NH_3–N concentration in the influent was 3.19 mg/L. Effluent concentration was observed to be 0.01 mg/L. Like $cBOD_5$, NH_3–N concentrations dissipated rapidly, achieving background concentrations in approximately 150 m. Pathogen removal by the wetland was also observed to occur quickly in

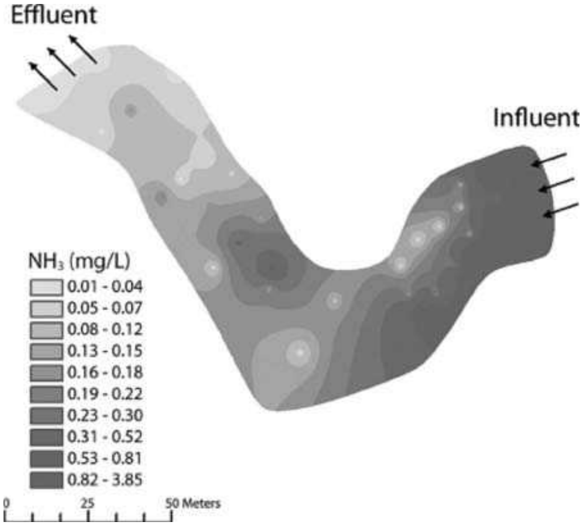


FIGURE 10 Concentration gradients of NH_3–N in the Paulatuk treatment wetland.

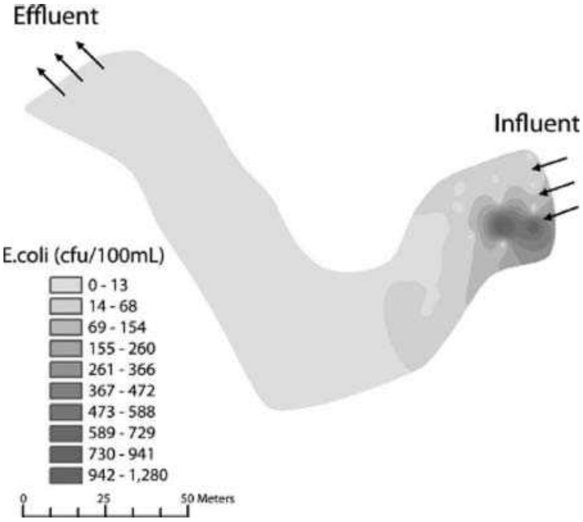


FIGURE 11 Concentration gradients of *E. coli* in the Paulatuk treatment wetland.

the pretreated wastewaters entering the wetlands in Paulatuk. Concentrations of *E. coli* were observed to be low at the influence of the wetland: 2.85E3 cfu/100 mL. *E. coli* was observed to be quickly removed, only observed in trace amounts within a third of the wetland distance. Variation throughout the wetland is likely explained by natural background concentrations from local wildlife (Figure 11).

Discussion

Microbial activity and plant growth influenced by soil temperature are likely the most important mechanisms for the treatment of wastewater in Arctic wetlands.[36] Arctic soil is known to be an excellent sink of organic matter and nutrients, immobilizing nutrients within the frozen matrix and within the microbial community.[37] Phosphorus has been shown to be bound to soils in the Arctic, rendering it unavailable for plant uptake.[38] However, it is unknown how much the soil matrix is responsible for "treatment" by locking nutrients. Fertilization studies in various Arctic habitats, including wet-sedge tundra, have shown that in nutrient-limiting conditions, plant communities respond to increased nutrient input based on small nutrient additions,[39–41] especially when nutrients were added simultaneously,[27] as would be the case with wastewater. Some Arctic plants have even demonstrated the ability to uptake organic forms of N because mineralization of organic material is slow due to low soil temperatures.[39] Plants in tundra treatment wetlands, such as those presented here, may be utilizing the readily available nutrients in such a manner, which may explain low values of inorganic N in a similar study conducted on Arctic treatment wetlands.[28] A study by Yates et al. showed that plants may play a significant role in treatment.[11] However, actual nutrient uptake rates in these systems have not been studied to determine rate or percentage of nutrients discharged into the system or taken up by the plant community. The microbial community may also play an equal role in the uptake of available nutrients in wastewater in these Arctic wetlands. Similar to plant communities, microbial activity is generally limited by temperature and available nutrients. Arctic microbial species are more efficient at lower temperatures than temperate microbial species; Arctic species continue to transform nutrients throughout the winter.[42,43] Hobbie and Chapin also suggested that microbial activity is able to utilize nutrients in soils at temperatures as low as −5°C.[36] Nutrient uptake at low temperatures was recently validated by Edwards and Jefferies.[16] These observations likely contribute to the rapid increase in wetland performance from late June to early July due to increases in microbial populations as a result of additional nutrient availability in still semi-frozen soils observed by Yates et al.[28] Whether winter microbial activity is sufficient to continue to mineralize organic matter and nutrients is unknown. It is likely that the microbial community would not be able to significantly consume the excess nutrient and organic loads at the top of the wetland, resulting in the gradual infilling of organics at the influence.

Natural ultraviolet radiation plays an important role in the disinfection of wastewater in surface wetlands and lagoon systems in more temperate systems. The long exposure of sunlight in the Arctic during the summer months in theory should promote increased disinfection. However, if water temperatures are not optimal, lysis of bacteria may not occur, as cold temperatures appear to stabilize populations, at least in lagoon environments.[2] In wetlands that do not maintain large areas of open water, solar radiation cannot penetrate the water column because of the plant canopy,[44] which is often the case in tundra wetlands that maintain dense stands of *Carex*.

Finally, sedimentation of solids occurs on the wetland surface, in various preferential flow channels throughout the wetlands, and also through entrapment in vegetation. Personal observations in field notes show accumulations of organic matter in many of the wetlands surveyed throughout the Arctic. As discussed earlier, decomposition rates by the microbial community are not as high as deposition rates. Although it was observed that water quality was low in organic load throughout much of the wetland, it would be expected that deposition would occur further down the wetland in the future.

Potential for CWs

The authors have also studied numerous other tundra wetland wastewater treatment systems throughout the Canadian Arctic observing similarly positive results. Examples of preliminary results can be found on Chesterfield Inlet's tundra treatment wetland,[28] as well as a subsurface flow CW in Baker Lake,[45] the latter being examined in some detail in the following discussion. Data from other communities that the authors have studied are included below and exemplify in/out performance of tundra treatment wetlands (Table 1).

Current findings in Arctic wetlands correspond with the results presented by Andersson et al. and Kadlec, demonstrating that natural wetlands can effectively polish pretreated wastewater, often in a much small area than was originally calculated by mass balance equations.[13,46] The performance results observed further prove the ability of wetlands, specifically natural systems, to treat wastewater in a cold climate, but more significantly, these early results demonstrate the resilience of wetlands to produce low-concentration effluents following approximately 9 months of frozen conditions. Although early studies by the authors did not aim to determine which mechanisms are largely responsible for tundra wetlands' high efficacy, specialized bacteria and macrophytes that have evolved in the low-temperature conditions of the Canadian Arctic are likely candidates. More importantly, these results indicate that ecotechnologies such as CWs could potentially be an appropriate technology for extreme cold climate wastewater treatment.

CWs or engineered wetlands have been applied around the world in numerous climates.[26,47–49] Most definitions of CWs simply acknowledge a CW as a man-made structure that emphasizes the natural characteristics of wetlands to transform and absorb contaminants.[30] Vymazal provides a similar definition: "CWs are engineered systems that have been designed and constructed to utilize natural processes involving wetland vegetation, soils, and the associated microbial assemblages to assist in treating wastewaters."[47] Throughout this entry, CWs are defined in the following manner: CWs are engineered systems that are lined to prevent significant exfiltration of wastewater into the underlying ground prior to passing through the system and maintain mechanisms to control influent and effluent flow. The preceding definition is used because various measures have been used to manipulate flow regimes in tundra wetlands in the Canadian Arctic. For example, some tundra systems have made use of some engineered structures, such as berms and inflow/outflow pipes, or make use of natural liners, such as bedrock, to contain or control flow. These systems are described as augmented natural wetlands.

Early belief was that cold climate conditions would not allow wetlands to optimally treat wastewater, and therefore, treatment wetlands would not find a place in cold climate wastewater treatment.[48] Studies from both North America and Scandinavia have largely shown that this has not been the case, and in most instances, only minor impediments to treatment have been observed.[50–52] Despite early convictions, CWs have shown great promise as alternative low-cost technologies to treat wastewater in remote, economically challenged regions and small communities even in cold climatic conditions.[53–55] However, despite their extensive successful use in cold temperate climates around the world,[48,51,52] they have yet to be tested in extreme cold climate conditions, like the Canadian Arctic. Communities in the Canadian Arctic in theory make excellent candidates for alternative wastewater treatment technologies, because of limited economic resources, physiographic characteristics, and trained personnel to operate and maintain more conventional mechanical treatment facilities.[18] For a number of decades, communities in the Canadian Arctic have been using tundra wetlands to treat their wastewater.[1,56] Although our knowledge is growing and understanding of treatment performance and mechanisms of the tundra wetlands is limited, some evidence has shown excellent (sometimes orders of magnitude below regulatory standards) removal for regulated wastewater effluent parameters during the summer months.[1,28] Because of the socioeconomic conditions and the extensive use of tundra wetlands in the Canadian Arctic, CWs warrant experimentation in this region. However, testing of these systems in the Canadian Arctic will require research trials on different design specifications to account for low soil/water temperatures, short frost-free period, slow rate of decomposition of organic matter, and therefore slow mineralization rates of various nutrients. Engineered designs to optimize existing tundra wetlands, augmented natural wetlands to increase hydraulic residency time (HRT), and increase active treatment zones (decrease areal loading rates) have been adopted in a few instances in the Arctic; Cambridge Bay, Nunavut, is an example of one such system.[1] However, the Arctic does provide an environment that in theory provides optimal treatment conditions, for example, 24 hours of sunlight in summer, and plants and bacteria that have evolved in a cold and nutrient-limited environment, giving them characteristics that allow them to utilize excess nutrients efficiently.

As tundra wetlands are extensively used in the Canadian Arctic, CWs have excellent potential to act as low-cost technologies for Arctic communities. The authors studied the performance of the first known experimental engineered HSSF system in the Canadian Arctic. The system demonstrated very promising results in its first year (2009) of operation despite high loading rates; observed reductions of wastewater concentrations were 25%, 31%, 52%, 99.3%, 99.3%, and 5% for cBOD_5, COD, TSS, *E. coli*, total coliforms, and TP, respectively.[45] In 2010, the system was operated with lower loading rates. It was expected that the system would achieve greater reductions, but this was not the case. Concentrations in the wetland effluent were observed to be greater than the effluent. Based on these observations, it was concluded that high organic loading prior to biofilm and plant establishment and high organic loading during the first year of study saturated the system with organics, resulting in the release of solids and unmineralized nutrients into a less concentrated influent. Overall, the HSSF system did not perform as expected but did demonstrate the potential for use in remote Arctic communities. Further investigations of other CW designs should be undertaken in the future.

The issues discussed to this point have largely pertained to wastewater management in the Canadian Arctic. However, they do reflect similar conditions in remote communities elsewhere in the world, cold and warm climates alike. Similar scenarios have been recently described by Jenssen in Greenland, which have similar demographics to the Canadian Arctic.[56] In the mid-1990s, rural communities in Estonia were also facing similar challenges with insufficient treatment facilities because of shortcomings in economic resources.[57] Denny and Kivaisi both describe how CWs could have potential for developing countries, specifically those in Africa.[55,58] As suggested by the previous examples, the discussion of wastewater management for developing countries is prevalent in the literature, yet little attention until this point has been directed towards remote underdeveloped communities in the Canadian Arctic. It is important to note that many common themes run throughout all of these regions that are not dissimilar to those described in the Canadian Arctic; these include but are not limited to low-economic capacity, absence of skilled labor, and complex sociocultural environments. The recommendations made for wastewater management in the Canadian Arctic contribute to knowledge development in all remote regions globally. Also, a similar set of approaches that include the use of an adaptable management framework, and accounting for differences in understanding from the experts in the field (e.g., engineers and planners) designing systems to those adopting the technology (e.g., wastewater operators) in the communities, could be easily adopted or tested outside the Canadian Arctic. The continued optimized use of wetlands, particularly CWs, is one approach that could be more extensively explored in all cases.

Modeling Treatment Wetlands

To support the United Nations Millennium Development Goals of reversing the loss of environmental resources as well as halving by 2015 the proportion of people without access to safe drinking water and sanitation, the United Nations Environment Programme promotes and facilitates environmentally sound technologies (ESTs). Natural wetlands and CWs are considered to be ESTs for wastewater treatment. Design and use of treatment wetlands are facilitated with the use of models to replicate processes and provide practical tools for sanitation.

SubWet, software for the design of subsurface horizontal flow artificial wetlands for water quality improvement and treatment, was originally developed by United Nations Environment Programme–Division of Technology, Industry and Economics–International Environmental Technology Centre (UNEP–DTIE–IETC). After being successfully used as a design tool in 15 cases in Tanzania, it was felt that the model should be upgraded for cold climate application. The Centre for Alternative Wastewater Treatment of Fleming College further developed a new version in collaboration with UNEP–DTIE–IETC, creating SubWet 2.0 to accommodate temperate and cold climatic conditions including summer Arctic and temperate winter conditions.[58] The model simulates removal of nitrogen (including nitrogen in ammonia, nitrate, and organic matter), phosphorus, and BOD_5 in milligrams per liter and the corresponding removal efficiencies in percentage. Design inputs in the model for the wetland are as follows: width, length, depth, slope, % particulate matter, precipitation factor (total flow rate included precipitation/flow rate of treated water), hydraulic conductivity, and selected flow rate (in cubic meters per day).

The forcing functions applied for the model are the following: volume of wetland; flow of water; porosity; input of BOD_5, ammonium, nitrate, total phosphorus, and organic nitrogen; fraction of BOD_5, phosphorus, and organic N as suspended matter; average oxygen concentration; and average temperature. The length of model simulations must be indicated as number of days. The model uses

16 parameters, for example, the temperature coefficient of nitrification, and has a version for warm climate and a version for cold climate. The difference between the two versions is only the set of parameter values. The model uses 25 differential process equations. The model calculates the values of BOD_5-out, nitrate-out (NIT-out), ammonium-out (AMM-out), total phosphorus-out (TPO-out), and organic nitrogen-out (ORN-out) and can be shown in the form of both tables and graphs.

SubWet 2.0 has been developed to support decision-making processes by assisting experts and water managers in the design and evaluation of CWs to improve water quality and treat domestic wastewater. Furthermore, SubWet 2.0 can also be used as a tool to improve the efficiency of low- or non-performing systems. Due to its characteristics, this software is very useful for training technicians and students who are interested in modeling natural and artificial wetlands used for wastewater treatment and can also be used as a tool by engineers and regulators.

Finally, some other technologies have emerged to try solving the challenge effectively and economically that are worth noting including chemical coagulation, UV, and peracetic acid disinfection in Greenland[59] and Bioelectrochemical Anaerobic Sewage Treatment, or BEAST. However, little data have been published on the success of these technologies to date in small Arctic communities.

Conclusion

Wastewater treatment in extreme cold climates will continue to be a challenge into the future until more knowledge is gained in the field and new technological advancements are tested in these regions. Socioeconomic barriers will likely persist even longer and will only become exacerbated in the short term as regions such as Canada's northern territories cope with rapid growth. However, as the authors have described, tundra wetlands can provide preferred levels of treatment of domestic wastewater in remote Canadian Arctic communities. The successful use of natural systems indicates that further experimentation of the use of CWs and engineered wetlands in extreme cold climates is warranted. Finally, much study is still required in tundra wetlands to identify the primary mechanistic functions responsible for the removal of nutrients and organic matter from wastewater in these systems.

References

1. Kadlec, R.H; Johnson, K. Cambridge Bay, Nunavut, Wetland planning study. *J. North. Territories Water Waste Assoc.* **2008**, *9* (Fall), 30–34.
2. Prince, D.S.; Smith, D.W.; Stanley, S.J. Intermittent-discharge lagoons for use in cold regions. *J. Cold Reg. Eng.* **1995**, *9* (4), 183–194.
3. Heinke, G.W.; Smith, D.W.; Finch, G.R. Guidelines for the planning and design of waste-water lagoon systems in cold climates. *Can. J. Civ. Eng.* **1991**, *18* (4), 556–567.
4. Dillon Consulting Ltd. *Hamlet of Paulatuk: Operation and Maintenance Manual, Sewage and Solid Waste Disposal*. Report 04-3332, 2004, available at http://www.docstoc.com/docs/45286151/Appendix-H---Operations-and-Main-tenance-Manual-Sewage-and-Solid-Waste (accessed March 2012).
5. Environment Canada. *Sewage lagoons in cold climates*. Report EPS 3/NR/1. Government of Canada. 1985. 98–&.
6. Miyamoto, H.K.; Heinke, G.W. Performance evaluation of an arctic sewage lagoon. *Can. J. Civ. Eng.* **1979**, *6*, 324–328.
7. Johnson, K; Wilson, A. *Sewage treatment systems in communities and camps of the Northwest Territories and Nunavut Territory*. 1st Cold Regions Specialty Conference of the Canadian Society for Civil Engineering, Regina, Canada, June 2–5, 1999.
8. Doku, I.A; Heinke, G.W. Potential for greater use of wetlands for waste treatment in Northern Canada. *J. Cold Reg. Eng.* **1995**, *9* (2), 75–88.

9. Dubuc, Y.; Janneteau, P.; Labonte, R.; Roy, C.; Briere, F. Domestic waste-water treatment by peatlands in a northern climate—A water-quality study. *Water Resour. Bull.* **1986**, *22* (2), 297–303.
10. Hayward, J.; Jamieson, R.; Boutilier, L.; Goulden, T.; Lam, B. Treatment performance assessment and hydrological characterization of an arctic tundra wetland receiving municipal wastewater. *Ecol. Eng.* **2014**, *73*, 786–797.
11. Yates, C.; Wootton, B.; Varickanickal, J.; Cousins, S. Testing the ability to enhance nitrogen removal at cold temperatures with *C. aquatilis* in a horizontal subsurface flow wetland system. *Ecol. Eng.* **2016**, *94*, 344–351.
12. Hammer, D.A.; Burckhard, D.L. Low temperature effects on pollutant removals at Minot's wetland. In *Natural Wetlands for Wastewater Treatment in Cold Climates*; Mander, U., Jenssen, P., Eds.; WIT Press: Boston, MA, 2002.
13. Kadlec, R.H. The Houghton Lake wetland treatment project. *Ecol. Eng.* **2009**, *35* (9), 1285–1286.
14. Shaver, G.R.; Chapin, F.S. Long-term responses to factorial, NPK fertilizer treatment by Alaskan wet and moist tundra sedge species. *Ecography* **1995**, *18* (3), 259–275.
15. Arens, S.J.T.; Sullivan, P.F.; Welker, J.M. Nonlinear responses to nitrogen and strong interactions with nitrogen and phosphorus additions drastically alter the structure and function of a high arctic ecosystem. *J. Geophys. Res., [Biogeosci.]* **2008**, *113* (G3), G03S09.
16. Edwards, K.A.; Jefferies R.L. Nitrogen uptake by *Carex aquatilis* during the winter–spring transition in a low Arctic wet meadow. *J. Ecol.* **2010**, *98* (4), 737–744.
17. Ritter, T.L. Sharing environmental health practice in the North American Arctic: A focus on water and wastewater service. *J. Environ. Health* **2007**, *69* (8), 50–55.
18. Johnson, K. The social context of wastewater management in remote communities. *Environ. Sci. Eng. Mag.* **2010**, (2), 28–30.
19. Daley, K.; Jamieson, R.; Rainham, D.; Hansen, L. Wastewater treatment and public health in Nunavut: A microbial risk assessment framework for the Canadian Arctic. *Environ. Sci. Pollut. Res.* **2018**, *25* (33), 32860–32872.
20. Daley, K.; Castleden, H.; Jamieson, R.; Furgal, C.; Ell, L.; Water systems, sanitation, and public health risks in remote communities: Inuit resident perspectives from the Canadian Arctic. *Soc. Sci. Med.* **2015**, *135*, 124–132.
21. Berkes, F.; Berkes, M.K.; Fast, H. Collaborative integrated management in Canada's north: The role of local and traditional knowledge and community-based monitoring. *Coastal Manage.* **2007**, *35* (1), 143–162.
22. Environment Canada. *National Consultation Report: Wastewater Consultations*. Government of Canada, 2009, available at http://www.ec.gc.ca/eu-ww/default.asp?lang=En&n=4B94EB4A-1 (accessed March 2012).
23. Dawson, R.N.; Grainge, J.W. Proposed design criteria for waste-water lagoons in arctic and subarctic regions. *J. Water Pollut. Control Fed.* **1969**, *41* (2P1), 237–246.
24. Grainge, J.W. Arctic heated pipe water and waste water systems. *Water Res.* **1969**, *3* (1), 47–71.
25. Johnson, K. *Inuit Position Paper Regarding the CCME Canada-wide Strategy for the Management of Municipal Wastewater Effluent and Environment Canada's Proposed Regulatory Framework for Wastewater.* Inuit Tapiriit Kanatami: Ottawa, 2008.
26. Vymazal, J. Constructed wetlands for wastewater treatment: Five decades of experience. *Environ. Sci. Technol.* **2011**, *45* (1), 61–69.
27. Gough, L.; Wookey, P.A.; Shaver, G.R. Dry heath arctic tundra responses to long-term nutrient and light manipulation. *Arctic Antarctic Alpine Res.* **2002**, *34* (2), 211–218.
28. Yates, C.N.; Wootton, B.; Jprgensen, S.E.; Santiago, V.; Murphy, S.D. Natural Wetlands for Treating Municipal Wastewater in the Canadian Arctic: A Case Study of the Chesterfield Inlet, Nunavut Wetland. *Proceedings of the 12th International Conference on Wetland Systems for Water Pollution Control*, Venice, Italy, October 12–17, 2010.

29. Mander, Ü.; Jenssen, P.D. *Constructed Wetlands for Wastewater Treatment in Cold Climates*; WIT Press: Boston, MA, 2003.
30. Kadlec, R.H.; Wallace, S. *Treatment Wetlands,* 2nd Ed.; CRC Press: Baton Rouge, LA, 2009.
31. Statistics Canada. 2012. *Paulatuk, Northwest Territories (Code 6101014) and Region 1, Northwest Territories (Code 6101) (Table).* Census Profile. 2011 Census. Statistics Canada Catalogue no. 98-316-XWE. Ottawa. Released February 8, 2012, available at http://www12.statcan.ca/cen-sus-recensement/2011/dp-pd/prof/index.cfm?Lang=E (accessed March 2012).
32. Statistics Canada. 2017. *Paulatuk, HAM [Census subdivision], Northwest Territories and Northwest Territories [Territory] (table).* Census Profile. 2016 Census. Statistics Canada Catalogue no. 98-316-X2016001. Ottawa. Released November 29, 2017.
33. Wootton, B.; Durkalec, A.; Ashley, S. *Canadian Council of Ministers of the Environment Draft Canada-wide Strategy for the Management of Municipal Wastewater Effluent: Inuvialuit Settlement Region Impact Analysis*; Inuit Tapiritt Kanatami: Ottawa, 2008.
34. Yates, C.N.; Wootton, B. *Treatment Wetland Performance and Physical Characterization: Paulatuk, NWT and Pond Inlet, NT;* Environment Canada—Wastewater Technology Centre: Burlington, 2010; 1–103.
35. Environment Canada. Canadian Climate Normals. 2010, available at http://www.climate.weatheroffice.gc.ca/climate_normals/index_e.html (accessed July 2010).
36. Hobbie, S.E.; Chapin, F.S. Winter regulation of tundra litter carbon and nitrogen dynamics. *Biogeochemistry* **1996**, *35* (2), 327–338.
37. Schmidt, I.K.; Jonasson, S.; Michelsen, A. Mineralization and microbial immobilization of N and P in arctic soils in relation to season, temperature and nutrient amendment. *Appl. Soil Ecol.* **1999**, *11* (2–3), 147–160.
38. Mack, M.C.; Schuur, E.A.G.; Bret-Harte, M.S.; Shaver, G.R.; Chapin, F.S. Ecosystem carbon storage in arctic tundra reduced by long-term nutrient fertilization. *Nature* **2004**, *431* (7007), 440–443.
39. Chapin, F.S.; Moilanen, L.; Kielland, K. Preferential use of organic nitrogen for growth by a nonmycorrhizal arctic sedge. *Nature* **1993**, *361* (6408), 150–153.
40. Hobbie, S.E.; Gough, L.; Shaver, G.R. Species compositional differences on different-aged glacial landscapes drive contrasting responses of tundra to nutrient addition. *J. Ecol.* **2005**, *93* (4), 770–782.
41. Edwards, K.A.; McCulloch, J.; Kershaw, G.P.; Jefferies, R.L. Soil microbial and nutrient dynamics in a wet Arctic sedge meadow in late winter and early spring. *Soil Biol. Biochem.* **2006**, *38* (9), 2843–2851.
42. Larsen, K.S.; Grogan, P.; Jonasson, S.; Michelsen, A. Dynamics and microbial dynamics in two subarctic ecosystems during winter and spring thaw: Effects of increased snow depth. *Arctic Antarctic Alpine Res.* **2007**, *39* (2), 268–276.
43. MacIntyre, M.E.; Warner, B.G.; Slawson, R.M. *Escherichia coli* control in a surface flow treatment wetland. *J. Water Health* **2006**, *4* (2), 211–214.
44. Yates, C.N.; Wootton, B.; Jørgensen, S.E.; Santiago, V.; Murphy, S.D. Annak, Ijji, Imaqsuk: Early Findings from The First Experimental Constructed HSSF Wetland in the Canadian Arctic. *Proceedings of the 12th International Conference on Wetland Systems for Water Pollution Control*, Venice, Italy, October 12–17, 2010.
45. Andersson, J.L.; Wittgren, H.B.; Kallner, S.; Ridderstolpe, P.; Hagermark, I. Wetland Oxelosund, Sweden—The first five years of operation. In *Natural Wetlands for Wastewater Treatment in Cold Climates;* Mander, U., Jenssen, P., Eds.; WIT Press: Boston, 2002.
46. Vymazal, J. Constructed wetlands for wastewater treatment. *Ecol. Eng.* **2005**, *25* (5), 475–477.
47. Wittgren, H.B.; Maehlum, T. Wastewater treatment wetlands in cold climates. *Water Sci. Technol.* **1997**, *35* (5), 45–53.

48. Maehlum, T.; Stalnacke, P. Removal efficiency of three cold-climate constructed wetlands treating domestic wastewater: Effects of temperature, seasons, loading rates and input concentrations. *Water Sci. Technol.* **1999**, *40* (3), 273–281.
49. Giaever, H. Treatment performance of a multistage constructed wetland at 69 degrees north, Norway. *J. Environ. Sci. Heal. A.* **2000**, *35* (8), 1377–1388.
50. Jenssen, P.D.; Maehlum, T.; Krogstad, T.; Vrale, L. High performance constructed wetlands for cold climates. *J. Environ. Sci. Heal. A.* **2005**, *40* (6–7), 1343–1353.
51. Wallace, S; Parkin, G.; Cross, C. Cold climate wetlands: Design and performance. *Water Sci. Technol.* **2001**, *44* (11–12), 259–265.
52. O'Hogain, S. Reed bed sewage treatment and community development/participation. In *Wastewater treatment, Plant Dynamics and Management in Constructed and Natural Wetlands*; Springer: New York, 2008.
53. Merlin, G.; Pajean, J.L.; Lissolo, T. Performances of constructed wetlands for municipal wastewater treatment in rural mountainous area. *Hydrobiologia* **2002**, *469* (1–3), 87–98.
54. Kivaisi, A.K. The potential for constructed wetlands for wastewater treatment and reuse in developing countries: A review. *Ecol. Eng.* **2001**, *16* (4), 545–560.
55. Wootton, B.; Yates, C.N. Wetlands: Simple and effective wastewater treatment for the north. *Meridian Newsl.* **2010**, (Fall/Winter), 12–16.
56. Jenssen, P. *Wastewater treatment in cold/arctic climate with a focus on small scale and onsite systems*. 28th Alaska Health Summit, Anchorage, Alaska, January 10–13, 2011.
57. Tenson, J. Wastewater treatment in Estonia: Problems and opportunities. *Recycl. Res.: Proc. Sec. Int. Conf. Ecol. Eng. Wastewater Treat.* **1996**, *5* (6), 65–71.
58. United Nations Environment Program. 2009. Download software here: http://www.unep.or.jp/Ietc/Publications/Water_Sanitation/SubWet2/index.asp (accessed March 2012).
59. Chhetri, R. K.; Klupsch, E.; Andersen, H. R.; Jensen, Pernille, E. Treatment of Arctic wastewater by chemical coagulation, UV and peracetic acid disinfection. *Environ. Sci. Pollut. Res.* **2018**, *25* (33), 32851–32859.

55

Wastewater Treatment: Biological

Shaikh Ziauddin Ahammad, David W. Graham, and Jan Dolfing

Introduction

Human activities and population growth have placed the environment under increasing stress. Furthermore, indiscriminate use of natural resources is accompanied by increased local and global pollution levels, which are reflected in imbalances in our ecosystems. The generation of large quantities of wastewater with a high organic content and toxicants is one obvious product of excessive consumption. It has been known for many years that environmental discharges of high loads of organic matter can result in oxygen depletion in receiving waters due to stimulated microbial activity. This oxygen depletion and the presence of trace toxicants found in wastes also negatively influence ecosystems, including reduced biodiversity and environmental health. Therefore, negative environmental impacts have driven our need to understand the effect of pollution on water bodies and develop proper measures to reduce discharges, including treatment processes.

Different technologies are available to treat wastes. However, biological wastewater treatment methods are most valuable because their economic benefits are high, especially when coupled with waste stabilization and resource recovery. The optimal treatment processes depend on the waste type and treatment goals. Wastewater generally originates from two sources: 1) domestic wastewater from gray water, toilets, and other domestic activities; and 2) industrial wastewater, generated by industries during the normal course of activity, which often rely on the local sewerage systems for waste processing. Therefore, the composition of wastewater, including quantity and constituents, varies considerably from place to place, depending on suite of sources, social behavior, the type and number of industries within a catchment, climatic conditions, water consumption, and the nature of the wastewater collection system. Given this variety, wastewater treatment processes must be innately versatile, but also sometimes must be tailored to the specific waste and conditions. The purpose of this entry is to describe different biological treatment methods and then discuss their relative capacities to treat different wastes on the basis of waste characteristics and the desire for resource recovery.

Wastewater Treatment Options

Special handling and treatment of wastes have been performed for thousands of years in response to their perceived importance, although approaches have changed as perceptions have changed over history. In 4th century B.C. in Greece, the *Athenian Constitution* written by Aristotle[1] proscribed provisions for the appropriate handling of sewage. Concern was based on aesthetics, probably odors, because relationships between domestic wastes and health were not yet known. It was not until the mid-1800s that links between wastes and human health became more apparent, which led to a progression of waste management approaches and technologies to address health concerns.

Treatment technologies evolved slowly over time, including physical, chemical, and biological approaches, many of which are still used in different sectors. Physical methods are based on the application of physical forces, such as screening, mixing, flocculation, sedimentation, flotation, filtration, and gas transfer. Alternately, chemical processes treat contaminants by adding chemicals or by stimulating specific chemical reactions. Precipitation, adsorption, and disinfection are common examples of chemical treatment methods. Physical and chemical methods are often combined, especially in industrial treatment scenarios. In contrast to physiochemical processes, biological processes remove organic contaminants (e.g., biodegradable organic material) largely through microbiological activity. Commonly used biological treatment methods include aerobic treatment in ponds, lagoons, trickling filters, and activated sludge plants,[2] and anaerobic treatment[3,4] in similar reactor systems. Processes that combine anaerobic and aerobic unit operations are also common.[5]

The best overall treatment approach depends on the source and nature of waste, such as production rates, constituents, and relative concentrations. As such, optimal process trains and designs should be as simple as possible in design and operation, while being efficient in removing key pollutants and minimizing energy consumption and negative by-products. More complex operations are only used when absolutely necessary.

Within a typical treatment plant, each type of treatment has a different purpose. For example, the main objective of biological treatment is to treat soluble organic matter in the wastes, which often requires physical pretreatment to remove solids before biological treatment.[2] For domestic wastewater, the main objective is to reduce the organic content and, in growing numbers of cases, secondary nutrients (nitrogen, N; phosphorus, P). For industrial wastewaters, the objective is usually to remove or reduce the concentration of organic compounds, especially specific toxicants that can be present in some wastes, which is why chemical processes are also included in industrial treatment systems. However, biological processes are almost always used when possible.

Biological degradation of organics is accomplished through the combined activity of microorganisms, including bacteria, fungi, algae, protozoa, and rotifers. To maintain the ecological balance in the receiving water, regulatory authorities have set standards for the maximum amount of the undesirable compounds present in the discharge water. In a typical wastewater treatment plant, the following steps are carried out to achieve the desired quality of the effluent before it can be safely discharged into the receiving water.

Pretreatment/Preliminary Treatment

Pretreatment is primarily used to protect pumping equipment and promote the success of subsequent treatment steps. Pretreatment devices such as screen and/or grit removal systems are designed and implemented to remove the larger suspended or floating solids, or heavy matter that can damage pumps. Sometimes, froth flotation is also used to remove excessive oils or grease in the wastes.

Primary Treatment

Most of the settleable solids are removed from the wastewater by simple sedimentation, a purely physical process. In this process, the horizontal velocity of the water through the settle is maintained at a level that provides solids adequate time to settle and floatable material be removed from the surface. Therefore, primary treatment steps consist of settling tanks, clarifiers, or flotation tanks, which send separated solids to digestion units and supernatant to subsequent, typically microbiological, treatment units.

Secondary Treatment

Secondary treatment uses microbial communities, under varying growth conditions, to biochemically decompose organic compounds in the waste that have passed from primary treatment units. An array of reactors are employed for biological treatment, which include suspended biomass, biofilm, fixed-film reactors, and pond or lagoon systems.

Secondary Clarification

Most biological treatment processes produce excess biomass through the conversion of waste carbon to new cells. As such, before the final treatment steps, such as disinfection or nutrient removal, solids must be separated from the secondary treatment effluents. This is usually by settling, but membranes are also employed. The separated solids are either recycled back to the head of the process train or sent to digesters for solids reduction and processing, depending on the type of the digester system.

Tertiary/Advanced Treatment

Advanced or tertiary treatment consists of processes that are designed to achieve higher effluent quality than attainable by conventional secondary treatment methods. These include polishing steps such as activated carbon adsorption, ion exchange, reverse osmosis, electrodialysis, chemical oxidation, and nutrient removal. Although not technically a tertiary process, final effluent disinfection is often performed after secondary or tertiary treatment using chlorination, ultraviolet methods, ozonation, and other methods designed specifically to kill residual organisms in the wastewater after all previous treatment steps.

Biological Treatment Options

Biological processes are classified according to the primary metabolic pathways present in the dominant different microorganisms active in the treatment system. As per the availability and utilization of oxygen, the biological processes are classified as aerobic, anoxic, and anaerobic.

Aerobic Processes

Treatment processes that occur in the presence of molecular oxygen (O_2) and use aerobic respiration to generate cellular energy are called aerobic processes. They are most metabolically active, but also generate more residual solids as cell mass.

Anoxic Processes

These are processes that occur in the absence of free molecular oxygen (O_2) and generate energy through anaerobic respiration. Microorganisms use combined oxygen from inorganic material in the waste (e.g., nitrate) as their terminal electron acceptor. Anoxic processes are common biological nitrogen removal systems through denitrification.[2]

Anaerobic Processes

These are the processes that occur in the absence of free or combined oxygen, and result in sulfate reduction and methanogenesis. They usually produce biogas (i.e., methane) as a useful by-product and tend to generate lower amounts of biosolids through treatment.

Apart from a classification based on microbial metabolism and/or oxygen utilization, biological wastewater treatment processes also can be classified based on the growth conditions in the reactor (see Figure 1). In this case, the two main categories are suspended growth and attached growth processes.

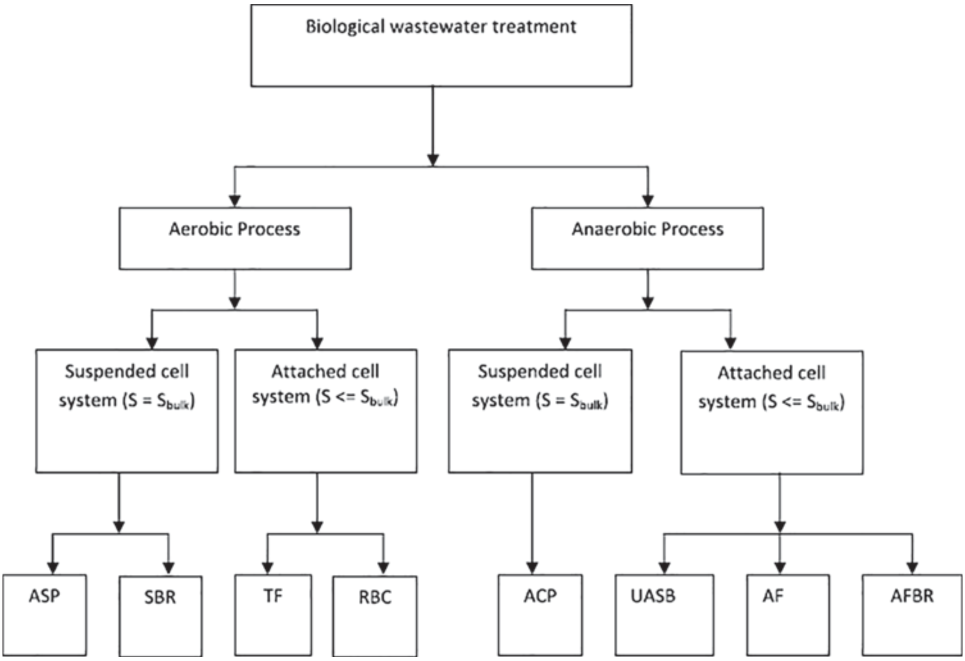


FIGURE 1 Different biological treatment processes. S, substrate concentration available to microorganisms; S_{bulk}, substrate concentration in the bulk of the liquid; ASP, activated sludge process; SBR, sequencing batch reactor; TF, trickling filter; RBC, rotating biological contactor; ACP, anaerobic contact process; AF, anaerobic filter; UASB, upflow anaerobic sludge blanket; AFBR, anaerobic fluidized bed reactor.

Suspended Growth Processes

In these processes, the microorganisms, which are responsible for the conversion of waste organic matter to simpler compounds and biomass, are maintained in suspension within the liquid phase. However, there are different types of aerobic and anaerobic suspended growth processes. Aerobic processes include activated sludge, aerated lagoons, and sequencing batch reactors, whereas anaerobic processes include bag digesters, plug-flow digesters, stirred-tank reactors, and baffled reactors with organisms primarily in the liquid phase.

Attached Growth Process

In these processes, the microorganisms responsible for degrading the waste are attached to surfaces (e.g., stones, inert packing materials), or are self-immobilized on flocs or granules in the system. Attached growth processes can be aerobic or anaerobic. Aerobic attached growth processes include trickling tilters, roughing filters, rotating biological contactors, and packed-bed reactors. Anaerobic systems include upflow packed-bed reactors, down-flow packedbed reactors, anaerobic rotating biological contactors, anaerobic fluidized bed reactors, upflow anaerobic sludge blanket (UASB) reactors, and various hybrid anaerobic reactors (HAR). UASBs are widely used reactors for the anaerobic treatment of industrial and domestic wastewater.

Aerobic Biological Waste Treatment Processes

Typical aerobic waste treatment systems provide a location where microbes are exposed to molecular oxygen (O_2) to oxidize complex organics present in the waste, producing carbon dioxide, simple organics, and new cell biomass. The activated sludge process (ASP) is very well known and the most widely used biological treatment process in developed countries.

Activated Sludge Process

Classic ASPs are aerobic suspended cell systems. Mineralization of waste organic compounds is accompanied by the formation of new microbial biomass and sometimes the removal of inorganic compounds, such as ammonia and phosphorus, depending on the particular process design. Activated sludge processes were first conceived in the early 1900s with the word "activated" referring to solids that catalyze the degradation of the waste. It was subsequently discovered that the "activation" part of the sludge was a complex mixture of microorganisms. The liquid in activated sludge systems is called the "mixed liquor," which includes both wastewater and the resident organisms.

There have been several incarnations of the ASP. The most common designs use conventional, step aeration, and continuous-flow stirred-tank reactors.[2] A conventional ASP consists of standard pretreatment steps, an aeration tank, and a secondary clarifier, an example of which is shown in Figure 2. The aeration tank can be aerated by subsurface or surface aerators designed to supply adequate dissolved oxygen to the water for the microorganisms to thrive. The wastewater flows through the tank and resident microorganisms consume organic matter in the wastewater. The aeration tank effluent flows to the clarifier where the microorganisms are removed. The clarifier supernatant is then transferred to disinfection or treatment units, and then ultimately discharged to the receiving water. Biosolids from the settler are recycled back to the head of the treatment system or sent to digesters for further processing.

Aeration Tanks

Aeration tanks are usually designed uncovered, open to the atmosphere. Air is supplied to the microorganisms by two primary methods: mechanical aerators or diffusers. Mechanical aerators, such as

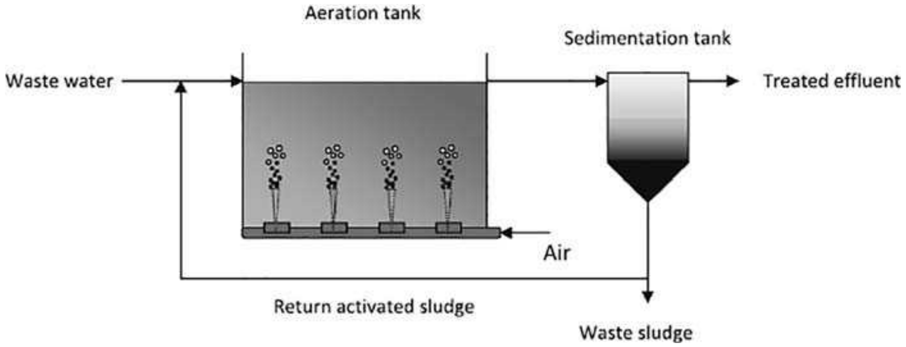


FIGURE 2 Activated sludge process.

surface aerators and brush aerators, aerate the surface of the water mechanically and promote diffusion of oxygen to water from the atmosphere. The concentration of dissolved oxygen in the liquid can be controlled by adjusting the speed of the rotors. Both mechanical aerators and diffusers are the largest energy consumers in aerobic biological wastewater treatment processes. Diffusers bubble air directly into the tank at depth and are usually preferred because of higher oxygen transfer efficiencies.

As previously indicated, aeration provides O_2 to the microorganisms and also serves to mix the liquor in the tank. Although complete mixing is desired, there are usually “dead zones” in the tank where anaerobic/anoxic conditions develop in poorly mixed areas. It is desirable to keep these zones to a minimum to minimize undesired odors and also problems with sludge bulking, which can reduce settling efficiency in secondary clarifiers.

Secondary Clarifiers

Clarifiers are used to separate the biomass and other solids coming out of the aeration tank by means of gravity settling. The flow rate of the liquid is maintained in such a way that the upflow velocity of the liquid is less than the settling velocity of the biosolids present in the liquid. As noted, some of the settled biosolids are returned back to the aeration tank to increase the solids’ contact time with the wastes and also maintain the desired biomass levels in the aeration tank.

Important Operating Parameters in Activated Sludge Systems

Key operating parameters and typical values for activated sludge systems are provided in Table 1. All parameters ultimately are used to guide and pseudo-control biosolids levels, and they profoundly affect process performance. The total suspended solids in the aeration tank are known as mixed-liquor suspended solids (MLSS). This term refers to the amount of solids in a certain volume of the water (usually milligram of solids per liter). The actual biomass fraction of the solids is estimated as the solids that can be volatilized at 550°C. The volatile fraction is known as mixed-liquor volatile suspended solids (MLVSS). Therefore, MLVSS is frequently used as a proxy for the active biomass treating the waste. MLVSS ranges from about 70% to 90% of the MLSS concentration in most activated sludge systems.[6]

Solid Retention Time

The most important design parameter in activated sludge systems is the mean cell residence time of cells in the reactor, also known as the sludge age or solid retention time (SRT). The SRT can be controlled by manipulating the rate at which excess sludge is wasted and is influenced by hydraulic flow conditions through the reactor. It is the ratio of the total solids in the system and the total solids leaving the system.

TABLE 1 Typical Design Parameters for ASP

Process Components or Variables	Typical Values	Reference
Aeration tank		[2]
Depth (m)	5–8	
Width (m)	7–12	
SRT (day)	5–15	[2]
MLSS (kg/m3)	1500–4000	[9]
SVI (kg/m3)	40–150	[7]
F/M	0.2–0.4	[6]
Organic loading rate (kg COD/m3day)	20–60	[6]
Oxygen requirement (kg/kg COD removed)	1.4–1.6	[9]

$$\mathrm{SRT} = VX/\left(QX_{\mathrm{e}} + Q_{\mathrm{w}}X_{\mathrm{w}}\right)$$

where SRT is the mean cell residence time (day); V is the volume of aeration basin (e.g., L); X is the mixed liquor suspended solids concentration (mg/L); Q is the volumetric flow rate (e.g., L/day); X_e is the effluent suspended solids concentration (mg/L); Q_w is the waste sludge flow rate (e.g., L/day); and X_w is the waste sludge suspended solids concentration (mg/L).

Sludge Volume Index

The sludge volume index (SVI) is another key parameter and used to describe the settling characteristics of the sludge. The SVI is expressed as the volume occupied by 1 g of sludge (mL/g) after 30 min of settling time. Well- settled sludge normally yields a clear separation between the water and the sludge. However, if the sludge has any problems, such as bulking, pinpoint floc formation of tiny, poorly settling floc, or ashing, the interface between the sludge and the water may not be seen clearly. Such conditions usually result from problems in the aeration tank and cause reduced effluent quality because of poor settling in the clarifier.

Dissolved Oxygen Concentration

Microorganisms in an activated sludge system require adequate oxygen to oxidize organics in the waste. The basic oxidation reaction for organics degradation can be approximated as (stoichiometry not provided)

$$\mathrm{CHON} + \mathrm{O_2} + \text{microorganisms} \rightarrow \mathrm{CO_2} + \mathrm{H_2O} + \mathrm{NH_3} + \text{more microorganisms}$$

Organics are consumed by microorganisms, and new microbial cells are synthesized with ratio of organisms produced relative to the organics consumed being the sludge yield. As noted, oxygen is supplied by mechanical aerators or diffusers in the aeration tank. Required oxygen levels in the system depend on the process, but the design goal is to minimize oxygen addition due to energy costs. The dissolved oxygen concentration can be controlled by either adjusting the speed of the air pump or throttling the air pipes. Air pumps are more widely used to aerate the wastewater because of their lower operational and maintenance costs.

Food-to-Microorganism Ratio

The food-to-microorganism ratio (F/M) is a good indicator for designing and regulating the operation of the aeration tank.[7] The F/M ratio is expressed as the amount of organic biodegradable material [milligrams of 5-day biological oxygen demand (BOD_5)] available for the amount of microorganisms present (mg MLVSS) per day.

$$F/M = \left(QS_o\right)/X$$

where F/M is the food-to-microorganism ratio (day^{-1}); S_o is the influent BOD_5 concentration (mg/L); X is the MLVSS concentration (mg/L); and Q is the volumetric flow rate (L/day).

The targeted F/M ratio for any treatment system varies depending on the design of the system, and values can range widely. However, since influent BOD cannot be controlled, MLVSS is typically modulated by varying the return activated sludge rate from the secondary clarifier, the goal being to maintain an optimum F/M ratio for specific activated sludge design.

Organic Loading Rate

The amount of organic matter in wastewater is commonly measured by BOD5, chemical oxygen demand (COD), or the total organic carbon content.[8,9] If there are excess organics in the influent or inadequate organisms in the aeration tank, incomplete treatment will result.

Common Microorganisms in Activated Sludge Systems

Activated sludge is a complex mixture of broadly differing microorganisms.[10] Major categories are as follows: bacteria, fungi, algae, protozoa (e.g., flagellates, ciliates, and rotifers), and viruses. Viruses and pathogenic bacteria are often present in wastewater, which is the primary reason for having post-biological disinfection steps in treatment plants.

Attached Growth Processes

Attached growth processes, such as trickling filters (Figure 3), can achieve similar treatment objectives as activated sludge systems. Conversion processes in these systems are typically mass transport limited: microorganisms in the outer layers of the biofilm contribute most to the overall substrate removal. The support material in trickling filters is chosen to provide sufficiently large pore spaces to allow air through the trickling filter regardless of biofilm growth and water trickling down the filter. Wastewater is distributed using rotary arms at the top and then trickles down the filter. Trickling filters are mainly used for the oxidation of carbon and ammonia, but can also achieve denitrification when convection of air through the system is optimized.[11]

Anaerobic Wastewater Treatment Processes

Anaerobic treatment technologies are widely practiced in different industries on the basis of their requirement and suitability. The processes have some advantages and disadvantages in treating different wastes, and few of them are summarized in Table 2. Under anaerobic conditions, organic matter is degraded through the sequential and syntrophic metabolic interactions of various trophic groups of prokaryotes, including fermenters, acetogens, methanogens, and sulfate-reducing bacteria (SRB).[12,13] Metabolic interactions between these microbial groups lead to the transformation of complex organic compounds to simple compounds such as methane, carbon dioxide, hydrogen sulfide, and ammonia.[14] The digestion process is essentially accomplished in four major reaction stages involving different microorganisms in each stage.[15,16]

Stage 1: Hydrolysis—The organic waste material mainly consists of carbohydrates, proteins, and lipids. Complex and large substances are broken down into simpler compounds by the activity of the microbes

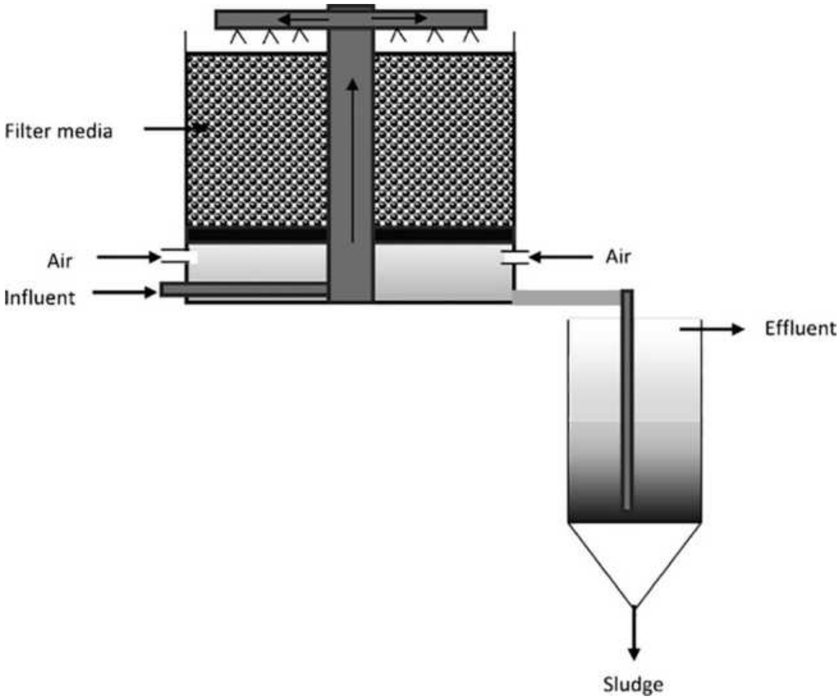


FIGURE 3 Aerobic trickling filter.

TABLE 2 Advantages and Disadvantages of Anaerobic Wastewater Treatment

Advantages

High efficiency: Good removal efficiency can be achieved in the system, even at high loading rates and low temperature.

Simplicity: The construction and operation of these reactors are relatively simple.

Flexibility: Anaerobic treatment can easily be applied on either a very large or a very small scale.

Low energy consumption: As far as no heating of the influent is needed to reach the working temperature and all plant operations can be done by gravity, the energy consumption of the reactor is almost negligible.

Energy recovery: Energy is produced during the process in the form of methane.

Low sludge production: Sludge production is low, well stabilized, and has good dewatering property.

Low nutrient and chemical requirement: Especially in the case of sewage, an adequate and stable pH can be maintained without addition of chemicals.

Disadvantages

Low pathogen and nutrient removal: Pathogens and nutrients are partially removed and hence post-treatment is needed.

Long start-up: Due to low growth rate of methanogenic organisms, the start-up takes longer time.

Possible bad odor: Hydrogen sulfide is produced. Proper handling of biogas is required to avoid bad smell.

Necessity of post-treatment: Post-treatment of the anaerobic effluent is generally required to reach the discharge standards for organic matter and pathogen.

Source: Data from Seghezzo et al.[22]

and the extracellular enzymes released by these microbes. The hydrolysis or solubilization is mainly done by hydrolytic microbes such as *Bacteroides, Bifidobacterium, Clostridium,* and *Lactobacillus.* These organisms hydrolyze complex organic molecules (cellulose, lignin, proteins, lipids) into soluble monomers such as amino acids, glucose, fatty acids, and glycerol. These hydrolysis products are used by the fermentative acidogenic bacteria in the next stage.[14,17]

Stage 2: Acidogenesis—Fermentative acidogenic bacteria convert simple organic materials such as sugars, amino acids, and long-chain fatty acids into short-chain organic acids such as formic, acetic, propionic, butyric, valeric, isobutyric, isovaleric, lactic, and succinic acids; alcohols and ketones (ethanol, methanol, glycerol, and acetone); carbon dioxide; and hydrogen. Generally, acidogenic bacteria have high growth rates and are the most abundant bacteria in any anaerobic digester.[18] The high activity of these organisms implies that acidogenesis is never the rate-limiting step in the anaerobic digestion process.[19] The volatile acids produced in this stage are further processed by microorganisms characteristic for the acetogenesis stage.

Stage 3: Acetogenesis—In this stage, acetogenic bacteria, also known as obligate hydrogen-producing acetogens, convert organic acids and alcohols into acetate, hydrogen, and carbon dioxide, which are subsequently used by methanogens and SRB. There is a strong symbiotic relationship between acetogenic bacteria and methanogens. Methanogens and SRB use hydrogen, which helps achieve the low hydrogen pressure conditions required for acetogenic conversions.[20]

Stage 4: Methanogenesis—It is the final stage of anaerobic digestion where methanogenic archaea convert the acetate, methanol, methylamines, formate, and hydrogen produced in the earlier stages into methane. The growth rate of methanogens is very low, and therefore, in most cases, this step is considered as the rate-limiting step of the anaerobic process, although there are also examples where hydrolysis is rate limiting.[21]

UASB Reactors

The most common and widely used anaerobic reactor is the UASB reactor.[22] It is an attached, self-immobilized cell system, which consists of a bottom layer of packed sludge bed (sludge blanket) and an upper liquid layer, as shown in Figure 4.[23]

Wastewater flows upward through a sludge bed consisting of bacterial aggregates floating blanket, and the microbes present in the sludge bed convert the complex organic materials to methane, carbon dioxide, and hydrogen.[24] The granular sludge (1–5 mm in diameter) has high biomass content (MLVSS) and specific activity, and good settling properties. The upward flow of the liquid inside the reactor is obtained by means of effluent recirculation. Because of the high density of biomass present in the self-immobilized granular sludge, the reactor is able to support a high SRT, which is diverse from the hydraulic retention time (HRT) and require no support material. The major drawback of the UASB is the requirement of high HRT to achieve desired biodegradation. Maintenance of high HRT demands huge reactor volume. These problems are overcome by using HAR where the advantages of AFBR are coupled with UASB operation by maintaining a high upflow velocity (4–8 m/hr) inside the reactor.[25] With higher upflow velocity, better mass transfer is obtained in the reactor, which reflects on the higher degradation with less HRT operation. The main purpose of these reactors is to achieve better degradation of waste and increase the production of biogas (methane) in a substantially reduced-size anaerobic reactor.

Important Operating Parameters in Anaerobic Reactors

Different operating parameters such as pH, temperature, HRT, and nutrients, among others, and their disturbances can manifest in case of industrial wastewaters treatment in anaerobic reactors, even under normal operational conditions.[26,27] Some of these factors are discussed below.

pH

The optimum degradation is achieved when the pH value of wastewater in the digester is maintained between 6.5 and 7.5. In the initial period of fermentation, as large quantities of organic acids are produced by acidogens and acetogens, a drop in pH occurs inside the digester. This low pH condition inhibits methanogens and subsequently reduces methane production. As the digestion proceed, the pH

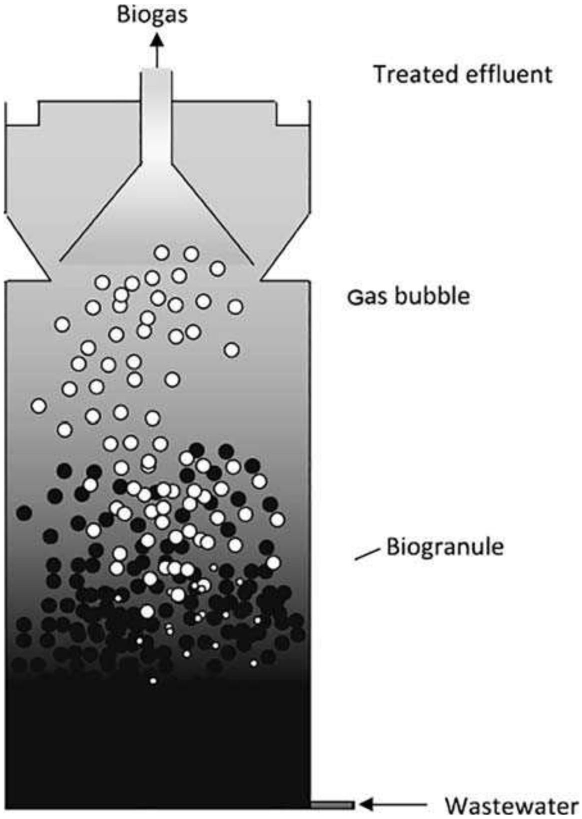


FIGURE 4 Upflow anaerobic sludge blanket reactor.

increases owing to the conversion of organic nitrogen to NH_4. When the methane production level is stabilized, the pH range remains buffered between 7.2 and 7.8.[28,29]

Waste Composition

To attain optimum degradation, wastewaters have to be nutritionally balanced in terms of carbon (C), nitrogen (N), phosphorous (P), and sulfur (S). The C/N/P ratio of 700:5:1 is recommended for efficient anaerobic digestion.[30] A fairly high concentration of acetate is required to prevent SRB outcompeting methanogens for acetate and hydrogen.[31]

Temperature

Methanogens are inactive at extremely high and low temperatures.[32] Few psychrophilic methanogens have been discovered, which can grow at a temperature range of 4–6°C.[33] Most of the methanogens can grow well from 25°C to 65°C temperatures.[34] The optimum temperature for the growth of the mesophilic methanogens is 35–37°C.[34] When the ambient temperature goes down to 10°C, gas production virtually stops. Satisfactory gas production takes place in the mesophilic range, from 30°C to 40°C.

Loading Rate

High organic loading rate may lead to acid accumulation and reduction of methane production. Similarly, if the plant is underfed, the gas production will also be low.[15]

Retention Time

The retention time depends on the growth rate of the microbial population and reactor configuration (attached cell or suspended cell system), waste strength, and waste composition.

Toxicity

The presence of toxicants in the wastewater, such as oxygen (lethal to obligate anaerobes), ammonia, chlorinated hydrocarbons, aromatic hydrocarbons, heavy metals, and long-chain fatty acids, among several others, may also result in occasional failures of anaerobic digesters.[15] The presence of trace amount of metals (e.g., nickel, cobalt, molybdenum) also stimulates the growth of microbes. Excess volatile fatty acid (VFA) concentrations are reported to inhibit the growth of several microbial species.[35] The undissociated forms of VFA can diffuse across the cell membrane and dissociate intracellularly, which results in reduction in growth rate.[35,36] The 50% inhibition of acetoclastic-methanogenesis in granular sludge was observed at a concentration of 13,000, 3,500, and 15,000 mg/L of acetate, propionate, and butyrate, respectively.[37] Small amounts of sulfide, a vital sulfur source, are beneficial for methanogens.[38] Acetoclastic methanogens are the most sensitive in terms of sulfide inhibition. Fifty percent inhibition was observed at total sulfide concentrations of 220–980 mg/L over the pH range 6.5–8.0.[39]

Granule Deterioration

Lipids present in the wastewater creates problem by forming long-chain fatty acids during hydrolysis in the anaerobic reactor. Long-chain fatty acid imparts toxic effect to acetogenic and methanogenic microbes. It also becomes adsorbed onto the sludge, inducing sludge flotation and resulting in washout.[40] Some long-chain fatty acids also act as surfactant at neutral pH and obstruct the floc formation by lowering the surface tension between water and the hydrophobic bacteria and promote their washout.[41] Addition of polyelectrolytes (calcium salts) may prevent inhibition to some extent, but it does not prevent flotation.[42]

Biological Removal of Nitrogen

The conventional biological nitrogen removal is a two- step process, nitrification followed by denitrification. The process is slow due to low microbial activity and yield. Nitrification involves a chemolithoautotrophic oxidation of ammonia to nitrate under strict aerobic conditions. This oxidation is a result of two sequential oxidative stages: ammonia to nitrite (ammonia oxidation) and nitrite to nitrate (nitrite oxidation). Different microorganisms involved in these stages use molecular oxygen as an electron acceptor and carbon dioxide as carbon source. The oxidation of ammonia to nitrite is performed by nitrifier microorganisms such as *Nitrosomonas, Nitrosococcus, Nitrosospira, Nitrosovibrio,* and *Nitrosolobus.* In the nitrite oxidation stage, *Nitrobacter, Nitrospira, Nitrospina, Nitrococcus,* and *Nitrocystis* are known to be involved in the production of nitrate.[10,43] Ammonia uptake rate varies according to reactor configuration, substrate type, and influent ammonium concentration. Denitrification is the second stage of the nitrogen removal process. It is a heterotrophic bioconversion process carried out by the heterotrophic denitrifiers under anoxic conditions. The oxidized nitrogen compounds (NO_2^- and NO_3^-) are reduced to nitrogen gas by the denitrifiers that use nitrite and/or nitrate as terminal electron acceptors and organic matter as carbon and energy source. *Pseudomonas, Alcaligenes, Paracoccus, Thiobacillus,* and *Halobacterium* are commonly found in dentrification systems.[44]

Few advanced processes, including partial nitrification, anaerobic ammonium oxidation (Anammox) and autotrophic nitrogen removal (Canon) are also being practiced in different treatment plants according to the characteristics of the wastewater. A combined system of partial nitrification and Anammox is advantageous as no extra carbon addition is needed, a negligible amount of sludge is produced, and less energy and oxygen are required compared with the conventional two-stage process.[45]

Sharon Process

The Sharon (single-reactor high-activity ammonium removal over nitrite) process is used for removal of ammonia through nitrite formation.[45,46] In this process, both autotrophic nitrification and heterotrophic denitrification take place in a single reactor with intermittent aeration. The denitrification in the Sharon process is achieved by adding methanol as a carbon source. Although the process is not suitable for all wastewaters due to a high temperature dependency, the Sharon process is suitable for removing nitrogen from waste streams having high ammonia concentrations (>0.5 g/L).

Anaerobic Ammonium Oxidation

Anaerobic ammonium oxidation (Anammox) is a highly exergonic, lithoautotrophic biological conversion process where ammonia becomes converted to nitrogen by the activity of a group of planctomycete bacteria.[47] These microorganisms use CO_2 as the sole carbon source and have a capability to oxidize ammonia to gaseous nitrogen by using nitrite as the electron acceptor in an anoxic condition.

Combined Nitrogen Removal

Ammonia-rich wastewater can be treated by Anammox, which requires nitrite as precursor. Thus, before feeding into the Anammox process, ammonia has to be preoxidized to nitrite. Thus, a partial Sharon process can be used before the Anammox process to improve the nitrogen removal efficiency. Partial nitritation (conversion of 55%–60% of ammonium to nitrite) is achieved in the Sharon process without heterotrophic denitrification. Nitrite-rich waste is then treated in an Anammox reactor. In the partial Sharon-Anammox digester, overall 83% ammoniacal nitrogen removal can be obtained from the waste stream has a total nitrogen load of 0.8 kg N/m^3/day.[48]

Canon Process

The Canon (completely autotrophic nitrogen removal over nitrite) process is also the combination of partial nitritation and Anammox processes. In this process, two groups of aerobic and anaerobic microorganisms (e.g., *Nitrosomonas* and planctomycetes) perform two sequential reactions in a single and aerated reactor. The nitrifiers consume oxygen and oxidize ammonia to nitrite. Consumption of oxygen creates an anoxic condition the Anammox process needed. The performance of the Canon process is very much dependent on operational parameters such as dissolved oxygen, biofilm thickness, nitrogen-surface load, and temperature.[49]

Biological Phosphorus Removal

The removal of phosphorus from the wastewater by the biological means is known as biological removal of phosphorus. The groups of microorganisms that are largely responsible for phosphorus removal are known as the polyphosphate-accumulating organisms (PAOs). These organisms are able to store phosphate as intracellular polyphosphate, leading to phosphorus removal from the bulk liquid phase through PAO cell removal in the waste activated sludge. Enhanced biological phosphorus removal can be achieved through the ASP by recirculating sludge through anaerobic and aerobic conditions.[50] Unlike most other microorganisms, PAOs can take up carbon sources such as VFAs under anaerobic conditions, and store them intracellularly as carbon polymers, namely poly-β-hydroxy-alkanoates (PHAs). The energy for this biotransformation is mainly generated by the cleavage of polyphosphate and release of phosphate from the cell. Reducing power is also required for PHA formation, which is produced largely through the glycolysis of internally stored glycogen.[51] The principal advantages of biological phosphorous removal are reduced chemical costs and less sludge production as compared with chemical precipitation.

The different types of bacteria used in biological phosphorus removal are *Acinetobacter, Pseudomonas, Microlunatus phosphovorus, Aeromonas,* and *Lampropedia. Acinetobacter calcoaceticus* has a very high capacity to intracellularly accumulate polyphosphate from various activated sludges. It can accumulate phosphate of an amount of 0.9%–1.9% of dry cell weight.

Conclusion

Biological treatment processes have a proven track record of dealing adequately with various kinds of wastes generated by human activities. They mimic natural processes occurring in streams and rivers. Waste treatment processes are increasingly engineered in such a way that they perform this task efficiently with a minimal input of energy. Traditionally, treatment has relied on technological approaches designed to mimic aerobic processes occurring in the water column of streams and rivers. To become truly sustainable, however, we must move away from energy-consuming aerobic processes and switch to anaerobic treatment processes, again mimicking natural processes, but now those occurring in the anaerobic sediments of the aforementioned streams and rivers. For example, there is a new focus in the water industry to integrate these two processes into systems where the waste is initially digested in an anaerobic step followed by an aerobic polishing step. Only by integrating these two processes, and variants thereof such as partial nitrification and Anammox wastewater treatment, will waste treatment become truly energy efficient and sustainable. Finally, it should be noted that anaerobic digestion to methane is not the only sustainable option. Great strides are now being made in microbial fuel cell technology within waste treatment with chemical energy from wastes being captured as electricity. All told, we are finally beginning to see again that wastes are not problems to be solved but are valuable resources, and new technologies continue to be developed to capture this capacity.

References

1. Van de Kraats, J. Editorial. Eur. Water Pollut. Control **1997**, *7*, 3–4.
2. Tchobanoglous, G.; Burton, F.L.; Stensel, H.D.; Metcalf and Eddy, Inc. Fundamentals of biological treatment. In *Wastewater Engineering: Treatment, Disposal, Reuse,* 6th Ed.; McGraw-Hill: New York, 2003.
3. Lettinga, G. Anaerobic digestion and wastewater treatment systems. Antonie van Leeuwenhoek **1995**, *67,* 3–28.
4. Lettinga, G. Sustainable integrated biological wastewater treatment. Water Sci. Technol. **1996**, *33,* 85–98.
5. Jewell, W.J. Resource-recovery wastewater treatments with biological systems. Proceedings of the Workshop on Sustainable Municipal Wastewater Treatment Systems, ETCWASTE, Leusden, the Netherlands, 1996; 67–101.
6. Qasim, S.R. *Wastewater Treatment Plants: Planning, Design and Operation*; Holt, Rinehart and Winston: New York, 1985.
7. WPCF, Natural Systems for Wastewater Treatment. *Manual for Practice*, prepared by Task Force on Natural Systems, Sherwood C. Reed, chairman, Water Pollution Control Federation, Alexandria VA, 1990.
8. American Public Health Association. *Standard Methods for the Examination of Water and Wastewater,* 17 th Ed.; American Public Health Association: Washington, DC, 1989.
9. Water Environment Federation. *Design of Municipal Wastewater Treatment Plants,* 4th Ed.; Water Environment Federation: Alexandria, VA, 1998; Manual of Practice No. 8.
10. Rittmann, B.E.; McCarty, P.L. *Environmental Biotechnology: Principles and Applications*; McGraw-Hill: New York, 2001.
11. Henze, M.; van Loosdrecht, M.C.M.; Ekama, G.A.; Brdjanovic, D., Eds. *Biological Wastewater Treatment. Principles, Modelling and Design*; IAW Publishing: London, 2008.

12. Zehnder, A.J.B., Ed. *Biology of Anaerobic Microorganisms*. John Wiley and Sons: New York, 1988.
13. Colleran, E.; Finnegan, S.; Lens, P. Anaerobic treatment of sulphate-containing waste streams. Antonie van Leeuwenhoek **1995**, *67* (1), 29–46.
14. Polprasert, C. *Organic Wastes Recycling;* Wiley: Chichester, 1989.
15. Stafford, D.A.; Wheatley, B.I. *Anaerobic Digestion;* Applied Science Pub. Ltd.: London, 1979.
16. Bitton, G. *Wastewater Microbiology,* 4th Ed.; Wiley- Blackwell: New York, 1994.
17. Speece, R.E. Anaerobic biotechnology for industrial wastewater treatment. Environ. Sci. Technol. **1983**, *17,* 416A–427A.
18. Zeikus, J.G. Microbial populations in digesters. In *Anaerobic Digestion;* Stafford, D.A., Wheatley, B.I., Hughes, D.E., Eds.; Applied Science Pub. Ltd.: London, U.K., 1980; 61–89.
19. Gujer, W.; Zehnder, A.J.B. Conversion processes in anaerobic digestion. Water Sci. Technol. **1983**, *15,* 127–167.
20. Dolfing J. Acetogenesis. In *Biology of Anaerobic Microorganisms;* Zehnder A.J.B., Ed.; John Wiley and Sons: New York, 1988; 417–468.
21. Gavala, H.N.; Angelidaki, I.; Ahring, B.K. Kinetics and modelling of anaerobic digestion process. Adv. Biochem. Eng. Biotechnol. **2003**, *81,* 57–93.
22. Seghezzo, L.; Zeeman, G.; van Lier, J.B.; Hamelers, H.V.M.; Lettinga, G. A review: The anaerobic treatment of sewage in UASB and EGSB reactors. Bioresour. Technol. **1998**, *65*, 175–190.
23. Lettinga, G.; van Velsen, A.F.M.; Hobma, S.W.; de Zeeuw, W.; Klapwijk, A. Use of up-flow sludge blanket (USB) reactor concept for biological wastewater treatment, especially for anaerobic treatment. Biotechnol. Bioeng. **1980**, *22*, 699–734.
24. Schink, B. Principles and limits of anaerobic degradation: Environmental and technological aspects. In *Biology of Anaerobic Microorganisms;* Zehnder, A.J.B., Ed.; John Wiley and Sons: New York, 1988; 771–846.
25. Ahammad, Sk. Z.; Gomes, J.; Sreekrishnan, T.R. A comparative study of two high cell density methanogenic bioreactors. Asia Pac. J. Chem. Eng. **2011**, *6*, 95–100.
26. Dolfing J. Granulation in UASB reactors. Water Sci. Technol. **1986**, *18* (12), 15–25.
27. Punal, A.; Lema, J.M. Anaerobic treatment of wastewater from a fish-canning factory in a full-scale upflow anaerobic sludge blanket (UASB) reactor. Water Sci. Technol. **1999**, *40* (8), 57–62.
28. Gerardi, M.H. *The Microbiology of Anaerobic Digesters;* John Wiley and Sons, Inc., Hoboken, New Jersey, USA, 2001.
29. Kim, In S.; Hwang, Moon H.; Jang, Nam J.; Hyun, Seong H.; Lee, S.T. Effect of low pH on the activity of hydrogen utilizing methanogen in bio-hydrogen process. Int. J. Hydrogen Energy **2004**, *29,* 1133–1140.
30. Sahm, H. Anaerobic wastewater treatment. Adv. Biochem. Eng. Biotechnol. **1984**, *29,* 84–115.
31. Lawrence, A.W.; McCarty, P.L.; Guerin, F.J.A. The effects of sulfides on anaerobic treatment. Air Water Int. J. **1966**, *110,* 2207–2210.
32. McHugh, S.; Carton, M.; Collins, G.; O'Flaherty, V. Reactor performance and microbial community dynamics during anaerobic biological treatment of waste waters at 16–37°C. FEMS Microbiol. Ecol. **2004**, *48,* 369–378.
33. Nozhevnikova, A.N.; Zepp, K.; Vazquez, F.; Zehnder, A.J.B.; Holliger, C. Evidence for the existence of psychrophilic methanogenic communities in anoxic sediments of deep lakes. Appl. Environ. Microbiol. **2003**, *69* (3), 1832–1835.
34. Bergey D.H.; Holt, J.G.; Krieg, N.R.; Sneath, P.H.A. *Bergey's Manual of Determinative Bacteriology,* 9th Ed.; Lippincott Williams and Wilkins: Philadelphia, PA, 1994.
35. van den Heuvel, J.C.; Beeftink, H.H.; Verschuren, P.G. Inhibition of the acidogenic dissimilation of glucose in anaerobic continuous cultures by free butyric-acid. Appl. Microbiol. Biotechnol. **1988**, *29* (1), 89–94.
36. Gyure, R.A.; Konopka, A.; Brooks, A.; Doemel, W. Microbial sulfate reduction in acidic (pH-3) strip-mine lakes. FEMS Microbiol. Ecol. **1990**, *73* (3), 193–201.

37. Dogan, T.; Ince, O.; Oz, N.A.; Ince, B.K. Inhibition of volatile fatty acid production in granular sludge from a UASB reactor. J. Environ. Sci. Health A **2005**, *40* (3), 633–644.
38. Daniels, L.; Belay, N.; Rajagopal, B.S. Assimilatory reduction of sulfate and sulfite by methanogenic bacteria. Appl. Environ. Microbiol. **1986**, *51* (4), 703–709.
39. O'Flaherty, V.; Colohan, S.; Mulkerrins, D.; Colleran, E. Effect of sulphate addition on volatile fatty acid and ethanol degradation in an anaerobic hybrid reactor. II: Microbial interactions and toxic effects. Bioresour. Technol. **1999**, *68* (2), 109–120.
40. Hwu, C.S.; van Beek, B.; van Lier, J.B.; Lettinga, G. Thermophilic high-rate anaerobic treatment of wastewater containing long-chain fatty acids: Effect of washed out biomass recirculation. Biotechnol. Lett. **1997**, *19* (5), 453–456.
41. Daffonchio, D.; Thaveesri, J.; Verstraete, W. Contact angle measurement and cell hydrophobicity of granular sludge from upflow anaerobic sludge bed reactors. Appl. Environ. Microbiol. **1995**, *61,* 3676–3680.
42. Hanaki, K.; Matsuo, T.; Nagase, M. Mechanisms of inhibition caused by long chain fatty acids in anaerobic digestion process. Biotechnol. Bioeng. **1981**, *23,* 1591–1560.
43. Teske, A.; Alm, E.; Regan, J.M.; Toze, S.; Rittmann, B.E.; Stahl, D.A. Evolutionary relationship among ammonia- and nitrite-oxidizing bacteria. J. Bacteriol. **1994**, *176,* 6623–6630.
44. *Zumft W.G. The denitrifying prokaryotes. In* The Prokaryotes. A Handbook on the Biology of Bacteria: Ecophysiology *Isolation Identification Applications;* Balows, A, Truper, H.G., Dworkin, M., Harder, W., Schleifer, K.H., Eds.; 2nd Ed.; Springer-Verlag: New York, 1992; Vol. 1, 554–582.
45. Jetten, M.S.M.; Schmid, M.; Schmidt, I.; Wubben, M.; Van Dongen, U.; Abma, W. Improved nitrogen removal by application of new nitrogen-cycle bacteria. Rev. Environ. Sci. Biotechnol. **2002**, *1,* 51–63.
46. Hellinga C.; Schellen A.A.J.C; Mulder J.W.; van Loosdrecht, M.C.M. The Sharon process: An innovative method for nitrogen removal from ammonium-rich wastewater. Water Sci. Technol. **1998**, *37,* 135–142.
47. Jetten, M.S.M.; Wagner, M.; Fuerst, J.; van Loosdrecht, M.C.M.; Kuenen, J.G.; Strous, M. Microbiology and application of the anaerobic ammonium oxidation (anammox) process. Curr. Opin. Biotechnol. **2001**, *12,* 283–288.
48. Jetten, M.S.M.; Horn, S.J.; Van Loosdrecht, M.C.M. Towards a more sustainable municipal wastewater treatment system. Water Sci. Technol. **1997**, *35*, 171–180.
49. van Loosdrecht M.C.M. *Recent Development on Biological Wastewater Nitrogen Removal Technologies,* In Proceedings of the International Conference on Wastewater Treatment for Nutrient Removal and Reuse (ICWNR'04); Bangkok, Thailand, 2004.
50. Barnard J.L. Biological nutrient removal without addition of chemicals. Water Res. **1975**, *9* (5–6), 485–490.
51. Mino, T.; van Loosdrecht, M.C.M.; Heijnen, J.J. Microbiology and biochemistry of the enhanced biological phosphate removal process. Water Res. **1998**, *32* (11), 3193–3207.

56

Wastewater Treatment: Conventional Methods

Sven Erik Jørgensen

Pollution Problems Associated with Wastewater

The water pollution problems associated with municipal and industrial wastewaters include their content of

- Nutrients (nitrogen and phosphorus) causing eutrophication
- Biodegradable organic matter causing oxygen depletion
- Bacteria and viruses affecting the sanitary quality of water, which is of particular importance, when the water is used for bathing, swimming, and drinking purposes
- Heavy metals, mainly lead, zinc, and cadmium from gutters, heavy metals from fungicides and other agricultural chemicals, and a wide range of other heavy metals in minor concentrations
- Refractory organic matter, originating from industries, hospitals, the use of pesticides, and the use of a wide spectrum of household entries

Wastewater Treatment

Tables 1 and 2 provide an overview of the wide range of conventional wastewater treatment methods, their efficiencies, and costs. Clearly, there is, with good approximations, a method available to virtually any of the aforementioned problems. This overview presents the conventional methods. To give a complete overview of all available methods, it is necessary to supplement this overview with a summary of low-cost methods and recently developed methods of wastewater treatment. Moreover, ecological engineering methods are not included. Particularly, the application of constructed wetland offers an attractive alternative to some of the methods presented here.

Industrial wastewaters can cause the same water pollution problems as municipal wastewater. In addition, they may contain higher concentrations of toxic organic and/ or inorganic compounds (particularly heavy metals and persistent organic pollutants). However, it is necessary in most cases to solve the problems associated with industrial wastewater at the source. They can hardly be solved by municipal wastewater treatment methods. It is also the general legislation all over the world today that industries are obliged to treat wastewater before it is discharged to the public sewage system. In many countries, the practice of the polluter-has-to-pay system has forced industries to solve their pollution problems to keep their production costs low. The major portion of toxic substances is therefore today removed by the industries themselves, at least in most industrialized countries. They are only partially removed, if at all, at municipal wastewater treatment plants, as this could contaminate the sludge produced at municipal wastewater treatment plants, thereby eliminating the possibility of the use of the sludge as a soil conditioner.

Application of the methods identified in Tables 1 and 2 gives only approximate results, and the indications should therefore be used with caution. However, first estimates such as those shown in the tables are useful for evaluations of various alternative solutions to wastewater pollution problems. The mentioned biological treatment may either be an activated sludge plant or a trickling filter of different design.

The cost of treating 100 m^3 of wastewater is based on approximate estimations, as the included cost (labor, electricity, and so on) varies from place to place, and is furthermore highly dependent on the size of the waste-water treatment plant. The costs are calculated as the running costs (electricity, labor, chemicals, and maintenance) plus 10% of the investment to cover interest and annual appreciation. The annual water consumption of one person in an industrialized country corresponds to approximately 100 m^3. For comparison, the treatment of municipal wastewater on waste stabilization ponds (WSPs) amounts to \$3–\$12/100 m^3, and that on constructed wetlands amounts to \$6–\$18/100 m^3. In most cases, WSPs cannot achieve a BOD_5 reduction above 85%–88%. In some cases, more than 90% efficiency is required. Constructed wetlands can, in most cases, with the right design, offer a fully acceptable water quality, particularly when they are used after WSPs.

The removal of high concentrations of biodegradable organic matter at the source is most often strongly recommended, since it is usually much more cost-effective to remove these pollutants, at least partially, when they are present in high concentrations. High concentrations of biodegradable organic matter are found in wastewater from slaughterhouses, starch factories, fish industries, dairies, and canned food industries.

The listed methods are often used in combinations of two or more steps to obtain the overall removal efficiency required by the most cost-moderate solution. The methods can of course also be applied in combination with cleaner technology or ecotechnology. Because wastewater treatment is often costly, it is recommended in the planning phase to examine *all* possible combinations of treatment options in order to identify the most feasible and appropriate one. For a more comprehensive presentation of all the features of the methods listed in Tables 1 and 2.[1]

A combination of treatment methods is needed in most cases, and the most applied combinations for treatment of municipal wastewater are presented in the entry "Municipal Waste Water."

Upgrading of Existing Wastewater Treatment Plants

Many existing municipal wastewater treatment plants were constructed years or decades ago, and may not meet today's higher standards. Nevertheless, upgrading existing wastewater treatment plants is possible and may be more cost-moderate than building new ones.[2,3] Because the funding allocated to pollution abatement is often limited, the overall effect of upgrading wastewater treatment plants that can be upgraded with sufficient efficiency will be to the benefit of the environment. An attractive solution is often to introduce *tertiary treatment* by chemical precipitation and flocculation in an existing mechanical–biological treatment plant, with the addition of chemicals and flocculants before the primary sedimentation phase. The installation costs for this solution are minor, and the additional running costs are limited to the costs of chemicals. The result is an 85%–95% removal of phosphorus at low cost. Similarly, nitrification and denitrification, ensuring an 80%–85% removal of nitrogen, can be realized with the installation of additional capacity for biological treatment (the overall water retention time in the plant is increased by 4–18 hr, depending on the standards and composition of the wastewater), which is considerably less costly than the installation of a completely new treatment plant.[4,5]

TABLE 1 Survey of Generally Applied Wastewater Treatment Methods

Method	Pollution Problem	Efficiency	Costs ($/100 m3)
Mechanical treatment	Suspended matter removal	0.75–0.90	3–5
	BOD5 reduction	0.20–0.35	
Biological treatment	BOD5 reduction	0.70–0.95	25–40
Flocculation	Phosphorus removal	0.3–0.6	6–9
	BOD5 reduction	0.4–0.6	
Chemical precipitation	Phosphorus removal	0.65–0.95	10–15
Al2(SO4)3 or FeCl3	Reduction of heavy metal concentrations	0.40–0.80	
	BOD5 reduction	0.50–0.65	
Chemical precipitation	Phosphorus removal	0.85–0.95	12–18
Ca(OH)2	Reduction of heavy metal concentrations	0.80–0.95	
	BOD5 reduction	0.50–0.70	
Chemical precipitation and flocculation	Phosphorus removal	0.9–0.98	12–18
	BOD5 reduction	0.6–0.75	
Ammonia stripping	Ammonia removal	0.70–0.95	25–40
Nitrification	Ammonium is oxidized to nitrate	0.80–0.95	20–30
Active carbon adsorption	COD removal (toxic substances)	0.40–0.95	60–90
	BOD5 reduction	0.40–0.70	
Denitrification after nitrification	Nitrogen removal	0.70–0.90	15–25
Ion exchange	BOD5 reduction (e.g., proteins)	0.20–0.40	40–60
	Phosphorus removal	0.80–0.95	70–100
	Nitrogen removal	0.80–0.95	45–60
	Reduction of concentrations	10–25	
Chemical oxidation (e.g., with Cl2)	Oxidation of toxic compounds	0.90–0.98	60–100
Extraction	Heavy metals and other toxic compounds	0.50–0.95	80–120
Reverse osmosis	Removes pollutants with high efficiency, but is expensive		100–200
Disinfection methods	Reduction of microorganisms	High, can hardly be indicated	6–10
Ozonation + active carbon adsorption	Removal of refractory compounds	0.5–0.95	100–120

Selection of Treatment Method

Any removal efficiency of pertinent parameters (BOD$_5$, nutrients, bacteria, viruses, toxic organic compounds, color, taste, heavy metals) can be obtained with a suitable combination of the available treatment methods, as is evident from Tables 1 and 2. However, which removal efficiencies are needed in a considered case? Because wastewater treatment is costly, the maximum allowable concentrations should not be set significantly lower than the lake or reservoir receiving the effluents can tolerate. The ban of phosphate detergents to decrease phosphorus concentrations in municipal wastewater treatment plant effluents is a point to consider in this context, as the treatment costs can be reduced considerably by the introduction of phosphorus- free detergents. On the other hand, it might be even more expensive to install an insufficient treatment plant. Thus, the potential effects of a wide range of possible pollutant inputs on water quality and on the receiving aquatic ecosystem should be assessed as the basis for selecting an acceptable option. This requires a quantification of the impacts of various possible pollutant inputs and the consideration of a wide range of solutions. All processes and components affected significantly by the impacts should be included in the quantification. It is usually very helpful to develop a water quality/ecosystem model to assist in the selection of specific

TABLE 2 Efficiency Matrix Relating Pollution Parameters and Wastewater Treatment

	Suspended Matter	BODs	COD	Total Phosphorus	Ammonium Nitrogen	Total Nitrogen	Heavy Metals	E. coli	Color	Turbidity
Mechanical treatment	0.75–0.90	0.20–0.35	0.20–0.35	0.05–0.10	~0	0.10–0.25	0.20–0.40	–	0.80–0.98	
Biological treatment[a]	0.75–0.95	0.65–0.90	0.10–0.20	0.05–0.10	~0	0.10–0.25	0.30–0.65	Fair	~0	–
Chemical precipitation	0.80–0.95	0.50–0.75	0.50–0.75	0.80–0.95	~0	0.10–0.60	0.80–0.98	Good	0.30–0.70	0.80–0.98
Ammonia stripping	~0	~0	~0	~0	0.70–0.96	0.60–0.90	~0	~0	~0	~0
Nitrification	~0	~0	~0	~0	0.80–0.95	0.80–0.95	~0	Fair	~0	~0
Active carbon adsorption®	–	0.40–0.70	0.40–0.70	~0.1	High[b]	High[b]	0.10–0.70	Good	0.70–0.90	0.60–0.90
Denitrification after nitrification	~0	–	–	~0	–	0.70–0.90	~0	Good	~0	–
Ion exchange	–	0.20–0.50	0.20–0.50	0.80–0.95	0.80–0.95	0.80–0.95	0.80–0.95	Very good	0.60–0.90	0.70–0.90
Chemical oxidation	–	Corresponding to oxidation	~0	~0	~0	~0	~0	~0	0.60–0.90	0.50–0.80
Extraction	–	Corresponding toextraction of toxic compounds	~0	~0	~0	~0	0.50–0.95	~0	~0	~0
Reverse osmosis[a]		See Table 1								
Disinfection methods	–	Much corresp. to appi, of chlorine, ozone, etc.						Very high	0.50–0.90	0.30–60

[a] Depends on the composition.

[b] As chloramines.

environmental treatment methods. It is important to emphasize that a model has an uncertainty in all its predictions that must be considered in making a final decision. Thus, it is essential to use safety factors to the benefit of the environment, in order to ensure that the selected treatment methods will have the anticipated effects. If the uncertainty is not taken into account for the sake of economy, as is unfortunately often done, the investment may be wasted because the foreseen recovery of the ecosystem will not be realized.

A problem in many developing countries is the relatively high cost of wastewater treatment. This cost justifies the application of "soft technology" or "ecotechnology," but proper planning at an early phase and the consideration of all predictable problems will always offer a wider range of cost-effective possibilities and may allow the prevention of pollution problems before they occur.

Corrections at a later stage, when pollution has already degraded the water quality and associated ecosystems, are possible, but will always be more expensive than the costs of proper wastewater treatment at an early stage. This is due in part to the fact that the accumulation of pollutants in an ecosystem over time will always cause additional problems and therefore result in additional costs. Thus, preventing pollution at an early stage is better than curing pollution at a later stage. Removal of phosphorus from wastewater at an early stage, for example, is always beneficial since the surplus phosphorus will accumulate in the sediments to a large extent, allowing its remobilization back into the water column under certain chemical conditions in the water body.

Completely new approaches have emerged in regard to sustainable development. For instance, serious consideration is being given to separation toilets in some locations, which collect urine separately from feces, thereby allowing utilization of the septic urine as fertilizer.

The selection of proper wastewater treatment methods for point sources of pollution is summarized in the following points:

- Develop models for the impacts of the wastewater on freshwater ecosystems, considering the impacts on the water quality and the entire lake ecosystem.
- Apply the model to identify the maximum allowable pollutant concentration in the treated wastewater. Any uncertainty associated with the model predictions should be reflected in identifying the lower maximum allowable concentrations.
- Select the combination of available treatment methods able to meet the standards at the lowest costs without impacting the proper operation of the plant.
- If the investment needed for a proper solution to a problem cannot be provided, the application of cost-moderate technology that will reduce the accumulation of pollutants in the aquatic ecosystem should be considered. Any measures taken at an early stage will inevitably reduce the costs at a later stage.

References

1. Jørgensen, S.E. *Principles of Pollution Abatement*; Elsevier: Amsterdam, 2000; 520 pp.
2. *Novotny, V.; Somlyódy, L., Eds.* Remediation and Management of Degraded River Basins with Emphasis on Central and Eastern Europe*; Springer Verlag: Berlin, Germany, 1995.*
3. Loosdrecht M. van. Upgrading of waste water treatment process for integrated nutrient removalxThe BCFS process. Water Sci. Technol. **1998**, *37* (9), 37–46.
4. Hahn, H.H.; Muller, N. Factors affecting water quality of (large) rivers past experiences and future outlook. In *Remediation and Management of Degraded River Basins with Emphasis on Central and Eastern Europe*; Novotny V., Somlyódy L., Eds.; Springer Verlag: Berlin, Germany, 1995; 385–426.
5. Henze, M.; Ødegaard, H. Wastewater treatment process development in Central and Eastern Europe—Strategies for a stepwise development involving chemical and biological treatment. In *Remediation and Management of Degraded River Basins with Emphasis on Central and Eastern Europe;* Novotny, V., Somlyódy, L., Eds.; Springer Verlag: Berlin, Germany, 1995; 357–384.

57

Water and Wastewater: Filters

Sandeep Joshi

Introduction

Windhoek, Namibia, a city in an arid region, suffered chronic freshwater scarcity. In the late 1960s, due to prolonged droughts, the only alternative for survival was to use treated wastewater for drinking purposes. In 1968, it became the first city in the world to directly supplement its drinking-water supplies with treated wastewater.[1] Windhoek has approximately 2,00,000 residents, with yearly increase of the population of about 5%. Since 1973, an epidemiological study of Windhoek residents has shown no adverse health impacts associated with drinking reclaimed wastewater. It provides 30% of the demand.

In 2002, this treatment plant was rebuilt, and the wastewater effluent is now the primary source for drinking water. This plant treats 24,000 m^3/day and has preozonation, flocculation, and filtration units comprising a rapid sand filter, a biological activated carbon filter, and two granular activated carbon (GACs) filters in a series, followed by ultrafiltration (UF).[2] This is a classic example of significance of filtration systems in water and wastewater treatment.

Water treatment commonly refers to the actions taken to purify water for domestic and industrial purposes. Generally, processes used for water treatment are physicochemical. Wastewater or sewage treatment processes are mainly biological. Most of the treatment plants purifying surface waters might have two stage processes—filtration and disinfection. For the wastewaters, normally, two more processes are added—pre- and postclarification. Water quality as far as suspended solids are concerned is achieved by filtration. Most of the filters are composed of granular media such as sand, anthracite, soil, etc. These filters have the ability to produce high-quality water when perfectly operated. Clarification and filtration processes are complimentary to each other.

Traditionally, settling of solids was the main process of clarifying water. Filtration was also practiced in ancient times. The earliest filters seem to be very simple—infiltration wells excavated near the banks or lakeshores. Wells on the seashore usually were dug tapping sweet water aquifers flowing from the land to the sea. The technology is modernized with addition of sophisticated UF or reverse osmosis for better-quality water. Filters can be classified based on the following: 1) filtration rate—slow, rapid, and pressured; 2) direction of flow—downflow, upflow, and biflow; 3) filtration medium—sand, soil, and synthetic media; 4) layers of media—single, dual, and multimedia; 5) input water quality—water, sewage, industrial effluents, etc.

Water or wastewater treatment is the process of removing undesirable chemical or biological contaminants. These can be gaseous or solid constituents. Concentration of gaseous contaminants is dependent on factors like solubility, temperature, partial pressure, etc. They can be removed physically by diffusion or temperature increment Size of solid contaminants may range from millimeters to angstroms—less than a nanometer. These contaminants are different types of solids that can be categorized based on their sizes or physicochemical properties. Origin of these solids can be nonbiological or biological. They can be divided into two—depending on their chemical composition—organic and inorganic, as shown in Figure 1. Suspended or dissolved solids having different sizes are shown in Figure 2. Solids larger than suspended solids are termed as floating solids like plastic, paper, cloth, etc. They can be easily removed by putting a screen across the path of water flow, but smaller particles need finer sieves.

Treatment of Water and Wastewater

Treatment includes physical, chemical, and biological processes to remove physical, chemical, and biological contaminants. Its objective is to produce an environmentally safe fluid waste stream (or treated effluent) and solid waste (or treated sludge, http://en.wikipedia.org/wiki/Sludge) suitable for disposal or

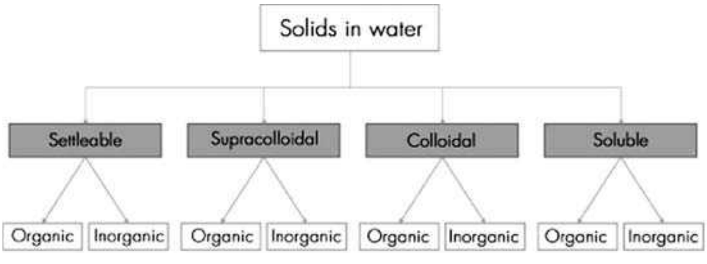


FIGURE 1 Classification of solids in water.

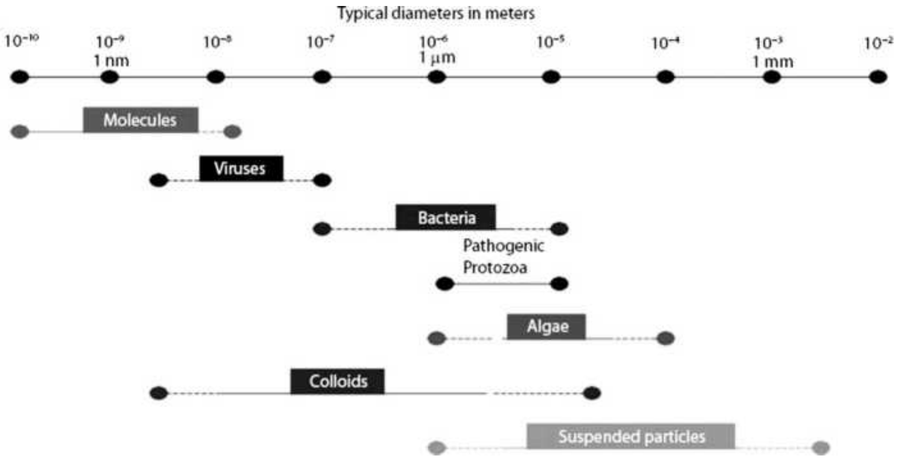


FIGURE 2 Different sizes of solid contaminants in water.

reuse (usually as farm fertilizer). Using advanced technology, it is now possible to reuse sewage effluent for drinking water. Singapore has advanced to convert sewage into drinking water on mass scale.

In general, a conventional water treatment plant usually consists of physical treatment (screening, sedimentation, flotation, and/or filtration) and chemical treatment (pH adjustment, coagulation-flocculation process, oxidation- reduction process, adsorption process).[3] The degree of the complexity of the treatment plant also depends on the quality of raw water and treated water requirement. In industrial processing, water is used in numerous applications requiring likewise different qualities of water. Examples of different uses are cooling water, water for rinsing and chemical production, boiler feed water, purified water, and water for injection. The growth in population, the increasing costs of treatment and distribution, contamination of fresh water sources, and the sophistication of end users somehow force the development of better water treatment technology.[4,5]

Filtration Process

Filtration means a process of solid-liquid separation using some form of physical barrier. In sedimentation or floatation, no barrier is used. Depending on other supportive mechanisms, the filters can be divided into mechanical or biological filters keeping the built-in solid–liquid stage. In 1685, Porzio, an Italian physician, used a filter to protect the health of soldiers in the Austro–Turkish war of that year. In 1790, James Peacock was granted the first patent for an upflow filter with graded support gravel and reverse- flow wash. James Simpson built a 1-acre slow sand filter, and it was commissioned satisfactorily in 1829. In 1858, the use of a rake was proposed to break the schmutzdecke (surface clogging layer). Holly, in the United States in 1871, patented the reverse-flow wash downflow filter, and in 1877, Cook developed the concept of a battery filter for keeping standby options during the washing of used filters.[6] In India, similarly, an underground filtration flow system was developed in Pune in 18th century to cater the people's needs for fresh and clean water.[7]

It has been noticed that very small particles and large particles are predominantly removed by diffusion and straining mechanisms, respectively. Five different mechanisms other than simple straining[8]

Generally, in conventional water or wastewater treatment plants, filtration is considered as the core of the process in which the solids are substantially removed. Incoming flow to the filters is clearly non-potable or turbid, but after filtration, it becomes clean and transparent. If the water is disinfected, then it can be used for human consumption.

The process of filtration is composed of passing water through a granular bed of sand or any other suitable medium at a designed flow rate. To achieve the desired quality of filtered water, the filtration system has to remove particles, and the sizes of the openings are crucial. It has been noticed that very small particles and large particles are predominantly removed by diffusion and straining mechanisms, respectively. Five different mechanisms other than simple straining[9] are identified in the filtration process when particles in a flow of water pass through a filter, viz., interception, sedimentation, diffusion, hydrodynamic action, and inertia (Figure 3).

Particles moving uniformly may collide with a grain of filter media; their transport is intercepted. Efficient interception is possible by decreasing the size of the media or increasing the size of the particle to be removed. Larger particles are removed more efficiently. Sedimentation is a result of gravitational forces acting on the particles depending on the ratio of settling velocity and velocity of the flow approaching the media. Diffusion as a result of Brownian motion is effective only for small particles. It becomes more pronounced with increasing temperature. Hydrodynamic action arising from the velocity gradient separates small particles from the flow. This is comparatively nonsignificant. Inertia (impaction) is useful only when hydrodynamic diversion is less. At less velocity, inertia is not significant. In addition to these mechanisms, coagulation, flocculation, adsorption, and absorption play major roles in the effectiveness of a filtration system.

Filtration with a granular bed is composed of indivisible grains that rest on each other, leaving voids in between. Dirt is removed when water flows through these voids. The real filter media are composed of irregular grains-often unpolished rounded edges. Sands are the most readily available material from natural sources such as the sea coast and riverbeds. Granular media can be obtained by crushing materials such as basalt, slag, and anthracite. The voidage increases with angularity of the material, e.g., crushed anthracite has more voidage (45%) than sand (38%) after gentle tapping. Filtration media retain solids while permitting water having turbidity less than 0.2 NTU (nephelometric turbidity unit) to pass.

Solids are collected on the surface of the filter bed, forming a surface mat or “mud cake.” This needs to be regularly cleaned to avoid resistance to flow. A granular bed filter having a mass of leading edges and space for accumulation of deposits helps in retaining the grains that are no longerin the line of flow. They are like silt on the sand bed of a river. These materials are comparatively cheaper than fibrous materials that are used for large-scale water or wastewater applications. If the grain does not dissolve in

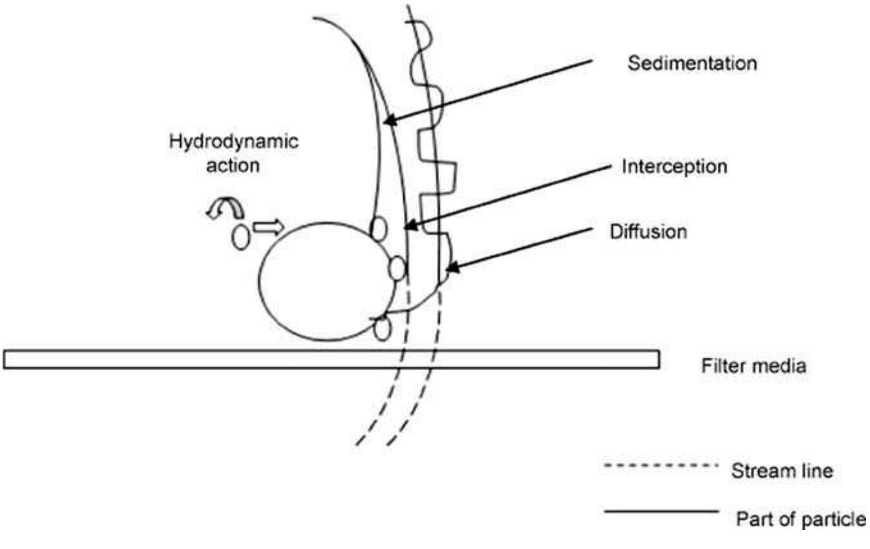


FIGURE 3 Filtration mechanisms.

water, then its chemical nature is not important. Therefore, structurally weak materials such as activated carbon (granular or powdered) can be used for filtration purposes.

Filters operate best normally at a low concentration of solids as a result of limited volume of pores (about 40% of the bed volume). A granular bed filter having a mass of leading edges and space for accumulation of deposits helps in retaining the grains that are no longer in the line of flow. Therefore, such filters operate at 0.2 bar (2 m head), but media such as a cloth, paper, porous sinter, or membrane tend to accumulate deposits, and then it requires high pressure (up to 7 bars) to maintain its permeability. The depth in simple granular filters may be 600–1000 mm for its desired effectiveness. Failure of the filters is associated with cleaning the mechanism either manually or mechanically with water or air backwash.

Types of Granular Filters

There are three basic types of granular filters, viz., sand filters, rapid gravity (RG) filters, and pressure filters. Slow sand filters—operating at low loading rates—are the oldest variety of filters. These filters made up of fine sand have two processes—physical straining and biological action. Rapid filters function at higher loading rates using coarse media with a higher permeability. In addition to sand, GAC is used to adsorb chemicals dissolved in water. Coagulation helps in effective removal of solids by RG filters. Simple RG filters use single-medium sand, but a multimedia version uses two or more types of media. These are cleaned by reversing the flow of water to wash out the dirt. Pressure filters are the variety operating under pressure in large closed vessels. These are normally used for treating groundwater sources, which are pumped from boreholes. Various types of filters are shown in Figure 4.

Slow Sand Filter

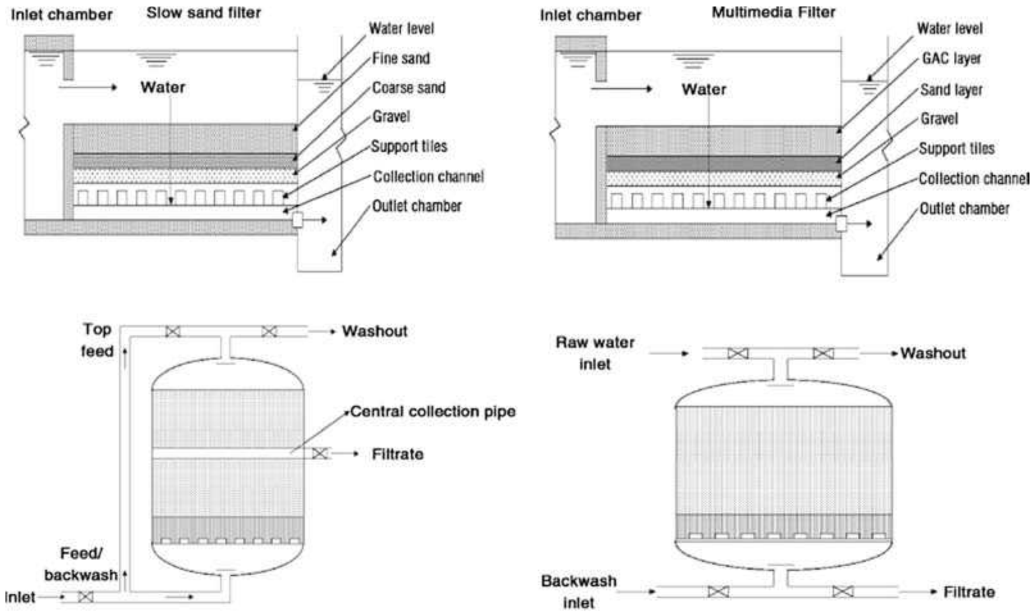
Slow sand bed filters are very-low-flow-rate filters. They are used in rural areas or where the water flow rate required is low. A slow sand filter can even remove turbidity and suspended solids. Pathogens are also removed, additionally producing water with very reasonable taste and no color. An operational advantage of slow sand filters is that no backwashing is done. They require few operator skills. Sometimes, slow sand filters are used in the treatment of sewage as a final polishing stage for the treated effluent and as a bed to dewater the sludge produced during treatment.

Rapid Gravity Filter with Coagulant Aid

Purification of water for drinking purposes is done using chemicals to flocculate as much particulate matter to be removed as quickly as possible in the filter. Coagulation of the particulate matter is achieved by adding highly charged cations, such as aluminum sulfate (alum) and small amounts of charged polymer chains. Coagulation and flocculation processes require contact time and agitation in tanks before the water is filtered so that a reasonably sized floc can be formed. This process is pH dependent. Thus, it is necessary to adjust the pH of the water to ensure efficient removal of water. A rapid gravity sand filter strains out the floc and the particles, including bacteria trapped within them. The medium of the filter has grain sizes to allow the water to pass through the filter rapidly. This filter is backwashed with clean water on a regular basis as the pores are clogged by particulate matter (flocs), indicated by a rapid drop-off in the flow rate of water through the filter. This backwash water is run into settling tanks so that the floc can settle out, and it is then disposed of as residue. Supernatant water is then run back into the treatment process or disposed of as a wastewater stream.

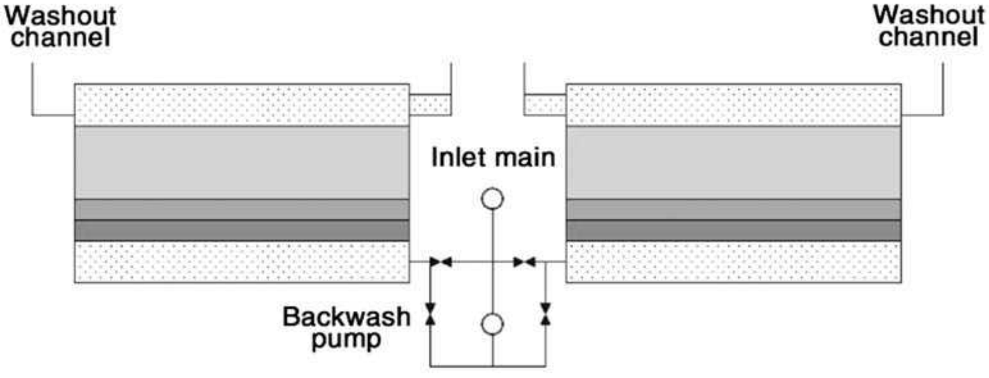
Pressure Sand Filter

Applying pressure to the water passing through a sand filter will give a greater flow rate, while using smaller grains of sand in a filter allows a greater surface area of filtration medium to remove



Central feed/collection sand filter

Gravity sand filter with backwash system



Sand filter in a series/parallel with backwash system

FIGURE 4 Different types of filters.

particulate matter. Smaller grain size requires more pressure to drive the water through the sand bed. Sand grain size in pressure filters is typically 0.6–1.2 mm, and large particulates (>100 pm) are removed. This causes rapid blinding of the filter with floc and the need for too-frequent backwashing. Pressure sand filters are typically 0.6–1.8 m in depth and operate under a maximum flow rate of about $9m^3/m^2/hr$ under a feed pressure of between 2 and 5 bars. Buildup of particulate solids causing an increase in pressure across the bed for a given flow rate will need to be backwashed when the pressure drop is around 0.5 bar. Particulate solids are then washed away with the backwash water. It is diverted to a settling tank to separate the solids from water. A small amount of sand can be lost in the backwashing process. Then, the sand bed may need to be topped up from time to time. Most pressure filters in industry or large water works employ an automated multiport valve together with sophisticated pressure and flow sensors.

Cartridge Filter

This type of filter has a removable housing, into which different types of filtration elements can be placed. A domestic cartridge filter element will often be rated at 30–50 μm or larger, whereas specialist industrial filter elements may be rated at 5 μm or less. Some cartridge filters are made effective by using some resins that are designed to remove specific contaminants from the water, such as nitrates, fluoride, or lead. Granulated activated carbon is used in filters to remove color, odor, volatile organics, and chlorine from water.

Biofilters

Biological treatment or biofiltration has become the mainstay of wastewater treatment systems over a period of time because of simplicity. Odor control using biofiltration has been patented from the 1950s for huge trickling filter plants and soil filters. Biofilters have primary applications for odor control at wastewater treatment plants and composting operations. They are being used also in the treatment of volatile organic compounds (VOCs), as an innovative method to treat toxic air emissions from industrial processes. The chronological development of biofilters is as shown below:

- 1923: To treat odorous compounds, Bach thought of using a biologically active biofilter to control emissions of H_2S from a wastewater treatment plant.
- 1955: A biofilter was applied to treat odorous emissions in low concentrations in Germany.
- 1959: A soil bed was installed at a sewage treatment plant in Nuremberg for the control of odors from an incoming sewer main.
- 1960s: Biofiltration was first used for the treatment of gaseous pollutants both in Germany and the United States.
- 1970s: Biofiltration became widespread in Germany.
- 1980s: Biofiltration is used for the treatment of toxic emissions and VOCs from industry.

Biological Aerated Filters

Biological aerated filters or biofilters combine filtration with biological carbon removal and nitrification or denitrification processes. A biologically aerated filter usually includes a reactor filled with a filter medium, which is either in suspension or supported by a gravel layer at the foot of the filter. The dual purpose of this medium is to support highly active biomass that is attached to it and to filter suspended solids. Carbon reduction and ammonia conversion occur in aerobic mode and are sometimes achieved in a single reactor, while nitrate conversion occurs in anoxic mode. Biological aerated filters are operated either in up- flow or downflow configuration.

Trickling Filters

The first trickling filter became operational in 1893. The concept of trickling filters was developed from contact filters—which were watertight basins filled with stones. Clogging, long resting period, and relatively low loading rate were some of the limitations of contact filters. The modern trickling filter is composed of a permeable medium (rock or plastic media) for the attachment of bacteria. Through this biofilter, wastewater percolates—trickles down. The filter medium is either rock (size varies from 25 to 100 mm) or slag or plastic packing materials. The depth of the rock medium is about 1–2.5 m. Wastewater is distributed on the top of the rock bed by rotary distributor. Plastic media are either round or square or irregularly shaped. Depth of plastic media varies from 4 to 12 m. Depending on packing, the trickling filter can be categorized into three groups, viz., vertical flow, cross-flow, and random.

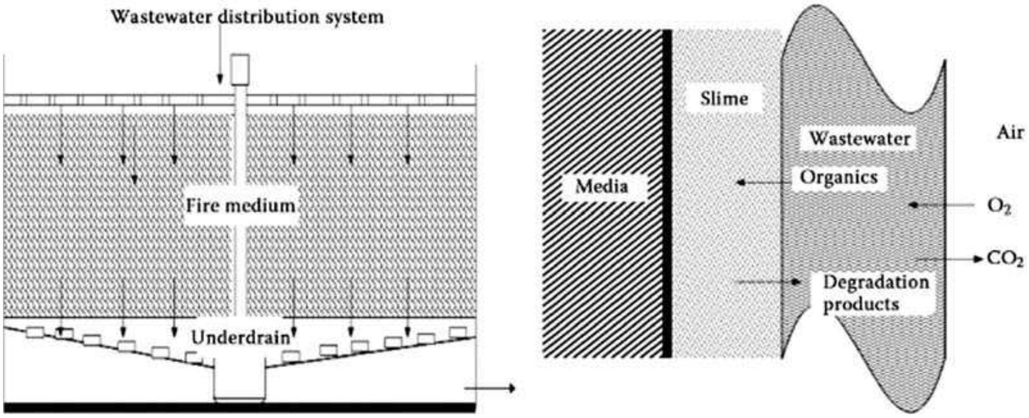


FIGURE 5 Trickling filter and slime layer.

The components of a trickling filter are the following:

- A bed of filter medium on which a layer of microbial slime is developed
- A container that houses the bed of filter medium
- Distribution of wastewater over the filter medium
- Removal and disposal of any sludge from the treated effluent

Filters are provided with underdrains for collection of treated wastewater and biosolids detached from the media. Underdrain systems are important as far as air circulation is concerned. Organic matter is degraded by the microorganisms present in the biological film or slime layer attached to the filter media (Figure 5). The outer layer of biofilm, about 0.1–0.2 mm, is aerobic, and the layer near the surface of filter media becomes anaerobic due to diffused oxygen and thus cannot penetrate the full depth of the slime layer. The inner layer is starved of nutrition of pollutants as the slime layer thickens. Endogenous degradation of bacteria in the inner layer leads to sloughing due to shear velocity of trickling water. Facultative microorganisms dominate the biological communities in the slime layer of the trickling filter. With these organisms, fungi, algae, protozoans, worms, insect larvae, and sometimes snails also dwell in the trickling filter.[9,10] Their presence in the filter is useful to maintain the bacterial population in a state of high growth or high rate of food consumption.

An excellent material for trickling filter is that having more surface area per unit volume, low cost, durability, and no tendency for clogging. Until the 1960s, a commonly used material was high-quality granite or blast furnace slag. River gravel or crushed stones were also used. All underdrains are developed to maintain and facilitate air ventilation. Sufficient airflow in the system is a key to successful operation of the trickling filter. In case of natural ventilation, the temperature difference between the ambient air and air in the pores of media is the key reason for air circulation. Good design may provide airflow at least 0.3 $m^3/minm^3$ of the filter area in either direction. Trickling filters can be classified on the following bases: 1) frequency of hydraulic and organic loading—slow rate, high rate, batch, intermittent, continuous; 2) nature of filtering media—natural, synthetic; 3) direction of inflow—downflow, up- flow, sectionalized; 4) internal environment of the filter— aerobic, anaerobic, anoxic; 5) depth of media—shallow, deep; 6) method of ventilation—natural, forced; and 7) extent of treatment expected—complete, roughing, polishing.

Membrane Filtration

The major breakthrough in the development of membrane technology was recorded in the late 1950s. The membrane bioreactor (MBR) has developed quite rapidly to become this important. The idea of combining sludge digestion with a very fine filter was first developed in the mid-1960s by Dorr Oliver, as a system with flat-plate membranes in a sidestream loop. This development was made possible by the appearance on the market of commercial-scale microfiltration (MF) and UF membranes. Membrane bioreactors have served industries that can afford an expensive system. The submerged membrane was innovated in Japan in 1989, which led to a Japanese government initiative to find better ways of wastewater treatment. Submerged or immersed MBRs used less energy than the sidestream version. This version entered the European market in the mid-1990s. Since 1990, the numbers of MBR installations have grown exponentially.

Technologies of secondary biological treatment systems for municipal wastewater generally involve microorganisms suspended in the wastewater to treat it. They have some deficiencies, including the difficulty of growing the exact types and numbers of microorganisms and the physical requirement of a large site. Application of MFMBRs surmounts many limitations of conventional systems. These membrane filtration systems combine the suspended biological growth with solids removal via filtration (Figure 6).

The basic MBR consists of two processing steps—a bioreactor, in which aerobic bacteria digest organic material in the presence of dissolved oxygen, and a membrane module, through which relatively pure water separates from the suspension of organic matter and bacteria. These two units may be set up to run in succession (i.e., the liquid flows first through the bioreactor, where it is held for as long as necessary for the reaction to be completed, and then through a loop containing the membrane separation stage), with a recycle of some of the separated sludge to the bioreactor. This is often called sidestream (or external) operation. Membranes are suspended in the slurry in the bioreactor, which is appropriately partitioned to achieve the correct airflows, with the surplus sludge withdrawn from the base of the bioreactor at a rate to give the required sludge retention time. This is termed as a submerged (or immersed) MBR.

Membranes can be designed for and operated in small spaces and with high removal efficiency of contaminants such as total suspended solids, nitrogen, phosphorus, biochemical oxygen demand (BOD), and bacteria. The membrane filtration system in effect can replace the secondary clarifier and sand filters in a typical activated sludge treatment system for large sewage treatment plants. Membrane filtration allows a higher biomass concentration to be maintained, thereby allowing smaller bioreactors to be used.

Membrane filtration involves the cross-flow of water containing pollutants across a membrane. Waste matters left behind do not accumulate at the membrane surface. They are removed from the system by gravity for disposal. Water passing through the membrane is called the permeate or filtrate, while water with more concentrated pollutants is called the concentrate, retentate, or reject. Membranes are composed of cellulose or other polymers having a number of pores. The requirement is that the membranes

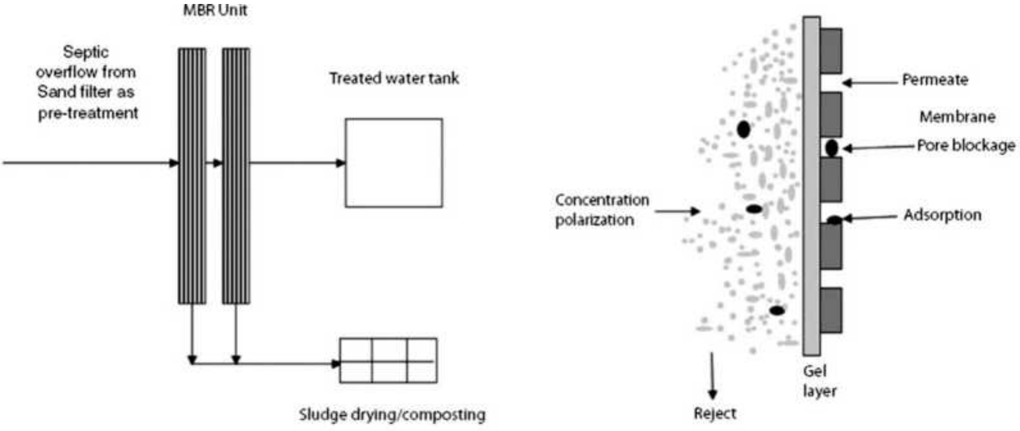


FIGURE 6 Typical flow chart of membrane filtration for sewage treatment and filtration through membrane.

filter out particles and microorganisms more than 1 μm (0.001 mm) in size. Two configurations most often used are hollow fibers grouped in bundles or flat plates for MBR designs.

Membrane filtration is an alternative to remove waterborne pathogens in water treatment, replacing the disinfection step in conventional water treatment plants to avoid risks and side effects of chemicals. Membrane processes can be classified in many ways, i.e., based on their nature, structure, or driving force. Hydrostatic pressure differences are used in MF and nanofiltration (NF), as well as in reverse osmosis (RO) and gas separation as a driving force for mass transport through the membrane. The MBR takes the place of the whole secondary stage of conventional activated sludge processing—and does it better and in a much smaller space. The excess biodegradation products can be removed for subsequent treatment.

The sludge settlement stage of the conventional secondary process is a slow one, so the removal of the clear liquid from the slurry by membrane filtration is a better option. A major advantage of the MBR system is that it can operate at a much higher concentration of solids in the bioreactor than that of a conventional activated sludge plant. An MBR plant can work effectively at concentrations of mixed liquor suspended solids typically in the range of 8000–12,000 mg/L (or 0.8%–1.2%) and can be extended up to 3%, whereas conventional activated sludge plants work at about 2000–3000 mg/L, because of the limitations on settling. This high sludge concentration capability enables an MBR system to deal effectively with strong industrial wastes, especially in places where water is short and factories are seeking to close their water cycles. It also results in a much smaller digestion tank, and thus a much smaller footprint for the whole plant. The smaller space footprint of an MBR plant will make it much more attractive for construction in developed urban areas.

The longer sludge retention times in the MBR permit the destruction of molecules difficult to biodegrade, such as detergents. With proper system design, nitrogen and phosphorus content can also be significantly reduced. The quality of water produced as effluent by an MBR is excellent and usually well under the local consent limits. Figures of less than 2 mg/L for BOD and less than 0.5 mg/L for total suspended solids are readily achievable, along with ammoniacal nitrogen of less than 0.5 mg/L and turbidity of less than 0.2 NTU. Particularly relevant are the capabilities of UF membranes (the most common form used) for removing pathogens such as protozoa, bacteria, and most viruses.

Most membrane filtration processes require quite a high transmembrane pressure in order to produce an acceptable permeate flow rate. However, an MBR operates with a low differential, about 0.5 bar. This can be provided by a vacuum pump, sucking on the permeate discharge line, through a receiver, or by the hydrostatic head of a deep bioreactor tank, or by a low level of pressurization of the tank.

Future Developments of MBR

The MBR (including its most common form, submerged MBR—not to be confused with the moving bed bioreactor) is a fascinating combination of two steps of waste treatment: the bacterial digestion of waste organic matter and the separation of the treated effluent from residual suspended solids. Its prime advantage is that it can constitute the whole of the secondary phase of the traditional three- phase municipal sewage treatment process and can eliminate some parts of the tertiary phase.

In the long history of wastewater treatment (over 110 years), the MBR is quite a recent invention, and so, not surprisingly, it is still in a period of intense development. Key areas of system investigation include the following:

- Aeration system design, to ensure minimal energy consumption and maximum mixing and scouring
- Use of enriched air or pure oxygen for aeration
- The possible use of non-membrane MF media
- The design and manufacture of MBR membranes truly defying fouling, enabling a realistic “fit-and-forget” situation

- Development of an MBR capable of continuous use
- Use of MBR "in reverse" to add nutrients in a controlled fashion to remove gases from bioreactor
- Extraction of specific pollutants from bioreactor zone
- Development of an MBR for use in gray water recycling
- Conversion of MBR from an aerobic process to anaerobic operation, thus producing more energy (in the form of methane) than the plant consumes, the anaerobic digestion UF system

A major disadvantage of MBR systems is the higher capital and operating costs than conventional systems for the same quantum of wastewater. Operation and maintenance costs include membrane cleaning, fouling control, and eventual replacement of membrane Energy costs are also higher because of the need for air scouring to control bacterial growth on the membranes. Additionally, the waste sludge from such a system may have a low settling property.

Applicability

Membrane bioreactor systems are also well suited for some industrial and commercial applications. The high- quality treated water produced by MBRs makes them particularly applicable to reuse applications and for surface water discharge applications requiring extensive nutrient (nitrogen and phosphorus) removal. Designers of MBR systems require only basic information about the wastewater characteristics (e.g., influent characteristics, effluent requirements, flow data) to design an MBR system. Depending on effluent requirements, certain supplementary treatment steps can be clubbed with the MBR system, e.g., chemical coagulation before the primary settling tank or before the secondary settling tank, before the MBR or final filters for phosphorus removal.

Ultrafiltration

Ultrafiltration is a fractionation technique that can simultaneously concentrate macromolecules or colloidal substances in a process stream.[11] It can be described as simultaneous purification, concentration, and fractionation of macromolecules or fine colloidal suspensions.

Ultrafiltration membranes are developed from both organic (polymer) and inorganic materials. The preference of a given polymer as a membrane material is dependent on very specific properties such as molecular weight, chain flexibility, and chain interaction. Some of these materials are polysulfone, polyethersulfone, polyvinylidene fluoride, polyacrylonitrile, cellulosics, polyimide, polyetherimide, and aliphatic polyamides. Inorganic materials have also been used such as alumina and zirconia.[9] The structure of a UF membrane can be symmetric or asymmetric. Thickness of symmetric membrane ranges from 10 to 200 μm. Resistance to mass transfer is determined by the thin top layer. Membrane material should have chemical, thermal, and mechanical stability. Specification of UF is determined based on molecular weight cut-off—a measure of membrane pore dimensions—to describe the retention capabilities of UF membrane. It is the molecular mass of a macrosolute (typically, polyethylene glycol, dextran, or protein) for which the membrane has a retention capability greater than 90%.

Performance of UF

One of the critical factors determining the overall performance of aUF system is the rate of solute or particle transport in the feed side. Steady state conditions are achieved when the convective transport of solute to the membrane is equal to the sum of the permeate flow plus the diffusive back-transport of the solute.[12] The accumulation of solutes/particles at the membrane surface can affect the permeate flux in two ways.[13] First, the accumulated solute can generate an osmotically driven fluid flow back across the membrane from the permeate side toward the feed side, thereby reducing the net rate of solvent transport. Second, the solutes/particles can irreversibly foul the membrane due to specific physical and/or chemical interactions, thereby generating an additional hydraulic resistance. These interactions can be attributed to one or more of the following mechanisms: 1) adsorption; 2) gel layer formation; and 3) plugging of the membrane pores. The severity depends on the membrane material, the nature of solutes, and other variables such as pH, ionic strength, solution temperature, and operating pressure.[14]

Ultrafiltration System Design

Ultrafiltration is a low-pressure system operating at transmembrane pressures of 0.5–5 bars. A UF system comprises series/parallel modules operating according to various modes, ranging from an intermittent single-stage system to a continuous multistage system.[15] Ultrafiltration membranes can be fabricated essentially in tubular or flat sheet forms. Two major types of UF modules can be used, i.e., hollow fibers (capillary) and spiral wound. Other modules are plate and frame, tubular, rotary modules, vibrating modules, and Dean vortices. Operation of a UF membrane can be performed in two different service modes, viz., dead-end flow and cross-flow. The dead-end flow approach allows optimal recovery of feed water in about a 95%–98% range, but is generally limited to feed streams of low suspended solids (<1 NTU). The cross-flow mode differs from the dead-end mode in that there is an additional flow—the concentrate. The cross-flow mode of operation typically results in lower recovery of feed water, about 90%–95%.[16]

Ultrafiltration in Water Treatment

Ultrafiltration can be used to replace clarification steps, i.e., coagulation, sedimentation, and filtration, in a conventional water treatment plant. Thus, it can be defined as a clarification and disinfection membrane operation. All particulate biological contaminants, viz., viruses and bacteria, are rejected. The main advantages of low- pressure UF membrane processes compared with conventional clarification processes are no need for chemicals and size-exclusion filtration for constant quality of treated water in terms of particle and microbial removal.

Source water quality directly impacts UF membrane performance. Depending on the quality of raw water, UF can be operated as single operation or in combination with another process (coagulation, adsorption, etc.), or hybrid membrane system (UF/MF). In water application, UF can be the main process or used as pretreatment, for example, in an RO system.[17] It is a preferred alternative to conventional technology to remove water-borne pathogens in the preparation of drinking water.[16]

Ultrafiltration alone is not very effective for removing dissolved substances in general. It has limited capability in removing organic matter. Thus, the use of powdered activated carbon (PAC) in combination with a UF membrane is attracting increasing interest for the removal of organic compounds in drinking-water treatment.[17] It is found that the combination of UF with PAC/GAC could improve the removal of organics and other micropollutants such as agrochemicals. Another potential application of UF is to produce ultrapure water, acting as a pretreatment of RO unit. Microfiltration/UF can be considered as an alternative water treatment in aquaculture.[18] A sufficient supply of good-quality water is essential to any aquaculture operation. Ultrafiltration generates highly pathogen-free water.

Reverse Osmosis System

Reverse osmosis is a hi-tech filtration method that removes many types of large molecules and ions from solutions by applying pressure to the solution. Dissolved solids are not removed by any other filtration technique. The result is that solute is retained on the pressurized side of the membrane, and the pure solvent is allowed to pass to the other side. Selectively, this membrane does not allow large molecules or ions through the pores (holes) but allows smaller components of the solution (such as the solvent) to pass freely. The process of osmosis through semipermeable membranes was first observed in 1748 by Jean Antoine Nollet. For the following 200 years, osmosis was only a phenomenon observed in the laboratory. In 1949, the University of California at Los Angeles first investigated desalination of seawater using semipermeable membranes.

In the normal osmosis process, the solvent naturally moves from an area of low solute concentration, through a membrane, to an area of high solute concentration. The movement of a pure solvent to equalize solute concentrations on each side of a membrane generates pressure, i.e., osmotic pressure. Applying an external pressure to reverse the natural flow of pure solvent is the RO. The process is very similar to membrane filtration, but the only difference is between flow of solvents and filtration. Reverse osmosis involves a diffusive mechanism so that separation efficiency is dependent on solute concentration, pressure, and water flux rate.[19] The membranes used for RO have a dense barrier layer in the polymer matrix where most separation occurs. In most cases, the membrane is designed to allow only water to pass through the dense layer, while preventing the passage of solutes such as salt ions. This process requires high pressure, usually 2–17 bars (30–250 psi) for fresh and brackish water and 40–70 bars (600–1000 psi) for seawater (http://en.wikipedia.org/wiki/Reverse_osmosis-cite_note-2http://en.wikipedia.org/wiki/Reverse_osmosis-cite_note-water-0).

One of the first membrane applications for the utilization of membrane technology was the conversion of seawater into drinking water by RO. AnRO system separates dissolved solutes (includes single charged ions, such as Na^+, Cl^-) from water. Reverse osmosis can be described as a diffusion-controlled process. Physical holes may not exist in an RO membrane, which differentiates it from other filtration systems. An RO membrane is very hydrophilic; therefore, water will be able to readily diffuse into and out of the polymer structure of the membrane. An RO membrane is capable of rejecting contaminants as small as 0.001 μm.[20]

Four types of modules are used for RO membrane, i.e., plate and frame, tubular, hollow fiber, and spiral wound. However, the spiral-wound element is the most common by far for the production of drinking water. Reverse osmosis configurations include single-stage, two-stage, and two-pass systems. A two-stage system is common for brackish water use, where it is necessary to increase the overall recovery ratio.[21] Nowadays, the RO system has become a popular water treatment technology in industry requiring desalination and in residential units to improve the taste of water and to remove unhealthy contaminants. Reverse osmosis has increased the water supply by making possible the use of brackish waters. Recent advances in membrane materials and pretreatment have made RO desalination economically attractive even for seawater. The scale of membrane applications is now very large; plants with capacity in excess of 19 million liters per day (MLD) are common.[22] An RO desalination plant with a capacity of 100 MLD is being developed in Chennai, India.[23]

Reverse osmosis membranes are comparatively more sensitive than thermal desalination processes to scaling, fouling, and chemobiological attacks. The susceptibility to fouling is one of the major shortcomings of the RO membrane. Hence, though RO is an energy-efficient alternative to thermal processes, it still continues to face competition due to the requirements of pretreatment. Pretreatment is important when working with RO and NF membranes due to the nature of their spiral-wound design (Figure 7). Material is engineered in such a fashion as to allow only one-way flow through the system. Since accumulated material can-not be removed from the membrane surface systems, they are highly susceptible to fouling. Therefore, pretreatment is a necessity for any RO or NF system.

Pretreatment has four major components:

- *Screening of solids.* Suspended solids to be removed to prevent fouling of the membranes by fine particles or biological growth to reduce the risk of damage to high- pressure pump components.
- *Cartridge filtration.* String-wound polypropylene filters used to filter particles between 1 and 5 mm.
- *Dosing.* Oxidizing biocides, such as chlorine added to kill bacteria, followed by bisulfite dosing to deactivate the chlorine, which otherwise can destroy a thin-film composite membrane.
- *Prefiltration pH adjustment.* If the pH, hardness, and alkalinity in the feedwater result in a scaling tendency (estimated using the Langelier saturation index), acid is dosed to maintain carbonates in their soluble carbonic acid form. Use of antiscalant is recommended[24] (http://en.wikipedia.org/wiki/Reverse_osmosis - cite_note-8).

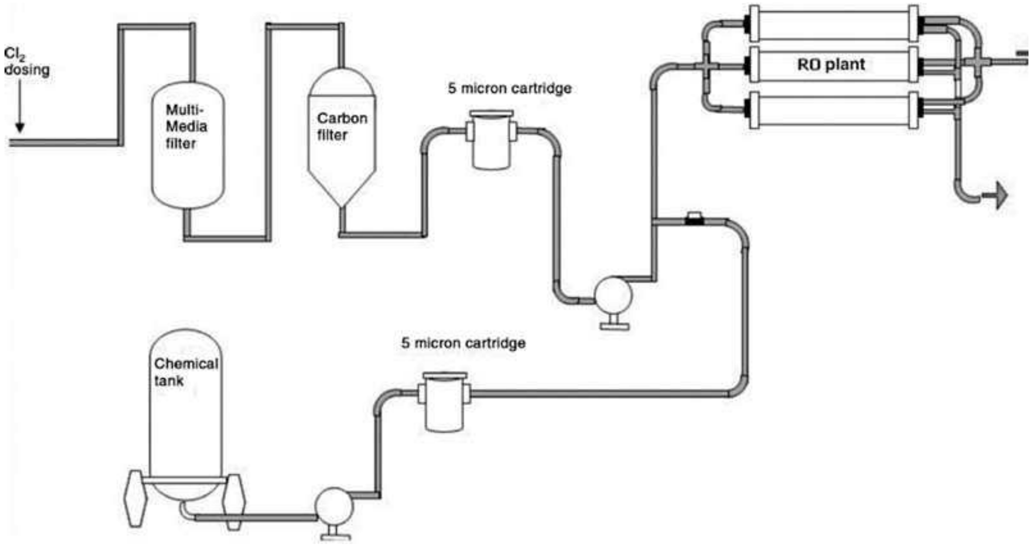


FIGURE 7 Pretreatment for RO system.

Integrated Membrane System

The integrated membrane system (IMS) design approach to water treatment systems has some significant advantages over RO systems designed with conventional pretreatment. The pretreatment of feed water prior to RO is intended to lower the silt density index, remove excessive turbidity or suspended solids, and adjust and control the pH.[25] The choice of IMS system depends on the fouling properties of the feed water, which may necessitate additional (pre)treatment based on the local circumstances. The combination of UF with a precoagulation at low dose helps in controlling the UF membrane fouling and providing filtered water in a steady-state condition.[26] Operation of an RO seawater unit at a higher flux and recovery rate enables optimization of the RO process and reduction of water cost.[27]

Water and Wastewater Purification

Rainwater collected from storm drains can be purified using reverse osmosis. Treated water can be used for landscape irrigation and industrial cooling. It works out to be an acceptable solution to the problem of water shortages. It is also used to purify the effluent and brackish groundwater. The effluent in larger volumes should be treated in an effluent treatment plant first, and then the clear effluent is subjected to reverse osmosis system. The process of reverse osmosis can be used for the production of deionized water. The RO process for water purification does not require thermal energy. Flow through an RO system can be regulated by a high-pressure pump. In 2002, the Singapore government developed NEWater using reverse osmosis to treat domestic wastewater before discharging the NEWater back into the treated water storages.

Eco-Filtration Systems

Zero-electricity systems with comparatively negligible maintenance can be developed using the ecosystem approach. Constructed wetlands, soil scape filters (vertical eco-filtration), green bridge (horizontal eco-filtration), etc., are useful in absorbing the pollutants and rendering them nontoxic through various degradative pathways to such an extent that they become nutrients for various groups of organisms.

Various applications like in situ treatment are evolving to cope with the pollution from nonpoint sources for which there is no economically viable solution in conventional systems.

Constructed Wetlands

Constructed wetlands can either be surface-flow or subsurface-flow and horizontal- or vertical- flow systems.[28] They may include engineered reedbeds and plant beds and belong to the family of phytorestoration and eco-technologies. A high degree of biological improvement takes place due to bioassimilation of pollution. These systems can act as a primary, secondary, and sometimes tertiary treatment. They are known to be highly productive systems as they copy natural wetlands, called the "kidneys of the earth" for their fundamental recycling capacity of the hydrological cycle in the biosphere. Robust and reliable, their treatment capacities improve as time goes by, the opposite of conventional treatment plants whose machinery ages with time. They are being increasingly used, although adequate and experienced designs are more fundamental than for other systems, and space limitation may impede their use (Figure 8).

Soil Biotechnology

Some advancement is being developed by researchers in erstwhile soil-based technologies for odor control in the 20th century. A process called soil biotechnology (SBT), developed in the Indian Institute of Technology in Mumbai, has encouraging results in process efficiency enabling total water reuse, due to extremely low operating power requirements of less than 50 J per kg of treated water.[32] Typically, SBT systems can achieve chemical oxygen demand (COD) levels less than 10 mg/L from sewage input of 400 mg/L COD. Soil biotechnology plants exhibit high reductions in COD values and bacterial counts as a result of the very high microbial densities available in the media. Unlike conventional treatment plants, SBT plants produce insignificant amounts of sludge, precluding the need for sludge disposal areas that are required by other technologies.[29]

Potential and Scope of Eco-Technology

Engineering applications of ecological principles and succession of biological communities are very useful inconsuming organic and inorganic pollutants from the water and bioconverting them into nontoxic form.[30,31] The consortia of organisms at different trophic levels utilize pollutants as nutrients.

FIGURE 8 Flourishing artificial wetland wastewater treatment system (courtesy of SERI, Pune, India).

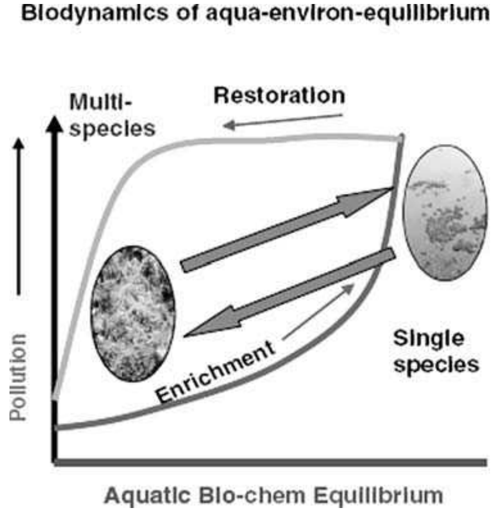


FIGURE 9 Aquatic biochemical equilibrium.

These eco-transformations, eco-conversions, and degradation or bioutilization of pollutants to nutrients are part of ecological and biogeochemical cycles. An attempt has been made to apply natural flora and fauna in a well-designed manner to develop technologies like Green Bridge, Green Lake Eco-Systems, Green Channel, BIOX (Biological Oxidation) and Stream Eco-Systems.

Many growing cities need cost-effective and less energy-intensive treatment methodology to control their pollution emanating from point and nonpoint sources. The eco-technological treatment systems, vertical filtration systems for point sources, and horizontal filtration systems promise results with minimal electricity as compared with conventional aerobic or anaerobic treatment systems.

Eco-filtration (vertical and horizontal) systems are based on the principle of living systems in action for pollution treatment. Efficiency of treatment is dependent on expression of multispecies intelligence—adaptability of organisms to changing environmental conditions while continuing the degradation or conversion of pollutants into utilizable form. When changes occur in the natural systems due to external inputs, biogeochemical cycles and food chains are reorganized and balanced. A new dynamic order emerges in the eco-filtration systems suitable to the environmental changes superimposing on it (Figure 9).

Detritus-feeding organisms in the eco-filtration systems consume the pollutants and wastes (biodegradation products) generated from this process and are useful for green plants. Secondly, the green plants absorb carbon dioxide from the atmosphere and produce oxygen, which is transferred to the eco-reactor in the rootzone area. Thus, the pollutants are transferred to natural cycles, i.e., biogeochemical cycles, and carbon gets stored in vegetation and subsequently in the soil. Plants store carbon in the form of live biomass. Once plants die, the biomass becomes a part of the food chain again and eventually enters the soil as soil carbon. This is a natural process that does not need electricity at all or any chemical for coagulation or activated carbon for adsorption. This makes the eco-filtration system very economical and routine operations very simple.

Soil Scape Filter

This is the simulation of natural filtration of wastewater through the ecofert—mixed cultures and fragmented rock materials. A soil scape filter contains layers of bioactive (i.e., biologically activated) organic material—bacterial mixtures derived from nontoxic and nonhazardous materials. This vertical

FIGURE 10 Sewage treatment using soil scape filter (courtesy of SERI, Pune, India).

eco-filtration process harnesses the ecological principles of interactions and interrelationships of biota with their environment and eco-transformations of substrates into assimilable form by treating, transforming, and detoxifying the pollutants using solar energy[33] (Figure 10).

Green Bridge Technology

Green Bridge technology uses filtration power of biologically originated cellulosic/fibrous material in combination with sand, gravels, root systems of green plants, and microorganisms.[34] It is an innovative approach to minimize the cost of pollution treatment when the cellulosic/ fibrous materials like coconut coir, dried water hyacinth, or aquatic grasses are compacted and woven to form a bridge/porous wall-like structure strengthened by stones and sand (Figure 11). All the floatable and suspended solids are trapped in this biological bridge, and the turbidity of flowing water is reduced substantially. The green plants growing there help in absorption of soluble substances, including heavy metals.

Other Emerging Filtration Techniques

Disc Filters

With disc filters, the divergence in woven filtration media and flow pattern makes each application unique. Both inside-out and outside-in configurations perform to expectations, considering that these devices have been marketed for the relatively clean waters of municipal tertiary applications. Inside-out technique provides better sustainability and life cycle cost. One major drawback of these systems is the biological growth that occurs on the inside face (the filtrate side) of the filter cloth. It is common with unchlorinated filter feed/secondary effluents. Another problem of these systems is the need to shut down the process in order to accomplish the washing of cloth/filtering medium. These negative aspects result in lengthy rinsing schedule, sizeable washwater volumes, and redundancy requirements to ensure continuous wastewater treatment.

FIGURE 11 Green Bridge installation for ecological restoration of Ahar River of Udaipur, India (courtesy of SERI, Pune, India).

Pleated Filtration Panel innovation

Inside-out and outside-in techniques rely on filtration media configured in a flat panel design. The effective filtration area is therefore the available area that is perpendicular to the direction of flow and should not account for any impervious surfaces of supporting structures such as frames, panel boxes, and metal works.

The disc filter with inside-out configuration brings a pleated design to the filtration panels, increasing the effective filtration area by 40% over flat panel designs. This also augments the overall strength of the filtration panel, allowing higher operating headloss. The pleated configuration makes the filtration panel less prone to deformities of the cloth media. Combination of higher filtration surface area and depth of the disc filter leads to high throughput capacity. The internal assembly of the disc/drum has a sliding cover of stainless steel for ease of inspection and maintenance. This is beneficial during installations in any weather conditions.

Conclusion

Technology can be related on the basis of—type of solids—colloidal or dissolved—chemical or biological species—capital and operational costs. Various types of sand filters are useful in treating surface or subsurface waters for drinking purposes. They can be supplemented with GAC or membrane filtrations to yield high-quality water required for industrial processes and production. Trickling filters, biofilters, and eco-technological filters have applications in treating domestic wastewaters before being discharged into the receiving water bodies or for land irrigation. These wastewater filters can be used as polishing systems also to improve the quality of secondarily treated water by reducing the expenditures on tertiary treatment. Eco-technological filters are found to be useful in treating industrial wastewaters by reducing the ecological toxicity of polluted waters. Disc filters or pleated filters are in the development stage. All these filters can be selected to treat waters or wastewaters depending on their source, the desired quality of treated water, land availability, electricity, skilled manpower requirements, etc.

References

1. WHO. Water Sanitation and Health. Available at http://www.who.int/water_sanitation_health/sanitproblems/en/index5.html (accessed on August 16, 2011).
2. IEES. International Ecological Engineering Society. Reclaiming wastewater in Windhoek, Namibia. Available at http://www.iees.ch/cms/index.php?option=com_content&task=view&id=24&Itemid=64 (accessed on August 16, 2011).
3. Kurita Water Industries, Ltd. *Kurita Handbook of Water Treatment*; Kurita Water Industries, Ltd, Tokyo, 1985. pp. 4.1–4.25.
4. Anselme, C.; Jacobs, E.P. Ultrafiltration. In *Water Treatment Membrane Processes*; Mallevialle, J., Odendaal, P.E., Wiesner, Eds.; American Water Works Association Research Foundation. Lyonnaise des Eaux, Water Research Commission of South Africa: McGraw-Hill, New York, 1996.
5. Clever, M.; Jordt, F.; Knauf, R.; Rabiger, N.; Rudebusch, M.; Hilter-Scheibel, R. Process water production from river water by ultrafiltration and reverse osmosis. In Proceedings of the Conference on Membranes in Drinking and Industrial Water Production; Desalination Publications: L'Aquila, Italy, 2000; Vol. 1.
6. Stevenson, D.G. *Water Treatment Unit Processes.* Imperial College Press. London. 1998.
7. Joshi, S. Unpublished study. 2000.
8. James, M. Montgomery Consulting Engineers Inc. *Water Treatment Principles and Design*; John Wiley and Sons, New York, 1985.
9. Hawkes, H.A. *The Ecology of Wastewater Treatment*; MacMillan: New York, 1985.
10. Higgins, I.J.; Burns, R.G. *The Chemistry and Microbiology of Pollution*; Academic: London, 1975.
11. Cheryan, M. *Ultrafiltration Handbook*; Technomic Publishing Company, Inc., Lancaster, Basel, 1986.
12. Mulder, M. *Basic Principles of Membrane Technology.* Kluwer Academic Publishers, Dordrecht, 1996.
13. Aimar, P.; Meireles, M.; Bacchin, P.; Sanchez, V. Fouling and concentration polarization in ultrafiltration and microfiltration. Paper presented at the ASI NATO Meeting, Curia, Portugal, March 1993.
14. Jonsson, G. Boundary layer phenomena during ultrafiltration of dextran and whey protein solutions. Desalination **1984**, *51,* 61–77.
15. Jönsson, A.S.; Trägardh, G. Fundamental principles of ultrafiltration. Chem. Eng. Process **1990**, *27,* 67–81.
16. Aptel, P.; Buckley, C.A. Categories of membrane operations. *Water Treatment Membrane Processes*; Mallevialle, J., Odendaal, P.E., Wiesner, M.R., Eds.; American Water Works Association Research Foundation, Lyonnaise des Eaux, Water Research Commission of South Africa: McGraw-Hill, New York, 1996.
17. Bates, W.T. Capillary UF as RO pretreatment. Hydranau- tics; 1999, available at http://www.membranes.com.
18. Laîné, J.M.; Vial, D.; Moulart, P. Status after 10 years of operation—Overview of UF technology today. In Proceedings of the Conference on Membranes in Drinking and Industrial Water Production; Desalination Publications, 2000; Vol. 1, 17–25.
19. Côté, P.; Cadera, J.; Adams, N.; Best, G. Monitoring and maintaining the integrity of immersed ultrafiltration membranes used for pathogen reduction. Water Supply **2002**, *2* (5–6), 307–311.
20. Campos, C., I. Baudin, J.M. Laîné. 2001. Adsorption performance of powdered activated carbon—Ultrafiltration systems. Abstract. Water Supply **2001**, *1* (5–6), 13–19.

21. Wenten, I.G. 2004. Industrial membrane application in Indonesia. Case study: Clean production in cassava starch industry. In Regional Symposium on Membrane Science and Technology Johor, Malaysia, 2004.
22. Crittenden, J.; Trussell, R.; Hand, D.; Howe, K.; Tchobano- glous, G. *Water Treatment Principles and Design,* 2nd Ed; John Wiley and Sons: New Jersey, 2005.
23. Taylor, J.S.; Jacobs, E.P. Reverse osmosis and nanofiltration. *Water Treatment Membrane Processes;* Mallevialle, J., Odendaal, P.E., Wiesner, M.R., Eds.; American Water Works Association Research Foundation, Lyonnaise des Eaux, Water Research Commission of South Africa: McGraw-Hill, New York, 1996, pp. 9.1–9.70.
24. Fawzi, N.; Al-Enezi, G. Design consideration of RO units: Case studies. Desalination **2002**, *153,* 281–286.
25. Buckley, C.A.; Hurt, Q.E. Membrane applications: a contaminant-based perspective. *Water Treatment Membrane Processes*; Mallevialle, J., Odendaal, P.E., Wiesner, M.R., Eds.; American Water Works Association Research Foundation, Lyonnaise des Eaux, Water Research Commission of South Africa: McGraw-Hill, New York, 1996.
26. Anon. Chennai desal project will open new opportunities: IVRCL in The Hindu. Business Line. Available at http://www.thehindubusinessline.com/todays-paper/tp-corporate/article1000069.ece (accessed on August 2011).
27. Malki, M. Optimizing scale inhibition costs in reverse osmosis desalination plants, Int. Desalin. Water Reuse Q. **2008**, *17* (4), 28–29, available at http://www.membranechemicals.com/english/Optimizing%20Operational%20Costs%20in%20Rerse%20Osmosis%20Desalination%20Plants%20%202008.pdf.
28. Ebrahim, S.; Abdel-Jawad, M.; Bou-Hamad, S.; Safar, M. Fifteen years of R&D program inseawater desalination at KISR. Part I. Pretreatment technologies for RO systems. Desalination **2001**, *135,* 141–153.
29. Brehant, A.; Bonnelye, V.; Perez, M. Comparison of MF/ UF pretreatment with conventional filtration prior to RO membranes for surface seawater desalination. Desalination **2002**, *144,* 353–360.
30. Glueckstern, P.; Priel, M.; Wilf, M. Field evaluation of capillary UF technology as a pretreatment for large seawater RO systems. Desalinatination **2002**, *147*, 55–62.
31. USEPA. *Constructed Wetlands: Treatment of Municipal Wastewater*, EPA/625/R-99/010; Office of Research and Development: Ohio, 2000.
32. Kadam, A.; Ozaa, G.; Nemadea, P.; Duttaa, S.; Shankar, H. Municipal wastewater treatment using novel constructed soil filter system. Chemosphere **2008**, *71* (5), 975–981.
33. Joshi, S. Ecotechnological treatment for industrial wastewater containing heavy metals. J. Indian Assoc. Environ. Manage. **2000**, *27* (2), 98–102.
34. Joshi, S.; Joshi, S. Ecotechnological applications for the control of lake pollution. In Proceedings of TAAL 2007: the 12th World Lake Conference held at Jaipur organized by Ministry of Environment and Forests of India and ILEC, New Delhi. Ed. Sengupta, B. and Dalwani, R.; 2008; pp. 864–867.

58

Wetlands: Constructed Subsurface

Jan Vymazal

Introduction

Constructed wetlands (CWs) with subsurface flow have been used to treat various types of wastewater for more than four decades.[1] In general, there are two types of CWs with subsurface flow: horizontal flow (HF) and vertical flow (VF).[2–4]

In HF CWs, the wastewater is continuously fed in at the inlet and flows slowly through the porous medium under the surface of the porous filtration material planted with emergent vegetation to the outlet where it is collected before leaving via a water level control structure.[4,5] As a consequence of the continuous saturation, the majority of the processes responsible for pollutant removal are either anoxic or anaerobic.[3] VF CWs are fed intermittently on the wetland surface. Wastewater then percolates down through the bed and is collected by a drainage network at the bottom. The bed drains completely and that allows air to refill the bed. Thus, VF CWs provide greater oxygen transfer into the bed, thus allowing for aerobic processes, such as nitrification, to occur.[4]

Horizontal and vertical CWs could be combined in several stages in so-called hybrid systems. These systems use the advantages of both types of systems to enhance removal of pollution, especially nitrogen. Hybrid systems may comprise various types of CWs but the combination of VF and HF CWs is the most common.[6]

During the 1980s and the early 1990s, the CWs with subsurface flow were used primarily in Europe for treatment of municipal or domestic wastewater.[7] During the late 1990s and 2000s, subsurface flow CWs spread throughout the world and were used to treat also various industrial and agricultural effluents, landfill leachate, and various stormwater runoff waters.[2,3]

Technology Background

HF Systems

Horizontal subsurface flow CWs are lined trenches or beds about 0.8 m deep, filled with porous materials such as pea gravel (Figure 1) and planted with wetland emergent macrophytes. The filtration material should be coarse enough to sustain subsurface flow. In the early systems designed by Seidel, coarse sand was proposed as filtration material.[8] The later concept proposed by Kickuth[9] used heavy cohesive soils that provided high level of treatment but failed to sustain subsurface flow due to clogging problems.[10] Based on the experience in the United Kingdom in the late 1980s,[4] the HF CWs have used gravel since then, usually with size fraction between 5 and 20 mm. During passage through the filtration material, the wastewater comes into contact with a network of aerobic, anoxic, and anaerobic zones. Most of the bed is anoxic/anaerobic due to permanent saturation of the beds. The aerobic zones occur around roots and rhizomes that leak oxygen into the substrate.[4,5]

The macrophytes growing in CWs have several properties in relation to the treatment process that make them an essential component of the design.[11] The macrophytes, among other functions, provide oxygen to the rhizosphere through the radial oxygen loss,[12] insulate filtration bed during winter periods,[13] provide substrate for the growth of attached bacteria,[14] can contribute to nutrient removal when harvested,[3,15] and exudate various antimicrobial compounds.[16] Therefore, they should 1) be tolerant of high organic and nutrient loadings; 2) have rich belowground organs (i.e., roots and rhizomes) in order to provide substrate for attached bacteria and oxygenation (even very limited) of areas adjacent

FIGURE 1 Constructed wetlands with horizontal subsurface flow. Left: filtration bed sealed with plastic liner, right: filtration bed filled with crushed rock with inflow and outflow zones filled with stones.
Source: Photos courtesy of Jan Vymazal.

to roots and rhizomes; and 3) have high aboveground biomass for winter insulation in cold and temperate regions and nutrient removal via harvesting.[17,18] The most frequently used plant for HF CWs is Common reed (***Phragmites australis***) (Figure 2).[3]

A large number of wetland systems have shown an exponential decrease of pollutant concentration level with the distance through the wetland from inlet to outlet. The observation is consistent with a first-order removal model, with the removal rate being proportional to the pollutant concentration. The removal can thus empirically be described with first-order plug-flow kinetics:[2,4,19]

$$\left(\ln C_{\mathrm{i}} - \ln C_{\mathrm{o}}\right) = A\ k_A / Q \tag{1}$$

$$C_{\mathrm{o}} = C_{\mathrm{i}}\ \exp^{(-A\ k\mathrm{A}/Q)} \tag{2}$$

or for the *k*-*C** model:[20]

$$\ln\left[\left(C_{\mathrm{i}} - C^{*}\right) / C_{\mathrm{o}} - C^{*}\right)\right] = A\ k_{\mathrm{A}} / Q \tag{3}$$

where

A = area of the bed (m^2)
Q = average flow rate (m^3 day^{-1})
C_{i} = inflow concentration of the pollutant (mg L^{-1})
C_{o} = outflow concentration of the pollutant (mg L^{-1})
k_{A} = first-order areal rate constant (m day^{-1})
C* = background concentration of a pollutant (mg L–1),

which is a concentration caused by an internal release of particulate and dissolved biomass to the water.[20]

As BOD (biochemical oxygen demand) was the primary target, the value of first-order areal rate constant k_{A} used for the design was k_{BOD}. The formerly proposed value of 0.19 m day^{-1} by Kickuth[9,19] resulted in too small area of the bed and consequently lower treatment effect. The extensive field measurements in Denmark in the mid-1980s found the average $\mathrm{k}_{\mathrm{BOD}}$ value of 0.083 ± 0.017 m day^{-1}.[21] Cooper et al.[4] reported values between 0.067 and 0.1 m day^{-1} in the United Kingdom and Vymazal and Kröpfelová[3] reported an average k_{BOD} value of 0.087 m day^{-1} for 31 systems in the Czech Republic. This generally

FIGURE 2 Common reed (*Phragmites australis*) growing in a HF constructed wetland.
Source: Photo courtesy of Jan Vymazal.

TABLE 1 First-Order Areal Rate Constant kA (m day–1) for Total Suspended Solids (TSS), Total Phosphorus (TP), Total Nitrogen (TN) and Ammonia-Nitrogen (NH4-N) Reported by Various Authors in the Literature

Reference	TSS	TP	TN	NH4-N
[3]	0.085	0.026	0.025	0.024
[10]		0.025	0.033	
[22]	2.74	0.033	0.074	0.093

means that the value of A is about 5 m^2 per one population equivalent (PE) for mechanically pretreated municipal or domestic wastewater. However, HF CWs may be designed with other target parameters than BOD$_5$ and therefore appropriate constants were developed for other pollutants as well (Table 1). When a CW is designed with more target parameters at the same time, e.g., BOD5, TSS, nitrogen, and phosphorus, then it is necessary to select the largest area according to all four calculations.

At present, the state-of-the-art *k-C** model seems to be best available design tool if the designer makes sure that all the assumptions are fulfilled and if he or she is aware of many pitfalls in the model.[22] One of the most important uncertainties comes from the fact that background concentrations inevitably change over the operation period as the system matures. In addition, Stein et al.[23] suggested that C^* is temperature dependent. Kadlec and Knight[20] suggested C^* values (in mg L^{-1}) for BOD$_5$ ($3.5+0.053C_i$), TSS ($7.8+0.063C_i$), TP (0.02), and TN (1.5).

Recently, the so-called *P-k-C** model has been developed:[2]

$$\left(C_o - C^*\right)/\left(C_i - C^*\right) = 1/\left(1 + k_A/Pq\right)^P \tag{4}$$

where

k_A = modified first-order areal constant (m day^{-1})
P = apparent number of TIS (tanks in series)
q = hydraulic loading rate (HLR) (cm day–1)

In HF CWS, pollution is removed through a complex of aerobic, anoxic, and anaerobic processes that occur simultaneously in the filtration bed with aerobic processes being restricted only to the thin area adjacent to macrophyte roots.[3,24,25] The treatment efficiency of HF CWs has recently been thoroughly evaluated by Vymazal and Kröpfelová.[3] Organic compounds are removed primarily by microbial processes; particulate and colloid organics may be filtered out through the wastewater passage through the bed. Suspended solids are removed in HF CWs via filtration and sedimentation. Systems that have been designed with a specific area of about 5 m^2 PE^{-1} have usually very high treatment efficiency for organics and suspended solids. The removal efficiency in most cases equals the efficiency of conventional treatment systems. However, contrary to conventional treatment systems, e.g., activated sludge systems, HF CWs can handle wastewaters with very low concentrations of organics such as wastewaters from combined sewer systems.[26,27]

Removal of nitrogen is usually low due to lack of oxygen in filtration bed, which limits nitrification of ammonia-nitrogen.[28] On the other hand, HF CWs provide suitable conditions for denitrification due to anoxic/anaerobic conditions in the filtration beds. Plant uptake is responsible for removal of only small part of the inflow nitrogen loading and usually does not exceed 10% of the inflow.[29,30] For municipal sewage, the removal of ammonia-N usually varies between 20% and 40%.[3]

HF CWs are seldom built with phosphorus being the primary target of the treatment, and therefore, materials with relatively low sorption capacity but high hydraulic conductivity such as river gravel or crushed rock are commonly used.[3] In this case, removal efficiency for municipal sewage is usually between 20% and 50%. In order to achieve high phosphorus removal, it is necessary to select materials with high P adsorption capacity. Such materials may include minerals with reactive Fe or Al hydroxide or oxide groups on their surfaces, or calcareous materials that can promote precipitation of Ca-phosphate,[31–33] lightweight aggregates,[34] industrial by-products and waste products such as electric arc furnace steel slags[35,36] or fly ash.[31] By using materials with high sorption capacity, the phosphorus removal efficiency may exceed 90%, but it is necessary to keep in mind that the sorption capacity is always limited and the material will have to be replaced in the future in order to keep a high rate of P removal.

Microbial pollution is removed through a complex of physical, chemical, and biological factors, which all participate in the reduction of the number of bacteria of anthropogenic origin. Physical factors include mechanical filtration, straining, adsorption, and sedimentation. Chemical factors include oxidation, exposure to biocides excreted by some plants, and adsorption to organic matter; biological removal factors include antibiosis, predation by nematodes, protozoa and zooplankton, attack by lytic bacteria and viruses, and natural die-off.[3,37–40]

One of the major threats for the long-run treatment performance of HF CWs is clogging and subsequent surface flow[41,42] and therefore efficient mechanical pretreatment should always be installed before the wastewater enters the filtration beds. For small systems, septic tank is usually used; for larger systems, Imhoff tank is suitable. In case of combined sewer systems, i.e., systems where wastewater is mixed with stormwater runoff, grit chamber should be included (Figure 3).

VF Systems

The earliest form of VF CW is that of Seidel in Germany in the 1960s where vertical filters were used to oxygenate anaerobic septic tank effluents.[43] Interest in the particular process seemed to wane, but it has been revived in the 1990s because of the need to produce beds that nitrify.[4]

FIGURE 3 Mechanical pretreatment: screens and horizontal grit chamber. (**Source:** Photo courtesy of Jan Vymazal).

VF CWs comprise a flat bed of sand or graded gravel planted with macrophytes (Figure 4). Contrary to HF CWs, vertical systems are fed intermittently with large batches on the bed surface. The distribution system should allow for even wastewater distribution across the bed (Figure 5). The intermittent feeding creates conditions where the filter completely drains, thus enabling oxygen to diffuse into the bed. On the other hand, VF CWs do not provide suitable conditions for denitrification to complete conversion to gaseous nitrogen forms that escape to the atmosphere. The major purpose of macrophyte presence in VF CWs is to help maintain the hydraulic conductivity of the bed.[4]

The early VF systems have usually been composed of several stages with two to four beds in the first stage that were fed with wastewater in rotation. Such VF systems are now called first-generation VF systems.[44] The early VF systems were frequently the first stage of the hybrid systems.[45,46] Recently, VF systems with only one bed have been used. These systems are called second-generation VF CWs or compact VF beds.[47,48]

FIGURE 4 Distribution of wastewater on the surface of a vertical flow constructed wetlands. Experimental station, Carrión de los Céspedes, Spain. (**Source:** Photo courtesy of Jan Vymazal).

FIGURE 5 Vertical flow CWs at Saint Thomé, France. (**Source:** Photo courtesy of Jan Vymazal).

VF CWs require pretreatment in order to prevent clogging, but in France, hundreds of systems are fed directly with raw sewage onto the first stage, allowing for easier sludge management in comparison to dealing with primary sludge from an Imhoff tank.[49,50] While compact VF systems are usually small, VF systems treating raw sewage in France are often much larger. The calculation of VF bed area has evolved in years, but at present, the area varies between 2 m^2 PE^{-1} (split into 1.2 m^2 and 0.8 m^2 for the first and second stage, respectively) in France[49] and 4 m^2 PE^{-1} in Austria and Germany.[51–53] Danish design recommendations are set for 4 m^2 PE^{-1}.[54] One of the major threats of good performance of VF systems is clogging of the filtration substrate,[55,56] and therefore, it is important to properly select the filtration material and the HLR and distribute the water evenly across the bed surface. HLRs reported in the literature vary greatly between about 3 cm day^{-1} up to 150 cm day^{-1},[47,57] but to avoid flooding due to clogging, HLRs should probably be kept below 80 cm day^{-1}.[58]

VF CWs provide high removal of organics and suspended solids and are also effective in removing ammonia- N. However, they have very limited capacity to denitrify; therefore, the removal of total nitrogen is about the same as for HF CWs. Removal of phosphorus is also within the same range as for HF CWs unless special media with high sorption capacity are used.[3]

Hybrid Systems

Various types of CWs may be combined in order to achieve higher removal efficiency, especially for nitrogen. Many of these systems are derived from original hybrid systems developed by Seidel at the Max Planck Institute in Krefeld, Germany.[43] The design consists of two stages, several parallel VF beds followed by two or three HF beds in series. The VF wetland is intended to remove organics and suspended solids and to provide nitrification, while in the HF wetland, denitrification and further removal of organics and suspended solids occur.

In the early 1980s, several hybrid systems of Seidel's type were built in France with the systems at Saint Bohaire and Frolois being the best described.[59] A similar system was built in 1987 in the United Kingdom at Oaklands Park,[60] and in the 1990s and the early 2000s, VF–HF systems were built in many countries such as Norway,[61] Ireland,[62] Estonia,[63] or Thailand.[64]

In the mid-1990s, Brix and Johansen[65] introduced a HF–VF hybrid system with a large HF bed placed first and a small VF bed as the second stage. In the first bed, removal of organics and suspended solids takes place while nitrification takes place in the second VF wetland. In order to remove nitrate produced during nitrification, the water have to be recirculated to the front end of the system where denitrification can take place in the less aerobic HF bed using the raw feed as a source of carbon needed for denitrification.[65] The hybrid systems of this type have been reported from other European countries,[66,67] North America,[68] or Asia.[69]

Use for Various Types of Wastewater

Subsurface flow CWs have predominantly been used to treat municipal (secondary as well as tertiary treatment) or domestic wastewater.[1–4] However, recently, they have been used to treat a variety of wastewaters including industrial and agricultural effluents, landfill leachate, or stormwater runoff (Table 2).

When organics and suspended solids are the primary target of the treatment, the HF CWs are usually used. These systems perform effectively and the operation and maintenance costs are very low as compared to conventional treatment systems. When removal of ammonia is required, VF CWs are more effective. These systems usually require electricity, but despite that, the operation and maintenance costs are still relatively low. However, to remove both ammonia- and nitrateN, hybrid CWs provide a suitable and effective option. Hybrid CWs are also often used to treat more complex wastewaters (Table 2).

TABLE 2 Examples of the Use of Subsurface CWs for Various Types of Wastewater

Type of Wastewater	Location	Type of CWs	Reference
Petrochemical	United States, China	HF	[70,71]
	Pakistan	VF	[72]
Chemical industry	United Kingdom	HF	[73]
	Portugal, United States, Germany	VF	[74–76]
Paper and pulp wastewaters	United States	HF	[77]
Abattoir	Mexico, Ecuador	HF	[78,79]
	Poland	VF–HF	[80]
Textile industry	Australia	HF	[81]
Tannery industry	Portugal	HF	[82]
Mixed industrial	Slovenia	HF–VF–HF	[83]
Food industry	Slovenia, Italy	HF	[84,85]
Distillery and winery	India, Italy	HF	[86,87]
Pig farm	Australia, Lithuania	HF	[88,89]
	Thailand	VF–HF	[90]
Fish farm	Canada, Germany	HF	[91,92]
Dairy	United States, Germany	HF	[93,94]
	The Netherlands	VF	[95]
	France, Japan	VF–HF	[96,97]
Highway runoff	United Kingdom	HF	[98]
Airport runoff	United States	HF	[99]
	Canada	VF	[100]
Nursery runoff	Australia	HF	[101]
Greenhouse	Korea	HF–VF–HF	[102]
Landfill leachate	Poland	HF	[103]
	Australia	VF	[104]
	Slovenia	VF–HF	[105]
Compost leachate	France	VF–HF	[94]

Conclusions

Constructed treatment wetlands have evolved during the last five decades into a reliable treatment technology applicable to all types of wastewater including sewage, industrial and agricultural wastewaters, landfill leachate, and stormwater runoff. The subsurface CW technology started in Germany but spread quickly to all continents. At present, subsurface constructed treatment wetlands provide a viable alternative to conventional treatment systems.

The subsurface technology has substantially improved over the years, but there are still gaps that need to be solved. Among others, the low-cost filtration medium that provides high removal of phosphorus is still to be determined, and the problem of system longevity and filtration material replacement or cleaning are under investigation. Recently, also more attention has been paid to dual-purpose or multipurpose services of constructed treatment wetlands such as flood control, carbon sequestration, or wildlife habitat. Another gap where a lot of work has been done recently is the modeling of HF CWs.[106]

Acknowledgments

The study was supported by grant No. QH81170 "Complex Evaluation of the Effect of Area Protection on Important Water Management Localities" from the Ministry of Agriculture of the Czech Republic.

References

1. Vymazal, J. Constructed wetlands for wastewater treatment: Five decades of experience. Environ. Sci. Technol. **2011,** *45* (1), 61–69.
2. Kadlec, R.H.; Wallace, S.D. *Treatment Wetlands,* 2nd Ed.; CRC Press: Boca Raton, Florida, 2008.
3. Vymazal, J.; Kröpfelová, L. *Wastewater Treatment in Constructed Wetlands with Horizontal Sub-Surface Flow*; Springer: Dordrecht, the Netherlands, 2008.
4. Cooper, P.F.; Job, G.D.; Green, M.B.; Shutes, R.B.E. *Reed Beds and Constructed Wetlands for Wastewater Treatment*; WRc Publications: Medmenham, Marlow, U.K., 1996.
5. Brix, H. Treatment of wastewater in the rhizosphere of wetland plants—The root zone method. Water Sci. Technol. **1987,** *19* (10), 107–118.
6. Vymazal, J. Horizontal sub-surface flow and hybrid constructed wetlands systems for wastewater treatment. Ecol. Eng. **2005,** *25*, 478–490.
7. Vymazal, J.; Brix, H.; Cooper, P.F.; Green, M.B.; Haberl, R., Eds. *Constructed Wetlands for Wastewater Treatment in Europe*; Backhuys Publishers: Leiden, the Netherlands, 1998.
8. Seidel, K. Phenol-Abbau in Wasser durch *Scirpus lacustris* L. wehrend einer versuchsdauer von 31 Monaten. Naturwissenschaften **1965,** *52* (13), 398–406.
9. Kickuth, K. Abwasserreinigung im Wurzelraumverfahren. Wasser, Luft Betr. **1980,** *11*, 21–24.
10. Brix, H. Denmark. In *Constructed Wetlands for Wastewater Treatment*; Vymazal, J., Brix, H., Cooper, P.F., Green, M. B., Haberl, R., Eds.; Backhuys Publishers: Leiden, the Netherlands, 1998; 123–152.
11. Brix, H. Do macrophytes play a role in constructed treatment wetlands? Water Sci. Technol. **1997,** *35* (5), 11–17.
12. Brix, H. Gas exchange through the soil-atmosphere interface and through dead culms of *Phragmites australis* in a constructed wetland receiving domestic sewage. Water Res. **1990,** *24*, 377–389.
13. Mander, Ü.; Jenssen, P., Eds. *Constructed Wetlands for Wastewater Treatment in Cold Climates*; WIT Press: Southampton, U.K., 2003.
14. Vymazal, J.; Ottová, V.; Balcarová, J.; Doušová, H. Seasonal variation in fecal indicators removal in constructed wetlands with horizontal subsurface flow. In *Constructed Wetlands for Wastewater Treatment in Cold Climates*; Mander, Ü., Jenssen, P., Eds.; WIT Press: Southampton, U.K., 2003; 239–258.
15. Adcock, P.; Ganf, G.G. Growth characteristics of three macrophyte species growing in natural and constructed wetland system. Water Sci. Technol. **1994,** 29, 95–102.
16. Vincent, G.; Dallaire, S.; Lauzer, D. Antimicrobial properties of roots exudate of three macrophytes: *Mentha aquatica* L., *Phragmites australis* (Cav.) Trin. and *Scirpus lacustris* L. In Proceedings of the 4th International Conference Wetland Systems for Water Pollution Control, Guangzhou, China, November 6–10, 1994; ICWS '94 Secretariat, 1994; 290–296.
17. Čížková-Končalová, H.; Květ, J.; Lukavská, J. Response of *Phragmites australis*, *Glyceria maxima* and *Typha latifo-lia* to addition of piggery sewage in a flooded sand culture. Wetlands Ecol. Manage. **1996,** *4*, 43–50.
18. Květ, J.; Dušek, J.; Husák, Š. Vascular plants suitable for wastewater treatment in temperate zones. In *Nutrient Cycling and Retention in Natural and Constructed Wetlands*; Vymazal, J., Ed.; Backhuys Publishers: Leiden, the Netherlands, 1999; 101–110.
19. Kickuth, R. 1981. Abwasserreinigung in Mosaikmatritzen aus aeroben und anaeroben Teilbezirken. In *Grundlagen der Abwasserreinigung;* Moser, F., Ed.; Verlag Oldenburg: Munchen, Wien, 1981; 639–665.
20. Kadlec, R.H.; Knight, R.L. *Treatment Wetlands;* CRC Press: Boca Raton, Florida, 1996.
21. Brix, H. Constructed wetlands for municipal wastewater treatment. In *Global Wetlands: Old World and New*; Mitsch, W.J., Ed.; Elsevier: Amsterdam, 1994; 325–333.

22. Rousseau, D.P.L.; Vanrolleghem, P.A.; De Pauw, N. Model-based design of horizontal subsurface flow constructed wetlands: a review. Water Res. **2004,** *38,* 1484–1493.
23. Stein, O.R.; Biederman, J.A.; Hook, P.B.; Allen, W.C. Plant species and temperature effects on the *k-C** firstorder model for COD removal in batch-loaded SSF wetlands. Ecol. Eng. **2006,** *26*, 100–112.
24. Imfeld, G.; Braeckevelt, M.; Kuschk, P.; Richnow, H.H. Monitoring and assessing processes of organic chemicals removal in constructed wetlands. Chemosphere **2009,** *74,* 349–362.
25. Garcia, J.; Rousseau, D.; Morato, J.; Lesage, E.; Matamoros, V.; Bayona, J.M. Contaminant removal processes in subsurface-flow constructed wetlands: A review. Crit. Rev. Environ. Sci. Technol. **2010,** *40,* 561–661.
26. Vymazal, J. Horizontal sub-surface flow constructed wetlands Ondřejov and Spálené Poříčí in the Czech Republic—15 years of operation. Desalination **2009,** *246,* 226–237.
27. Green, M.B.; Martin, J.R. Constructed reed beds clean up stormwater overflows on small wastewater treatment works. Water Environ. Res. **1996,** *68,* 1054–1060.
28. Vymazal, J. Nitrogen removal in constructed wetlands with horizontal sub-surface flow-can we determine the key process? In *Nutrient Cycling and Retention in Natural and Constructed Wetlands*; Vymazal, J., Ed.; Backhuys Publishers: Leiden, the Netherlands, 1999; 1–17.
29. Obarska-Pempkowiak, H.; Gajewska, M. The dynamics of processes responsible for transformation of nitrogen compounds in hybrid wetlands systems in a temperate climate. In *Wetlands—Nutrients, Metals and Mass Cycling*; Vymazal, J., Ed.; Backhuys Publishers: Leiden, the Netherlands, 2003; 129–142.
30. Browning, K.; Greenway, M. Nutrient removal and plant biomass in a subsurface flow constructed wetland in Brisbane, Australia. Water Sci. Technol. **2003,** *48* (5), 183–189.
31. Drizo, A.; Frost, C.A.; Grace, J.; Smith, K.A. Physico-chemical screening of phosphate removing substrates for use in constructed wetlands. Water Res. **1999,** *33*, 3595–3602.
32. Molle, P.; Liénard, A.; Grasmick, A.; Iwema, A. Phosphorus retention in subsurface constructed wetlands: Investigations focused on calcareous materials and their chemical reactions. Water Sci. Technol. **2003,** *48* (5), 75–83.
33. Ádám, K.; Krogstad, T.; Vråle, L.; Søvik, A.K.; Jenssen, P. Phosphorus retention in the filter materials shellsand and Filtralite®—Batch and column experiments with synthetic P solution and secondary wastewater. Ecol. Eng. **2007,** *29,* 200–208.
34. Jenssen, P.D.; Krogstad, T. Design of constructed wetlands using phosphorus sorbing lightweight aggregate (LWA). In *Constructed Wetlands for Wastewater Treatment in Cold Climates*; Mander, Ü., Jenssen, P., Eds.; WIT Press: Southampton, U.K., 2003; 260–271.
35. Mann, R.A.; Bavor, H.J. Phosphorus removal in constructed wetlands using gravel and industrial waste substrate. Water Sci. Technol. **1993,** *27,* 107–113.
36. Xu, D.; Xu, J.; Wu, J.; Muhammad, A. Studies on the phosphorus sorption capacity of substrates used in constructed wetland systems. Chemosphere **2006,** *63,* 344–352.
37. Decamp, O.; Warren, A.; Sanchez, R. The role of ciliated protozoa in subsurface flow wetlands and their potential as bioindicators. Water Sci. Technol. **1999,** *40* (3), 91–98.
38. Neori, A.; Reddy, K.R.; Čížková-Končalová, H.; Agami, M. Bioactive chemicals and biological–biochemical activities and their functions in rhizospheres of wetland plants. Bot. Rev. **2000,** *66*, 350–378.
39. Stevik, T.K.; Geir Ausland, K.A.; Hanssen, J.F. Retention and removal of pathogenic bacteria in wastewater percolating through porous media: A review. Water Res. **2004,** *38*, 1355–1367.
40. Seidel, K. Macrophytes and water purification. In *Biological Control of Water Pollution*; Tourbier, J., Pierson, R.W., Eds.; Pennsylvania University Press: Philadelphia, 1976; 109–122.
41. Knowles, P.; Dotro, G.; Nivala, J.; Garcia, J. Clogging in sub-surface flow treatment wetlands: Occurrence and contributing factors. Ecol. Eng. **2011,** *37,* 99–112.

42. Cooper, D.; Griffin, P.; Cooper, P. Factors affecting the longevity of sub-surface horizontal flow systems operating as tertiary treatment for sewage effluent. Water Sci. Technol. **2005,** *51* (9), 127–135.
43. Seidel, K. Neue Wege zur Grundwasseranreicherung in Krefeld.Vol. II. Hydrobotanische Reinigungsmethode. GWF Wasser/Abwasser **1965,** *30,* 831–833.
44. Cooper, P.F. The performance of vertical flow constructed wetland systems with special reference to the significance of oxygen transfer and hydraulic loading rates. Water Sci. Technol. **2005,** *51* (9), 81–90.
45. Burka, U.; Lawrence, P. A new community approach to wastewater treatment with higher water plants. In *Constructed Wetlands in Water Pollution Control*; Cooper, P.F., Findlater, B.C., Eds.; Pergamon Press: Oxford, U.K., 1990; 359–371.
46. Lienard, A.; Boutin, C.; Esser, D. Domestic wastewater treatment with emergent helophyte beds in France. In *Constructed Wetlands in Water Pollution*; Cooper, P.F., Findlater, B.C., Eds.; Pergamon Press: Oxford, U.K., 1990; 183–192.
47. Weedon, C.N. Compact vertical flow reed bed system—First two years performance. Water Sci. Technol. **2003,** *48* (5), 15–23.
48. Arias, C.A.; Brix, H. Initial experience from a compact vertical flow constructed wetland treating single household wastewater. In *Natural and Constructed Wetlands: Nutrients, Metals and Management*; Vymazal, J., Ed.; Backhuys Publishers: Leiden, the Netherlands, 2005; 52–64.
49. Molle, P.; Liénard, A.; Boutin, C.; Merlin, G.; Iwema, A. How to treat raw with constructed wetlands: An overview of French systems. Water Sci. Technol. **2005,** *51* (9), 11–21.
50. Paing, J.; Voisin, J. Vertical flow constructed wetlands for municipal septage treatment in French rural areas. Water Sci. Technol. **2005,** *51* (9), 145–155.
51. Langergraber, G.; Prandstetten, C.; Pressl, A.; Rohrhofer, R.; Haberl, R. Removal efficiency of subsurface vertical flow constructed wetlands for different organic loads. In Proceedings of the 10th International Conference Wetland Systems for Water Pollution Control; MAOTDR: Lisbon, Portugal, 2006; 587–597.
52. ÖNORM B 2505. *Bepflanzte Bodenfilter* (*Pflanzenkläran-lagen*)*—Anwendung, Bemessung, Bau und Betrieb* (Subsurface-flow constructed wetlands—Application, dimensioning, installation and operation); Österreichisches Normungsinstitut: Vienna, Austria, 2005 (in German).
53. DWA. *Grundsätze für Bemessung, Bau und Betrieb von Pflanzenkläranlagen mit bepflanzten Bodenfiltern zur biologischen Reinigung kommunalen Abwassers*; Arbeitsblatt DWA-A 262; DWA—Deutsche Vereinigung für Wasserwirtschaft, Abwasser und Abfall e.V., Hennef, Germany, 2006 (in German).
54. Brix, H.; Johansen, N.H. *Retningslinier for etablering af beplantede filteranlæg op til 30 PE* (Guidelines for vertical flow constructed wetland systems up to 30 PE); Økologisk Byfornyelse og Spildevandsrensning N. 52, Miljøstyrelsen, Miljøministeriet (in Danish), 2004.
55. Chazarenc, F.; Merlin, G. Influence of surface layer on hydrology and biology of gravel bed vertical flow constructed wetlands. Water Sci. Technol. **2005,** *51* (9), 91–97.
56. Platzer, C.; Mauch, K. Soil clogging in vertical-flow reed beds—Mechanisms, parameters, consequences, and solutions? Water Sci. Technol. **1997,** *35* (5), 175–181.
57. Cooper, P.F.; Smith, M.; Maynard, H. The design and performance of a nitrifying vertical-flow reed bed system. Water Sci. Technol. **1997,** *35* (5), 215–222.
58. Johansen, N.H.; Brix, H.; Arias, C.A. Design and characterization of a compact constructed wetland system removing BOD, nitrogen and phosphorus from single household sewage. In Proceedings of the 8th International Conference Wetland Systems for Water Pollution Control; University of Dar es Salaam: Tanzania, 2002; 47–61.
59. Boutin, C. Domestic wastewater treatment in tanks planted with rooted macrophytes: Case study, description of the system, design criteria, and efficiency. Water Sci. Technol. **1987,** *19* (10), 29–40.

60. Burka, U.; Lawrence, P. A new community approach to wastewater treatment with higher water plants. In *Constructed Wetlands in Water Pollution*; Cooper, P.F., Findlater, B.C., Eds.; Pergamon Press: Oxford, U.K., 1990, 359–371.
61. Mæhlum, T.; Stålnacke, P. Removal efficiency of three cold-climate constructed wetlands treating domestic wastewater: Effects of temperature, seasons, loading rates and input concentrations. Water Sci. Technol. **1999,** *40* (3), 273–281.
62. O'Hogain, S. The design, operation and performance of a municipal hybrid reed bed treatment system. Water Sci. Technol. **2003,** *48* (5), 119–126.
63. Öövel, M.; Tooming, A.; Mauring, T.; Mander, Ü. School-house wastewater purification in a LWA-filled hybrid constructed wetland in Estonia. Ecol. Eng. **2007,** *29,* 17–26.
64. Kantawanichkul, S.; Neamkam, P. Optimum recirculation ratio for nitrogen removal in a combined system: Vertical flow vegetated bed over horizontal flow sand bed. In *Wetlands: Nutrients, Metals and Mass Cycling*; Vymazal, J., Ed.; Backhuys Publishers: Leiden, the Netherlands, 2003; 75–86.
65. Brix, H.; Johansen, N.H. Treatment of domestic sewage in a two-stage constructed wetland—Design principles. In *Nutrient Cycling and Retention in Natural and Constructed Wetlands*; Vymazal, J., Ed.; Backhuys Publishers: Leiden, the Netherlands, 1999; 155–163.
66. Masi, F.; Martinuzzi, N. (2009). Constructed wetlands for the Mediterranean countries: Hybrid systems for water reuse and sustainable sanitation. Desalination **2009,** *215,* 44–55.
67. Obarska-Pempkowiak, H.; Gajewska, H.; Wojciechowska, E. Application, design and operation of constructed wetland systems. In Proc. Workshop Wastewater Treatment in Wetlands. Theoretical and Practical Aspects; I. Toczylowska, I., Guzowska, G., Eds.; Gdańsk University of Technology Printing Office: Gdansk, Poland, 2005; 119–125.
68. Belmont, M.A.; Cantellano, E.; Thompson, S.; Williamson, M.; Sánchez, A.; Metcalfe, C.D. Treatment of domestic wastewater in a pilot-scale natural treatment system in central Mexico. Ecol. Eng. **2004,** *23,* 299–311.
69. Singh, S.; Haberl, R.; Moog, O.; Shrestha, R.R.; Shrestha, P.; Shrestha, R. Performance of an anaerobic baffled reactor and hybrid constructed wetland treating high-strength wastewater in Nepal—A model for DEWATS. Ecol. Eng. **2009,** *35*, 654–660.
70. Wallace, S.D. On-site remediation of petroleum contact wastes using subsurface-flow wetlands. In *Wetlands and Remediation II*; Nehring, K.W., Brauning, S.E., Eds.; Battelle Press: Columbus, OH, 2002; 125–132.
71. Ji, G.; Sun, T.; Zhou, Q.; Sui, X.; Chang, S.; Li, P. Constructed subsurface slow wetland for treating heavy oil-produced water of the Liaohe Oilfield in China. Ecol. Eng. **2002,** *18*, 459–465.
72. Aslam, M.M.; Malik, M.; Baig, M.A.; Qazi, I.A.; Iqbal, J. Treatment performance of compost-based and gravel-based vertical flow wetlands operated identically for refinery wastewater treatment in Pakistan. Ecol. Eng. **2007,** *30*, 34–42.
73. Sands, Z.; Gill, L.S.; Rust, R. Effluent treatment reed beds: Results after ten years of operation. In *Wetlands and Remediation*; Means, J.F., Hinchee, R.E., Eds.; Battelle Press: Columbus, OH, 2000; 273–279.
74. Machate, T.; Noll, H.; Behrens, H.; Kettrup, A. Degradation of phenanthrene and hydraulic characteristics in a constructed wetland. Water Res. **1997,** *31,* 554–560.
75. Tan, K.; Jackson, W.A.; Anderson, T.A.; Pardue, J.H. Fate of perchlorate-contaminated water in upflow wetlands. Water Res. **2004,** *38,* 4173–4185.
76. Novais, J.M.; Martins-Dias, S. Constructed wetlands for industrial wastewater treatment contaminated with nitro-aromatic organic compounds and nitrate at very high concentrations. In Proc. Conf. The Use of Aquatic Macrophytes for Wastewater Treatment in Constructed Wetlands; Dias, V., Vymazal, L., Eds.; ICN and INAG: Lisbon, Portugal, 2003; 277–288.

77. Thut, R.N. Feasibility of treating pulp mill effluent with a constructed wetland. In *Constructed Wetlands for Water Quality Improvement*; Moshiri, G.A., Ed.; Lewis Publishers: Boca Raton, FL, 1993; 441–447.
78. Poggi-Varaldo, H.M.; Gutiérez-Saravia, A.; Fernández-Villagómez, G.; Martínez-Pereda, P.; Rinderknecht-Seijas, N. A full-scale system with wetlands for slaughterhouse wastewater treatment. In *Wetlands and Remediation II*; Nehring, K.W., Brauning, S.E., Eds.; Battelle Press: Columbus, OH, 2002; 213–223.
79. Lavigne, R.L.; Jankiewicz, J. Artificial wetland treatment technology and its use in the Amazon River forests of Ecuador. In Proc. 7th Internat. Conf. Wetland Systems for Water Pollution Control; University of Florida: Gainesville, 2000, 813–820.
80. Soroko, M. (2005). Treatment of wastewater from small slaughterhouse in hybrid constructed wetlands system. In Proceedings of the Workshop Wastewater Treatment in Wetlands. Theoretical and Practical Aspects; Toczylowska, I., Guzowska, G., Eds.; Gdańsk University of Technology Printing Office: Gdansk, Poland, 2005; 171–176.
81. Davies, T.H.; Cottingham, P.D. The use of constructed wetlands for treating industrial effluent. In Proceedings of the 3rd International Conference Wetland Systems in Water Pollution Control; IAWQ and Australian Water and Wastewater Association: Sydney, NSW, Australia, 1992; 53.1–53.5.
82. Calheiros, C.S.C.; Rangel, A.O.S.S.; Castro, P.K.L. Constructed wetland systems vegetated with different plants applied to the treatment of tannery wastewater. Water Res. **2007,** *41,* 1790–1798.
83. Zupančič Justin, M.; Vrhovšek, D.; Stuhlbacher, A.; Griessler Bulc, T. Treatment of wastewater in hybrid constructed wetland from the production of vinegar and packaging of detergents. Desalination **2009,** *246,* 100–109.
84. Vrhovšek, D.; Kukanja, V.; Bulc, T. Constructed wetland (CW) for industrial waste water treatment. Water Res. **1996,** *30,* 2287–2292.
85. Mantovi, P.; Marmiroli, M.; Maestri, E.; Tagliavini, S.; Piccinini, S.; Marmiroli, N. Application of a horizontal subsurface flow constructed wetland on treatment of dairy parlor wastewater. Bioresour. Technol. **2003,** *88*, 85–94.
86. Billore, S.K.; Singh, N.; Ram, H.K.; Sharma, J.K.; Singh, V.P.; Nelson, R.M.; Das, P. Treatment of a molasses based distillery effluent in a constructed wetland in central India. Water Sci. Technol. **2001,** *44* (11/12), 441–448.
87. Masi, F.; Conte, G.; Martinuzzi, N.; Pucci, B. Winery high organic content wastewaters treated by constructed wetlands in Mediterranean climate. In Proc. 8th Internat. Conf. Wetland Systems for Water Pollution Control; University of Dar-es-Salaam, Tanzania and IWA, 2002; 274–282.
88. Finlayson, M.; Chick, A.; von Oertzen, I.; Mitchell, D. Treatment of piggery effluent by an aquatic plant filter. Biol. Wastes **1987,** *19,* 179–196.
89. Strusevičius, Z.; Strusevičiene, S.M. Investigations of wastewater produced on cattle-breeding farms and its treatment in constructed wetlands. In Proceedings of the International Conference Constructed and Riverine Wetlands for Optimal Control of Wastewater at Catchment Scale; Mander, Ü., Vohla, C., Poom, A., Eds.; University of Tartu, Institute of Geography: Tartu, Estonia, 2003; 317–324.
90. Kantawanichkul, S.; Neamkam, P. Optimum recirculation ratio for nitrogen removal in a combined system: Vertical flow vegetated bed over horizontal flow sand bed. In *Wetlands: Nutrients, Metals and Mass Cycling*; Vymazal, J., Ed.; Backhuys Publishers: Leiden, the Netherlands, 2003; 75–86.
91. Comeau, Y.; Brisson, J.; Réville, J.P.; Forget, C.; Drizo, A. Phosphorus removal from trout farm effluents by constructed wetlands. Water Sci. Technol. **2001,** *44* (11/12), 55–60.
92. Schulz, C.; Gelbrecht, J.; Rennert, B. Treatment of rainbow trout farm effluents in constructed wetland with emergent plants and subsurface horizontal water flow. Aquaculture **2003,** *217,* 207–221.

93. Drizo, A.; Twohig, E.; Weber, D.; Bird, S.; Ross, D. Constructed wetlands for dairy effluent treatment in Vermont: two years of operation. In Proceedings of the 10th International Conference Wetland Systems for Water Pollution Control; MAOTDR 2006; Lisbon, Portugal, 2006; 1611–1621.
94. Kern, J.; Brettar, I. Nitrogen turnover in a subsurface constructed wetland receiving dairy farm wastewater. In *Treatment Wetlands for Water Quality Improvement*; Pries, J., Ed.; CH2M Hill Canada Limited: Waterloo, Ontario, 2002; 15–21.
95. Veenstra, S. the Netherlands. In *Constructed Wetlands for Wastewater Treatment in Europe;* Vymazal, J., Brix, H., Cooper, P.F., Green, M.B., Haberl, R., Eds.; Backhuys Publishers: Leiden, the Netherlands, 1998; 289–314.
96. Kato, K.; Koba, T.; Ietsugu, H.; Saigusa, T.; Nozoe, T.; Ko-bayashi, S.; Kitagawa, K.; Yanagiya, S. Early performance of hybrid reed bed system to treat milking parlour wastewater in cold climate in Japan. In Proceedings of the 10th International Conference Wetland Systems for Water Pollution Control; MAOTDR 2006; Lisbon, Portugal, 2006; 1111–1118.
97. Reeb, G.; Werckmann, M. First performance data on the use of two pilot-constructed wetlands for highly loaded non-domestic sewage. In *Natural and Constructed Wetlands: Nutrients, Metals and Management*; Vymazal, J., Ed.; Backhuys Publishers: Leiden, the Netherlands, 2005; 43–51.
98. Revitt, D.M.; Shutes, R.B.E.; Jones, R.H.; Forshaw, M.; Winter, B. The performance of vegetative treatment systems for highway runoff during dry and wet conditions. Sci. Total Environ. **2004,** *334–335,* 261–270.
99. Karrh, J.D.; Moriarty, J.; Kornue, J.J.; Knight, R.L. Sustainable management of aircraft anti/deicing process effluents using a subsurface-flow treatment wetland. In *Wetlands and Remediation II*; Nehring, W., Brauning, S.E., Eds.; Battelle Press: Columbus, OH, 2002; 187–195.
100. McGill, R.; Basran, D.; Flindall, R.; Pries, J. Vertical-flow constructed wetland for the treatment of glycol-laden stormwater runoff at Lester B. Pearson International Airport. In Proceedings of the 7th International Conference Wetland Systems for Water Pollution Control; University of Florida and IWA: Lake Buena Vista, FL, 2000; 1080–1081.
101. Headley, T.R.; Huett, D.O.; Davison, L. The removal of nutrients from plant nursery irrigation runoff in subsurface horizontal-flow wetlands. Water Sci. Technol. **2001,** *44* (11–12), 77–84.
102. Seo, D.C.; Hwang, S.H.; Kim, H.J.; Cho, J.S.; Lee, H.J.; DeLaune, R.D.; Jagsujinda, A.; Lee, S.T.; Seo, J.Y.; Heo, J.S. Evaluation of 2- and 3- stage combinations of vertical and horizontal flow constructed wetland for treating greenhouse wastewater. Ecol. Eng. **2008,** *32*, 121–132.
103. Wojciechowska, E.; Obarska-Pempkowiak, H. Performance of reed beds supplied with municipal landfill leachate. In *Wastewater Treatment, Plant Dynamics and Management in Constructed and Natural Wetlands*; Vymazal, J., Ed.; Springer: Dordrecht, the Netherlands, 2008; 251–265.
104. Headley, T.R.; Davison, L.; Yeomans, A. Removal of am-monia-N from landfill leachate by vertical flow wetland: A pilot study. In Proceedings of the 9th International Conference Wetland Systems for Water Pollution Control; ASTEE: Lyon, France, 2004; 143–150.
105. Bulc, T.G. Long term performance of a constructed wetland for landfill leachate treatment. Ecol. Eng. **2006,** *26,* 365–374.
106. Langergraber, G.; Giraldi, D.; Mena, J.; Meyer, D.; Peňa, M.; Toscano, A.; Brovelli, A.; Korkusuz, E.A. Recent developments in numerical modeling of subsurface flow constructed wetlands. Sci. Total Environ. **2009,** *407,* 3931–3943.

59

Wetlands: Sedimentation and Ecological Engineering

Timothy C. Granata
and J.F. Martin

Problems of Sedimentation in Wetlands

Sediment Accretion

The accretion of particulate material in wetlands limits the lifetime of these systems, advancing their succession to a terrestrial ecosystem as a function of sedimentation rates. Three sources of particulate material can be identified: suspended particles derived from the inflow, production of microbial particles in the wetland, and the production of detritus from macrovegetation. The latter two increase with increased primary production in wetlands. The former is a function of hydraulic loading, defined as the flow, Q, per unit area of wetland, A, i.e., QA^{-1}. Natural wetlands are prominent components of riparian ecosystems that exist either as floodplains, receiving periodic submersion, or as part of the riparian corridor, receiving continuous, though varying, inflow. In both cases, flow transports suspended sediments to wetlands. Suspended material is often deposited as current velocities decrease in broad, shallow expanses of wetlands. Sediment transport is a function of total flow (and thus rainfall intensity) in a watershed but also depends on the slope of the landscape and the susceptibility of soils to erosion. Steep slopes with highly unconsolidated soils will contribute to large suspended sediment loads during high runoff events. For low-flow events suspended sediment concentrations do not exceed 20 mg/L while for high flows they can exceed 80 mg/L.[1] In extreme cases of accretion, sedimentation rates of 26.67 cm/yr can bury emergent grasses because sediment accumulation outpaces plant growth.[2] In contrast, constructed wetlands are designed to treat wastewater having steady and high suspended solid concentrations, ranging from 20 to 75 mg/L, and volumetric ratios of settable solids from 5 to 20 mL/L.[3]

Reduced Conductivity of the Bed

Wetlands can experience reduced infiltration rates as particulate matter clogs bottom substrates. Generally, a reduction of hydraulic conductivity occurs with increasing hydraulic loading to the wetland. As a consequence of lower current velocities and higher sedimentation rates with increasing distance from the inlet, the hydraulic conductivity is lowest near the inlet and highest near the outlet. A general design criterion for wetlands is a hydraulic conductivity 1500 m/ day based on bed of 1.25–2.5 cm diameter gravel.[4]

Phosphorus Accumulation

The accretion of organic material on the bed results in wetlands with a high ion exchange capacity. The beds are partially responsible for removal of phosphates by sorption until they become saturated. The movement of phosphorus through the bed is slower than the hydraulic conductivity as a result of the storage (exchange) capacity of the bed.[5] Plant uptake of phosphates is limited because a reduced hydraulic conductivity decreases infiltration and thus transport of phosphates to plant roots.[6] A more consistent process for phosphate removal is by adsorption onto suspended sediments and their subsequent deposition.[7] Phosphorus (P) loading is directly proportional to suspended sediment concentrations and hydraulic loading. It is estimated that 40 mg P accompanies each gram of suspended solids entering wetlands.[1] Higher loading rates occur for flows over landscapes rich in P, such as agricultural lands. During periods of inundation, wetlands are characterized by aerobic and anaerobic soil zones. In anaerobic zones the decreased redox potential leads to an increase in the solubility of particulate phosphorus, which may be discharged from the wetlands. This dynamic feedback between removal of P by sedimentation and release of P after particulate accretion often leads to variable retention of P by wetlands. Despite mineralization of particulate P, wetlands are generally sinks for P, with removal as high as 97%.[8] The retention of P, however, decreases as P loading increases[9] and often decreases through the life of the wetland.[10]

Particle Types and Sedimentation Processes in Wetlands

The size distribution of particles in a wetland varies from nanometer-sized colloidal material through micrometer phytoplankton and heterotrophic organisms to millimeter- and centimeter-sized leaf detritus and sediments. For discrete particles, settling results from the force balance between particle weight (F_w) and drag on the particle such that

$$F_w = F_D = \rho C_D A_{pr} \frac{w^2}{2} = \left(\rho_p - \rho\right) V_g \tag{1}$$

where C_D is the drag coefficient on a 2D surface, A_{pr} is the projected area of the particle, w is the settling speed, ρ_p and ρ are the particle and water densities, respectively, V is the volume of the particle, and g is the gravitational acceleration (9.8 m/sec^2). Solving for settling rate, w, gives

$$w = \sqrt{2 \frac{\left(\rho_p - \rho\right) V_g}{\rho A_{pr} C_D}} \tag{2}$$

Thus, the settling rate of a particle is proportional to the excess density of the particles over that of water and the length scale of the particle defined by VA_{pr}^{-1}. Providing flow is not turbulent, specified as a Reynolds number Re <0.1, and particle shape is unimportant to the drag; C_D can be approximated as

$$C_D = \frac{24}{\mathrm{Re}} = \frac{24\nu}{2Rw} \tag{3}$$

where R is the radius of the particle. Assuming a spherical particle, the above equation reduces to

$$w = \frac{2\rho' g R^2}{9\nu} \tag{4}$$

the celebrated Stokes' Law for sinking particles. Here the excess density is written as

$$\rho' = \frac{\rho_p - \rho}{\rho} \tag{5}$$

For biological material, excess density is less than 0.1%, and size is the determining factor. For this reason small microbial organisms will sink slower than larger detrital material. The excess density of suspended sediments can reach 260%, indicating that the density is responsible for the high settling rates of suspended sediments.

Sedimentation in a wetland will occur if the settling speed is greater than the surface loading rate per area of wetland, or $w > QA^{-1}$. Given the large spectrum of particle types in natural flows, w is usually calculated based on the size and density of the particle type to be removed. Dividing the depth by the sinking rate gives the retention time in the wetland. Particles that are vertically well mixed as they enter a wetland will have a uniform concentration distribution comprised of the spectrum of particle sizes. If the settling rate of the particle, w_p, is greater than w, the particle will be deposited in the wetland cell. For particles with $w_p < w$, the fraction of particles that will be removed is X_p. Thus, the total particle removal by a wetland can be calculated as

$$\text{FR} = (1 - X_c) + QA^{-1} \sum X_r \Delta x \tag{6}$$

where $1–X_c$ is the fraction of particles removed with sinking speeds $w_p > w$ and $QA^{-1}\Sigma X_r \Delta x$ is the fraction with a settling speed $w_p < w$ that are removed over a distance Δx, along the path of the flow.[3] The crucial step in determining the sedimentation rate in wetlands is first determining the settling rate of each concentration fraction, either by settling columns or size-concentration measurements.

Often aggregations of particles form in wetlands,[11] a process called flocculation. Sedimentation rates of aggregates are dominated by organic particles of fractal dimension.[12] Aggregates tend to settle faster than their discrete component particles, and the only way to determine removal rate is by carefully transferring flocs to a settling column and determining the percent removal as a function of the height of the column and time.[3]

In wetlands with a high concentration of suspended solids, settling is affected by the contact between particles. This would occur in the bottom sediments of most wetlands where two settling processes can be identified: hindered and compression settling.[3] The hindered zone is marked by a large gradation in particle concentrations, which is less than the total concentrated in the compression zone. By plotting the concentration of particles over the height of a settling column as a function of time, a break point can be found in the time from hindered settling to compression settling.[3]

Because large, heavy particles settle out first, the highest sedimentation rates in wetlands are near the inlets. As mentioned above, the hydraulic conductivity through the bottom sediment decreases as clogging occurs. Because clogging is also a function of distance from the inlet, the most accurate estimate of the hydraulic conductivity is to measure the change in water surface elevation with increasing distance from the inlet. However, as Kadlec and Watson[11] show, this is not a simple function of Darcy's Law but must include evapotranspiration in the wetland. For newly constructed wetlands, the hydraulic conductivity of the bottom substrate (gravel, sand, and mud) cannot be used because clogging will result from sediment accretion.[4]

Ecological Engineering Solutions to Sedimentation

To keep suspended particles entrained by flows from accruing in wetlands and clogging bottom sediments, one or more settling basins can be included between the inlet channel and the wetland cell. This would have two effects: first, to collect all but the finest and least dense suspended particles and, second, to remove phosphorus from the inflow.

To overcome the problem of detrital accumulation in wetlands, woody plants could be substituted for grasses and periodically harvested. This would not only reduce the amount of biomass accrued in the wetland but would also increase the efficiency of nutrient uptake in the unclogged root zone and provide a potentially marketable resource. Because phytoplankton are the most abundant nutrient filters in a wetland and have intrinsically low sinking rates, a wetland could be designed with low retention of suspended plankton to further reduce sedimentation while improving nutrient removal. A settling basin could then be sized to accumulate these nutrient-rich particles for harvesting before nutrients are remineralized. To keep flocs of particles from forming in wetlands, cells could be mixed with aerators or inlet pumps. This would break up flocs, which would be exported as discrete particles.

Conclusions

Because wetlands are always shallow to promote macrophtye growth, they essentially act as flat plate collectors of sinking particles. The dominant particles in wetlands are suspended particles in the inflow and biomass of vegetation resulting from growth in the wetlands. The processes of sedimentation can reduce a wetland's storage capacity, its efficiency to retain nutrients, and its lifetime. Future efforts should be aimed at enhancing nutrient removal while reducing sedimentation. This can be done with engineered structures, such as settling basins and grit chambers, or by less costly technologies such as stilling wells. More innovative green solutions are on the horizon.

References

1. Porter, K.S. Nitrogen and phosphorous. In *Food Production, Waste, and the Environment*; Ann Arbor Science Publishers Inc.: Ann Arbor MI, 1975; 372.
2. Chung, C.-H. Ecological engineering of coastlines with salt-marsh plantations. In *Ecological Engineering: An Introduction to Ecotechnology;* Mitsch, W.J., Jørgenson, S.E., Eds.; John Wiley and Sons: New York, 1989; 472.
3. Metcalf, Eddy. *Wastewater Engineering: Treatment, Disposal, Reuse;* McGraw Hill: New York, 1991; 1334.
4. Watson, T.J.; Choate, K.D. Hydraulic conductivity of onsite constructed wetlands. 9th National Symposium of Individual and Small Community Sewage Systems, Fort Worth, Texas, Mar 11–14, 2001.
5. Kadlec, R.H.; Knight, R.L. *Treatment Wetlands;* CRC Press/Lewis Publishers: Boca Raton, FL, 1996; 839.
6. Davies, T.H.; Cottingham, P.D. Phosphorus removal from wastewater in a constructed wetland. In *Constructed Wetlands for Water Quality Improvement;* Moshiri, G.A., Ed.; Lewis Publishers: Boca Raton, FL, 1993; 632 pp.
7. Richardson, C.J.; Craft, C.B. Efficient phosphorus retention in wetlands: fact or fiction? In *Constructed Wetlands for Water Quality Improvement;* Moshiri, G.A., Ed.; Lewis Publishers: Boca Raton, FL, 1993; 632.
8. Mitsch, W.J.; Gosselink, J.G. *Wetlands;* Van Nostrand Reinhold Co.: New York, 2000; 920 pp.
9. Mitsch, W.J.; Reeder, B.C.; Klarer, D.M. Wetlands for nutrient control: Western Lake Erie. In *Ecological Engineering: An Introduction to Ecotechnology;* Mitsch, W.J., Jørgenson, S.E., Eds.; John and Wiley Sons: New York, 1989; 472 pp.
10. White, J.S.; Bayley, S.E.; Curtis, P.J. Sediment storage of phosphorus in a northern prairie wetland receiving municipal and agro-industrial wastewater. Ecol. Eng. **2000**, *14,* 127–138.
11. Kadlec, R.H.; Watson, J.T. Hydraulics and solids accumulation in a gravel bed treatment wetland. In *Constructed Wetlands for Water Quality Improvement;* Moshiri, G.A., Ed.; Lewis Publishers: Boca Raton, FL, 1993; 632 pp.
12. Kilps, J.R.; Logan, B.E.; Alldregde, A.L. Fractal dimensions of marine snow determined from image analysis of in situ photographs. Deep-Sea Res. **1994**, *41* (8), 1159–1169.

60

Wetlands: Treatment System Use

Constructed vs. Natural Wetlands

Kyle R. Mankin

How Do Wetlands Work?

Wetlands can be used to treat wastewater because they process contaminants. However, they treat wastewater more slowly than traditional treatment plants. Oxygen, and the manipulation of oxygen levels, is a primary concern for wastewater treatment because many of the necessary biological and chemical treatment processes require oxygen. Traditional treatment plants can easily manipulate oxygen levels by pumping air into the wastewater. Oxygen enters wetlands by slower, natural processes. Increasing oxygen concentration, by increasing wastewater contact with air, plant roots, or photosynthetic algae, often can enhance the processing ability of wetlands.

When considering wastewater treatment by constructed wetlands, five contaminant groups are of primary importance: sediments, organic matter, nutrients, pathogenic microbes, and metals. Wetlands slow down water movement, allowing sediments to settle out of the water. Organic matter can be processed, or decomposed, by highly competitive microbes. Less competitive microbes called nitrifiers process nitrogen. Both microbe types require oxygen. Because the nitrifiers are less competitive, oxygen levels become very important to insure that both organic matter and nitrogen are fully processed. The other two, pathogenic microbes and metals, are more situational, related to the specific waste being treated. Wetlands treat pathogenic microbes by detaining them until they naturally die off, are eaten by other predatory organisms in the wetland, or are exposed to UV radiation near the water surface. Metals are processed by being adsorbed to other particles and settling out of the water.

The remainder of this entry further explains wetland processes and design considerations. References are provided for more in-depth information.

Treatment Wetland Types

Constructed vs. Natural Wetlands

Wetlands constructed as treatment systems differ from natural wetlands in several important ways. Constructed wetlands usually are built with uniform depths and shapes designed to provide consistent detention times and maximize contaminant removal. In contrast, natural wetlands are irregular in depth and shape, which causes irregular flow, allows water to by-pass the shallow treatment zones by moving through the deeper channels, and leads to less effective treatment. In addition, water-quality regulations in the United States dictate that if a natural wetland is associated with an existing water body of the United States, as most are, wastewater discharges into the wetland must meet specific quality standards, similar to other water bodies. Wetlands constructed as wastewater treatment systems typically are located in uplands where wetlands did not exist before and are not subject to inflow water-quality regulations. Natural wetlands are *not* recommended for use as treatment wetlands.

Constructed wetlands increasingly are being used for wastewater treatment in a variety of applications (Table 1). Examples can be found of wetlands being used to treat municipal sewage, urban runoff, onsite residential wastewater, animal feedlot and barnyard runoff, cropland runoff, industrial wastewater, mine drainage, and landfill leachate. Each application takes advantage of a combination of physical, chemical, and biological processes characteristic of natural wetlands to reduce the concentration of contaminants in water. Such contaminants include sediments, organic materials, nutrients (particularly nitrogen and phosphorus), metals, microbial pathogens, and pesticides.

Free-Water vs. Submerged-Bed Wetlands

Constructed wetlands have two common types. Free-water surface (FWS) wetlands (also called surface-flow wetlands) have plants that grow in a shallow layer of water over a soil substrate (Figures 1 and 2). The location of the plants in the system can vary: the plants can float on the water surface with their roots suspended in the water (free-floating macrophyte systems); they can be rooted in the soil with the

TABLE 1 North American Wetlands as of 1994

Wastewater Type	Quantity	Size (ha) Minimum	Median	Maximum
Agricultural	58	0.0004	0.1	47
Industrial	13	0.03	10	1093
Municipal	159	0.004	2	500
Stormwater	6	0.2	8	42
Other	7	3	376	1406

Source: Kadlec et al.[1]

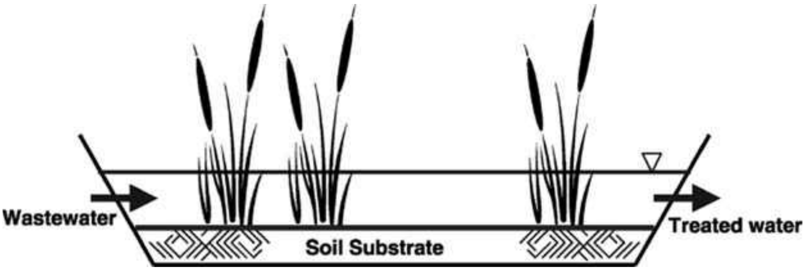


FIGURE 1 FWS wetland with emergent macrophytes.

FIGURE 2 A three-cell, FWS wetland for treating dairy wastewater. This system is in its first year of operation; plants were recently established.
Source: Photo by Peter Clark.

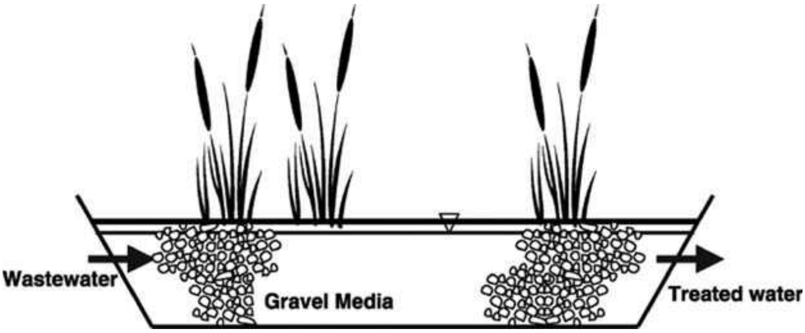


FIGURE 3 VSB wetland with emergent macrophytes.

entire leaves and stems below the water surface (submerged-macrophyte systems); they can be rooted in the soil having leaves and stems that rise above the water surface (emergent macrophyte systems); or the wetland may use a combination of planted and open-water zones. About two-thirds of existing wetlands as of 1994 were FWS.[1] In vegetated submergedbed (VSB) wetlands (also called subsurface flow wetlands or rock-plant filters), plants are rooted in a porous media, such as sand or gravel, and water flows through the media in either horizontal or vertical direction (Figures. 3 and 4). About one-quarter of treatment wetlands were VSB systems.[1] However, these systems are currently used in thousands of smaller-scale, onsite residential applications in the United States that do not appear in this database.

Treatment Processes

Many wastewaters entering constructed wetlands must be pretreated to avoid excessive contaminant loading, particularly of mineral and organic solids. Pretreatment technologies include septic tanks for onsite systems or anaerobic lagoons for animal waste, municipal, or mine-drainage treatment systems. In each case, the anaerobic condition in the pretreatment process reduces production of additional algae solids. Typical contaminant levels entering treatment wetlands are summarized in Table 2.

The wetland type impacts the processes used to retain or remove contaminants. In a VSB system, wastewater flows through pore spaces of the media and comes into direct contact with the roots of plants. In a FWS system, water flows across the media surface and contacts plant stems and leaves. In either system, solid particles, including sediments (clay and silt particles and colloids) and organic matter (manure particles, organic residues, and algae or other phytoplankton), settle out of the water column or are trapped or filtered as water passes through a wetland. Contaminants that are adsorbed

FIGURE 4 VSB wetland for treating onsite residential wastewater. This system uses gravel media and variety of wetland plants.
Source: Photo by Barbara Dallemand.

TABLE 2 Wetland Influent Concentrations

Wastewater Type	bod$_5$ (mg/L)	TSS (mg/L)	TN (mg/L)	nh$_4$-n (mg/L)	no$_3$-n (mg/L)	TP (mg/L)	FC (per 100 mL)
Residential-septic tank[2]	129–147	44–54	41–49	28–34	0–0.9	12–14	105.4–106.0
Municipal-primary[2]	40–200	55–230	20–84	15–40	0	4–15	105.0–107.0
Municipal-pond[2]	11–35	20–80	8–22	0.6–16	0.1–0.8	3–4	1008–1056
Livestock13 [avg.]	263	585	254	122	3.6	24	1.6 × 105
Livestock13 [median]	81	118	274	60	1.1	20	1.7 × 103
Landfill leachate[4]	312–729	241–7840	287–670	254–2074	0–3	0.9	

Note: BOD5 = 5-day biochemical oxygen demand, TSS = total suspended solids, TN = total nitrogen, NH4-N = ammonium nitrogen, NO3-N = nitrate nitrogen, TP = total phosphorus, FC = fecal coliform bacteria. FWS wetlands

to sediments (e.g., P, NH4, fecal bacteria) or absorbed within organic solids (e.g., nutrients) are also removed. However, these constituents can be re-suspended or desorbed back into the wetland water. This natural cycling of materials is an important function of wetlands, although it makes system design and interpretation of treatment complex.

Once entrapped, organic materials and associated contaminants are decomposed in wetlands by microbial and chemical transformations. In the degradation process, microbes use oxygen. The amount of oxygen used is related to the amount of organic material in the water. The controlled measurement of biochemical oxygen demand (BOD) is a common way to illustrate the amount of organic matter in water. When wastewater lacks oxygen, or is anaerobic, it requires the addition of oxygen to degrade organic matter. Oxygen is also required for transformation of ammonium to nitrite and nitrate (nitrification), whereas anaerobic conditions are required for transformation of nitrate to nitrogen gas (denitrification). Aerobic wetland conditions often remove metals by aerobic oxidation of iron; subsequently iron hydroxides and other metals precipitate in the wetland.[5] Although some oxygen diffuses into a wetland from the air, a common assumption is that oxygen also is transported through wetland plants and made available to microbes in close proximity to leaky roots.[6] This mechanism may be less important than once thought, though.[2] Treatment wetlands are thought to function effectively because they combine anaerobic zones in the water column with aerobic zones near the water interfaces with air and roots. However, because the microbes that break down organic carbon can out compete nitrifiers for oxygen, nitrogen removal in higher strength wastewaters is often low.

Design Considerations

Design and resulting effectiveness of constructed wetlands (Table 3) depend upon many factors: climate (precipitation, temperature, growing season, evapotranspiration), wastewater characteristics (constituents, loading, flow rate, and volume), topography, and wildlife activity. Wetland designs must specify total area; the number, depth, and size of wetland cells; hydraulic retention times; vegetation types and coverage; inlet and outlet configuration and location; and internal flow patterns.[2] Details for design can be found in numerous references[2,7–11] and some elements are discussed here.

VSB Wetlands

Properly designed VSB systems can achieve high removal rates. Treatment in a VSB wetland is governed by system residence time and wastewater contact with media and plant-root surfaces. Because of this, depth is a critical dimension and is often chosen according to the rooting depth of the selected plant (e.g., cattails: 30 cm; reeds: 40 cm; bulrush: 60 cm). Once depth is chosen, crosssectional area (and thus wetland width) is selected to assure adequate flow rates. Then, volume (and thus wetland length) is determined from the retention time needed to treat the wastewater to the desired quality. Proper design of inlet and outlet control structures helps maintain uniform flow patterns and depth, avoids problems with clogging and freezing, and minimizes system operation and maintenance (O and M) problems. High loading from influent solids and clogging can lead to surface flows and poor treatment. VSB systems must receive influents that are pretreated to remove solids (e.g., septic tank and effluent filter or anaerobic lagoon).

Properly designed FWS systems also can achieve high removal rates. Design typically follows one of two methods. The areal loading approach allows a designer to select the wetland surface area according to the influent load and the desired effluent quality.[13] Another approach allows a designer to select the wetland area by knowing the biological reaction rate, wastewater concentration, and flow rate along with selected water depth and target outflow water quality.[7,8] Again, depth is a critical dimension and is governed by plant tolerance to standing water and treatment objectives.

TABLE 3 Wetland Treatment (%)

Wastewater Type	BOD5	TSS	TN	nh_4-n	TP	FC	Metals
Municipal[11] [avg.]	74	70	53	54	57	—	—
Livestock[13] [avg.]	65	53	42	48	42	92	—
Landfill leachate[12] [range]	11–90	45–97	7–45	13–88	—	—	8–95+

Note: BOD5 = 5-day biochemical oxygen demand, TSS = total suspended solids, TN = total nitrogen, NH_4–N = ammonium nitrogen, TP = total phosphorus, FC = fecal coliform bacteria, metals = Fe, Cu, Pb, Ni, or Zn.

FWS vs. VSB Systems

Selection of the most appropriate wetland system depends on wastewater characteristics, treatment requirements, and site constraints. VSB systems generally require less land area, are less susceptible to freezing and mosquito problems, and have no exposed wastewater at the surface (avoiding contact-related health problems). FWS systems are less expensive to construct (without the cost of media), have greater potential for wildlife habitat, and are easier to maintain if solids accumulate.

Operation and Maintenance

O and M of treatment wetlands are relatively simple. The goal of an O and M plan is to assure that the wetland system continues to operate as planned, designed, and constructed. Several sources provide specific O and M guidance,[14] and most design manuals also contain such guidelines. Operation should be consistent with treatment objectives while maintaining structural integrity of the system, uniform flow conditions, and healthy vegetation as well as minimizing odors, nuisance pests and insects. Most maintenance plans require such items as checking water levels, checking for evidence of leaks or wildlife damage, and maintaining plant health on a weekly or monthly basis.

Conclusion

Constructed wetlands are complex natural-treatment systems that are well suited for many applications. They are low in cost and maintenance, provide significant reductions of many contaminants, and offer an aesthetic appearance. More work is needed to characterize treatment processes in constructed wetlands and improve design procedures to account for variability in wastewater and climate.

Acknowledgments

The author extends his gratitude to Kristina M. Boone, Associate Professor of Agricultural and Environmental Communications, Kansas State University for her significant review contribution to this entry, and to those at Kansas State University who have supported wetlands research and education.

References

1. Kadlec, R.H., Knight, R.L., Reed, S.C., Ruble, R.W., Eds. *Wetlands Treatment Database (North American Wetlands for Water Quality Treatment Database);* EPA/600/C-94/002; U.S. Environmental Protection Agency: Cincinnati, OH, 1994.
2. US EPA. *Constructed Wetlands Treatment of Municipal Wastewaters,* EPA/625/R-99/010; U.S. Environmental Protection Agency: Cincinnati, OH, 2000.
3. *CH2M Hill and Payne Engineering. Constructed Wetlands for Livestock Wastewater Management: Literature Review, Database, and Research Synthesis;* Gulf of Mexico Program— Nutrient Enrichment Committee: Montgomery, AL, 1997.
4. Kadlec, R.H. Constructed wetlands for treating landfill leachate (Ch. 2). In *Constructed Wetlands for the Treatment of Landfill Leachates;* Mulamoottil, G., McBean, E.A., Rovers, F., Eds.; Lewis Publishers: Boca Raton, FL, 1999.
5. US EPA. *Engineering Bulletin—Constructed Wetlands Treatment;* EPA/540/S-96/501; U.S. Environmental Protection Agency: Cincinnati, OH, 1996.
6. Brix, H. Do macrophytes play a role in constructed treatment wetlands? WaterSci. Technol. **1997**, *35* (5), 11–17.
7. Kadlec, R.H.; Knight, R.L. *Treatment Wetlands;* Lewis Publishers: Boca Raton, FL, 1996.

8. Reed, S.C.; Crites, R.W.; Middlebrooks, E.J. *Natural Systems for Waste Management and Treatment*, 2nd Ed.; McGraw-Hill: New York, 1995.
9. Steiner, G.R., Watson, J.T., Eds. *General Design, Construction, and Operation Guidelines: Constructed Wetlands Wastewater Treatment System for Small Users Including Individual Residences, Technical Report TVA/ MW-93/10;* Tennessee Valley Authority: Chattanooga, TN, 1993.
10. Powell, G.M.; Dallemand, B.L.; Mankin, K.R. *Rock-Plant Filter Design and Construction for Home Wastewater Systems*; MF-2340, http://www.oznet.ksu.edu/library/h20ql2/mf2340.pdf Kansas State University Cooperative Extension Service: Manhattan, KS, 1998.
11. Sievers, D.M. *Design of Submerged Flow Wetlands for Individual Homes and Small Wastewater Flows*; Special Report 457; Missouri Small Wastewater Flows Education and Research Center, 1993.
12. Mulamoottil, G., McBean, E.A., Rovers, F., Eds. *Constructed Wetlands for the Treatment of Landfill Leachates*; Lewis Publishers: Boca Raton, FL, 1999; 69, 86, 157, 212.
13. USDA-NRCS. *Constructed Wetlands for Agricultural Wastewater Treatment: Technical Requirements*; U.S. Department of Agriculture-Natural Resource Conservation Service: Washington, DC, 1991.
14. Powell, G.M.; Dallemand, B.L.; Mankin, K.R. *Rock-Plant Filter Operation, Maintenance, and Repair*, MF-2337; Available at http://www.oznet.ksu.edu/library/h20ql2/mf2337.pdf Kansas State University Cooperative Extension Service: Manhattan, KS, 1998.

VII

NEC: Natural Elements and Chemicals Found in Nature

61

Cyanobacteria: Eutrophic Freshwater Systems

Anja Gassner and
Martin V. Frey

Factors Leading to Cyanobacterial Dominance

The taxonomic composition of phytoplankton communities, the abundance and the relative dominance of the different species and groups present, undergo seasonal changes. This process of continuous community change is termed succession. Under undisturbed conditions, most phytoplankton populations are of relatively short duration. Typically, the growth and decline cycle of one specific population lasts, on average, 4 to 8 weeks. The "seasonal paradigm" of phytoplankton succession[1] describes the typical pattern of phytoplankton succession corresponding to the prevailing nutrient cycle in temperate, undisturbed lakes: a spring maximum of diatoms, sometimes followed by a second maximum in the autumn, an early summer maximum of Chlorophyceae (green algae) and a late summer maximum of Cyanophyta (blue-green algae).

It is generally accepted that with excess nutrients in the water column, in particular phosphorus, the phytoplankton flora deviates from the traditional seasonal community pattern with a shift toward cyanobacterial dominance. However, it must be stressed that nutrient limitation does not, in itself, provide cyanobacteria with the ability to become dominant; it is the combination of a multitude of abiotic and biotic factors. Enrichment experiments demonstrated that the maximum biomass of temperate lakes is ultimately limited by the phosphorus supply.[2] Increasing supplies of phosphorus lead to an increase of phytoplankton growth until other essential nutrients become limited. The first nutrient to become limited after phosphorus is usually nitrogen. Cyanobacteria are the only species that are able to fix atmospheric nitrogen. Whereas other algae become nitrogen limited, the ascendancy of nitrogen-fixing cyanobacteria is favored.

Apart from their ability of fixing atmospheric nitrogen, cyanobacteria feature some adaptations that enable them to outcompete other species. Eutrophic conditions result in large suspended stocks of phytoplankton, which reduce light penetration. Cyanobacteria possess gas vacuoles to control buoyancy. When subjected to suboptimal light conditions, they respond by increasing their buoyancy (regulated by the rate of photosynthesis) and move nearer to the surface and hence to the light.[2] Additionally, the possession of chlorophyll *a* together with phycobiliproteins allows them to harvest light efficiently and to grow in the shade of other species. Cyanobacteria are supposed to be more tolerant of high pH conditions and have an additional selective advantage at times of high photosynthesis because of their ability to use CO_2 as carbon source.[3] Some genera are able to offset the effects of photoinhibiting UV radiation encountered by near-surface populations. The resistance to photoinhibition is achieved by producing increased amounts of carotenoid pigments, which act as "sunscreens."[4] Once established, cyanobacteria are able to inhibit the growth of other algae by producing secondary metabolites that are toxic to species of other genera.[5]

Consequences of Cyanobacterial Blooms

Like any phytoplankton, bloom proliferation of blue-green algae reduces water quality in terms of human water use but also results in a reduction in diversity of the aquatic species assemblage at all trophic levels. The presence of "pea soup green" water, the accumulation of malodorous decaying algal cells, and the buildup of sediments rich in organic matter lead to user avoidance with the associated problems and implications for water quality management. The most obvious sign of an advanced blue-green algae bloom is the formation of green "scum," which leads to deoxygenation of underlying waters, subsequent fish kills, foul odors, and lowered aesthetic values of affected waters.[6] In addition, certain genera and species produce taste and odor compounds, typically geosmin and 2-methyl iso-borneol, which cause non-hazardous but unpleasant problems for suppliers and users of potable water.[4]

The most serious public health concerns associated with cyanobacteria arise from their ability to produce toxins. Since the first published reported incidence of mammal deaths related to a toxic cyanobacterial bloom in 1978, more than 12 species belonging to nine genera of blue-green algae have been implicated in animal poisoning.[7] For human exposure, routes are the oral route via drinking water, the dermal route during recreational use of lakes and rivers, or consumption of algal health food tablets. Toxins produced in a random and unpredictable fashion by cyanobacteria are called cyanotoxins and classified functionally into hepatotoxins, neurotoxins, and cytotoxins. Additionally, some cyanobacteria produce the lesser toxic lipopolysaccharides (LPS) and other secondary metabolites that may be of potential pharamacological use.[8] One of the most tragic encounters of humans with cyanobacterial toxins led to the deaths of 60 dialysis patients due to contaminated water supply used in a hemodialysis unit.[9] Presently, a drinking water guideline of $1\,\text{ng L}^{-1}$ of toxin has been developed and implemented only for microcystin-LR.[4] Haider et al.[8] stress that the biggest challenge for water treatment procedures for the removal of cyanobacterial toxins is that one is faced with soluble and suspended substances. Thus, the most common treatment, chlorination, in general has been found not to be an effective process in destroying cyanotoxins.

Monitoring and Management of Algal Blooms

Drinking water treatment strategies are not always successful in removing algal toxins. Thus, detection of early-stage (emergent) blooms of cyanobacteria, especially if the bloom has not started to produce toxins, is important to allow municipalities and recreation facilities to implement a response plan. It has been shown that remote sensing technology can be used to estimate the concentration and distribution of cyanobacteria through measurement of the concentration of the pigment phycocyanin.[10]

Once detected, the growth of nuisance algae is prevented by the use of chemicals; the commonest is copper sulfate. Other algicides include phenolic compounds, amide derivatives, quaternary ammonium compounds, and quinone derivatives. Dichloronaphthoquinone is selectively toxic to blue-greens. The inherent problem of algicides is that on cell lysis, toxins contained in the algae cell are released into the surrounding water. In 1979, almost 150 people had to be hospitalized for treatment of liver damage after a reservoir contaminated with *Cylindrospermopsis* was treated with copper sulfate.[4] Biological control by zooplankton is, in principle, possible, although not always practical or effective because of the low nutrient adequacy, toxicity, and inconvenient size and shape of most blue-green algae.

The only zooplankton reported to successfully graze on blue greens is *Daphnia* sp., but it tends to decrease with increasing nutrient content of the water[11] More effective is the use of microorganisms, as certain chytrids (fungal pathogens) and cyanophages (viral pathogens) specifically infest akinetes and other heterocysts, whereas Myxobacteriales (bacterial pathogens) can affect rapid lysis of a wide range of unicellular and filamentous blue-greens, although heterocysts and akinetes remain generally unaffected.[12]

The consensus regarding the management of blue-green algal blooms is the management of excess nutrient loads into receiving water bodies.[13,14] Management options can be divided into two broad categories: catchment management (decrease of nutrient export) or lake management (decrease of internal nutrient supply). Catchment options are, e.g., management of urban and agricultural runoff, biological and chemical treatment of wastewater, nutrient diversion, and implementation of legislation. Lake management options are dredging, chemical sediment treatment, and biomanipulation.[13]

Conclusions

Cyanobacteria pose a serious threat to ecosystem health and human livelihood. From a human perspective, the most serious threat associated with blue-greens are their toxins. Routes for human exposure are the oral route via drinking water, the dermal route during recreational use of lakes and rivers, or consumption of algal health food tablets. Removal of these algae and their toxins from water bodies poses a great logistical problem. However, it is important to understand that the proliferation of blue-greens and thus the presence of their toxins is a response to human-induced "cultural" eutrophication. Increasing awareness of the need of proper watershed management is urgently needed among municipalities and stakeholders, especially because chlorination has been shown not to be very effective in removing toxins from the water.

References

1. Reynolds, C.S. *The Ecology of Freshwater Plankton*; Cambridge University Press: Cambridge, 1983.
2. Havens, K.E.; James, R.T.; East, T.L.; Smith, V.H. N:P ratios, light limitation, and cyanobacterial dominance in a subtropical lake impacted by non-point source nutrient pollution. Environ. Pollut. **2003**, *122* (3), 379–390.
3. Shapiro, J. Current beliefs regarding dominance by blue-greens: the case for the importance of CO_2 and pH. Proceedings of the International Association of Theoretical and Applied Limnology, Verh. Int. Verein. Theor. Angew. Lim-nol. **1990**, *24,* 38–54.
4. *Chorus, I.; Bartram, J., Eds.; WHO (World Health Organization) Toxic Cyanobacteria in Water: A Guide to Their Public Health Consequences, Monitoring and Management, 1st Ed. E and F Spon: London, 1999.*
5. Carr, N.G.; Whitton, B.A. The biology of blue-green algae. Bot. Monogr. Vol. 9. University of California Press, Berkeley.
6. Codd, G.A. Cyanobacterial toxins, the perception of water quality, and the prioritisation of eutrophication control. Ecol. Eng. **2000**, *16* (1), 51–60.
7. Carmichael, W.W. Health effects of toxin-producing cyanobacteria: "The cyanoHABs." Hum. Ecol. Risk Assess. **2001**, *7,* 1393–1407.
8. Haider, S.; Naithani, V.; Viswanathan, P.N.; Kakkar, P. Cyanobacterial toxins: a growing environmental concern. Chemosphere **2003**, *52* (1), 1–21.
9. Pouria, S.; de Andrade, A.; Barbosa, J.; Cavalcanti, R.; Barreto, V.; Ward, C. Fatal microcystin intoxication in haemo-dialysis unit in Caruaru, Brazil. Lancet **1998**, *352,* 21–26.
10. Richardson, L.L. Remote sensing of algal bloom dynamics. BioScience **1996**, *44*, 492–501.
11. Epp, G.T. Grazing on filamentous cyanobacteria by *Daphnia pulicaria.* Limnol. Oceanogr. **1996**, *41* (3), 560–567.

12. Philips, E.J.; Monegue, R.L.; Aldridge, F.J. Cyanophages which impact bloom-forming cyanobacteria. J. Aquat. Plant Manage. **1990**, *28*, 92–97.
13. Harding, W.R.; Quick, A.J.R. Management options for shallow hypertrophic lakes, with particular reference to Zeekoevlei, Cape Town. S. Afr. J. Aquat. Sci. **1992**, *18* (1/2), 3–19.
14. Herath, G. Freshwater algal blooms and their control: comparison of the European and Australian experience. J. Environ. Manage. **1997**, *51,* 217–227.

62

Estuaries

Claude
Amiard-Triquet

Introduction

In the whole world, estuaries are the seat of the same paradox: they are among the most productive ecosystems while being strongly impacted by anthropogenic activities. Considering the value of the world's ecosystem service and natural capital, estuaries are among the most productive systems.[1] More than 76% of all commercially and recreationally important fish and shellfish species are estuarine dependent.[2] By using the specific search term "estuary" in the Marine Pollution Bulletin database, 553 titles emerged to date (July 2011), highlighting the importance of pollution threats to these ecosystems. This estuarine paradox is well exemplified in Figure 1, showing the mouth of the Loire estuary, one of the big estuaries along the northwestern Atlantic coast. On the north bank, the town of Saint Nazaire (65,000 inhabitants) is highly industrialized, being particularly famous for its shipyards. Between Saint Nazaire and the urban community of Nantes (580,000 inhabitants) 60 km upstream, industrial areas receive large plants, such as a company producing fertilizers, an oil refinery, and an electrothermal plant. Agricultural areas destined to meadows and livestock farming, open fields, and specialized cultures (market gardening, horticulture, wine yards, etc.) are inserted between industrial areas. The south bank is less urbanized, less industrialized, and consequently less artificialized. As shown in the picture, at low tide, large mudflats emerge. They are home to rich invertebrate fauna, including small crustaceans, bivalves, and worms, which are at the bases of food chains leading to different fish species (flounder, eel, sole, sea bass, etc.) of economic interest and to vast populations of wading birds.

Worldwide, there are some 1200 major estuaries (discharges of 10 m^3/s), with a total area of approximately 50 million ha.[1] Tidal flats and estuaries have been widely degraded and lost. Sea traffic has required the fitting of dikes, the dredging of channels, and land reclamation to support harbors and industry with drastic ecological consequences such as the reduction of mudflats and also wetlands, which are important for the protection of water quality and are ecosystems of floristic and faunistic interest.

FIGURE 1 The Loire estuary near its mouth showing the contrast between the north bank, urbanized and industrialized, and the south bank, with intertidal mudflats that are crucial for feeding of fish and wading birds.

Origins of the Problems of Estuaries

Drivers of loss and change to wetland ecosystems have been listed in the Millennium Ecosystem Assessment.[1] Influx of nutrients due to river transport is at the basis of the biological wealth of estuarine areas, but on the other hand, the fluxes of reactive (biologically available) nitrogen and high phosphorus loading from anthropogenic sources result in eutrophication.[3] The degradation of primary producers responsible for eutrophication is oxygen consuming as is also the degradation of anthropogenic influx of NH_4 (present in wastewater treatment effluents) into nitrates. Thus, episodes of hypoxia may be observed, and in many estuaries they are also favored by the presence of a natural entrapment zone [estuarine turbidity maximum (ETM)] where the accumulation of suspended organic and mineral matter enhances bacterial activity.[4] In estuarine environments, bacteria are generally attached to particles, and their behavior is related to the dynamics of suspended matter.[5] Bianchi[3] reports that the abundance and production of virus-like particles (viriobenthos) in sediments are considerably higher than those found for water column virus particles (virioplankton).

The processes controlling movement of sediments in estuaries have been described by Bianchi.[3] The location of the ETM is generally considered to be controlled by tidal amplitude, volume of river flow, and channel bathymetry. Due to rapid and high sedimentation rates in the ETM, the accumulation of particles in the benthic boundary layer can result in the formation of mobile and fluid muds, [3] but resuspension episodes frequently occur, for instance, during spring tide and storm events.

Urban and industrial activities are responsible for the input of contaminated effluents and microbes in the aquatic environment, whereas agriculture results in diffuse inputs of pesticides and fertilizers. Contaminants are a source of concern for the growth and reproduction of cultivated species as well as a risk for the health quality of seafood products, particularly for pollutants able to biomagnify in food chains (e.g., methylmercury, polychlorinated biphenyls [PCBs]). In industrial areas, the role of estuaries as waterways leads to dredging activities in order to keep a convenient bathymetry. Because dredged sediments are loaded with a cocktail of many classes of contaminants, dredging activities are a serious source of concern.

The *Eisler's Encyclopedia of Environmentally Hazardous Priority Chemicals*[6] provides 142 instances when using the search terms "estuary" or "estuarine." Bianchi[3] also provides a review of the literature dealing with nutrients and contaminants entering estuaries: trace metals, either essential at low doses (e.g., Cu, Zn) or nonessential and toxic even at low doses (e.g., Hg, Pb); PCBs, very stable so they are still present despite their ban; and polycyclic aromatic hydrocarbons (PAHs). Bianchi[3] also reports that inputs of these multiple stressors can interact to reduce, enhance, and/or mask the individual effects of each. Emerging contaminants of concern in coastal and estuarine environments have been reviewed by Hale and La Guardia.[7] They include brominated flame retardants, natural and synthetic estrogens, alkylphenol ethoxylates, and associated degradation products, pharmaceuticals, and personal care products. The recent advances in nanotechnology and the increasing use of nanomaterials in every sector of society have resulted in uncertainties regarding environmental impacts. Deposit-feeding benthic estuarine organisms may be particularly at risk given the likelihood of nanoparticles to agglomerate, aggregate, and settle.[8]

The public and scientists realized environmental problems of estuaries first because of the disastrous depletion of populations of migratory fish, especially salmon and eels. This blindness is partly explained by the erroneous confidence in the following: 1) the power of dilution of pollutant fluxes in huge volumes of seawater at river mouths; and 2) the ability of estuarine species to cope with chemical stress as they are able to cope with dramatic changes in ecological conditions (salinity, temperature, turbidity, oxygen). On the contrary, organisms living at the limits of their tolerance to natural variations are generally more sensitive to any additional stress.[9,10]

Environmental Assessment of Estuarine Ecosystems

Environmental assessment is classically based on a triad of analysis including the following: 1) chemical analyses in environmental matrices; 2) bioassay methods for the evaluation of toxic effects of contaminants; and 3) biological responses at the level of community structure. This approach was particularly well developed as the Sediment Quality Triad[12] since in the aquatic environment, sediment is the main store for most contaminants entering ecosystems. In estuaries, all assessment techniques need to consider the unique and complex dynamic processes at work, in particular, salinity effects, but also key estuarine parameters such as temperature, pH, dissolved O_2, redox potential, hydrodynamics and sedimentary processes, and their spatiotemporal changes.[13,14] These parameters, indeed, strongly influence estuarine biogeochemical processes, controlling contaminant exchanges between sediment and the water column as well as contaminant bioavailability and toxicity.[13] They also govern the biological features of estuarine communities according to the concept of the estuarine quality paradox.[15,16]

Legislation has been adopted on a worldwide scale to determine the ecological integrity of surface waters including estuaries (United States' Clean Water Act [CWA], 1972; European Community Water Framework Directive, [WFD]).[17] The so-called ecological quality status (EcoQS)[17] must be assessed by comparison with undisturbed conditions, but at the beginning of the third millennium, the formulation of reference values is highly questionable. Anyway, effects at the community level become significant only after severe environmental degradation has already occurred, thus leading to expensive remediation processes. In this context, the methodology of biomarkers provides predictive tools applicable much earlier in any environmental degradation process.[18,19]

Chemical Analyses in Environmental Matrices

Knoery and Claisse[20] have described the experience of the French marine chemical monitoring network over the last 33 years. They highlight the rationale for using different environmental matrices, namely, water, sediment, and biota (suspension feeders such as *Mytilus* sp. and *Crassostrea gigas,* as recommended in NAS[21]), to describe spatiotemporal variations of chemical pollution (PAHs, PCBs, organochlorinated pesticides, metals). Despite being recognized in the WFD, water is not a suitable matrix to answer this aim due to analytical problems at environmentally very low levels of contaminants combined with the low spatial and temporal representativeness of water samples collected at insufficient frequencies (even if recent progress in the field of passive samplers allows the measurement of time-weighted average concentrations; see, for instance, Mills et al.[22]). Using analyses in organisms, which reflect past exposure in the environment, with higher contaminant levels due to bioaccumulation, tremendously increases the reliability of data. This strategy was complemented by monitoring chemical contaminants in sediments for integration over several years. The public can access the data as time series plots through Ifremer's Web site at http://wwz.ifremer.fr/envlit/resultats/surval__1 and as maps at http://wwz.ifremer.fr/var/envlit/storage/documents/parammaps/contaminants-chimiques/index.html. The main estuaries are clearly among the most contaminated sites.

However, the gross concentrations of contaminants in any of the compartments of the environment have poor ecotoxicological significance. A greater knowledge of the biogeochemical cycling in estuaries, which involves the transformation, fate, and transport of chemical substances, is critical in understanding the effects of contaminants.[3] The role of bacteria in biogeochemical cycles is well established, and in the field of pollution studies, a focus has been made on the role of bacteria in the methylation of mercury, a process that is responsible for a higher bioavailability and a higher toxicity of this element for aquatic organisms.[6] More generally, there is clearly a need to put more "bio" into biogeochemical cycles (Ouddane et al. in Ref. [14]).

Bioassay Methods for the Evaluation of Toxic Effects of Contaminants

In their book devoted to *Coastal and Estuarine Risk Assessment,* Newman, Morris, and Hale[7] underlined the predominant freshwater focus for risk assessment and discussed the use of more abundant data for freshwater species to predict consequences to saltwater species. They concluded that improving the current practice effectively will require substantially more data for marine species. In an inventory of marine biotest methods, estuarine species, belonging to different taxa, are listed among relevant biological models.[23] On the copepod *Acartia tonsa*—known as highly tolerant toward salinity and temperature fluctuations—combined stresses from suboptimal temperature and salinity and toxic stress were tested.[24] However, toxicological stress altered the established response pattern of *A. tonsa* toward temperature and salinity, confirming that organisms living in stressful or extreme environments may experience profound interaction effects of anthropogenic and natural stress factors. On the other hand, a review performed by the European Centre for Ecotoxicology and Toxicology of Chemicals[25] suggests a reasonable correlation between the ecotoxicological responses of freshwater and saltwater biota.

Bioassays are used for the determination of predicted no-effects concentrations (PNECs) as described in detail in the Technical Guidance Document on Risk Assessment (TGD) in support of European Commission regulations.[26] The determination of the PNECs takes into account the number of bioassay results available for different species representative of different trophic levels and the availability of long-term toxicity data. However, the notion of long term in the framework of the methodology of bioassays[23] differs from the notion of long term when considering ecological consequences at higher levels of biological organization.

The relevance of the TGD[26] for in situ assessment in estuarine environments was tested by determining local PNECs for atrazine in a highly polluted estuary, the Seine estuary, France. For downstream, middle, and upstream sections of this estuary, key species important for local food chains were selected. Because the number of ecotoxicological data was restricted at the local scale compared with the global scale, a higher security factor needed to be used. In fact, differences between European Union PNECs[27] and local PNECs were limited.[28]

Estuarine Community Structure

Estuarine Quality Paradox

In agreement with Dauvin[15] and Elliott and Quintino[16] describe this paradox as follows: "the dominant estuarine faunal and floral community is adapted to and reflects the high spatial and temporal variability of highly naturally-stressed areas. However, this community has features very similar to those found in anthropogenically-stressed areas, thus making it difficult to detect anthropogenically-induced stress in estuaries. Furthermore, as estuaries are naturally organically rich, the biota thus is similar to anthropogenically-organic rich areas. Because of this, there is a danger that any indices based on these features and used to plan environmental improvements will be flawed." Although numerous bioindicators and indices are used to define the EcoQS of coastal waters, very few of them were developed specifically for environments with a mosaic of conditions and salinity levels. Natural and anthropogenic stressors make estuaries highly heterogeneous environments, which may generate thousands of potential combinations (Figure 2). According to the WFD, the elements of biological quality must be determined by comparison with reference conditions. Because of habitat heterogeneity in estuaries, it has been proposed that reference conditions should be habitat specific in order to reflect natural gradients.[29]

Tolerance and the Monitoring of Communities

This topic has been recently reviewed by Berthet et al.[11] A biological community is composed of different species, the inherent sensitivity of which toward a given toxicant is highly variable. Thus, in a contaminated environment, the most sensitive organisms are lost, whereas tolerant organisms are maintained. The variability of interspecific responses is used in the concept of the pollution-induced community tolerance developed by Blanck, Wängberg, and Molander[31] and revisited by Tlili and Montuelle.[11] Depending on the functional role of different species in the community, different indirect effects of tolerance may be expected. If a sensitive species is a prey or a host species, its extinction will lead to a depletion of its predator or symbiont populations. On the other hand, the extinction of a sensitive species that is a competitor or a predator of a tolerant species will favor the latter.

Determining ecological integrity to answer the needs of legislations (CWA, WFD) typically emphasizes analyses of phytoplankton, macroalgae, angiosperms, benthic macrofauna, and ichthyofauna. In this aim, existing data were revisited, whereas new data were obtained, for instance, in the framework of the WISER Project (Water bodies in Europe: Integrative Systems to assess Ecological status and

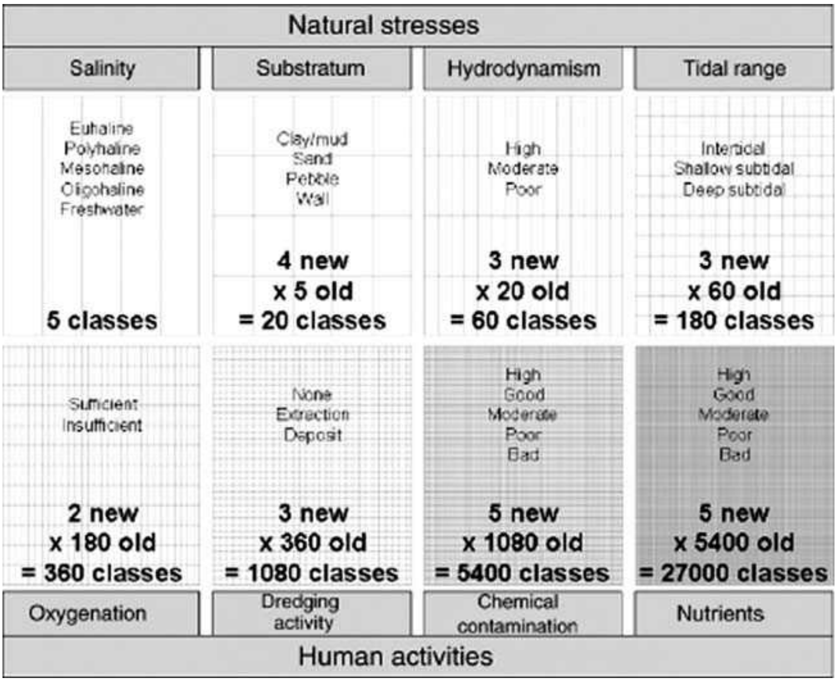


FIGURE 2 Heterogeneity of estuarine conditions that can potentially result from interactions between natural stressors and anthropogenic activities.
Source: Adapted from Dauvin et al.,[30] with permission.

Recovery, http://www.wiser.eu/programme/). Only a few results had already been published in international journals, but many reports have been prepared, and an inventory was made by Courrat, Foussard, and Lepage[32] for the different categories of biotic communities in the framework of BEEST (for good ecological status of large estuaries), a research project funded by the French ministry of environment, http://seine-aval.crihan.fr/web/pages.jsp?currentNodeId=7).

Concerning phytoplankton, the WISER project aims at developing assemblage phytoplankton metrics, including the potential use of pigment data, taking into account sources of uncertainty on the determination of biomass and community composition due to spatiotemporal heterogeneity. In the ETM region of macrotidal estuaries, high light attenuation results in low primary production, at least as phytoplankton in the water column.[33,34] Thus, it has been proposed to explore the feasibility of using microphytobenthos, an important component in primary estuarine production.[35] The spatiotemporal variability of microphytobenthos on intertidal mudflats is largely unknown because traditional techniques based upon spot sampling techniques do not allow a sound assessment of large-scale distribution. It is only recently that remote sensing techniques have been applied to obtain synoptic information on microphytobenthos distribution in estuarine intertidal zones.[36,37]

Due to phytoplankton paucity in the ETM region, it is interesting to propose another indicator for the water column. In estuaries, the omnivorous copepod *Eurytemora affinis* is especially prevalent in ETM regions worldwide(Forget-Leray et al. in Ref. [14]).[33] Many references quoted in Souissi and Devreker[38] show that the life cycle of certain species is well described and responds to the presence of pollutants or toxins. These authors also highlight that estuarine copepods, restricted to water masses with a well-defined salinity, are not affected by spatiotemporal natural fluctuations as much as benthic species. This limitation of confounding factors favors clear responses to stress. An international inquiry about the potential use of zooplankton as an indicator of estuarine water quality (22 participants from 13 countries in Europe, North America, and Asia) was made in the framework of BEEST.[38] Because of its abundance and key role in the estuarine trophic web, zooplankton was recognized as a good candidate.

Indicators and index approaches based on benthic macroinvertebrate communities are the most consistently emphasized methods for the environmental assessment of aquatic ecosystems. According to Dauvin, Bellan, and Bellan-Santini,[39] this is because macrobenthic organisms are "relatively non-mobile and therefore useful for studying the local effects of physical and chemical perturbations; some of these species are long-lived; their taxonomy and their quantitative sampling is relatively easy; and there is extensive literature on their distribution in specific environments and on the effect of the various stresses that these organisms could encounter." In the framework of the WISER project, 13 single metrics (abundance, number of taxa, and several diversity and sensitivity indices) and 8 of the most common indices used within the WFD for benthic assessment were tested. The different indices are largely consistent in their response to pressure gradient (preliminary classification based on professional judgment), but in transitional waters, inconsistencies between indicator responses were most pronounced.[40]

Fish-based indices have been recently reviewed.[32,41] In most European countries, sound scientific bases exist for regional assessment (Basque country, [42] the United Kingdom, [43] Belgium, [44] France[45,46]).

Amiard-Triquet and Rainbow[14] have underlined that on estuarine intertidal mudflats, bacteria, microalgae, and meiofauna (including foraminiferans, nematodes, and harpacticoid copepods) are relatively abundant compared with macrofaunal taxa. They also tend to exhibit higher diversity than macrofauna, increasing the range of potential responses to pollutants and potential sensitivity for the detection of an impact. Thus, data can be obtained with small samples and minimized impact on the study site while having high information content that facilitates statistical analysis. In addition, because of short generation times and low dispersal characteristics of most of these taxa, rapid population responses to environmental perturbations may be expected. On the other hand, such microscopic taxa often require specialist taxonomic skills, expertise that is of decreasing availability throughout the world.

Question of the Reference Site

All index developers invest a large amount of effort on the formulation of reference values, that is, the quality or conservation value given to pristine, undisturbed, condition, or reference status.[41] At the best, sites as clean as possible may be used as references, but of course, they will not be found among large estuaries, strongly occupied by human populations, but only in small estuaries, less affected by human impacts, but which also differ from large ones by many natural factors.[14] These confounding factors, which cannot be controlled (salinity, granulometry, organic content, hydrodynamics, food availability, etc.), must at least be measured in parallel with ecological and ecotoxicological determinations to mitigate the interpretation of indicators and indices. In this context, the WFD provides a list of hydromorphological, chemical, and physicochemical parameters that must be monitored as sustaining biological parameters.

The notions of biological or ecological integrity and the good chemical and ecological status, which are, respectively, at the basis of the U.S. CWA and the WFD, are still topics of discussion. They are based on an ecocentric point of view including the conservation of biodiversity with reference to a nearly undisturbed situation, in agreement with the concepts of climax and resilience, which have been developed in the 1960s but today are reexamined by scientific ecologists.[47,48] For nonscientists involved in ecology, reaching the good status may be more or less perceived as the recovery of a kind of Garden of Eden, the so-called pristine areas that in fact no longer exist at the present time. Sociological inquiries carried out in France in the framework of the BEEST project showed that for users and inhabitants living in estuarine areas, points of view differ depending on their relation to estuaries: fishing, social and recreational activities, aesthetic relationship, etc.

Concurrently to the ecocentric approach, an anthropocentric approach (Figure 3), based on the conservation of goods and services, may be proposed. It leads to the concept of good ecological potential

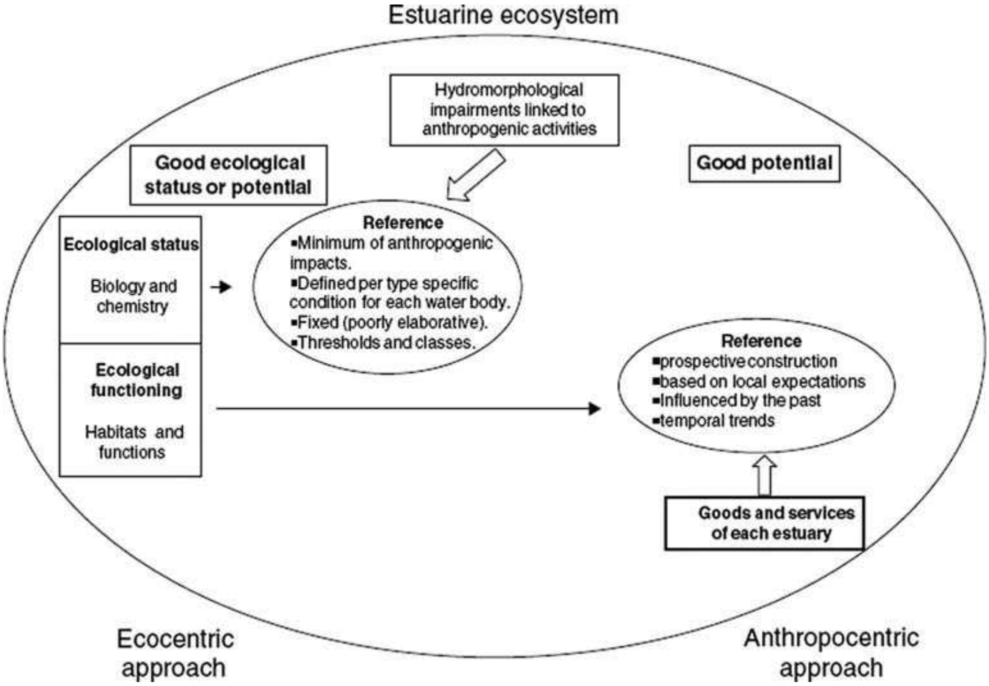


FIGURE 3 From the ecological integrity (CWA) or good ecological status (WFD) to the good potential of estuaries. **Source:** Adapted from Synthèse du projet BEEST Vers une approche multicritère du Bon Etat des grands ESTuaires atlantiques.[49]

rather than good ecological status, allowing a prospective construction including, for instance, expectations and demands of the public or uncertainties like global change. It is also in agreement with the WFD, which mentions limitations that must be taken into account (environmental and social needs, excessive costs, maintaining people's security, higher general interest) and explicitly imposes the information, consultation, and involvement of the public, including users.

Assessing the good potential of an estuary needs tools other than those useful to describe only the situation at a given time (Figure 3). It is indispensable to characterize the key elements of ecological functioning through the following: 1) by considering the concept of functional habitat; 2) by using indicators allowing a better integration of functional aspects of estuarine ecosystems; and 3) by searching out evolutionary trends rather than quality thresholds, which are not really relevant in this highly variable context.[49]

New Tools

Ecosystem Functioning or Ecosystem Structure?

According to Elliott and Quintino,[16] "functional characteristics either as well as or rather than structural ones should be used in detecting environmental perturbations in estuaries." Indicators described above may be used for a better characterization of ecosystem functioning. Phytoplankton and microphytobenthos may serve as proxies for primary production, and in addition, Underwood and Kromkamp[35] underscore that microphytobenthic biofilms may play an important role in biogeochemical cycles (exchanges between sediment deposits and water phase, effects on bacterial processes). Because of its central place in the trophic web, zooplankton is a useful indicator of good ecological functioning rather than an indicator of structure.[38] Fish-based indices are generally constructed by using ecological and trophic guilds, thus allowing the description and explanation of transitional waters' community structure and estuarine functionalities.[46] For instance, taking into account juvenile marine fish can allow the characterization of the role of a given estuary as nursery.

Biomarkers

Despite biomarkers having been defined by Depledge in 1994[50]: "A biochemical, cellular, physiological or behavioral variation that can be measured in tissue or body fluid samples or at the level of whole organisms that provides evidence of exposure to and/or effects of, one or more chemical pollutants (and/or radiations)," they are still "new" since their use is limited or ignored in environmental regulations (CWA, WFD) even if they are well incorporated in some monitoring procedures, for instance, under the OSPAR Convention. They suffer mainly from three kinds of criticisms: 1) their responsiveness to confounding factors; 2) their insufficient specificity of response toward a given class of chemical; and 3) their lack of ecological relevance.

As for other parameters reviewed above, biomarkers may be responsive to spatiotemporal changes of natural conditions. In addition, they also respond to biological parameters such as the size, sex, and reproductive status of organisms. Thus, before using a sentinel species, it is needed to determine the sources of fluctuations[51,52] to limit them and, for those that cannot be controlled, to determine correction factors.[53]

"Specific" and "nonspecific" biomarkers may complement each other to assess the health status of estuarine ecosystems.[14] Among nonspecific biomarkers, those that are linked directly or indirectly to the success of reproduction and the sustainability of populations have been termed "biomarkers of ecological relevance."[54] They include changes in lysosomal stability, immunotoxicity, energy metabolism, behavior, endocrine disruption, and genotoxicity. When such impairments are demonstrated, it is necessary to identify the main classes of contaminants which may be responsible for such impairments by using a multibiomarker approach based on "more specific biomarkers" (e.g., metallothioneins induced by different metals, delta aminolevulinic acid dehydratase, specific to lead toxicity; imposex of gastropods in the presence of tributyltin). Then, the quantification and characterization (specific fractions, metabolites) of chemicals in water, sediments, and biota may allow the validation of hypotheses based on biomarker studies.

How Can We Solve the Problem?

Preventing and controlling pollution must be based on a combined approach aiming at reducing pollution at source by determining emission limit values and environmental quality standards (EQSs, e.g., WFD). Environmental management aiming at the improvement of chemical and ecological quality in estuaries must be based on robust risk assessments. Retrospective risk assessments are performed when sites have potentially been impacted in the past. When they show a degradation of environmental quality, the restoration of degraded habitats and ecosystems must be addressed. Prospective or predictive risk assessments aim at assessing the future risks of anthropogenic pressure such as climate change or chemical releases into the environment.

Environmental Quality Standards and Guidelines

Establishing EQSs (concentration in water, sediment, or biota that must not be exceeded) is a major tool to protect the aquatic environment and human health. The procedures recommended for their determination[17] are similar to those described for PNECs in the TGD,[26] and very often, EQSs are identical to PNECs. In Europe, EQSs for priority substances and certain other pollutants in surface waters have been recently published.[27] However, for persistent, bioaccumulative, and toxic pollutants, it is not possible to ensure protection against indirect effects and secondary poisoning by EQS for surface water alone. It is therefore appropriate to establish EQS for sediment and biota. The PNECoral is the quality standard for biota tissue with respect to secondary poisoning of top predators as an objective of protection. According to the OSPAR/International Council for the Exploration of the Sea (ICES) experts,[55] the risk to marine predators is calculated as the ratio between the concentration in their food (marine fish) and the no-effect concentration for oral intake (PNECoral-predator) with the help of bioaccumulation models (Abar-nou in Ref. [14]). However, in estuaries, the presence of top predators is limited, and the major risk is for human consumers, which is a reason why oyster culture or fishing is forbidden in certain sites as a consequence of the overrun of quality standards for human health.

At this stage, no EQSs were set in sediments under the WFD partly because the significance of gross concentrations of potentially toxic chemicals in sediments is not easily established, and this is particularly difficult in estuaries where the mobility and bioavailability of contaminants is deeply affected by salinity changes.[13] In the United States, the Washington State Department of Ecology has published sediment management standards.[56] Bioavailability may be examined by using the methods described by Hansen et al.[57] for PAH mixtures and U.S. Environmental Protection Agency (EPA)[58] for metals. When EQSs are not available, sediment quality guidelines (SQGs) are commonly used by official organisms. such as the effects-range low (ERL) and the effects-range median (ERM) values that were derived for the National Oceanic and Atmospheric Administration (NOAA) with numerous modeling, laboratory, and field studies performed in marine and estuarine sediments.[59] Australian sediment quality guidelines were adapted from ERL/ERM.[60] The threshold effect level (TEL) and the probable effect level (PEL) initially described by MacDonald et al.[61] are also used by the NOAA.[62]

Mean SQG quotients can be calculated by dividing chemical concentrations in sediments by their respective SQGs and calculating the mean of quotients for individual chemicals. The resulting index provides a method of accounting for the presence and concentrations of multiple chemicals in sediments, relative to their effects-based guidelines.[56]

Sediment quality guidelines are also of great use in the regulation of dredging activities. This activity is particularly well regulated at the international level, and all these regulations share the same philosophy. For details, the reader can refer to the revised OSPAR guidelines for the management of dredged material.[63]

Environmental monitoring is indispensable to verify if EQSs are honored or if it is necessary to adopt decisions in order that the environment and human health may be adequately protected. Once again, the procedures are comparable in different areas in the world, at least in economically developed countries, and for instance, the Coordinated Environment Monitoring Programme (CEMP) provides a monitoring manual Web page that is regularly updated to take into account scientific and technical advances provided by the OSPAR commission in collaboration with the ICES (http://www.ospar.org/content/content.asp?menu=00900301400135_000000_000000).

How the Biota Will Respond to Restoration

From a large number of examples examined by Borja et al.,[64] including the recovery from a list of different stressors, it appears that only in a few cases, recovery can take less than 5 years, but more frequently, a minimum of 15–25 years is needed to attain the original biotic composition, and diversity may lag far beyond that period. Similarly, Hering et al.[65] conclude that it cannot be expected that European aquatic ecosystems will fully recover within 15 or even 30 years from over a century of degradation. Considering such situations, the common saying "an ounce of prevention is worth a pound of cure" makes sense.

Predictive Risk Assessment of New and Existing Chemicals

Predictive risk assessment aims at assessing the future risks from chemical releases into the environment. In the United States, the Federal Toxic Substances Control Act gives the EPA authority to regulate, and even ban, the manufacture, use, and distribution of both new and existing chemicals. In Europe, a significant improvement has occurred recently with a new chemical policy, REACH (Registration, Evaluation, and Authorization of Chemicals).[66] The pros and cons of REACH have been analyzed, with some details in Hansen[24] and Verdonck et al.[67]

Risk assessment under REACH will be carried out mainly by applying existing methodologies.[26] The risk quotient approach needs the determination of predicted environmental concentrations (PECs) and PNECs. Very simplistically, if the PEC/PNEC ratio is greater than 1, the substance is considered to be of concern, and risk reduction measures must be envisaged. According to Verdonck et al., [67] uncertainty is hidden and concealed in risk quotient numbers that appear to be certain and, therefore, create a false sense of certainty and protectiveness. These authors propose strategies to improve uncertainty analysis in risk assessment of chemicals. Management obstacles have been summarized by Hansen, [24] who considers that "there are few areas of applied science where the connection between science and policy is as difficult and complicated as in the area of risk management of chemicals." She underlines the lack of communication of uncertainties between assessors and decision makers. Even in the cases where there is an important set of data on a given chemical, there may be little consensus regarding what the actual risk is. Quoting Chapman,[68] Hansen[24] bewails that the effect of all this uncertainty is generally the continued production and use of possible harmful chemicals. Even with "old" contaminants very well studied, the complexity of field situations often leads to a lack of decision. This is well illustrated by Hameedi et al.:[62] despite substantive scientific evidence that points to copper as a contaminant of concern in the St. Lucie Estuary, it is debatable as to how much the observed levels of copper in the estuary are adversely affecting the flow of products (such as fish) or services (such as recreation) of the estuary. This approach poses the problem of what is environmental risk for different actors, as previously mentioned in the section "The Question of the Reference Site."

Conclusions

In aquatic environments, an important set of methods and techniques is available to provide decision makers with tools allowing the prevention, control, and remediation of pollution inflows and impacts. However, because estuaries are highly dynamic and complex systems, the application of these methods is particularly complicated, generating a high degree of uncertainty, whatever the category of assessment tools considered among the classical triad of analysis (chemistry, bioassays, community structure) as well as other assessment strategies such as biomarkers. Many authors have underlined that only a few toxicity data are available for estuarine species compared with marine and, moreover, freshwater species. Thus, toxic impacts on estuarine biota are often derived from toxicological parameters determined in non-estuarine species.

Among tolerant organisms, some are keystone species, with important roles in ecosystem functioning, on which numerous species will depend. Depending on the harshness of natural conditions in different ecosystems, the number of species able to fulfill the same functional role is strikingly variable. As estuarine species are much less numerous than freshwater or marine species, it is evident that the extinction of even a small number of species is sufficient to affect ecosystem functioning.

The state of knowledge would allow considerable improvement of risk assessment and monitoring. The integration of scientific advances in regulations and management needs decade(s). Some of the reasons for that are the higher technicity and the higher cost of certain new tools. Is the society ready to pay for better conservation of the environment? Furthermore, for what kind of conservation: recovery of the Garden of Eden or sustainability of estuaries for the benefits of nature to human beings?

References

1. MEA (Millennium Ecosystem Assessment). *Ecosystems and Human Well-Being: Synthesis*. Island Press, World Resources Institute: Washington, DC, 2005.
2. Scott, G.; Klaine, S.J. Marine and estuarine toxicology and chemistry. Environ. Toxicol. Chem. **2001**, *20* (1), 1–2.
3. Bianchi, T.S. *Biogeochemistry of Estuaries*. Oxford University Press: Oxford, U.K., 2007.
4. Uncles, R.J.; Joint, I.; Stephens, J.A. Transport and retention of suspended particulate matter and bacteria in the Humber–Ouse estuary, United Kingdom, and their relationship to hypoxia and anoxia. Estuaries **1998**, *21* (4A), 597–612.
5. Berthe, T.; Touron, A.; Leloup, J.; Deloffre, J.; Petit, F. Faecal-indicator bacteria and sedimentary processes in estuarine mudflats (Seine-France). Mar. Pollut. Bull. **2008**, *57* (1–5), 59–67.
6. Eisler, R. Eisler's Encyclopedia of Environmentally Hazardous Priority Chemicals; Elsevier: Amsterdam, The Netherlands, 2007.
7. Newman, M.C.; Morris, H.R., Jr.; Hale R.C. *Coastal and Estuarine Risk Assessment*; Lewis Publishers: Boca Raton, FL, USA, 2002.
8. Klaine, S. J.; Alvarez, P. J. J.; Batley, G. E.; Fernandes, T. F.; Handy, R. D.; Lyon, D. Y.; Mahendra, S.; McLaughlin, M. J.; Lead, J. R. Nanomaterials in the environment: Behavior, fate, bioavailability, and effects. Environ. Toxicol. Chem. **2008**, *27* (9), 1825–1851.
9. Hummel, H.; Bogaards, R.; Bek, T.; Polishchuk, L.; Amiard-Triquet, C.; Bachelet, G.; Desprez, M.; Strelkov, P.; Sukhotin, A.; Naumov, A.; Dahle, S.; Denisenko, S.; Gantsevich, M.; Sokolov, K.; De Wolf, L. Sensitivity to stress in the bivalve *Macoma balthica* from the most northern (Arctic) to the most southern (French) populations: Low sensitivity in Arctic populations because of genetic adaptations? Hy0drobiologia **1997**, *355* (1–3), 127–138.
10. Heugens, E.H.W.; Hendriks, A.J.; Dekker, T.; van Straalen, N.; Admiraal, W. A review of the effects of multiple stressors on aquatic organisms and analysis of uncertainty factors for use in risk assessment. Crit. Rev. Toxicol. **2001**, *31* (3), 247–284.
11. Amiard-Triquet, C.; Rainbow, P.S.; Romeo, M., Eds. 2011. *Tolerance to Environmental Contaminants*; CRC Press: Boca Raton, FL, 2011.
12. Chapman, P.M. The Sediment Quality Triad approach to determining pollution-induced degradation. Sci. Total Environ. **1990**, *97–98*, 815–825.
13. Chapman, P.M.; Wang, F. Assessing sediment contamination in estuaries. Annual review. Environ. Toxicol. Chem. **2001**, *20* (1), 3–22.

14. Amiard-Triquet, C.; Rainbow, P.S., Eds. *Environmental Assessment of Estuarine Ecosystems. A Case Study;* CRC Press: Boca Raton, FL, 2009.
15. Dauvin, J.C. Paradox of estuarine quality: Benthic indicators and indices in estuarine environments, consensus or debate for the future. Mar. Pollut. Bull. **2007**, *55* (1–6), 271–281.
16. Elliott, M.; Quintino, V. The estuarine quality paradox, environmental homeostasis and the difficulty of detecting anthropogenic stress in naturally stressed areas. Mar. Pollut. Bull. **2007**, *54* (6), 640–645.
17. WFD (2000). Directive 2000/60/EC of the European Parliament and Council of 23 October 2000 establishing a framework for policy in the field of water (JO L 327 of 22 December 2000); European Commission: Brussels.
18. Allan, I.J.; Vrana, B.; Greenwood, R.; Mills, G.A., Roig, B.; Gonzalez, C. A "toolbox" for biological and chemical monitoring requirements for the European Union's Water Framework Directive. Talanta **2006**, *69* (2), 302–322.
19. Hagger, J.A.; Jones, M.B.; Lowe, D.; Leonard, P.; Owen, R.; Galloway, T.S. Application of biomarkers for improving risk assessments of chemicals under the Water Framework Directive: A case study. Mar. Pollut. Bull. **2008**, *56* (6), 1111–1118.
20. Knoery, J.; Claisse, D. Insights from 30 years of experience in running the French marine chemical monitoring network. ICES-CM 2010 / F08.
21. NAS. *The International Mussel Watch;* National Academy of Sciences: Washington, DC, 1980.
22. Mills, G.A.; Greenwood, R.; Allan, I.J.; Lopuchin, E.; Brümmer, J.; Knutsson, J.; Vrana B. Application of passive sampling techniques for monitoring the aquatic environment. In *Analytical Measurements in Aquatic Environments;* Namiesnik, J., Szefer, P., Eds.; CRC Press: Boca Raton, FL, USA, 2010; 41–68.
23. Nendza, M. Inventory of marine biotest methods for the evaluation of dredged material and sediments. Chemo-sphere **2002**, *48* (8), 865–883.
24. Hansen, M.E. Risk Assessment—The Role of Combined Environmental and Toxic Stress. MS Thesis, Roskilde University, 2006.
25. ECETOC. *Risk Assessments in Marine Environments,* Technical report no. 82; European Centre for Ecotoxicology and Toxicology of Chemicals (ECETOC): Brussels, 2001.
26. TGD. Technical Guidance Document on Risk Assessment in support of Commission Directive 93/67/EEC on Risk Assessment for new notified substances, Commission Regulation (EC) n° 1488/94 on Risk Assessment for existing substances and Directive 98/8/EC of the European Parliament and of the Council concerning the placing of biocidal products on the market; European Commission. Joint Research Centre. EUR 20418 EN/2, 2003.
27. EC 2008. Directive 2008/105/EC of the European Parliament and Council of 16 December 2008 on environmental quality standards in the field of water policy, amending and subsequently repealing Council Directives 82/176/EEC, 83/513/EEC, 84/156/EEC, 84/491/EEC, 86/280/EEC and amending Directive 2000/60/EC of the European Parliament and of the Council (JO L 348 of 24 December 2008); European Commission: Brussels.
28. Bocquené, G. Risque environnemental d'origine chimique dans l'estuaire de Seine, 2011. Available at http://seine-aval.crihan.fr/web/attached_file/componentId/kmelia324/attachmentId/26836/lang/fr/name/_Contamination%20de%20l%27estuaire_n5_reduit.pdf
29. De Paz, L.; Patricio, J.; Marques, J.C.; Borja, A.; Laborda, A.J. Ecological status assessment in the lower Ebro estuary (Spain). The challenge of habitat heterogeneity integration: A benthic perspective. Mar. Pollut. Bull. **2008**, *56* (7), 1275–1283.
30. Dauvin, J.C.; Bachelet, G.; Barillé, A.L.; Blanchet, H.; de Montaudouin, X.; Lavesque, N.; Ruellet, T. Benthic indicators and index approaches in the three main estuaries along the French Atlantic coast (Seine, Loire and Gironde). Mar. Ecol. **2009**, *30*, 228–240.

31. Blanck, H.; Wängberg, S.A.; Molander, S. 1988. Pollution-induced community tolerance—A new tool. In *Function Testing of Aquatic Biota for Estimating Hazards of Chemicals;* Cairns, J.J., Pratt, J.R., Eds.; American Society for Testing and Materials: Philadelphia, PA, USA, 1988; 219–230.
32. Courrat, A.; Foussard, V.; Lepage, M. Les indicateurs DCE estuariens. Etat des lieux à l'échelle européenne en décembre 2010, http://seine-aval.crihan.fr/web/attached_file/componentId/kmelia324/attachmentId/24968/lang/fr/name/axe1_rapport_synthese_indicateur_mario.pdf.
33. Lloyd, S.S. Zooplankton Ecology in the Chesapeake Bay Estuarine Turbidity Maximum, with Emphasis on the Calanoid Copepod Eurytemora affinis. PhD thesis, University of Maryland, 2006.
34. Scholle, J.; Dau, K. Reference Conditions of Biological Quality Components in Accordance with the EU Water Framework Directive in Coastal and Transitional Waters in Nl, D and DK. Status of Development and Comparison with the Targets of the Trilateral Wadden Sea Plan, Harbasins report; 2007; 65 pp., http://www.harbasins.org/fileadmin/inhoud/pdf/Final_Products/WP1/1.2/WFD_Implementation_Status_in_NL-D-DK_Nov.2007.pdf.
35. Underwood, G.J.C.; Kromkamp, J. Primary production by phytoplankton and microphytobenthos in estuaries. Adv. Ecol. Res. **1999**, *29,* 93–153.
36. Lerouxel, A.; Blandin, E.; Rosa, P.; Launeau, P.; Rincé, Y.; Barillé, L. Cartographie du microphytobenthos de l'estuaire de la Loire par télédétection infra-rouge; Project BEEST, Rapport; 2010, http://seine-aval.crihan.fr/web/attached_file/componentId/kmelia324/attachmentId/24963/lang/fr/name/axe%203_rapport_microphytobenthos.pdf.
37. van der Wal, D.; Wielemaker-van den Dool, A.; Herman P.M.J. Spatial synchrony in intertidal benthic algal biomass in temperate coastal and estuarine ecosystems. Ecosystems **2010**, *13* (2), 338–351.
38. Souissi, S.; Devreker, D. Le zooplankton peut-il être utilisé comme indicateur de la qualité des eaux estuariennes?; Rapport Projet BEEST, 2010, http://seine-aval.crihan.fr/web/attached_file/componentId/kmelia324/attachmentId/24978/lang/fr/name/axe3_rapport_zooplancton.pdf.
39. Dauvin, J.C.; Bellan, G.; Bellan-Santini, D. Benthic indicators: From subjectivity to objectivity—Where is the line? Mar. Pollut. Bull. **2010**, *60* (7), 947–953.
40. Borja, A.; Barbone, E.; Basset, A.; Borgensen, G.; Brkljacic, M.; Elliott, M.; Garmendia, J.M.; Marques, J.C.; Mazik, K.; Muxika, I.; Neto, J.M.; Norling, K.; Rodriguez, J.G.; Rosati, I.; Rygg, B.; Teixeira, H.; Trayanova, A. Response of single benthic metrics and multimetric methods to anthropogenic pressure gradients, in five distinct European coastal and transitional ecosystems. Mar. Pollut. Bull. **2011**, *62* (3), 499–513.
41. Perez-Domínguez, R.; Maci, S.; Courrat, A.; Borja, A.; Neto, J.; Elliott, M. *Review of Fish-Based Indices to Assess Ecological Quality Condition in Transitional Waters*, WISER report; 2010; 31 pp., http://www.wiser.eu/down-load/D4.4-1.pdf.
42. Uriarte, A.; Borja, A. Assessing fish quality status in transitional waters, within the European Water Framework Directive: Setting boundary classes and responding to anthropogenic pressures. Estuar. Coast. Shelf Sci. **2009**, *82* (2), 214–224.
43. Coates, S.; Waugh, A.; Anwar, A.; Robson, M. Efficacy of a multi-metric fish index as an analysis tool for the transitional fish component of the Water Framework Directive. Mar. Pollut. Bull. **2007**, *55* (1–6), 225–240.
44. Breine, J.; Quataert, P.; Stevens, M.; Ollevier, F.; Volckaert, F.A.M.; Van den Bergh, E.; Maes, J. A zone-specific fish-based biotic index as a management tool for the Zee-schelde estuary (Belgium). Mar. Pollut. Bull. **2010**, *60* (7), 1099–1112.
45. Courrat, A.; Lobry, J.; Nicolas, D.; Laffargue, P.; Amara, R.; Lepage, M.; Girardin, M.; Le Pape, O. Anthropogenic disturbance on nursery function of estuarine areas for marine species. Estuar. Coast. Shelf Sci. **2009**, *81* (2), 179–190.

46. Delpech, C.; Courrat, A.; Pasquaud, S.; Lobry, J.; Le Pape, O.; Nicolas, D.; Boët, P.; Girardin, M.; Lepage, M. Development of a fish-based index to assess the ecological quality of transitional waters: The case of French estuaries. Mar. Pollut. Bull. **2010**, *60* (6), 908–918.
47. Lévêque, C.; Mounolou, J.C. *Biodiversity.* Wiley: Chichester, 2004.
48. Hillebrand, H.; Matthiessen, B. Biodiversity in a complex world: Consolidation and progress in functional biodiversity research. Ecol. Lett. **2009**, *12* (12), 1405–1419.
49. Lévêque, C. Synthèse du projet BEEST Vers une approche multicritère du Bon Etat des grands ESTuaires atlantiques; http://seine-aval.crihan.fr/web/pages.jsp?currentNodeId=157, 2011.
50. Depledge, M.H. The rational basis for the use of biomarkers as ecotoxicological tools. In *Non Destructive Biomarkers in Invertebrates;* Fossi, M.C., Leonzio, C., Eds.; Lewis Publishers: Boca Raton, 1994; 261–285.
51. Kalman, J.; Buffet, P.E.; Amiard, J.C.; Denis, F.; Mouneyrac, C.; Amiard-Triquet, C. Assessment of the influence of confounding factors (weight, salinity) on the response of biomarkers in the estuarine polychaete *Nereis diversicolor*. Biomarkers **2010**, *15* (5), 461–469.
52. Fossi Tankoua, O.; Buffet, P.E.; Amiard, J.C.; Amiard-Triquet, C.; Mouneyrac, C.; Berthet; B. Potential influence of confounding factors (size, salinity) on biomarkers in the sentinel species *Scrobicularia plana* used in programmes monitoring estuarine quality. Environ. Sci. Pollut. Res. **2011**, *18,* 1253–1263, doi: 10.1007/s11356-011-0479-3.
53. Mouneyrac, C.; Linot, S.; Amiard, J.C.; Amiard-Triquet, C.; Métais, I.; Durou, C.; Minier, C.; Pellerin, J. Biological indices, energy reserves, steroid hormones and sexual maturity in the infaunal bivalve *Scrobicularia plana* from three sites differing by their level of contamination. Gen. Comp. Endocrinol. **2008**, *157* (2), 133–141.
54. Mouneyrac, C., Amiard-Triquet, C. Biomarkers of ecological relevance. In *Encyclopedia of Aquatic Ecotoxicology*; Blaise, C., Férard, J.F., Eds.; Springer: Berlin, 2012 (in press).
55. OSPAR Commission. Revised OSPAR guidelines for the management of dredged material. Reference 2004-08. Convention for the Protection of the Marine Environment of the North-East Atlantic; 2004; 30 pp., http://www.dredging.org/documents/ceda/downloads/environ-ospar-revised-dredged-material-guidelines.pdf.
56. Long., E.R.; Ingersoll, C.G.; MacDonald, D.D. Calculation and uses of mean sediment quality guideline quotients: A critical review. Environ. Sci. Technol. **2006**, *40* (6), 1726–1736.
57. Hansen, D.J.; Di Toro, D.M.; McGrath, J.A.; Swartz, R.C.; Mount, D.R.; Spehar, R.L.; Burgess, R.M.; Ozretich, R.J.; Bell, H.E.; Reiley, M.C.; Linton, T.K. *Procedures for the Derivation of Equilibrium Partitioning Sediment Benchmarks (ESBs) for the Protection of Benthic Organisms: PAH Mixtures,* EPA/600/R-02/013; Office of Research and Development, U.S. Environmental Protection Agency: Nar-ragansett, RI, 2003.
58. USEPA. Procedures for the Derivation of Equilibrium Partitioning Sediment Benchmarks (ESBS) for the Protection of Benthic Organisms: Metal Mixtures (Cadmium, Copper, Lead, Nickel, Silver and Zinc), EPA/600/R-0211. US Environmental Protection Agency: Washington, DC, 2005.
59. Long, E.R.; MacDonald, D.D.; Smith S.L.; Calder, F.D. Incidence of adverse biological effects within ranges of chemical concentrations in marine and estuarine sediments. Environ. Manage. **1995**, *19* (1), 81–97.
60. McCready, S.; Birch, G.F.; Long, E.R.; Spyrakis, G.; Greely, C.R. An evaluation of Australian sediment quality guidelines. Arch. Environ. Contam. Toxicol. **2006**, *50* (3), 306–315.
61. MacDonald, D.D.; Carr, R.S.; Calder, F.D.; Long, E.R.; Ingersoll, C.G. Development and evaluation of sediment quality guidelines for Florida coastal waters. Ecotoxicology **1996**, *5* (4), 253–278.
62. Hameedi, M.J.; Johnson, W.E.; Kimbrough, K.L.; Browder, J.A. Sediment contamination, toxicity and infaunal community composition in St. Lucie estuary, Florida based upon measures on the Sediment Quality Triad, 2006. Available at http://www.ccma.nos.noaa.gov/publications/sle_report.pdf.

63. OSPAR/ICES 2004. Workshop on the evaluation and update of background reference concentrations (B/RCs) and ecotoxicological assessment criteria (EACs) and how these assessment tools should be used in assessing contaminants in water, sediment and biota, The Hague, Feb 9–13 2004; Final rep. Moffat, C., Pijnenburg, J., Traas, T., Eds. http://www.ospar.org/documents/dbase/publications/p00214_BRC%20EAC%20Workshop.pdf.
64. Borja, A.; Dauer, D.M.; Elliott, M.; Simenstad, A. Medium-and long-term recovery of estuarine and coastal ecosystems: Patterns, rates and restoration effectiveness. Estuar. Coasts **2010**, *33* (6), 1250–1260.
65. Hering, D.; Borja, A.; Carstensen, J.; Carvalho, L.; Elliott, M.; Feld, C.K.; Heiskanen, A. S.; Johnson, R.K.; Moe, J.; Pont, D.; Solheim, A.L.; van de Bund, W. The European Water Framework Directive at the age of 10: A critical review of the achievements with recommendations for the future. Sci. Total Environ. **2010**, *408* (19), 4007–4019.
66. CEC Commission of the European Communities. Proposal for a regulation of the European Parliament and of the Council concerning the registration, evaluation, authorisation and restrictions of chemicals (REACH) establishing a European Chemical Agency and amending directive 1999/45/EC and regulation (EC) (on persistent organic pollutants). COM 644; 2003.
67. Verdonck, F.A.M.; Souren, A.; Van Asselt, M.B.A.; Van Sprang, P.A.; Vanrolleghem, P.A. Improving uncertainty analysis in European Union risk assessment of chemicals. Integr. Environ. Assess. Manage. **2007**, *3* (3), 333–343.
68. Chapman, A. Regulating chemicals—From risks to riskiness. Risk Anal. **2006**, *26* (3), 603–616.

63

Everglades

Kenneth L. Campbell, Rafael Munoz-Carpena, and Gregory Kiker

Introduction

The Everglades of south Florida was originally a broad, shallow "River of Grass"[1] that extended from the south shore of Lake Okeechobee to Florida Bay at the southern tip of the state, east to the Coastal Ridge, and west to the Immokalee Ridge. Historically, the area was a vast sawgrass marsh, dotted with tree islands and interspersed with wet prairies and sloughs covering an area about 40 mi wide by 100 mi long. One of the unique regions of the world, it has steadily decreased in size and declined in health during the past century. Half its wetland area has been lost to agriculture and urban development and the remaining segments are impacted by lack of a clean, dependable water supply. Natural water flows have been diverted for irrigation, drinking water, and flood protection. The conveyance system of canals, levees, structures, and pumps developed for flood control has altered natural patterns of water flow and storage, adversely affecting food webs that supported a diverse ecosystem. Nutrient runoff from urban and agricultural sources is transported by the conveyance system to the remaining natural wetland areas, causing undesirable changes in flora and fauna. Hydroperiod changes have altered natural fire patterns and stimulated invasion of exotic species. A multi-agency state and federal task force has developed a Comprehensive Everglades Restoration Plan (CERP)[2] to address and reverse these major changes to this unique wetland ecosystem. The major hydrologic modifications to be addressed in the Everglades restoration include: 1) regain lost storage capacity; 2) restore more natural hydropatterns; 3) improve timing and quantities of fresh water deliveries to estuaries; and 4) restore water quality conditions. The Comprehensive Plan, considered the world's largest such project, includes more than 60 components proposed for implementation over a period of four decades with an estimated investment approaching $8 billion. State and federal legislation provides for a 50/50 cost share between the federal and state governments to implement the plan.

Everglades Water Management—Past, Present, and Future

History

Primitive canals were dug in portions of the Everglades as early as the late 1800s in attempts to reclaim fertile swampland for agriculture.[3] Early promoters and developers led people to believe that a productive subtropical agriculture was possible in the entire Everglades region. These early attempts at land reclamation were largely unsuccessful until the 1920s when a period of less than normal rainfall helped dry the region around Lake Okeechobee for farming. Following severe hurricane damage in the region in the late 1920s and again in 1947, the focus was shifted from land reclamation to flood protection and the Central and Southern Florida Flood Control Project was authorized and implemented beginning in 1948. Over the next 15 years, this project resulted in a perimeter dike around Lake Okeechobee and the extensive conveyance system of canals, levees, structures, and pumps currently in place. It also allowed development of the Everglades Agricultural Area (EAA), a highly productive, 700,000-acre region of organic soils in the northern Everglades used primarily for sugar cane and winter vegetable production.[4]

Environmental Issues

By the mid-1960s, concerns were already growing about conservation issues and adverse environmental impacts. Additional areas along the eastern border of the Everglades have since experienced urban encroachment. A total of about 1 million acres, roughly 50% of the Everglades wetlands, have been transformed for human uses during the past half-century. The 1700 mi of canals and levees in the region have interrupted connections between the central Everglades and the adjacent wetlands, resulting in over-drainage in some areas and excessive flooding in others. This system provides water supply, flood protection, water management, and other benefits to south Florida, but it must be modified to reduce the negative impacts on the environment. The current canal system works very effectively, discharging an average of 1.7 billion gal of water per day to the ocean and gulf. This discharge must be reduced if future urban, agricultural, and environmental demands for water are to be met.

Today's remaining Everglades have been significantly affected by the current water management system. Wading birds and other wildlife populations are greatly decreased. Tree islands, with their unique combination of wetland and terrestrial vegetation and wildlife, are considered to be an excellent indicator of the overall health of the Everglades. Many of these tree islands have disappeared from the northern Everglades over the past 50 years, and many others have been taken over by exotic vegetation. These effects are mainly due to changes in the quantity, quality, timing, and distribution of water that have occurred over the years as a result of changed water management. Water depth, duration, and timing are important to both wildlife and vegetation. The sawgrass wetlands of the Everglades developed under very low nutrient conditions with rainfall as the main source of phosphorus. Nutrient inflows, especially phosphorus, as a result of development and modified water management have influenced changes in vegetation type.[5] Where phosphorus concentrations have increased, sawgrass and spike rush have been replaced by cattail causing undesirable changes in the ecosystem. Native vegetation remains healthy where phosphorus concentrations are low.

Restoration

Restoration of the remaining Everglades depends upon a knowledge and understanding of the original conditions. Efforts are focusing on improving upstream water quality and the distribution, timing, depth, and flow of surface water into and through the Everglades. Early historical information sources, combined with further interpretation and analysis, are being used to estimate original drainage patterns and soil, topographic and vegetation conditions before canal drainage began in the late 1800s. Results of these studies indicate that the predrainage landscape of the Everglades probably was configured in subtle ridges and sloughs with two major flow pathways: a flow path southeastward to the Atlantic Ocean, and a southwestward flow path along Shark Slough to the Gulf of Mexico.[6] These flow patterns may have influenced the ridge and slough landscape configuration that is important to the health of the ecosystem. Redevelopment of these flow patterns and landscape configuration will be important to the restoration process. About 70% less water flows through the Everglades today compared to the historic Everglades system.

The main goal of Everglades restoration is to deliver the correct amount of water, with the correct quality, to the correct locations, and at the correct time.[7] Most of the water currently lost to the ocean or gulf will be stored in surface and subsurface storage areas until needed, when 80% of it will be allocated to the environment and 20% to increase urban and agricultural water supplies. Water to be stored for future use will be routed through surface storage reservoirs and wetland-based stormwater treatment areas to improve its quality. Additional water quality improvements can be expected from comprehensive integrated water quality planning efforts currently in progress. To restore water flow paths, more than 240 mi of canals and levees will be removed in the Everglades. This will allow more natural overland water flow in the remaining natural areas of the Everglades. Water held and released will be managed to match natural discharge patterns more closely. Operational plans will be developed in some areas to simulate natural rainfall patterns with water releases to improve the timing of water flowing through the Everglades ecosystem. These strategies are all being designed to enhance not only ecosystem restoration, but also urban and agricultural water supply and flood protection as part of the process of moving toward a more sustainable south Florida.

Conclusion

The Everglades landscape is a unique combination of subtropical wetlands and uplands, including sawgrass marshes, sloughs, wet prairies, tree islands, tropical hardwood hammocks, pinelands, and mangroves. It provides important habitat for many threatened and endangered species. Water management for flood control and water supply purposes has caused some areas to become drier and others to become wetter than normal. More than half of the original wetland area has been lost to agricultural and urban development. The introduction of increased nutrients resulting from this development has caused undesirable shifts in vegetation communities. Hydrologic changes have altered the extent of naturally occurring fires and promoted the growth of exotic species. While the current water management system performs well for flood protection it must be modified to reduce adverse environmental impacts and conserve more fresh water to meet a variety of needs. A Comprehensive Everglades Restoration Plan received initial authorization in 2000 to begin the restoration of the south Florida ecosystem and provide for water-related needs of the region. This plan addresses the quantity, quality, distribution, and timing of water to the Everglades. A large amount of additional information regarding the Everglades is available on the web at http://www.sfwmd.gov/koe_section/2_everglades.html and http://www.evergladesplan.org/.

The following quote from the Comprehensive Everglades Restoration Plan web site[7] conveys the importance of the Everglades and the current restoration program. The significance of the remaining Everglades to the nation and the world has been affirmed time and again. Congress established Everglades National Park. The Everglades have also been designated an International Biosphere Reserve, a World Heritage Site, and a Wetland of International Significance. Identified as one of the world's major ecosystem types, the Everglades are home to 68 threatened or endangered plant and animal species. The benefits and functions of these plants and animals may never be known if we do not restore and protect their habitat. Saving the Everglades requires us to save the entire south Florida ecosystem. The ecological and cultural significance of the Everglades is equal to the Grand Canyon, the Rocky Mountains, or the Mississippi River. As responsible stewards of our natural and cultural resources, we cannot sit idly by and watch any of these disappear. The Everglades deserves the same recognition and support.

References

1. Douglas, M.S. *The Everglades: River of Grass;* Hurricane House, Coconut Grove, FL, 1947.
2. U.S. Army Corps of Engineers, Jacksonville District. *Central and Southern Florida Project Comprehensive Review Study Final Integrated Feasibility Report and Programmatic Environmental Impact Statement,* April 1999; The Corps: Jacksonville, FL, South Florida Water Management District: West Palm Beach, FL, 1999; 10 Vols.; http://www.evergladesplan.org/pub/restudy_eis.shtml (accessed March 2002).

3. Blake, N.M. *Land into Water—Water into Land: A History of Water Management in Florida*; University Presses of Florida: Tallahassee, FL, 1980.
4. Bottcher, A.B.; Izuno, F.T. *Everglades Agricultural Area (EAA): Water, Soil, Crop, and Environmental Management*; University Press of Florida: Gainesville, FL, 1994.
5. Reddy, K.R., O'Connor, G.A., Schelske, C.L., Eds.; *Phosphorus Biogeochemistry in Subtropical Ecosystems*; Lewis Publishers: Boca Raton, FL, 1999.
6. South Florida Water Management District. *Watershed Management: Everglades/Florida Bay;* http://www.sfwmd.gov/org/wrp/wrp_evg/ (accessed March 2002).
7. U.S. Army Corps of Engineers, Jacksonville District, South Florida Water Management District. *Rescuing an Endangered Ecosystem—the Journey to Restore America's Everglades;* http://www.evergladesplan.org/ (accessed March 2002).

64

Water Quality: Range and Pasture Land

Thomas L. Thurow

Introduction

Livestock and clean water are two products that can be simultaneously obtained from range and pasture lands. This requires that ecological and hydrological principles be applied when crafting a grazing management strategy that is compatible with *predetermined* water quality goals. Making protection of water quality the starting point of land use planning is a philosophical foundation of 1972 U.S. Clean Water Act and subsequent amendments. This goal is operationalized by management agencies establishing total maximum daily load (TMDL) standards for waterways. A TMDL is a calculation of the maximum amount of a pollutant from all contributing point sources (a specific location such as a confined animal feedlot operation [CAFO]) and non-point sources (pollution that occurs over a wide area such as may originate from grazing).

Physical Characteristics

Suspended Sediment

Suspended sediment is the most pervasive non-point source pollutant from grazing lands. All waterways naturally contain some suspended sediment attributable to the geologic (natural) erosion influenced by stream type (primarily determined by the geology, topography, and location within the watershed) and ecological factors (e.g., climate, vegetation, soil). Therefore, formulation of TMDL suspended sediment standards must be catchment specific so that geologic erosion can be differentiated from accelerated erosion associated with human activities such as grazing management.

Grazing management can effect the erosion rate of a site primarily by influencing the degree to which livestock impact the soil and vegetation.

Livestock Impacts on Soil

Soil structure is the arrangement of soil particles and intervening pore spaces. The size of soil particles (aggregation) and their stability when wetted determines the porosity of the soil, which governs the rate at which water will enter the soil (infiltration). If the rainfall rate is greater than the infiltration rate, water will run off the site, carrying sediment with it.

Livestock trampling compacts the soil, increasing the bulk density (i.e., the pore volume is reduced resulting in decreased infiltration rate). The degree of damage associated with trampling at a particular site depends on soil type, soil water content, seasonal climatic conditions, and the intensity of livestock use.[1] Compacted trails form on sites where livestock traffic is concentrated. The density of trails tends to increase as the number of pastures is increased within an intensive rotation grazing system. Another common reason for trail formation is repeated movement to and from limited sources of water, mineral supplements, or shelter. The low infiltration rate of trails results in concentrated runoff, which may eventually create gullies. Roads across hilly range and pasture lands are also a serious erosion source, especially since they are often poorly designed and maintained.[2]

Another way livestock trampling causes surficial problems is by churning dry soil to dust. This is very detrimental to infiltration because the disaggregated soil particles are carried by water and lodge in the remaining soil pores making them smaller or sealing them completely. This “washed in” layer where clay particles clog soil pores is a common way that soil crusts are formed. Soil crusts can reduce infiltration by 90%, thereby dramatically increasing runoff and sediment transport.[3] Trampling a crusted soil does break the crust and incorporates mulch and seeds into the soil. However, this benefit is short lived because the subsequent impact of falling raindrops re-seals the soil surface after several minutes. To effectively address a soil- crusting problem, livestock grazing systems must concentrate on addressing poor aggregate stability, which is the cause of crusting. This requires protecting the soil surface from direct raindrop impact through maintaining vegetation cover and facilitating organic matter buildup in the soil via litter deposition.

Livestock Impacts on Vegetation

Direct raindrop impact on soil represents the greatest potential erosive force on grazing land; therefore it is very important that raindrop energy be dissipated by striking some form of cover before reaching the soil.[4] The amount of cover is positively associated with vegetation litter deposition. Litter slows overland flow, resulting in reduced ability to transport sediment. Litter also aids formation of stable aggregates (associated with high infiltration and low erosion rates) by binding soil particles together with adhesive byproducts produced by decaying litter and microbial synthesis.[5]

Grazing impacts on the vegetation community may be manifest by physical removal of standing vegetation through herbivory or through a gradual change in the composition of vegetation. As grazing pressure increases, the amount of cover and the amount of organic matter returned to the soil is reduced, resulting in an increased likelihood of runoff and erosion. Cover and infiltration rate tends to be greatest under trees and shrubs, followed in decreasing order by bunchgrass, shortgrass, and bare ground.[6] There is little impact on species composition with moderate or light grazing but composition change is great in response to heavy grazing, regardless of grazing strategy.[7] Often the change in species composition associated with heavy grazing is toward dominance by annuals or shortgrass species that have more runoff and erosion associated with them.[8] By the time erosion becomes obvious it may be too late to implement economically viable conservation options. Early recognition of a developing degradation pattern requires knowledge of range ecology, for the first signs of an impending erosion problem almost invariably are manifest by changes in plant density, composition, and vigor.[9]

Chemical Characteristics

Dissolved Chemicals

Nutrient loss from grazing lands via leaching or runoff is normally negligible, i.e., less than the input of nutrients from rainfall.[10] Most of the dissolved chemical constituents in runoff are contributed from the soil. Nutrients and organic matter adsorbed to the soil particles are also lost via erosion. Therefore, the most important role of a grazing system in nutrient loss is manifest through land use activities that alter the volumes or timing of runoff and erosion.[11]

Most of the nitrogen in urine is lost via volatilization, and most of the nitrogen in feces is sequestered by microorganisms or eventually transferred to soil organic matter. Nitrate is very mobile during heavy rain periods but loss by leaching is probably insignificant on most grasslands.[12] Feces contain almost all of the phosphorus excreted by livestock. Phosphorus is very resistant to leaching as it is rapidly precipitated or absorbed by other soil minerals. Nitrogen or phosphorus contamination of waterways is only of imminent concern when livestock are allowed to congregate near waterways.[13] Because of this concern, the U.S. Environmental Protection Agency interpretation of the Clean Water Act has deemed location of feedlots near waterways an unacceptable practice.

Dissolved Oxygen

Dissolved oxygen decreases when organic matter, such as animal manure, is added to water. This decrease occurs because biological decomposition processes consume available oxygen, as does oxidation of other reduced compounds such as ammonium. Excessive additions to surface water of nutrients such as nitrogen or phosphorus lead to eutrophication, often expressed by enhanced growth of aquatic plants and reduced water transparency (especially due to increases in algae). As the aquatic plants decay the microbes consume oxygen, lowering the concentration of oxygen available needed to support higher forms of aquatic life such as macroinvertrebrates and fish.

Biological Characteristics

The primary types of pathogens associated with livestock and wildlife feces are bacteria (e.g., *Campylobacter jejuni, Escherichia coli, Leptospira interrogans, Salmonella* spp.) and water-borne protozoa (e.g., *Cryptosporidia parvum, Giardia duodenalis*). These infectious pathogens can pose potential health risks to human drinking water supplies. Environmental fluctuation in temperature and soil moisture of grazing land creates a harsh environment for bacteria and the oocycsts of protozoa. Fecal coliforms can survive for several months in soil but can survive for up to a year within feces.[14] There is a rapid mortality of most oocysts when feces are deposited on land,[15] however, viable oocysts can be transported overland, especially when fresh feces are washed by an intense storm.[16] Once pathogens reach a water body, the threat of contamination may last from days to months,[17] with freshwater sediments being the site of greatest concentration and survival.[18]

Few detailed studies have explicitly studied the link between livestock grazing and water-borne pathogens. Much of the research has relied upon indicator coliforms that are more easily cultured but have been shown to be poorly correlated with some types of pathogenic bacteria.[14] Furthermore, many wildlife species harbor the same pathogens that livestock do, thus the natural occurrence of pathogens must be considered when analyzing water quality and making the relationship to livestock use of an area. The greatest threat of pathogen contamination of waterbodies occurs when livestock are allowed to concentrate along streams.[19] In situations where risk of bacteriological contamination is unacceptable, it is necessary to restrict livestock access to streams or riparian areas. Livestock use of these sensitive sites can be significantly reduced through development of water supply away from streams.[20]

Conclusions

Two broad objectives must be achieved to protect water quality associated with range and pasture grazing.

Limit Runoff and Erosion

Suspended sediment is the most common pollutant associated with grazing. Best management practices (BMPs) to limit runoff and erosion rely on maintenance of soil structure. Vegetation provides the organic matter necessary to enhance formation of stable aggregates and provides the cover to dissipate the erosive force of direct raindrop impact. Appropriate range and pasture grazing systems are designed to maintain vegetation cover and composition by adjusting intensity, frequency, and season of use. Flexibility needs to be built into grazing systems to adjust for unexpected fluctuation in the climate or market prices. The underdevelopment of climate and market risk management planning and policy regarding grazing plans is perhaps the most formidable threat to progress in improving water quality since these variables continue to be used as an excuse for water quality deterioration and/or the lack of progress in improving it.[21]

Limit Direct Livestock Use of Waterways and Sensitive Riparian Areas

Contamination of waterways by nutrients and pathogens is a predominant concern only on sites that allow livestock to congregate near water. On sites with limited water distribution, livestock tend to stay in the vicinity of water so long as forage is available. This increases the likelihood of excrement being deposited directly into the waterway. It also causes deterioration of the soil structure and plant community near the waterway, resulting in accelerated runoff and erosion. Streambanks and moist soil around springs and streamside meadows are particularly susceptible to erosion damage and compaction. Livestock impacts to streams and riparian sites can be limited by providing water, mineral supplements, and shelter at locations away from natural water sources. Special fencing or livestock herding may also be needed to protect sensitive areas from excessive use at critical times. Another reason for protecting wetland or riparian sites is that they serve as vegetation buffer strips that slow runoff and trap sediment before it reaches a waterway.

References

1. Warren, S.D.; Blackburn, W.H.; Taylor, C.A., Jr. Effects of season and stage of rotation cycle on hydrologic condition of rangeland under intensive rotation grazing. J. Range Manag. **1986**, *39*, 486–491.
2. Sutterlund, D.R. *Wildand Watershed Management*; Ronald Press: New York, 1972.
3. Boyle, M.; Frankenberger, W.T., Jr.; Stolzy, L.H. The influence of organic matter on soil aggregation and water infiltration. J. Prod. Agric. **1989**, *2*, 290–299.
4. Hudson, N. *Soil Conservation*; Cornell University Press: Ithaca, New York, 1981.
5. Thurow, T.L.; Blackburn, W.H.; Taylor, C.A., Jr. Infiltration and interrill erosion responses to selected livestock grazing strategies, Edwards plateau, Texas. J. Range Manag. **1988**, *41*, 296–302.
6. Thurow, T.L.; Blackburn, W.H.; Taylor, C.A., Jr. Hydrologic characteristics of vegetation types as affected by livestock grazing systems, Edwards plateau, Texas. J. Range Manag. **1986**, *39*, 505–509.
7. Ellison, L. Influence of grazing on plant succession of rangelands. Bot. Rev. **1960**, *26*, 1–78.
8. Thurow, T.L.; Blackburn, W.H.; Taylor, C.A., Jr. Some vegetation responses to selected livestock grazing strategies, Edwards plateau, Texas. J. Range Manag. **1988**, *41*, 108–114.
9. Thurow, T.L. Hydrology and erosion. In *Grazing Management: An Ecological Perspective*; Heitschmidt, R.K., Stuth, J.W., Eds.; Timber Press: Portland, Oregon, 1991; 141–159.

10. Menzel, R.G.; Rhoades, E.D.; Olness, A.E.; Smith, S.J. Variability of annual nutrient and sediment discharges in runoff from Oklahoma cropland and rangeland nutrient and bacterial pollution. J. Environ. Qual. **1978**, *7,* 401–406.
11. Robbins, W.D. Impact of unconfined livestock activities on water quality. Trans. Am. Soc. Agric. Eng. **1979**, *22*, 1317–1323.
12. Woodmansee, R.G.; Allis, I.; Mott, J.J. Grassland nitrogen. Ecol. Bull. **1981**, *33*, 443–462.
13. Khaleel, R.; Keddy, D.R.; Overcash, M.R. Transport of potential pollutants in runoff water from land areas receiving animal wastes: a review. Water Resour Res. **1979**, *14,* 421–436.
14. Bohn, C.C.; Buckhouse, J.C. Coliforms as an indicator of water quality in wildland streams. J. Soil Water Conserv. **1985**, *40,* 95–97.
15. Walker, J.J.; Montemagno, C.D.; Jenkins, M.B. Source water assessment and nonpoint sources of acutely toxic contaminants: a review of research related to survival and transport of *Cryptosporidium parvum*. Water Resour. Res. **1998**, *34,* 3383–3392.
16. Tate, K.W.; Atwill, E.R.; George, M.R.; McDougald, N.K.; Larsen, R.E. *Cryptosporidium parvum* transport from cattle fecal deposits on California rangelands. J. Range Manag. **2000**, *53,* 295–299.
17. Stephenson, G.R.; Street, L.V. Bacterial variations from a southwest Idaho rangeland watershed. J. Environ. Qual. **1978**, *7,* 150–157.
18. Burton, G.A.; Gunnison, D.; Lanza, G.R. Survival of pathogenic bacteria in various freshwater sediments. Appl. Environ. Microbiol. **1987**, *53,* 633–638.
19. Larsen, R.E.; Miner, J.R.; Buckhouse, J.C.; Moore, J.A. Water quality benefits of having cattle manure deposited away from streams. Biores. Technol. **1994**, *48,* 113–118.
20. Miner, J.R.; Buckhouse, J.C.; Moore, J.A. Will a water trough reduce the amount of time spent in the stream? Rangelands **1992**, *14,* 35–38.
21. Thurow, T.L.; Taylor, C.A., Jr. Viewpoint: the role of drought in range management. J. Range Manag. **1999**, *52,* 413–419.

65

Water: Drinking

Marek Biziuk
and Matgorzata
Michalska

Introduction

Water is a source and a basis of life on earth, and is a substance vital and necessary for humans and their well-being. It is a main part of our daily diet, transporting many necessary macro- and microelements in the human body. Excluding fat, water composes approximately 70% of the human body by mass. We drink 2 to 3 L of water every day; thus, the quality of drinking water is important. Humans can survive for several weeks without food, but for only a few days without water. Additionally, clean water is becoming a much desirable and expensive product. Desertification of large areas of the world, increasing industrialization, and intensification of agriculture result in a large input of toxic organic and inorganic compounds, significantly reducing the quality of surface waters, which are a source of drinking water for a large part of the world population. Other dangers include biological pollution by bacteria and viruses, human fecal pathogens, and parasites, causing waterborne diseases. Over large parts of the world, humans have inadequate access to potable water and use sources contaminated with unacceptable levels of dissolved chemicals or bacteria and viruses. Thus, water, especially drinking water, can create a serious hazard to human health and life, as well as to the life and health of flora and fauna. Sources of drinking water include spring water and groundwater; surface water from rivers, streams, lakes, and glaciers; water from precipitation (rain, snow); and seawater after desalination. All industrialized countries have 100% access to safe drinking water (data from the United Nations Children's Fund from 2000); however, some countries, especially those from Africa (Kenya, Uganda), have limited access.[1] The population percentages from chosen countries with limited access to safe drinking water are listed in Table 1. It was estimated that in 2006, about 1.1 billion people lacked proper drinking water. Waterborne diseases were estimated to cause 1.8 million deaths each year.[2]

Unfortunately, quality water from groundwater supply, which is practically non-renewable, is diminishing. In addition, anthropogenic groundwater pollution has become a fact of life. The majority of anthropogenic water pollutants are toxic not only to humans but also to animals and plants. These pollutants should be removed or destroyed during water treatment, and determined and monitored in surface and tap water. Water pollutants can be divided into physical, biological, radioactive, inorganic, and organic.

TABLE 1 Percentage of Population with Access to Safe Drinking Water (2000)

Country	%	Country	%	Country	%
Algeria	89	Indonesia	78	South Africa	86
Azerbaijan	78	Iran	92	South Korea	92
Brazil	87	Iraq	85	Sudan	67
Chile	93	Kenya	57	Syria	80
China	75	Mexico	88	Turkey	82
Cuba	91	Morocco	80	Uganda	52
Egypt	97	Peru	80	Venezuela	83
India	84	Philippines	86	Zimbabwe	83

Source: United Nations Children's Fund (UNICEF).[1]

For many millennia, the basic criteria of water suitability for drinkable purposes was organoleptic analysis—checking the taste, smell, and appearance of water. These criteria are included, even now, in guidelines for drinking water quality; however, they had to be dramatically verified when in 1854 it was established in London that water from the well at Brad Street, very popular because of its taste, smell, and appearance, was the source of a cholera epidemic.[3] Threshold limit values for bacteriological contamination of water (the so-called coliform count) were the first standard introduced for controlling drinking water quality (in 1914 in the United States). The presence of coliform bacteria is the most convenient for a class of harmful fecal pathogens. This standard is still valid today. Microbial pathogenic parameters are very important because of their immediate health risk. The following were included in guidelines for drinking water quality: metals and anions since 1925, organic compounds since 1942, and radioactive substances since 1962. The Public Health Service and later the Environmental Protection Agency (EPA) have prepared such guidelines for the United States.[4,5] The World Health Organization (WHO)[6] worldwide and the European Community (EC)[7,8] in Europe also elaborate and publish such guidelines. On these guidelines, each country defines its maximum admissible concentrations (MACs) of contaminants in drinking water. Each member state is responsible for establishing the required policing measures to ensure that the legislation is implemented.[9]

Drinking water should fulfill the following requirements:

- It should be clear, without smell, colorless, and refreshing in taste
- It should not contain pathogenic bacteria; animal parasites and their larvae or eggs; toxic compounds; and excessive quantities of calcium, magnesium, iron, or manganese compounds
- It should be easily accessible, be always of good quality, and occur in appropriate amounts
- It should be permanently protected from contamination.
- It should contain substances that are necessary for human life in the proper amounts

Inorganic Components and Pollutants of Drinking Water

The problem of the presence of mineral components and anions in drinking water is very complicated because some elements are necessary for human life.[10–12] They play a significant role in physiological processes and metabolic functions. We can divide these into two groups:

- Macroelements, such as carbon, hydrogen, oxygen, nitrogen (the main components of organic matter), as well as sulfur, potassium, magnesium, calcium, sodium, phosphorus, and chlorine
- Microelements, such as boron, cobalt, chromium, copper, fluorine, iron, lithium, molybdenum, manganese, nickel, rubidium, selenium, silicon, vanadium, and zinc (which also participate in the metabolic function and are components of enzymes), as well as hormones, vitamins, etc

Table 2 shows some macro- and microelements, their recommended daily intake, and the effects of deficiency in these elements. Some of these elements are toxic at higher concentrations, and very often

TABLE 2 Macro- and Microelements, Their Recommended Daily Intake, and the Effects of Deficiency

Element	Recommended Concentration in Drinking Water in Poland (mg/L)	Recommended Daily Intake (mg)	Effects of Deficiency
Ca	—	900–1200	Osteoporosis, rickets, neurological disorders
Fe	0.2	14–18	Anemia
P	—	700–900	Muscular weakness, bone aches, lack of appetite
Mg	30–125	300–370	Gastrointestinal problems, muscular cramps
Na	200	1000	Gastrointestinal problems, diarrhea, weakness, headache
K	—	2–3.5	Muscular weakness, neurological disorders
Zn	—	13–16	Hair loss, skin lesions, reduced resistance to infection
Cu	2	2–2.5	Anemia, heart arrhythmia, blood vessel rupture
F	1.5	1.5–4	Tooth decay
I	—	0.15–0.16	Underactive thyroid, goiter
Se	0.01	0.06–0.075	Cancer, heart diseases, reduced resistance to microbiological and viral infection, reduced antibody production

the difference between an indispensable and a toxic dose is very small. Insufficient or excessive daily intake of these elements can cause harmful effects (see Tables 2 and 3).[9,11,12] In drinking water, we can find all elements and inorganic substances present in nature and introduced by industry, some not needed by organisms and toxic to humans. The most dangerous are lead, mercury, cadmium, and arsenic (the so-called Big Four), but the following are hazardous as well: aluminum, antimony, asbestos, barium, boron, chromium, cobalt, copper, cyanides, manganese, nickel, silver, tin, vanadium, and zinc.[12–49] The harmfulness of metals is mainly due to their persistence in the environment and biomagnification in the food chain, as well as their tendency to accumulate in selected tissues. Metals can be transported with water and air, and thus can endanger humans also through the water they drink. In Table 3, the toxicity of the most dangerous elements, their daily intake, and their typical and recommended concentrations in drinking water are listed.[4–9,23–34]

Underground and even mineral water, used as a source of drinking water, are not as clean as we would expect. Very often, overly high natural concentrations of manganium, iron, fluorides, ammonia, and hydrogen sulfide occur, and these waters should be treated before consumption. Volatile compounds such as ammonium and hydrogen sulfide can be removed by aeration, whereas manganium and iron can be removed by precipitation in the form of hydroxides and filtration or biotechnological methods. Fluorides present the biggest problem.[50] The best solution is to mix high-fluoride water with water containing no or very small concentrations of fluorides. An additional problem concerns pipes made from asbestos-cement materials or lead, used in the past in water distribution systems. Lead can be diluted in water, particularly when water is treated by chlorination or ozonization. Asbestos can contaminate drinking water as a result of corrosion, and thereby accumulate in lung tissues, resulting in symptoms that can develop many years after exposure. Asbestos fibers in water are usually shorter than 10 pm, and can cause cancer of the gastrointestinal and urinary systems. When they are transferred from water to air, after water spillage and drying, asbestos fibers can cause lung and throat cancer upon inhalation. Such an episode was described by Webber,[51] when in Woodstock, New York, in water transported by eroded asbestos-cement pipes in the distribution system; 10,000 asbestos fibers per liter was detected in some samples and up to 0.12 fiber per milliliter was detected in household air.

All these elements and substances and their recommended MACs are listed in the WHO and EC Guidelines for Drinking Water Quality, or in the EPA and national MACs,[4–9] and have to be determined

TABLE 3 Toxicity of the Most Dangerous Elements, Their Daily Intake, and Typical and Recommended Concentrations in Drinking Water

	Concentration in Drinking Water (mg/L)					
		Recommended MAC				
Element	Typical	EC	WHO	Poland	Daily Intake (mg)	Toxicity
Al	0.02–0.1		0.1–0.2	0.2	20–45	Senile dementia of the Alzheimer's type
As	<0.01	0.01	0.01	0.01	0.01–0.02	Respiratory and gastrointestinal tract, skin, and central nervous system effects leading to coma and death, peripheral artery disease, cardiovascular disease, diabetes mellitus, various cancers, neurological effects; the inorganic form of arsenic (As^{3+}) is the most toxic, whereas elemental as well as organic arsenic compounds are considered to be virtually non-toxic
B	0.1	1	0.5	1	1.3–4.5	Dry skin and gastric disorders
Cd	0.002	0.005	0.003	0.003	0.03–0.04	Proteinuria, renal dysfunction, damage to the blood vessel system, lung cancer, low reproduction functions, emphysema, hypertension
Cr	0.002	0.05	0.05	0.05	0.04–0.08	Dermatitis, skin ulcers, respiratory and gastrointestinal problems, liver and kidney damage
Co	0.005	—	—	—	0.04–0.05	Gastrointestinal problems, heart dilation, secondary thrombosis, neurotoxicological symptoms
Cu	<0.1	2		2.0	2–4	Congenital disorder causing accumulation of copper in the liver, brain, and kidney resulting in hemolytic anemia and neurological abnormalities
Fe	0.001–0.5	—	—	0.2	12–16	Gastrointestinal problems, fibrosis, heart disease, abnormal glucose metabolism
Pb	0.002–0.01	0.01	0.01	0.025	0.1–0.5	Carcinogenic, anemia, hypertension, renal dysfunction, insomnia, weakness
Mn	0.02–0.05		0.4	0.05	3.7	Apathy, anorexia, insomnia, speech disturbance, central nervous system effects
Hg	0.0002–0.0003	0.001	0.006	0.001	0.6	Organic forms of mercury (methyl- and dimethylmercury)— the most toxic of the discussed metals; central nervous damage; DNA deformation; immunological impairment; mutagenic, carcinogenic, and teratogenic effects; coronary disease; brain damage
F	0.5	1.5	1.5	1.5	0.5–2	Osteoporosis, fluorosis, central nervous system effects, carcinogenic (osteosarcoma)
Ni	0.001–0.002	0.02	0.07	0.02	0.1–0.9	Dermatoses, potential carcinogenity
Se	<0.01	0.001	0.01	0.01	0.15	Atrophy and decompensation of the heart, renal glomerulonephritis, liver cirrhosis, anemia
V	4–222	—	—	—	2	Irritation of skin and eyes, discoloration of the tongue and oral mucosa

in drinking water. In Tables 2 and 3, recommended MACs for some metals from the WHO Guidelines for Drinking Water Quality and Polish standards are listed as examples. Excessive concentrations of pollutants, above maximum contaminant levels, not only warn consumers but also signal the existence of uncontrolled discharge of wastes, improperly operating treatment plant, lack of enforcement of legislation dealing with water management, or other violations of environmental laws. To be able to monitor environmental pollutants properly and effectively, analysts need a variety of methods for the isolation and determination of these contaminants, taking into account both physical-chemical properties of individual compounds and groups of compounds, as well as the characteristics of the matrices in which these compounds occur.

Consequently, there is a need for continuous monitoring of the degree of pollution of potable and surface waters by all these elements and substances. The following have been used mainly for final analysis: flame atomic emission spectrometry; flame atomic absorption spectrometry; electrothermal atomic absorption spectrometry; inductively coupled plasma (ICP) coupled with atomic emission, atomic absorption, or mass spectrometry (MS); and electrochemical methods (anodic stripping voltammetry).[12,24,52–63]

Organic Pollutants of Drinking Water

The majority of organic compounds in water are of natural origin. They represent a large and diversified group of mostly unidentified compounds such as humic and fulvic acids, tannins, peptides, amino acids, etc. Their total concentration varies from 0 to 1 mg/L in groundwater, from 1 to 5 mg/L in most surface waters, and from 20 to 25 mg/L in some waters with the highest concentration of organic matter.[20,52,53,63–67] These compounds are predominantly non-toxic, but they can be precursors of toxic compounds in the process of water treatment. The main hazard to life and health of humans and flora and fauna comes from anthropogenic organic compounds. The number of known organic compounds is now estimated to be about 16 million,[3] 2 million of which are produced by synthesis alone. Every year, approximately 250,000 new compounds are synthesized, about 1000 of which are manufactured on an industrial scale.[68] Presently, ca. 70,000 organic compounds are commercially available with an annual global production of 100–200 million tons. Approximately one- third of all organic compounds produced end up in the environment, including water. More than 700 chemical compounds, including more than 600 organic compounds,[64,69] many of which are biologically active, have been detected in some water supply samples. Using gas chromatography (GC)-MS, Coleman et al.[69] identified approximately 460 organic compounds in Cincinnati drinking water.

Particularly important pollutants among organic compounds are volatile organohalogen compounds and pesticides as a result of their common use, persistence in the environment, and toxicity. They are primarily anthropogenic. Volatile organohalogen compounds are used mainly as solvents, cleaning and degreasing agents, blowing agents, polymerization modifiers, and heat-exchange fluids. It is estimated that the annual global production of organohalogen solvents alone amounts to several million tons.[64] The most important source of organohalogen compounds in drinking water, particularly volatile ones, is water disinfection by chlorination[14,63–65,70–75] for killing pathogenic bacteria. It is a paradox that during this process, harmless and naturally occurring humic and fulvic compounds in water (the so-called precursors) are converted into organohalogen compounds, which are dangerous to human life and health (see Figure 1). The largest group of compounds formed during chlorination comprises trihalomethanes (THMs), i.e., trichloromethane (chloroform, the most abundant compound), bromodichloromethane, dibromochloromethane, and tribromomethane. Organobromine compounds are formed when the water being chlorinated contains a large amount of bromides or when the chlorine used for disinfection is contaminated with bromine. Hypobromous acid formed in the reaction of bromide ions with hypochlorous acid reacts with an organic matrix about 200 times faster than hypochlorous acid does.[76] The amount and kind of organohalogen compounds formed depend on the water pH, the amount of chlorine used, and the content of organic matrix (total organic carbon, TOC) in chlorinated water.

Apart from THMs, other organohalogen compounds are formed in the course of chlorination, in smaller quantities. Volatile organohalogen compounds such as tetrachloromethane, chloroethylene, 1,1-dichloroethylene, 1,1,2-tri-chloroethylene, tetrachloroethylene, 1,1,1-trichloroethane, and 1,2-dichloroethane are also commonly found in chlorinated water. In addition, more than 100 other derivatives after chlorination of humic substances have already been identified, including chlorinated

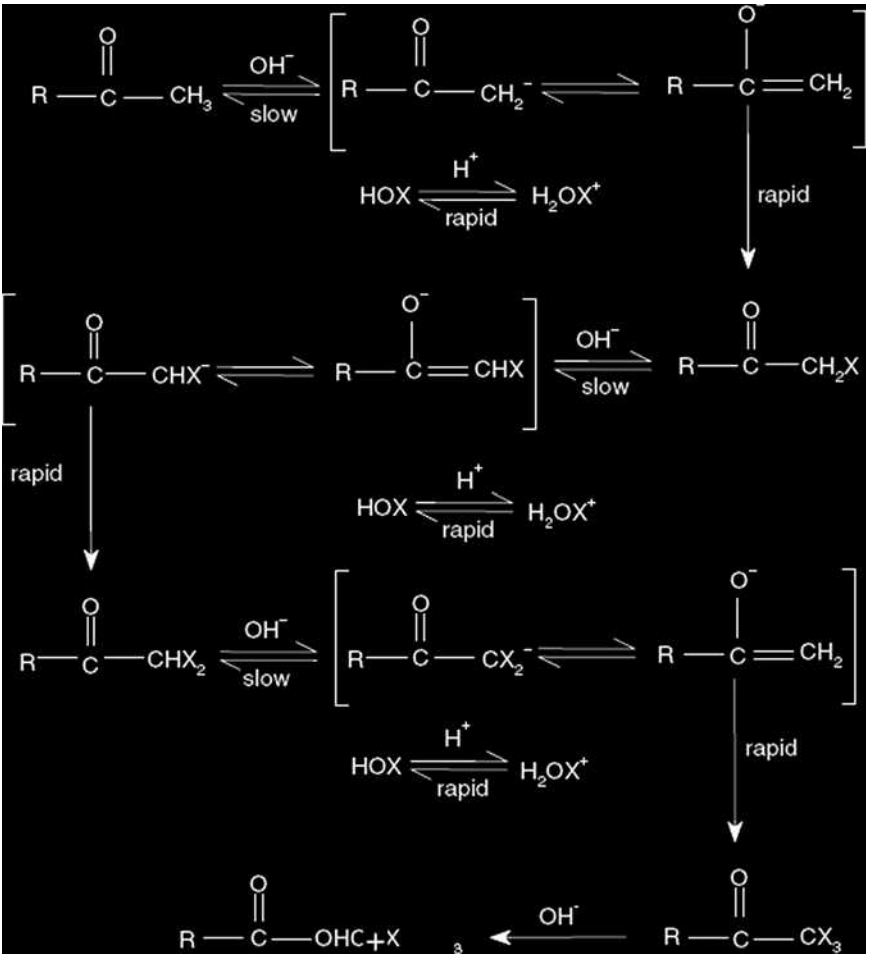


FIGURE 1 The haloform reaction—THMs formation.

acetone, chlorinated acetonitrile, chloropicrin, chloral, chloroacetic acids, chlorinated ethers, chlorophenols, chlorinated ketones, etc.[71] Koch and Krasner[74] estimated that among organohalogen compounds formed during chlorination of water, 77% are THMs, 15% are haloacetic acids, 3% are halonitriles, 4% is trichloroacetaldehyde hydrate, and 1% are the remaining compounds. Typical concentrations of trichlorometh-ane (chloroform), the most frequently present compound in chlorinated water, varied from 1 to 30 μg/L. Volatile organohalogen compounds have been determined primarily in tap water;[3,63–65,68–76] however, they also occur in surface water, groundwater, rainwater,[65,70,77,78] or even in the water and ice of polar regions.[65] Several volatile organohalogen compounds fall into the category of known or suspected carcinogens. In Table 4, the carcinogenic potential of organohalogen compounds occurring in drinking water and recognized by the EPA[2,4,5] as toxic or carcinogenic is presented. The MAC for drinking water varied from 0.3 μg/L for 1,1-dichloroethene (WHO, Norway) to 100 μg/L (EU, WHO, EPA, United Kingdom) and 350 μg/ L (Canada) for the total THMs. The MAC in the directives of WHO, Poland, and Great Britain for chloroform in tap water is 30 μg/L. Besides the groups of compounds discussed above, large amounts of other toxic pollutants, including pesticides, polychlorinated biphenyls (PCBs), polycyclic aromatic hydrocarbons (PAHs), crude oil derivatives, phenols, endocrine disruptors, etc., can be present in drinking water. Pesticides enter water either directly

TABLE 4 Carcinogenic Potential of Organohalogen Compounds Occurring in Drinking Water and Recognized by the U.S. EPA as Toxic or Carcinogenic

Compound	Carcinogenic Potential (μg/kg body/day)	Certainty of Hazard
Chloromethane	2.5	S
Dichloromethane	2.2	3
Trichloromethane (causes hepatocellular carcinomas and liver and kidney tumor in mice, and renal tumor in rats after chronic exposure)	11	CA
Tetrachloromethane (induces preneoplastic changes in the liver of rats, and liver tumors in male rats and mice)	0.4	CA
Bromodichloromethane	55	S
Dibromochloromethane	250	S
Tribromomethane	1100	S
Dichloroiodomethane	2.2	S
1,2-Dichloroethane	9.4	CA
1,1,1-Trichloroethane	20	S
Vinyl chloride	0.017	CU
Trichloroethylene (causes increases in the incidence of hepatocellular carcinomas, pulmonary carcinoma, and malignant lymphomas in mice)	1.6	CA
Tetrachloroethene (induces high incidence of nephropathy in mice and rats, and hepatocellular carcinomas in mice)	0.73	CA
1,4-Dichlorobenzene	1.4	S
1,3-Dichlorobenzene	88	S
1,2-Dichlorobenzene	6.9	S
1,2,4-Trichlorobenzene	4.4	S
2,4-Dichlorophenol	1.7	S
Pentachlorophenol	40	S
Polychlorinated biphenyls	53	S

Note: CA, animal carcinogen; CU, human carcinogen; S, suspected carcinogen.

by their application for mosquito control, or indirectly from drainage of agricultural lands, permeation through soil, erosion, wastewater from pesticide production, municipal wastes (fungicides and bactericides), etc. Consequently, a number of papers have been published confirming the presence of pesticides and PCBs not only in surface waters but also in potable water, particularly in water from wells existing in agricultural areas, in rainwater, as well as in water and ice from polar regions.[66,79–81] For humans, the major route of exposure to these pollutants is the gastrointestinal system, mainly by way of food (because of bioaccumulation and biomagnification in the food chain), but also through drinking water. Pesticides, PCBs, phenols, and some PAHs are carcinogenic, mutagenic, and teratogenic, and cause cardiovascular, neurological, and other diseases;[14,20,64,66,69,82] thus, there is a need for continuous monitoring of the degree of pollution by these compounds of potable and surface waters. Owing to the possibility of bioaccumulation of organic compounds, even low concentrations can result in poisoning of an organism. WHO, the EC, and most countries, including Poland, introduced a MAC for some toxic organic compounds present in tap water.[4–9,83] In Table 5, the MACs for toxic organic compounds recommended by WHO and those established by Poland are listed.[6,9]

Consequently, there is a need for continuous monitoring of the degree of pollution of potable and surface waters by toxic anthropogenic organic compounds. To be able to monitor environmental pollutants properly and effectively, analysts need a variety of methods for the isolation and determination of these contaminants, taking into account both physical-chemical properties of individual compounds

TABLE 5 WHO Guideline Values for Organic Compounds That Are of Health Significance in Drinking Water (In Milligrams per Liter) and the Established Macs for Drinking Water in Poland

	Concentration in Drinking Water (mg/L)			Concentration in Drinking Water (mg/L)	
	Recommended MAC			Recommended MAC	
Compound	WHO	Poland	Compound	WHO	Poland
Acrylamide	0.0005	0.0001	Edetic acid (EDTA)	0.6	
Alachlor	0.02	0.0001	Endrin	0.0006	0.0001
Aldicarb	0.01	0.0001	Epichlorohydrin	0.0004	0.0001
Aldrin and dialdrin	0.00003	0.0001	Ethylbenzene	0.3	
Atrazine	0.002	0.0001	Fenoprop	0.009	0.0001
Benzene	0.01	0.001	Hexachlorobutadiene	0.0006	0.0001
Benzoja jpyrene	0.0007	0.00001	Isoproturon	0.009	0.0001
Bromodichloromethane	0.06	0.015	Lindane	0.002	0.0001
Bromoform	0.1		2-Methyl-4-chlorophenoxyacetic acid	0.002	0.0001
Carbofuran	0.007	0.0001	Mecoprop	0.01	0.0001
Carbon tetrachloride	0.004	0.002	Metoxychlor	0.02	0.0001
Chlordane	0.0002	0.0001	Metolachlor		0.0001
Chloroform	0.3	0.03	Microcystin LR	0.001	0.001
Chlorotoluron	0.03	0.0001	Molinate	0.006	0.0001
Chlorpyrifos	0.03	0.0001	Monochloramine	3	
Cyanazine	0.0006	0.0001	Monochloracetate	0.02	
2,4-Dichlorophenoxyacetic acid	0.03	0.0001	Nitrilotriacetic acid	0.2	
Dichlorodiphenylotrichloroethane and metabolites	0.001	0.0001	N-Nitrosodiethylamine	0.1	
Di(2-ethylhexyl)phthalate	0.008	0.02	Pendimethalin	0.02	0.0001
Dibromoacetonitrile	0.07		Pentachlorophenol	0.009	
Dibromochloromethane	0.1		Permethrin	0.3	0.0001
1,2-Dibromo-3-chloropropane	0.001		Perlproxylen	0.3	
1,2-Dibromomethane	0.0004		Simazine	0.002	0.0001
Dichloroacetate	0.05		Styrene	0.02	
Dichloroacetonitrile	0.02		2,4,5-Trichlorophenoxyacetic acid	0.009	
1,2-Dichlorobenzene	1		Terbuthylazine	0.007	0.0001
1,4-Dichlorobenzene	0.3		Toluene	0.7	
1,2-Dichloroethane	0.03	0.003	Trichloroacetate	0.2	
1,2-Dichloroethene	0.05		Tetrachloroethene	0.04	0.01
Dichloromethane	0.02		Trichloroethene	0.02	
1,2-Dichloropropane	0.04		2,4,6-Trichlorophenol	0.2	0.2
1,3-Dichloropropene	0.02		Trichloromethanes, total	1	0.15
Dichloroprop	0.1	0.0001	Vinyl chloride	0.0003	0.0005
Dimethoate	0.006	0.0001	Xylenes	0.5	
1,4-Dioxane	0.05		Pesticides, total		0.0005

and groups of compounds and the characteristics of the matrices in which these compounds occur. Volatile organohalogen compounds, pesticides, and other organic pollutants occur in water at relatively low concentration levels; thus, there is a need for the isolation of organic compounds from a complex aqueous matrix and their preconcentration.

Methods of isolation and preconcentration of organic compounds from water are closely associated with the kind of analytes, their volatility, polarity, stability, water solubility, solubility in organic solvents, etc. Numerous techniques for the isolation and enrichment of organic analytes have been developed, the most common ones being solvent extraction, solid sorbent extraction, techniques utilizing the distribution of solute among the liquid and the gaseous phase (headspace, purging), as well as less commonly used techniques such as freezing out, lyophilization, vacuum distillation, steam distillation, and membrane processes (reverse osmosis, ultrafiltration, dialysis).[53–58,61–70,84–91] A schematic diagram of utilization of various isolation techniques for the determination of organic compounds in water is shown in Figure 2. The only method of determination of organohalogen compounds in water that avoids the isolation and preconcentration step uses direct injection of an aqueous sample onto a GC column and an electron capture detector (direct aqueous injection-electron capture detection, DAI-ECD).[63,65,70,92–97] This method requires a special injector allowing cold on-column injection and special capillary columns (see Figure 3). Cooling of the injector prevents sample evaporation from the needle of a syringe before its withdrawal. The DAI-ECD technique has been successfully used for the determination of volatile organohalogen compounds containing one or two carbon atoms (the products

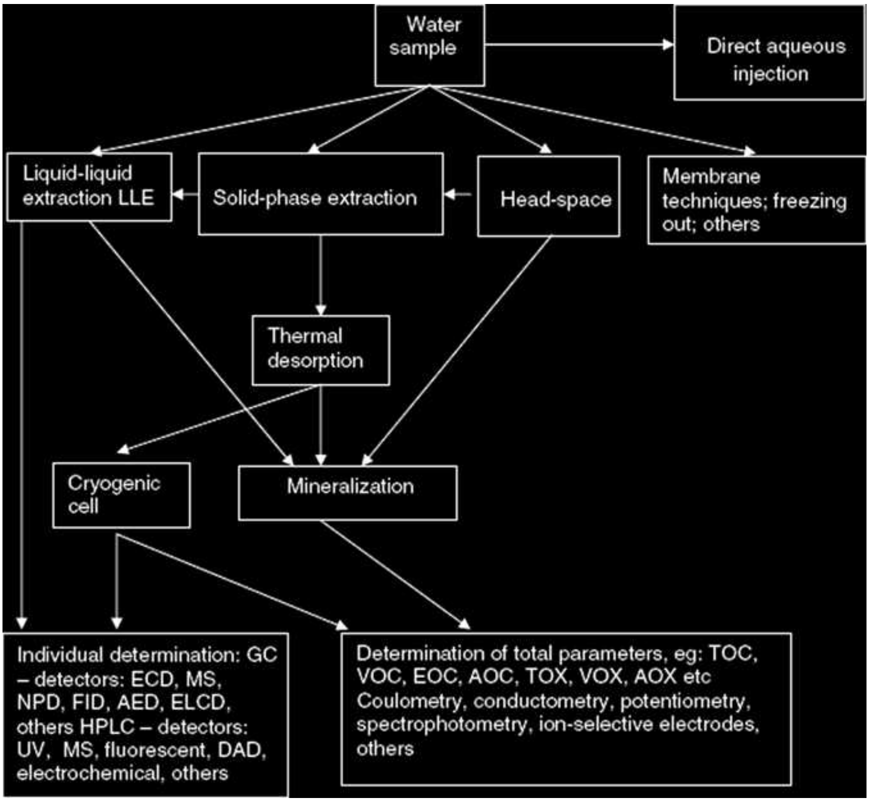


FIGURE 2 A general procedure for the analysis of trace organic pollutants in water.

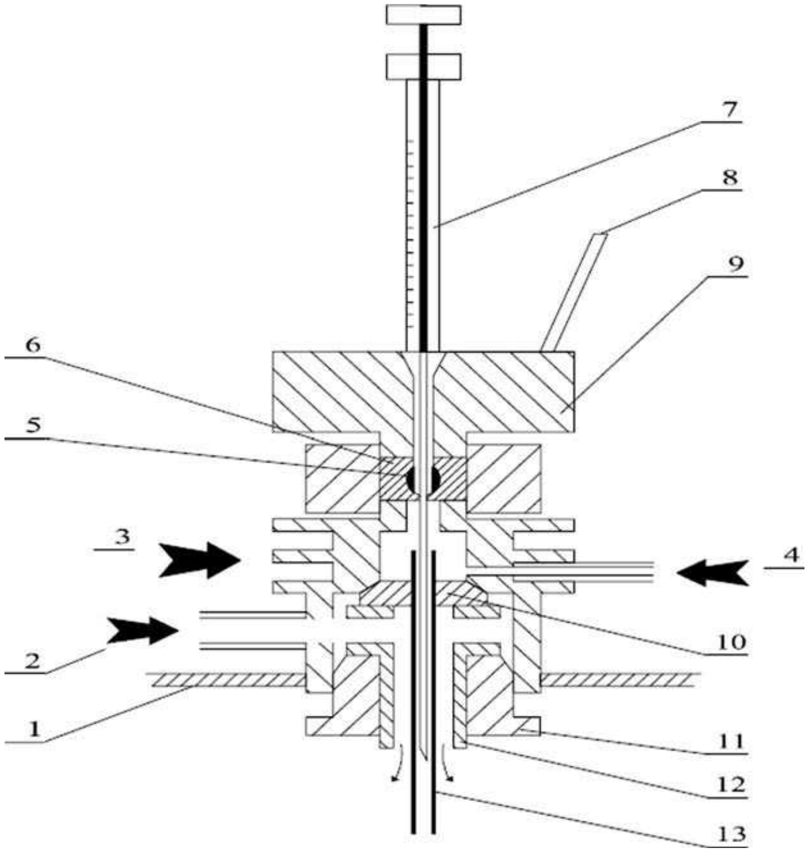


FIGURE 3 Schematic diagram of non-vaporizing septum-less, cold on-column injector: 1—the oven wall; 2—the secondary air cooling; 3—the principal air cooling system; 4—the carrier gas inlet; 5—the stainless steel rotating valve; 6—the valve seal; 7—the syringe; 8—the valve lever; 9—the upper part of injector; 10—the column seal; 11—the bottom part of injector; 12—the cooling jacket; 13—the capillary column.

of chlorination of humic substances; THMs) in tap water and surface water.[63,65,90,92–96] The method avoids the problems associated with incomplete recovery of the analytes during their isolation from the aqueous phase, the effect of potential contaminants when using solvent or solid-phase extraction, and losses of the analytes during the enrichment step. The main advantages of DAI-ECD include its simplicity (no isolation and preconcentration methods are necessary), repeatability, reduction of the possibility of sample contamination, and low detection limits (0.015–0.6 μg/L depending on the percentage of halogen in the compound). The detection limit of the method is related to the amount of the analyte in the sample injected onto the column and the volatility of the analytes.

For final determination, GC with specific detectors such as ECD, nitrogen-phosphorus detector (NPD), flame photometric detector, atomic emission detector (AED), and electrolytic conductivity detector (ELCD), or GC with universal detectors such as flame ionization detector (FID) or mass spectrometry (MS) is used. Liquid chromatography methods with ultraviolet (UV), MS, diode-array (DAD), electrochemical, or fluorescent detectors are also often used.[53–58,62,63,65–68,92–96] In Figure 2, two main ways of determining organic compounds in water are shown: determination of individual compounds or of total parameters. Total parameters such as total organic carbon or halogen (TOC, TOX), dissolved organic carbon or halogen (DOC, DOX), and suspended organic carbon or halogen (SOC, SOX) are used to characterize the content of organic compounds in water. Other parameters are defined in terms of the method of isolation of an organic fraction from water: volatile (VOC, VOX) or purgeable (POC, POX), extractable (EOC, EOX), and adsorbable (AOC, AOX) organic carbon or halogen. However, total parameters measuring the carbon content in an organic fraction are not particularly suitable as an estimate of anthropogenic water pollutants and their hazard to human health since a decisive majority of organic compounds in water are biogenic.

Characteristics of Some Water Bacteria and Pathogens in Drinking Water

Natural non-contaminated groundwater and surface water contain relatively low numbers of microorganisms. In general, groundwater has fewer bacteria than surface water. Microbial abundance and species composition depend on the depth of water-bearing stratum, soil type, and the geological structure and sanitary conditions in the given area. Aquatic bacteria are both autochthonic and allochtonic organisms. Autochthonic organisms, the so-called local microbes, reside permanently in the waters of a given water body, while allochtonic (alien) organisms are transferred into the water basin from the atmosphere and soil and through municipal and industrial wastewaters. Autochthonic microflora consists of autotrophic bacteria that are capable of photo- and chemosynthesis. It also includes heterotrophic organisms that cannot synthesize their own food. Purple as well as green bacteria are among the photosynthesizing autotrophs. Chemosynthetic bacteria are represented by nitrifying bacteria, iron- and sulfur-oxidizing bacteria, and hydrogen bacteria. Many species of autochthonic bacteria are heterotrophic organisms that feed exclusively on non-organic matter of allochthonous origin. Typical heterotrophic bacteria belong to the genera *Pseudomonas, Spirocheta, Vibrio,* and *Aeromonas.* Allochthonic organisms are mostly heterotrophic. They consist of saprophytes (feeding on organic compounds that are the products of decomposing remnants of plants and animals) as well as parasites (feeding at the expense of other live organisms). The most abundant saprophytic group is represented by a rod-shaped species such as *Escherichia coli*, *Aerobacter aerogenes*, and *Serratia marcescens*. Moreover, the rods from the genera *Bacillus* and *Clostridium* that become transferred from soil into the natural waters during rainfall can be classified as allochtonic aquatic bacteria. Water, which is not the natural environment of disease-causing microorganisms, plays a very important role in the process of their transfer. In cases when infectious diseases occur among the general public or when people become disease carriers, it is possible that potable water can become contaminated with the excrements of infected humans. The use of such water for drinking or food preparation and the exposure to it during washing may result in infection. The survival times of disease-causing organisms in the aquatic environment depend on many factors, inter alia, the pH of water, UV radiation, the presence of chemical compounds, and the level of water contamination. The main source of disease-causing microorganisms in natural waters is wastewater of human and animal origin. Pathogenic bacteria that become transferred into the water bodies through excrements, wastewater, and runoff from the fields are mainly related to digestive tract disorders in humans and animals.[98–100] The most typical bacteria of this kind come from the rod-shaped genus *Salmonella* and cause a variety of digestive system infections. The main genera of disease-causing bacteria that are transferred by water and the diseases caused by these bacteria are presented in Table 6. In water delivered by the water supply system, there can be microorganisms that have no influence on human health but form a film or membrane on the inside of the pipeline walls. The growth of such bacteria has a detrimental effect on water quality. A typical manifestation of biological fouling inside the water supply system is the altered color, turbidity, and undesirable taste and smell of water. In addition, the parameters of water exploitation may worsen owing to microbiological corrosion, plugging of the water-bearing stratum, and the lowered efficiency of the well.

TABLE 6 Main Kinds of Pathogenic Bacteria Found in Water and the Diseases They Cause

No.	Pathogenic Bacteria	Diseases
1	Rod-shaped Salmonella	Abdominal typhoid fever, Paratyphoid fever, salmonellosis
2	Rod-shaped Shigella	Bacterial dysentery
3	Curved rods Vibrio cholerae	Cholera
4	Francisella tularensis	Tularemia
5	Spirochaeta from Leptospira genus	Spirochetal jaundice, leptospirosis
6	Rod-shaped Escherichia coli	Acute dysentery, infection sticks of colon
7	Legionella pneumophila	Acute pneumonia, Legionnaires' disease
8	Aerobic, spore-forming bacilli Bacillus	Corneal abrasion, ankylosing spondylitis, feed poisoning, anthrax
9	Rod-shaped Mycobacterium	Tuberculosis
10	Yersinia enterocolitica	Yersiniosis

Microbiological Research Methods for Analysis of Drinking Water

The danger of humans contracting infections from water necessitates the permanent sanitary control of drinking water by bacteriologists and hygienists. Simple methods that can indicate the presence of pathogenic organisms are used because indirect methods are laborious and time consuming. Currently, in routine studies, intermediate methods based on indicative bacteria are used. The organism indicative of contamination has to comply with the following conditions:

- It has to be a permanent inhabitant of the intestine that occurs in large numbers
- It cannot be a pathogenic bacterium
- It cannot form spores
- It cannot multiply in water
- should be easy to detect
- It has to exist longer than pathogens in the natural environment (water, soil)
- It should be removed during the water conditioning in a similar way as the pathogen organisms

Water samples for microbiological analysis must always be collected in a sterile container. The samples should be refrigerated and transported to the testing laboratory as soon as possible. Fecal contamination of water can be determined by the presence of fecal coliforms or enterococci in a water sample by the multiple-tube technique. The bacteria can also be detected by the membrane filtration technique. The membrane filtration technique is based on the entrapment of the bacterial cells by a membrane filter (pore size, 0.22 or 0.45 μm). After filtration, the membrane with the microorganisms is placed on a petri dish containing the appropriate medium and incubated. After incubation at the correct temperature and time, characteristic colonies on the membrane surface are counted. Confirmatory tests are carried out where necessary. The membrane filtration method is especially useful for testing drinking water because large volumes of sample can be analyzed in a short time. The main indicating organisms of water fecal contaminations are *E. coli,* lacto-positive bacteria, coliform bacteria, thermotolerant *E. coli* types, fecal enterococcus, and *Clostridium* species. Several of the above-mentioned indicator conditions are fulfilled by *E. coli* and, to a lesser extent, by coliform bacteria and thermotolerant *E. coli* types. *E. coli* belongs to the family Enterobacteriaceae. It is a small, non-spore-forming, gram-negative, rod-shaped bacterium. It lives in the large intestine of humans and animals, where its density reaches a maximum of 10^9 cells per gram of fecal matter. These bacteria are grown at temperatures of 44–45°C on complex culture media. Usually, they are able to ferment lactose and mannitol with formation of acid and aldehyde, and they are able to produce indole from tryptophan. *E. coli* possesses two enzymes: β-galactosidase and β-glucoronidase; however, it does not form oxidase and does not hydrolyze urea. Lacto-positive bacteria are able to form colonies in aerobic conditions at 36±3°C on culture medium with lactose, with acid formation. Coliform bacteria are non-spore-forming, oxidase-negative, gram-negative, rod-shaped bacteria that are able to ferment, at 35–37°C, lactose to acids and aldehydes within 24–48 hr. Thermotolerant *E. coli* types are able to ferment lactose at 44–45°C. *Enterobacter, Citrobacter,* and *Klebsiella* also belong to this group. Coliform bacteria do not have to be directly connected with fecal contamination or with the occurrence of pathogens in drinking water. They could be present in drippings, in nutrient-rich water, in soil, and in decayed vegetable residues. These bacteria cannot occur in conditioned drinking water. Their presence in water suggests improper conditioning of water, secondary contamination, and excessive content of nutrient substances in conditioned water. These bacteria could be exploited as an indicator of the effectiveness of water conditioning. The frequency of indicative bacteria is determined to set the coliform titer—the lowest volume of water (milliliters) in which these bacteria are detected. According to ISO standards, the final Result is given as an index of coliform bacteria of *E. coli* in 100 mL of water tested. Fecal enterococci are spherical or oval, catalase-negative, gram-positive bacteria, occurring as a short chain. They can reduce 2,3,5-triphenyltetrazolic chloride to formazine and hydrolyze esculin at 44°C. They express D Lancefield antigen. The term *fecal Enterococcus* refers to those bacteria that occur in droppings of humans. Fecal enterococci that are not found in droppings of humans but occur in droppings of animals do not belong to the fecal enterococcus group. They die quickly in the external environment, faster than many pathogenic bacteria. The presence of fecal enterococcus in water can indicate fresh contamination of water; thus, it also proves the potential menace of pathogenic bacteria. Because these bacteria are resistant to drying, they could be helpful in routine control of water quality carried out after new water supply systems are built or after their repair. After *E. coli*, they could be the second fecal enterococcus that can be used as an indicative organism of contamination. Clostridial bacteria are sulfite-reducing, gram-positive, spore-forming bacilli that can survive in soil and water for a long time (even for many years). They are usually found in droppings, but they can originate from other sources. These bacteria are grown at 37°C. Breeding of these species can indicate a very "old" water contamination, after the death of all pathogenic bacteria. The spores of *Clostridium* are resistant to disinfection; thus, their presence in disinfected water may indicate a shortcoming in water conditioning. These bacteria are not recommended for the routine monitoring of water quality or detecting fresh contamination because of their length of survival. *Clostridium* might be present for a long time after the appearance of the contamination and far away from the place of this contamination. In sanitary analysis of drinking water, besides the presence of indicative bacteria, the general number of psychrophilic and mesophilic bacteria in 1 mL of water is determined. Culturing of psychrophilic bacteria is carried out on a solid agar culture medium for 72 hr at 22°C—the optimum conditions for these bacteria. A large quantity of these bacteria can indicate the inflow of organic substances to water, which creates favorable conditions for the growth of saprophytic bacteria. Culturing of mesophilic bacteria is performed on a solid agar culture medium for 24 h or 48 h at 37°C. The thermal optimum for these bacteria is the temperature of the human body, 37°C. The overall quantity of mesophilic bacteria consists of effluent bacteria and some soil bacteria. Their presence in water is a proof that contamination with domestic sewage and industrial waters has occurred.

References

1. United Nations Children's Fund (UNICEF). *Excerpt from Progress since the World Summit for Children: A Statistical Review*; UNICEF: New York, 2001.
2. U.S. Centers for Disease Control and Prevention. *Safe Water System: A Low-Cost Technology for Safe Drinking Water Fact Sheet*; World Water Forum 4 Update: Atlanta, GA, 2006.
3. *Water Treatment. Principles and Design*; Montgomery, J.M., Ed.; John Wiley and Sons: New York, 1985.
4. *Drinking Water Contaminants;* EPA 816-F-09–0004; EPA; May 2009.
5. EPA. *National Primary Drinking Water Regulations*; Code of Federal Regulations, 40 CFR Part 141; EPA; Electronic Code for Federal Regulations: June 3, 2010.

6. *Guidelines for Drinking-Water Quality,* 3rd Ed., incorporating first and second addenda; WHO: Geneva, Switzerland, 2006.
7. Kaika, M. The water framework directive: A new directive for a changing social, political and economic European Framework. Eur. Plann. Stud. **2003**, *11,* 299–316.
8. Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for community action in the field of water policy, Official Journal of the European Communities; 1.327/1: 22.12.2000.
9. Rozporzadzenie Ministra Zdrowia z dnia 29 marca 2007 r. w sprawie jakożci wody przeznaczonej do spozycia przez ludzi; Dziennik Ustaw Nr 61; Poz 417.
10. Synowiecki, J. Skladniki mineralne (Mineral components). In *Chemia Żywności* (*Food Chemistry*); Sikorski, Z., Ed.; WN-T: Warsaw, Poland, 2000 (in Polish).
11. *Kompendium wiedzy o żywności, żywieniu i zdrowiu;* Gawecki, J., Mossor-Pietraszewska, T., Eds.; PWN: Warsaw, Poland, 2004 (in Polish).
12. Biziuk, M.; Kuczyńska, J. Mineral components in food— Analytical implications (Ch. 1). In *Mineral Components in Foods;* Szefer, P., Nriagu, J., Eds.; CRC Press: Boca Raton, FL, 2007; 1–31.
13. Eichler, W. *Gift in unserer Nahrung*; Kilda: Greven, Germany, 1982.
14. Bull, R.J. Carcinogenic and mutagenic properties of chemicals in drinking water. Sci. Tot. Environ. **1998**, *47,* 385–413.
15. Hartwig, A. Carcinogenicity of metal compounds: Possible role of DNA repair inhibition. Toxicol. Lett. **1998**, *102,* 235–239.
16. Kasprzak, K.S. Oxidative DNA and protein damage in metal-induced toxicity and carcinogenesis. Free Radic. Biol. Med. **2002**, *32,* 958–967.
17. Pourahmad, J.; O'Brien, P.J.; Jokar, F.; Daraei, B. Carcinogenic metal induced sites of reactive oxygen species formation in hepatocytes. Toxicol. In Vitro **2003**, *17,* 803–810.
18. Schwerdtle, T.; Walter, I.; Hartwig, A. Arsenite and its biomethylated metabolites interfere with the formation and repair of stable BPDE-induced DNA adducts in human cells and impair XPAzf and Fpg. DNA Repair **2003**, 1449–1463.
19. Tully, D.B.; Collins, B.J.; Overstreet, J.D.; Smith, C.S.; Dinse, G.E.; Mumtaz, M.M.; Chapin, R.E. Effects of arsenic, cadmium, chromium, and lead on gene expression regulated by a battery of 13 different promoters in recombinant HepG2 cells. Toxicol. Appl. Pharmacol. **2000**, *168,* 79–90.
20. Connell, D.W. *Bioaccumulation of Xenobiotic Compounds;* CRC: Boca Raton, FL, 1990.
21. Zukowska J.; Biziuk M. Methodological evaluation of method for dietary metal intake. J. Food Sci. **2008**, 73, 21–29.
22. Villaescusa, I.; Bollinger, J.C. Arsenic in drinking water: Sources, occurrence and health effects (a review). Rev. Environ. Sci. Biotechnol. **2008**, *7,* 307–323.
23. Ferguson, L.R. Natural, man-made mutagens and carcinogens in the human diet. Mutat. Res. **1999**, *443,* 1–10.
24. Moore, J.W. *Inorganic Contaminants of Surface Water. Research and Monitoring Priorities*; Springer: New York, 1991.
25. Nriagu, J.O.; Pacyna, J. Quantitative assessment of worldwide contamination of air, water and soils by trace metals. Nature **1988**, *333*, 134–139.
26. Rojas, E.; Herrera, L.A.; Poirier, L.A.; Ostrosky-Wegman, P. Are metals dietary carcinogens. Mutat. Res. **1999**, *443,* 157–181.
27. Kabata-Pendias, A.; Pendias, H. *Trace Elements in Soil and Plants,* 3rd Ed.; CRC Press: Boca Raton, FL, 2001.
28. Kabata-Pendias, A.; Mukherjee, A.B. *Trace Elements from Soil to Human;* Springer: Berlin, 2007.
29. Waalkes, M.P.; Misra, R.R. Cadmium carcinogenicity and genotoxicity. In *Toxicology of Metals*; Chang, L.W., Ed.; CRC Press: Boca Raton, FL, 1996; 231–243.
30. Cohen, M.D.; Kargacin, B.; Klein, C.B.; Costa, M. Mechanisms of chromium carcinogenicity and toxicity. Crit. Rev. Toxicol. **1993**, *23,* 255–281.

31. McLaughlin, M.J.; Parker, D.R.; Clarke, J.M. Metals and micronutrients—Food safety issues. Field Crops Res. **1999**, *60,* 143–163.
32. Pershagen, G. The carcinogenity of arsenic. Environ. Health Perspect. **1981**, *40,* 93–100.
33. Pocock, S.J.; Delves, H.T.; Ashby, D.; Shaper, A.G.; Clayton, B.E. Blood cadmium concentration in the general population of British middle-aged men. Hum. Toxicol. **1988**, *7,* 95–103.
34. Verougstraete, V.; Lison, D.; Hotz, P. Cadmium, lung a prostate cancer: A systematic review of recent epidemiological data. J. Toxicol. Environ. Health, Part B **2003**, *6*, 227–255.
35. Coogan, T.P.; Latta, D.M.; Snow, E.T.; Costa, M. Toxicity and carcinogenicity of nickel compounds. CRC Crit. Rev. Toxicol. **1989**, *19,* 341–384.
36. Soo, Y.O.; Chow, K.M.; Lam, C.W.; Lai, F.M.; Szeto, C.C.; Chan, M.H.; Li, P.K. A whitened face woman with nephrotic syndrome. Am. J. Kidney Dis. **2003**, *41*, 250–253.
37. Strenzke, N.; Grabbe, J.; Plath, K.E.; Rohwer, J.; Wolff, H.H.; Gibbs, B.F. Mercuric chloride enhances immunoglobulin E-dependent mediator release from human basophils. Toxicol. Appl. Pharmacol. **2001**, *174,* 257–263.
38. Ellingsen, D.G.; Bast-Pettersen, R.; Efskind, J.; Thomassen, Y. Neuropsychological effects of low mercury vapor exposure in chloralkali workers. Neurotoxicology **2001**, *22*, 249–258.
39. Godfrey, M.; Wojeik, D.; Cheryl, K. Apolipoprotein-E as a potential biomarker for mercury neurotoxicity. J. Alzheimer's Dis. **2003**, *5*, 189–195.
40. Mutter, J.; Naumann, J.; Sadaghiani, C.; Schneider, R.; Walach, H. Alzheimer disease: Mercury as a pathogenetic factor and apolipoprotein-E as a moderator. Neuroendocrinol. Lett. **2004**, *25*, 275–283.
41. Leonard, A.; Jacquet, P. Mutagenicity and teratogenicity of mercury compounds. Mutat. Res. **1983**, *114,* 1–18.
42. Zahir, F.; Rizwi, S.J.; Haq, S.K.; Khan, R.H. Low dose mercury toxicity and human health. Environ. Toxicol. Pharmacol. **2005**, *20,* 351–360.
43. Duker, A.A.; Carranza, E.J.M.; Hale, M. Arsenic geochemistry and health. Environ. Int. **2005**, *31,* 631–641.
44. Robson, M. Methodologies for assessing exposures to metals: Human host factors. Ecotoxicol. Environ. Saf. **2003**, *56,* 104–109.
45. Gochfeld, M. Cases of mercury exposure, bioavailability, and absorption. Ecotoxicol. Environ. Saf. **2003**, *56,* 174179.
46. Apostoli, P. Elements in environmental and occupational medicine. J. Chromatogr. B **2002**, *778,* 63–97.
47. Gebel, T. Genotoxicity of arsenical compounds. Int. J. Hyg. Environ. Health **2001**, *203,* 249–262.
48. Egan, S.K.; Tao, S.S-H.; Pennington, J.A.T.; Bolger, P.M. U.S. Food and Drug Administration's Total Diet Study: Intake of nutritional and toxic elements, 1991–1996. Food Addit. Contam. **2002**, *19,* 103–125.
49. Lee, H-S.; Cho, Y-H.; Park, S-O.; Kye, S-H.; Kim, B-H.; Hahm, T-S.; Kim, M.; Lee, J.O.; Kim, C-I. Dietary exposure of the Korean population to arsenic, cadmium, lead and mercury. J. Food Compos. Anal. **2006**, *19,* S31–S37.
50. Robberecht, H.; Van Cauwenbergh, R.; Bosscher, D.; Cornelis, R.; Deelstra, H. Daily dietary total arsenic intake in Belgium using duplicate portion sampling and elemental content of various foodstuffs. Eur. Food Res. Technol. **2002**, *214*, 27–32.
51. Ozsvath, D.I. Fluoride and environmental health: A review. Rev. Environ. Sci. Biotechnol. **2009**, *8*, 59–79.
52. Webber, J.S.; Syrotynski, S.; King, M.V. Asbestos-contaminated drinking water: Its impact on household air. Environ. Res. **1988**, *46,* 153–167.
53. Nemerow, N.L. *Stream, Lake, Estuary and Ocean Pollution;* Van Nostrand Reinhold Co.: New York, 1985.
54. *Handbook of Water Analysis;* Nollett, L.M.L., Ed.; Marcel Dekker: New York, 2000.

55. Fresenius, W.; Quentin, K.E.; Schneider, W., Eds. *Water Analysis;* Springer-Verlag: Berlin, Germany, 1988.
56. Leenheer, J.A. *Water Analysis;* Academic Press: Oxford, U.K., 1984.
57. Hewitt, C.N., Ed. *Instrumental Analysis of Pollutants;* Elsevier: London, 1991.
58. Hunt, D.T.E.; Wilson, A.L., Eds. *The Chemical Analysis of Water;* Royal Society of Chemistry: London, 1988.
59. Willard, H.H.; Merritt, L.L.; Dean, J.A.; Settle, F.A., Eds. *Instrumental Methods of Analysis;* Wadsworth: Belmont, CA, 1988.
60. Taylor, L.R.; Pap, R.B.; Pollard, B.D. *Instrumental Methods for Determining of Elements;* VCH: New York, 1994.
61. Bersier, P.M.; Howell, J.; Bruntlett, C. Advanced electroanalytical techniques versus atomic absorption spectrometry, inductively coupled plasma atomic emission spectrometry and inductively coupled plasma mass spectrometry in environmental analysis. Analyst **1994**, *119*, 219–233.
62. Wang, J. *Stripping Analysis. Principles, Instrumentation and Applications*; VCH: Deerfield Beach, FL, 1985.
63. Hellmann, H. *Analysis of Surface Waters*; Ellis Horwood: New York, 1987.
64. Biziuk, M.; Michalska, M. Analysis of drinking water. In *Methods of Analysis of Food Components and Additives;* Otles, S., Ed.; CRC Taylor & Francis: Boca Raton, FL 2005; 31–57.
65. de Kruijf, H.A.M.; Kool, J.H., Eds. *Organic Micropollutants in Drinking Water and Health;* Elsevier: Amsterdam, the Netherlands, 1985.
66. Biziuk, M.; Przyjazny, A. Methods of isolation and determination of volatile organohalogen compounds in natural and treated waters. J. Chromatogr. A **1996**, *733,* 417–448.
67. Biziuk, M.; Przyjazny, A.; Czerwiriski, J.; Wiergowski, M. Occurrence and determination of pesticides in natural and treated waters. J. Chromatogr. A **1996**, *754,* 103–123.
68. Moore, J.W.; Ramamoorthy, S. *Organic Chemicals in Natural Waters. Applied Monitoring and Impact Assessment*; Springer Verlag: New York, 1984.
69. Bruchet, A.; Legrand, M.F.; Anselme, C.; Mallevialle, J. Perspectives actuelles pour l'analyse des composés organiques volatils et non volatils dans les eaux naturelles ou en cours de traitement. Water Supply **1989**, *7,* 77–84.
70. Coleman, W.E.; Melton, R.G.; Kopfler, F.C.; Barone, K.A.; Aurand, T.A.; Jellison, M.G. Identification of organic compounds in a mutagenic extract of a surface drinking water by computerized gas chromatography-mass spectrometry system. Environ. Sci. Technol. **1980**, *14,* 576–588.
71. Biziuk, M. Gas chromatography by direct aqueous injection in environmental analysis. In *Encyclopedia of Analytical Chemistry;* Mayers, R.A., Ed.; Wiley: Chichester, U.K., 2000; Vol. 3, 2549–2587.
72. de Leer, W.B. *Aqueous Chlorination Products. The Origin of Organochlorine Compounds in Drinking and Surface Waters*; Delft University Press: Delft, the Netherlands, 1987.
73. Jolley, R.L.; Cotruvo, J.A.; Cummings, R.B.; Matice, J.S.; Jacobs, V.A., Eds. *Water Chlorination. Environmental Impact and Health Effects;* Ann Arbor Sc. Pub. Inc.: Ann Arbor, MI, 1980.
74. Rebhun, M.; Manka, J.; Zilberman, A. Trihalomethane formation in high-bromide Lake Galilee water. J. AWWA **1988**, *June,* 84–89.
75. Koch, B.; Krasner, S.W. Occurrence of disinfection byproducts in a distribution system. Proc. Water Qual. Technol. Conf. AWWA, Los Angeles, June 18–22, 1989; J. AWWA: Denver, 1989; 1203–1230.
76. Singer, P.C. DBPs in drinking water: Additional scientific and policy consideration for public health protection. J. AWWA **2006**, *October*, 73–80.
77. Peleg, M.; Yungster, H.; Shuval, H. *Developments in Arid Zone Ecology and Environmental Quality;* Balaban ISS: Philadelphia, PA, 1981; 271 pp.
78. Hirata, T.; Nakasugi, O.; Yoshioka, M.; Sumi, K. Groundwater pollution by volatile organochlorines in Japan and related phenomena in subsurface environment. Water Sci. Technol. **1992**, *25*, 9–16.

79. Stevens, A.A.; Dressman, R.C.; Sorrell, R.K.; Brass, H.J. Organic halogen measurements: Current uses and future prospects. J. AWWA **1985**, *April,* 146–152.
80. Beyer, A.; Biziuk, M. Environmental fate and global distribution of polychlorinated biphenyls Rev. Environ. Contam. Toxicol. **2009**, *201,* 137–158.
81. Dias-Cruz, M.S.; Llorca, M.; Barcelo, D. Organic UV filters and their photodegradates, metabolites and disinfection byproducts in the aquatic environment Trends Anal. Chem. **2008**, *27,* 873–887.
82. Eichhorn, P.; Knepper, T.P.; Ventura, F.; Diaz, A. The behavior of polar aromatic sulfonates during drinking water production: A case study on sulfophenyl carboxylates in two European waterworks. Water Res. **2002**, *36*, 2179–2186.
83. Biziuk, M.; Bartoszek, A. Environmental contamination of food. In *Carcinogenic and Anticarcinogenic Food Components;* Baer-Dubowska, W., Bartoszek, A., Malejka- Giganti, D.; CRC: Boca Raton, FL, 2006; 113–136.
84. Quevauviller, P.; Borchers, U.; Gawlik, B.M. Coordinating links among research, standardisation and policy in support of water directive chemical monitoring requirements. J. Environ. Monit. **2007**, *9*, 915–923.
85. Jakubowska, N.; Zygmunt, B.; Polkowska, Z.; Zabiegala, B.; Namiessnik, J. Sample preparation for gas chromatographic determination of halogenated volatile organic compounds in environmental and biological samples. J. Chromatogr. A **2009**, *1216,* 422–441.
86. Demeestere, K.; Dewulf, J.; De Witte, B.; Van Langenhove, H.Sample preparation for the analysis of volatile organic compounds in air and water matrices. J. Chromatogr. A **2007**, *1153,* 130–144.
87. Ras, M.R.; Borrull, F.; Marce, R.M. Sampling and preconcentration techniques for determination of volatile organic compounds in air samples. Trends Anal. Chem. **2009**, *28,* 347–361.
88. Tobiszewski, M.; Mechliriska, A.; Zygmunt, B.; Namiesnik, J. Green analytical chemistry in sample preparation for determination of trace organic pollutants. Trends Anal. Chem. **2009**, *28,* 943–951.
89. Namiesnik, J.; Górecki, T.; Biziuk, M. Isolation and preconcentration of volatile organic compounds from water. Anal. Chem. Acta **1990**, 237, 1–60.
90. Pawliszyn, J. *Solid Phase Microextraction—Theory and Practice*; Wiley: New York, 1997.
91. Pawliszyn, J. *Application of Solide Phase Microextraction*; The Royal Society of Chemistry: Cambridge, 1999.
92. Biziuk, M.; Czerwiriski, J. *Determination of dichloromethane, trichloromethane, tetrachloromethane, bromodichloromethane, dibromochloromethane, tribromomethane, trichloroethylene, tetrachloroethylene, 1,1,1-trichloroethane, I, 1,2,2-tetrachloroethane in water by gas chromatography using direct injection of the sample;* Polish Standard Method-PN-C-04549–1; Polish Committee for Standardization: Warsaw, Poland, 1998.
93. Astel, A.; Biziuk, M.; Przyjazny, A.; Namiesnik, J. Chemometrics in monitoring spatial and temporal variations in drinking water. Water Res. **2006**, *40,* 1706–1716.
94. Astel, A.; Astel, K.; Biziuk, M.; Namiesnik, J. Classification of drinking water samples using the Chernoff's faces visualization approach. Pol. J. Environ. Stud. **2006**, *15,* 691–697.
95. Biziuk, M.; Polkowska, Z.; Gorlo, D.; Janicki, W.; Namiessnik, J. Determination of volatile organic compounds in water intakes and tap water by purge and trap and direct aqueous injection-electron capture detection techniques. Chem. Anal. (Warsaw) **1995**, *40,* 299–307.
96. Biziuk, M.; Namiessnik, J.; Czerwinsski, J.; Gorlo, D.; Makuch, B.; Janicki, W. Occurrence and determination of organic pollutants in tap and surface waters of the Gdanssk district. J. Chromatogr. A **1996**, *733*, 171–184.
97. Starostin, L.; Witkiewicz, Z. Environmental water samples preparation for chemical analysis. Chem. Anal. (Warsaw) **1994**, *39,* 263–279.
98. Hench, K.R.; Bissonnette, G.K.; Sexstone, A.J.; Coleman, J.G.; Garbutt, K.; Skousen, J.G. Fate of physical, chemical, and microbial contaminants in domestic wastewater following treatment by small constructed wetlands. Water Res. **2003**, *37,* 921–927.

99. Golas, I.; Filipkowska, Z.; Lewandowska, D. Zmyslowska, I. Potentially pathogenic bacteria from the family *Enterobacteriaceae, Pseudomonas* sp. and *Aeromonas* sp. in waters designated for drinking and household purposes. Pol. J. Environ. Stud. **2002**, *11,* 325–330.
100. Nawrocki, J.; Bilozor, S. Uzdatnianie wody. *Procesy Chemiczne i Biologiczne;* PWN: Warsaw, Poland, 2000 (in Polish).

66

Water: Surface

Victor de Vlaming

Introduction

In numerous aquatic ecosystems across the United States losses of biodiversity and significant population declines of multiple species have been documented over the past 40 years. These aquatic ecosystems are imperiled due to human activities such as habitat destruction, damming and diversion of waters, introduction of exotic species, and release of toxic concentrations of chemicals, including pesticides, into surface waters.

Pesticide Use

Quantities of pesticides used in the United States have increased approximately 50-fold since the 1960s such that more than a billion pounds of pesticide active ingredients are used each year.[1] Hundreds of different pesticides have been developed for application in agricultural and other settings to control weeds, insects, fungus, and other pests. The characteristics of pesticides have evolved over the past 30 years from chemicals such as the organochlorines (OC) that are very persistent in the environment but tend to have relatively low toxicity, to chemicals such as organophosphorus (OP) and carbamate insecticides that are toxic at very low concentrations (e.g., parts per trillion) but generally are not very persistent. The majority of pesticide use is agricultural (70%–80% of total), but there is also use in urban areas for gardens, lawns, homes, and buildings, as well as in forestry, along roads and railways, and in various industrial and commercial situations.

Pesticide Transport and Patterns of Occurrence

Unfortunately pesticides do not always remain where they are applied. Off-site movement occurs irrespective of applications made according to label instructions. Off-site movement can occur by aerial drift of sprays, by evaporation, in storm and irrigation water runoff, and by seepage. Surface waters are vulnerable to pesticide contamination because most agricultural and urban areas drain into streams and rivers. Since pesticides are designed to be lethal to organisms, they pose a significant risk to biota when they enter aquatic ecosystems. Many of the currently used pesticides are so lethal that runoff of less than 1% of the quantities applied in a watershed into surface waters can have profound effects on aquatic biota.

The most extensive program for monitoring surface waters for chemical contaminants in the United States is the National Water-Quality Assessment (NAWQA) Program of the U.S. Geological Survey. Initiated in 1991, the focus in NAWQA has been on major watersheds distributed throughout the United States, encompassing 60%–70% of national water use. Pesticides have been one of NAWQA's top priorities. Much of the information summarized below was collected in the NAWQA program.[1–3]

Pesticides have been detected in every region of the United States where surface waters were analyzed for these chemicals. Although no individual study analyzed for every pesticide, a wide variety of pesticides including insecticides, herbicides, and fungicides have been identified in surface waters throughout the United States. The distribution of pesticides in surface waters generally follows geographic and seasonal agricultural patterns and also the influence of urban areas. That is, frequency of detections and highest concentrations are recorded in streams and rivers where agriculture is a major land use and pesticide use is intense. In most agricultural areas, the highest concentrations of pesticides occur as seasonal pulses, with duration of a few weeks to several months. Frequency of occurrence and highest concentrations also are associated with seasonal application patterns, being greatest coincident with or after applications. In urban-dominated streams, seasonal patterns are less obvious, pesticide concentrations being elevated for longer periods. In these urban streams, insecticides are detected at higher frequencies and concentrations than in streams draining agricultural areas. The largest areas where high quantities of pesticides are applied to crops occur in California, Florida, the Midwest, the lower Mississippi River Valley, and the coastal areas of the Southeast.

Herbicides are detected more frequently, and at the highest concentrations, than other pesticides in surface waters. Considering data collected from across the United States, herbicides occurring most frequently and at the highest concentrations are atrazine, simazine, metolachlor, prometon, DEA, alachlor, and cyanozine. The insecticides measured most frequently and at the highest concentrations are diazinon, chlorpyrifos, carbaryl, carbofuran, and malathion.

Pesticides are encountered most often and at elevated concentrations in streams and rivers draining agricultural areas just prior to and during the growing season. Also, a greater number of pesticides occur in surface waters coincident with these periods. These mixtures of pesticides have a potential for additive and synergistic adverse impacts on aquatic ecosystem health. An exposure pattern that is developing is one of long-term exposure to relatively low concentrations of pesticide mixtures punctuated with seasonal pulses of high concentrations, the effects of which are not currently known.

Off-site movement is a critical issue with regard to pesticide contamination of aquatic ecosystems. The relationship between quantities of pesticides applied (amount per unit area and total area of application) and detection frequency, as well as concentrations in surface waters was stated above. While this is a general principle, it does not apply to all pesticides and situations. Some pesticides, which are used rather extensively in agriculture, are seldom detected in surface waters. This low detection rate in surface waters relates to physical/chemical and/or degradation properties. Some such properties can result in pesticides adsorbing to particles (organic or soil) that may reduce off-site movement while different properties of other pesticides favor rapid degradation.

One might conclude that use of pesticides with a lower potential for off-site movement would reduce risk of impacting aquatic ecosystem health. While there is some truth in this idea, such physical/chemical properties often render pesticides more persistent. Furthermore, organic and soil particles to which pesticides are adsorbed can be transported by erosion (associated with rainfall, irrigation, or wind) into streams and rivers. Pesticides adsorbed to particles can settle into surface water sediments and have deleterious effects on bottom-dwelling organisms. Such physical/chemical properties also tend to result in bioaccumulation by aquatic organisms. These pesticides can bioaccumulate in aquatic species to levels that are detrimental and/or biomagnified to adverse levels in the food chain as contaminated organisms are eaten.

Examination of fish and bivalve tissues reveal that they are being exposed to a variety of bioaccumulable pesticides in both agricultural- and urban-dominated waterways. Residues of some pesticides in fish tissues can be such that they are harmful to human health. Organochlorine (e.g., DDT, chlordane), pyrethroid, and other hydrophobic insecticides are examples of pesticides that adsorb to organic materials, are persistent, bioaccumulate, and biomagnify. Unfortunately, there is a paucity of information on what pesticide tissue residue levels are deleterious to organisms or to other species that eat them.

Effects of Pesticides

The occurrence of pesticides in surface waters of the United States is widespread. What is the significance of this phenomenon? Adverse effects of chemicals, including pesticides, are determined by concentration, as well as by duration and frequency of exposure. The federal Clean Water Act (CWA) was enacted to protect human and aquatic organism health, requiring that no chemical can occur in surface waters at toxic concentrations. CWA requirements are implemented through enforceable water quality standards for specific chemicals and toxicity. Water quality standards and criteria have been established by various agencies for only a few pesticides. These standards and criteria are an estimate of a chemical concentration in water below which detrimental effects are not expected to occur. Comparing concentrations of a chemical in surface water to such standards and criteria provides an indication of potential antagonistic impacts on aquatic biota.

Standards for human health apply to treated drinking water supplied by community agencies. Therefore, the standards do not apply directly to most surface waters. While these standards do not pertain to concentrations of pesticides in surface waters, they do afford a benchmark to which measured pesticide concentrations can be compared. Chemical analyses of surface water samples do not include all pesticides; however, few pesticides included in analyses were detected at concentrations exceeding any drinking water standards. The pesticides most often exceeding standards are the triazine and acetanilide herbicides, atrazine, alachlor, cyanazine, and simizine.[1] Some, but not all, treatments of water to be used for drinking destroy or remove these herbicides.

For the more than 120 pesticides detected in surface waters there exist only 13 U.S. Environmental Protection Agency criteria developed for the protection of aquatic ecosystem health. For most currently used insecticides and for all herbicides there are no criteria for the protection of aquatic life. Canada has a larger number of aquatic life criteria, which are more stringent than U.S. criteria. U.S. pesticide aquatic life criteria are commonly exceeded in streams and rivers collecting from agricultural lands and/ or urban areas. Aquatic life criteria of four OP insecticides, azinphos-methyl, chlorpyrifos, diazinon, and malathion are the most frequently exceeded by concentrations in surface waters. Data collected from across the United States indicate that azinphos-methyl and chlorpyrifos exceed aquatic life criteria for more days per year than other insecticides monitored.[1,3] Major concerns regarding OP insecticides are that they are toxic to aquatic species at very low concentrations, different OPs repeatedly co-occur in surface waters, and their toxicity is additive.

In some regions of the country, where their use is high, carbamate insecticides are threats to aquatic biota. More than 20 years after being banned, OC insecticides continue to be detected in surface waters and sediments at concentrations that exceed aquatic life criteria. Especially in streams and rivers draining agricultural areas, but also in urbandominated streams, two or more pesticides often cooccur at concentrations that exceed their respective aquatic life criteria. Several studies from across the country reported high occurrences of diseased, deformed, and highly parasitized fish, as well as fish with a high incidence of tumors in surface waters where pesticide concentrations exceed aquatic life criteria and/or are elevated.

Measuring pesticide concentrations in surface waters does not furnish direct information of bioavailability (the percentage of the analytically measured amount of a chemical that produces toxic effects) or toxicity to aquatic biota. Toxicity testing of surface waters provides a direct measure of capacity of these waters to support healthy aquatic organisms. Standardized toxicity tests with aquatic species are available that measure lethal and sublethal (e.g., inhibition of reproduction, growth, etc.), effects. As a diagnostic tool for assessing water quality, toxicity testing has several merits. Toxicity tests afford an integrative measure of adverse effects of chemicals on organism health and viability, as well as the bioavailabity of chemicals.[4] Also, results of these toxicity tests have been reliable predictors of impacts on aquatic ecosystem biota. Chemical analyses of water cannot provide such information.

Statewide monitoring programs in California have disclosed that, on a seasonal basis, agricultural- and urban-influenced streams and rivers are lethal to test species.[4,5] Complex toxicological, chemical, and physical procedures (toxicity identification evaluation-TIE) that specifically identify the chemical(s) causing toxicity demonstrated that the pesticides most commonly responsible for the surface water toxicity are diazinon and chlorpyrifos. Carbofuran, malathion, carbaryl, methyl-parathion, thiobencarb, diuron, and molinate also have been shown to be causes of toxicity in California's surface waters. These surface water toxicity testing programs are not common in other regions of the United States. Such surface water toxicity is likely to occur in most urban streams and in surface waters collecting runoff from areas where agriculture is the predominant land use and where pesticides are used intensively.

Gaps in Knowledge

Pesticide adverse impacts on aquatic ecosystem health throughout the United States are most likely underestimated for several reasons. 1) The number of U.S. streams and rivers thoroughly monitored is very limited so that the distribution and extent of pesticide contamination is unknown. 2) Most investigations have been incomplete for one or more reasons, including too few sampling sites, sample collection was in frequent, and study duration was abbreviated. 3) None of these monitoring investigations included the complete range of pesticides that could impact aquatic biota. 4) Analytical detection limits for pesticides have been a problem in assessing detrimental effects on aquatic ecosystem health. In most monitoring projects analytical detection limits for many pesticides were higher than known toxic effects on aquatic species, as well as above aquatic life criteria. 5) Sublethal and delayed effects of pesticides, including bioaccumulation and biomagnification responses, generally are not evaluated. For example, several pesticides, including alachlor, atrazine, 2,4-D, metribuzin, trifluralin, aldicarb, carbaryl, parathion, some pyrethroids, benomyl, mancozeb, maneb, zineb, and ziram, commonly used in U.S. agriculture, have been shown to disrupt endocrine systems in some aquatic species.[6,7] Existing aquatic species toxicity screening procedures are inadequate for assessing endocrine disruption. 6) Seldom are indirect effects considered. For example, direct adverse effects on zooplankton during a period when they are critical food for larval fish could indirectly impact fish populations. 7) As stated above, aquatic organisms are exposed to pesticide mixtures in many watersheds across the United States. Assessments of impacts infrequently involve analysis of exposures to multiple chemicals.

Pesticides, especially insecticides, are having widespread impacts on surface water quality and aquatic ecosystem health throughout the United States. To reduce risks of pesticide impacts on aquatic ecosystem health, measures should be identified, developed, and implemented to eliminate or reduce off-site movement of these chemicals. More extensive and thorough monitoring of pesticides and of toxicity caused by these chemicals is advisable to assess the extent of pesticide-caused water quality degradation, as well as the effectiveness of remediation projects.

References

1. Larson, S.J.; Capel, P.D.; Majewski, M.S. *Pesticides in Surface Waters: Distribution, Trends, and Governing Factors;* Ann Arbor Press, Inc.: Chelsea, MI, 1997; 373.
2. Larson, S.J.; Gilliom, R.J.; Capel, P.D. Pesticides in Streams of the United States—Initial Results from National Water-Quality Assessment Program. *Water Resources Investigations Report 98–4222;* U.S. Geological Survey: Sacramento, CA, 1999; 92.
3. Gilliom, R.J.; Barbash, J.E.; Kolpin, D.W.; Larson, S.J. Testing water quality for pesticide pollution. Environ. Sci. Technol. **1999**, *33,* 164A–169A.

4. de Vlaming, V.; Connor, V.; DiGiorgio, C.; Bailey, H.C.; Deanovic, L.A.; Hinton, D.E. Application of whole effluent toxicity test procedures to ambient water quality assessment. Environ. Toxicol. Chem. **2000**, *19,* 42–62.
5. Kegley, S.; Neumeister, L. *Disrupting the Balance: Ecological Impacts of Pesticides in California,* Californians for Pesticide Reform, San Francisco, 1999; 99.
6. *Environmental Endocrine Disruptors: A Handbook of Property Data*; Lawrence, H.K., Ed.; John Wiley and Sons, Inc.: New York, 1997; 1232.
7. Moore, A.; Waring, C.P. Sublethal effects of the pesticide diazinon on olfactory function in mature male Atlantic salmon Parr. J. Fish Biol. **1996**, *48,* 758–775.

67

Wetlands

Ralph W. Tiner

Introduction

Wetland is a universal term used to describe the collection of flooded or saturated environments that have been referred to as marshes, swamps, bogs, fens, salinas, pocosins, mangroves, wet meadows, sumplands, salt flats, varzea forests, igapo forests, bottomlands, sedgelands, moors, mires, potholes, sloughs, mangals, palm oases, playas, muskegs, and other regional and local names. It has been defined as a basis for inventorying these natural resources, for conducting scientific studies, and, in some countries, for regulating uses of these areas. Given that wetlands include a diverse assemblage of ecosystems, classification schemes have been developed to separate and describe these different systems and to group similar habitats. Wetlands provide a number of functions that are considered valuable to society (e.g., surface water storage to minimize flood damages, sediment retention and nutrient transformation to improve water quality, shoreline stabilization, streamflow maintenance, and provision of vital habitat for fish, shellfish, wildlife, and plants that yield food and fiber for people). Because of these values and the widespread recognition of wetlands as important natural resources, numerous wetland definitions and classification systems have been developed to inventory these resources around the globe. The purpose of this entry is to provide readers with an understanding of what wetlands are (wetland definition), how they vary globally (wetland types), and their extent as determined by various inventories. This entry should serve as a starting point for learning about wetlands, with the listed references being sources of more detailed information.

Wetland Definitions

Wetlands are aquatic to semiaquatic ecosystems where permanent or periodic inundation or prolonged waterlogging creates conditions favoring the establishment of aquatic life. Wetlands are often located between land and water and have, therefore, been referred to as ecotones (i.e., transitional communities). However, many wetlands are not ecotones between land and water, since they are not associated with a river, lake, estuary, or stream.[1] Wetlands may derive water from many sources, including groundwater, river overflow, surface water runoff, precipitation, snowmelt, tides, melting permafrost, and seepage from impoundments or irrigation projects.

While the term "wetland" has many definitions, all definitions have common elements (see Table 1; some definitions even include deepwater habitats.). The presence of water in wetlands may be permanent or temporary. Their water may be salty or fresh. Wetlands may be natural habitats or artificially created.

TABLE 1 Examples of Wetland Definitions Used for Inventories

Country/Organization	Wetland Definition [Source]
International/Ramsar	"areas of marsh, fen, peatland, or water, whether natural or artificial, permanent or temporary, with water that is static or flowing, fresh, brackish, or salt, including areas of marine water the depth of which at low tide does not exceed 6 m ……. may incorporate riparian and coastal zone adjacent to wetlands, and islands or bodies of marine water deeper than 6 m at low tide lying within the wetlands."[2]
Australia	"areas of seasonally, intermittently, or permanently waterlogged soils or inundated land, whether natural or artificial, fresh or saline, e.g., waterlogged soils, ponds, billabongs, lakes, swamps, tidal flats, estuaries, rivers, and their tributaries.[12]
Canada	"land that is saturated with water long enough to promote wetland or aquatic processes as indicated by poorly drained soils, hydrophytic vegetation, and various kinds of biological activity which are adapted to a wet environment."[3]
U.S.	"lands transitional between terrestrial and aquatic systems where the water table is usually at or near the surface or the land is covered by shallow water." Wetland attributes include hydrophytic vegetation, undrained hydric soil, or saturated or flooded substrates.[4]

Source: Photos courtesy of U.S. Fish and Wildlife Service.

They range from shallow water environments to temporarily wet (i.e., flooded or saturated) areas. All are wet long enough and often enough to, at least, periodically support hydrophytic vegetation and other aquatic life (including anaerobic microbes), to create hydric soils or substrates, and to activate biogeochemical processes associated with wet environments.

Wetland Types

Differences in climate, soils, vegetation, hydrology, water chemistry, nutrient availability, and other factors have led to the formation of a multitude of wetland types around the globe. In general, wetlands are characterized by their hydrology (e.g., tidal vs. nontidal, inundation vs. soil saturation, frequency and duration of wetness), the presence or absence of vegetation (vegetated vs. nonvegetated), the type of vegetation (forested or treed, shrub, emergent, or aquatic bed), and soil type (e.g., organic vs. inorganic, peatland vs. nonpeatland). Table 2 presents brief descriptions of some North American types and Figure 1 shows examples of vegetated wetlands.

Various countries have devised classification systems for describing differences among their wetlands and for categorizing wetlands for natural resources inventories. Scientists have created systems to organize certain wetlands into meaningful groups for analysis and management (e.g., peatland classifications; see Tiner[1] for details). In 1998, the Ramsar Convention Bureau published a multinational classification system to provide consistency for inventorying wetlands and designating wetlands of international importance.[2] This system includes 11 types of marine or coastal wetlands (i.e., shallow water and intertidal habitats: permanent shallow marine waters; marine subtidal aquatic beds; coral reefs; rocky marine shores; sand; shingle or pebble shores; estuarine waters; intertidal mud, sand, or salt flats; intertidal marshes; intertidal forests; coastal brackish/saline lagoons; coastal freshwater lagoons). This system also includes 19 inland wetland types (i.e., permanently flooded aquatic habitats to intermittently flooded sites are represented: permanent inland deltas; permanent rivers/streams/creeks; seasonal, intermittent, or irregular rivers/streams/creeks; permanent freshwater lakes; seasonal or intermittent freshwater lakes; seasonal or intermittent saline/brackish/alkaline lakes and flats; permanent saline/brackish/alkaline marshes and pools; seasonal or intermittent saline/brackish/alkaline marshes and pools; permanent freshwater marshes and pools; seasonal or intermittent freshwater marshes and pools; nonforested peat-lands; alpine wetlands including meadows and temporary snowmelt waters; tundra wetlands; shrubby-dominated wetlands; freshwater tree-dominated wetlands on inorganic soils; forested peatlands; freshwater springs and oases; geothermal wetlands; and subterranean karst and cave hydrological systems). Lastly, nine man-made wetland types (aquaculture ponds; ponds; irrigated land

TABLE 2 Brief Nontechnical Descriptions of Some Wetland Types in North America

Wetland Type	General Description
Marsh	Herb-dominated wetland with standing water through all or most of the year, often with organic (muck) soils
Tidal marsh	Herb-dominated wetland subject to periodic tidal flooding
Salt marsh	Herb-dominated wetland occurring on saline soils, typically in estuaries and interior arid regions
Swamp	Wetland dominated by woody vegetation and usually wet for extended periods during the growing season
Mangrove swamp (Mangal)	Tidal swamp dominated by mangrove species
Peatland, mire, moor, or muskeg	Peat-dominated wetland
Bog	Nutrient-poor peatland, typically characterized by ericaceous shrubs, other woody species, and peat mosses
Fen	More or less nutrient-rich peatland, often represented by sedges and/or calciphilous herbs and woody species
Wet meadow	Herb-dominated wetland that may be seasonally flooded or saturated for extended periods, often with mineral hydric soils
Bottomland	Riverside or streamside wetland, usually on floodplain
Flatwood	Forested wetland with poorly drained mineral hydric soils located on broad flat terrain of interstream divides, common on coastal plains and glaciolacustrine plains
Farmed wetland	Wetland cultivated for rice, cranberries, sugar cane, mints, or other crops

Note: These types may be defined differently in other regions.

including rice paddies; seasonally flooded agricultural land; salt exploitation sites; water storage areas including impoundments generally more than 8ha; excavations; wastewater treatment areas; and canals and drainage channels) are also included in the system.

In North America, the Canadian and United States wetland classification systems were developed by government agencies interested in wetland conservation and management. The Canadian system emphasizes wetland origin (class), form, and vegetation in describing the wetland types.[3] Five wetland classes are recognized: bog, fen, marsh, swamp, and shallow water. Within each class, different forms and types are characterized. Eight general vegetation types are defined by the presence or absence of vegetation (treed, shrub, forb, graminoid, moss, lichen, aquatic bed, and nonvegetated). These types may be subdivided into other types (e.g., treed into coniferous or deciduous, shrub into tall, low, and mixed, graminoids into grass, reed, tall rush, low rush, and sedge, aquatic bed into floating and submerged). The U.S. Fish and Wildlife Service's wetland and deepwater habitat classification[4] is the official federal system used for mapping wetlands and for reporting the status and trends of wetlands in the U.S. The features separating wetlands include general ecological and physical factors and specific features such as vegetation, soil/substrate composition, hydrology, water chemistry, and human alterations. Classification follows a hierarchical approach with five main levels designated: ecological system (marine, estuarine, lacustrine, riverine, and palustrine), subsystem, class (vegetated: forested, scrub-shrub, emergent, and aquatic bed; nonvegetated: unconsolidated shore, rocky shore, streambed, and reef), subclass, and modifiers. The modifiers are used to describe a wetland's hydrology (water regime), pH and salinity (water chemistry), soils, and the influence of humans and beaver (special modifiers). Common types include estuarine intertidal emergent wetlands (e.g., salt and brackish marshes), estuarine intertidal unconsolidated shore (e.g., tidal flats and beaches), palustrine emergent wetlands (e.g., marshes, fens, and wet meadows), palustrine forested wetlands, and palustrine scrub-shrub wetlands (e.g., shrub bogs and shrub swamps).

A hydrogeomorphic approach (HGM) to wetland classification has also been developing in the U.S.[5] The HGM system emphasizes abiotic features important for assessing wetland functions. Seven

FIGURE 1 Some examples of North American wetlands: (a) tidal salt marsh, (b) inland marsh, (c) pothole marsh, (d) wet meadow/shrub swamp, (e) northern peatland, (f) bottomland swamp, (g) hardwood swamp, and (h) flatwood wetland.
Source: Photos courtesy of U.S. Fish and Wildlife Service.

hydromorphic classes are identified: riverine, depressional, slope, mineral soil flats, organic soil flats, lacustrine fringe, and estuarine fringe. The U.S. Fish and Wildlife Service has adapted the HGM approach to provide additional modifiers to its classification system on a pilot basis. These HGM-type descriptors include landscape position (i.e., lotic, lentic, terrene, estuarine, and marine), landform (i.e., slope, basin,

TABLE 3 Estimates of the Current Extent of Wetlands in Different Regions of the World

Region/Country	Wetland Extent (ha)	Source
Africa	121,321,683–124,686,189	[13]
Asia	211,501,790–224,117,790	[14]
Central America		
Mexico	3,318,500 (very incomplete)	[15]
Europe		
Eastern	225,849,930	[16]
Western	28,821,979	[17]
Middle East	7,434,790	[18]
Neotropics	414,996,613	[19]
North America		
Canada	127,199,000–150,000,000	[20]
U.S.	114,544,800	[1]
Oceania	35,748,853	[21]
South America		
Tropical region	200,000,000	[22]

interfluve, floodplain, flat, island, and fringe), and water flow path (i.e., inflow, outflow, throughflow, bidirectional flow, isolated, and paludified).[6] These descriptors provide the required information to aid the evaluation of functions of wetlands across watersheds and large geographic areas.

Extent of Wetlands

Comprehensive wetland inventories do not exist in most countries. There are many inconsistencies among the inventories (e.g., different levels of effort, focus on particular types, and artificial wetlands such as rice paddies are often not included in wetland inventories).[7] Consequently, comparative analysis is fraught with problems. Nonetheless, Table 3 provides some perspective on the extent of wetlands in many regions. Most of the data came from a series of reports produced for the Bureau of the Ramsar Wetlands Convention.[8] Globally, estimates for wetlands range from about 750 million ha[7] to about 1.5 billion ha. Ten countries have over 2 million ha of peat-lands alone, with Canada leading at nearly 130 million ha (represents about 18% of the country) followed by the former U.S.S.R. at 83 million ha.[9,10] About a third of Finland is covered by peatlands (10 million ha). The Pantanal of South America, perhaps the largest wetland in the world, reportedly covers about 200,000 km^2 (or 2 million ha) during the wet season.[11]

References

1. Tiner, R.W. *Wetland Indicators: A Guide to Wetland Identification, Delineation, Classification, and Mapping;* Lewis Publishers, CRC Press: Boca Raton, FL, 1999; 392 pp.
2. *Ramsar Convention Bureau.* Information Sheet on Ramsar Wetlands: Gland, Switzerland, 1998.
3. National Wetlands Working Group. *Wetlands of Canada;* Ecological Land Classification Series No. 21, Land Conservation Branch, Canadian Wildlife Service; Environment Canada: Ottawa, Ont., 1988; 452 pp.
4. Cowardin, L.M.; Carter, V.; Golet, F.C.; LaRoe, E.T. *Classification of Wetlands and Deepwater Habitats of the United States;* FWS/OBS-79/31; U.S. Department of the Interior, Fish and Wildlife Service: Washington, DC, 1979; 131 pp.
5. Brinson, M.M. *A Hydrogeomorphic Classification for Wetlands;* Wetlands Research Program Tech. Rep. WRP-DE-4; U.S. Army Waterways Expt. Station: Vicksburg, MS, 1993; 103 pp.

6. Tiner, R.W. *Keys to Waterbody Type and Hydrogeomorphic-Type Wetland Descriptors for U.S. Waters and Wetlands, Operational Draft;* U.S. Department of the Interior, Fish and Wildlife Service; Northeast Region: Hadley, MA, 2000; 20 pp.
7. Finlayson, C.M.; Davidson, N.C. Global review of wetland resources and priorities for wetland inventory: summary report. In *Global Review of Wetland Resources and Priorities for Wetland Inventory;* Finlayson, C.M.; Spiers, A.G., Eds.; Supervising Scientist Report 144; Canberra, Australia, 1999; 9 pp.
8. Finlayson, C.M., Spiers, A.G., Eds.; *Global Review of Wetland Resources and Priorities for Wetland Inventory*; Supervising Scientist Report 144; Canberra, Australia, 1999.
9. Taylor, J.A. Peatlands of the British Isles. In *Mires: Swamp, Bog, Fen, and Moor; Regional Studies;* Gore, A.J.P., Ed.; Elsevier: Amsterdam, the Netherlands, 1–46.
10. Botch, M.S.; Massing, V.V. Mire ecosystems in the U.S.S.R. In *Mires: Swamp, Bog, Fen, and Moor; Regional Studies*; Gore, A.J.P., Ed.; Elsevier: Amsterdam, the Netherlands; 95–152.
11. Swarts, F.A., Ed. *The Pantanal of Brazil, Bolivia, and Paraguay;* Waterland Research Institute; Hudson MacArthur Publishers: Gouldsboro, PA, 2000; 287 pp.
12. Wetland Advisory Committee. *The Status of Wetlands Reserves in System Six*; Report of the Wetland Advisory Committee to the Environmental Protection Authority: Australia, 1977.
13. Stevenson, N.; Frazier, S. Review of wetland inventory information in Africa. In *Global Review of Wetland Resources and Priorities for Wetland Inventory;* Finlayson, C.M., Spiers, A.G., Eds.; Supervising Scientist Report 144; Canberra, Australia, 1999; 94 pp.
14. Watkins, D.; Parish, F. Review of wetland inventory information in Asia. In *Global Review of Wetland Resources and Priorities for Wetland Inventory*; Finlayson, C.M., Spiers, A.G., Eds.; Supervising Scientist Report 144; Canberra, Australia, 1999; 26 pp.
15. Olmstead, I. Wetlands of Mexico. In *Wetlands of the World: Inventory, Ecology, and Management Volume I*; Whigham, D.F., Dykyjova, D., Heiny, S., Eds.; Kluwer Academic Publishers: Dordrecht, the Netherlands, 1993; 637–677.
16. Stevenson, N.; Frazier, S. Review of wetland inventory information in Eastern Europe. In *Global Review of Wetland Resources and Priorities for Wetland Inventory*; Finlayson, C.M., Spiers, A.G., Eds.; Supervising Scientist Report 144; Canberra, Australia, 1999; 53 pp.
17. Stevenson, N.; Frazier, S. Review of wetland inventory information in Western Europe. In *Global Review of Wetland Resources and Priorities for Wetland Inventory*; Finlayson, C.M., Spiers, A.G., Eds.; Supervising Scientist Report 144; Canberra, Australia, 1999; 57 pp.
18. Frazier, S.; Stevenson, N. Review of wetland inventory information in the Middle East. In *Global Review of Wetland Resources and Priorities for Wetland Inventory*; Finlayson, C.M., Spiers, A.G., Eds.; Supervising Scientist Report 144; Canberra, Australia, 1999; 19 pp.
19. Davidson, I.; Vanderkam, R.; Padilla, M. Review of wetland inventory information in the Neotropics. In *Global Review of Wetland Resources and Priorities for Wetland Inventory*; Finlayson, C.M., Spiers, A.G., Eds.; Supervising Scientist Report 144; Canberra, Australia, 1999; 35 pp.
20. Davidson, I.; Vanderkam, R.; Padilla, M. Review of wetland inventory information in the North America. In *Global Review of Wetland Resources and Priorities for Wetland Inventory;* Finlayson, C.M., Spiers, A.G., Eds.; Supervising Scientist Report 144; Canberra, Australia, 1999; 35 pp.
21. Watkins, D. Review of wetland inventory information in Oceania. In *Global Review of Wetland Resources and Priorities for Wetland Inventory*; Finlayson, C.M., Spiers, A.G., Eds.; Supervising Scientist Report 144; Canberra, Australia, 1999; 26 pp.
22. Junk, W.J. Wetlands of tropical South America. In *Wetlands of the World: Inventory, Ecology, and Management Volume* I; Whigham, D.F., Dykyjova, D., Heiny, S., Eds.; Kluwer Academic Publishers: Dordrecht, the Netherlands, 1993; 679–739.

VIII

PRO: Basic Environmental Processes

68

Eutrophication

Sven Erik Jørgensen

Eutrophication Problem

Many aquatic ecosystems suffer from eutrophication: lakes, reservoirs, estuaries, lagoons, fjords, and bays. It is a worldwide problem and, together with the oxygen depletion problem, is probably the most serious pollution problem of aquatic ecosystems.

The word eutrophy is generally taken to mean "nutrient rich." In 1919, Nauman introduced the concepts of oligotrophy and eutrophy, distinguishing between oligotrophic lakes containing little planktonic algae and eutrophic lakes containing much phytoplankton. The eutrophication of aquatic ecosystems all over the world has increased rapidly during the last decade due to increased urbanization and the consequently increased discharge of nutrients. The discharges in industrialized countries are about 1.8 kg N/yr and 0.5 kg P/yr. The production of fertilizers has grown exponentially in this century, and the concentration of phosphorus in many lakes reflects this.

The word eutrophication is used increasingly in the sense of the artificial addition of nutrients, mainly nitrogen and phosphorus, to waters. Eutrophication is generally considered to be undesirable, but this is not always true. The green color of eutrophied lakes makes swimming and boating less safe due to the increased turbidity, and from an aesthetic point of view, the chlorophyll concentration should not exceed 100 mg/m^3. However, the most critical effect from an ecological point of view is the reduced oxygen content of the hypolimnion, caused by the decomposition of dead algae, particularly in the fall. Eutrophic aquatic ecosystems sometimes show a high oxygen concentration (at the surface during the summer time) but a low concentration of oxygen in the hypolimnion, which may be lethal to fish. The oxygen depletion in the hypolimnion will often imply that the eutrophication is more difficult to abate, because anaerobic sediment will more easily release its content of phosphorus. Iron (III) is reduced to iron (II) by anaerobic conditions. As iron (III) has a very insoluble phosphorus salt, while iron (II) phosphate is readily soluble, the phosphorus release by anaerobic conditions is dependent on the composition of the sediment, particularly of course the iron content. One of the most applied lake restoration methods is the pumping of air or oxygen to the hypolimnion, which is used to significantly reduce the release of phosphorus from the sediment.

Water makes up 75%–90% of the total wet weight of plant tissue. It means that except for oxygen and hydrogen, the relative composition on dry weight basis would be 4–10 times higher. For phytoplankton, the carbon, nitrogen, and phosphorus content on dry weight basis are, respectively, approximately 40%–60%, 6%–8%, and 0.75%–1.0%. Phosphorus is considered the major cause of eutrophication in lakes, as it was formerly the growth-limiting factor for algae in the majority of lakes. Its use has increased tremendously during the last decade. Nitrogen is limiting in a number of East African lakes as a result of the nitrogen depletion of soils by intensive erosion in the past. Nitrogen is furthermore often limiting in the coastal ecosystems, at least for part of the year. However, today, nitrogen may become limiting in lakes as a result of the tremendous increase in the phosphorus concentration caused by the discharge of wastewater, which contains relatively more phosphorus than nitrogen. While algae use 5–10 times more nitrogen than phosphorus (see the content of these two elements in phytoplankton above), wastewater generally contains only 3 times as much nitrogen as phosphorus. In lakes, a considerable amount of nitrogen is furthermore lost by denitrification (nitrate → N_2). Lakes that have received wastewater for a longer period may therefore be limited by nitrogen. In environmental management, the key question is, however, not which element is the limiting factor but which element can be most easily controlled as a limiting factor. As phosphorus generally can be removed more easily and cheaply, it is often the most effective abatement of eutrophication to remove phosphorus effectively from wastewater discharge to the lake. Further details are given below.

It is a good management strategy for an abatement of eutrophication in aquatic ecosystems to find quantitatively all the sources of nitrogen and phosphorus. In most cases, based on this information, it is easy to find possible solutions and the corresponding costs. In this context, it is beneficial to apply an ecological model. Eutrophication models have been developed and applied for all these ecosystems. Many eutrophication models have been applied for environmental management, particularly for lakes and reservoirs. The *Handbook of Ecological Models Used in Ecosystem and Environmental Management*[1] gives a good overview of the available ecological models, including eutrophication models.

Growth of Phytoplankton

The growth of phytoplankton is the key process of eutrophication, and it is therefore important to understand the interacting processes that regulate growth. Primary production has been measured in great detail in a number of aquatic ecosystems and presents the synthesis of organic matter. The overall process can be summarized as follows:

$$\text{Light} + 6CO_2 + 6H_2O \rightarrow C_6H_{12}O_6 + 6O_2 \tag{1}$$

The composition of phytoplankton is not constant. The composition of phytoplankton and plants in general reflects to a certain extent the concentration of the water. If, e.g., the phosphorus concentration is high, the phytoplankton will take up relatively more phosphorus—this is called luxury uptake. Phytoplankton consists mainly of carbon, oxygen, hydrogen, nitrogen, and phosphorus: without these elements, no algal growth will take place. This leads to the concept of the limiting nutrient, which has been developed by Liebig as the law of the minimum. However, the concept has been considerably misused due to oversimplification. First of all, growth might be limited by more than one nutrient. The composition as mentioned above is not constant but varies with the composition of the environment. Furthermore, growth is not at its maximum rate until the nutrients are used, at which point growth stops, but the growth rate slows down as the nutrients become depleted.

The sequence of events leading to eutrophication has often been described as follows: Oligotrophic waters will have a ratio of N to P greater than or equal to 10, which means that phosphorus is less abundant than nitrogen for the needs of phytoplankton. If sewage is discharged into the aquatic ecosystem, the ratio will decrease, since the N-to-P ratio for municipal wastewater is 3:1, and consequently, nitrogen will be less abundant than phosphorus relative to the needs of phytoplankton, as indicated above. In this situation, however, the best remedy for the excessive growth of algae is not necessarily the removal of nitrogen from the sewage, because the mass balance might then show that nitrogen-fixing algae will give an uncontrollable input of nitrogen into the system. It is particularly the case if the aquatic ecosystem is a lake or reservoir. It is necessary to set up mass balances for each of the nutrients as already pointed out, and these will often reveal that the input of nitrogen from nitrogen-fixing blue-green algae, precipitation, and tributaries is contributing too much to the mass balance for the removal of nitrogen from the sewage to have any effect. On the other hand, the mass balance may reveal that the phosphorus input (often more than 95%) comes mainly from sewage, which means that it is better management to remove phosphorus from the sewage than nitrogen. Thus, what is important in environmental management is not which nutrient is most limiting, but rather, which nutrient can most easily be made to limit the algal growth.

The conceptual diagram in Figure 1 shows the nitrogen cycle, and Figure 2 shows the phosphorus cycle of aquatic ecosystems and illustrates the processes behind the cycling of these nutrients. As clearly pointed out above, it is always beneficial to use a mass balance to choose the most important components, forcing functions, and state variables to be considered in the management context. Let us illustrate the needed mass balance considerations by use of an example. Let us anticipate that it is an open question whether birds should be included in the selected management strategy (and in the eutrophication model). Birds may contribute considerably to the inputs of nutrients by their droppings. If the nutrients—nitrogen and phosphorus—coming from the birds' droppings are insignificant compared with the amounts of nutrients coming from drainage water, precipitation, and wastewater, inclusion of birds in the management strategy is only an unnecessary complication that only would contribute to the uncertainty. There are, however, a few cases where birds may contribute as much as 25% or at least more than 5% of the total inputs of nutrients. In such cases, it is of course important to include birds in the management and as a model component or at least as an important forcing function. A mass balance is always needed to uncover the main sources of a pollution problem—in the example of the inputs of nutrients. It is, however, rare (a good guess based on the author's experience is about 1% of aquatic ecosystems) that it is necessary to consider the droppings of birds as a significant source of eutrophication.

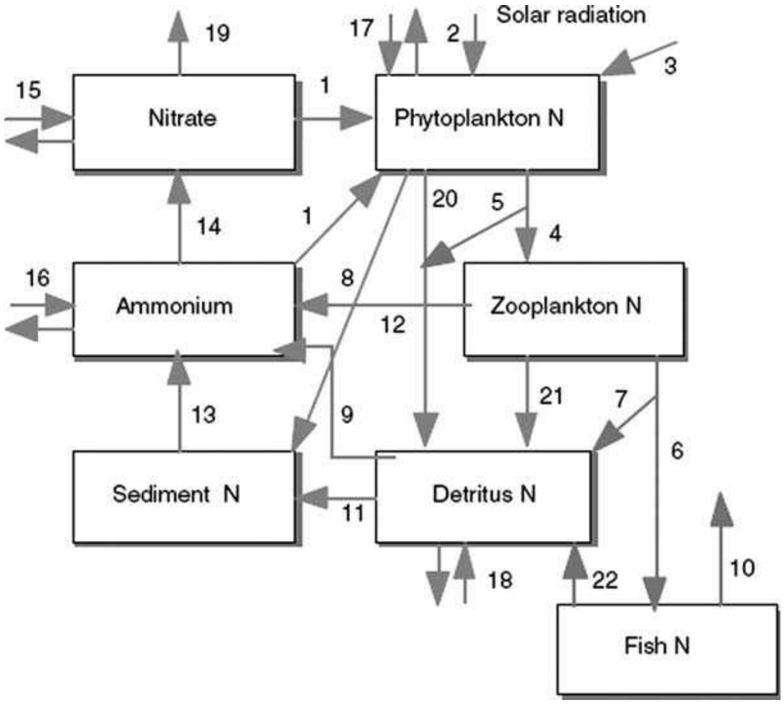


FIGURE 1 The conceptual diagram of a nitrogen cycle in an aquatic ecosystem. The processes that connect the state variables and forcing functions are the following: 1) uptake of nitrate and ammonium by algae; 2) photosynthesis; 3) nitrogen fixation; 4) grazing with loss of undigested matter; 5–7), predation and loss of undigested matter; 8) settling of algae; 9) mineralization; 10) fishery; 11) settling of detritus; 12) excretion of ammonium from zooplankton; 13) release of nitrogen from the sediment; 14) nitrification; 15–18) inputs/outputs; 19) denitrification; and 20–22) mortality of phytoplankton, zooplankton, and fish.

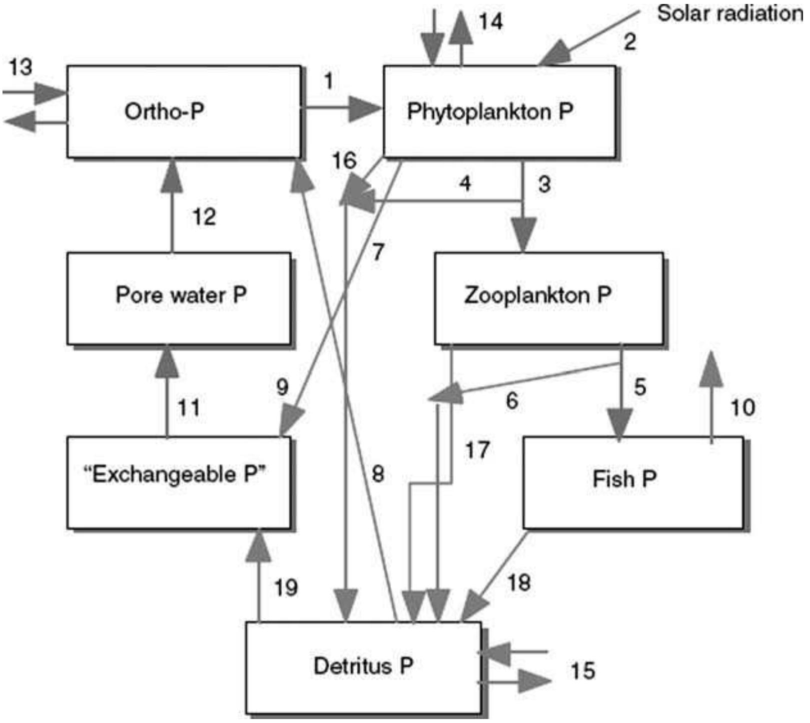


FIGURE 2 The phosphorus cycle. The processes are the following: 1) uptake of phosphorus by algae; 2) photosynthesis; 3) grazing with loss of undigested matter; 4–5) predation with loss of undigested material; 6, 7, 9) settling of phytoplankton; 8) mineralization; 10) fishery; 11) mineralization of phosphorous organic compounds in the sediment; 12) diffusion of pore water P; 13–15) inputs/outputs; 16–18) mortalities; and 19) settling of detritus.

The so-called Michaelis–Menten equation can be applied to describe the growth of phytoplankton:

$$\mathrm{gr} = \mathrm{grmax\ NS}/(\mathrm{kn}+\mathrm{NS}) \tag{2}$$

Or

$$\mathrm{gr} = \mathrm{grmax\ PS}/(\mathrm{kp}+\mathrm{PS}) \tag{3}$$

dependent on which nutrient is limiting, N or P. grmax, kn, and kp are parameters. If both nutrients are limiting in different periods of the year, the formulation is

$$\mathrm{gr} = \mathrm{grmax\ min}\big(\mathrm{NS}/(\mathrm{kn}+\mathrm{NS}),\ \mathrm{PS}/(\mathrm{kp}+\mathrm{PS})\big) \tag{4}$$

The product or average of several limiting factors has also been proposed and applied.

For the influence of the temperature, there are two possible formulations:

$$K_{\mathrm{t}}^{(\mathrm{TEMP}-20)}(\text{the so} - \text{called Arrhenius equation}) \tag{5}$$

Or

$$\exp\big(A^{*}(\mathrm{TEMP}-\mathrm{OPT})/(\mathrm{TEMP}-\mathrm{MAXTEMP})\big) \tag{6}$$

K_t is a parameter, which in most cases is between 1.04 and 1.06 and is on average 1.05. OPT is the optimum temperature for phytoplankton growth, and MAXTEMP is the maximum temperature. OPT, MAXTEMP, and A are all parameters that are different for different phytoplankton species.

The description of phytoplankton growth by equations using the constant stoichiometry approach is a simplification, because the phytoplankton growth is in reality a two-step process, as illustrated in Figure 3 and applied in the non-constant approach. The first step is uptake of nutrients, and the second step is growth of phytoplankton (increase of the biomass). The more correct description can be formulated mathematically by the following equations:

$$\text{Uptake rate } P = \mathrm{d}\text{PA}/\mathrm{d}t = \text{PA}^{*}\text{maxupp}^{*}\big(\text{PS}/\big(\text{kp}+\text{PS}\big)\big)$$

$$^{*}\ \big((\text{PAMAX}-\text{PA})/(\text{PAMAX}-\text{PAMIN})\big)$$

$$\text{Parallel for uptake rate } N \tag{7}$$

$$\text{Uptake rate } C = \mathrm{d}\text{CA}/\mathrm{d}t = \text{CA}^{*}\text{maxupc}^{*}\big(\text{CS}/\text{kc}+\text{CS}\big)^{*}$$

$$\big((\text{CAMAX}-\text{CA})/(\text{CAMAX}-\text{CAMIN})\big)^{*}\ \big((L/\text{KL}+L)-\text{RESP}$$

$$\text{IF } L\big\langle \text{L1}, L \text{ is used; if } \text{L2}\big\rangle L > \text{L1}, \text{L1 is used;}$$

$$\text{if } L > \text{L2}, \text{L1}+\text{L2}-L \text{ is used for } L \tag{8}$$

PA, NA, and CA are state variables that cover the amount of phosphorus, nitrogen, and carbon, in the form of phytoplankton, expressed as mg P, N, or C per liter of water. Notice that the unit is mg in 1 L of water. maxupp, maxupn, macups, kp, kn, kc, PAMAX, PAMIN, NAMAX, NAMIN, CAMAX, CAMIN, KL, L1, and L2 are all parameters. PAMAX, PAMIN, NAMAX, NAMIN, CAMAX, and CAMIN are,

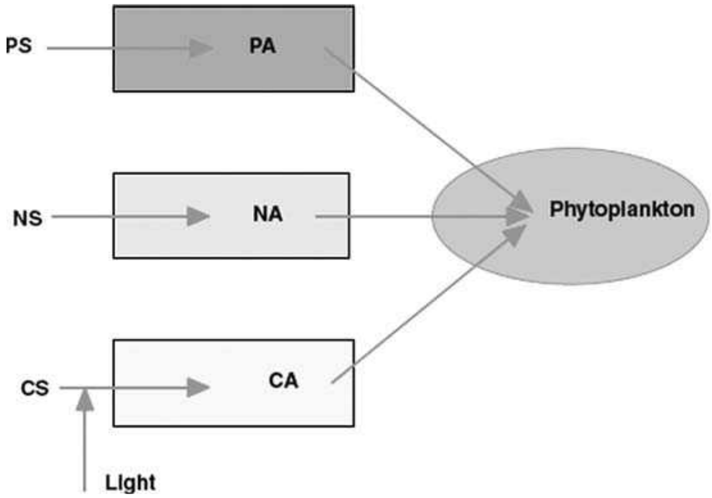


FIGURE 3 The two-step model of phytoplankton growth. The first step is uptake of nutrients PS, NS, and CS, followed by a growth of phytoplankton dependent on the nutrient concentrations in the phytoplankton cells, PA, NA, and CA. The uptake of carbon is dependent on light, while the uptake of phosphorus and nitrogen can take place even in darkness. It is a more physiologically correct description of phytoplankton growth than the (constant stoichiometry) approach.

however, known fairly well. They are the phytoplankton concentration times, respectively, 0.025, 0.005, 0.12, 0.05, 0.6, and 0.4 with good approximations. It is of course more difficult to calibrate the two-step growth equations than the constant stoichiometry approach due to the higher number of parameters in the NC equations, although the approximate knowledge that is available to the six parameters, PAMAX, PAMIN, NAMAX, NAMIN, CAMAX, and CAMIN, facilitates the calibration slightly. Notice that the uptakes of phosphorus, nitrogen, and carbon are, according to the equations, dependent both on the concentrations of the nutrients in the water and on the concentrations of nutrients in the cells.

The closer the nutrient concentrations in phytoplankton are to the minimum, the faster is the uptake. When a nutrient concentration, on the other hand, has reached the maximum value, the uptake stops. The carbon uptake opposite the uptake of phosphorus and nitrogen is dependent also on light, according to a Michaelis–Menten expression that includes the light prohibition. Finally, RESP covers the respiration. Of course, only carbon is involved in the respiration.

The growth process is quantified by the following equation:

$$\begin{aligned}\text{Growth} = \text{grmax}^{*}\ \text{phytoplankton}^{*}\ (min((PA - PAMIN)/\\(PAMAX - PAMIN)), ((NA - NAMIN)/\\(NAMAX - NAMIN)), ((CA - CAMIN)/\\(CAMAX - CAMIN)))\end{aligned} \tag{9}$$

grmax is a parameter in line with the corresponding parameter in Eq. 1. Eq. 9 indicates that the higher the nutrient concentrations are compared with the minimum levels, the faster the growth is.

The phytoplankton growth model based on this approach has four state variables: PA, NA, CA, and phytoplankton. They are all assumed to be expressed in mg/L. As the minimum and maximum values are presumed to be a parameter times the phytoplankton concentration, they are also expressed in mg/L.

The two-step description is of course more difficult to calibrate, validate, and use generally, which of course raises the following question: when should the two-step description be applied instead of the more easily applicable constant stoichiometric approach? Di Toro and Conolly[2] have revealed that the need for the two-step description increases with the shallowness and eutrophication of the aquatic ecosystem; see Figure 4. It is definitely recommended for shallow, very eutrophied aquatic ecosystems to apply the two-step description, while it is hardly needed for deep mesotrophic or oligotrophic aquatic ecosystems.

Solutions to the Eutrophication Problem

It is usually necessary to apply a wide spectrum of methods to solve eutrophication problems, because the nutrients have many sources. The amount of nutrients discharged with the wastewater can be reduced by several wastewater treatment methods; see for instance the overview of conventional wastewater treatment methods and municipal wastewater and its treatment. Nutrients can be discharged by drainage water, including drainage water from agriculture, which may have high nutrient concentration; this can be removed or reduced by a number of ecological engineering methods, for instance, wetlands; see these entries.

It is possible to reduce the nutrient concentrations in the aquatic ecosystems by use of environmental technological methods, ecotechnological methods, cleaner technology, and environmental legislation. There are a number of cases where a consequent environmental strategy has solved the problems, but there are also many examples of insufficient environmental management, which has led to only a partial solution or no solution at all of the eutrophication problem. From experience, it can be concluded that quantitative nutrient balances, including all sources of nutrients, are the best starting point for a good

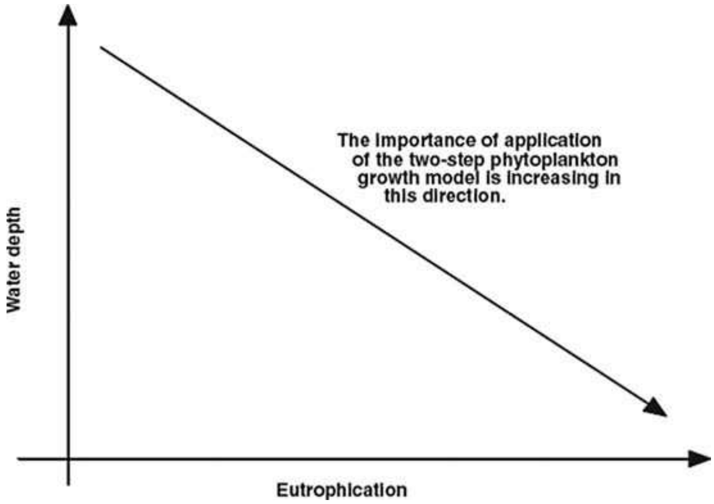


FIGURE 4 The need for the two-step description of phytoplankton growth increases with nutrient concentration and decreases with depth.

environmental management strategy, because the nutrient balances show clearly which sources are important to eliminate or reduce, and they facilitate a comparison of the costs for various management strategies. For further details about selection of an environmental management strategy, see Jørgensen[3] and Jørgensen et al. [4]

References

1. Jørgensen, S.E. *Handbook of Ecological Models Used in Ecosystem and Environmental Management*; CRC; Boca Raton, FL, 2011; 620 pp.
2. Di Toro, D.M.; Conolly, J.F. Jr. Mathematical models of water quality in large lakes, Part II: Lake Erie. U.S. Environmental Protection Agency, Ecological Research Series, EPA-600/3–80-065, 1980.
3. Jørgensen, S.E. *Principles of Pollution Abatement;* Elsevier: Amsterdam and Oxford, 2000; 520 pp.
4. Jørgensen, S.E.; Löffler, H.; Rast, W.; Straskraba, M. *Lake and Reservoir Management*; Elsevier: Amsterdam and Oxford, 2004; 503 pp.

69

Wastewater Use in Agriculture

Manzoor Qadir,
Pay Drechsel, and
Liqa Raschid-Sally

Introduction

The volume of wastewater generated by domestic, industrial, and commercial water use has increased with population, urbanization, improved living conditions, and economic development. With the increase in wastewater generation, its productive use in agriculture has increased, particularly downstream of cities where often farmers have no or limited alternative sources of reliable irrigation water.[1–3] In developing countries, treated, partly treated, diluted, or untreated wastewater is used in urban and peri-urban areas to produce vegetables as a market-ready product. In addition, rice, wheat, fodder, and industrial crops are produced. In developed countries, wastewater is treated and mostly used for landscape irrigation, particularly in cities though extensive and well-regulated use in agriculture occurs as well.[4]

Despite official restrictions and potential health implications, farmers in many freshwater-scarce developing countries tend to opt for wastewater irrigation where available because (1) wastewater is a reliable or often the only water source available for irrigation throughout the year; (2) wastewater irrigation often reduces the need for fertilizer application as it is a source of nutrients; (3) wastewater use involves less energy cost even when pumping, if the alternative clean water source is from deep groundwater; and (4) wastewater generates additional benefits such as greater income generation from cultivation and marketing of high-value crops such as vegetables, which create year-round employment opportunities.[5,6]

As wastewater irrigation is in most instances part of the informal irrigation sector, authorities face challenges controlling or regulating the practice. The protection of consumer and farmer health and environment is especially the main concern.[7] Thus, sustainable use of wastewater must address three major aspects: pertinent policies, regulations, and institutional arrangements; wastewater treatment per intended reuse option; and risk management practices that eliminate or minimize the health and environmental impacts, particularly when wastewater treatment is limited. This entry addresses on-farm management practices with a focus on crop, water, and soil management interventions.

Crop Selection and Diversification

Where wastewater is used for irrigation, crop selection is determined by more factors than usual freshwater irrigation conditions. While the market potential of the crop is always an important criterion, the suitability of the crop to its biophysical environment requires more attention than is usually given. Even more important is the potential risk to consumers, which becomes a decisive factor influencing enforcement of regulations and crop restrictions. Especially if irrigated crops are eaten raw, as with salad greens, microbial contamination can be high; hence, different crops should be chosen. In guidelines addressing safe use of wastewater, it is advised to avoid cultivating crops eaten uncooked and prefer cereal and fodder crops, and trees.[8]

A particular challenge for farmers is the salinity of the wastewater as considerable variation exists among crops in their ability to tolerate saline conditions.[9] An appropriate selection of plant species capable of producing adequate biomass is vital while using saline water for irrigation.[10] Such selection is generally based on the ability of the species to withstand elevated levels of salinity in irrigation water and soil[9] while also providing a salable product or one that can be used on-farm.[11,12] The salt tolerance of a crop is not an exact value because it depends on several soil, crop, and climatic factors. This diversity can be exploited to identify local crops that are better adapted to saline and/or sodic soil conditions.[13,14]

Based on the linear response equation proposed by Maas and Hoffman,[9] crops can be characterized for their salt tolerances. The following are two parameters obtained from the Maas–Hoffman equation: (1) the maximum allowable soil salinity for a crop without yield reduction (the threshold soil salinity) and (2) the percent decrease in crop yield per unit increase in salinity beyond the threshold salinity level for the crop (the slope). Both values can be used to calculate relative yield for any given soil salinity exceeding the threshold level (Table 1).

Research efforts have led to the identification of several field crops, forage grasses and shrubs, biofuel crops, and fruit-tree and agroforestry systems, which can suit a variety of salt-affected environments.[15] Such systems linked to secure markets should support farmers in finding the most suitable

TABLE 1 Yield Potentials of Some Grain, Forage, Vegetable, and Fiber Crops as a Function of Average Root Zone Salinity

		Yield Potential (%) at Specified Salinity (dS/m)		
Common Name	Tolerance Based On	50%	80%	100%
Durum wheat	Grain yield	19	11	6
Barley	Grain yield	18	12	8
Cotton	Seed cotton yield	17	12	8
Rye	Grain yield	16	13	11
Sugar beet	Storage root	16	10	7
Wheat	Grain yield	13	9	6
Sorghum	Grain yield	10	8	7
Alfalfa	Shoot dry weight	9	5	2
Spinach	Top fresh weight	9	5	2
Broccoli	Shoot fresh weight	8	5	3
Eggplant	Fruit yield	8	4	1
Rice paddy	Grain yield	7	5	3
Potato	Tuber yield	7	4	2
Maize	Ear fresh weight	6	3	2

Source: Based on the salt tolerance data of different crops and percentage decrease in yield per unit increase in root zone salinity in terms of dS/m as reported by Maas and Grattan.[14]

Note: These data serve only as a guideline to relative tolerances among crops. Absolute tolerances can vary between varieties also depending on climate, soil conditions, and cultural practices.

and sustainable crop diversifying systems to mitigate any perceived production risks, while ideally also enhancing the productivity per unit of saline wastewater and protecting the environment.

Aside from excess salts in the water, there is an increasing possibility of heavy metal contamination where industrial effluent contributes to the water that farmers use. In terms of potential toxicity of the metals and metalloids, Hamilton et al.[16] classified them into four groups based on their retention in soil, translocation in plants, phytotoxicity, and potential risk to the food chain. They categorized cadmium, cobalt, selenium, and molybdenum as posing the greatest risk to human and animal health because they may appear in wastewater-irrigated crops at concentrations that are not generally phytotoxic; that is, farmers cannot rely on their plants dying before concentrations reach levels not recommended for humans.

Guidelines for maximum allowable levels of metals and metalloids in irrigation water are summarized in Table 2. Where these limits are exceeded, Simmons et al.[17] describe available methods for damage control.

TABLE 2 Recommended Maximum Concentrations (RMCs) of Selected Metals and Metalloids (mg/L) in Irrigation Water

Element	RMC[a]	Remarks
Aluminum	5.00	Can cause non-productivity in acid soils ($pH < 5.5$), but more alkaline soils at $pH > 7.0$ will precipitate the ion and eliminate any toxicity
Arsenic	0.10	Toxicity to plants varies widely, ranging from 12 mg/L for Sudan grass to less than 0.05 mg/L for rice
Beryllium	0.10	Toxicity to plants varies widely, ranging from 5 mg/L for kale to 0.5 mg/L for bush beans
Cadmium	0.01	Toxic at concentrations as low as 0.1 mg/L in nutrient solution for beans, beets, and turnips. Conservative limits recommended
Chromium	0.10	Not generally recognized as an essential plant growth element. Conservative limits recommended
Cobalt	0.05	Toxic to tomato plants at 0.1 mg/L in nutrient solution. It tends to be inactivated by neutral and alkaline soils
Copper[b]	0.20	Toxic to a number of plants at 0.1–1.0 mg/L in nutrient solution
Iron[b]	5.00	Non-toxic to plants in aerated soils, but can contribute to soil acidification and loss of availability of phosphorus and molybdenum
Lithium	2.50	Tolerated by most crops up to 5 mg/L. Mobile in soil. Toxic to citrus at low concentrations with recommended limit of <0.075 mg/L
Manganese[b]	0.20	Toxic to a number of crops at a few tenths to a few milligrams per liter in acidic soils
Molybdenum	0.01	Non-toxic to plants at normal concentrations in soil and water. Can be toxic to livestock if forage is grown in soils with high concentrations of available molybdenum
Nickel	0.20	Toxic to a number of plants at 0.5–1.0 mg/L; reduced toxicity at neutral or alkaline pH
Lead	5.00	Can inhibit plant cell growth at very high concentrations
Selenium	0.02	Toxic to plants at low concentrations and toxic to livestock if forage is grown in soils with relatively high levels of selenium
Zinc[b]	2.00	Toxic to many plants at widely varying concentrations; reduced toxicity at $pH \geq 6.0$ and in fine textured or organic soils

Source: Modified from Ayers and Westcot.[22]

[a] The maximum concentration is based on a water application rate that is consistent with good irrigation practices (10,000 m^3/ha/yr). If the water application rate greatly exceeds this, the maximum concentrations should be adjusted downward accordingly. No adjustment should be made for application rates less than 10,000 m^3/ha/yr. The values given are for water used on a long-term basis at one site.

[b] Synergetic action of copper and zinc, and antagonistic action of iron and manganese have been reported in certain plant species' absorption and tolerance of metals after wastewater irrigation. If irrigation water contains high concentrations of copper and zinc, copper concentrations in the tissue may increase greatly. In plants irrigated with water containing a high concentration of manganese, manganese uptake in plants may increase, and consequently, the concentration of iron in the plant tissue may be reduced considerably. Generally, metal ion concentrations in plant tissue increase with concentrations in irrigation water. Metal ion concentrations in roots are higher than those in leaves[26] and metal ion concentrations such as cadmium concentrations in leafy vegetables are higher than those in non-leafy species.[27]

The examples show that the selection of crops for irrigation with wastewater has to consider a range of factors uncommon in freshwater irrigation. Finding the best compromise between the various criteria while targeting profits (i.e., high market demand) limits the choice of crop significantly in most cases.

Irrigation Management

In addition to the choice of the most appropriate crop(s), irrigation water management (ranging from water access to the type of irrigation, application rates, and scheduling) offers a variety of important management practices to address the particularities of wastewater irrigation.[15,17,18]

There are different ways in which crops are irrigated with wastewater, such as surface or flood irrigation, manual irrigation with watering cans, furrow irrigation, sprinkler irrigation, and micro-irrigation such as drip or trickle irrigation. There are also different ways of water access, from manual water fetching to pumping and gravity flow. Each method has a different level of risk for farmers (through farmers' contact with the irrigation water) and consumers [through the contact of the harvested (and eventually consumed) crop part and the water]. Several parameters for the evaluation of commonly used irrigation methods in relation to risk reduction are given in Table 3.[7,19] Key criteria are health risks, costs, and water use efficiency.

Flood irrigation is usually a low-cost method with, however, also low water use efficiency. Health protection is limited for the farmers and—if the crops are low growing—for the consumers as well. With medium level of health protection, furrow irrigation needs more soil preparation and is suitable when

TABLE 3 Parameters for Evaluation of Selected Irrigation Methods in Relation to Risk Reduction for Crops and Humans

Evaluation Parameter	Irrigation Method			
	Furrow Irrigation	Flood Irrigation	Sprinkler Irrigation	Drip Irrigation
Farmer exposure to pathogens	Low to medium	Medium to high	Low when sprinkler is off	Very low
Crop exposure to pathogens	Low if planted on ridge	High only for low-growing crops	High	Very low
Possibility of leaf damage from salts resulting in poor yield	No foliar injury as the crop is planted on the ridge	Some bottom leaves may be affected, but the damage is not so serious as to reduce yield	Severe leaf damage can occur resulting in significant yield loss	No foliar injury likely
Root zone salt accumulation with repeated applications	Salts tend to accumulate in the ridge, which could harm the crop	Depending on soil texture, salts might move vertically downwards or can accumulate in the root zone	Salt drainage is limited as water amounts are low. Surface crusting possible	Salt movement is radial along the direction of water movement. A salt wedge is formed between drip points; clogging of pipes can occur
Ability to maintain high soil water potential (risk of soil moisture stress)	Plants may be subject to stress between irrigations	Plants may be subject to water stress between irrigations	Not possible to maintain high soil water potential throughout the growing season	Possible to maintain a high and well-targeted soil moisture content throughout the growing season
Suitability to handle brackish wastewater without significant yield loss	Fair to medium. With good management and drainage, acceptable yields are possible	Fair to medium. Good irrigation and drainage practices and conditions can produce acceptable levels of yield	Poor to fair. Crops might suffer from leaf damage resulting in low yields	Excellent to good. Almost all crops can be grown with very little reduction in yield, unless the pipes clog

Source: Modified from Pescod.[19]

TABLE 4 On-Farm Options for Pathogen Reductions

Control Measure	Pathogen Reduction (Log Units)	Notes
Alternative safe water source	>6	Depends on availability of safe groundwater and/or alternative farm land
Crop restrictions	>6	Acceptance depending on controls and profit margin of the alternative crop
Drip irrigation	2–4	2-log unit reduction for low-growing crops, and 4-log unit reduction for high-growing crops
Pathogen die-off	0.5–2 per day	Die-off after last irrigation before harvest (rate depends on climate, crop type, etc.)
Slow sand filter	1–3	Depends on appropriate particle size
Furrow irrigation	1–2	Might reduce cropping density and yield
Reduced splashing	1–2	Splashing adds contaminated soil particles onto the crop, which can be avoided
Allow sedimentation in ponds and dugouts	1–2	During dry season via natural die-off. Reduction of helminths to less than 1 egg over 2–3 days possible

Source: Modified from WHO.[8]

there is a greater leaching need to remove high levels of salts. Irrigation with sprinklers is medium to high cost, does not require soil preparation, but has medium-level water use efficiency. Sprinkler irrigation systems have the advantage of reducing the amounts of water and salts applied to soil and crop.[20,21] The disadvantage compared to flood and furrow irrigation is that sprinklers distribute any water contaminants straight on the top of the plants. The same applies to the use of watering cans unless they are directed towards the roots. Overhead irrigation may also cause leaf burn under direct sunlight, from salts absorbed directly through wetted leaf surfaces,[22] which can be avoided by irrigating at night.[14,23,24]

Drip irrigation systems are costly, but highly efficient in water use along with the highest levels of health protection for farmers and consumers. The clogging of drippers on the other hand may limit the use of drip irrigation systems for many types of wastewater. Therefore, prior filtration is needed to prevent clogging of emitters.[25]

Irrigation frequency is an ambivalent issue in wastewater irrigation. Because soluble salts reduce the availability of water in almost direct proportion to their concentration, irrigation frequency should generally be high. This will help in maintaining the moisture content and salinity of irrigated soils at acceptable levels, which is important especially during seedling establishment.[23] On the other hand, a low frequency of irrigation—if possible even the cessation of irrigation for several days before harvest—supports the natural die-off of pathogens and is an important low-cost measure for health risk reduction.[7]

Other options for on-farm interventions addressing risks from pathogens are summarized in Table 4. Options to cope with salinity or heavy metals were recently summarized by Qadir and Drechsel[10] and Simmons et al.[17]

Soil- and Health-based Interventions

Good management practices play a crucial role in the preservation of key soil properties while irrigating with wastewater. Soil-based interventions are important, particularly in case of inorganic contaminants, which usually accumulate in the upper part of the soil because of strong adsorption and precipitation phenomena. For moderate levels of metals and metalloids in wastewater, there is no particular management needed if the soils are calcareous, i.e., contain appreciable levels of calcite that renders most metals immobile. However, metal ions may be a problem in acid soils, which need specific management measures such as liming, avoiding use of fertilizers with acidic reactions, and selection of crops that do

TABLE 5 Contribution of Irrigation with Recycled Wastewater in Terms of Nutrient Addition to the Soil

		Fertilizer Contribution (kg/ha)	
Nutrient	Concentration (mg/L)	Irrigation at 3,000 m^3/ha	Irrigation at 5,000 m^3/ha
Nitrogen	16–62	48–186	80–310
Phosphorus	4–24	12–72	20–120
Potassium	2–69	6–207	10–345
Calcium	18–208	54–624	90–1040
Magnesium	9–110	27–330	45–550

Source: Lazarova and Bahri.[20]

not accumulate the metals of concern.[22] In case of irrigation with wastewater containing elevated levels of sodium, care should be taken to avoid soil structure deterioration. Application of a source of calcium such as gypsum is desirable. Procedures to determine the rate of gypsum application to mitigate the effects of sodium resulting from sodic wastewater irrigation are available.

The quality and depth of groundwater prior to wastewater irrigation determine the detrimental effects of salts, nitrates, and metals reaching groundwater. The deeper the groundwater, the longer it will take to have such effects. In case of shallow groundwater or coarse-textured soils, i.e., sandy soils, which are highly permeable, care must be taken to prevent groundwater pollution.

Although the fertilizer value of undiluted wastewater, in particular, is of great importance as nutrients in wastewater contribute to crop requirements, periodic monitoring is required to estimate the nutrient loads in wastewater and adjust fertilizer applications accordingly.[20] Excessive nutrients not only can cause nutrient imbalances, undesirable vegetative growth, and delayed or uneven maturity, but can also reduce crop quality while polluting groundwater and surface water. The amount of nutrients applied via wastewater irrigation can vary considerably if it is raw, treated, or diluted with stream water. The contribution in terms of nutrient addition to the soil from irrigation with recycled wastewater is given in Table 5.

It becomes obvious that irrigation with wastewater has a variety of implications steering farmers' decision-making process in view of crop selection as well as soil and water management. Most of them are limiting the choice of options. However, while a wrong choice, for example, in view of salinity management or any other phytotoxic hazard usually results in a quick learning process for the farmer, hazards affecting farmers' and especially consumers' health might remain less obvious and hidden among various confounding factors, such as poor sanitation at home. With economic interests being of paramount importance in most cases, it is not surprising that farmers might opt for those options that yield highest returns while keeping investments as low as possible. The result is the common picture of high-value exotic vegetables irrigated with low-cost watering cans or via flood irrigation, both imposing high risks for human health. Increasing awareness about these risks and providing incentives and regulations to encourage alternative crop choices in spite of possible disadvantages will create the conditions that would favor the utilization of farm-based interventions for the safe and productive use of wastewater in irrigated agriculture.

Conclusions

In the arid and semi-arid areas of the world, wastewater is mainly used for agriculture because of the competition between agriculture and municipal sectors. While wastewater is managed safely and used in treated form mostly for landscape irrigation in developed countries, the situation is different in developing countries, where wastewater is rarely used in treated form but mostly in partly treated, diluted, or untreated forms to produce a range of crops, mostly vegetables as a market-ready product. Such wastewater use in developing countries continues to offset local water scarcity despite official restrictions and potential health implications. Thus, it is important to consider a range of risk management practices to eliminate or minimize the health and environmental impacts, particularly in situations where wastewater is used for irrigation in partly treated, diluted, or untreated forms. On-farm management practices can play a major role to eliminate or minimize the health and environmental impacts. The following aspects are the key features and have to be considered: 1) selection and diversification of wastewater-irrigated crops to reduce possible health risks, accounting for their market value and tolerance against ambient stress from salts, metals, and metalloids; 2) irrigation water management addressing water access, on-farm treatment, type of irrigation, application rates, and irrigation scheduling; and 3) soil- and health-based interventions considering soil characteristics, soil preparation practices, and application of fertilizers and amendments. Increasing awareness about health and environmental risks, and knowledge and implementation of best farm-based interventions by the wastewater-irrigating farmers will help create the conditions that would favor the safe and productive use of wastewater in agriculture.

References

1. Qadir, M.; Wichelns, D.; Raschid-Sally, L.; Minhas, P.S.; Drechsel, P.; Bahri, A.; McCornick, P. Agricultural use of marginal-quality water—Opportunities and challenges. In *Water for Food, Water for Life: A Comprehensive Assessment of Water Management in Agriculture*; Molden, D., Ed.; Earthscan: London, 2007; 425–457.
2. Jiménez, B.; Asano, T. Water reclamation and reuse around the world. In *Water Reuse: An International Survey of Current Practice, Issues and Needs*; Jimenez, B., Asano, T., Eds.; IWA Publishing: London, 2008; 3–26.
3. Drechsel, P.; Scott, C.A.; Raschid-Sally, L.; Redwood, M.; Bahri, A., Eds. *Wastewater Irrigation and Health: Assessing and Mitigating Risks in Low-Income Countries*; Earthscan–International Development Research Centre (IDRC)–International Water Management Institute (IWMI), Earth-scan: London, 2010.
4. Asano, T. *Wastewater Reclamation and Reuse*; Water Quality Management Library Series. Vol. 10; Technomic Publishing Co., Inc.: Lancaster, PA, 1998.
5. Keraita, B.N.; Jimenez, B.; Drechsel, P. Extent and implications of agricultural reuse of untreated, partly treated and diluted wastewater in developing countries. *CAB Rev.* 3 (058), 2008; 1–15.
6. Qadir, M.; Scott, C.A. Non-pathogenic tradeoffs of wastewater irrigation. In *Wastewater Irrigation and Health: Assessing and Mitigating Risks in Low-Income Countries*; Drechsel, P., Scott, C.A., Raschid-Sally, L., Redwood, M., Bahri, A., Eds.; Earthscan–International Development Research Centre (IDRC)–International Water Management Institute (IWMI), Earthscan: London, 2010; 101–126.
7. Jiménez, B.; Drechsel, P.; Koné, D.; Bahri, A.; Raschid-Sally, L.; Qadir, M. Wastewater, sludge and excreta use in developing countries: An overview. In *Wastewater Irrigation and Health: Assessing and Mitigating Risks in Low-Income Countries*; Drechsel, P., Scott, C.A., Raschid-Sally, L., Redwood, M., Bahri, A., Eds.; Earthscan–International Development Research Centre (IDRC)–International Water Management Institute (IWMI), Earthscan: London, 2010; 3–27.
8. WHO. *Guidelines for the Safe Use of Wastewater, Excreta and Grey Water: Wastewater Use in Agriculture*; WHO, FAO, UNEP: Geneva, 2006; Vol. 2.
9. Maas, E.V.; Hoffman, G.J. Crop salt tolerance—Current assessment. *J. Irrig. Drainage Div.* **1977**, *103*, 115–134.
10. Qadir, M.; Drechsel, P. Managing salts while irrigating with wastewater. *CAB Rev.* 5 (016), 2010; 1–14.
11. Qadir, M.; Oster, J.D. Crop and irrigation management strategies for saline-sodic soils and waters aimed at environmentally sustainable agriculture. *Sci. Tot. Environ.* **2004**, *323*, 1–19.

12. Hamilton, A.J.; Stagnitti, F.; Xiong, X.; Kreidl, S.L.; Benke, K.K.; Maher, P. Wastewater irrigation: The state of play. *Vadose Zone J.* **2007**, *6* (4), 823–840.
13. Shannon, M.C. Adaptation of plants to salinity. *Adv. Agron.* **1997**, *60*, 76–120.
14. Maas, E.V.; Grattan, S.R. Crop yields as affected by salinity. In *Agricultural Drainage*; Skaggs, R.W., van Schilfgaarde, J., Eds.; ASA-CSSA-SSSA: Madison, 1999; 55–108.
15. Qadir, M.; Tubeileh, A.; Akhtar, J.; Larbi, A.; Minhas, P.S.; Khan, M.A. Productivity enhancement of salt-affected environments through crop diversification. *Land Degrad. Dev.* **2008**, *19*, 429–453.
16. Hamilton, A.J.; Boland, A.-M.; Stevens, D.P.; Kelly, J.; Radcliffe, J.; Dillon, P.J.; Paulin, B. Position of the Australian horticultural industry with respect to the use of reclaimed water. *Agric. Water Manage.* **2005**, *71*, 181–209.
17. Simmons, R.W.; Qadir, M.; Drechsel, P. Farm-based measures for reducing human and environmental health risks from chemical constituents in wastewater. In *Wastewater Irrigation and Health: Assessing and Mitigating Risks in Low-Income Countries*; Drechsel, P., Scott, C.A., Raschid-Sally, L., Redwood, M., Bahri, A., Eds.; Earthscan–International Development Research Centre (IDRC)–International Water Management Institute (IWMI), Earthscan: London, 2010; 209–238.
18. Keraita, B.; Konradsen, F.; Drechsel, P. Farm-based measures for reducing microbiological health risks for consumers from informal wastewater-irrigated agriculture. In *Wastewater Irrigation and Health: Assessing and Mitigation Risks in Low-Income Countries*; Drechsel, P., Scott, C.A., Raschid-Sally, L., Redwood, M., Bahri, A., Eds.; Earthscan–IDRC–IWMI, Earthscan: London, 2010; 189–207.
19. Pescod, M.B., Ed. *Wastewater Treatment and Use in Agriculture*; Irrigation and Drainage Paper No. 47; Food and Agriculture Organization of the United Nations: Rome, Italy, 1992.
20. Lazarova, V.; Bahri, A. *Water Reuse for Irrigation: Agriculture, Landscapes, and Turf Grass*; CRC Press: Boca Raton, FL, 2005.
21. Qadir, M.; Raschid-Sally, L.; Drechsel, P. Wastewater use in agriculture: Agronomic considerations. In *Encyclopedia of Water Science*; Trimble, S.W., Ed.; Taylor & Francis: New York, 2008; 1296–1299.
22. Ayers, R.S.; Westcot, D.W. *Water Quality for Agriculture*; Irrigation and Drainage Paper 29 (Rev. 1); FAO: Rome, 1985.
23. Rhoades, J.D. Use of saline drainage water for irrigation. In *Agricultural Drainage*; Skaggs, R.W., van Schilfgaarde, J., Eds.; American Society of Agronomy (ASA)–Crop Science Society of America (CSSA) –Soil Science Society of America (SSSA): Madison, WI, 1999; 615–657.
24. Qadir, M.; Minhas, P.S. Wastewater use in agriculture: Saline and sodic waters. In *Encyclopedia of Water Science*; Trimble, S.W., Ed.; Taylor & Francis: New York, 2008; 1307–1310.
25. Minhas, P.S.; Samra, J.S. *Wastewater Use in Peri-urban Agriculture: Impacts and Opportunities*; Central Soil Salinity Research Institute: Karnal, 2004.
26. Kalavrouziotis, I.K.; Drakatos, P.A., 2002. Irrigation of certain Mediterranean plants with heavy metals. *Int. J. Environ. Pollut.* 2004, *18*, 294–300.
27. Qadir, M.; Ghafoor, A.; Murtaza, G. Cadmium concentration in vegetables grown on urban soils irrigated with untreated municipal sewage. *Environ. Dev. Sustain.* **2000**, *2*, 11–19.

70

Wetlands: Biodiversity

Jean-Claude
Lefeuvre and
Virginie Bouchard

Role of Hydrology

Wetlands differ significantly in their water source and seasonal hydrologic regime. Hydrological patterns (i.e., flooding frequency, duration and hydroperiod) influence physical and chemical characteristics (e.g., salinity, oxygen and other gas diffusion rates, reduction–oxidation potential, nutrient solubility) of a wetland. In return, these internal parameters and processes control flora and fauna distribution as well as ecosystem functions. Plants, animals and microbes are often oriented in predictable ways along the hydrological gradient (Figure 1). Conversely, the biotic component affects the hydrology by eventually modifying flow or water level in a wetland.[1,2] Species also influence nutrient cycles and other ecosystem functions.[3]

A Landscape Perspective

Although the hydrology is part of the ecological signature of an individual wetland, wetlands are neither considered as aquatic nor terrestrial systems. They have characteristics from both systems and are defined as ecotones placed under this dual influence.[4] Because wetlands are located at the interface of multiple systems, they assure vital functions (e.g., wildlife habitats) beneficial at the landscape level. Reduction of wetland area often reduces biodiversity in the land-scape.[2,5] Increases in biodiversity occur when wetlands are created or restored in a disturbed landscape.[6]

Wetland Flora

Wetland plants are adapted to a variety of stressful abiotic conditions (e.g., immersion, wave abrasion, water level fluctuation, low oxygen conditions). Identical adaptations to common environmental features have led taxonomically distinct species to sometimes look similar in terms of morphology, life cycle and life forms.[7] Traditionally, wetland plants have been classified into groups of different life forms, primarily in relation with hydrological conditions. Helophytes are defined as plant species with over-wintering buds in water or in the submerged bottom.[8] They are differentiated from hydrophytes in that their vegetative organs are partially raised above water level.[8] Hejny's classification is based on relatively stable vegetative features that determine the ability of wetland plants to survive two unfavorable conditions, cold and drought.[7] This classification uses the types of photosynthetic organs present in both the growth and flowering phases. Other classifications include both life form and growth form. When a species has a range of growth form, it is classified under the form showing the greatest achievement of its potential.[7]

Plant communities in wetlands can be more or less homogeneous, mosaic-like, or distributed along a gradient resulting in a clear zonation of species. A gradient exists if one or several habitat parameters change gradually in space. This phenomenon is common in fresh water marshes that present a gradient of water depth and water saturated soils (Figure 1). Such a gradient is often accompanied by differences in peat accumulation that is influenced by waves or currents. General principles of the zonation of aquatic plants have been largely described.[7,9,10] Littoral vegetation can belong to several types of communities, which derived from the general principle that, from deeper water to the shore, we may expect successively submergent, floating, and emergent macrophytes. The most important habitat factor is water depth, depending on slope and peat accumulation.[10] Other factors may be poor irradiance caused by high turbidity or exposure to waves or flow.[7,10]

Riparian ecosystems are found along streams and rivers that occasionally flood beyond confined channels or where riparian sites are created by channel meandering in the stream network. Riparian or bottomland hardwood forests contain unique tree species that are flood tolerant. Species distribution is associated with floodplain topography, flooding frequency, and flooding duration.[11,12] In southeastern U.S. bottomland, seasonally flooded forests are colonized by *Platanus occidentalis* (sycamore), *Ulmus americana* (American elm), *Populus deltoides* (cottonwood) and are flooded between 2% and 25% of the growing season (Figure 1). Other species such as *Fraxinus pennsylvanica* (green ash), *Celtis laevigata* (sugarberry) and *Carya aquatica* (water hickory) colonized bottomlands that are flooded by less than 2% of the growing season.[11] Freshwater marshes are dominated with emergent macrophytes rooted in

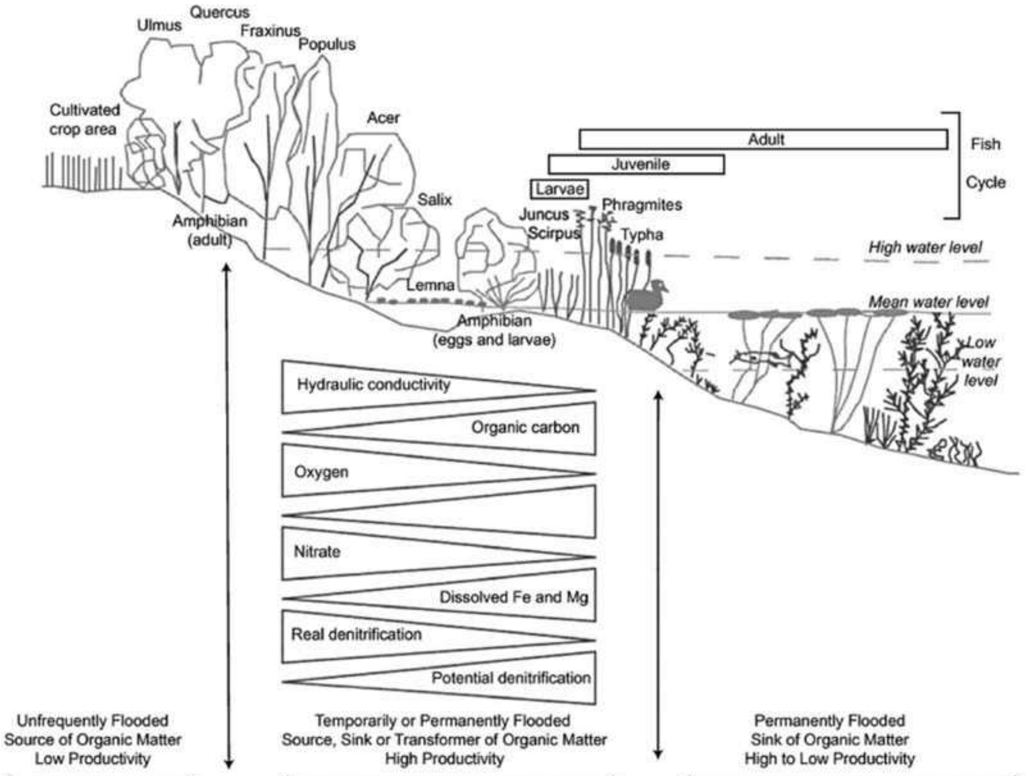


FIGURE 1 Species distribution along the hydrological gradient in a freshwater marsh.

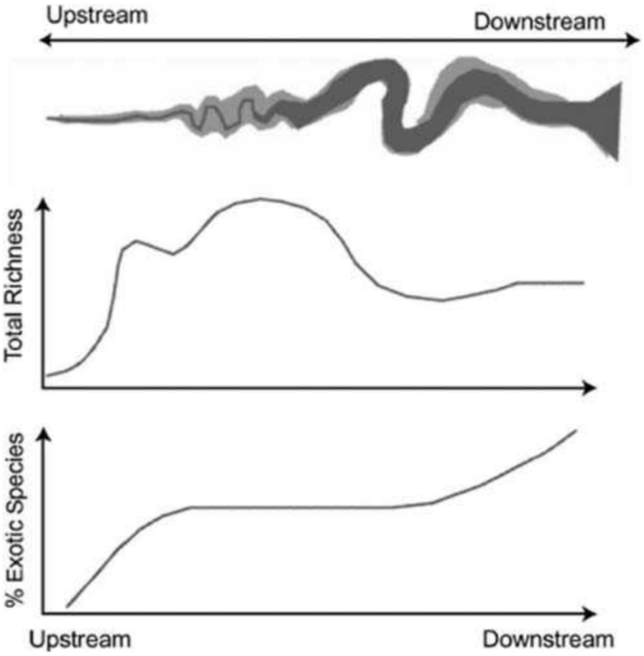


FIGURE 2 Total richness and invasive species distribution along a stream longitudinal gradient.
Source: Planty-Tabacchi.[14]

the bottom with aerial leaves (i.e., helophytes). Species such as *Typha* (cattail), *Phragmites* (reed grass) and *Scirpus* (bulrush) are often clonal. A plant community is usually organized in sequence of patches that are dominated by one species. The second plant groups are the rooted plants with floating leaves (*Nymphaeid*). Lotus (*Nelumbo*) and water lilies (*Nymphea*) have very identical morphology (i.e., similar leaves and flowers) but a genetic analysis showed that lotus is more closely related with plane-tree than with water lilies.[10] Submerged plants include elodeids (i.e., cauline species whose whole life cycle can be completed below the water surface or where only the flowers are emergent) and isoetids (i.e., species growing on the bottom whose whole life cycle can be completed without contact with the surface). Submerged species include species such as coontail (*Ceratophyllum demersum*) and water milfoil (*Myriophyllum* spp.). Plant species found in salt marshes are called halophytes (i.e., plants which complete their cycle in saline environments). A saltmarsh can be divided into low, middle, and upper marsh, according to flooding frequency and duration. Each zone is dominated by different plant species according to their tolerance to saline immersion.[10]

A dominant competitive species—often a clonal species—can modify the theoretical zonation. Change in water chemistry (i.e., eutrophication) or hydrology may favor a particular species over the natural plant community. Highly competitive species are often invasive and aggressive in displacing native species. The expansion of *Phragmites australis* into tidal wetlands of North America causes a reduction in biodiversity as many native species of plants are replaced by a more cosmopolitan species.[13] In riparian ecosystems, biodiversity is usually higher in the intermediate zones, whereas it is lower upstream and downstream (Figure 2). The percentage of exotic species is low in upstream areas, but can represent up to 40 in downstream zones (Figure 1).[14]

Wetland Fauna

Diversity of vertebrate and invertebrate species is the result of a diverse community composed of resident and transient species, which use the space differently and at various times of day and year. The density and variety of animal populations at a particular wetland site is also explained by climatic events that affect geographic areas on a large scale. For example, the population of waterfowl during winter is largely dependent on climate variations in the northern part of the continents.[15]

Resident species are often dependent on the type of vegetation. Animal communities are generally distributed along a zonation pattern, parallel to the plant communities, which is driven by the hydrological gradient. A few species depend entirely on a single plant species for their survival. One beetle (*Donacia* spp.) may depend on reeds, at least during its larval stage where another beetle (*Gale-rucella* spp.) uses only water lilies as their habitat and diet. For many other residents, their habitats extend to several plant communities during their life span. It is generally the case for many vertebrates such as amphibians, rodents, passerines, and waterfowl.

Many amphibian species depend on wetland or riparian zones for reproduction and larval stage. Likewise, wet meadows are necessary as a reproduction zone and nursery for a number of freshwater fishes. About 220 animals and 600 plant species are threatened by a serious reduction of wetlands in California, and the state's high rate of wetland loss (91% since the 1780s) is partly responsible.[15] Many waterfowl species are sensitive to areas of reduction, patch size and distribution, wetland density, and proximity to other wetlands.[16] When the Marais Poitevin (France), one of the principal wintering and passages sites for waterfowl in the Western Europe, underwent agricultural intensification in the 1980s, the population of ducks and waders declined tremendously. This decline was partly due to a 50% reduction of wet meadows between 1970 and 1995, primarily caused by the conversion to arable farmlands.[17] Thus, maintenance of biodiversity depends on the existence of inter connections between wetlands, and between aquatic and terrestrial ecosystems. In fact, some authors have pointed out that increase in biodiversity occurs when wetlands are created or restored in a disturbed landscape.[6,18]

Conclusions

Despite the importance of wetlands, conservation efforts have ignored them for a long time. It is urgent to conserve the existing wetlands, and also to restore and create wetland ecosystems. A wide range of local, state, federal, and private programs are available to support the national policies of wetland "No Net Loss" in the U.S., and around the world. From a biodiversity perspective, on-going wetland protection policies may not be working because restored or created wetlands are often very different from natural wetlands.[19] Created wetlands often result in the exchange of one type of wetland for another, and result in the loss of biodiversity and functions at the landscape level.[19] We know now that it takes more than water to restore a wet- land,[20] even if an important place should be given to the ability of self-design of wetland ecosystems.[21]

References

1. Naiman, R.J.; Johnston, C.A.; Kelley, J.C. Alteration of north American streams by beaver. Bioscience **1988**, *38,* 753–762.
2. National Research Council. *Wetlands: Characteristics and Boundaries;* National Academy Press: Washington, DC, 1995; 306 pp.
3. Naiman, R.J.; Pinay, G.; Johnston, C.A.; Pastor, J. Beaver influences on the long-term biogeochemical characteristics of boreal forest drainage networks. Ecology **1994**, *75,* 905–921.
4. Hansen, A.J; Di Castri, F; Eds.; Landscape boundaries. In *Consequences for Biotic Diversity and Ecological Flows*; Ecological Studies; Springer: Berlin, 1983; 452 pp.
5. Gibbs, J.P. Wetland loss and biodiversity conservation. Conser. Biol. **2000**, *14,* 314–317.
6. Weller, M.W. *Freshwater Marshes,* 3rd Ed.; University of Minnesota Press: Minneapolis, US, 1994; 192 pp.
7. Westlake, D.F.; Kvet, J.; Szczepanski, A.; Eds.; *The Primary Ecology of Wetlands;* Cambridge University Press, 1998; 568 pp.

8. Raunkiaer, C. *The Fife Forms of Plants and Statistical Plant Geography*; Oxford University Press: Oxford, England, 1934; 632 pp.
9. Grevilliot, F.; Krebs, L.; Muller, S. Comparative importance and interference of hydrological conditions and soil nutrient gradients in floristic biodiversity in flood meadows. Biodivers. Conserv. **1998**, *7*, 1495–1520.
10. Mitsch, W.J.; Gosselink, J.G. *Wetlands,* 3rd Ed.; Wiley: New York, 2000; 920 pp.
11. Clark, J.R.; Benforado, J.; Eds.; *Wetlands of Bottomland Hardwood Forest*; Elsevier: Amsterdam, the Netherlands, 1981; 401 pp.
12. Keogh, T.M.; Keddy, P.A.; Fraser, L.H. Patterns of tree species richness in forested wetlands. Wetlands **1999**, *19,* 639–647.
13. Chambers, R.M.; Meyerson, L.A.; Saltonstall, K. Expansion of Phragmites Australis into Tidal wetlands of North America. Aquat. Bot. **1999**, *64,* 261–273.
14. Planty-Tabacchi, A.M. Invasions Des Corridors Riverains Fluviaux Par Des Espèces Végétales D'origine Étrangère. These University Paul Sabatier Toulouse III, 1993; France, 177 pp.
15. Hudson, W.E., Ed.; *Landscape Linkages and Biodiversity;* Island Press: Washington, DC, 1991.
16. Leibowitz, S.C.; Abbruzzese, B.; Adams, P.R.; Hughes, L.E.; Frish, J. A Synoptic Approach to Cumulative Impact Assessment: A Proposed Methodology, 1992; EPA/600/R-92/167.
17. Duncan, P.; Hewison, A.J.M.; Houte, S.; Rosoux, R.; Tour-nebize, T.; Dubs, F.; Burel, F.; Bretagnolle, V. Long-term changes in agricultural practices and wildfowling in an internationally important wetland, and their effects on the guild of wintering ducks. J. Appl. Ecol. **1999**, *36,* 11–23.
18. Hickman, S. Improvement of habitat quality for resting and migrating birds at the des plaines river wetlands demonstration project. Ecol. Engng. **1994**, *3*, 485–494.
19. Whigham, D.F. Ecological issues related to wetland preservation, restoration, creation and assessment. Sci. Total Environ. **1999**, *240,* 31–40.
20. Zedler, J.B. Progress in wetland restoration ecology. Trends Ecol. Evol. **2000**, *15* (10), 402–407.
21. Mitsch, W.J.; Wu, X.; Nairn, R.W.; Weihe, P.E.; Wang, N.; Deal, R.; Boucher, C.E. Creating and restoring wetlands: a whole ecosystem experiment in self-design. BioScience **1998**, *48,* 1019–1030.

71

Wetlands: Carbon Sequestration

Virginie Bouchard
and Matthew
Cochran

Introduction

The increase in the concentration of "greenhouse gases" in the troposphere and its relation to human activities is now well-documented.[1] Carbon fixation via photosynthesis and carbon release during decomposition have always been two important processes regulating the concentrations of CO_2 and CH_4 in the atmosphere, even in prehistoric and pre-industrial times. In addition to these internal processes, wetland ecosystems receive and release organic carbon through hydrologically-driven mass fluxes. Thus, the carbon pool within any wetland ecosystem is in balance between primary production, microbial decomposition, and carbon fluxes within interconnected ecosystems (Figure 1). Wetlands play a particularly complex role in controlling greenhouses gases, as these ecosystems are intimately associated with all aspects of the production and consumption of both CO_2 and CH_4 (Figure 2).

Carbon Sequestration with Primary Production

The concept of primary productivity is directly related to the ideas of energy flow in ecosystems. A portion of the photosynthetically active radiation (PAR), which is radiation in the 400–700 nm wave band, received by an ecosystem is absorbed by autotrophic organisms (photosynthetic plants and microorganisms). The absorbed energy is reradiated, lost as latent heat, or stored by the activity of photosynthesis in organic substances. This last flow of energy corresponds to net primary production (NPP). Fundamental ecological questions relating to the global carbon budget, the location of the missing carbon sink, and predictions of global climate change rely on obtaining good estimates of NPP.

Wetland ecosystems can have very high standing biomass values and correspondingly high NPP.[2,3] The annual aboveground NPP of macrophytes is reported to be up to 5 kg dry matter in the most productive sites.[2] This production varies according to species, wetland type, and latitude[2,3] and is often well correlated with the maximum aboveground standing crop. Belowground production is much more difficult to estimate because roots and rhizomes grow and die at different rates and times and because materials are translocated to and from shoots. The ratio of belowground to aboveground production can vary from 0.2 to 2.5 according to various studies.[2] The connectivity of a wetland to hydrological fluxes

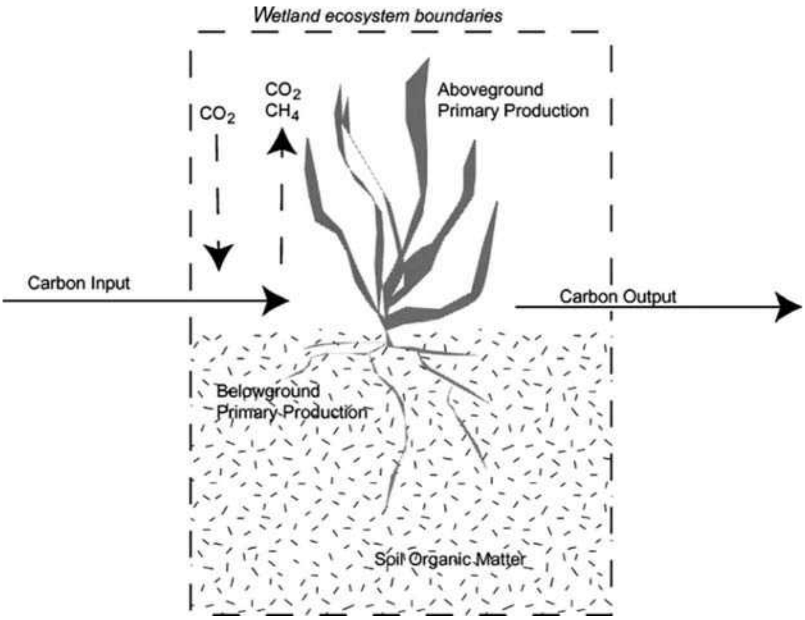


FIGURE 1 Conceptual model showing the fluxes of organic carbon in wetlands.

FIGURE 2 View of a freshwater wetland.

(i.e., saltmarshes flooded by tides, coastal freshwater marshes flooded by seiche, riparian bottomland forests flooded by floods) is one of the most important factors enhancing primary productivity.[4]

Net carbon capture from herbaceous vegetation is minimal compared to the long-term accumulation of carbon in bottomland hardwood forests. At the end of the growing season, most of the carbon trapped in the biomass of herbaceous macrophytes is found in dead plant litter and is released as CO_2 or CH_4 during decomposition or exported as dissolved or particulate organic matter to adjacent systems.

Bottomland hardwood forests store carbon in tree biomass for a much longer period of time. The productivity of bottomland forest ranges between 200 g and 2000 g dry organic matter per year.[5]

Carbon Sequestration in Soils

Carbon fixed in photosynthesis either remains in the sediment as carbon accretion or is decomposed to CO_2 and CH_4 by a suite of fermentative microbes involved in soil organic matter decomposition. A diversity of biological, chemical, and physical mechanisms is also known to selectively “protect” different pools of soil organic matter from decomposition by soil microorganism. Most of the recalcitrant carbon destined for sediment accretion is derived from heavily lignified biomass. Organic soil is a result of the anaerobic conditions created by standing water or poorly drained conditions.

Carbon is even better protected in acidic environments, in marine sediments and under low temperatures. Peat accumulation is a result of reduced oxidation of the biomass produced in wetlands. The northern peatlands have accumulated 25–38 Pmol C since the last glaciation, equivalent to 50%–70% of the total amount of carbon currently present in the atmosphere.[6] Peat accumulation is greater in bogs than in fens due to differences in nutrient availability. Decomposition rates are enhanced in minerotrophic fens receiving nutrients from adjacent mineral soils relative to ombrotrophic bogs, which are fed only by rainwater.

Canadian wetland soils store about 154 Gt C, representing an average of 60% more carbon than what is stored in Canadian forests (95 Gt C in biomass and soils), which in turn is two orders of magnitude larger than the agricultural soil carbon pool[7] (Table 1). Peatlands are a sink for between 20 and 30 g Cm^{-2} yr^{-1}, which means that Canadian peatlands sequester about 0.03 Gt Cyr^{-1} (Table 1). The sequestration rate of peatlands is currently smaller than the rate computed for boreal forests. However, the major difference between forests and peatlands is that a forest sink is transient and cannot be sustained continuously. If peatlands also have a theoretical limit to growth, it is reached only after many thousands of years.[8] However, the fluxes of CO_2 and CH_4 from peatlands vary highly according to seasons, years, species and sites. Inter-annual measurement schemes—such as those that have been developed for forest ecosystems with the Ameriflux and Euroflux programs—are needed are needed to establish more precise fluxes of CO_2.

Impact of Human Activities on Carbon Storage

Rates of plant production are limited by phosphorus supply in freshwater systems and by nitrogen supply in saltwater and terrestrial systems.[9,10] An excess of nutrients leads to an increase of both primary production and decomposition in terrestrial systems.[11] Decomposition rates increase, however, at a slower rate than primary production, leading to a potential increase of carbon storage in soil.[11] In wetlands, little is known about the effect of excess nutrients on the denitrification process, which is directly linked to the amount of soil organic matter.[12]

Wetlands have been drained on all continents due to human development. Rapid changes occur in organic soils after drainage as aerobic microbial decomposition is enhanced, releasing large amounts of CO_2 into the atmosphere. Carbon losses can lead to subsidence of the soil profile, changes in bulk density, and decreases in carbon content of the remaining soil.

TABLE 1 Summary of Carbon Stocks and Fluxes in Canadian Wetlands/Peatlands, Forests and Agricultural Fields

Land Use	Area (km^2)	C Stock (Gt C)	Fluxes (Gt C yr^{-1})	Potential Duration (Years)
Peatlands/wetlands soils	1.24 × 106	154	−0.03	1000 years (?)
Forests	5.15 × 106	95	−0.2	50
Agriculture fields	0.45 × 106	6	0.0034	50

Source: Adapted from Roulet.[7]

Creation of Wetlands to Sequester Carbon

Wetlands—particularly bottomland hardwood forests— are considered as potentially excellent "carbon sinks" because they take carbon dioxide out of the atmosphere and store it in living plant tissues and soil organic matter. Wetland creation might provide an opportunity to positively address wetland habitat losses while also addressing global warming [8,13] However, a question, still remains: are we capable of constructing wetlands that function in similar ways as natural wetlands? How well these created and constructed wetlands mimic natural wetlands is still being debated.[14,15]

Various studies have indicated that created wetlands might not function in the same capacity as adjacent reference wetlands.[16] In created *Spartina alterniflora* salt marshes, vegetation rapidly achieves 100% cover, although soil nitrogen and organic matter are slow to accumulate. Salt marshes constructed in North Carolina 25 years ago have lower soil organic carbon (SOC) and lower total N reservoirs than a 2000-year-old natural marsh.[16]

Their C accumulation rates are similar to those of reference sites, but N accumulation rates are higher, thus C/N ratios have declined over time. In Oregon, 95 restored freshwater marshes had lower soil organic matter than natural marshes and no evidence of accumulation.[17]

Because bottomland hardwood forests combine both a long-term carbon storage in tree biomass and a slow release of soil organic carbon under flooded conditions, they are considered as excellent ecosystems that could potentially be used to sequester carbon. Restoration of bottomland hardwood forests on marginal agricultural lands in the Mississippi River Valley offers significant net carbon sequestration in the south-central region of the U.S.[13] However hardwood plantings have problems with slow initial growth and excessive early mortality rates. The use of herbaceous weed control and bedding may offer the potential to overcome these difficulties.[13] Under the initiative of the U.S. Department of Energy, an 80 acre pilot study is currently underway in Louisiana. Abandoned marginally productive agricultural fields were planted in January 1997 with seven bottomland hardwood species. Following planting, the site will be monitored for planting success and survival. When the trees attain heights >4.5 ft, growth data and carbon sequestration per acre will be calculated.[13]

Conclusions

The primary production in wetland ecosystems is greatly enhanced by hydrological fluxes, which leads to high biomass, primary production and accretion rates. The carbon that is either fixed stored in standing biomass, released by soil decomposing microorganisms, or stored in soil sediments. With the continual destruction of wetlands worldwide, more of this sequestered carbon is released to the atmosphere. The protection of wetlands will preserve the amount of carbon already stored in these ecosystems.

The creation and restoration of new wetlands will certainly contribute to sequestration of carbon and should be considered as one way to mitigate greenhouse emissions.

References

1. Houghton, J.T., Meira Filho, L.G., Bruce, J., Hoesung, L., Callander, B.A., Haites, E., Harris, N., Maskell, K., Eds.; IPCC (Intergovernemental Panel on Climate Change), climate change 1994. In *Radiative Forcing of Climate Change and an Evaluation of the IPCC IS92 Emission Scenarios;* Cambridge University Press: Cambridge, U.K., 1995.
2. Westlake, D.F., Kvet, J., Szczepanski, A., Eds.; *The Primary Ecology of Wetlands;* Cambridge University Press: Cambridge, 1998; 568 pp.
3. Mitsch, W.J.; Gosselink, J.G. *Wetlands,* 3rd Ed.; Wiley: New York, 2000; 920 pp.
4. Odum, W.E.; Odum, E.P.; Odum, H.T. Natures pulsing paradigm. Estuaries **1995**, *18* (4), 547–555.
5. Conner, J. Effects of forest management practices on southern forested wetland productivity. Wetland **1994**, *14,* 27–40.

6. Gorham, E. Northern peatlands: role in the carbon cycle and probable responses to climatic warming. Ecol. Appl. **1991**, *1*, 182–195.
7. Roulet, N.T. Peatlands, carbon storage, greenhouse gases, and the kyoto protocol: prospects and significance for Canada. Wetlands **2000**, *20* (4), 605–615.
8. Clymo, R.S. Models of peat growth. SUO **1993**, *43,* 127–136.
9. Tamm, C.O. *Nitrogen in Terrestrial Ecosystems;* Springer: Berlin, 1991.
10. Vitousek, P.M.; Howarth, R.W. Nitrogen limitation on land and in the sea: how can it occur? Biogeochemistry **1991**, *75,* 87–115.
11. Vitousek, P.M.; Aber, J.D.; Howarth, R.W.; Likens, G.E.; Matson, P.A.; Schindler, D.W.; Schlesinger, W.H.; Tilman, D.G. Human alterations of the global nitrogen cycle: sources and consequences. Ecol. Appl. **1997**, *7* (3), 737–750.
12. Davidsson, T.E.; Stahl, M. The influence of organic carbon on nitrogen transformation in five wetland soils. Soil Sci. Soc. Am. J. **2000**, *64,* 1129–1136.
13. Williams, J.R. Addressing global warming and biodiversity through forest restoration and coastal wetlands creation. Sci. Total Environ. **1999**, *240,* 1–9.
14. Mitsch, W.J.; Wu, X.; Nairn, R.W.; Weihe, P.E.; Wang, N.; Deal, R.; Boucher, C.E. Creating and restoring wetlands: a whole ecosystem experiment in self-design. Bio Science **1998**, *48,* 1019–1030.
15. Zedler, J.B. Progress in wetland restoration ecology. Trends Ecol. Evol. **2000**, *15* (10), 402–407.
16. Craft, C.; Reader, J.; Sacco, J.N.; Broome, S.W. Twenty-five years of ecosystem development of constructed Spartina Alterniflora (Loisel) marshes. Ecol. Appl. **1999**, *9* (4), 1405–1419.
17. Shaffer, P.; Ernst, T. Distribution of soil organic matter in freshwater emergent/open water wetlands in the Portland, Oregon metropolitan area. Wetlands **1999**, *19,* 505–516.

Index

Page numbers followed by f and t indicate figures and tables, respectively.

SECOND EDITION
ENVIRONMENTAL MANAGEMENT HANDBOOK
VOLUME V

Managing Air Quality and Energy Systems

edited by
Brian D. Fath
Sven E. Jørgensen

CRC Press
Taylor & Francis Group

Managing Air Quality and Energy Systems

Environmental Management Handbook, Second Edition

Edited by
Brian D. Fath and Sven E. Jørgensen

Volume 1
Managing Global Resources and Universal Processes

Volume 2
Managing Biological and Ecological Systems

Volume 3
Managing Soils and Terrestrial Systems

Volume 4
Managing Water Resources and Hydrological Systems

Volume 5
Managing Air Quality and Energy Systems

Volume 6
Managing Human and Social Systems

Managing Air Quality and Energy Systems

Second Edition

Edited by
Brian D. Fath and Sven E. Jørgensen

Assistant to Editor
Megan Cole

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

Cover photo: Hounslow, United Kingdom, N. Fath

Second edition published 2021
by CRC Press
6000 Broken Sound Parkway NW, Suite 300, Boca Raton, FL 33487-2742

and by CRC Press
2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

First edition published by CRC Press 2013

CRC Press is an imprint of Taylor & Francis Group, LLC

Library of Congress Cataloging-in-Publication Data
Names: Fath, Brian D., editor. | Jørgensen, Sven Erik, 1934- editor.
Title: Environmental management handbook / edited by Brian D. Fath, Sven Erik Jorgensen.
Other titles: Encyclopedia of environmental management.
Description: Second edition. | Boca Raton : CRC Press, 2020. | Series: Applied ecology and environmental management | Includes bibliographical references and index. | Contents: volume 1. Managing global resources and universal processes — volume 2. Managing biological and ecological systems — volume 3. Managing soils and terrestrial systems — volume 4. Managing water resources and hydrological systems — volume 5. Managing air quality and energy systems — volume 6. Managing human and social systems. | Summary: "The Environmental Management Handbook is an excellent resource for finding basic knowledge on environmental systems. It reflects an extensive coverage of the field and includes the most important problems and solutions posed to environmental management today. In a very practical way, the handbook demonstrates the key processes and provisions for enhancing environmental management. The experience, evidence, methods, and models used in studying environmental management is presented here in six stand-alone volumes, arranged along the major environmental systems. The chapters are contributed by leading experts from around the globe"— Provided by publisher.
Identifiers: LCCN 2020015980 (print) | LCCN 2020015981 (ebook) | ISBN 9781138342637 (v. 1 ; hardback) | ISBN 9781138342644 (v. 2 ; hardback) | ISBN 9781138342651 (v. 3 ; hardback) | ISBN 9781138342668 (v. 4 ; hardback) | ISBN 9781138342675 (v. 5 ; hardback) | ISBN 9781138342682 (v. 6 ; hardback) | ISBN 9780429346132 (v. 1 ; ebook) | ISBN 9780429346170 (v. 2 ; ebook) | ISBN 9780429346255 (v. 3 ; ebook) | ISBN 9781003045045 (v. 4 ; ebook) | ISBN 9781003043461 (v. 5 ; ebook) | ISBN 9781003053514 (v. 6 ; ebook)
Subjects: LCSH: Environmental management—Handbooks, manuals, etc.
Classification: LCC GE300 .E54 2020 (print) | LCC GE300 (ebook) | DDC 628.03—dc23
LC record available at https://lccn.loc.gov/2020015980
LC ebook record available at https://lccn.loc.gov/2020015981

ISBN: 978-1-138-34267-5 (hbk)
ISBN: 978-0-367-48980-9 (pbk)
ISBN: 978-1-003-04346-1 (ebk)

Typeset in Minion
by codeMantra

Contents

SECTION I APC: Anthropogenic Chemicals and Activities

SECTION II COV: Comparative Overviews of Important Topics for Environmental Management

SECTION III CSS: Case Studies of Environmental Management

SECTION IV DIA: Diagnostic Tools: Monitoring, Ecological Modeling, Ecological Indicators, and Ecological Services

SECTION V ENT: Environmental Management Using Environmental Technologies

SECTION VI PRO: Basic Environmental Processes

Preface

Given the current state of the world as compiled in the massive Millennium Ecosystem Assessment Report, humans have changed ecosystems more rapidly and extensively during the past 50 years than in any other time in human history. These are unprecedented changes that need certain action. As a result, it is imperative that we have a good scientific understanding of how these systems function and good strategies on how to manage them.

In a very practical way, this multi-volume *Environmental Management Handbook* provides a comprehensive reference to demonstrate the key processes and provisions for enhancing environmental management. The experience, evidence, methods, and models relevant for studying environmental management are presented here in six stand-alone thematic volumes, as follows:

VOLUME 1 – Managing Global Resources and Universal Processes
VOLUME 2 – Managing Biological and Ecological Systems
VOLUME 3 – Managing Soils and Terrestrial Systems
VOLUME 4 – Managing Water Resources and Hydrological Systems
VOLUME 5 – Managing Air Quality and Energy Systems
VOLUME 6 – Managing Human and Social Systems

In this manner, the handbook introduces in the first volume the general concepts and processes used in environmental management. The next four volumes deal with each of the four spheres of nature (biosphere, geosphere, hydrosphere, and atmosphere). The last volume ties the material together in its application to human and social systems. These are very important chapters for a wide spectrum of students and professionals to understand and implement environmental management. In particular, the features include the following:

- The first handbook that demonstrates the key processes and provisions for enhancing environmental management.
- Addresses new and cutting-edge topics on ecosystem services, resilience, sustainability, food–energy–water nexus, socio-ecological systems, etc.
- Provides an excellent basic knowledge on environmental systems, explains how these systems function, and gives strategies on how to manage them.
- Written by an outstanding group of environmental experts.

Since the handbook covers such a wide range of materials from basic processes, to tools, technologies, case studies, and legislative actions, each handbook entry is further classified into the following categories:

APC: Anthropogenic chemicals—the chapters cover human-manufactured chemicals and activities
COV: Indicates that the chapters give comparative overviews of important topics for environmental management

CSS: The chapters give a case study of a particular environmental management example
DIA: Means that the chapters are about diagnostic tools—monitoring, ecological modeling, ecological indicators, and ecological services
ELE: Focuses on the use of legislation or policy to address environmental problems
ENT: Addresses environmental management using environmental technologies
NEC: Natural elements and chemicals—the chapters cover basic elements and chemicals found in nature
PRO: The chapters cover basic environmental processes.

Overall, these volumes will be a valuable resource for all libraries supporting programs in environmental science and studies, earth science, geography, and policy.

In this volume, #5, the focus is on managing air quality and the closely related topic of energy systems, as represented in over 50 entries. Energy basics and physics for conventional and alternative sources are considered. Specific impacts such as global climate change, acid rain, and ozone are covered. New entries include specific tools to measure road traffic emissions, the importance of managing micropower grids, and the role of individual and household behavior in emission scenarios. Case studies look at energy conversion and the impact of wind farm noise. This volume contains a number of entries on air pollution control strategies.

Brian D. Fath
Brno, Czech Republic
December 2019

Editors

Brian D. Fath is Professor in the Department of Biological Sciences at Towson University (Maryland, USA) and Senior Research Scholar at the International Institute for Applied Systems Analysis (Laxenburg, Austria). He has published over 180 research papers, reports, and book chapters on environmental systems modeling, specifically in the areas of network analysis, urban metabolism, and sustainability. He co-authored the books *A New Ecology: Systems Perspective* (2020), *Foundations for Sustainability: A Coherent Framework of Life–Environment Relations* (2019), and *Flourishing Within Limits to Growth: Following Nature's Way* (2015). He is also Editor-in-Chief for the journal *Ecological Modelling* and Co-Editor-in-Chief for *Current Research in Environmental Sustainability*. Dr. Fath was the 2016 recipient of the Prigogine Medal for outstanding work in systems ecology and twice a Fulbright Distinguished Chair (Parthenope University, Naples, Italy, in 2012 and Masaryk University, Czech Republic, in 2019). In addition, he has served as Secretary General of the International Society for Ecological Modelling, Co-Chair of the Ecosystem Dynamics Focus Research Group in the Community Surface Modeling Dynamics System, and member and past Chair of Baltimore County Commission on Environmental Quality.

Sven E. Jørgensen (1934–2016) was Professor of environmental chemistry at Copenhagen University. He received a doctorate of engineering in environmental technology and a doctorate of science in ecological modeling. He was an honorable doctor of science at Coimbra University (Portugal) and at Dar es Salaam (Tanzania). He was Editor-in-Chief of *Ecological Modelling* from the journal inception in 1975 until 2009. He was Editor-in-Chief for the *Encyclopedia of Environmental Management* (2013) and *Encyclopedia of Ecology* (2008). In 2004, Dr. Jørgensen was awarded the Stockholm Water Prize and the Prigogine Medal. He was awarded the Einstein Professorship by the Chinese Academy of Sciences in 2005. In 2007, he received the Pascal Medal and was elected a member of the European Academy of Sciences. He had published over 350 papers, and has edited or written over 70 books. Dr. Jørgensen gave popular and well-received lectures and courses in ecological modeling, ecosystem theory, and ecological engineering worldwide.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Contributors

Kalyan Annamalai
Paul Pepper Professor of Mechanical Engineering
Texas A&M University
College Station, Texas

Senthil Arumugam
Enerquip, Inc.
Medford, Wisconsin

Muhammad Asif
School of the Built and Natural Environment
Glasgow Caledonian University
Glasgow, United Kingdom

Lu Aye
Renewable Energy and Energy Efficiency Group
Department of Infrastructure Engineering
Melbourne School of Engineering
University of Melbourne
Melbourne, Victoria, Australia

Udo Becker
Chair of Transport Ecology
Faculty of Transportation Sciences "Friedrich List"
Institute of Transport Planning and Road Traffic
Dresden University of Technology
Dresden, Germany

John O. Blackburn
Professor Emeritus of Economics
Duke University
Maitland, Florida

Pascal Boeckx
Faculty of Agricultural and Applied Biological Sciences
University of Ghent
Ghent, Belgium

Jens Borken-Kleefeld
Air Quality and Greenhouse Gases
International Institute for Applied Systems Analysis
Laxenburg, Austria

Felix Creutzig
Mercator Research Institute on Global Commons and Climate Change (MCC)
and
Technical University of Berlin
Berlin, Germany

Bernd Delakowitz
Faculty of Mathematics and Natural Sciences
University of Applied Sciences
Zittau, Germany

Harvey E. Diamond
Energy Management International
Conroe, Texas

Ibrahim Dincer
Faculty of Engineering and Applied Science
University of Ontario Institute of Technology (UOIT)
Oshawa, Ontario, Canada

Svetlana Drozdova
Institute of Chemical Technologies
and Analytics
Vienna University of Technology
Vienna, Austria

Sarina J. Ergas
Department of Civil and Environmental
Engineering
University of Massachusetts
Amherst, Massachusetts

Sara Evangelisti
Interuniversity Research Center for Sustainable
Development (CIRPS)
Sapienza University of Rome
Rome, Italy

Andrew Fang
Humphrey School of Public Affairs
University of Minnesota
Minneapolis, Minnesota

Stefan Fraenzle
Department of Biological and Environmental
Sciences
Research Group of Environmental Chemistry
International Graduate School Zittau
Zittau, Germany

Regine Gerike
Chair of Integrated Transport Planning and
Traffic Engineering
Faculty of Transportation Sciences
"Friedrich List"
Institute of Transport Planning and
Road Traffic
Dresden University of Technology
Dresden, Germany

Fredric S. Goldner
Energy Management & Research Associates
East Meadow, New York

Dan Golomb
Department of Environmental, Earth and
Atmospheric Sciences
University of Massachusetts—Lowell
Lowell, Massachusetts

Alex E.S. Green
Professor Emeritus
University of Florida
Gainesville, Florida

Eliane Tigre Guimarães
Experimental Air Pollution Laboratory
Department of Pathology
School of Medicine
University of Sao Paulo
Sao Paulo, Brazil

Ryan Hanna
School of Global Policy and Strategy
University of California San Diego
La Jolla, California

Chris Hanning
Sleep Medicine
University Hospitals of Leicester
Leicester, United Kingdom

Fabian Heidegger
Chair of Integrated Transport Planning and
Traffic Engineering
Faculty of Transportation Sciences
"Friedrich List"
Institute of Transport Planning and Road Traffic
Dresden University of Technology
Dresden, Germany

Steven D. Heinz
Good Steward Software
State College, Pennsylvania

Arif Hepbasli
Faculty of Engineering
Department of Energy Systems Engineering
Yaşar University
Bornova, Izmir, Turkey

Nathan E. Hultman
University of Maryland
College Park, Maryland

Keith E. Idso
Center for the Study of Carbon Dioxide and
Global Change
Tempe, Arizona

Sherwood Idso
Center for the Study of Carbon Dioxide and Global Change
Tempe, Arizona

Kui Jiao
Department of Mechanical Engineering
University of Waterloo
Waterloo, Ontario, Canada

Sven Erik Jørgensen
Institute A, Section of Environmental Chemistry
Copenhagen University
Copenhagen, Denmark

Gholamreza Karimi
Department of Mechanical Engineering
University of Waterloo
Waterloo, Ontario, Canada

Janey Kaster
Yamas Controls West
San Francisco, California

Ronald L. Klaus
VAST Power Systems
Elkhart, Indiana

Ewa Klugmann-Radziemska
Chemical Faculty
Gdansk University of Technology
Gdansk, Poland

Milivoje M. Kostic
Department of Mechanical Engineering
Northern Illinois University
DeKalb, Illinois

Umesh Kulshrestha
School of Environmental Sciences
Jawaharlal Nehru University
New Delhi, India

Xianguo Li
Department of Mechanical Engineering
University of Waterloo
Waterloo, Ontario, Canada

Bernd Markert
Environmental Institute of Scientific Networks (EISN)
Haren-Erika, Germany

Rudolf Marloth
San Diego State University
San Diego, California

Andrea Micangeli
Interuniversity Research Center for Sustainable Development (CIRPS)
Sapienza University of Rome
Rome, Italy

Adnan Midili
Department of Mechanical Engineering
Faculty of Engineering
Nigde University
Nigde, Turkey

Luisa T. Molina
Massachusetts Institute of Technology
Cambridge, Massachusetts

Martin A. Mozzo, Jr.
M and A Associates Inc.
Robbinsville, New Jersey

Tariq Muneer
School of Engineering
Napier University
Edinburgh, United Kingdom

Leila Niamir
Mercator Research Institute on Global Commons and Climate Change (MCC)
Berlin, Germany

and

University of Twente
Enschede, The Netherlands

K.E. Ohrn
Cypress Digital Ltd.
Vancouver, British Columbia, Canada

Bohdan W. Oppenheim
U.S. Department of Energy Industrial Assessment Center
Loyola Marymount University
Los Angeles, California

Andrea Nunes Vaz Pedroso
Nucleus Research in Ecology
Institute of Botany
Sao Paulo, Brazil

Soyuz Priyadarsan
Texas A&M University
College Station, Texas

Anu Ramaswami
Humphrey School of Public Affairs
University of Minnesota
Minneapolis, Minnesota

Marc A. Rosen
Faculty of Engineering and Applied Science
University of Ontario Institute of Technology
Oshawa, Ontario, Canada

Erwin Rosenberg
Institute of Chemical Technologies and Analytics
Vienna University of Technology
Vienna, Austria

Danilo Sbordone
Interuniversity Research Center for Sustainable Development (CIRPS)
Sapienza University of Rome
Rome, Italy

Wolfram Schmidt
Chair of Transport Ecology
Faculty of Transportation Sciences "Friedrich List"
Institute of Transport Planning and Road Traffic
Dresden University of Technology
Dresden, Germany

Daniel Shepherd
Auckland University of Technology
Auckland, New Zealand

Leslie A. Solmes
LAS and Associates
Mill Valley

Amanda Staudt
Climate Scientist
National Wildlife Federation
Reston, Virginia

Therese Stovall
Oak Ridge National Laboratory
Oak Ridge, Tennessee

John M. Sweeten
Texas A&M University
Amarillo, Texas

Bob Thorne
Massey University
Palmerston North, New Zealand

Jill S. Tietjen
Technically Speaking, Inc.
Greenwood Village, Colorado

Wayne C. Turner
Industrial Engineering and Management
Oklahoma State University
Stillwater, Oklahoma

Robert E. Uhrig
Department of Nuclear Engineering
University of Tennessee
Knoxville, Tennessee

Oswald Van Cleemput
Faculty of Agricultural and Applied Biological Sciences
University of Ghent
Ghent, Belgium

George F. Vance
Department of Ecosystem Sciences and Management
University of Wyoming
Laramie, Wyoming

Waldemar Wardencki
Department of Chemistry
Gdansk University of Technology
Gdansk, Poland

Eric A. Woodroof
Profitable Green Solutions
Plano, Texas

Simone Wuenschmann
Environmental Institute of Scientific Networks (EISN)
Haren-Erika, Germany

Kazuyuki Yagi
National Institute for Agro-Environmental Sciences
Ibaraki, Japan

Andrew R. Zimmerman
Department of Geological Sciences
University of Florida
Gainesville, Florida

I

APC: Anthropogenic Chemicals and Activities

1

Genotoxicity and Air Pollutions

Eliane Tigre Guimarães and Andrea Nunes Vaz Pedroso

Air Pollutants

Air pollution can be generated by natural and anthropogenic sources. The natural sources, such as electrical discharge, decomposition of organic matter, volcano eruption, and natural fires, do not depend on human actions and emit large amounts of pollution, usually in restricted and sparsely populated areas.

The anthropogenic sources can be stationary or mobile. Stationary sources are mainly industries that cause local problems of air contamination. Their pollution emissions are determined by the characteristics of the manufacturing processes, which include the sort of raw materials and fuels used and the products furnished, as well as by the efficiency of the industrial processes and the control measurements adopted.

Mobile sources consist of automotive vehicles, trains, airplanes, ships, and motorboats, which release pollutants into the atmosphere due to incomplete burning of fossil fuels. However, the automotive vehicles are the main mobile sources.

The atmosphere in large cities is usually contaminated by a range of pollutants from stationary and mobile sources. The pollutant emissions of mobile sources are difficult to be controlled, mainly because of the increasing number of automotive vehicles in the last 50 years. This number increased tenfold during this period.[1] The emissions of air pollutants in urban centers have been causing great concerns all over the world and have been causing harmful effects on living organisms.

The main pollutants in urban centers are gases such as carbon, nitrogen, and sulfur oxides, and organic compounds such as hydrocarbons, volatile organic compounds (VOCs), and particulate matter. They will be described below.

The gas carbon monoxide is an odorless, colorless, and tasteless gas formed during the incomplete combustion of carbon-containing fuels.[2]

The main source of sulfur dioxide (SO_2) is the combustion of fuels. Fossil fuels have 1%–5% sulfur in their composition. During combustion, the sulfur is converted to SO_2. Nowadays, in developed countries, a large quantity of the sulfur is removed from motor fuels during the refining process and gas emission from chimneys. However, in developing countries, unabated burning of coal and the use of fuel oils and automotive diesel with higher sulfur content are major sources of SO_2.[2]

The nitrogen derived from the combustion process of fossil fuels is converted to nitrogen oxides[2] such as nitrogen monoxide (NO) and nitrogen dioxide (NO_2). They are considered the precursors of tropospheric ozone (O_3) formation. Nitrogen dioxide diffuses into the atmosphere, where it is usually oxidized and can react with water to form acid rain, causing corrosion in materials and damage to human beings.[3]

The nitrogen oxides and VOCs are considered precursors and produce by photochemical reactions many secondary pollutants, among them O_3 and peroxyacetyl nitrate,[1,3] which compose so-called photochemical smog.

When O_3 is formed in an atmosphere without pollutants, it is consumed within minutes by the photostationary equilibrium between NO and NO_2. Nevertheless, in a polluted atmosphere, NO is converted to NO_2 and can be consumed by RO_2 (organic radical), and, as a consequence, O_3 is accumulated.[3,4] The O_3 is considered one of the most damaging gaseous pollutants to human health and plants, because it forms the reactive oxygen species (ROS) such as superoxide, hydrogen peroxide, and hydroxyl, among others. Reactive oxygen species are oxidative and affect lipids, proteins, and nucleic acids; the cell membranes, composed by polyunsaturated fatty acids, represent the initial target of ROS, changing their permeability and triggering lipid peroxidation,[5] amino acid oxidation, and inactivation of enzymes.[6]

The particulate matter (PM) is a mixture of solid or liquid particles suspended in the air, including smoke, fumes, soot, and other combustion by-products, besides natural particles such as wind-blown dust, sea salt, pollen, and spores.[7] These components can be characterized by their size and composition.[1] Based on the aerodynamic diameter, which ranges from 0.002 to 100 μm, the particulate matter is classified into three categories: 1) coarse particles, ranging from 2.5 to 100 μm; 2) fine particulate matter, below 2.5 μm; 3) ultrafine particles, below 0.1 μm.[8]

Air quality is now regulated by standard concentrations established by laws, based on experiments on humans and/or animals and epidemiological investigations. The standards for air pollutants in Europe are proposed by the European Commission and in the United States by the Environmental Protection Agency (EPA). In Brazil, the standard values for air quality control are defined by a resolution proposed by the National Council of Environment (CONAMA; Table 1).[9]

Although these standards are often revised in order to protect the human health, the World Health Organization states that around 2.4 million people still die each year due to causes related to air pollution.

TABLE 1 Standard Values for Pollutants Established by Environmental Agencies (European Commission, EPA, and CONAMA)

Pollutant	Sampling Time	European Commission	EPA	CONAMA
CO (carbon monoxide)	1 hr	26 ppm	35 ppm	35 ppm
	8 hr	10 mg/m^3	–	–
NO_2	8 hr	–	–	9 ppm
	1 hr	200 μg/m^3	100 ppb	320 μg/m^3
PM_{10} (particulate matter)	Annual	40 μg/m^3	53 ppb	100 μg/m^3
	24 hr	50 μg/m^3	150 μg/m^3	150 μg/m^3
$PM_{2,5}$ (particulate matter)	Annual	25 μg/m^3	150 μg/m^3	50 μg/m^3
	24 hr	–	35 μg/m^3	–
O_3	8 hr	120 μg/m^3	80 ppb	120 μg/m^3
	24 hr	–	120 ppb	125 μg/m^3
SO_2	24 hr	125 μg/m^3	140 ppb	125 μg/m^3
	Annual	–	75 ppb	80 μg/m^3

Source: Adapted from “Air Quality Standards,”[16] “National Ambient Air Quality Standards,”[17] and “Qualidade do Ar.”[18]
–, Limit not defined.

About 1.5 million deaths are attributable to indoor air pollution (estimated deaths).[2] Epidemiological studies suggest that Americans and Europeans have high rates of deaths from cardiopulmonary diseases arising from air pollution.[10] Worldwide, the number of deaths per year caused by pollution is greater than that caused by car accidents.[11] The individual response to air pollutants depends of the type of pollutant, the degree of exposure, the health conditions and the individual genetics[12] and still, socioeconomic profile.[13,14]

Air pollutants affect the vital molecules of human beings, such as nucleic acids, causing genotoxic effects, among numerous other health problems. Thus, this will be the main focus of this entry from here on. The genotoxic effects most commonly reported in the literature and bioassays proposed for prognosis of genotoxic risks will be reported. Finally, other effects to human health will be mentioned at the end of this entry.

Genotoxic Effects of Air Pollutants

Genotoxicity is defined as every alteration occurring in genetic material that causes loss of cellular integrity.[15]

The literature defines that mutations can modify the amino acid sequence of the gene encoding the protein or damage in the DNA molecule. They may occur over the life span of living beings. This process is of extreme importance for the evolution of species. Among gene mutations, we may find the following cases: 1) when a single nucleotide base is replaced by another; and 2) when extra base pairs are added or deleted from the DNA. These are also referred to as insertions and deletions, respectively. These mutations can be devastating, because the messenger RNA is translated into new groups of three nucleotides and the protein produced can cause serious damage.

The mutations that encompass larger portions of DNA are called macrolesions, which change the structure of chromosomes, resulting in damage of the genotype and phenotype of the organism, such as translocations, inversions, deletions, and duplications. They occur most frequently during meiosis.[19]

The mutation may result from exposure to different environmental or chemical agents. Throughout the numerous and successive divisions, the cell can accumulate a large number of mutations and trigger the loss of division control and contribute to the initial stage of tumor development.[20] Therefore, the genotoxic agents are considered mutagenic or carcinogenic,[15] and there is a slight difference between these classifications: 1) mutagenic effect is an alteration in the genetic material of the cell of a living organism that is more or less permanent and that can be transmitted to the cell's descendants; and 2) carcinogenic effects are caused by genotoxic substances that can produce tumors, abnormal tissue growths caused by a loss of control in cell replication. Nevertheless, repairing mechanisms of DNA are present in all organisms, and their complexity is directly proportional to the complexity of the organism.[21] The mechanisms can be classified into direct reversal, damage excision, or recombination.[22]

In human beings, most of the pollutants studied are particulate matter.[7,23,24] In the previous section, we said that smaller fractions of PM, i.e., $<2.5\,\mu m$, are more harmful, and this was confirmed by Rossner and colleagues,[25] who analyzed the organic fractions of PM of polluted sites and found DNA damage. A similar result was found by Coronas and colleagues;[26] they employed cells of lymphocytes and buccal mucosa cells to assess the genotoxic potential of PM in people living and working near refinery oil. The authors used two genotoxicity assays—comet assay and micronucleus assay—and pointed out that the comet assay was more sensitive. These tests will be detailed in the next section.

Other studies show organic extracts of particulate matter to cause genotoxicity. Roubicek and colleagues[27] found that in regions polluted in Mexico City, organic extracts of PM containing Cd and PAHs induce micronucleus formation in human epithelial cells. In several European cities were collected organic extracts mixtures of PAHs and particulate matter, confirmed this genotoxic, using HepG2 cells have the metabolic capacity for PAHs similar to human hepatocytes and represent the best in vitro model for investigating the genotoxic potential of complex mixtures containing PAHs.[28]

Another study found DNA damage in human lung cells when exposed to particulate matter.[29] Still, Gilli and colleagues[30] obtained positive correlations statistically between PM_{10} and mutagenicity, bioavailable iron, sulfates, and nitrates.

Genotoxicity Tests

Currently, there are numerous protocols with prokaryotic and eukaryotic organisms, which evaluate the mutagenic effects of different substances in order to identify risks that living organisms are exposed to. The genotoxic tests most used to detect genotoxicity of air pollutants will be mentioned below.

The Ames test is also known as the *Salmonella* mutagenicity test. This test was developed by Bruce Ames and colleagues and aimed to evaluate the carcinogenic potential of different substances using mutant strains of auxotrophic *Salmonella typhimurium* with respect to histidine. It detects mutagens that cause the displacement of the reading frame (frameshift) or substitution of base pairs of DNA.[31,32]

Other tests use eukaryotic organisms, aiming to evaluate the mutagenic potential of different substances by means of numerical and/or structural chromosomal abnormalities involving at least 10 million base pairs (10 Mb). Among them, we may include the in vitro cytogenetic test in mammalian cells (mouse lymphoma assay) that quantifies the genetic changes that affect the expression of the TK gene of the enzyme thymidine kinase (tk) cells in cultures of L5178Y tk+/tk– lymphoma mice. Although the mammalian cells present locus heterozygosity, only one gene copy is functional in this mouse strain. Loss of locus heterozygosity of the enzyme thymidine kinase, when the gene is affected by a mutagen, causes the resistance of cells mutated to supplementation of medium with trifluorothymidine (TFT). The TFT causes inhibition of metabolism, preventing cell division and leading to death of cells that have the entire way of nucleotides recovery. This test was developed by Clive and colleagues in 1979[33] and modified by Cole and colleagues in 1990.[34]

The comet assay is also a well-known test. It is used to detect not chromosomal mutations but genomic lesions. The test is based on the technique of gel electrophoresis, which detects DNA damage. Since DNA is negatively charged, the electric current causes migration of small broken pieces through the gel, faster and farther than larger pieces. As a result, the damaged cell looks like a comet, with the pieces of damaged DNA forming the tail. The smaller the pieces of DNA, the more they migrate from the cell body. Therefore, a longer tail with smaller pieces implies a greater genetic damage. The content and fragment length of the tail are directly proportional to the amount of DNA damage.

Among mutagenic tests, the micronucleus test is widely used because it is applicable with different eukaryotic organisms. By definition, the micronucleus is a small nucleus, regarded as a product of breakage of genomic DNA of eukaryotic cells. During cell division, genetic material is duplicated and distributed equally between two daughter cells. Radiation and chemicals can cause chromosomal breakage or damage, affecting the distribution of genetic material between daughter cells. Parts or fragments of chromosomes resulting from this damage can be distributed to any of the daughter cells. It is not incorporated into the new core; they may be presented in the form of micronuclei clearly observable on optical microscope.[35]

The micronucleus test in erythrocytes of bone marrow of rodents was developed by Matter and Schmid[36] and modified by Heddle and Salamone[37] in the following years and more recently by Mavournin and colleagues.[38] The test is based on the fact that the effect of genotoxic agent is observed in polychromatic erythrocytes. In addition, the micronuclei are easily observed, and the frequency of micronuclei is dependent on the sampling time.

The micronucleus test in human peripheral blood lymphocytes is also used to detect the mutagenic potential of substances. Nonetheless, some technical problems occur due to the fragility of the cell and the variability in the process of mitotic lymphocytes. These problems were solved by cell hypotonization[39] and radioactive labeling with cytochalasin B to identify the cells that suffered mitosis by inhibiting cytokinesis without blocking mitosis.[40]

Although studies on animals used to detect the effects of air pollution are successful,[41] simple and more efficient analyses to investigate the environmental risks and to determine the genotoxicity induced by pollution are needed. The bioassays with plants are generally more sensitive than other systems for this purpose.[42] Several studies with genus *Tradescantia* have been considered since 1960 as effective biomarkers for determining the genotoxic potential of air pollutants.[43–50]

The *Tradescantia* micronucleus bioassay is the quantification of micronuclei formed in meiotic prophase I, better seen in the young tetrad stage.[44] Among the genus *Tradescantia,* we may highlight the 4430 clone (hybrid between *Tradescantia hirsutiflora* and *T. subacaulis*).[44,48] A cultivar of *T. pallida* Purpurea from Mexico and Honduras has also been used for the micronucleus test since 1999.[47,49]

Nonetheless, in urban areas, gaseous and particulate pollutants interact with each other, thereby enhancing the genotoxic effects on the living organisms. Some studies with *Tradescantia* showed a significant increase in genotoxic potential in plants exposed in the urban environment compared with the rural environment,[49,51,52] also to detect the genotoxic potential of water-soluble fraction of PM_{10}[47] and a dose of 60 ppb ozone in fumigation chambers.[53]

Furthermore, other tests are also conducted with micronuclei in different plant species such as *Allium cepa* and *Vicia faba*; however, the micronuclei are formed from errors in mitotic division in those species.[50]

Another test used to evaluate the risks caused by mutagens also using the genus *Tradescantia* is the *Tradescantia* stamen hair bioassay.[54] It was developed by Arnold H. S. Sparrow based on the fact that the stamen hair cells of plants are heterozygous for color, making it possible to detect mutations based on the change in pigmentation from blue (dominant) to pink (recessive).[55] A pink mutant cell can continue to divide, giving rise to a series of contiguous pink cells, representing a single mutation event. Two mutant cells separated by blue cells are considered two mutation events.[54,56]

In addition to the genotoxic tests using plant species, there is another test known as pollen abortion, which has the ability to detect lethal mutations in haploid cells (microsporous) that end up affecting the development of pollen grains.[57,58] This was confirmed by Micieta and Murín,[59] who evaluated approximately 40 species of native vegetation in Slovakia subjected to industrial pollution, and they observed a positive relationship, i.e., high rate of pollen abortion in polluted area.

Other plant species used in genotoxic tests, like *Nicotiana tabacum*, in different degrees of polluted environments show the largest amount of DNA damage in high pollutant concentrations.[52]

The genotoxic effects caused by air pollutants are studied with more emphasis on the respiratory system in human beings. Effects as changes in pulmonary functions, modification of biochemical and cellular functions, or secretions could happen in the respiratory epithelium. Pacini and colleagues[60] observed a higher amount of DNA damage in people living in the polluted region in Florence, when compared with those living in less polluted areas in Sardinia, both in Italy. In Suwon, Korea, the genotoxic potential of organic extract of $PM_{2,5}$ in lung bronchial epithelial cells was also detected.[61] In addition, the same genotoxic effect can be observed when human beings were exposed to different concentrations of NO_2.[62] Additionally, Tova-lin and colleagues[63] noted the severity of DNA damage in workers in a large urban center, due to the combination of air pollutants VOC, PM_{25}, and O_3.

A considerable amount of tests to detect the mutagenicity of different substances is available, although only the most used were described. A review on urban air mutagenicity and experimental systems reported that 50% of the studies apply the *Salmonella* assay (Ames test); about 30% apply the plant systems (micronucleus tests, chromosomal aberrations, among others); and the other 20% of the studies used other bioassays (such as damage in DNA), animals, and other combinations of studies.[64]

Air Pollutants and Other Health Effects

Historically, the harmful effects of air pollution on human health have become evident. The association between high levels of pollution and diseases manifested by the population exposed has been long detected. Perhaps the most known episode occurred in London in December 1952, where a thermal inversion was responsible for 4000 deaths in 2 weeks.[65] Other similar episodic events confirm this evidence, for example, Meuse Valley in Belgium in 1930,[66] Donora in Pennsylvania in 1948.[67] and St. Louis in 1985.[68]

There is evidence to support the concept that particulate matter causes human mortality, morbidity,[69,70] and genotoxic effects.[71] There is an association between particulate matter and alteration in the respiratory system with restricted activity and severe breathing conditions (acute bronchitis and asthma), resulting in difficulty in breathing and insomnia in adults,[72] emergency room visits, hospital admissions,[73] and pulmonary vasoconstriction.[74,75] It affects lung growth in rats after chronic exposure[76] and significant association between lung cancer and long-term exposure to fine particles, reinforcing the role of fine particulates in the pathogenesis of lung cancer.[77]

Still, in the circulatory system, the fine particulate matter reaches the alveolar regions, transposes the alveolar capillary barrier, and, as a consequence, intensifies the risk of functional abnormalities,[78] such as acute vascular dysfunction, increases thrombus formation,[79,80] arrhythmia, and sudden death.[81]

The toxicity of carbon monoxide has been widely investigated and is well known. Studies show that a major change in humans is the formation of a stable complex between CO and hemoglobin, called carboxyhemoglobin. It decreases the release transport of oxygen to the tissues via blood.[82] In relation to nitrogen dioxide, epidemiological studies have shown that it affects the respiratory system of humans when inhaled[83] and, in high levels of concentration, can be correlated with increased symptoms of asthmatic bronchitis and reduced lung function in children.[2]

In humans, high concentrations of O3 are associated with reduced forced expiratory volume in 1 sec and forced expiratory flow at 50% and 75% of forced vital capacity.[83] Regions with higher ozone concentrations present a higher incidence of asthmatic patients.[83,84] Additionally, according to Pereira et al.,[85] this pollutant showed a positive correlation with the incidence of lung and larynx cancer. Animal studies suggest that O_3 may damage the ciliate cells of the epithelium with changes in the air–blood barrier permeability, causing an inflammatory response.[86]

The complex mixtures of pollutants may affect the human circulatory system, with changes in the levels of fibrinogen, increases erythrocyte count, and plasma viscosity.[87–89] Some studies evidenced deleterious effects on lung defense mechanism, causing inflammatory changes in the airway and distal lung parenchyma.[90–92] Others confirm the carcinogenic effect resulting from air pollution, which acts as a promoter and/or initiator of pulmonary tumor in mice.[93,94]

Over the past few years, some evidence focused particularly on male fertility and pregnancy, showing the negative effects of urban air pollutants on reproductive health in humans.[95–97] Moreover, few studies have been able to demonstrate an association between air pollution and changes in fertility in women, probably due to multiple factors involved in female reproductive function.[98] Male and female mice exposed to urban pollution in São Paulo, Brazil, show changes in the genus distribution in their offspring, suggesting that air pollution can change the proportion of XY sperm in exposed animals.[99]

Previous studies showed that air pollution has a significant impact on female reproductive function in mice. Exposure to fine particulate matter has been implicated in disruption of the pattern of segregation of inner cell mass and trophectoderm cell lineages at the blastocyst stage,[100,101] an important marker of embryo viability and development potential.[102] A retrospective epidemiological study confirmed the increased risk of early pregnancy loss, which was already observed in experimental studies in women exposed to air pollution. In addition, an association between brief exposure to high levels of environmental particles during pre-conception and early pregnancy loss was found, independently from conception method (natural or after in vitro fertilization treatment), and the risk of miscarriage increased 2.6-fold.[100] Furthermore, a positive association between air pollution and intrauterine mortality was found in a study conducted in São Paulo, southeastern Brazil, suggesting that pollution in São Paulo may promote adverse effects on fetuses.[103]

Conclusion and Remarks

Air pollution is a reality in megacities, and it is intensified mainly by a huge number of vehicles circulating. According to the literature, the major air pollutants that human populations are exposed to in urban areas have a number of substances with carcinogenic activity. These substances can cause mutations and trigger neoplasias.

Therefore, the high levels of vehicular emissions require complex demands and strict actions from the government. These solutions can be taken individually with proper maintenance of vehicles; indeed, need the improvement of public transport, which means efficient and rapid tubes, trains, trams, and buses. Alternative measurements such as rodizio have been implemented in large cities such as São Paulo, Mexico City, and Beijing, aiming to reduce the daily and enormous traffic and, consequently, the direct emissions from sources.

Although not emphasized in this entry, industrial air pollution also should not be underestimated. However, the emissions of industries are more easily controlled and must be applied, followed by periodic monitoring, employment of cleaner technologies, and cleaning up of filters on the chimneys, which should be constructed with enough height to guarantee adequate dispersion of pollutants. Furthermore, it is important to choose raw materials that produce waste with lower pollution potential at the end of the production process. It is mandatory to apply more resources to develop new technologies, which can be developed at universities, research institutes, and private companies. The use of cleaner energy sources, such as wind and biogas, is desirable.

Nevertheless, the pollutants produced locally may not stay in the same place and its surrounding area. It can also damage remote areas of the planet due to long-term transportation. Each country has its own formulation of public policies that impose limits on environment pollution levels. Finally, not only signed agreements among governments will bring significant changes to reduce emissions of pollutants. Individual and local actions organized by nongovernmental organizations, schools, and media are equally important to preserve life.

References

1. Fenger, J. Air pollution in the last 50 years—From local to global. Atmos. Environ. **2009**, *43,* 13–22.
2. WHO—World Health Organization Air quality guidelines; Copenhagen, Denmark, 2005.
3. Freedman, B. *Environmental Ecology;* Academic Press: San Diego, 1995.
4. Kley, D.; Kleinmann, M.; Sanderman, H.; Kruppa, S. Photochemical oxidants: State of the science. Environ. Pollut. **1999**, *100,* 19–42.
5. Nohl, H.; Kozlov, A.V.; Gille; L.; Staniek, K. Cell respiration and formation of reactive oxygen species: Facts and artefacts. Biochem. Soc. Trans. **2003**, *31* (6), 1306–1311.
6. Kovacic, P.; Somanathan, R. Biomechanisms of nanoparticles (toxicants, antioxidants and therapeutics): Electron transfer and reactive oxygen species. J. Nanosci. Nanotechnol. **2010**, *10,* 7919–7930.
7. Dockery, D.W. Health effects of particulate air pollution. Ann. Epidemiol. **2009**, *19*(4), 257–263.
8. Seinfeld, J.H.; Pandis, S.N. *Atmospheric Chemistry and Physics—From Air Pollution to Climate Changes,* 1st Ed; John Wiley and Sons, Inc.: New York, 1949.
9. CETESB—Companhia de Tecnologia de Saneamento Ambiental, Relatório de qualidade do ar do Estado de São Paulo de 2008. Série Relatórios, São Paulo, 2009.
10. Zanobetti, A.; Woodhead, M. Air pollution and pneumonia: The "old man" has a new "friend." Am. J.Respir. Crit. Care Med. **2010**, *181,* 5–6.
11. Enander, R.T.; Gagnon. R.N.; Hanumara, R.C.; Park, E.; Armstrong, T.; Gute, D.M. Environmental health practice: Statistically based performance measurement. Am. J. Pub. Health **2007**, *97,* 819–824.

12. Kaplan, C. Indoor air pollution from unprocessed solid fuels in developing countries. Rev. Environ. Health **2010**, *25,* 221–242.
13. Martins, M.C.H.; Fatigati, F.L.; Véspoli, T.C.; Martins, L.C.; Pereira, L.A.A.; Martins, M.A.; Saldiva, P.H.N.; Braga, A.L.F. Influence of socioeconomic conditions on air pollution adverse health effects in elderly people: An analysis of six regions in São Paulo, Brazil. J. Epidemiol. Commun. Health **2004**, *58,* 41–46.
14. Peled, R. Air pollution exposure: Who is at high risk? Atmos. Environ. **2011**, *45* (12), 1781–1785.
15. Weisburger, J.H. Carcinogenicity and mutagenicity testing, then and now. Mutat. Res. **1999**, *437* (2), 105–112.
16. Air Quality Standards, available at http://ec.europa.eu/envi-ronment/air/quality/standards.htm, (accessed May 3, 2011).
17. National Ambient Air Quality Standards, available at http://www.epa.gov/air/criteria.html, (accessed May 3, 2011).
18. Qualidade do Ar, available at http://www.cetesb.sp.gov.br/ar/qualidade-do-ar/31-publicacoes-e-relatorios, (accessed May 3, 2011).
19. Alberts, B.; Alexander, J.; Lewis, L.; Raff, M.; Roberts, K.; Walter, P. *Molecular Biology of the Cell*, Garland Science: New York, 2002.
20. Cohen, S.M.; Arnold, L.L. Chemical carcinogenesis. Toxicol. Sci. **2011**, *120,* 76–92.
21. Evans, M.D.; Dizdaroglu, M.; Cooke, M. S. Oxidative DNA damage and disease: Induction, repair and significance Mutat. Res. **2004**, *567,* 1–61.
22. Oga, S.; Camargo, M.M.A.; Batistuzzo, J.A.O. *Fundamentos da Toxicologia,* 3rd Ed; Atheneu Editora: São Paulo, 2008.
23. Oh, S.M.; Kim, H.R.; Park Y.J.; Lee, S.Y.; Chung, K.H. Organic extracts of urban air pollution particulate matter (PM2.5)–induced genotoxicity and oxidative stress in human lung bronchial epithelial cells (BEAS-2B cells). Mutat. Res. **2011**, *723*, 142–151.
24. Risom, L.; Møller, P.; Loft, S. Oxidative stress-induced DNA damage by particulate air pollution. Mutat. Res. **2005**, *592,* 119–137.
25. Rossner, P.; Topinka, J.; Hovorka, J.; Milcova, A.; Schmuczerova, J.; Krouzek, J.; Sram, R.J. An acellular assay to assess the genotoxicity of complex mixtures of organic pollutants bound on size segregated aerosol. Part II: Oxidative damage to DNA. Toxicol. Lett. **2010**, *198* (3), 312–316.
26. Coronas, M.V.; Pereira, T.S.; Rocha, J.A.V.; Lemos, A.T.; Fachel, J.M.G.; Salvadori, D.M.F.; Vargas, V.M.F. Genetic biomonitoring of an urban population exposed to mutagenic airborne pollutants. Environ. Int. **2009**, *35*, 1023–1029.
27. Roubicek, D.A.; Gutierrez-Castillo, M.E.; Sordo,M.; Cebrian- Garcia, M.E.; Ostrosky-Wegman, P. Micronuclei induced by airborne particulate matter from Mexico City. Mutat. Res., Genet. Toxicol. Environ. Mutagen. **2007**, *631* (1), 9–15.
28. Sevastyanova, O.; Binkova, B.; Topinka, J.; Sram, R.J.; Kalina, I.; Popov, T.; Novakova, Z.; Farmer, P.B. In vitro genotoxicity of PAH mixtures and organic extract from urban air particles—Part II: Human cell lines. Mutat. Res., Fundam. Mol. Mech. Mutagen. **2007**, *620,* 123–134
29. Bonetta, Sa.; Gianotti, V.; Bonetta, Si.; Gosetti, F.; Oddone, M.; Gennaro, M.C.; Carraro, E. DNA damage in A549 cells exposed to different extracts of $PM_{2.5}$ from industrial, urban and highway sites. Chemosphere **2009**, *77,* 1030–1034.
30. Gilli, G.; Traversi, D.; Rovere, R.; Pignata, C.; Schilirò, T. Chemical characteristics and mutagenic activity of PM_{10} in Torino, a Northern Italian City. Sci. Total Environ. **2007**, *385,* 97–107.
31. Ames, B.N.; Yamasaki, E. The detection of chemical mutagens with enteric bacteria. In *Chemical Mutagens: Principles and Methods for their Detection*; Hollander, A., Ed.; Plenum Press: New York, 1971.
32. Mortelmans, K.; Zeiger, E. The Ames *Salmonella/micro-* some mutagenicity assay. Mutat. Res. **2000**, *455*, 29–60.

33. Clive, D.; Johnson, K.O.; Spector, J.F.S.; Batson, A.G.; Brown, M.M.M. Validation and characterization of the L5178Y/TK+/− mouse lymphoma mutagen assay system. Mutat. Res. **1979**, *589,* 61–108.
34. Cole, J.; McGregor, D.B.; Fox, M.; Thacker, J.; Garner, R.C. Gene mutation assays in cultured mammalian cells. In *Mutagenicity Tests*; Kirkland, D.J., Ed.; Ed Basic, Cambridge Press: Cambridge, U.K., 1990.
35. Ribeiro, L.R.; Salvadori, D.M.F.; Marques, E.K. *Mutagênese ambiental,* 1st Ed; Editora ULBRA: Canoas, Brasil, 2003.
36. Matter, B.; Schmid, W. Treminon-induced chromosomal damage in bone narrow cells of six mammalian species evaluated by the micronucleus test. Mutat. Res. **1971**, *12,* 417–425.
37. Heddle, J.A.; Salamone, M.F. The micronucleus assay in vivo. In Proceedings of the international workshop on short-term tests for chemical carcinogens; Stich, H., San, R.H., Eds.; Springer-Verlag: New York, 1981.
38. Mavournin, K.H.; Blakey, D.H.; Cimino, M.C.; Salamone, M.F.; Heddle, J.A. The in vivo micronucleus assay in mammalian bone marrow and peripheral blood. A report of the U.S. Environmental Protection Agency Gene-Tox Program. Mutat. Res. **1990**, *239,* 29–80.
39. Iskander, O. An improved method for the detection of micronuclei in human lymphocytes. Stain Technol. **1989**, *54,* 221–223.
40. Fenech, M.; Morley, A.A. Measurement of micronuclei in lymphocytes. Mutat. Res. **1985**, *147,* 29–36.
41. Saldiva, P.H.N.; Böhm, G.M. Animal indicator of adverse effects associated with air pollution. Ecosyst. Health **1998**, *4*, 230–235.
42. EPA. Current status of bioassays in genetic genotoxicology (Gene-Tox), 1980.
43. Sparrow, A.H.; Underbrink, A.G.; Rossi, H.H. Mutation induced in *Tradescantia* by small doses of x-rays and neutrons: Analysis of dose–response curves. Science **1972**, *176,* 916–918.
44. Ma, T.H. *Tradescantia* micronuclei (Trad-MCN) test for environmental clastogens. In *In Vitro Toxicity Testing of Environmental Agent. Current and Future Possibilities. Part A: Survey of Test Systems;* Kolber, A.R., Wong, T.K., Grant, L.D., DeWoskin, R.S., Hughes, T.J., Eds.; Plenum Press: New York, 1983; 191–214.
45. Rodrigues, G.S.; Ma, T.H.; Pimentel, D.; Weinstein, L.H. *Tradescantia* bioassays as monitoring systems for environmental mutagenesis—Review. Crit. Rev. Plant Sci. **1997**, *16* (4), 325–359.
46. Arutyunyan, R.M.; Pogosyan, V.S.; Simonyan, E.H.; Atoyants, A.L.; Djigardjian, E.M. *In situ* monitoring of the ambient air around the chloroprene rubber industrial plant using the *Tradescantia-stamen-hair* mutation assay. Mutat. Res. **1999**, *426* (2), 117–120.
47. Batalha, J.R.F.; Guimarães, E.T.; Lobo, D.J.A.; Lichtenfels, A.J.F.C.; Deur, T.; Carvalho, H.A.; Alves, E.S.; Domingos, M.; Rodrigues, G.S.; Saldiva, P.H.N. Exploring the clastogenic effects of air pollutants in São Paulo (Brazil) using the *Tradescantia* micronuclei assay. Mutat. Res. **1999**, *426,* 229–232.
48. Monarca, S.; Feretti, D.; Zanardini, A.; Falistocco, E.; Nardi, G. Monitoring of mutagens in urban air samples. Mutat. Res. **1999**, *426,* 189–192.
49. Guimarães, E.T.; Domingos, M.; Alves, E.S.; Caldini Jr., N.; Lobo, D.J.A.; Lichtenfels, A.J.F.C.; Saldiva, P.H.N. Detection of the genotoxicity of air pollutants in and around the city of São Paulo (Brazil) with the *Tradescantia*-micronucleus (Trad-MCN) assay. Environ. Exp. Bot. **2000**, *44*, 1–8.
50. Misík, M.; Ma, T.H.; Nersesyan, A.; Monarca, S.; Kim, J.K..; Knasmueller, S. Micronucleus assays with *Tradescantia* pollen tetrads: an update. Mutagenesis **2011**, *26* (1), 215–221.
51. Prajapati, S.K.; Tripathi, B.D. Assessing the genotoxicity of urban air pollutants in Varanasi City using *Tradescantia* micronucleus (Trad-MCN) bioassay. Environ. Int. **2008**, *34* (8), 1092–1096.

52. Villarini, M.; Fatigoni, C.; Dominici, L.; Maestri, S.; Ederli, L.; Pasqualini, S.; Monarca, S.; Moretti, M. Assessing the genotoxicity of urban air pollutants using two in situ plant bioassays. Environ. Pollut. **2009**, *157* (12), 3354–3356.
53. Lima, E.D.; de Souza, S.R.; Domingos, M. Sensitivity of *Tradescantia pallida* (Rose) Hunt. 'Purpurea' Boom to genotoxicity induced by ozone. Mutat. Res. **2009**, *675* (1–2), 41–45.
54. Ichikawa, S. *Tradescantia* stamen-hair system as an excellent botanical tester of mutagenicity: Its responses to ionizing radiation and chemical mutagens, and some synergistic effects found. Mutat. Res. **1992**, *270* (1), 3–22.
55. Ma, T.H.; Cabrera, G.L.; Cebulska-Wasilewska, A.; Chen, R.; Loarca, F.; Vandenberg, A.L.; Salamone, M.F. *Tradescantia* stamen hair mutation bioassay. Mutat. Res. **1994**, *310* (2), 211–220.
56. Takahashi, C.S. Teste dos pêlos estaminais de *Tradescantia.* In *Mutagênese, Carcinogênese e Teratogênese: Métodos e critérios de avaliação;* Rabelo-Gay, M.N., Rodrigues, M.A.L.R., Monteleone-Neto, R., Eds.; Rev. Brasileira de Genética—Ribeirão Preto, Sociedade Brasileira de Genética, 1991; 61–66.
57. Malallah, G.; Afzal, M.; Murin, G.; Murin, A.; Abraham, D. Genotoxicity of oil pollution on some species of Kuwaiti flora. Biologia **1997**, *52* (1), 61–70.
58. Micieta, K.; Kunová, K. Phytoindication of genotoxic deterioration of polluted environment. Biológia **2000**, *55*, 75–79.
59. (56) Micieta, K.; Murín, G. Microspore analysis for genotoxicity of a polluted environment. Environ. Exp. Bot. **1996**, *36,* 21–27.
60. Pacini, S.; Giovanelli, L.; Gulisano, M.; Peruzzi, B.; Polli, G.; Boddi, V.; Ruggiero, M.; Bozzo, C.; Stomeo, F.; Fenu, G.; Pezzatini, S.; Pitozzi, V.; Dolara, P. Association between atmospheric ozone levels and damage to human nasal mucosa in Florence, Italy. Environ. Mol. Mutagen. **2003**, *42* (3), 127–135.
61. Oh, S.M.; Kim, H.R.; Park, Y.P.; Lee, S.Y.; Chung, K.H. Organic extracts of urban air pollution particulate matter ($PM_{2.5}$)–induced genotoxicity and oxidative stress in human lung bronchial epithelial cells (BEAS-2B cells). Mutat. Res., Genet. Toxicol. Environ. Mutagen. **2011**, doi:10.106/j.mrgentox.2011.04.003.
62. Koehler, C.; Ginzkey, C.; Friehs, G.; Hackenberg, S.; Froelich, K.; Scherzed, A.; Burghartz, M. Aspects of nitrogen dioxide toxicity in environmental urban concentrations in human nasal epithelium. Toxicol. Appl. Pharmacol. **2010**, *245* (2), 219–225.
63. Tovalin, H.; Valverde, M.; Morandi, M.T.; Blanco, S.; Whitehead, L.; Rojas, E. DNA damage in outdoor workers occupationally exposed to environmental air pollutants. Oc-cup. Environ. Med. **2006**, *63,* 230–236.
64. Claxton, L.D.; Woodall, G.M.. A review of the rodent carcinogenicity and mutagenicity of ambient air. Mutat. Res. **2007**, *636* (1–3), 36–94.
65. Logan, W.P.D. Mortality in the London fog incident, 1952. Lancet **1953**, *1,* 336–338.
66. Firket, J. Fog along the Meuse Valley. Trans. Faraday Soc. **1936**, *32,* 1192–1197.
67. Helfand, W.H.; Lazarus, J.; Theerman, P. Dondora, Pennsylvania: An environmental disaster of the 20th century. Am. J. Pub. Health **2001**, *91* (4), 553.
68. Dockery, D.W.; Schwartz, J.; Spengler, J.D. Air pollution and daily mortality: Associations with particulates and acid aerosol. Environ. Res. **1992**, *59* (2), 362–373.
69. Dockery, D.W. Epidemiologic evidence of cardiovascular effects of particulate air pollution. Environ. Health Perspect. **2001**, *109* (4), 483–486.
70. Lin, C.A.; Pereira, L.A.A.; Conceição, G.M.S.; Kishi, H.S.; Milani, R., Jr.; Braga, A.L.F.; Saldiva, P.H.N. Association between air pollution and ischemic cardiovascular emergency room visits. Environ. Res. **2003**, *92* (1), 57–63.
71. Avogbe, P.H.; Ayi-Fanou, L.; Autrup, H.; Loft, S.; Fayomi, B; Sanni, A.; Vinzents, P.; Møller, P. Ultrafine particulate matter and high-level of benzene urban air pollution in relation to oxidative DNA damage. Carcinogenesis **2005**, *26* (3), 613–620.

72. Ostro, B.D.; Rothschild, S. Air pollution and acute respiratory morbidity: An observational study of multiple pollutants. Environ. Res. **1989**, *50* (2), 238–247.
73. Ostro, B. The association of air pollution and mortality: Examining the case for inference. Arch. Environ. Health **1993**, *48(5),* 336–342.
74. Batalha, J.R.F.; Saldiva, P.H.N.; Clarke, R.W.; Coull, B.A.; Stearns, R.C.; Lawrence, J.; Krishna Murthy, G.G.; Koutrakis, P.; Godleski, J.J. Concentrated ambient air particles induce vasoconstriction of small pulmonary arteries in rats. Environ. Health Perspect. **2002**, *110,* 1191–1197.
75. Rivero, D.H.R.F.; Soares, S.R.C.; Lorenzi-Filho, G.; Saiki, M.; Godleski J.J.; Antonangelo, L.; Dolhnikoff; M.; Saldiva, P.H.N. Acute cardiopulmonary alterations induced by fine particulate matter of São Paulo, Brazil. Toxicol. Sci. **2005**, *85*, 898–905.
76. Mauad, T.; Rivero, D.H.R.F.; Oliveira, R.C.; Lichtenfels, A.J.F.C.; Guimarães, E.T.; Andre, P.A.; Kasahara, D.I.; Bueno, H.M.S.; Saldiva, P.H.N. Chronic exposure to ambient levels of urban particles affects mouse lung development. Am. J.Respir. Crit. Care Med. **2008**, *178,* 721–728.
77. Pope, C.A. III; Burnett, R.T.; Thun, M.J.; Calle, E.E.; Krewski, D.; Ito, K.; Thurston, G.D. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. J. Am. Med. Assoc. **2002**, *287,* 1132–1141.
78. Schwartz, J. Air pollution and blood markers of cardiovascular risk. Environ. Health Perspect. **2001**, *109* (3), 405–409.
79. Lucking, A.J.; Lundback, M.; Mills, N.L.; Faratian, D.; Barath, S.L.; Pourazar, J.; Cassee, F.R.; Donaldson, K.; Boon, N.A.; Badimon, J.J.; Sandstrom, T.; Blomberg, A.; Newby, D.E. Diesel exhaust inhalation increases thrombus formation in man. Eur. Heart J. **2008**, *29* (24), 3043–3051.
80. Törnqvist, H.; Mills, N.L.; Gonzalez, M.; Miller, M.R.; Robinson, S.D.; Megson, I.L.; Macnee, W.; Donaldson, K.; Söderberg, S.; Newby, D.E.; Sandström, T.; Blomberg, A. Persistent endothelial dysfunction in humans after diesel exhaust inhalation. Am. J. Respir. Crit. Care Med. **2007**, *176* (4), 395–400.
81. Stone, P.H.; Godleski, J.J. First steps toward understanding the pathophysiologic link between air pollution and cardiac mortality. Am. Heart J. **1999**, *138,* 804–807.
82. Roughton, F.J.W.; Darling, R.C. The effect of carbon monoxide on the oxyhemoglobin dissociation curve. Am. J. Physiol. **1944**, *141,* 17–31.
83. Schmitzberger, R.; Rhomberg, K.; Buchele, H.; Puchegger, R.; Schmitzberger-Natzmer, D.; Kemmler, G.; Panosch, B. Effects of air pollution on the respiratory tract of children. Pediatr. Pulmonol. **1993**, *15* (2), 68–74.
84. Sousa, S.I.V.; Ferraz, C.; Alvim-Ferraz, M.C.M.; Martins, F.G.; Vaz, L.G.; Pereira, M.C. Spirometric tests to assess the prevalence of childhood asthma at Portuguese rural areas: Influence of exposure to high ozone levels. Environ. Int. **2011**, *37* (2), 474–478.
85. Pereira, F.A.C.; Assunção, J.V.; Saldiva, P.H.N.; Pereira. L.A.A.; Mirra, A.P.; Braga, A.L.F. Influence of air pollution on the incidence of respiratory tract neoplasm. J. Air Waste Manage. Assoc. **2005**, *55* (1), 83–87.
86. Yang, W.; Omaye, S.T. Air pollutants, oxidative stress and human health. Mutat. Res. **2011**, *674,* 45–54.
87. Baskurt, O.K.; Levi, E.; Caglayan, S.; Dikmenoglu, N.; Kutman, M.N. Hematological and hemorheological effects of air pollution. Arch. Environ. Health **1990**, *45* (4), 224–228.
88. Seaton, A.; Soutar, A.; Crawford, V.; Elton, R.; McNerlan, S.; Cherrie, J.; Watt, M.; Agius, R.; Stout, R. Particulate air pollution and the blood. Thorax **1999**, *54,* 1027–1032.
89. Gardner, S.Y.; Lehmann, J.R.; Costa, D.L. Oil fly ash–induced elevation of plasma fibrinogen levels in rats. Toxicol. Sci. **2000**, *56*, 175–180.
90. Böhm, G.M.; Saldiva, P.H.N.; Pasqualucci, C.A.G.; Massad, E.; Martin, M.A.; Zin, W.A.; Cardoso, W.V.; Criado, P.M.P.; Komatsuzaki, M.; Sakae, R.S.; Negri, E.M.; Lemos, M.; Capelozzi, V.D.; Crestana, C.; Silva, R. Biological effects of air pollution in São Paulo and Cubatao. Environ. Res. **1989**, *49,* 208–216.

91. Saldiva, P.H.N.; King, M.; Delmonte, V.L.C.; Macchione, M.; Parada, M.A.C.; Daliberto, M.L.; Sakae, R.S.; Criado, P.M.P.; Silveira, P.L.P.; Zin, W.A.; Böhm, G.M. Respiratory alterations due to urban air pollution: An experimental study in rats. Environ. Res. **1992**, *57* (1), 19–33.
92. Lemos, M.; Lichtenfels, A.J.F.C.; Amaro Jr, E.; Macchione, M.; Martins, M.A.; King, M.; Böhm, G.M.; Saldiva, P.H.N. Quantitative pathology of nasal passages in rats exposed to urban levels of air pollution. Environ. Res. **1994**, *66* (1), 87–95.
93. Reymão, M.S.F.; Cury, P.M.; Lichtenfels, A.J.F.C.; Lemos, M.; Battlehner, C.N.; Conceição, G.M.S.; Capelozzi, V.L.; Montes, G.S.; Júnior, M.F.; Martins, M.A.; Böhm, G.M.; Saldiva, P.H.N. Urban air pollution enhances the formation of urethane-induced lung tumors in mice. Environ. Res. **1997**, *74* (2), 150–158.
94. Cury, P.M.; Lichtenfels, A.J.F.C.; Reymão, M.S.F.; Con- ceição, G.M.S.; Capelozzi, V.L.; Saldiva, P.H.N. Urban levels of air pollution modifies the progression of urethane- induced lung tumours in mice. Pathol. Res. Pract. **2000**, *196*, 627–633.
95. Lee, B.E.; Ha, E.H.; Park, H.S.; Kim, Y.J.; Hong, Y.C.; Kim, H.; Lee, J.T. Exposure to air pollution during different gestational phases contributes to risks of low birth weight. Hum. Reprod. **2003**, *18* (3), 638–643.
96. Rubes, J.; Selevan, S.G.; Evenson, D.P.; Zudova, D.; Vozdova, M.; Zudova, Z.; Robbins, W.A.; Perreaut, S.D. Episodic air pollution is associated with increased DNA fragmentation in human sperm without other changes in semen quality. Hum. Reprod. **2005**, *20* (10), 2776–2783.
97. Silva, I.R.R.; Lichtenfels, A.J.F.C.; Pereira, L.A.A.; Saldiva, P.H.N. Effects of ambient levels of air pollution generated by traffic on birth and placental weights in mice. Fertil. Steril. **2008**, *90* (5), 1921–1924.
98. Tomei, G.; Ciarrocca, M.; Fortunato, B.R.; Capozzella, A.; Rosati, M.V.; Cerratti, D.; Tomao, E.; Anzelmo, V.; Monti, C.; Tomei, F. Exposure to traffic pollutants and effects on 17-beta-estradiol (E2) in female workers. Int. Arch. Occup. Environ. Health **2006**, *80* (1), 70–77.
99. Lichtenfels, A.J.F.C.; Gomes. J.B.; Pieri, P.C.; Miraglia, S.G.E.K.; Hallak, J.; Saldiva, P.H.N. Increased levels of air pollution and a decrease in the human and mouse male-to-female ratio in São Paulo, Brazil. Fertil. Steril. **2007**, *87,* 230–232.
100. Perin, P.M.; Maluf, M.; Januário, D.N.; Saldiva, P.H. Effects of short-term exposure of female mice to diesel exhaust particles on in vitro fertilization and embryo development. Fertil. Steril. **2008**, *90* (1), S206.
101. Maluf, M.; Perin, P.M.; Januário, D.A.N.F.; Saldiva, P.H.N. In vitro fertilization, embryo development, and cell lineage segregation after pre- and/or postnatal exposure of female mice to ambient fine particulate matter. Fertil. Steril. **2009**, *92* (5), 1725–1735.
102. Kovacic, B.; Vlaisavljevic, V.; Reljic, M.; Cizek-Sajko, M. Developmental capacity of different morphological types of day 5 human morulae and blastocysts. Reprod. BioMed. **2004**, *8* (6), 687–694.
103. Pereira, L.A.A.; Loomis, D.; Conceição, G.M.S.; Braga, A.L.F.; Arcas, R.M.; Kishi, H.S.; Singer, J.M.; Böhm, G.M.; Saldiva, P.H.N. Association between air pollution and intrauterine mortality in São Paulo, Brazil. Environ. Health Perspect. **1998**, *106* (6), 325–329.

2

Methane Emissions: Rice

Kazuyuki Yagi

Introduction

The atmospheric concentration of methane (CH_4) has increased rapidly in recent years. Because it is a radiative trace gas and takes part in atmospheric chemistry, the rapid increase could be of significant environmental consequence. Of the wide variety of sources, rice fields are considered an important source of atmospheric CH_4, because the harvest area of rice has increased by about 70% during last 50 years and it is likely that CH_4 emission has increased proportionally. Recent estimates suggest that global emission rates of CH_4 from rice fields account for about 4%–19% of the emission from all sources.[1] Due to the large amount of the global emission from rice cultivation, reduction of CH_4 emission from this source is very important in order to stabilize atmospheric concentration. In addition, because of the possibility of controlling the emission by agronomic practices, rice cultivation must be one of the most hopeful sources for mitigating CH_4 emission.

Processes Controlling CH_4 Emissions from Rice Fields

Table 1 provides a summary of measured methane emissions at a number of specific research sites around the world.[2] It should be noted that methane fluxes from rice fields show pronounced diel and seasonal variations and vary substantially with different climate, soil properties, agronomic practices, and rice cultivars.

Processes involved in CH_4 emission from rice fields are illustrated in Figure. 1. Like other biogenic sources, CH_4 is produced by the activity of CH_4 producing bacteria, or methanogens, as one of the terminal products in the anaerobic food web in paddy soils. Methanogens are known as strict anaerobes that require highly reducing conditions. After soil is flooded, the redox potential of soil decreases rapidly by sequential biochemical reactions.

Flooded paddy soils have a high potential to produce CH_4, but part of CH_4 produced is consumed by CH_4 oxidizing bacteria, or methanotrophs. In rice fields, it is possible that a proportion of CH_4 produced in the anaerobic soil layer is oxidized in the aerobic layers, such as the surface soil–water interface and the rhizosphere of rice plants.

The emission pathways of CH_4 that is accumulated in flooded paddy soils is: diffusion into the flood water, loss through ebullition, and transport through the aerenchyma system of rice plants. In the

TABLE 1 Methane Emission from Rice Fields in Various World Locations[a]

			Season Total	
Country	Daily Average (g/m^2 day)	Flooding Period (days)	Average (g/m^2)	Range (g/m^2)
China	0.19–1.39	75–150	13	10–22
India	0.04–0.46	60	10	5–15
Italy	0.10–0.68	130	36	17–54
Japan	0.01–0.39	110–130	11	3–19
Spain	0.10	120	12	
Thailand	0.04–0.77	80–110	16	4–40
U.S.A.	0.05–0.48	80–100	25	15–35

[a] The data are for the fields without organic fertilizer.

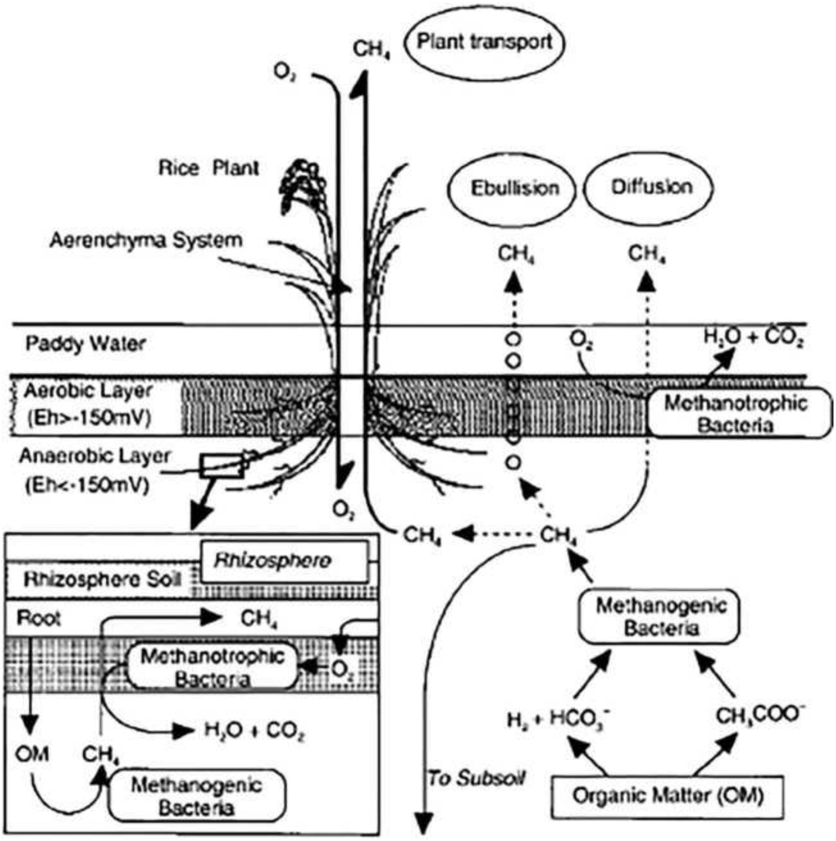


FIGURE 1 Production, oxidation and emission of CH4 in rice paddy fields.
Source: Conrad[3] and Knowles.[4]

temperate rice fields, more than 90% of CH_4 is emitted through plants,[5] while significant amounts of CH_4 may evolve by ebullition, in particular during the early part of the season in the tropical rice fields.[6] Therefore, it is concluded that possible strategies for mitigating CH_4 emission from rice cultivation can be made by controlling either production, oxidation, or transport processes.

Options for Mitigating CH_4 Emission

Water Management

Mid-season drainage (aeration)in flooded rice fields supplies oxygen into soil, resulting in a reduction of CH_4 production and a possible enhancement of CH_4 oxidation in soil.[7,8] A study using an automated sampling and analyzing system clearly showed that short-term drainage had a strong effect on CH_4 emission, as shown in Figure. 2. Total emission rates of CH_4 during the cultivation period were reduced by 42%–45% by short-term drainage practices compared with continuously flooded treatment.[9] These results indicate that improvement in water management can be one of the most promising mitigation strategies for CH_4 emission from rice fields. Increasing the rate of water percolation in rice fields by installing underground pipe drainage may also have an influence on CH_4 production and emission.

Soil Amendments and Mineral Fertilizers

The progress of soil reduction can be retarded by adding one of several electron acceptors in the sequential soil redox reactions. Sulfate is one of the most promising candidates for this strategy because it is commonly used as a component of mineral fertilizer and soil amendment. Field measurements have shown that CH_4 emission rate decreased by at most 55%–70% by application of ammonium sulfate or gypsum.[10,11]

Additions of other oxidants, such as nitrate and iron-containing materials, may influence CH_4 emission from rice fields. As well as adding oxidants, dressing paddy fields with other soils that contain a large amount of free iron and manganese may decrease CH_4 emission. Other chemical candidates are nitrification inhibitors and acetylene releasing materials.

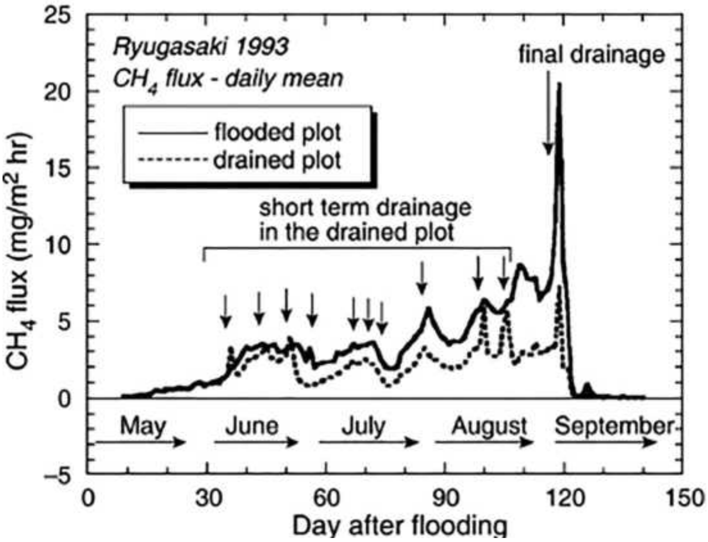


FIGURE 2 Effect of water management on CH_4 emission from a rice paddy field. The arrows indicate period of midseason drainage in the intermittent irrigation plot and the timing of final drainage in both of the plots.

Organic Matter Management

In rice cultivation, fresh organic matter and animal wastes are often applied as fertilizers. In the fields, a proportion of the biomass of previous crops and weeds remains in soils at the start of rice cultivation. Such organic matter is decomposed in soils and acts as a substrate for fermentation reactions. Many researchers have demonstrated that incorporation of rice straw and green manure into rice paddy soils dramatically increases CH_4 emission.[7,10,12] The impact of organic amendments on CH_4 emissions can be described by a dose–response curve which adopts correction factors for composted and fermented organic matter.[6] Mitigation of CH_4 emission requires that the quantities of organic amendments be minimized.

Field experiments also indicated that composted or fermented organic matter increased CH_4 emission much less than fresh organic matter, due to a lower content of easily decomposable carbon.[6,7] Therefore, stimulation of composting organic amendments appears to be a promising mitigation option. Plowing the fields during the fallow period and promoting aerobic degradation of organic matter is also likely to reduce CH_4 emission.

Others

Different tillage and cropping practices change the physical, chemical, and microbiological properties of the plow layer soil and may reduce CH_4 emission. These include deep tillage, no tillage, and flooded rice-upland crop rotation.

Selecting and breeding rice cultivars that emit lower CH_4 is a desirable approach because it is easy to adopt. There are four points to consider for selecting cultivars: 1) they should exude low levels carbon from their roots; 2) they should have a low level of CH_4 transport and a high level of CH_4 oxidation in the rhizosphere; 3) they should have a higher harvest index, in order to reduce organic matter input into soil after harvest; and 4) they should be suitable and have a high productivity when other mitigation options are performed.

Problems and Feasibility of the Options

If the above mitigation options could be applied to world's rice cultivation, global CH_4 emission from rice fields could decrease significantly. However, there are several formidable obstacles to adopting the mitigation options into local rice farming. Table 2 summarizes the problems and feasibility of the individual mitigation options along with the efficiency of the options.

Application of some options is limited to specific types of rice fields. In particular, altering water management practices may be limited to rice paddy fields where the irrigation system is well equipped. Long midseason drainage and short flooding may cause possible negative effects on grain yield and soil fertility. Improving percolation by underground pipe drainage requires laborious engineering work. The increased water requirement is another problem in the water management options because water is a scarce commodity in many regions.

Cost and labor are serious obstacles for applying each option to local farmers. Most of the mitigation options will decrease profitability and the farmer net returns in the short run. To overcome these obstacles, an effort to maximize net returns by joining CH_4 mitigation and increased rice production will be needed, as well as political support.

It is recognized that the mitigation options should not have any significant trade-off effects, such as decreased rice yield, a decline in soil fertility, or increased environmental impact by nitrogen compounds. The development of anaerobic conditions in soil by flooding decreases decomposition rates of soil organic matter compared with aerobic soils, resulting in soil fertility being sustained for a long time. Flooded rice cultivation shows very little growth retardation by continuous cropping. Some mitigation options may reduce these advantages of rice fields. Application of sulfate-containing fertilizer may cause a reduction in rice yield due to the toxicity of hydrogen sulfide. Mid-season aeration and soil amendments may induce nitrogen transformation resulting in enhanced N_2O emissions.[16,17]

TABLE 2 Evaluation of the Mitigation Options for Methane Emission from Fields

		Problem for Application							
	CH_4 Mitigation Efficiency	Applicability		Economy		Effects On		Time Span	Other Trade-off Effects
		Irrigated	Rain-fed	Cost	Labor	Yield	Fertility		
Water management									
Midseason drainage	☐	o	•	~	↑	+	~	o	May promote N_2O emission
Short flooding	☐	o	•	~	~	–	–	o	May promote N_2O emission
High percolation	☐	o	•	↑	↑	+	~	o	May promote nitrate leaching
Soil amendments									
Sulfate fertilizer	☐	o	o	↑	~	Δ	–	o	May cause H_2S injury
Oxidants	☐	o	o	↑	↑	Δ	–	o	
Soil dressing	o	o	o	↑	↑	–	–	o	
Organic matter									
Composting	☐	o	o	↑	↑	+	+	o	
Aerobic	☐	o	o	~	↑	~	~	o	
decomposition	☐								
Burning	o	o	o	~	↑	~	~	o	Causing atmospheric pollution
Others									
Deep tillage	o	o	o	↑	↑	–	–	o	
No tillage	?	o	o	~	↓	–	~	o	
Rotation	o	o	Δ	~	↑	–	–	o	
Cultivar	o	o	o	~	~	~	~	•	

Source: Ranganathan et al.,[13] Neue et al.,[14] and Yagi et al.[15]

Key:
☐ Very effective
o Effective/applicable Δ Case by case
• Not applicable/require long time
? No information
↑ Increase
↓ Decrease
~ About equal to previous situation
+ Positive
– Negative

References

1. Prather, M.; Derwent, R.; Ehhalt, D.; Fraser, P.; Sanhueza, E.; Zhou, X. Other trace gases and atmospheric chemistry. In *Climate Change 1994, Radiative Forcing of Climate Change and an Evaluation of the IPCCIS92 Emission Scenarios*; Houghton, J.T., Meira Filho, L.G., Bruce, J., Lee, H., Callander, B.A., Haites, E., Harris, N., Maskell, K., Eds.; Cambridge University Press: Cambridge, England, 1995; 73–126.
2. International panel on climate change. *Greenhouse Gas Inventory Reference Manual*; IPCC Guidelines for National Greenhouse Gas Inventories, OECD: Paris, France, 1997; Vol. 3, 46–60.
3. Conrad, R. Control of methane production in terrestrial ecosystems. In *Exchange of Trace Gases Between Terrestrial Ecosystems and the Atmosphere*; Andreae, M.O., Schimel, D.S., Eds.; John Wiley and Sons Ltd.: New York, 1989; 39–58.

4. Knowles, R. Processes of production and consumption. In *Agricultural Ecosystem Effects on Trace Gases and Global Climate Change*; Harper, L.A., Mosier, A.R., Duxbury, J.M., Rolston, D.E., Eds.; American Society of Agronomy: Madison, WI, 1993; 145–156.
5. Cicerone, R.J.; Shetter, J.D. Sources of atmospheric methane: measurements in rice paddies and a discussion. J. Geo-phys. Res. **1981**, *86,* 7203–7209.
6. Denier van der Gon, H.A.C.; Neue, H.-U. Influence of organic matter incorporation on the methane emission from a wetland rice field. Global Biogeochem. Cycles **1995**, *9,* 11–22.
7. Yagi, K.; Minami, K. Effect of organic matter application on methane emission from some Japanese paddy fields. Soil Sci. Plant Nutr. **1990**, *36,* 599–610.
8. Sass, R.L.; Fisher, F.M.; Wang, Y.B.; Turner, F.T.; Jund, M.F. Methane emission from rice fields: the effect of flood-water management. Global Biogeochem. Cycles **1992**, *6*, 249–262.
9. Yagi, K.; Tsuruta, H.; Kanda, K.; Minami, K. Effect of water management on methane emission from a Japanese rice paddy field: automated methane monitoring. Global Biogeochem. Cycles **1996**, *10,* 255–267.
10. Schütz, H.; Holzapfel-Pschorn, A.; Conrad, R.; Rennenberg, H.; Seiler, W. A 3 years continuous record on the influence of daytime, season, and fertilizer treatment on methane emission rates from an Italian rice paddy. J. Geophys. Res. **1989**, *94,* 16405–16416.
11. Denier van der Gon, H.A.C.; Neue, H.-U. Impact of gypsum application on methane emission from a Wetland rice field. Global Biogeochem. Cycles **1994**, *8*, 127–134.
12. Sass, R.L.; Fisher, F.M.; Harcombe, P.A.; Turner, F.T. Mitigation of methane emission from rice fields: possible adverse effects of incorporated rice straw. Global Biogeochem. Cycles **1991**, *5*, 275–287.
13. Ranganathan, R.; Neue, H.-U.; Pingali, P.L. Global climate change: role of rice in methane emission and prospects for mitigation. In *Climate Change and Rice;* Peng, S., Ingram, K.T., Neue, H.-U., Ziska, L.H., Eds.; Springer-Verlag: Berlin, Germany, 1995; 122–135.
14. Neue, H.-U.; Wassmann, R.; Lantin, R.S. Mitigation options for methane emissions from rice fields. In *Climate Change and Rice;* Peng, S., Ingram, K.T., Neue, H.-U., Ziska, L.H., Eds.; Springer-Verlag: Berlin, Germany, 1995; 137–144.
15. Yagi, K.; Tsuruta, H.; Minami, K. Possible options for mitigating methane emission from rice cultivation. Nutr. Cycling Agro-Ecosys. **1997**, *49,* 213–220.
16. Cai, Z.; Xing, G.; Yan, X.; Xu, H.; Tsuruta, H.; Yagi, K.; Minami, K. Methane and nitrous oxide emissions from rice paddy fields as affected by nitrogen fertilizers and water management. Plant Soil **1997**, *196,* 7–14.
17. Bronson, K.F.; Neue, H.-U.; Singh, U.; Abao, E.B., Jr. Automated chamber measurements of methane and nitrous oxide flux in a flooded rice soil: I. Residue, nitrogen, and water management. Soil Sci. Soc. Am. J. **1997**, *61,* 981–987.

3

Petroleum: Hydrocarbon Contamination

Svetlana Drozdova
and Erwin
Rosenberg

Introduction

Historically, environmental analyses focused on monitoring compounds that pose a threat to humans and their environment. Petroleum hydrocarbon compounds are among them. Contamination of water, soil, and sediment samples by petroleum hydrocarbons is a common and severe environmental problem caused by improper handling, storage, transport, or use of petrochemical products or raw materials.

Petroleum products are the major source of energy for industry and daily life. Leaks and accidental spills occur regularly during exploration, production, refining, transport, and storage. In addition, natural processes can result in seepage of crude oil from geologic formations below the seafloor. The total input of crude oil and petroleum into the environment is estimated to be 1.3 million tons per year. To understand the potential effect of petroleum contaminations on the environment, it is important to understand the nature and distribution of sources and their inputs. Petroleum poses a range of environmental risks when released into the environment. Catastrophic and large-scale spills have a very severe physical impact in addition to the chemical pollution that they cause; chronic discharges and small releases can damage and eventually kill the exposed flora and fauna due to toxicity of many of the individual compounds contained in petroleum. Oil contamination in the environment is primarily assessed by measuring the chemical concentrations of petroleum products in the affected environmental compartment (e.g., sediment, biota, water).

This entry provides a discussion of the environmental relevance of petroleum hydrocarbons; the principal sources of petroleum contaminations in the environment; and the nature and composition of crude oil and petroleum products derived from it. The fate of petroleum hydrocarbons in the environment, possible effects from exposure to them, and their toxicity are discussed as well. The entry is concluded by an overview of analytical methods for determination of petroleum hydrocarbon contamination.

Petroleum Hydrocarbons and Their Environmental Relevance

Oil and gas resources are organic compounds, formed by the effects of heat and pressure on sediments trapped beneath the earth's surface over millions of years. The remains of animals and plants that lived millions of years ago in a marine environment were covered by layers of sand and silt over the years. Heat and pressure from these layers helped the remains turn into crude oil or petroleum.[1] The word "petroleum" means "rock oil" (from Greek: petra [rock] + Latin: oleum [oil])[2] or "oil from the earth." While ancient societies made some use of these resources, the modern petroleum age began less than a century and a half ago, when in 1859, Colonel Drake discovered oil in Oil Creek in Titusville, Philadelphia, United States. From that time on, the world's demand for fossil fuel and the production of oil have continuously increased. From the 1980s, in particular, after the second oil crisis of 1979, the petroleum business has developed into a high-technology industry. Advances in technology have greatly improved the ability to find and extract oil and gas and to convert them to efficient fuels and useful consumer products. About 100 countries produce crude oil. Russia, Saudi Arabia, the United States, Iran, and China are the top five producing countries in 2009 (Table 1).[3] In the United States, the oil and gas industry employs 1.4 million people and generates about 4% of U.S. economic activity. It is larger than the domestic automobile industry and larger than education and social services, the computer industry, and the steel industry combined.[4] At a refinery, different fractions of the crude oil are separated into useable petroleum products. Various sources of information provide a good overview of the different processes in petroleum refining.[5–7] Petroleum products are used worldwide for energy production, as fuel for transport, and as a raw material for many chemical processes. The United States is the biggest consumer of oil in the world (Table 1). Although there exist well-developed alternatives to the use of oil (particularly for energy production and transportation), our societies are still strongly dependent on oil, which is an environmental burden, an economic problem, and a political hazard. However, at the current time, the economic situation still favors the use of petroleum and petroleum products for these applications rather than its alternatives, which at the moment are not competitive from an economic point of view.

Petroleum poses a range of environmental risks when it is released into the environment (whether by catastrophic spills or through chronic discharges). In addition to the physical impact of large spills, the toxicity of many of the individual compounds contained in crude oils or petroleum products is significant. Information on how petroleum hydrocarbons enter and diffuse in the environment is abundant.[8,9] The sources of petroleum input to the environment, particularly to the sea, are diverse. They can be categorized effectively into four major groups, namely, natural seeps, petroleum extraction, petroleum transportation, and petroleum consumption.

Natural seeps are frequently encountered phenomena that occur when crude oil seeps from the geologic strata beneath the seafloor to the overlying water column as a natural process.[10] Recognized by geologists for decades as indicating the existence of potentially exploitable reserves of petroleum, these seeps release vast amounts of crude oil annually. Yet these large volumes are released at a rate low enough that the surrounding ecosystem can adapt and even thrive in their presence; which is not true in case of the catastrophic and accidental impact of a tanker or oil well spill. Natural processes are, therefore, responsible for over 45% of the petroleum entering the marine environment worldwide (Table 2).[11]

As result of human activities, about 700,000 tons of petroleum is released annually into the sea worldwide. Processes such as petroleum extraction, transportation, and consumption can cause soil and groundwater contamination in case of equipment failure or operation errors and other reasons. Petroleum extraction can result in release of both crude oil and refined products as a result of human activities associated with efforts to explore and produce petroleum. The nature and size of these releases are highly variable—see Table 3 for the largest oil spills observed until 2010[12]—and can include accidental spills of crude oil from platforms and blowouts such as that of the oil rig Deepwater Horizon in the Gulf of Mexico in April 2010 or slow chronic releases of water produced from oil- or gas- bearing formations during extraction. Under current industry practices, this "produced water" is treated to separate from crude oil and either injected back into the reservoir or discharged overboard. Produced water is

TABLE 1 Annual Production and Consumption of Oil by the Top 10 Industrial Nations and by the Top 10 Countries in the European Union

Oil production by country					Oil consumption by country in the world				
Rank	Country	Amount bbl/day	Date	Percentage %	Rank	Countries	Amount bbl/d	Date	Percentage %
1	Russia	10,120,000	2010	11.9	1	United States	18,690,000	2009	22.6
2	Saudi Arabia	9,764,000	2009	11.5	2	China	8,200,000	2009	9.9
3	United States	9,056,000	2009	10.7	3	Japan	4,363,000	2009	5.3
4	Iran	4,172,000	2009	4.9	4	India	2,980,000	2009	3.6
5	China	3,991,000	2009	4.7	5	Russia	2,740,000	2010	3.3
6	Canada	3,289,000	2009	3.9	6	Brazil	2,460,000	2009	3.0
7	Mexico	3,001,000	2009	3.5	7	Germany	2,437,000	2009	2.9
8	United Arab Emirates	2,798,000	2009	3.3	8	Saudi Arabia	2,430,000	2009	2.9
9	Brazil	2,572,000	2009	3.0	9	Korea, South	2,185,000	2010	2.6
10	Kuwait	2,494,000	2009	2.9	10	Canada	2,151,000	2009	2.6
Total		84,764,555			**Total**		82,769,370		
Oil production by EU member states					Oil consumption by EU member states				
Rank	Countries	Amount bbl/d	Date	Percentage %	Rank	Countries	Amount bbl/d	Date	Percentage %
1	United Kingdom	1,502,000	2009	60.4	1	Germany	2,437,000	2009	16.2
2	Denmark	262,100	2009	10.5	2	France	1,875,000	2009	12.5
3	Germany	156,800	2009	6.3	3	United Kingdom	1,669,000	2009	11.1
4	Italy	146,500	2009	5.9	4	Italy	1,537,000	2009	10.2
5	Romania	117,000	2009	4.7	5	Spain	1,482,000	2009	9.9
6	France	70,820	2009	2.8	6	Hungary	1,373,000	2009	9.1
7	Netherlands	57,190	2009	2.3	7	Netherlands	922,800	2009	6.1
8	Poland	34,140	2009	1.4	8	Belgium	608,200	2009	4.1
9	Spain	27,230	2009	1.1	9	Poland	545,400	2009	3.6
10	Austria	21,880	2009	0.9	10	Greece	414,400	2009	2.8
Total (EU, 27 countries):		2,485,550	2009		**Total (EU, 27 countries):**		15,012,050		
	Norway	2,350,000	2009			Norway	204,100		
	Turkey	52,980	2009			Turkey	579,500		

Source: Adapted from Energy Statistics: Oil-Production (Most Recent) by Country.[3]
bbl, barrel; EU, European Union. 1 bbl ≈ ca. 159 L.

TABLE 2 Petroleum Input to the Sea

	North America		Worldwide	
Source of Input	Tons	%	Tons	%
Natural seeps	160,000	61	600,000	46
Petroleum extraction	3,000	1	38,000	3
Petroleum transportation	9,100	4	150,000	12
Petroleum consumption	84,000	32	480,000	37
Other	3,900	2	32,000	2
Total input:	260,000 tons		1,300,000 tons	

Source: Adapted from *Oil in the Sea III Inputs, Fates, and Effects.*[11]

TABLE 3 Top 10 Oil Spills in the World as of 2010

	Incident	Location	Year	Type of Incident	Magnitude of Oil Spill (gallons)
1	Gulf War	Kuwait	1991	Oil spill due to war action and sabotage of oil drilling stations and pipelines, encompassing also the dumping of the charge of several oil tankers into the Persian Gulf by Iraqi troops during the Gulf War.	520,000,000
2	Deepwater Horizon	Gulf of Mexico	2010	Oil spill as a consequence of a methane blowout (which could not be prevented due to a technical problem) at the oil rig Deepwater Horizon, which caused an explosion and and the subsequent loss of the oil drilling platform. The well continued to leak for over 100 days.	172,000,000
3	Ixtoc I	Mexico	1979	After an unexpected blowout at the offshore oil rig Ixtoc 1 in the Gulf of Mexico, the platform exploded and collapsed. Oil escaped freely from the well for almost 1 year until the well could be capped.	138,000,000
4	Atlantic Empress/ Aegean Captain	Trinidad and Tobago	1979	Collision of two ships, the Aegean Captain and the supertanker Atlantic Empress, during a heavy storm in the Caribbean Sea. The Atlantic Empress exploded, sank, and lost its freight.	90,000,000
5	Fergana Valley/ Mingbulak	Russia	1992	Technical failure of an oil well in the Fergana Valley located between Kyrgyzstan and Uzbekistan from which oil blew out for a period of 8 months.	88,000,000
6	Nowruz Oil Field	Persian Gulf	1983	Collision of an oil tanker with an oil platform at the Nowruz Oil Field during the Iran–Iraq War. After the oil drilling platform collapsed, the wellhead was destroyed and leaked oil into the Persian Gulf for more than 6 months before being capped. A similar event at the same oilfield resulted directly form war action.	80,000,000
7	Castillo de Bellver	South Africa	1983	A fire at the tanker Castillo de Bellver caused the ship to drift and then break into two separate pieces. Relatively little damage was done to the South African coastline since the oil may have sunk into the sea or burned during the fire.	79,000,000
8	The Amoco Cadiz	France	1978	The crude oil carrier Amoco Cadiz ran aground off the French Atlantic coast and finally spilt into halves, whereby it lost its complete freight, which contaminated 200 km of the French coastline.	69,000,000
9	ABT Summer	Angola	1991	Following a fire aboard the oil tanker *ABT Summer*, it sank and all its freight either leaked to the sea or sank to the ground about 900 miles from the coast of Angola.	51,000,000
10	The MT Haven	Genova, Italy	1991	After unloading the oil tanker MT Haven, a fire broke out, followed by explosions after which the ship sank and continued to leak oil for 12 years.	45,000,000

Source: Adapted from Top 10 Worst Oil Spills.[12]

the largest single wastewater stream in oil and gas production. The amount of produced water from a reservoir varies widely and increases over time as the reservoir is depleted. Petroleum transportation can result also in releases of dramatically varying sizes of petroleum products (not just crude oil) from major incidents (mostly from tankers, such as the one in 1979 off the coast of Tobago, when two tankers collided and one of these, the Atlantic Empress, sank, losing all its freight) to relatively small operational releases that occur regularly, such as those from pipelines.

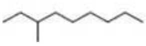
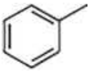
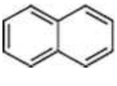
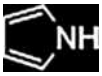
Releases that occur during the consumption of petroleum, whether by individual car and boat owners, non-tank vessels, or runoff from urban or industrial areas, are typically small but frequent and widespread and are responsible for the vast majority (70%) of petroleum introduced to the environment through human activity.

Because crude oil and petroleum products are a complex and highly variable mixture of hundreds to thousands of individual hydrocarbon compounds, characterizing the risks posed by petroleum-contaminated soil and water has proven to be difficult and inexact. It is very important to have an understanding of the toxicology, analytical science, environmental fate and behavior, risk, and technological implications of petroleum hydrocarbons in order to interpret, evaluate the risk of, and make decisions about the hazardous effect to and ensure the appropriate protection of the environment.

General Chemical Composition Features of Crude Oils and Petroleum Products

Crude oil is an extremely complex mixture of several thousands of different compounds; its compositions and physical properties vary widely depending on the source from which the oils are produced, the geologic environment, and location in which they migrated and from which they are extracted. The nature of the refining processes has an effect on crude oil compositions as well. As indicated in Table 4, petroleum and petroleum products contain primarily hydrocarbons, heteroatom compounds, and relatively small concentrations of (organo)metallic constituents.[13,14] The complexity of petroleum and petroleum products increases with carbon number of its constituents, so it is impossible to identify all components. Petroleum and petroleum products are typically characterized in terms of boiling range and approximate carbon number. Raw petroleum is usually dark brown or almost black, although some fields deliver a greenish or sometimes yellow petroleum. Depending upon the oil field and the way the petroleum composition was formed, the crude oil will also differ in viscosity. The composition of crude oil impacts certain physical properties of the oil, and it is these physical properties (e.g., density or viscosity) by which crude oils are generally characterized, classified, and traded. These physical properties can be used to classify crude oils as light, medium, or heavy. The American Petroleum Institute (API) gravity[15] is a measure of the specific gravity of a petroleum liquid compared with water (API = 10).

TABLE 4 Main Constituents of Petroleum Hydrocarbons and Representative Examples

Petroleum Hydrocarbon Compounds						
Aliphatics/Alicyclics				Aromatics		
Saturated hydrocarbons		Unsaturated hydrocarbons				
Alkanes(paraffins)	Cycloalkanes	Alkenes (olefins)	Alkynes (acetylenes)	Benzene and alkylbenzenes (BTEX)	Polynuclear aromatics (PAH)	Heterocyclic compounds
Single carbon bonds, straight and branched structure	Straight and cyclic structure	One or more double carbon bonds, straight, branched, or cyclic	One or more triple carbon bonds, straight, branched, or cyclic	Single aromatic ring or with attached functional group	Two or more aromatic rings fused together, can be with attached functional group	Aromatic ring structures with one or more heteroatom (N, S, O) in the ring
C_nH_{2n+2}	C_nH_{2n}	C_nH_{2n}	C_nH_{2n-2}			
 				 		
n-Decane 3-Methylnonane	Cyclohexane	1-Octene	1-Hexyne	Benzol Toluene	Naphthalene	Pyrrole

Light oils are defined as having an API < 22.3, heavy oils are those with API > 31.1, and medium oils have an API gravity between 22.3 and 31.1.

Regardless of the complexity, petroleum compounds can be separated into two major categories: hydrocarbons and non-hydrocarbons. Hydrocarbons (compounds composed solely of carbon and hydrogen) comprise the majority of the components in most petroleum products and are the compounds that are primarily (but not always) measured as total petroleum hydrocarbons (TPH).[16] The nonhydrocarbon components are heterocyclic hydrocarbons (compounds containing heteroatoms such as sulfur, nitrogen, or oxygen in addition to carbon and hydrogen). These heterocyclic hydrocarbons are typically present in oils at relatively low concentrations and can be found in most refined motor fuels as they are concentrated in the heavier fractions and residues during refining. Most organic nitrogen hydrocarbons in crude oils are present as alkylated aromatic heterocycles, mostly with a pyrrolic structure. Crude oils also contain small amounts of organometallic compounds (of nickel, vanadium, and other metals up to atomic number 42, with the exception of rubidium and niobium) and inorganic salts. Although, depending on the analytical method, sulfur-, oxygen-, and nitrogen-containing compounds are sometimes included in the value reported as TPH concentration, they do not fall under the definition of petroleum hydrocarbons in the strict sense.[16]

Depending on the structure of petroleum hydrocarbons, the individual compounds are grouped into aliphatic(saturated and unsaturated) hydrocarbons and aromatics. Saturated hydrocarbons are the major class of compounds found in crude oil. The common names of these types of compounds are alkanes and isoalkanes or, as used in petroleum industry, paraffins and isoparaffins, respectively. Unsaturated hydrocarbons have at least one multiple bond (double bond [alkenes] or triple bond [alkynes]), and they are typically not present in crude oil but can be formed during the cracking process. Aromatic hydrocarbons are based on the benzene ring structure and are further categorized depending on the number of rings. Benzene rings are very stable and therefore persistent in the environment, and particularly, the mono- and polycyclic aromatic compounds can have toxic effects on organisms. Aromatic hydrocarbons with one benzene ring and with one or more side chains are alkyl benzenes and include benzene; toluene; ethylbenzene; and o-, p-, and m-xylenes (BTEX). This class of compounds has significant water solubility and is more mobile in the environment. Polycyclic aromatic hydrocarbons (PAHs) are aromatic compounds with two or more fused aromatic rings. Occurrence of PAH compounds in oils is dominated almost completely by the C1- to C4-alkylated homologues of the parent PAH, in particular, for naphthalene, phenanthrene, dibenzothiophene (a sulfur-containing aromatic heterocycle), fluorine, and chrysene. These alkylated PAH homologues form the basis of chemical characterization and identification of oil spills.[17,18] A typical crude oil may contain 0.2% to more than 7% total PAHs. Of the hydrocarbon compounds common in petroleum, PAHs appear to pose the greatest toxicity to the environment.

Different crude oil sources usually have a unique hydrocarbon composition.[19,20] The actual overall properties of each different petroleum source are defined by the percentage of the main hydrocarbons found within petroleum as part of the petroleum composition. The percentages for these hydrocarbons can vary greatly. It gives the crude oil a quite-specific compound personality depending on geographic region. The typical percentage of hydrocarbons (although covering very wide ranges) is as follows: paraffins (15%–60%), naphthenes (30%–60%), aromatics (3%–30%), and asphaltenes making up the remainder. Furthermore, due to differences in refining technologies and refinery operating conditions, each refining process has a distinct impact on the hydrocarbon composition of the product.

Refined petroleum products are primarily produced through distillation processes that separate fractions from crude oil according to their boiling ranges. Production processes may also be directed to increase the yield of low-molecular-weight fractions, reduce the concentration of undesirable sulfur and nitrogen components, and incorporate performance-enhancing additives. Therefore, each petroleum product has its unique, product-specific hydrocarbon pattern. The petroleum products are composed of both aliphatic and aromatic hydrocarbons in a range of molecules that include C6 and greater. The different classes of compounds contained in various petroleum products are summarized in Table 5.[20,21] The main products are gasoline (benzene), naphtha/solvents, jet fuels, kerosene, diesel fuel, and lubricating (motor) oils. Due to the variety of components in petroleum, they are typically characterized using the boiling range of the mixture and the carbon number rather than individual components. For example, diesel is a fraction with boiling points between 200°C and 325°C and is represented as C10–C22.

While a physical property such as boiling range may establish the initial product specification, other finer specifications define their ultimate use in certain applications. A lighter, less dense, raw petroleum composition with a composition that contains higher percentages of hydrocarbons is much more profitable as a fuel source. On the other hand, other denser petroleum compositions with a less flammable level of hydrocarbons and containing higher levels of sulfur are expensive to refine into a fuel and are therefore more suitable for plastics manufacturing and other uses. In contrast to the ever-increasing demand, the world's reserves of light petroleum (light crude oil) are severely depleted, and refineries are forced to refine and process more and more heavy crude oil and bitumen.

Petroleum fractions are among the most complex samples an analyst can face in terms of the number of compounds present. The characterization of petroleum fractions is typically done by gas chromatography (GC). As can be seen in Figure 1, the petroleum products contain such a large number of hydrocarbon constituents that complete chromatographic separation is not possible. Even then, GC remains the most informative analytical technique, providing both quantitative information (deduced from the total signal recorded in a chromatogram) and qualitative information, which derives from the fact that the retention times in the chromatograms can be correlated with the boiling points of the compounds contained in the petroleum. To illustrate the complexity of chemical composition of petroleum products, Figure 2 shows the chromatograms for six different petroleum products, including a crude oil with API of 18.7 and the BAM (Bundesanstalt für Materialprüfung, Berlin) petroleum hydrocarbon standard. The BAM standard K-010 is a certified reference material for the determination of mineral hydrocarbons, which is a synthetic mixture of a diesel and a lubricating oil. It is evident that these six samples are very different according to their carbon ranges. The difference is clearly seen from the comparison of their chromatograms. The volatile fuel with a content of hydrocarbons with less than 10 C atoms (benzene and premium gasoline) has the majority of its constituents at the beginning of chromatogram (Figure 2c). The peaks in the chromatogram of diesel are shifted to the retention time window where hydrocarbons from C10 to C22 are eluted (Figure 2d). In turn, the chromatogram of motor oil shows a characteristic "bump" (because the fraction of saturated alkanes is very small) situated in the

TABLE 5 Overview of Petroleum Products with Respect to Boiling Point Ranges, Approximate Carbon Number, and Average Percentage Amount of Aliphatic and Aromatic Compounds

		Fractions				Hydrocarbons	
	Boiling Range	<C7 (% w/w)	C7–C10 (% w/w)	C10–C40 (% w/w)	>C40 (% w/w)	Aliphatic	Aromatic
Statfjord C (39.1)[a]		11.6	18.1	56.6	13.7		
Crude oil (API = 18.7)		0.9	3.0	63.2	32.8		
Grane[a]							
Normal benzene[b]	40–200°C	~100 (C5–C12)				~70%	20%–50%
Jet fuel[b]	150–300°C		~100 (C6–C14, C16)			80%–90%	10–20%
Kerosene[b]	150–300°C		~100 (C6, C9–C16)			60%–80%	5%–20%
Diesel[b]	200–325°C		~100 (C10–C22)			60%–90%	30%–40%
Light heating oil[b]	200–325°C		~100 (C10–C22)				
Lubricant or motor oil[b]	325–600°C			~100 (C20–C40)		70%–90%	10%–30%
Heavy heating oil[b]	325–600°C			~100 (C20–C50)			

[a] **Source:** Data from Crude Oil Assays.[20]

[b] **Source:** Data from Statoil Web site.[21]

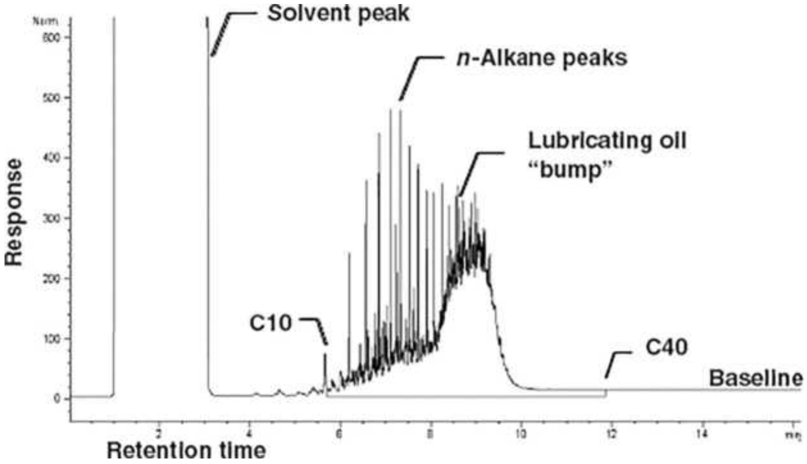


FIGURE 1 Chromatogram of a mixture of petroleum products (diesel and lubricating oil, 1:1), obtained by GC with FID.

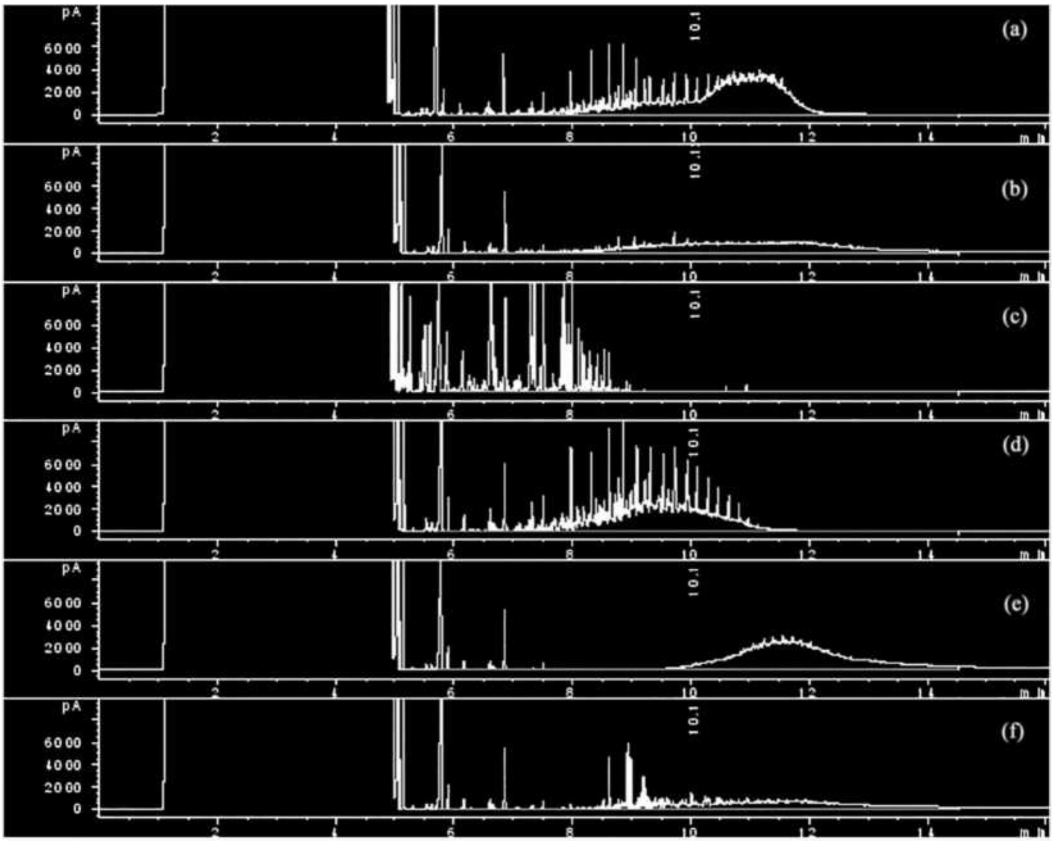


FIGURE 2 Comparison of chromatograms of different oil samples: (a) BAM; (b) crude oil (API = 18.7); (c) gasoline; (d) diesel; (e) motor oil; and (f) heavy heating oil with the same concentration (20 ppm oil in water) obtained by GC-FID method

region where heavier hydrocarbons C20–C40 are eluted (Figure 2e). Thus, GC- based methods provide important qualitative information, which in the ideal case even allows the assignment of the source of contamination. This is proven by the comparison of chromatograms of different oil samples (Figure 2).

Fate of Petroleum Hydrocarbons in the Environment

The effects of petroleum hydrocarbons entering the environment are a complex function of the magnitude and the rate of release; the nature of the released petroleum (its physicochemical properties and, in particular, the amount of toxic compounds it may contain); and the affected geographical, hydrogeological, and biological ecosystem. The fate of petroleum-type pollutants in the environment has been investigated in many studies.[22] Complex transformation and degradation processes of oil in the environment start from its first contact with the atmosphere, seawater, and soil. They depend on the physical properties (volatility, solubility, etc.), as well as on the chemical properties (chemical composition) of the oil. While the former are responsible for transport, or diffusion of the petroleum hydrocarbons in the environment, the latter are responsible for their chemical, photo-, and microbial degradation. The main processes affecting the environmental fate of petroleum hydrocarbons after their release to the environment are thus their volatilization, dissolution/dispersion and emulsification in water, adsorption to soil, oxidation, destruction, and biodegradation.[23,24] In addition to the parameters that characterize the oil's composition, reactivity, and toxicity, the environmental conditions, i.e., the meteorological and hydrological factors, also play an important role in the fate of petroleum hydrocarbons.

When petroleum hydrocarbons are released to the water column, certain fractions will float on top and form thin surface films. This process is controlled by the viscosity of the oil and the surface tension of water. A spill of 1 ton of oil can disperse over a radius of 50 m in 10 min, forming a slick 10 mm thick. Later, it spreads, gets thinner, and covers an area of up to 12 km^2.[11] It should be pointed out that much of the environmental and ecological damage caused by oil spills actually is due to this oil film that covers the surface of the sea, or the coastline, thus physically impairing birds and other animals and causing suffocation of fish as oxygen will not permeate the oil layers to a sufficient degree anymore. In the first days after the spill, the volatile compounds from oil evaporate. Only a small proportion of the hydrocarbon constituents of petroleum products are significantly soluble in water. Dissolution takes more time compared with evaporation, considering that most oil components are soluble in water only to a limited degree (although the degradation products typically are more polar and thus more soluble). Other heavier fractions (up to 10%–30%) will accumulate in the sediment at the bottom of the water, which may affect bottom-feeding fish and organisms. This happens mainly in the narrow coastal zone and shallow waters, where water is intensively mixing.

Crude oil released to the soil may percolate and reach the groundwater. Because petroleum has a lower specific gravity than water, free (undissolved) product and most dissolved contamination are usually concentrated near the top of the groundwater.[25] This may then lead to a fractionation of the original complex mixture, depending on the chemical properties of the compound. Some of these compounds will evaporate, while others will dissolve into the groundwater and be diffused from the release area. Other compounds will adsorb to soil or sediments and will remain there for a long period of time, while others will be metabolized by organisms found in the soil.[26,27]

While evaporation and dissolution redistribute the oil, photochemical oxidation and bacterial degradation transform it. Where crude oil is exposed to sunlight and oxygen in the environment, both photooxidation and aerobic microbial oxidation take place. The photochemical oxidation of hydrocarbons is dependent upon ultraviolet (UV) radiation and will therefore occur only in the upper surface layers. The aromatic hydrocarbons absorb UV radiation with high efficiency and are transformed mainly into hydrogen peroxides. Alkanes are much less efficient in absorbing UV radiation, and only small quantities are transformed by this process. The final products of oxidation (hydroperoxides, phenols, carboxylic acids, ketones, aldehydes, and others) usually have increased water solubility and toxicity. Where oxygen and sunlight are excluded in anoxic environments, anaerobic microbial oxidation takes place.[28,29]

Generally, saturated alkanes are more quickly degraded by microorganisms than aromatic compounds; alkanes and smaller-sized aromatics are degraded before branched alkanes, multiring and substituted aromatics, and cyclic compounds.[30,31] Polar petroleum compounds such as sulfur- and nitrogen-containing species are the most resistant to microbial degradation. Complex structures (e.g., branched methyl groups) and the stability of hydrocarbons decrease the rates of mineralization, which are likely a consequence of the greater stability of carbon–carbon bonds in aromatic rings than in straight-chain compounds. Emulsification also provides greater surface area for microorganisms to attach.

It has been shown in experiments that *n*-alkanes are among the most biodegradable hydrocarbons, and therefore, they are easily broken down and preferentially depleted from soil samples.[32] Also, it has been proven in simulation experiments of the biodegradation of two different samples of crude petroleum (paraffinic and naphthenic type) that microbial cultures that were isolated as dominant microorganisms from the surface of a wastewater canal of an oil refinery (most abundant species: *Phormidium foveolarum,* filamentous Cyanobacteria [blue-green algae] and *Achnanthes minutissima,* diatoms, algae) show a strongly differentiated degradation behavior with clear preference for the degradation of *n*-alkanes and isoprenoid aliphatic alkanes.[33] As can be seen in Figure 3, the largest degree of biodegradation was achieved in a medium containing the base nutrients $Ca(NO_3)_2.4H_2O$, $K_2HPO_4.7H_2O$, KCl, $FeCl_2$, and K_2SO_4, at pH ≈ 8 and exposed to light. Biodegradation activity is somewhat lower with the same medium in the dark. With a medium containing not only the nutrient broth but also organic compounds (tryptone, yeast extract, glucose, at pH ≈ 7), degradation occurs at a much lower rate, especially without light.

When crude oil or petroleum products are accidentally released to the environment, they are immediately subjected to a variety of weathering processes that lead to compositional changes and to the depletion of certain hydrocarbon compounds. Weathering processes include all previously mentioned physicochemical processes, such as dissolution, evaporation, photooxidation, polymerization, adsorptive interactions between hydrocarbons and the soil, and some biological factors. Furthermore, due to the fact that the degree of biodegradation is different for different types of petroleum hydrocarbons and varies depending on their nature, the weathering rate also depends on the type of petroleum contaminant. If we thus observe in the analysis of petroleum hydrocarbon contaminants changing patterns of hydrocarbons with time, this may be either due to the segregation of the oil according to the physical properties or due to the action of bacteria and microorganisms. As these are able to degrade only certain classes of compounds, or at least they exhibit a strong preference for some over other compounds, characteristic changes of the hydrocarbon pattern will result, as observed by GC (Figure 3).

Possible Toxic Effects from Exposure to Petroleum Hydrocarbons

As it was discussed earlier, crude oil and petroleum products are complex mixtures of groups of compounds. Many of the compounds are apparently benign, but many others are known to have toxic effects. Due to petroleum hydrocarbon toxicity, spilled hydrocarbons pose a threat that affects not only the sea and land but also the lakes, rivers, and groundwater and can be harmful for animals and human health.

Much of what is known about the impacts of petroleum hydrocarbons comes from studies of catastrophic oil spills and chronic seeps. Large oil spills usually receive considerable public attention because of the obvious environmental damage, oil-coated shorelines, and dead or moribund wildlife, including, in particular, oiled seabirds and marine animals. The acute toxicity of petroleum hydrocarbons to marine organisms is dependent on the persistence and bioavailability of specific hydrocarbons. The exposure to them may alter an organism's chances for survival and reproduction in the environment, and the narcotic effects of hydrocarbons on nerve transmission are a major biological factor in determining the ecologic impact of any release. Marine birds and mammals may be especially vulnerable to oil spills. In addition to acute effects such as high mortality, chronic, low-level exposure to hydrocarbons may affect reproductive performance and physiological impairment of seabirds and some marine mammals as well.[11] Petroleum contamination may also cause unfavorable impacts on nearby plants and animals. Plants growing in contaminated soils or water may die or appear distressed.

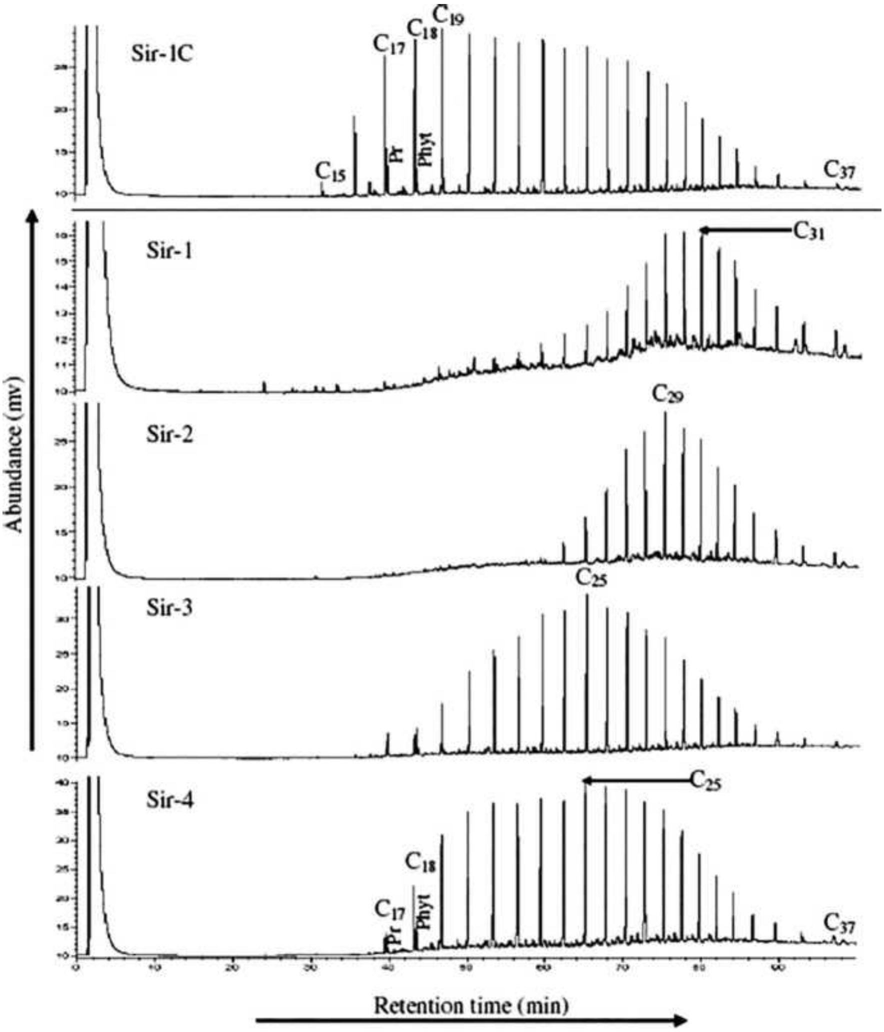


FIGURE 3 Gas chromatograms of the alkane fractions derived from crude oil Sirakovo (Sir, paraffinic type) after 90 days of simulated biodegradation with *Phormidium foveolarum* and *Achnanthes minutissima* with inorganic medium in the light (Sir-1), with inorganic medium in the dark (Sir-2), with organic medium in the light (Sir-3), and with organic medium in the dark (Sir-4), together with chromatogram of alkane fraction typical for the control experiments (Sir-1C), pristane (Pr), phytane (Phyt).
Source: Adapted from Antić et al.[33]

In turn, natural seeps, leaking pipelines, and production discharges release small amounts of oil over long periods of time, resulting in chronic exposure of organisms to oil and oil chemical compounds. The lower-molecular-weight compounds are usually the more water-soluble components of a product, and hence, attention has also been paid to the water-soluble fractions of petroleum and related products. Concentrations in the environment are usually comparatively low, and chronic effects are usually more significant.[26] The persistence of some compounds such as PAH in sediments, especially in urban areas, is also an example of chronic pollution and toxicity.

Nowadays, humans can be exposed to petroleum hydrocarbons through ingestion of contaminated drinking water and soil residues; inhalation of vapors and airborne soils; and contact of contaminants with skin (dermal exposure) from many sources, including gasoline fumes at the pump, spilled crankcase oil on pavement, chemicals used at home or work, or certain pesticides that contain petroleum hydrocarbon components as solvents. Most petroleum hydrocarbon constituents will enter the bloodstream rapidly when inhaled or ingested. Incorporated petroleum hydrocarbons are widely distributed by the blood throughout the body and quickly are metabolized into less harmful compounds. Others may be degraded into more harmful chemicals. Even other compounds are distributed by the blood to other parts of the body and do not readily break down but are accumulated instead in fat tissue. The resorption of petroleum compounds through dermal tissue is slower; that is why direct exposure of the skin to petroleum hydrocarbons is generally harmless when exposure is only occasional and of short duration.

Studies on animals have shown effects on the lungs, central nervous system, liver, kidney, developing fetus, and reproductive system from exposure to petroleum compounds, generally after breathing or swallowing the compounds. Health impacts of exposure to petroleum contamination may include lung irritation, headaches, dizziness, fatigue, diarrhea, cramps, and nervous system effects. Benzene and other chemicals found in petroleum products have been determined to be carcinogenic (cause cancer). More information regarding toxicity of petroleum chemicals is available, for example, from the Agency for Toxic Substances and Diseases Registry (ATSDR), an agency of the U.S. Department of Health and Human Services,[34] or from the European Chemicals Agency.[35]

Oil products are complex mixtures of hundreds of chemicals, with each compound having its own toxicity characteristics. There are many difficulties associated with assessing the health effects of such complex mixtures with regard to hazardous waste site remediation. This means that the traditional approach of evaluating individual components is largely inappropriate. Toxicity information is in the best case available for the pure product; however, once a petroleum product is released to the environment, it changes its composition as a result of weathering. These compositional changes may be reflected in changes in the toxicity of the product.

One approach for assessing the toxicity of oil products is to use toxicity information from studies conducted on the whole product. A second approach is to identify and quantify all components and then consider their toxicities. This approach produces data that theoretically could be compared with the known toxicity of each compound. The impracticality of this approach stems from its high analytical cost and the lack of toxicity data for many of the component chemicals found in hydrocarbon mixtures. A third approach is to consider a series of hydrocarbon fractions and determine appropriate tolerable concentrations and toxicity specific for those fractions. A number of groups have examined such an approach, but the most widely accepted and internationally used are the ones developed by the Total Petroleum Hydrocarbons Criteria Working Group (TPHCWG) and the Massachusetts Department of Environmental Protection (MA DEP) in the United States, although they have been subject to adjustments in many cases. For example, in the United Kingdom, the TPHCWG approach is modified and extended to consider heavier hydrocarbon fractions. It has been developed as part of the Environment Agency's[36] environment sciences program and published in documents related to petroleum hydro- carbons.[37]

The MA DEP introduced in 1994 the concept of petroleum hydrocarbon size-based fractions for use in evaluating the human health effects of exposure to complex mixtures of hydrocarbons[38,39] and provided oral toxicity values for each of the fractions. The toxicity value assigned for each fraction is used in dose–response evaluations. Cancer risks or hazard amounts are subsequently summed across the fractions to get the total values. The TPHCWG has developed and published a series of five monographs[16,40–42] detailing the data on petroleum hydrocarbons and, in addition, has developed tolerable intakes for a series of total hydrocarbon fractions. The TPHCWG independently identified largely similar groupings of hydrocarbon fractions with somewhat different toxicity values in 1997. Of the 250 individual compounds identified in petroleum by the TPHCWG, toxicity data were available for only 95. Of these 95, the TPHCWG concluded that there were sufficient data to develop toxicity criteria for only 25.

As there are differences in toxicity between different hydrocarbon compounds, it is impossible to accurately predict toxic effects of contamination for which only total hydrocarbon data are available. Health assessors often select surrogate or reference compounds (or combinations of compounds) to represent TPH so that toxicity and environmental fate can be evaluated. Correspondence dates relating the toxicologically derived hydrocarbon fractions and their toxicity values to the analytically defined reporting fractions (by MA DEP) are contained in Table 6 for ingestion and inhalation exposure. Inhaled or ingested volatile hydrocarbons have both general and specific effects. The toxicity values are represented as a reference dose (RfD), which is the U.S. Environmental Protection Agency's (EPA's) maximum acceptable oral dose of a toxic substance. Significant efforts have been undertaken by MA DEP to describe an approach for the evaluation of human health risks from ingestion exposure to complex petroleum hydrocarbon mixtures. The methods offered by MA DEP for determination of air-phase (APH), volatile (VPH), and extractable (EPH) petroleum hydrocarbons[43–45] are designed to complement and support the toxicological approach. The ranges of quantified hydrocarbons within each method and their reporting limits are shown in Table 7.

The components of petroleum can be generally divided into broad chemical classes: alkanes, cycloalkanes, alkenes, and aromatics. A review of Table 6 shows that a U.S. EPA RfD is available for only one alkane, *n*-hexane. In general terms, alkanes have relatively low acute toxicity, but alkanes having carbon numbers in the range of C5–C12 have narcotic properties, particularly following inhalation exposure to high concentrations, because of their relatively high volatility and low solubility in water. Repeated exposure to high

TABLE 6 Oral and Inhalation Toxicity Values by MA DEP for Petroleum Hydrocarbon Fractions and Individual Compounds Present in Petroleum Products

		Toxicity Value, RfD		
Carbon range	Compound	Inhalation mg/m^3	Oral mg/kg/day	Critical Effect
Aliphatic				
C5–C8		0.2	0.04	Neurotoxicity
	n-Hexane	0.2	0.06	
C9–C18		0.2	0.1	Neurotoxicity, hepatic, and hematological effects
C19–C32		NA	2	Liver granuloma
Aromatic				
C6–C8		Use individual RfCs for compounds in this range		
	Benzene	NA	0.03	
	Toluene	0.4	0.2	
BTEX	Ethylbenzene	1.0	0.1	
	Styrene	1.0	0.2	
	Xylene (o-, p-, m-)	NA	2	
C9-C18		0.05		Body weight reduction; hepatic, renal, and developmental effects
	Isopropylbenzene	0.4	0.1	
	Naphthalene	0.003	0.02	
	Acenaphthene	NA	0.06	
	Biphenyl	NA	0.05	
	Fluorene	NA	0.04	
	Anthracene	NA	0.3	
	Fluoranthene	NA	0.04	
	Pyrene	NA	0.03	
C9–C32			0.3	Neurotoxicity
C19–C32		NA		

Source: Adapted from The U.K. Approach for Evaluating Human Health Risks from Petroleum Hydrocarbons in Soil[37] and Interim Final Petroleum Report Development of Health-Based Alternative to the Total Petroleum Hydrocarbon TPH Parameter.[38]
NA, not applicable.

TABLE 7 The Ranges of Hydrocarbons Quantified within the Methods for Determination of APH, VPH, and EPH by MA DEP and Their Reporting Limits

	APH 28°C–218°C		VPH 36°C–220°C		EPH 150°C–265°C		
Aliphatic	C5–C8	C9–C12	1C5–C8	C9–C12 C9–C10	C9–C18	C19–C36	
Aromatic	C9–C10		***Reporting limits***		C11–C22		PAH
For the individual target analytes							
In air phase	2–5 g/m^3						
In soil			0.05–0.25 mg/kg		20 mg/kg		0.2–1 mg/kg
In water			1–5 µ g/L		100µ g/L		2–5 µg/L
For the collective							
hydrocarbon ranges							
In air phase	10–12 g/m^3						
In soil			5–10 mg/kg		20 mg/kg		
In water			100–150 µg/L		100 µg/L		

Source: Adapted from *Interim Final Petroleum Report Development of Health-Based Alternative to the Total Petroleum Hydrocarbon TPH Parameter.*[38]

concentrations, for example, of *n*-hexane (RfD, 0.06 mg/kg/day) may lead to irreversible effects on the nervous system. Hexane is considered to be the most toxic compound in the C5–C8 aliphatic fraction. No RfDs are available for other alkanes, nor for any cycloalkane or alkene. Alkenes exhibit little toxicity other than weak anesthetic properties. Alkanes and cycloalkanes are treated similarly and have similar toxic effects.

Aromatic compounds with less than nine carbon atoms (such as BTEX) are evaluated separately because the toxicity values for each are well supported and these compounds have a wide range of toxicity. However, most of the smaller aromatic compounds have low toxicity, with the exception of benzene, which is a known human carcinogen (RfD, 0.029 mg/kg/day). Most petroleum hydrocarbon mixtures contain very low concentrations of PAHs. The major concern regarding PAHs is the potential carcinogenicity of some of these. Benzo(*a*)pyrene and benz(*a*)anthracene are classified as probable human carcinogens. Benzo(*a*)pyrene is normally considered to be the most potent carcinogenic PAH, but the carcinogenic potency of most PAHs is not well characterized. In case of spills of petroleum products affecting water, PAHs are not usually a specific concern; however, this concern becomes more specific if these compounds are released into the soil due to a bioaccumulation of PAH in soil.

Different regulations and guidelines to protect public health have been developed. These public health statements tell as well about petroleum hydrocarbons and the effects of exposure. The U.S. EPA[46] identifies the most serious hazardous waste sites in the United States. The EPA lists certain wastes containing petroleum hydrocarbons as hazardous. It regulates certain petroleum fractions, products, and some individual petroleum compounds. General health and safety data are as well discussed by the Energy Institute,[7] which is the main professional organization for the energy industry within the United Kingdom that promotes the safe, environmentally responsible, and efficient supply and use of energy in all its forms and applications. The Occupational Safety and Health Administration and the Food and Drug Administration are other agencies that develop regulations for toxic substances in the United States. The information provided by all of them is regularly updated as more information becomes available. The Dutch National Institute for Public Health and the Environment (RIVM), has been involved in a number of studies on risk assessment for petroleum hydrocarbons which were commissioned by the Dutch government and the European Commission.[47] Also the U.K. Environment Agency, mentioned before, is the leading public body protecting and improving the environment in the United Kingdom, including protection from petroleum contaminations.

Total Petroleum Hydrocarbons and Analytical Methods for Determination of Petroleum Hydrocarbons in Environmental Media

Due to the compositional complexity of petroleum products, it is impossible to assess the extent of petroleum hydrocarbon contamination by directly measuring the concentration of each hydrocarbon contaminant. For this reason, at the present time, no single analytical method is capable of providing comprehensive chemical information on petroleum contaminants. Total petroleum hydrocarbon is one parameter and definition that is currently widely used for expressing the total concentration of nonpolar petroleum hydrocarbons in soil, water, or other investigated samples. In the United States, for example, there are no federal regulations or guidelines for TPH in general. Many states have standards for controlling the concentrations of petroleum hydrocarbons or components of petroleum products. These are designed to protect the public from the possible harmful health effects of these chemicals. Analytical methods are specified as well, many of which are considered to be methods for TPH. These generate basic information that is a surrogate for contamination, such as a single TPH concentration. Such data are not suitable for risk assessment. However, they are relatively quick and easy to obtain and can offer useful preliminary information.

The term TPH is widely used, but it is rarely well defined. In essence, TPH is defined by the analytical method—in other words, estimates of TPH concentration often vary depending on the analytical method used to measure it. Thus, the ATSDR defines the TPH as a term used to describe a broad family of several hundred chemical compounds that originally come from crude oil. In this sense, TPH is really a mixture of chemicals. As per the TPHCWG, TPH, also called “hydrocarbon index,” refers sometimes to mineral oil, hydrocarbon oil, extractable hydrocarbon, oil, and grease. The TPHCWG also says that the TPH measurement is the total concentration of the hydrocarbons extracted and measured by a particular method, and it depends on the analytical method used for determination. According to the MA DEP, the TPH is also a loosely defined parameter, which can be quantified using a number of different analyses, and this parameter is an estimate of the total concentration of petroleum hydrocarbons in a sample. Again, depending on the analytical method used to quantify TPH, the TPH concentration may represent the entire range of petroleum hydrocarbons from C9 to C36 or the sum of concentrations of a number of single compounds (for instance, BTEX) and groups of compounds (fractions, e.g., primarily aliphatics C9–C18, C19-C36, and aromatics C11–C20). Great improvements in the definition and analysis of TPH were finally introduced by the International Organization for Standardization (ISO)[48] in 2000, when it published the standard method ISO 9377-2:2000[49] for the quality control of water in which a method for the determination of the hydrocarbon oil index within the C10–C40 range in waters by means of GC is specified. The definition of “hydrocarbon oil index by GC-FID” was introduced, which defines the fraction of compounds extractable with a hydrocarbon solvent, boiling point between 36°C and 69°C, not adsorbed on Florisil, and which may be chromatographed with retention times between those of *n*-decane ($C10H22$) and *n*-tetracontane ($C40H82$). (Substances complying with this definition are long-chain or branched aliphatic, alicyclic, aromatic, or alkyl -substituted aromatic hydrocarbons.)

The TPHCWG and MA DEP evaluated the risk implications and arrived at the conclusion that TPH concentration data cannot be used for a quantitative estimation of the human health risk. The same concentration of TPH may represent very different compositions and very different risks to human health and the environment because the TPH parameter includes a number of compounds of differing toxicities and the health effects associated with exposure to particular concentrations of TPH cannot be determined. For example, two sites may have the same amount of TPH, but constituents at one site may include carcinogenic compounds while these compounds may be absent at the other site. If TPH data indicate that there may be significant contamination of environmental media, then fractionated measurements and the separate determination of BTEX compounds and PAHs are necessary so that potential risk to human health can be quantitatively assessed.[50] The hydrocarbon index is thus a good indicator of the (magnitude of the) relative contamination of oil; however, it will not be suitable to give a true representation of the actual concentration of TPH in the investigated sample. There are several reasons why TPH data do not provide the ideal information for investigated samples and do not establish target cleanup criteria. This is due to many factors including the complex nature of petroleum hydrocarbons, their interaction with the environment over time, and the non-specificity of some of the methods used. The scope of the methods used for TPH determination varies greatly. There are few, if any, methods that are capable of quantifying all hydrocarbons without interference from non-hydrocarbons. All methods are subject to interferences from non-hydrocarbons, some to a greater extent than others.

There are numerous established analytical methods that are available for detecting, measuring, or monitoring TPH and its metabolites. Analytical methods used for analysis of petroleum hydrocarbons in environmental media should provide a sufficient degree of robustness. At the current time, however, the correctness and precision of results for the petroleum hydrocarbon determination strongly depend on the proper choice of method and measurement parameters whose correct selection is left to the judgment of the analyst. Besides methods that measure the TPH concentration, two other types of methods can be distinguished. These are methods that measure the concentration of a group or fraction of petroleum compounds and methods that measure individual petroleum constituent concentrations. For product identification, the results of analyses of the petroleum groups or fractions can be useful because they separate and quantify different categories of hydrocarbons. Individual constituent methods quantify concentrations of specific compounds that might be present in petroleum-contaminated samples, such as BTEX and PAHs, which can be used to evaluate human health risk.

There are several basic steps related to the separation of analytes of interest from a sample matrix prior to their measurement, such as extraction, concentration, and cleanup. These steps are common to the analytical processes for all methods, irrespective of the method type or the environmental matrix. Each of these steps together with the sampling, which is also an important step in performing petroleum analyses, affects the final result and has a certain impact onto the measurement uncertainty.[51,52]

Sample taking and sample handling have been recognized as probably the most significant factors that contribute random errors and uncertainties in the analysis of offshore oil in produced water. There are some general guidelines available through a number of studies that have been carried out on this subject. To separate the analytes from the matrix, extraction is performed using one of the many available extraction methods. Heating of the sample or purging with an inert gas can be used in the analysis of volatile compounds; solid-phase extraction or extraction into a solvent is usually applied for water samples, the latter extraction method also being used for soil samples. For some types of solid samples, the extraction efficiency depends on the extraction method and time. However, ultrasonication and extraction by shaking are equally used for this purpose. It was demonstrated by some studies that extraction and cleanup are the most crucial steps in sample preparation procedures. According to the results. the most critical factors affecting TPH recovery are extraction solvent and type of cosolvent, extraction time, adsorbent and its mass, and the TPH concentration.[53] The results of a study where the occurrence of matrix effects in the gas chromatographic determination of petroleum hydrocarbons in soil was evaluated indicate that solid-phase extraction does not appear to be effective enough in removing interfering matrix components from the extract.[54]

Most of the methods for the determination of TPH involve a cleanup step using Florisil (a particular form of magnesium silicate) and sodium sulfate (anhydrous), which essentially aims at removing the polar, non-petroleum hydrocarbons of biological origin and remaining traces of water. It appears that the found hydrocarbon concentration strongly depends on the used cleanup technique. The efficiency of the cleanup procedure for removing polar compounds is not limited to heteroatomic substances like O-, N-, or Cl-containing compounds. Also, some hydrocarbons have a tendency to adsorb on Florisil, e.g., aromatic compounds with p- electrons or alkyl aromatics. The TPH recoveries after a cleanup procedure might depend on the composition of the oil investigated. Lower TPH recoveries may be expected for oils containing high concentrations of unsaturated hydrocarbons or PAHs. Also, lubricating oils often contain different amounts and types of (non-petrogenic) additives that may behave differently from the other compounds during the cleanup procedure.[55] The results demonstrate also that the ratio of Florisil amount and extract volume are of importance for the recovery of the purified extracts.[56]

The three most commonly used TPH testing methods include GC,[49,57–60] infrared absorption (IR),[61,62] and gravimetric analysis.[63–65] Conventional TPH methods are summarized in Table 8.

Methods based on solvent extraction followed by quantitative IR measurement (at a frequency of 2930 cm^{-1}, which corresponds to the stretching vibration of aliphatic CH2 groups) have been widely used in the past for TPH measurement because they are simple, quick, and inexpensive. However, the use of these methods has been discontinued, since the sale and use of Freons (required for the extraction of hydrocarbons from the sample) is no longer allowed, and Freons are generally phased out worldwide due to their ozone layer–destructing potential. Recently, a new IR-based method was introduced, based on Freon- free extraction. This method defines oil and grease in water and wastewater as the fraction that is extractable with a cyclic aliphatic hydrocarbon (for example, cyclohexane) and measured by IR absorption in the narrow spectral region of 1370–1380 cm^{-1} (which corresponds to the excitation frequency of the symmetrical deformation vibration of CH_3 groups) using mid-IR quantum cascade lasers.[62] The method also considers the volatile fraction of petroleum hydrocarbons, which is lost by gravimetric methods that require solvent evaporation prior to weighing, as well as by solventless IR methods that require drying of the employed solid-phase material prior to measurement. Similarly, a more complete fraction of extracted petroleum hydrocarbon is accessible by this method as compared with GC methods that use a time window for quantification, as petroleum hydrocarbons eluting outside these windows are also quantified. On the other hand, IR-based methods hardly provide any information on the chemical composition of the oil or the presence or absence of other relevant compounds (aromatics, PAHs). In contrast, they even detect compounds that are not typically considered as TPH, such as surfactants, which also may absorb IR radiation due to the presence of CH bonds. However, this statement is only partially true, since it depends mainly on the cleanup whether the IR method determines also compounds other than the TPH.

Gravimetric-based methods are also simple, quick, and inexpensive; they measure anything that is extractable by a solvent, not removed during solvent evaporation, and capable of being weighed. Consequently, they do not offer any selectivity or information on the type of oil detected. Gravimetric-based methods may be useful for oily sludges and wastewaters at high(er) concentrations but are not suitable for measurement of light hydrocarbons (less than C15), which will be lost by evaporation below 70–85°C.

Gas chromatography–based methods are currently the preferred laboratory methods for TPH measurement because they detect a broad range of hydrocarbons, they provide both sensitivity and selectivity, and they can be used for TPH identification as well as quantification. The potential of GC for producing information on the product-specific hydrocarbon pattern has been long recognized by researchers in the field of petroleum hydro-carbon analysis. [66–68]

Currently, there are several standard methodologies based on GC for different types of samples (water, soil, wastes). The ISO has published the standard ISO 93772:2000 for the quality control of water and specifies a method for the determination of the hydrocarbon oil index within the C10–C40 range in waters by means of GC. The method is suitable for surface water, wastewater, and water from sewage treatment plants and allows the determination of the hydrocarbon oil index in concentrations above 0.1 mg/L. Due to systematic differences, which became evident between the results from the DIN ISO method and those from the IR-based method, the GC-based method was subsequently modified.[69,70] As a result, the modified version of DIN ISO 9377-2:2000, the OSPAR (Oslo–Paris commission) reference method,[58] was published in 2005 and taken into force as a reference method in the field of petroleum production in January 2007. The OSPAR reference method is applicable for the determination of dispersed oil content in produced water and other types of wastewater discharged from gas, condensate, and oil platforms. It also allows the determination of the dispersed mineral oil content in concentrations above 0.1 mg/L and includes the determination of certain hydrocarbons within the C7–C10 range, with the TEX (toluene, ethylbenzene, and *o*-/*p*-/*m*-xylene) compounds being reported separately.

TABLE 8 Summary of Common TPH Methods

Analytical Method	Method Name	Matrix	Scope of Method	Carbon Range	Approximate Detection Limits	Advantages	Limitations	Reference
GC based	DIN ISO 9377-2:2000	Water	Solvent (hydrocarbon) extraction, cleanup using Florisil, evaporation, 1 µL injection, GC-FID	C10-C40	0.1 mg/L	Detects broad range of hydrocarbons; provides information (e.g., a chromatogram) for identification	Does not quantify below C10; chlorinated compounds can be quantified as TPH	[49]
	OSPAR (2007)	Water	n-Pentane extraction, cleanup using Florisil, 50 µL injection, GC-FID	C7–C40 + TEX compounds	0.1 mg/L	Does not need preconcentration step; detect broad range of hydrocarbons and polar hydrocarbons; provide information for identification	Does not quantify below C7	[57]
	DIN ISO 16703:2005–12	Soil	Acetone/n-heptane extraction, cleanup using Florisil, evaporation, GC-FID	C10–C40	10 mg/kg	Detects broad range of hydrocarbons; provide information (e.g., a chromatogram) for identification	Does not quantify below C10; chlorinated compounds can be quantified as TPH	[58]
	DIN EN 14039:2004	Wastes	Acetone/n-heptane extraction, cleanup using Florisil, evaporation, GC-FID	C10–C40	10 mg/kg			[59]
IR based	EPA 418.1 (1991/1992)	Water, soil	Freon extraction, silica gel treatment to remove polar compounds	Most hydrocarbons with exception of volatile and very high hydrocarbons	1 mg/mL in water, 10 mg/kg in soil	Technique is simple, quick, and inexpensive	Freon is banned now; low sensitivity; lack of specificity; prone to interference; provides quantitation only	[62]
	ASTM D7678-11 (2011)	Water, wastewater	Solvent (cyclic aliphatic hydrocarbon) extraction, cleanup using Florisil, IR Absorption in the region of 1370–1380 cm^{-1} (7.25–7.30 mm)	Most hydrocarbons with volatile	0.5 mg/mL	Technique is simple, very quick; a more complete fraction of extracted petroleum hydrocarbon is accessible		[61]

(*Continued*)

TABLE 8 (*Continued*) Summary of Common TPH Methods

Analytical Method	Method Name	Matrix	Scope of Method	Carbon Range	Approximate Detection Limits	Advantages	Limitations	Reference
Gravimetry	EPA 413.1 (1979) ASTM D4281–95(2005)el	Most appropriate for wastewater, sludge, sediment	Freon extraction, solvent evaporation	Anything that is extractable (with exception of volatiles which are lost)	5 mg/mL in water, 50 mg/kg in soil	Technique is simple, quick, and inexpensive	Freon is banned now; low sensitivity; lack of specificity not suitable for low boiling fractions; prone to interference (organic acids, phenols, and other polar hydrocarbons); provides quantitation only	[63,65]
	EPA 1664 (1999)	Most appropriate for water and wastewater	n-Hexane extraction, silica gel treatment to remove polar compounds, solvent evaporation	Anything that is extractable (with exception of volatiles which are lost)	5 mg/mL	Technique is simple, quick, and inexpensive	Low sensitivity; lack of specificity not suitable for low boiling fractions; prone to interference; provides quantitation only	[64]

Gas chromatography–based methods are based on the extraction of water samples with a nonpolar (hydrocarbon) solvent, the removal of polar substances by cleanup with Florisil, and capillary GC measurements using a nonpolar column and a flame ionization detector (FID), cumulating the total peak area of compounds eluted between *n*- decane (C10H22) and *n*-tetracontane ($C_{40}H_{82}$) for the DIN ISO 9377-2:2000 standard method and for the DIN ISO 16703:2005–12[59] standard method for soil samples. The OSPAR method was modified in order to include the determination of certain hydrocarbons with a boiling point between 98°C and 174°C (that is, from *n*-heptane to *n*-decane), with the TEX compounds being determined separately by integration and subtraction of their peak areas from the total integrated area. The GC-based methods usually cannot quantitatively detect compounds with a lower boiling point than *n*-heptane because these compounds are highly volatile and are interfered by the solvent peak. Furthermore, the EPA method 8240,[61] which is used to determine volatile organic compounds in a variety of waste matrices by GC/mass spectrometry (MS), exists. It can be used to quantitate most volatile organic compounds that have boiling points below 20°C and that are insoluble or slightly soluble in water. The estimated quantitation limit of the EPA 8240 method for an individual compound is approximately 5 μg/kg (wet weight) for soil/sediment samples, 0.5 mg/kg (wet weight) for wastes, and 5 μg/L for groundwater.

Gas chromatography–based methods are suitable for surface water, wastewater, and other types of wastewater discharged from gas, concentrate, and oil platforms and allow the determination of hydrocarbon oil concentration above 0.1 mg/L. To reach the required detection limit, the method according to DIN ISO 9377-2:2000 foresees preconcentration of the extracts by solvent evaporation, which bears the risk of losing the more volatile constituents of the sample. In contrast to this, the OSPAR method does not allow for any external apparatus for preconcentration, for which reason the GC must be equipped with an injection system that allows the injection of a volume of up to 100 μL of the extract. This is most easily realized with programmed-temperature vaporizer large-volume injectors. This technique can reduce the loss of volatile analytes, can increase sensitivity, and is a viable, fast, and automated alternative to an external preconcentration procedure.[71–73]

Petroleum products easily contain thousands of different compounds. Classical capillary GC cannot resolve such mixtures up to the level of individual compounds. A powerful analytical tool for separation of complex mixtures, such as petroleum hydrocarbons, is comprehensive two-dimensional GC (GC×GC or 2D-GC).[74–76] The use of 2D-GC with MS detection (GC×GC/MS) is expected to not only allow the separation of the various constituents of complex TPH samples but also to identify them based on MS detection (Figure 4b). It is known that a certain class of chemical compounds (a series of "homologues") forms a very distinct, clearly identifiable pattern in the two-dimensional space of the GC×GC separation. The diesel total ion (TIC) GC×GC/MS chromatogram, illustrated in Figure 4a, is characterized by very typical group-type patterns: saturated hydrocarbons, which present low second-dimension retention times, are followed by monocyclic and dicyclic aromatics; tri-and tetracyclic aromatics are the most retained on the secondary polar column.[77] Moreover, partial overlapping between chemical groups occurs, the monocyclic aromatics are situated in a rather narrow band, and the tri-and tetracyclic aromatics are hardly visible in the two-dimensional chromatogram. The analytical potential of such a two-dimensional system is great.

Conclusion

Due to the importance and widespread use of petroleum hydrocarbons for energy production, for transport, and as a raw material in the chemical industries, there are many routes for their inadvertent or accidental release into the environment. Thus, they do represent one of the most important sources of large-scale environmental pollution. While petroleum hydrocarbons also are introduced into the oceans from natural seeps, these continuous emissions of comparatively low intensity represent a less significant environmental problem since the resident flora and fauna have adapted to this continuous input of hydrocarbons and effects are limited to local scale. Large oil spills in contrast exceed the self-cleaning capacity

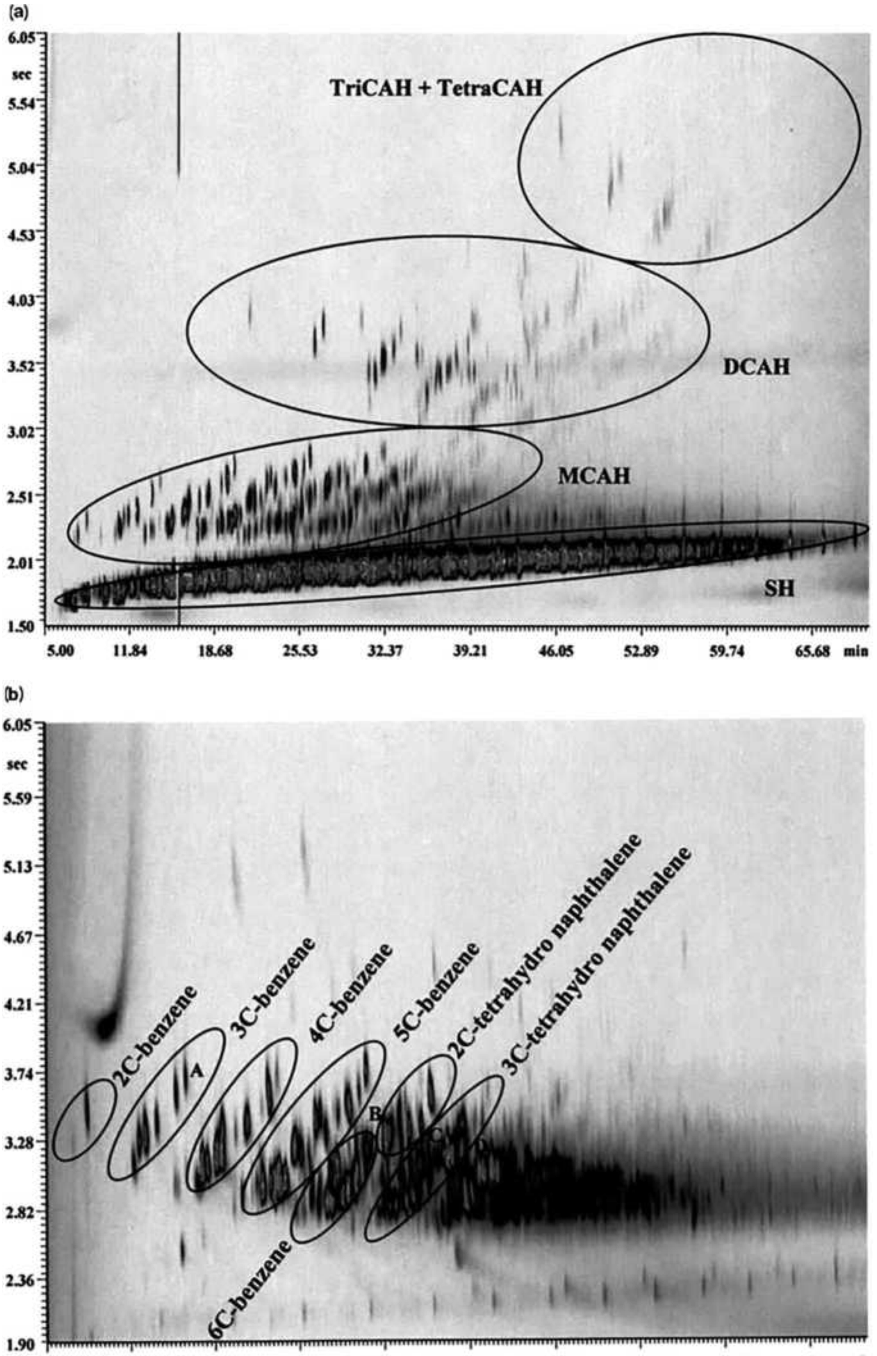


FIGURE 4 (a) TIC GC×GC-qMS (quadruple-mass spectrometry) chromatogram of diesel oil. SH, saturated hydrocarbons; MCAH, monocyclic aromatics; DCAH, dicyclic aromatics; TriCAH, tricyclic aromatics; TetraCAH, tetracyclic aromatics. (b) TIC LC-GC×GC- qMS chromatogram of the monocyclic aromatic fraction of diesel oil. A) Indane, B) 1,2,3,4-Tetrahydro-2,7-dimethyl naphthalene, C) 1-Cyclohexyl 3-methyl benzene, D) 1,2,3,4-Tetrahydro-2,5,8-trimethyl naphthalene.
Source: Adapted from Sciarrone et al.[77]

of the ecosystem, which cannot regenerate without human intervention to both physically and chemically immobilize, bind, and remove oil from the affected region. Although such techniques are available, large-scale oil spills always have caused severe damage to the environment, with the affected ecosystems recovering only slowly. Analytical methods are available for the qualitative and quantitative determination of the composition of oil samples and the assessment of pollution levels in various environmental compartments. Gas chromatographic techniques mostly have supplanted the former analytical standard method based on Freon extraction and mid-IR determination, but there is further research and development going on to develop either more powerful analytical methods—such as two-dimensional GC—or alternative detection methods, such as the ones based on mid-IR lasers as light sources.

Acknowledgments

This report was compiled within the frame of project 818084-16604 SCK/KUG of the Austrian Science Foundation (FFG), whose financial support is gratefully acknowledged.

References

1. Tissot, B.P.; Welte, D.H. *Petroleum Formation and Occurrence;* Springer-Verlag: Berlin, 1984.
2. Availableathttp://www.oxforddictionaries.com/view/entry/m_en_gb0623910#m_en_gb0623910 (accessed September 2011).
3. Available at http://www.nationmaster.com/graph/ene_oil_pro-energy-oil-production (accessed September 2011).
4. Available at http://www.eia.gov/petroleum/ (accessed September 2011).
5. Leffler, W.L. *Petroleum Refining in Nontechnical Language*, 4th Ed.; PennWell Corporation: Tulsa, OK, 2008.
6. Fahimm, M.A.; Al-Sahhaff, T.A.; Lababidii, H.M.S.; Elkilanii A. *Fundamentals of Petroleum Refining*, 1st Ed.; Elsevier: Oxford, U.K., 2010.
7. Available at http://www.energyinst.org (accessed September 2011).
8. Nancarrow, D.J.; Adams, A.L.; Slade, N.J.; Steeds, J.E. Land *Contamination: Technical Guidance on Special Sites: Petroleum Refineries*; R&D Technical Report P5–042/ TR/05; Environment Agency: Bristol, 2001.
9. *Toxicological Profile for Total Petroleum Hydrocarbons (TPH)*; U.S. Department of Health and Human Services, Public Health Service Agency for Toxic Substances and Disease Registry: Atlanta, GA, 1999.
10. Leifer, I.; Kamerling, M.J.; Luyendyk, B.P.; Douglas, S.W. Geologic control of natural marine hydrocarbon seep emissions, Coal Oil Point seep field, California. Geo-Mar. Lett. **2010,** *30,* 331–338.
11. *Oil in the Sea III Inputs, Fates, and Effects;* The National Academies Press: Washington, DC, 2003.
12. Available at http://www.toptenz.net/top-10-worst-oil-spills. php (accessed September 2011).
13. Weggen, K.; Pusch, G.; Rischmüller, H. Oil and gas. In *Ullmann's Encyclopedia of Industrial Chemistry;* Wiley-VCH: Weinheim, 2000.
14. Sanin, P.I. Petroleum hydrocarbons. Russ. Chem. Rev. **1976**, *45* (8), 684–700.
15. Available at http://api-ep.api.org/ (accessed September 2011).
16. Analysis of petroleum hydrocarbons in environmental media. In *Total Petroleum Hydrocarbon Criteria Working Group Series;* Weisman. W., Ed.; Amherst Scientific Publishers: Amherst, MA, 1998; vol. 1.
17. Wang, Z.; Fingas, M.F. Development of oil hydrocarbon fingerprinting and identification techniques. Mar. Pollut. Bull. **2003,** *47,* 423–452.
18. Wang, Z.; Fingas, M.; Pag.e, D.S. Oil spill identification. J. Chromatogr. A **1999,** *843,* 369–411.
19. Available at http://www.etc-cte.ec.gc.ca/databases/OilProperties/oil_prop_e.html (accessed September 2011).

20. Availableat http://www.statoil.com/en/ouroperations/tradingproducts/crudeoil/crudeoilassays/pages/default.aspx (accessed September 2011).
21. Availableathttp://www.castrol.com/liveassets/bp_internet/castrol/castrol_switzerland/STAGING/local_assets/ downloads/a/ABC_D_Mai_2009.pdf (accessed September 2011).
22. *Spills of Emulsified Fuels: Risks and Responses;* The National Academy of Sciences: Washington, DC, 2001.
23. *Petroleum Products in Drinking-Water;* Background document for development of WHO Guidelines for Drinking-water Quality, WHO/SDE/WSH/05.08/123; World Health Organization: Geneva, 2005.
24. *Fate of Spilled Oil In Marine Waters: Where Does It Go? What Does It Do? How Do Dispersants Affect It?;* An Information Booklet for Decision Makers, Publication 4691; American Petroleum Institute: Virginia, 1999.
25. Afifi, S.M. Petroleum hydrocarbon contamination of groundwater in Suez: causes severe fire risk. Proceedings 24th AGU Hydrology Days, March 10–12, 2004, pp. 1–9. Colorado State University (2004).
26. Vaajasaar, K.; Joutti, A.; Schultz, E.; Selonen, S.; Westerholm, H. Comparisons of terrestrial and aquatic bioassays for oil-contaminated soil toxicity. J. Soils Sediments **2002,***2* (4), 194–202.
27. *Guidelines for Assessing and Managing Petroleum Hydrocarbon Contaminated Sites in New Zealand;* Module 2—Hydrocarbon contamination fundamentals; Ministry for Environment: New Zealand, 1999.
28. Das, N.; Chandran, P. Microbial degradation of petroleum hydrocarbon contaminants: An overview. Biotechnol. Res. Int. **2011,** Article ID 941810, 13 pages.
29. Atlas, R.M. Microbial degradation of petroleum hydrocarbons: An environmental perspective. Microbiol. Rev. **1981,** *45,* 180–209.
30. Van Hamme, J.D.; Singh, A.; Ward, O.P. Recent advances in petroleum microbiology. Microbiol. Mol. Biol. Rev.**2003**,*67* (4), 503–549.
31. Kaplan, I.R.; Galperin, Y.; Lu, S.T.; Lee, R.P. Forensic environmental geochemistry: Differentiation of fuel-types, their sources and release time. Org. Geochem. **1997,** *2,* 289–317.
32. Šepieč, E.; LeskovŠek, H.; Trier, C. Aerobic bacterial degradation of selected polyaromatic compounds and n-alkanes found in petroleum. J. Chromatogr. A **1995,** *697,* 515–523.
33. Antić, M.P.; Jovaneičević, B.S.; Ilić, M.; Vrvić, M.M.; Schwarzbauer, J. Petroleum pollutant degradation by surface water microorganisms. Environ. Sci. Pollut. Res. **2006,** *13* (5), 320–327.
34. Available at http://www.atsdr.cdc.gov/ (accessed September 2011).
35. Available at http://echa.europa.eu/home_en.asp (accessed September 2011).
36. Available at http://www.environment-agency.gov.uk/ (accessed September 2011).
37. *The U.K. Approach for Evaluating Human Health Risks from Petroleum Hydrocarbons in Soil;* Science report P5–080/TR3; Environment Agency: Bristol, U.K., 2005.
38. MA DEP 1994. *Interim Final Petroleum Report Development of Health-Based Alternative to the Total Petroleum Hydrocarbon TPH Parameter*; Massachusetts Department of Environmental Protection: Boston, Massachusetts, 1994.
39. MA DEP 2003. *Updated Petroleum Hydrocarbon Fraction Toxicity Values for the VPH/EPH/APH*; Massachusetts Department of Environmental Protection: Boston, Massachusetts, 2003.
40. Edwards, D.A.; Andriot, M.D.; Amoruso, M.A.; Tummey, A.C.; Tveit, A.; Bevan, C.J.; Hayes, L.A.; Youngren, S.H.; Nakles, D.V. Development of fraction specific reference doses (RfDs) and reference concentrations (RfCs) for total petroleum hydrocarbons. In *Total Petroleum Hydrocarbon Criteria Working Group Series*; Amherst Scientific Publishers: Amherst, Massachusetts, 1997; Vol. 4.
41. Potter, T.L.; Simmons, K.E. Composition of petroleum mixtures. In *Total Petroleum Hydrocarbon Criteria Working Group Series;* Amherst Scientific Publishers: Amherst, Massachusetts, 1998; Vol. 2.
42. Vorhees, D.J.; Weisman, W.H.; Gustafson, J.B. Human health risk-based evaluation of petroleum release sites: implementing the working group approach. In *Total Petroleum Hydrocarbon Criteria Working Group Series*; Amherst Scientific Publishers: Amherst, Massachusetts, 1999; Vol. 5.

43. MA DEP 2003. *Method for the Determination of Air-Phase Petroleum Hydrocarbons (APH)*; Massachusetts Department of Environmental Protection: Boston, Massachusetts, 2009.
44. MA DEP 2004. *Method for the Determination of Volatile Petroleum Hydrocarbons (VPH);* Massachusetts Department of Environmental Protection: Boston, Massachusetts,2004.
45. MA DEP 2004. *Method for the Determination of Extract-able Petroleum Hydrocarbons (EPH);* Massachusetts Department of Environmental Protection: Boston, Massachusetts, 2004.
46. Available at http://www.epa.gov/ (accessed September 2011).
47. Verbruggen, E.M.J. *Environmental Risk Limits for Mineral Oil (Total Petroleum Hydrocarbons)*; RIVM report 601501021; National Institute for Public Health and the Environment: Bilthoven, the Netherlands, 2004.
48. Available a3t http://www.iso.org/iso/home.html (accessed September 2011).
49. DIN ISO 9377-2:2000. *Water Quality—Part 2, Method Using Solvent Extraction and Gas Chromatography*; International Organization for Standardisation: Geneva, 2000.
50. Pollard, S.J.T.; Duarte-Davidson, R.; Askari, K.; Stutt, E. Managing the risk from petroleum hydrocarbons at contaminated sites achievements and future research directions. Land Contam. Reclam. **2005,** *13* (2), 115–122.
51. Saari, E.; Perämäki, P.; Jalonen, J. Measurement uncertainty in the determination of total petroleum hydrocarbons (**TPH**) in soil by GC-FID. Chemom. Intell. Lab. Syst. **2008,** *92* (1), 3–12.
52. Becker, R.; Buge, H.G.; Bremser, W.; Nehls, I. Mineral oil content in sediments and soils: Comparability, traceability and a certified reference material for quality assurance. Anal. Bioanal. Chem. **2006,** *385* (3), 645–651.
53. Saari, E.; Perämäki, P.; Jalonen, J. Evaluating the impact of extraction and cleanup parameters on the yield of total petroleum hydrocarbons in soil. Anal. Bioanal. Chem. **2008,** *392* (6), 1231–1240.
54. Saari, E.; Perämäki, P.; Jalonen, J. Effect of sample matrix on the determination of total petroleum hydrocarbons (**TPH**) in soil by gas chromatography–flame ionization detection. Microchem. J. **2007,** *87* (2), 113–118.
55. Muijs, B.; Jonker, M.T.O. Evaluation of clean-up agents for total petroleum hydrocarbon analysis in biota and sediments. J. Chromatogr. A **2009,** *1216* (27), 5182–5189.
56. Koch, M.; Liebich, A.; Win, T.; Nehls, I. *Certified Reference Materials for the Determination of Mineral Oil Hydrocarbons in Water, Soil and Waste;* Forschungsbericht 272; Bundesanstalt für Materialforschung und - prüfung (BAM): Berlin, 2005.
57. OSPAR. *Reference Method of Analysis for Determination of the Dispersed Oil Content in Produced Water*; OSPAR Commission, ref. no. 2005–15: Malahide, published in 2005, *taken into force in 2007.*
58. DIN ISO 16703:2005–12. *Soil Quality—Determination of Content of Hydrocarbon in the Range C10 to C40 by Gas Chromatography*; International Organization for Standardisation: Brussels, 2005.
59. DIN EN 14039:2004. *Characterization of Waste—Determination of Hydrocarbon Content in the Range of C10 to C40 by Gas Chromatography*; German version; International Organization for Standardisation: Brussels, 2004.
60. EPA Method 8240. *Gas Chromatography/Mass Spectrometry for Volatile Organics, Test Methods for Evaluating Solid Wastes*; US Environmental Protection Agency: Washington, 1986, Vol. 1B.
61. ASTM Standard D7678111. *Standard Test Method for Total Petroleum Hydrocarbons (TPH) in Water and Wastewater with Solvent Extraction using Mid-IR Laser Spectroscopy*; ASTM International: West Conshohocken, PA, 2011; DOI: 10.1520/D7678–11.
62. EPA Method 418.1. *Total Recoverable Petroleum Hydrocarbons by IR, Groundwater Analytical Technical Bulletin*; Groundwater Analytical Inc.: Buzzards Bay, MA, 1991/1992.
63. EPA method 413.1. *Standard Test Method for Oil and Grease Using Gravimetric Determination*; issued in 1974, editorial revision in 1978 (withdrawn).
64. EPA method 1664. *Revisiow A: w-Hexawe Extractable Material (HEM; Oil awd Grease) awd Silica Gel Treated n-Hexawe Extractable Material (SGT-HEM: Now-polar Material) by Extractiow awd Gravimetry*; US Environmental Protection Agency: Washington, 1999.

65. ASTM D4281–95(2005)e1. *Standard Test Method for Oil awd Grease (Fluorocarbow Extractable Substawces) by Gravimetric Determiwatiow;* ASTM International: West Conshohocken, PA, 2005.
66. Blomberg, J.; Schoenmakers, P.J.; Brinkman, U.A.T. GC methods for oil analysis. J. Chromatogr. A **2002,** *972* (2), 137–173.
67. Beens, J.; Brinkman, U.A.T. The role of GC in compositional analyses in the petroleum industry. Trends Anal. Chem. **2000,** *19* (4), 260–275.
68. Saari, E.; Perämäki, P.; Jalonen, J. Evaluating the impact of GC operating settings on GC–FID performance for total petroleum hydrocarbon (TPH) determination. Microchem. J. **2010,** *94* (1), 73–78.
69. Thomey, N.; Bratberg, D.; Kalisz, C. A comparison of methods for measuring total petroleum hydrocarbons in soil. In Proceedings of the Petroleum Hydrocarbons and Organic Chemicals in Groundwater: Prevention, Detection and Restoration, November 15–17, 1989; National Water Well Association: Houston, Texas, 1989.
70. Xie, G.; Barcelona, M.J.; Fang, J. Quantification and interpretation of total petroleum hydrocarbons in sediment samples by a GC/MS method and comparison with EPA 418.1 and a rapid field method. Anal. Chem. **1999,** *71* (9), 1899–1904.
71. Hoh, E.; Mastovska, K. Large volume injection techniques in capillary gas chromatography. J. Chromatogr. A **2008,** *1186,* 2–15.
72. Miñones Vázqiez, M.; Vázquez Blanco, M.E.; Muniategui Lorenzo, S.; Loópez Mahía, P.; Fernández-Fernández, E.; Prada Rodríguez, D. Application of programmed-temperature split/splitless injection to the trace analysis of aliphatic hydrocarbons by gas chromatography. J. Chromatogr. A **2001,** *919,* 363–371.
73. Dellavedova, P.; Vitelli, M.; Ferraro, V.; Di Toro, M.; Santoro, M. *Applicatiow of enhawced large volume injectiow; an approach to the analysis* of *petroleum hydrocarbons iw water.* Chromatographia **2006,** *63,* 73–76.
74. Van De Weghe, H.; Vanermen, G.; Gemoets, J.; Lookman, R.; Bertels, D. Application of comprehensive two-dimensional gas chromatography for the assessment of oil contaminated soils. J. Chromatogr. A **2006,** *1137,* 91–100.
75. von Mühlen, C.; Alcaraz Zini, C.; Bastos Caramão, E.; Marriott, P. Applications of comprehensive two-dimensional gas chromatography to the characterization of petrochemical and related samples. J. Chromatogr. A **2006,** *1105,* 39–50.
76. van Deursen, M.M.; Beens, J.; Reijenga, J.C.; Lipman, P.J.L.; Camers, C.A.M.G.; Blomberg, J. Group-type identification of oil samples using comprehensive two-dimensional gas chromatography coupled to a time-of-flight mass spectrometer (GCxGC-TOF). J. High Resolut. Chromatogr. **2000,** *23* (7–8), 507–510.
77. Sciarrone, D.; Tranchida, P.Q.; Costa, R.; Donato, P.; Ragonese, P.; Dugo, P.; Dugo, G.; Mondello, L. Offline LC-GCxGC in combination with rapid-scanning quadrupole mass spectrometry. J. Sep. Sci. **2008,** *31,* 3329–3336.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

4

Road-Traffic Emissions

Fabian Heidegger,
Regine Gerike,
Wolfram Schmidt,
Udo Becker, and
Jens Borken-Kleefeld

Introduction

The transport sector is a major contributor of air pollution and greenhouse gas emissions worldwide. Governments around the world are under pressure to deal with the consequences of climate change while also striving to meet growing transport demands. The prime challenge is to find and apply measures that reduce the environmental impact of transport, i.e., emissions (including air pollution, greenhouse gases, noise and vibration), land use, separation effects, and effects on landscape.

This chapter focuses solely on the topic of emissions, which includes energy consumption, air pollutants and greenhouse gases of road traffic and their measurements, as well as the possibility to model and influence the reduction of emissions via traffic planning measures, traffic management, or economic policy and user-related measures. Off-road emissions from vehicles are not included in this article, nor are upstream or downstream processes (e.g., well-to-tank emissions). Noise also contributes to road-traffic emissions but is not discussed here; however, it is important to note that the impact of noise can have major consequences on humans and animals.

Particularly in emerging economies, the aviation and road transport sectors have contributed considerably to the increase in emissions worldwide (although the latter sector has a high potential for emission reduction). The problems caused by road-traffic emissions can be separated into three categories:

- Climate Impact (mainly CO_2 emissions): In 2016, the transport sector produced around 8.0 $GtCO_2$ (25% of the total emissions that year), which measured 71% higher than those in 1990 (International Energy Agency 2018); 74% of those emissions were produced by road transport (International Energy Agency 2018).

- Major Share in Oil Consumption: With 92% of transport-final energy demand being composed of oil products, the transport sector is the least diversified energy end-use sector (International Energy Agency 2017). Hence, limited fossil fuel resources ensure a large share of mobility needs, especially in rural but also in urban areas today. Despite future energy efficiency gains and a wider application of electric cars, fuel needs will be higher in 2040 than they are today as a direct result of an increase in car use (International Energy Agency 2018).
- Air Pollutant Emissions: Many substances directly impact human health; in particular, there is high emission density in urban areas, which drives cities to develop solutions such as anti-pollution schemes.

Emissions are released in gaseous or particulate form into the atmosphere and are subject to physical and chemical transformation processes as well as to the influence of meteorological parameters during the transmission process (from the source to the place with pollution damage). Depending on the location, pollution damage has an effect on human health, vegetation, soil, water, building materials, and the climate. These substances can be grouped by the following impact categories (VDI 2019):

- Climate relevant substances (e.g., CO_2, CH_4, N_2O, black carbon)
- Substances contributing to ozone formation (e.g., NO_x, non-methane hydrocarbons [NMHC], carbon monoxide [CO])
- Eutrophying substances (e.g., NO_x, SO_2)
- Acidifying substances (e.g., NH_3, NO_x)
- Toxic substances (e.g., black carbon, heavy metals, polycyclic aromatic hydrocarbons, CO)
- Carcinogenic substances (e.g., black carbon, heavy metals, lead [Pb], aromatic hydrocarbons, e.g., benzene [C_6H_6], polycyclic aromatic hydrocarbons).

In the context of emissions and pollutants, the substances PM_{10} and NO_x (NO and NO_2) present the biggest problems in transportation. Particulate matter (PM) includes particles <10 μm in diameter (PM_{10}) as well as particles <2.5 μm in diameter ($PM_{2.5}$) (Pfäfflin 2018). PM occurs in incomplete combustion processes or due to abrasion and resuspension. PM may cause respiratory and cardiovascular diseases, and the smaller the size of the particle, the more dangerous it is to human health. NO_x is formed during high-temperature and high-pressure combustion processes and can lead to respiratory disorders, inflammation, and bronchitis. With long-term exposure, the number of heart attacks and fatalities as a result of respiratory disease significantly increases (Pfäfflin 2018).

Table 1 illustrates thresholds for nitrogen dioxide (NO_2), PM_{10}, and $PM_{2.5}$ in Europe. It is clearly visible that the guideline values of the WHO (World Health Organization) are substantially lower than the legally effective values of the European Union.

The European Union legislation directed toward creating a framework for the compliance of motor vehicles and their components limits the mass and number of particles, NO_x, and additional air pollutants, as well as greenhouse gases for vehicles (DIRECTIVE 2007/46/EC 2007), (REGULATION No

TABLE 1 Thresholds for Nitrogen Dioxide (NO_2), PM_{10}, and $PM_{2.5}$

	NO_2 [μg/m_3]		PM_{10} [μg/m_3]		$PM_{2.5}$ [μg/m_3]	
EU	40	Annual mean of hourly measurements	40	Annual mean of hourly measurements	25	Annual mean of hourly measurements
	–	–	50	24-hour mean with 35 permitted exceedances	–	–
WHO	40	Annual mean of hourly measurements	20	Annual mean of hourly measurements	10	Annual mean of hourly measurements
	200	1-hour mean	50	24-hour mean	25	24-hour mean

Source: According to (DIRECTIVE 2008/50/EC 2008) and (WHO 2006).

715/2007/EC 2007). Mainly CO_2 emissions are relevant regarding the direct climate impact of road transportation. As greenhouse gases do not cause local environmental damages or injuries to health, there is no need for local greenhouse gas thresholds. In 2009, the European Union introduced mandatory targets for average CO_2 emissions of 130 g CO_2/km for 2015 and 95 g CO_2/km for 2021 to be applied to each newly manufactured vehicle fleet (REGULATION No 333/2014/EU 2014).

Road vehicles are certified according to exhaust emission standards. In the United States, the Environmental Protection Agency (EPA) manages emission standards nationally. The European Union has its own set of emission standards, which all new vehicles must meet: The latest *Euro 6* targets passenger cars, light commercial vehicles, and motorcycles, and the *Euro VI* is for large goods vehicles and coaches (DIRECTIVE 2007/46/EC 2007), (REGULATION No 715/2007/EC 2007). Many Asian countries adopted these European emission standards; however, since the specified driving cycles, e.g., the WLTC (Worldwide Harmonized Light Vehicles Test Cycle), do not adequately represent real driving performance, the official limit values for air pollutants[1] and greenhouse gases (CO_2) are below actual driving emissions (although the vehicles fulfill the requirements of the test) (Hooftman et al. 2018). Therefore, the official exhaust thresholds cannot be used for a true and reliable emission calculation. In recent years, the topic of air pollution has received remarkable attention from the public, as well as from environmental and consumer organizations. Legal actions have been enforced upon vehicle manufacturers (e.g., Volkswagen, Daimler, Fiat, and Chrysler) for emission frauds (i.e., not complying with the limit thresholds of their vehicles) and upon cities and governments due to noncompliance with air quality standards.

Reliable emission projections/inventories combined with an integrated and systematic reduction of emissions—preferably in the form of pro-active prevention—are greatly needed. Substantial energy use and high levels of exhaust emissions are an unwanted outcome of transport demand; these are a consequence of many individual and non-individual interconnected processes. Mode choice and route choice, for example, are behaviors which could be adjusted rather quickly toward more sustainable options, while the implementation of change on the density of transportation networks, transportation planning, land use, and other economic factors would require an extensive amount of time and planning. In addition, emission factors (how much each vehicle emits per km [g/km or g/s]) and fuel use factors [l/km] of the vehicles, depending on, for example, vehicle type choice, topography, and driving behavior, are influencing coefficients for exhaust emissions.

This chapter is organized as follows: “Modeling of Road-Traffic Emissions” section gives an overview of energy consumption, vehicle propulsion technologies, default parameters of emission calculation, and types of emissions; “Potential Influences on Emission Reduction and Their Impacts” section introduces the modeling of road-traffic emissions by presenting different approaches and model types, followed by an example of an emission model and validation measures of emissions; and “Conclusion” section is dedicated to the potential influences for emission reduction and their impact, which can also be found in anti-air pollution schemes and climate action plans of cities and countries.

Energy Consumption and Types of Emissions

Default Parameters for Emission Calculation

Figure 1 summarizes the default parameters (all boxes on top) for determining road-traffic emissions. Colored (straight) arrows indicate effects to the type of emissions (all boxes in the middle); for example, the roughness of road surface will have an impact on abrasion and resuspension emissions. Moreover, vehicle fleet composition and traffic mileage affect the sum of all traffic emissions (box at the bottom). Resultant effects are shown as black (arcuated) arrows which close the circuit in a mutually reinforcing manner: An increase in the release of emissions directly affects climate change and meteorological conditions, just as

---

[1] True notably for NO_x (from diesel cars), but not necessarily for other regulated pollutants.

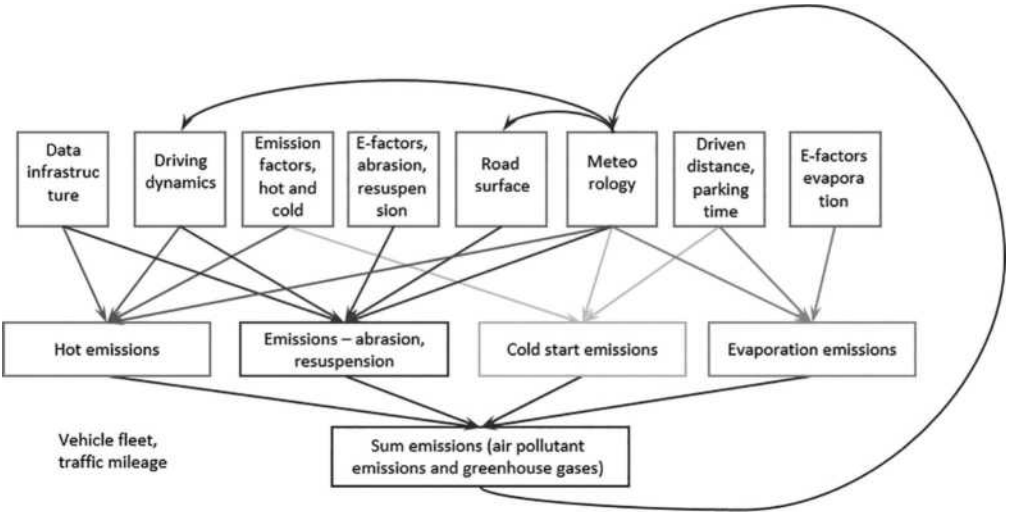


FIGURE 1 Default parameters (all boxes on top) for determining road-traffic emissions with resultant effects.

weather conditions affect driving dynamics (driving behavior) and road surfaces. Though not shown in Figure 1, emissions from trains, ships, airplanes, power stations and industry, domestic fuels and small businesses, mobile sources, agriculture, and biogenic sources are also part of the circuit.

Basics of Driving Dynamics

A vehicle moves or accelerates by converting externally added energy (fuel, electricity) into kinetic energy. When braking, kinetic energy is transferred to thermal energy. Some vehicle concepts are also able to recuperate and store electric energy for subsequent vehicle movement.

The level of energy consumption of a vehicle is dependent on the kinetic energy demand (to overcome driving resistances) as well as on the degree of efficiency of the engine and the use of auxiliary units such as air conditioning or lighting.

Kinetic energy demand arises from vehicle parameters and traffic situations (Pfäfflin 2018). Vehicle parameters include

- Acceleration Resistance: Overcoming vehicle mass inertia (including loading) and rotating masses (wheels, gearboxes, engine)
- Air Resistance: Force of displacement depending on vehicle size, drag coefficient, and speed of the vehicle
- Rolling Resistance: A result of tire contact with the road, the inner frictional resistance, and the weight force
- Gradient Resistances: Dependent on vehicle mass and topography.

During low speeds or stop-and-go traffic situations, the energy demand for the acceleration of vehicle mass is the decisive parameter. With a constant driving speed, rolling resistance and air resistance—especially at higher speeds—are the dominating factor for energy demand.

Energy Conversion in the Vehicle

As previously mentioned, kinetic energy is generated in either a combustion or an electric engine. Electric engines have a higher degree of efficiency than combustion engines. A hybrid vehicle obtains its energy from an electric engine and a fuel-based engine. Combustion engines are characterized by a cyclic, non-stationary, and non-optimal combustion process (Pfäfflin 2018). For a comparable vehicle type with similar engine power, the diesel engine has an advantage in terms of fuel consumption as compared to the gasoline engine. The spark ignition of the diesel leads to a more energy-efficient yield of fuel: This combustion process results in less fuel consumption and lowers CO_2 emissions; however, rising combustion temperatures lead to greater NO_x emissions.

Types of Emissions

The following is a list of the main types of emissions, divided by their components, as well as the chemical reactions in which they are released:

- Ammoniac (NH_3) and Dinitrogen (N_2O)—Released through catalytic converters
- Lead (Pb)—If the fuel contains lead
- Carbon Monoxide (CO) and Carbon Dioxide (CO_2)—During the combustion of fuel
- Hydrocarbons (HC)—During the combustion and evaporation of fuel
- PM (PM_{10}/$PM_{2.5}$)—In incomplete combustion or abrasion and resuspension
- Sulphur Dioxide (SO_2)—During the combustion of sulfurous fuel
- Nitrogen Oxide (NO_x)—During combustion process, especially from diesel engines.

There exist further emissions in road traffic, e.g., heavy metals, however, their effects on humans and the environment have yet to be sufficiently researched.

The following is a list of the main types of emissions, divided by point of origin:

- **Hot/Warm**

  These are emissions (CO, CO_2, NO_x, PM, HC, etc.) released when the engine and the exhaust reduction systems are hot, i.e., have reached their ideal operating temperature. Furthermore, hot emissions depend on traffic situations. Except for PM, hot/warm emissions have the highest share of total emissions on national inventories. Hot/warm emissions are given in units of [g/km].

- **Cold Start**

  When a vehicle has been parked for a minimum of 12 hours before the start of the engine, engine and exhaust reduction systems have reached cold start/ambient temperature; these technical systems are not at an ideal operating temperature. Upon start, an optimal operating temperature is first attained after a few kilometers. This produces cold-start supplement emissions, especially carbon monoxide (CO) and hydrocarbons (HC). Cold-start emissions are given as excess emissions (in [g/start]). A cool start arises in standstill times between 0.5 and 12 hours.

- **Abrasion and Resuspension**

  Abrasion from tires, brakes, clutch, and road surfaces are sources for non-exhaust particles (PMs). In addition, resuspension of generated and deposited particles from road surfaces occurs as a result of vehicle-induced airflow and wind. Technical improvements have reduced engine-based particles; thus, the emissions from abrasion and resuspension can exceed exhaust emissions.

- **Evaporation**

  Evaporative emissions in the form of HCs can be distinguished as follows:

  - Hot/warm soak emissions, given in [g/stop]—These emissions occur after switching off the engine when the engine is still hot or warm.
  - Diurnal evaporative emissions per vehicle given in [g/day].
  - Running losses—Generated as a result of vapor in the fuel tank during vehicle operation, given in [g/km].

Modeling of Road-Traffic Emissions

Overview of Approaches

In general, there are two different approaches for calculating road-traffic emissions for air pollutants and greenhouse gases: the top-down approach and the bottom-up approach. The selection of the suitable approach depends on the acceptable tolerance limit.

In the top-down approach, emissions are calculated based on "fuel sold" or "fuel used" (not including fuel exports to other countries) (European Environment Agency 2016). This approach is a simple means for calculating the (national) total quantity of emissions and is well suited for large-scale areas (in which the volume of fuel sold/fuel used within the country is known and inventoried), without, however, the use of temporal or spatial differentiation parameters or vehicle data. Moreover, the fuel-based approach is exact for CO_2 emissions (calculated from fuel quantity) but only approximate for some air pollutants.

All emission types for specific street sections as well as for partial or complete road networks on the meso- and microlevels can be calculated only with a bottom-up approach for all temporal and spatial differentiations on the basis of vehicle data, i.e., by means of (agent-based) traffic and emission models or by modeled vehicles. These two bottom-up approaches try to imitate or model driving behavior. Therefore, all the following remarks relate to the bottom-up approaches based on vehicle data only.

The bottom-up quantification of emissions can be implemented in different degrees. Potential spatial and temporal resolutions depend on the detail level of (available) input variables. Ericsson (2000) identifies cause–effect relationships (partly also interacting with each other) that influence variability of driving patterns: driver factors, vehicle factors, weather factors, traffic factors, street environment factors, and travel behavior factors (see Figure 2). Figure 2 allows a more detailed perspective into the top boxes of Figure 1 that can be used in a bottom-up approach.

Fontaras et al. (2017) add to Figure 2 driving style and vehicle maintenance (for driver factors), wind (for weather factors), aerodynamic and rolling resistance and auxiliary systems (for vehicle factors), road grade (for street environment factors), and none for travel behavior factors and traffic factors.

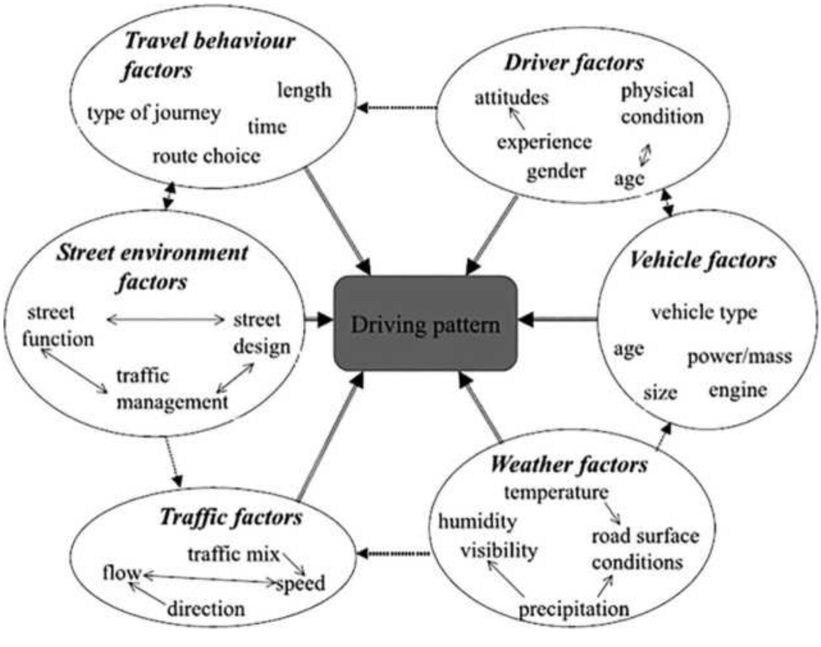


FIGURE 2 Cause–effect model of variability in driving patterns. [According to Ericsson (2000).]

Two technical possibilities for measuring emissions and one for modeling emissions are known. Emissions can be measured as a function of time or distance (known as emission factors):

- Either by a Portable Emission Measurement System (PEMS) at the exhaust of a test car that moves in traffic (exact, but costly)
- Or emissions can be determined in the laboratory with engine test facilities or chassis dynamometers. The necessary engine power to overcome driving resistances and the engine speed (according to a gearshift model) generate emission maps which allow for the calculation of emissions.

Emission models provide emission factors representative of vehicle classes/layers under specific driving situations. The model approach is ideal for the emission determination of large traffic networks, as it is less complex than individual measurements on each street. A traffic-supply model, a traffic-demand model, and an emission model are coupled. Data components for each model are discussed in the following sections.

Traffic-Supply Model

A traffic-supply model includes all infrastructure data of the road network (e.g., length of road sections, number of lanes, speed limits, gradients, traffic lights, and road surfaces). The result is a geo-referenced static traffic model allowing for the mapping of emissions back to their place of origin.

Traffic-Demand Model

The factors for the specification of driving dynamics (see Figure 2) are bundled in a traffic (demand) model. Traffic volume, traffic mileage (traffic volume multiplied by road section length), traffic flow, traffic composition, and traffic quality all serve as input variables for the energy demand of vehicles and emissions. Traffic volume is modeled either as Average Annual Daily Traffic (AADT) based on an average day of the year or in a higher temporal resolution with traffic flow by means of time series of a day, a week, or a year. Traffic flow is the number of vehicles which cross a specific road section over a certain period of time.

Traffic factors have to be quantified by indicators to make them useful for emission calculations in the study area. Ericsson (2000) identifies three parameter categories, which represent changes in driving patterns and have a high impact on fuel use and emissions:

- Level Indicators: Level measures describe, e.g., (average) speed, vehicle kilometers traveled, and number of vehicles. However, no information is given about variation or frequency of variation.
- Oscillation Indicators: Oscillation measures describe frequency, e.g., relative positive acceleration (RPA). RPA is defined as (see Eq. 1)

$$RPA = 1/x \int v * a^{+} \, dt \tag{1}$$

where x = total distance, v = speed, a^{+} = (positive) acceleration, and t = time.

- Distribution Indicators: Distribution measures indicate the distribution of a variable in a certain section, e.g., the percentage time share of acceleration ($a > 1.5\,\text{m/s}^2$).

According to Ericsson (2000), the five most important parameters for fuel consumption are

- Relative positive acceleration
- Number of accelerations per 100 m
- Percentage of time with $a > 1.5\,\text{m/s}^2$
- Percentage of time with $-2.5\,\text{m/s}^2 < a < -1.5\,\text{m/s}^2$
- Percentage of time with $v < 15$ km/h.

Data sources for quantifying traffic factors can be traffic detectors, radar, video detection, Floating car data (FCD), short-time counting, traffic messages, and data from police and traffic models with origin–destination matrices and assignment processes.

Traffic-Emission Model Types

The literature encompasses two general bottom-up approaches for calculating vehicle emissions by means of a traffic model: the macroscopic modeling approach with aggregated network parameters, which has rather low accuracy but enables faster computation of emission estimates, and the microscopic modeling approach with instantaneous vehicle emission rates, using either vehicle engine or vehicle speed/acceleration data (Jiang et al. 2018). According to Smit et al. (2010), there are five different types of traffic-emission models (see Table 2) with different input data and different interfaces between traffic volume and emission factors.

The first step for creating the model is to allocate road infrastructure data and traffic activity data into different categories (e.g., point sources, line sources, area sources) and into different temporal resolutions which refer back to each road section by means of geographical (interpolation) approaches and by assignment processes in traffic models. The second step is to select the emission factors of the model type that correspond as accurately as possible to the merged data of the road section. Subsequently, activity data (A) per road section are multiplied with specific emission factors (EF) per vehicle category, where E = emission (see Eq. 2):

$$E = A * EF \tag{2}$$

The total amount of emissions of a road network results in the sum of all mentioned vehicle emissions that are released on the road sections.

An emission model implies classed emission factors per substance, country, and year for all vehicle categories. These vehicle categories exist for the bottom-up approach (VDI 2019):

- Cars
- Light duty vehicles (≤ 3.5 tons permissible total weight, e.g. small buses, camper vans)
- Heavy duty vehicles (> 3.5 tons permissible total weight, e.g., trucks, truck trains, semitrailer trucks)
- Buses (regular buses, coaches)
- Motorized two-wheelers (motorcycles and mopeds)
- Per vehicle layer/segment: Vehicle category with the same propulsion technique and same size (mass or engine capacity) with the same emission performance to make emission calculation more manageable.

TABLE 2 Different Types of Traffic-Emission Models

Model Type	Input Data	Source	Example Models
Average speed models	Mean traveling speed	Traffic models or field measurements	COPERT, MOBILE or EMFAC
Traffic situation models	Vehicle kilometers traveled per traffic situation	Traffic models	ARTEMIS or HBEFA
Traffic variable models	Traffic-flow variables	Macroscopic and microscopic traffic models	TEE or Matzoros
Cycle variable models	Function of driving cycle variables at high resolution (seconds to minutes)	Obtained from vehicle movements (equipped with GPS) or microscopic traffic models	MEASURE or VERSIT+
Modal models	Function of driving cycle variables	Obtained via engine or vehicle operating models at the highest resolution (one to several seconds)	PHEM, CMEM, or VeTESS

Source: Own Representation Based on Smit, Ntziachristos, and Boulter (2010).

Vehicle-specific components depend predominantly on the following:

- Vehicle size (weight and motorization)
- Drive concept (induced engine [spark ignition engine], or autoignition [diesel engine], natural gas, hybrid)
- In-Engine technology-based reduction measures (e.g., exhaust gas recirculation [EGR]) or exhaust gas after-treatment systems (particulate traps, selective catalytic reduction [SCR], lean NO_x trap, three-way catalytic converter)
- Gear ratio
- Air resistance
- Rolling resistance of tires
- Vehicle conditions (age, mileage, maintenance).

Operational-specific components depend predominantly on the following:

- Engine and exhaust gas temperature
- Gear selection
- Operation of accessories (e.g., air conditioner, alternator)
- Cargo load
- Fuel quality
- Gradient of road
- Ambient temperature
- Height above sea level
- Driving behavior.

To manage the multiple variability of emission calculations, vehicles with similar technical attributes are summarized into the abovementioned vehicle categories with similar consumption and emission behavior, and then, emission factors are attached. The lowest level of detail are vehicle layers. Each vehicle is considered homogenous with respect to its emission characteristics and reasonably distinct from any other layer. Layer criteria are the abovementioned vehicle-specific components. The share of each vehicle layer in the vehicle category changes over time and place due to fleet modernization; therefore, the emission factors are updated yearly for the whole vehicle fleet. Traffic-mileage-related emission calculations need fleet compositions which are weighted accordingly, as average yearly traffic mileage may vary within the vehicle layers (e.g., cars with diesel engines have a higher traffic mileage than cars with gasoline engines).

The HBEFA Traffic Situation Model

The Handbook of Emission Factors for Road Transport (HBEFA) provides a traffic situation model which can be used for large-scale emission calculations while also taking driving dynamics into consideration.

HBEFA maps a comprehensive range of real driving specifications through 1190 classed driving cycles, simulated with the emission model Passenger Car and Heavy Duty Emission Model (PHEM). Kinematic parameters (velocity, percentage of standstill time, and RPA) are considered by traffic situations. They are differentiated by

- Area (urban area/rural area)
- Road type (motorway, arterial road, collecting roads, etc.)
- Speed limit
- Traffic conditions/levels of service (LOS): Fluent, dense, saturated, stop and go, etc.

HBEFA model generates emission factors for all current vehicle categories and concepts for different countries (Germany, Austria, Switzerland, Sweden, France, and Norway).

Figure 3 shows NOx emission factors for a HBEFA traffic situation with examples given for different vehicle categories, such as diesel cars, gasoline cars, light duty vehicles, heavy goods vehicles, and

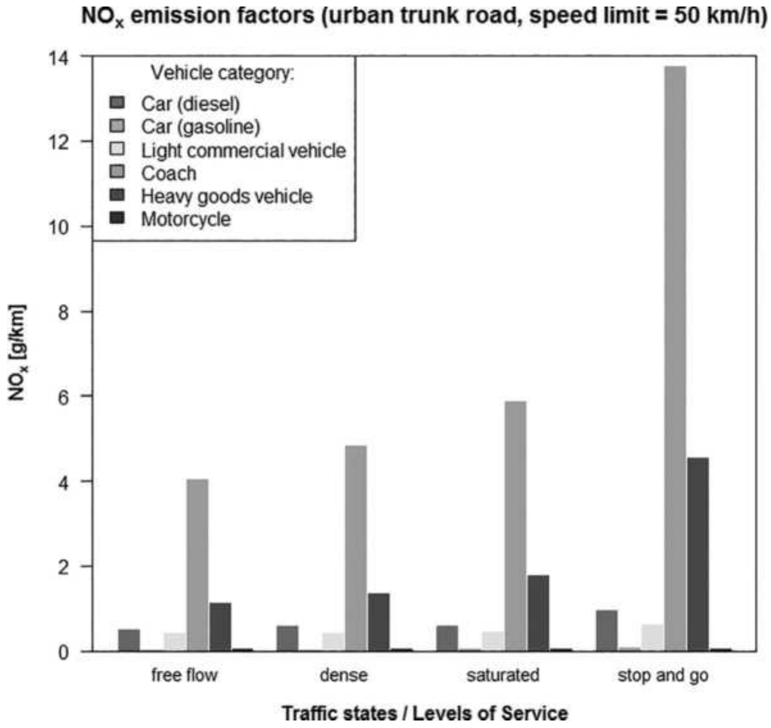


FIGURE 3 Emission factors produced by the HBEFA traffic situation model (urban area/speed limit = 50 km/h on a trunk road) for an average German car in 2019. (INFRAS 2017.)

motorcycles (all from the 2019 German vehicle fleet; ordered from left to right in figure 3), on an urban trunk road with speed limit = 50 km/h without gradient for different LOS (INFRAS 2017).

Validation of Emissions

According to Smit et al. (2010), the calculated emissions can be validated at different spatial scales (local, road, journey, area) by means of six different methods:

1. In *laboratory measurements*, driving conditions can be imitated under controlled parameters
2. *Onboard measurements* are able to test real driving conditions of a vehicle.
3. *Tunnel validation measures* at the tunnel entrance and exit, linked with tunnel features, e.g., tunnel airflow, cross-sectional areas, road length, and traffic density, measure emissions from a large sample of the on-road fleet.
4. Compared to the tunnel validation with relatively "controlled" conditions, *ambient concentration measurements* require results from combined emission and dispersion modeling (including background concentrations from households, industry, agriculture, and natural resources).
5. In the *remote sensing method*, an infrared/ultraviolet beam is directed from a pylon across the road to measure instantaneous ratios of various pollutants of a large sample of the fleet.
6. *Ambient mass balance studies* specify emission fluxes (kg/h) through the measurement of pollutant concentrations upwind and downwind of certain areas at different heights to compare these data with area-wide emission prognoses by the model for the same period.

In addition to the above methods, Poehler et al. (2019) use the *plume chasing method* to measure real vehicle emissions of a vehicle with a testing vehicle driven behind. The ratio of a pollutant (e.g., NO$_x$) to

CO_2 is independent from the dilution of the exhaust plume after correction of the background concentration. The ratio enables the calculation of the emission values of the vehicle.

Potential Influences on Emission Reduction and Their Impacts

As mentioned in "Energy Consumption and Types of Emissions" section, traffic demand and the resulting environmental impacts are embedded in complex interdisciplinary relationships. Therefore, systematic emission–reduction measurements, preferably implemented as pro-active prevention approaches, are needed—for example, by covering the dimensions of the four "E's". These include

- Engineering (e.g., technical reduction possibilities, traffic planning measures)
- Encouragement/Economy (e.g., pricing measures, incentives to change traffic behavior)
- Education (e.g., pedagogical and communicative measures, such as the promotion of environmentally-friendly modes)
- Enforcement (e.g., legislative measures, control, and monitoring).

The four "E's" merge in a bundle of vehicle-related, infrastructural, traffic-planning-related, traffic-management-related, and user-related measures, in which the impacts increase as more measures are implemented together (Schlag 1998).

Vehicle-Related Measures

Undesirable secondary outcomes emerge from combustion processes while driving. Their amount can be reduced by using high-quality fuel and in-engine technology as well as through the utilization of five different after-treatment systems of exhaust gases, as follows:

1. A *three-way catalytic converter* (for suction pipe gasoline engines) oxidizes the air pollutants carbon monoxide (CO), hydrocarbons (HC), and nitrogen oxides (NO_x) to CO_2 and inert N_2, if the exhaust temperature is sufficiently high.
2. An *oxidation catalytic converter* (for diesel engines and gasoline engines with direct injection) oxidizes HC, CO, and soluble particles.
3. *SCR* reduces nitrogen oxides with ammoniac (NH_3)—based on an aqueous urea solution—to nitrogen (N_2) and water (H_2O).
4. In *particulate reduction systems*, the exhaust gases flow over porous surfaces in open or closed filter systems.
5. *EGR* mixes exhaust gases back to the combustion chamber to reduce thermic NO_x (Pfäfflin 2018).

Infrastructural, Traffic Planning, Traffic Management-Related Measures

Traffic-planning measures—often coupled with traffic-management measures—reveal steering effects on traffic flows, which usually include, in addition to primary local (desired) impacts, secondary (unwanted) effects on the transportation system (e.g., induced traffic[2] [increasing traffic], spatial shift of traffic flows, etc.). Thus, traffic-planning measures are typically valued on a case-by-case basis. Often, the effectiveness of a planning measure is also dependent on framework conditions and the type of chosen implementation. Pfäfflin (2018) has outlined seven typical planning measures as follows:

1. *Measures of improvement of traffic efficiency/decongestion* include, for example, capacity expansions or construction of roundabouts (consider maximum thresholds of traffic volume) from an infrastructure perspective. From a traffic management perspective, traffic control centers can implement environmentally oriented traffic management based on the following:

---

[2] Hills (1996) explains the usage of this term in the context of transportation.

- Traffic models for an interpolation or prognosis of traffic scenarios and strategies
- Data archives to recognize traffic behavior patterns and causal relations
- Strategy management systems to run predefined strategies.

Traffic control centers use Intelligent Transportation Systems (ITS); these include sensor systems for traffic monitoring, devices for transfer systems for traffic incidents (e.g., road works, accidents, events), devices to control traffic flow (e.g., [virtual] variable message signs), and further information devices for transportation users. Further traffic management scenarios are optimization of traffic lights (green waves), ramp metering, dedicated lanes for vehicle with high occupancy (HOV lane), or high-occupancy toll lanes (HOT lanes), in which high-occupancy vehicles are without charge, and other vehicles are required to pay a variable, demand-adjusted fee.

All these measures lead—under the assumption of same boundary conditions—to an improvement in local traffic quality with less stop and go and, thus, less emissions; however, rebound effects must be considered. Lower travel times induce traffic (either from changed-route choice or from modal shift from public transport) which may lead to an increase in the total emissions in the network than before the implementation of the traffic planning measure.

2. *Construction of new bypass roads* leads to traffic shifts and/or induced traffic with a shift or increase of emissions. Dismantling of main through-roads reduces the risk of the abovementioned rebound effects.
3. *Access and transit prohibitions* restrict temporarily or permanently an entry of certain vehicle categories or vehicle groups to specified roads or areas, e.g., implemented as environmental zones or city-center tolls. An environmental zone can lead to an acceleration of vehicle-fleet modernization with less specific emissions—also outside of the environmental zone. In contrast, traffic shifts and a rise in vehicle kilometers traveled (due to longer trips to detour around the environmental zone) can lead to increased emissions.
4. *Parking management* includes the dismantling or creation of parking spaces as well as the establishment of parking guidance systems. Parking guidance systems reduce the emissions created by traffic in search of parking; yet, at the same time, the attractiveness of parking spaces has the counter effect of induced traffic (increasing traffic). Consistent parking-space management zones of sufficient size with homogenous prices will reduce variable attractiveness between parking options and, thus, avoid extended searches for parking.
5. The *reduction of permissible maximum speed* is a measure that does not directly affect emissions but rather driving behavior. A maximal reduction in emissions can be seen in the mid-speed range of 60–80 km/h. Emissions increase at faster speeds due to higher engine power demand; they also increase at speeds below 50 km/h as a result of external factors such as traffic control, traffic signal settings, and gradient. Therefore, an overarching statement concerning permissible maximum speed is not possible.
6. The *attraction of environmentally-friendly modes* includes modes of transportation with zero or less emissions than motorized traffic, e.g., walking, cycling, public transportation, or trains. Measures that increase attractiveness of these modes include a safe infrastructure for non-motorized traffic, an increase in network density and frequency, barrier-free vehicles, public relations (information and communication, marketing, consultations), educational measures (e.g., biking safety classes), carsharing, ridesharing, and intermodal mobility stations (mobility as a service). For the most part, reducing environmental impacts can only be achieved when pull measures are combined with push measures of motorized transportation (reduction of parking spaces, increase of parking prices, etc.).
7. *Measures without secondary effects to traffic quality and traffic volumes* include construction of noise-protection barriers for lowering emissions, planting vegetation, or installing photocatalytic surface coatings for the reduction of NO_2 concentrations. The reduction potential is very low compared to other measures.

User-Related Measures: Internalizing External Costs of Environmental Impact, Mobility Management, and Land Use Planning

The abovementioned traffic-planning measures have a predominantly local or regional impact on transportation. International policies must set incentives for all emission sectors to develop and implement a common cooperation strategy for reducing and eliminating emissions. This can be achieved through the realization of a climate-friendly legal framework which follows the *true-cost principle* (i.e., usage-related charges with no external effects on other regions, people, or time frames) combined with municipal and operational *mobility management* (a target-oriented approach for influencing individual mobility behavior by perceiving, revealing, and assessing mobility options for traffic prevention). A climate-friendly transportation policy and legislation also encourages the development of livable, green, and healthy cities / (small) towns composed of the following:

- Diversity: Buildings and public spaces serve multiple purposes, such as for employment, shopping, or recreation.
- Density: Critical for short distances and the (public) transit system, density concentrates trip origins or destinations, allowing for ride sharing to become practical as well as economical.
- Design: This provides an accessible, barrier-free, safe, and rewarding infrastructure with an easy-to-understand orientation system. Streets would not only have link functions (which minimize travel times) but also have place functions (which maximize the time traffic participants spend for social interaction in public spaces) (Cervero and Kockelman 1997).

A successive realization of the true-cost principle in transport with social compensation measures leads to traffic behavior changes, sustainability, and a shift toward eco-conscious transportation.

Conclusion

Transport emissions are a significant contributor to total emissions worldwide; in particular, road-traffic emissions have been increasing, and this trend will continue in the future. The substances PM_{10} and NO_x (NO and NO_2) present the greatest problems in transport, with the resultant CO_2 contributing directly to global warming. Fuel-based vehicles are the majority, but electric vehicles gain market shares. Emissions arise from combustion processes, abrasion, resuspension, and evaporation. Either the emissions can be measured with test vehicles or the emissions can be modeled with traffic and emission models for large networks. Several methods exist on different spatial scales to validate emissions. Emissions can be reduced most effectively with an integrated approach consisting of vehicle-related, infrastructural, traffic-planning-related, traffic-management-related, and user-related measures. (Rebound effects must also be considered.) Traffic and emission models can be used for environmentally oriented traffic management. A realization of the true-cost principle will support a shift toward eco-conscious transportation.

References

Cervero, Robert, and Kara Kockelman. 1997. “Travel demand and the 3Ds: Density, diversity, and design.” *Transportation Research Part D*, 9: 199–219.

DIRECTIVE 2007/46/EC. 2007. DIRECTIVE 2007/46/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 5 September 2007 establishing a framework for the approval of motor vehicles and their trailers, and of systems, components and separate technical units intended for such vehicles. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32007L0046&qid=1561907271004&from=EN.

DIRECTIVE 2008/50/EC. 2008. DIRECTIVE 2008/50/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 21 May 2008 on ambient air quality and cleaner air for Europe. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32008L0050&from=EN.

Ericsson, Eva. 2000. "Variability in urban driving patterns." *Transportation Research Part D*, 5: 337–354.

European Environment Agency. 2016. EMEP/EEA air pollutant emission inventory guidebook 2016 – Technical guidance to prepare national emission inventories, Vol. 21.

Fontaras, Georgios, Nikiforos-Georgios Zacharof, and Biagio Ciuffo. 2017. "Fuel consumption and CO2 emissions from passenger cars in Europe – Laboratory versus real-world emissions." *Progress in Energy and Combustion Science*, 60: 97–131.

Hills, Peter J. 1996. 'What is induced traffic?" *Transportation*, 2: 5–16. https://link.springer.com/article/10.1007/BF00166216.

Hooftman, Nils, Maarten Messagie, Joeri Van Mierlo, and Thierry Coosemans. 2018. "A review of the European passenger car regulations – Real driving emissions vs local air quality." *Renewable and Sustainable Energy Reviews*, 86: 1–21. doi:10.1016/j.rser.2018.01.012.

INFRAS. 2017. Handbook of Emission Factors for Road Transport, Version 3.3. www.hbefa.net.

International Energy Agency. 2018. CO2 emissions from fuel combustion – Highlights. Comp. Fatih Birol.

International Energy Agency. 2017. Energy Technology Perspectives 2017 – Catalysing Energy Technology Transformations. Comp. Fatih Briol.

International Energy Agency. 2018. World Energy Outlook 2018. Comp. Fatih Birol.

Jiang, Yan-Qun, Pei-Jie Ma, and Shu-Guang Zhou. 2018. "Macroscopic modeling approach to estimate traffic-related emissions in urban areas." *Transporation Research Part D*, 5: 41–55.

Pfäfflin, Florian. 2018. Hinweise zu Energie, luftbezogenen Emissionen und Immissionen im Straßenverkehr H EEIS. Translation of Title: Information on energy, air-related emissions and pollution in road traffic. Arbeitsgruppe Straßenentwurf (Working Group Road Design). Cologne, Germany: Road and Transportation Research Association (FGSV).

Poehler, Denis, Tobias Engel, Uli Roth, Joscha Reber, Martin Horbanski, Johannes Lampel, and Ulrich Platt. 2019. "Messung realer Fahrzeugemissionen mit dem 'Plume Chasing'-Verfahren [Measurement of Real Vehicle Emissions using the 'Plume Chasing' Method]." *Kolloquium Luftqualitaet an Straßen* 2019 (*Colloquium Air Quality at Roads* 2019). Bergisch Gladbach, Germany: Federal Highway Research Institute (bast), 63–78.

REGULATION No 333/2014/EU. 2014. REGULATION (EU) No 333/2014 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 11 March 2014 amending Regulation (EC) No 443/2009 to define the modalities for reaching the 2020 target to reduce CO2 emissions from new passenger cars. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32014R0333&from=EN.

REGULATION No 715/2007/EC. 2007. REGULATION (EC) OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 20 June 2007 on type approval of motor vehicles with respect to emissions from light passenger and commercial vehicles (Euro 5+6) and on access to vehicle repair and maintenance information.

Schlag, Bernhard. 1998. "Zur Akzeptanz von Straßenbenutzungsgebühren, Translation of Title: Acceptance of road pricing." *Internationales Verkehrswesen*, 50 (7/8): 308–312.

Smit, Robin, Leonidas Ntziachristos, and Paul Boulter. 2010. "Validation of road vehicle and traffic emission models – A review and meta-analysis." *Atmospheric Environment*, 44: 2943–2953.

VDI. 2019. "Umweltmeteorologie – Kfz-Emissionsbestimmung – Luftbeimengungen [Environmental meteorology – Determination of the emission from motor vehicles – Air quality]." VDI-Richtlinien (VDI Guidelines), Fachbereich Umweltmeteorologie (Department Environmental Meteorology), Verein Deutscher Ingenieure (The Association of German Engineers).

WHO. 2006. WHO Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide – Global update 2005 – Summary of risk assessment. World Health Organization.

II

COV: Comparative Overviews of Important Topics for Environmental Management

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

5

Alternative Energy

Bernd Markert,
Simone
Wuenschmann,
Stefan Fraenzle,
and Bernd
Delakowitz

Introduction

Regenerative resources like wind and solar energies, running and falling water (hydropower), biomass processed in some way and geothermic energy are in a position to replace crude oil, natural gas, coal, and uranium step by step depending on both technological innovations and political decisions. Here, technological innovations are not meant to require fundamental inventions to render some renewable resource useful for energy (electric current) "harvesting." Rather, this is about making existing technologies cheaper and overcoming specific material problems, like device corrosion and fast "blocking" of underground heat exchange pathways with geothermic energy, storage in an easy-to-handle form (methanol?) for both solar energy and biomass, or using less toxic and less brittle semiconductors in much thinner layers in photovoltaics [thin-film cells based on either copper indium dichalcogenides (chalcopyrites) or organic semiconductors].

Given the relativeness of time and the notorious "difficulty to make predictions which refer to the future," what does it say about classical fossil resources running out in the foreseeable future? Putting this into proper context means to distinguish between "reserve" and "resource": "reserve" just encompasses those deposits of energy carriers that are actually known and can be really accessed following both technical and economic criteria. In the pre-1973 world of %2.70 per barrel of crude oil (some $19 per ton), it would have been considered a fancy to try to extract oil from shales or sand and drill deep below the seafloor (now always performed in the Atlantic Ocean and the Mexican Gulf, off Brazil and Angola) or in remote arctic regions. Currently, the meaning of underground hard coal production or making access to very deep natural gas deposits is doubtful. While oil production from oil sands (Alberta Province, Canada) is now economically viable—but still an ecological disaster in a sensitive surrounding—other methods of accessing certain fossil resources will probably never be viable: it simply takes more energy to extract and process traces of ^{235}U dissolved in seawater than can be obtained from its fission afterwards. As the term "reserves" by definition ("share of total potential which can be mined and exploited economically reasonable by currently available technical means") includes both the present level of technology and current pricing, reserves are subject to changes other than due to ongoing prospection and exploitation. For example, when disregarding the environmental and safety issues associated with either kind of fossil energy carrier for this moment, both oil sands and shale gas are close to the lower limit of economic feasibility given the current crude and natural gas prices. However, available amounts are not settled with shale gas, not permitting to include it into the reserves, while oil sands—although a blueprint for ecological disaster—are economically and technically feasible to produce and hence are part of (Canada's and global) crude reserves. Matters are different once again if the technology to actually obtain energy from some fossil or far-spread reservoir around is not yet at hand or it is doubtful whether energy required to mine and concentrate that particular source might even exceed the energetic payback obtained thereafter: consider amounts/concentrations of deuterium, ^{6}Li, and ^{235}U in seawater. The former two (^{6}Li being a precursor to fusion "fuel" tritium) would combine to an inexhaustible source of energy if "only" net-energy-yield nuclear fusion would be at hand already, while there are just some 15 ng (!) of ^{235}U that can be actually extracted from 1 L of ocean water, providing a few hundred joules ($\approx 10^{-4}$ kWh) of electric energy in nuclear (now, fission) power plants thereafter. Even though the corresponding extraction was already demonstrated in Japan, using ion-exchange resins, the energy for producing the resins and running the device is so large that ocean-derived uranium is not a viable resource either and thus cannot be counted among the reserves.

"Resources," on the other hand, are those energy carriers either already discovered (not a single "elephant field" of crude oil was spotted after the 1960s anywhere in the world!) or *reasonably believed* to exist from geological arguments in Earth's crust, but cannot be exploited right now for either technological or economic reasons. Due to the difficulties and risks associated with drilling either below some 9 km underground or into active magma regions, most of the huge amounts of geothermic energy can never be actually used. Hence, for fossil resources as well, we are left with what is in the crust or the ocean (floor). Given this distinction, the residual economic lifetimes for traditional forms of energy carriers such as oil, natural gas, brown and hard coal, or uranium given in Figure 1 are obtained.

Energy Depletion of Fossil Fuels

Counting from year 2000 onwards, the German Federal Institute for Geosciences and Natural Resources estimated in 2002 that reserves of oil, natural, gas and uranium will last for just another 40 to 65 years. Reserves of both hard and brown coal will still last up to 200 years, whereas the resources of coal and gas and uranium are to last more than 200 years. Things are more critical with crude oil, reserves of which will be gone within 60 years while resources are estimated to last for 160 years at best. Data and predictions in Figure 1 do not cover and include the energy consumption of current growth regions [BRICS states like Brazil, Russia, India, China (PR), and other Latin American countries], leaving us with the

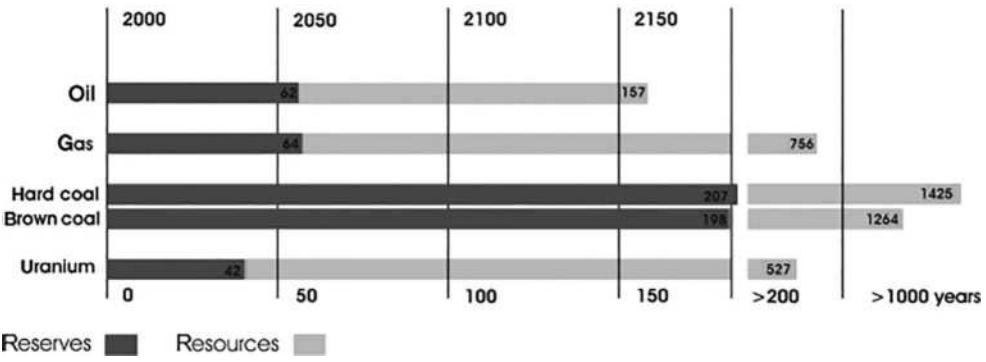


FIGURE 1 Predicted reserves/resources of fossil energy carriers from year 2000 onwards.
Source: German Federal Institute for Geosciences and Natural Resources, 2002.

conclusion that we are left with much less time to change our bases of energy supply altogether.[1–4] The resources of crude oil are much disputed. Apparently, most of the additional stockpiles—beyond established reserves still considered a few years ago—simply do not exist. Speculation on oil term in markets is considerably influenced by this insecurity of affairs while the big oil companies obviously produce an exaggerated picture of resources rather than make speculation go its way.

Another matter is the regional distribution of these reserves/resources all around the globe (Figure 2; Federal Institute for Geosciences and Raw Material Research.[5]): the problem is obvious with crude oil, and everybody is aware of it, but there are also biases/imbalances with hard coal. Considering the total energy stored in it, some 60% of it rests with brown or hard coal; the smaller part is included in liquid and gaseous hydrocarbons. Note that efficiencies of power plants differ considerably among these energy carriers, with hard coal and gas steam plants being superior to others ($\eta_{el} > 55\%$). Crude oil, which may be produced by current technologies, reasonably amounts to 6682 EJ, a little less than with natural gas (7136 EJ). 1000 EJ (exajoule) = 10^{21} J. Standard heats of formation of the compounds/mixture/combustion products involved are as follows:

CH_4	−50 kJ/mol
C (graphite)	zero [by definition (standard state of an element)]
"CH_2" (fraction of crude oil)	≈ −20 kJ/mol
CO_2	−394 kJ/mol
H_2O	−237 kJ/mol

Thus, 1 mol (44 g) of CO_2 produced from combustion of natural gas (CH_4, essentially), crude oil, and coal yields 818 kJ, 611 kJ, and 394 kJ, respectively. The annual global anthropogenic CO_2 output is of the order of gigatons (1015 g, Pg), with the atmosphere containing some 770 Pg of it now, considerably more than which is tied up in living biota (about 610 Pg). One gigaton of CO2 from hard coal, crude oil, and natural gas natural translates into somewhat less than 9, 13.9, and 18.6 EJ of thermal energy, respectively. Thus, actually combusting the above-estimated resources would leave us with some 50,000 gigatons, that is, 65 times (!) the present CO2 inventory of the atmosphere.

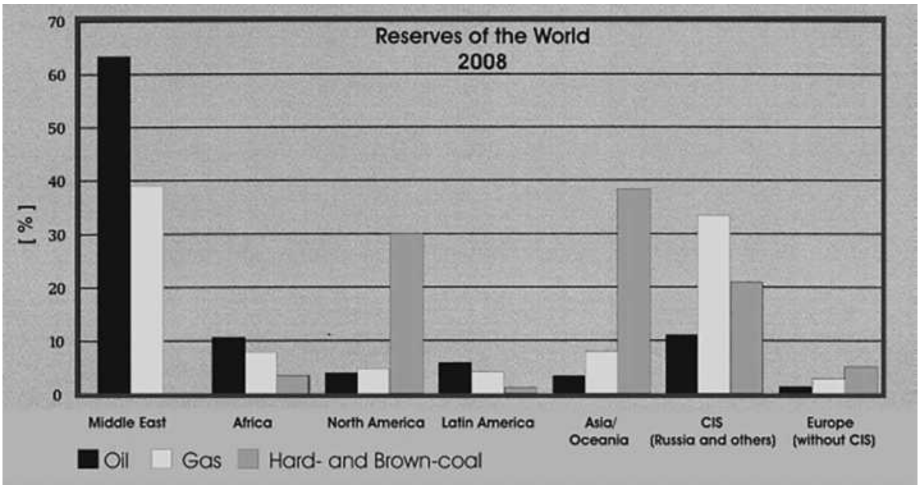


FIGURE 2 Estimate of the regional/continental distribution of oil/gas/coal reserves.[5] As for the Commonwealth of Independent States (former USSR except of Baltic states and Georgia), the biggest share is with just three of them: Russia (the current biggest oil producer in the world), Azerbaijan (both oil and gas), and Turkmenistan (almost gas only).

The total energy from conventional fossil energy carriers (resources combined) would be some 448,289 EJ, 97% of these resources being hard or brown coal. Besides the above-mentioned conventional energy carriers, there are substantial amounts of other energy carriers such as oil sands, oil shales, tars, and other kinds of the heaviest, most condensed oils, natural gas from dense storage sites, carbon deposits (adsorbed), or aquifer waters, plus gas (CH_4) hydrates on deep shelf (below some 350 m of seawater or in the uppermost sediments down there) reserves and resources that correspond to an energy equivalent of 2368 EJ and 116,270 EJ, respectively.[6]

Most industrial activities now depend on oil, as do almost all traffic systems, be it airplanes, ships, or cars. Oil getting scarce thus does not only cause prices to rise; there will also be ramifications on workforce in petrochemical branches and political implications. The largest share of oil (2005), some 742 billion barrels, is located in Central and Southern Asia (Bangladesh, Bhutan, India, Maldives, Nepal, Pakistan, Afghanistan, Sri Lanka, and parts of Iran).

Considering the global oil reserves according to British Petrol in 2005, this translates into 62% of global stockpiles, whereas North America commands just 5%, being one of the meta-regions scarcest in oil besides Asia/Pacific and Europe.

With oil getting scarce, a one-sided economic dependence on the Middle East poses increasing political risks, causing everybody to consider oil and gas resources located elsewhere and how they and additional geological goods might be obtained. Thus, there is a recent growing geopolitical interest in the Arctic region, which will become void of drifting ice during the next 20 or 30 years due to climate change. According to the U.S. Geological Survey Institute,[7] some 30% of natural gas and 13% of crude oil are located there. However, most of these are located far offshore [the remainder already exploited for decades in Russia (Taimyr Peninsula) and Alaska], with most of the gas belonging to the Russian Federation while oil is scattered among Canada, Alaska, and Greenland (still partly governed by Denmark). Though very large, experts say[7] that these deposits are not enough to consider relocation of most production activities from the Middle East. In addition, long transport distances (thus, costs) and still adverse weather and climate conditions pose grave problems. Everybody is still aware of what can happen with deep sea-based oil production considering the Deepwater Horizon catastrophe of April 2010, environmental concerns translating into larger political obstacles, higher insurance fees, and eventually less consumer acceptance.

That the so-called peak-oil level where global oil production has reached its maximum ever was already achieved in the beginning of the 21st century is evident from a depiction of development of oil production from 1930 until (predicted) 2050 (Figure 3): scarce oil means there will be no more cheap oil. Until the beginning of the 21st century, one barrel of oil commanded between $20 and a maximum

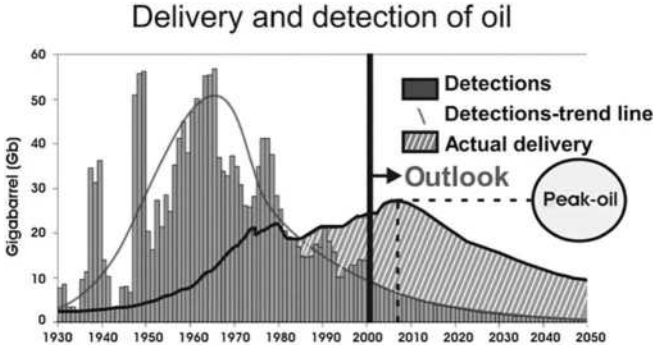


FIGURE 3 Oil findings and delivery rates from 1930 to 2050 with the outlook after Campbell (2006).[8] The delivery rate per year is calculated by the Association for Peak-Oil and Gas Studies, Germany. Data from ExxonMobil (2002). Peak oil: point of time when the maximum rate of global petroleum extraction is reached.
Source: Figure modified from Blum (2005).[9]

of \$40. The financial crisis made it (Brent) rise up to \$147 after 2007. Declining somewhat, it now (2011) stabilizes around \$100. In the long term, it is more likely to increase again, having severe drawbacks on economic conditions in industrialized countries especially.

Figure 1 tells us that hard and brown coal reserves are to last another about 200 years, rendering coal-fired power plants most attractive if it were not for coal being one of the most polluting sources of energy. Combustion of coal produces plenty of CO2, and therefore, it contributes to the anthropogenic part of greenhouse effect warming of our atmosphere.

Climate Protection

CO_2 is one of the most prominent greenhouse gases in the atmosphere, contributing to heating the atmosphere and thus the Earth's surface.[10,11] There is both a natural and an anthropogenic (share of) greenhouse effect. CO_2, produced by animal respiration, wildfires, and volcanoes, is a natural (<300 ppm) component of the atmosphere, heating our planet from a radiation equilibrium ("blackbody") value of some −18°C to a global average of +15°C, that is, by a considerable 33°C.

A global increase of average atmospheric temperatures is now seen for decades, with Figure 4 displaying the distribution of this effect for 2000–2009 as compared to the reference time frame of 1951–1980. Satellite measurements by the NASA Earth Observatory showed the largest increases of the average temperature in the Arctic parts of the Northern Hemisphere, besides some parts of Antarctica. Between the beginning and the end of the 20th century, the average increase was 0.74 ± 0.18°C.

Time is imminent for rethinking energy production with a focus on sustainability in all ecological, economic terms and social acceptance. Man-made radiative greenhouse forcing is established for more than 20 years now. It causes enhanced glacier meltdown rates in the Arctic (Greenland actually turns green again), the Alps, and other mountainous regions, as well as sea level rising (both due to meltdown waters and to thermal expansion of warmed surface water), with more frequent droughts causing limnetic waters to evaporate often completely. We also would be urged and obliged to seriously think about what we are doing if, as some scientists still maintain to argue, this would be a

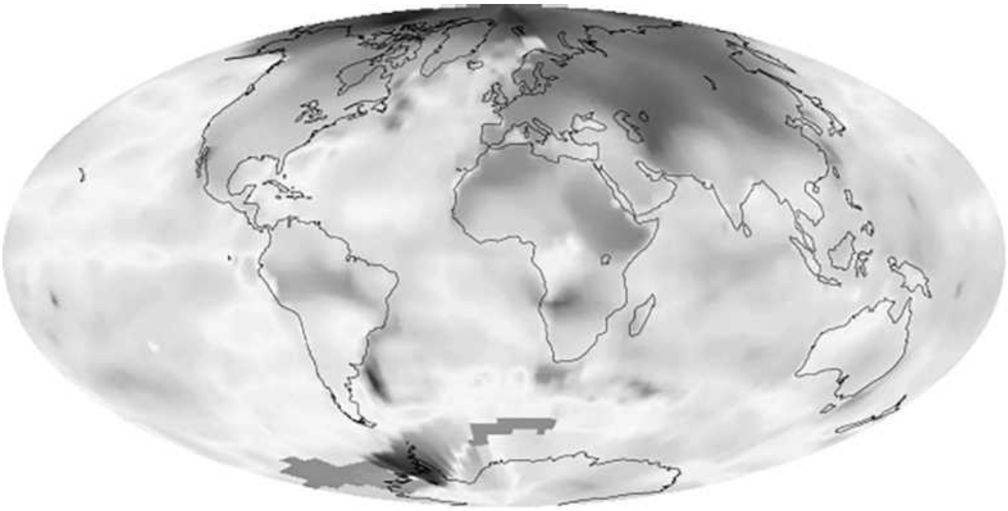


FIGURE 4 Regional distribution of relative greenhouse effect warming. The map shows the 10 years average (2000–2009) global mean temperature anomaly relative to the 1951–1980 mean. The largest temperature increases are in the Arctic and the Antarctic Peninsula. The darker the gray shades in the picture of NASA,[14] the larger the increase in local average temperatures.

Source: NASA Earth observatory.

normal climate excursion as occurred in the later Middle Ages. It does not take much proficiency in the natural sciences to acknowledge the problems associated with the industry emitting greenhouse gases[12] while striving to maintain and enhance our common well-being. The destruction of the environment concomitant with the exploitation of the developing Third World countries has already been an issue some 40 years ago, prompting the first Conference on the Human Environment at Stockholm (Sweden) on June 16, 1972, initiated and organized by the United Nations. Twenty years after, in 1992, the most comprehensive global conference on such issues ever held, the United Nations Conference on Environment and Development (UNCED), took place in Rio de Janeiro. One hundred years ago, Wilhelm Ostwald mentioned in his approach "Die Mühle des Lebens,"[13] "Do not waste energy, rather use it wisely," This is in opposite to the Kantian imperative at this time, and the first energetic step to a so-called sustainable chemistry of today.[15]

A total of 178 countries dedicated themselves to Agenda 21 to oblige the 118 more developed countries to embrace environmental restoration, preservation, and social development. Their aims are to meet the challenge of global warming, pollution, and biodiversity and to solve the interrelated social problems of poverty, health, and population. The Agenda 21 program furthermore encompasses constructing a network of "new and equitable global partnership through the creation of new levels of cooperation among States, key sectors of societies and people, while working towards international agreements which respect the interests of all and protect the integrity of the global environmental and developmental system."

In 1997, the World Climate Summit was held at Kyoto, Japan, focusing on the future climate protection policy and measures. The 158 states that had so far signed and ratified the Framework Convention on Climate Protection and 6 "observer states" had sent a total of almost 2300 delegates, plus another 3900 observers from non-governmental and other international organizations and 3700 journalists from global media, making up a total of close to 10,000.[16] The Kyoto process unleashed by this huge meeting now describes the attempts to maintain climate protection agreements beyond the Kyoto protocol running out in 2012 and to do so in a way that still is safeguarded by agreements of international law.

The industrialized states listed in Appendix 1 of the Kyoto Protocol agreed to reduce their collective greenhouse gas emissions by 5.2% from the 1990 levels by the year 2012. The common (purported) aim of the international community is to limit the increase of global average temperature within 2°C. By now, it has become most doubtful whether this "two-degree aim" can be kept at all. In addition, it was proven that even an aggressive forestation policy could not cope by photosynthetic absorption (assimilation) with the present upsurge of CO_2 produced. The progress made so far is modest at best, with real successes still to be waited for. The subsequent UN Climate Conference at Copenhagen (Denmark) in late 2009 saw just a communiqué of minimal consent without any mandatory aims in CO2 reduction although the "two-degree aim" was once again acknowledged to be worthwhile.

Role of Nuclear Power

The Fukushima (Japan) sequence of catastrophic events began on March 11, 2011 [the strongest earthquake ever recorded in Japan (Richter magnitude 9.1)], causing coolant pipes to break and starting a sequence of reactor core meltdowns in at least three adjacent nuclear power plants (NPPs). The area was then hit by a tsunami; any semblance of control of the preceding events was lost. Finally, the uncooled reactor systems exploded one after the other, releasing large amounts of radioactivity and visibly destroying reactor block 3. People reconsidered the risks of nuclear energy all over the world, although with grossly differing political consequences.

Now introduced into IAEA (International Atomic Energy Agency) accident level VII, the events at Fukushima were set equal to the Chernobyl disaster [April 1986, Ukraine (officially: Ukrainian Soviet Socialist Republic)] and considered worse than the explosion of a nuclear fission waste storage tank at Kyshtym in 1957 [near Chelyabinsk, Urals, then RSFSR (Russian Soviet Federative Socialist Republic), USSR (Union of Soviet Socialist Republic); level VI accident], which delivered very large amounts of

^{90}Sr [several EBq (exabecquerel: 10^{18} Bq)] that an area of some 15,000 km^2 had to be abandoned for any human use up to now. Japan is known to be tectonically most active, prompting the authorities and engineers to consider earthquakes up to Richter magnitude 7.9 in design. Two matters demand further consideration here:

a. What causes, e.g., main coolant pipes to break is acceleration and the mere amplitude of dislocation combined with inertia, that is, parameters rather covered by the "old-fashioned" Mercalli scale of earthquake intensity, with Mercalli 13 denoting accelerations larger than that of the Earth's gravity. Such quakes actually happened, e.g., the Easter quake of 1964 in Central Alaska (there is a very impressive movie showing cars and humans and even some homes losing contact with the ground during the quake). In contrast, Richter scaling gives the energy unleashed by a quake [or a landslide or an underground explosion (be it chemical or nuclear in origin)] in a logarithmic way: Depending on the depth of the epicenter, i.e., the amount of matter located above the vibrating or breaking sample of crustal matter, accelerations due to a given amount of tectonic energy can be quite different: the "fairly moderate" Haiti quake of February 2010 (Richter 7.1) became a real killer not because of its energy but since its epicenter was located just a few kilometers below the surface, producing a massive acceleration, thus aggravating the effects of non-adapted architecture. Inertia also is a matter of dislocation of the ground, which provides a kind of reference frame, physically speaking, to which any device that can vibrate, etc., will respond: this dislocation was about 2.5 m in Japan on March 11, 2011, that is, much larger than in the even stronger (more energetic) Chile 1960 (Richter 9.5) and Sumatra Christmas 2004 (Richter 9.3) earthquakes.
b. At least two NPPs, one of them in Japan, had already experienced earthquakes exceeding these limits of design: at Niigata (NW Japan) and at then Leninakan in Armenian Soviet Socialist Republic (SSR) in 1988. Hence, the design was obviously short of reasonable expectations [Central Europe, German, Swiss, and French NPPs next to the Upper Rhine Rift Valley do not take account of the fact that Basle City was almost completely destroyed by a quake in historic times (in 1356)].

The Fukushima earthquake was far stronger than this. The tragedy was worsened by a tsunami caused by this quake; it took thousands of lives in cities and villages near the coast, but most likely, the crucial damage to the Fukushima NPPs was caused by the earthquake, though the NPP area was inundated soon after. Certain radionuclides, mainly highly volatile ones were released; that core meltdowns had occurred was first contested and then confirmed by Japanese authorities only 2 1/2 mo later. These highly volatile radionuclides include ^{131}I, $^{134;137}$Cs and some noble gases, and traces of ^{132}Te but not Sr, Ba, and rare earth elements. Apparently, the temperatures during core meltdown did not yet suffice to vaporize alkaline earths or rare earth element compounds (cesium becomes highly volatile as a hydroxide, CsOH molecules sublimating as easily as NaOH and KOH do), neither did volatile chlorides form after NaCl from seawater pumped inside to cool the reactor cores solidified. Including those still missing, their whereabouts unaccounted for but unlikely still to be alive, there were some 23,000 fatalities due to the quake and tsunami, and several workers received massive radiation damages during cleanup, with a nuclear catastrophe still looming ahead. The number of people forced to abandon their homes and jobs (permanently) at Fukushima and Iitate provinces [40 km NW Fukushima (the NPP site, not the town of Fukushima)] is more than 80,000 now, that is, close to that of people who became homeless due to the Chernobyl accident. Responses from different countries were different: while politicians and citizens in certain countries started to abandon energetic uses of nuclear fission (Germany and Switzerland) or reaffirmed prior decisions of this kind [New Zealand, referendum in Italy with a 94% (2011) turnout, Belgium], others chose to keep a pronuclear course or even kept up decisions to build national first-ever NPPs (Poland).

On a global scale, the issue of nuclear energy is controversial in all developing countries, classical industrialized states, and "official" and "unofficial" nuclear-arms-possessing states. While the United States plans to commence building the first new NPPs after a moratorium of some third of a century,[17] Italy officially revoked the schedule to abandon nuclear energy adopted by a 1994 referendum (but only so in the old Berlusconi regime, the Monti administration being undecided on this issue), and the spring 2012 presidential election campaign in France triggered a broad discussion on the future of NPPs there for the first time. Several developing states extend their present NPP program apparently as a pretext to obtain more fissionable material for nuclear arms, like India and Iran, regardless of whether they signed the Non-Proliferation treaty (Iran), did not sign (India and Israel), or left the protocol some time later (North Korea). In "established" nuclear powers, on the other hand, NPP extension also is a means to get rid of excess amounts of weapons-grade or "special" nuclear fuels (United States, Russia, U.K.). The most interesting case is some 2 tons of ^{233}U left over from the thorium/uranium breeding cycle, which is/was used only in India for energy production and formerly in the United States for making arms. Now, the United States is left with a stockpile of this nuclide, a nuclide that actually would be usable by very-low-technology groups or states to produce "efficient" and reliable-yield warheads (different from that with any plutonium isotope [-mixture]!) and correspondingly must safeguard it to a level that would not be necessary with any other isotope—unless they "burn" it in reactors.

Hence, the ambiguity of nuclear technologies continues to influence decisions on introducing or maintaining it as an energy source, far beyond, and competing with, considerations of supply reliability (the number of states that provide/export significant amounts of uranium is considerably smaller than those that provide all natural gas, crude oil, and hard coal). Political decisions on increased or renewed use of nuclear energy in certain countries hence cannot be considered a signal that is globally significant, except for their drawbacks to the issue of proliferation.

The classical argument in favor of nuclear energy production is their releasing much less CO_2 as compared to oil, coal, or natural gas power plants, enabling economies to meet Kyoto ends more easily. For certain countries that produce nuclear fuels but command few other natural resources, there is the bonus of apparent independence; this includes Japan. In the EU, the Czech Republic presently is the only state to mine domestic uranium ores, globally large producers being differently organized and reliable in political terms [Australia, Canada, the United States, Kazakhstan, Niger, Gabon (Central Africa)].

Nuclear energy is rather "compact," a GW-class power plant taking a few hectares at most, which is important in countries as densely populated as Japan (or India, or Belgium), with Japan covering a total area hardly larger than that of Germany (372,000 km^2). Next to the United States (104) and France (59), Japan has the largest number of NPPs [53, from which 6 at Fukushima and 3 at another highly earthquake-prone site (scheduled for shutdown) must be subtracted; thus, effectively it is only 44] due to an experience from the 1970s: the OPEC oil crisis in 1973, rapidly increasing the prices of crude, of course hit all the already then industrialized countries except for those producing sufficient fossil fuels of their own (Canada, Australia) or relying on other sources mainly for different reasons (Norway, New Zealand, South Africa) but worst so in Japan: in fact, there were large-scale electricity shutdowns badly hurting Japan's economy. Hence, Japan—once again in its history—became eager for independent supply, planning to build yet more NPPs before Fukushima happened. By now, unlike France, Belgium, the United States, or formerly Lithuania, which obtains 60% or more of its electricity from NPPs, Japan's rate is 30%. During the period of Soviet dominance, Lithuania even had the largest NPP in the world, a 2400 MW$_{el}$ plant at Ignalina (now Visaginas) of the notorious (Chernobyl) RbmK (*reaktor bolshoy moshchnosti kanalniy:* high-power channel-type reactor) construction type. Relying on nuclear energy by >70% shortly after regaining independence, Lithuania had to close down this plant as a precondition for joining the EU, as Bulgaria had with Kozloduy. Now, Lithuanians plan to erect a new, similarly huge NPP at the same site next to the border triangle with Latvia and Belarus to supply all of the three Baltic states even though Latvia is now the EU's champion in regenerative energy supply, producing almost 40% of its (admittedly rather limited) demand mainly from hydropower, wind, and some biomass use, and suffers from financial crisis (much like Ireland and Greece but less perceived to do so since they still maintain their own currency, the Lat). Figure 5 shows the global pattern of distribution of NPPs that is almost complete [note that there are 1) considerable differences in electric outputs of individual plants and 2) some plants were primarily meant to produce radionuclides rather than electricity, although they

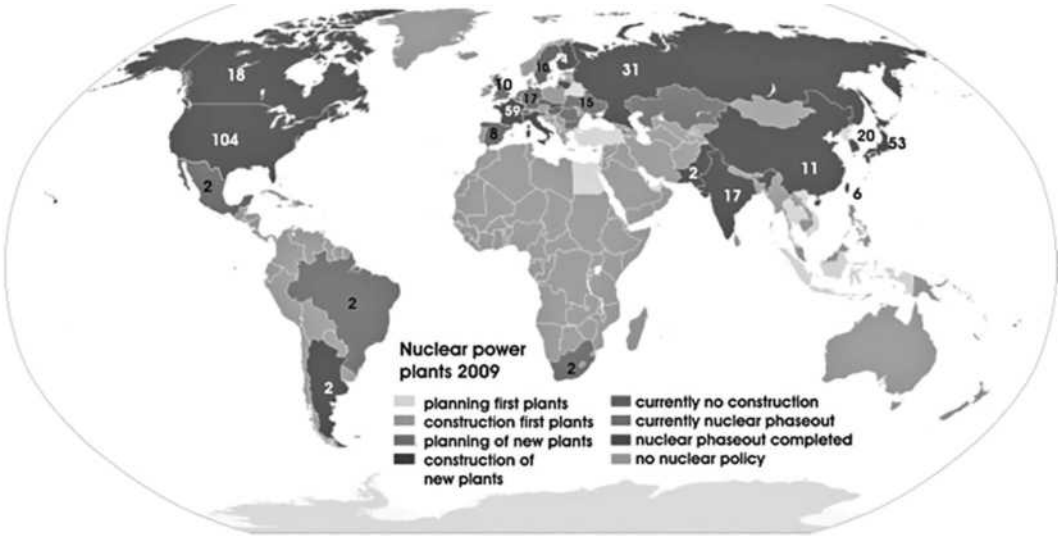


FIGURE 5 Nuclear power worldwide in 2009 (the figure was transferred from Wikipedia; the number of nuclear power plants has been added by collecting information separately by Wikipedia). The number of nuclear power plants is not complete.

are connected to the public current grids (e.g., Dimona in Southern Israel)]. The total number is about 450 NPPs.

As shown above (Figure 1), uranium reserves are to last for just another about 30 years from now (2011), adding another problem to safety concerns and disposal of radioactive wastes (see below) associated with nuclear (fission) energy. The increase in prices during the last 5 years even relatively surpassed that of oil, the process probably going on as demands for uranium did not yet reduce considerably on a global scale after Fukushima. Figure 6 shows a map indicating the 10 most prolific uranium-mining countries (by 2008).

Let us turn to the third issue associated with nuclear energy: radioactive wastes, which are produced by all nuclear fission, then located in "used" (irradiated) fuel rods, neutron impact on non-fissionable ^{238}U, and on construction materials from Zircaloy to concrete, and their first decay products, sometimes dissolved in activated wastewaters including dissolution residues (nitric acid) in nuclear fuel reprocessing (now banned in both the United States and Germany), must be disposed of and stored until radioactive towards stable products is essentially complete. A fuel rod in an NPP is a metal (Zr mainly) tube filled with cylindrical sintered pellets of UO2 or some mixture of UO2 and PuO2; metal alloys or other compounds (carbide UC2, hydride UH3) are used in minireactors only. After a while, a fuel rod is "spent," reducing the ^{235}U content from the original 3.5% to some 1.3% while 1–1.5% of plutonium—most of it fissionable also—were produced from ^{238}U in situ. Although use could be continued by further pulling out the control rods that absorb excess neutrons, it would be no longer safe to work with such rods. Thus, they are replaced, usually a third of them every year or so. The "spent" rods are stored for several decades, usually next to "their" reactor to get rid of the highly active short-lived isotopes of high yields, e.g., $^{141;144}$Ce and $^{103;106}$Ru and ^{91}Y, then either processed or put to a final depository (if there is one, by now only in Finland).

The case is not at all settled; there are no operating final deposit sites but only such ones meant to contain (withhold) the waste for a few decades. Obviously, radioactive waste solutions were discarded into the open sea or rivers both after accidents and routinely during nuclear reprocessing [Windscale/Sellafield (U.K.) and La Hague (France), Mayak near Chelyabinsk (Russian Federation)], not to mention simply dumping entire discarded reactors from nuclear-powered submarines into the sea, with or without the rest of the vessels … What can be done responsibly with nuclear waste instead?

FIGURE 6 The 10 states (dark gray) that produced the largest amounts of uranium ores (in 2008). Besides the spatially largest states outside of South America and except for China and India, most are located in Western Africa, including Gabon (Central Africa) whose ultra-high-grade (>60% U) deposits at Oklo and neighboring sites gave rise to natural reactors some 2 billion years ago (map from Wikipedia.de).

The periods of time over which radionuclides from fission reactors must be stored and safeguarded are outright unimaginable and far beyond any other time frame of political or economic planning: there are nuclides with half-lives of several million years (^{94}Nb, ^{129}I, ^{237}Np); thus, they will create a danger for 10^7 year or the like. With ^{237}Np (tons of which currently exist) and ^{243}Am (americium, tens of kilograms of which exist), there is an additional problem: during very long storage, enriched samples of either nuclide will spontaneously turn into fissionable materials (^{233}U and ^{239}Pu, respectively), causing heat and neutron release to increase after millennia.

Here are some examples:

Uranium	^{238}U	4.468 billion years
Uranium	^{235}U	704 million years
Iodine	^{129}I	15 million years
Neptunium	^{237}Np	2.144 million years
Plutonium	^{239}Pu	24,110 years

While actinides [from the third, protactinium (Pa, $Z = 91$) onwards] generally are extremely chemotoxic, this does only matter for long-lived nuclides like the natural uranium isotopes, ^{237}Np, $^{242;244}$Pu, or $^{247;248}$Cm. In the other cases, say at $T_{1/2} \leq 10^5$ years for α emitters with negligible spontaneous fission shares, radiotoxicity, i.e., the effects caused by particles of ionizing radiation emitted during decay, prevails even against this chemotoxicity. Among these nuclides, $^{239;240}$Pu (which are produced together in a reactor given there is substantial irradiation of a uranium sample) are peculiar in their radiotoxicity but are rather long lived as they do reside in the body at very sensitive points: the marrow and mucosa around or in bones and liver, while radioactive ("hot") particles may be inhaled and reside in the lung, exposing it to radiation. Note that some 6 tons (!) of ^{239}Pu that escaped fission during nuclear bomb tests were spread up into the stratosphere as an aerosol that still keeps on being deposited ("fallout") apart from some other transuranic nuclides: from analyses of fallout composition, 1100 kg of ^{240}Pu and substantial amounts of $^{241;243}$Am (produced in nuclear weapons containing ^{241}Pu or ^{242m}Am as fissionable nuclides, which have much smaller critical masses with fast neutrons than all $^{233;235}$U and ^{239}Pu) are determined. Estimated amounts of the latter are 25 kg ^{241}Pu, 70 kg of its decay daughter ^{241}Am, and 2.5 kg ^{238}Pu [most of it from breakup of a RIG (radioisotope thermoelectric generator) during re-entry in a 1960s space probe mission; data calculated from Breban et al.[18] Besides this, there are natural contributions of plutonium in the biosphere: some 5 kg of ^{244}Pu is still left over from the origins of the solar system owing to its 83-milion-year half-life and trace amounts (<1 g) of ^{238}Pu from the double-β-decay branch of ^{238}U. Most fallout radioactivity is still due to ^{241}Pu ($T_{1/2} = 14$ years), but this is less destructive in radiotoxicological terms because it does undergo β-decay, rather than α-decay, followed by each comparable activities of $^{239;240}$Pu and ^{241}Am ($T_{1/2} = 433$ years). This compromises sensitive tissues.

Cancer rates, e.g., bone skin sarcomas and their metastases, will thus increase after a Super-GAU (German: *größter anzunehmender Unfall* means worst case, meltdown) if plutonium is released, like in the 1951 Mayak accident (Russian Federation) or after a Pu-based nuclear fission bomb was destroyed 1) in an airplane crash at Thule (Greenland) in 1968 or 2) by mis-ignition ("fizzle") in the *Hardtack Quince* test (explosion yield but 0.02 kt TNT) at Runit Island, Enewetak Atoll (now some part of Republic of Marshall Islands) in 1958. Both latter events spread several kilograms each of ^{239}Pu over a very restricted area, which made effectively cleaning it impossible. The doubtful results of (2 out of 5) nuclear bomb tests in Pakistan (1998) and the first one in North Korea (2006)—neither yet precluding burst of Pu and fission products to the surface though these were underground test "fizzles"—must be added to this list. Runit Island (Pacific Ocean) is now used as a dumpsite for highly contaminated materials from nearby test craters, the hole capped by a concrete dome of some 100 m diameter. This is the gravest possible kind of catastrophe from an ecological and ecotoxicological point of view, going beyond even what happened at Fukushima or Chernobyl. The environment, including groundwater, is so polluted that access is strongly impeded for centuries or even longer.

As with Chernobyl and the Bikini Atoll test site, animals and plants apparently adapted to the harsh radiological conditions there. Starting soon after the Chernobyl accident, researchers noted that diverse animals (including wolves, foxes, lynx, moose, hares, and many kinds of birds)—and some humans—returned to the off-limits area around Chernobyl, including the so-called Red Forest and the decontamination lake. Among these, there are species that try to avoid man, prompting them to invade an area where there are (almost) no humans left behind. In addition, biodiversity is lower than in comparable areas, brain sizes tend to be reduced in both mammals and birds, and there are other malformations, tumors, and evidence of genetic alterations.[19] Likewise, there are lots of aquatic life in Bikini Atoll [except for the crater basin of the largest ever U.S. test, *Castle Bravo* (1954, some 15 MT)], but once again, there is reduced biodiversity.

At least since 9/11 some 10 years ago, but actually ever since the first threat to attack a NPP by a flying passenger airplane abducted before (in 1972 over Oak Ridge, Tennessee) citizens and public authorities of countries which make use of nuclear fission power: while older NPPs—all over the world—must be considered to be improperly protected against even the impact of a smaller airliner there is virtually none anywhere to withstand the perpendicular impact of a fighter plane. The jet engine compressor axle in a military plane is about 3–3.5 m long, made of dense metals such as niobium alloys, then hitting a concrete ($\rho \approx 2.8$ g/cm^3) shell at about the speed of sound. This will suffice to penetrate some 10 m (!) of concrete, as compared to an actual containment thickness of less than 2 m. In addition, global warming and eutrophication combine to increase the likelihood that river or ocean water cooling can no longer be taken for granted in summer at least: in 2003, rivers became so warm all over Central Europe (probably related to global warming) that NPPs from France to Lithuania had to be shut down for many weeks. In the same summer, cooling water entries at the Russian Sosnovy Bor plant were blocked by algae excessively growing in the adjacent Bight of Helsinki. Either problem is likely to occur more often in the future.

Transmutation[20,21] is discussed as a means of faster disposal of fissiogenic radionuclides, exposing the nuclide mixture obtained by fuel reprocessing to a high-energy proton beam. Fission products that are distinguished by a considerable excess of neutrons in their nuclei hence are brought closer to stability while actinides will either undergo fission directly or produce at least nuclides of much shorter lifetimes. It is completely unsettled and doubtful whether this can be done on a scale of tons of material.

Rethinking in the Way for Ecological Economics

The previous entry was mainly concerned with the problems of conventional and nuclear energy sources from an economic, ecological, and political point of view, which will challenge us more and more in the near future. There are several quite different—and mutually independent— reasons to switch to alternative,[4,22–251] that is, regenerative, resources of energy, including resources of fossil fuels becoming scarce(r), constraining greenhouse gas emissions, and striving for geopolitical risk management (spreading suppliers among most diverse regions and political systems, if you are not in a position to produce the materials related to your energy demands domestically). Although there are numerous incentives, this transformation will take several decades not only on a global scale but also in national dimensions, and it depends on both economic interest positions and technical innovations. While wind energy is, at present, technically "ripe," the size of rotors being no longer limited by mechanical problems with generators, etc., exposed to vibrations and bending along a horizontal axis, but by the size of available cranes required to erect them (maximum: some 200 m rotor diameter, 7–9 MW peak power), and photovoltaic devices that reasonably benefit from technologies of semiconductor processing—which is now advanced much beyond anything that would ever be required in solar energy conversion—becoming a large-scale-business as well [the combined area of highly integrated chips that make it to the market in computers and numerous other devices (some 70 of them alone in a common car) annually is in the square kilometer region also], extracting energy from the oceans actually still takes an engineer's ingenuity to make it reasonably work. For solar energy, the "surviving" technical innovation challenges are restricted to thin-layer and organic semiconductor systems and to conversion/storage in a convenient chemical storage form, like methanol.

The following part of this entry will deal with both those energy forms already present on the global market for energy conversion devices, plus a discussion of how energy can be harvested from the open sea, keeping in mind which will be the challenges for Germany or Italy who both decided either to abandon (D) the use of nuclear energy until 2022 or not to restart it (I). This prompts the question whether redesigning the energy supply of a medium-sized highly industrialized country, including consumers which are really demanding, within some 10 years, merits further consideration which is added, too.

Globally View of Renewable Energy

Available energy resources are "renewable" when they are sustainable and environmentally benign (which need not coincide) if

- They renew themselves in the short term by itself (e.g., biomass).
- Their use does not contribute to the depletion of the source (e.g., wind, sun, water).
- They do not have an impact on the environment.

Sustainability in that sense means a triad related to ecology, economy, and social affairs. Figure 7 gives an overview of different energy sources like sun, wind, water, geothermal energy, and biomass for the establishment of technical supported renewable energies.[4,22,24,26,27]

In 2010, the Global Status Report on Renewable Energies[28] was published, parts of which are now to be quoted. In its preface, El Ashry (UN Foundation) already noted more than a hundred states to have political agendas or strategies in 2010 in favor of spreading and widening the use of renewable energy sources, about a doubling from just 55 five years before. In 2009, the largest increases were observed with wind power and photovoltaics, investing some $150 billion into extending capacities and producing and implementing renewable energies vs. only $30 billion in 2004. Except sometimes for advancement of nuclear energy, more money is allocated into capacities and growth of renewable energy sources than into fossil types. By reaching substantial shares of energy input (particularly in some less-populated countries like Denmark, Latvia, or Mongolia), renewable energies did have a turning point, which renders them significant, as they also address the problems of climate change. As implied before, acceptance and broad

FIGURE 7 Renewable energy sources: onshore and offshore wind power, biomass, solar power, geothermal, hydropower. (a) Onshore wind turbines located outside of Palm Springs, California (photograph: Wikipedia, Tim1337); (b) sugar cane is a major supplier of biomass, which is used either as food or as energy supplier (photograph: Wikipedia, Culture sugar cane, Avaré, Sao Paulo. José Reynaldo da Fonseca, 2006); (c) newly constructed offshore windmills on the Thornton Bank, 28 km offshore, on the Belgian part of the North Sea (photograph: Wikipedia, Hans Hillewaert, 2008); (d) photovoltaic array near Freiberg (Germany) (photograph: Wikipedia, I, Eclipse.sx); (e) Krafla Geothermal Station (2006), North Iceland (photograph: Wikipedia, Mike Schiraldi); (f) Grand Coulee Dam is a hydroelectric gravity dam on the Columbia River in the U.S. state of Washington. It is the largest electric-power-producing facility in the United States (photograph: Wikipedia, (http://users.owt.com/chubbard/gcdam/html/photos/exteriors.html) U.S. Bureau of Reclamation). Larger dams are located in Russia, Brazil, and China.

application of renewables are not (no longer) restricted to highly developed industrialized countries but extend to developing nations now displaying more than half of implemented and operating renewable energy supply systems. However, it remains to be seen as to how far this is actually in favor of domestic development, better situations especially in rural areas, or larger autonomy towards globalized oil and coal markets or is another blueprint for export of goods to the North: growing sugar cane and oil palms for biofuel exports in Brazil or Indonesia threatens local primary forests if not even the food supply there. The Desertec blueprint (http://www.desertec.org) for producing lots of hydrogen from solar sources in Northern Africa revives existing supplier roles with countries like Algeria once again.

Figure 8 gives a more precise idea of the present relevance of renewables, their share being some 19% of total in 2008. Share of final energy means: at the point of end use, as electricity, heat, and directly used fuels. This method counts all forms of electricity equally, regardless of origin. The European Commission adopted this method in 2007 when setting the EU target of a 20% renewables share of energy by 2020. Thus, it could be called the "EC method."[29] This includes all traditional sources of biomass energy carriers (biogas, ethanol, wood, etc.), large hydropower plants, and the "novel" kinds of renewables such as small-scale hydropower plants, modern biomass sources, wind and solar energy, geothermics, and biofuels. To be exact, it should be noted that some of the "modern" ones actually are rather old, from small-scale water mills also delivering electricity up to biofuels [Rudolf Diesel demonstrated his engine to be run with peanut oil fuel already at the 1900 World Fair (EXPO) at Paris].

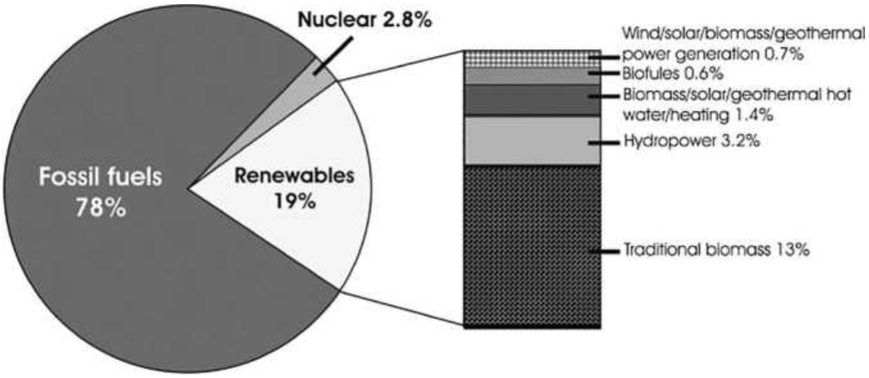


FIGURE 8 Global final energy consumption of conventional and renewable energy in 2008 after REN 21 (2010). Also consider footnote 327 of REN 21 (2010).[28]

Most of these 19% (13%, i.e., relatively 70% among the renewables) rests with traditional biomass used for cooking and heating. This, however, is subject to some change and even cutbacks, especially because there is virtually no possible firing wood (not even small branches) left over around (and this is to say, up to 50–70 km from the outer edge of “informal suburban settlements”!) megacities[30] all over developing and threshold countries, causing to replace with more advanced (e.g., solar cooking devices) or at least more efficiently use these resources. Hydropower represents 3.2% (17% out of renewables) and is growing modestly but from an already advanced level. The other renewable energies contribute 2.7% (relative share: some 14%) and undergo very fast growth in both industrialized and certain developing countries.

In developing countries, renewable energies have a particular role in national support and corresponding policies as well, contributing to their present share of renewable energy capacities of more than 50% [but compare this to both their share of global population, which, including the BRICS (Brazil, Russia, India, China) states except Russia, is >80%, on one hand, and their share of global energy consumption, on the other hand]. China now leads in several indicators of market growth. Among suppliers of wind energy, India now ranks No. 5 globally, owing to its substantial expansion of biogas and photovoltaic capacities especially in the countryside (India also plans to install another 15 NPPs owing to its being populated as densely as, e.g., Belgium or the Netherlands in total and citing growing industrial energy demands). Brazil is now a key supplier of ethanol from sugar cane, adding to other kinds of biomass and wind power in its renewable energy portfolio. Concerning the complete array of renewable energy sources, Argentina, Costa Rica, Egypt, Indonesia, Kenya, Tanzania, Thailand, Tunisia, Uruguay, and others boast high growth rates, often even surpassing their sometimes impressive gross economic growth rates.

In a nutshell, the geographical distribution and political significance of renewable energies have considerably been altered in favor of near-global distribution and application. As compared to quite a few countries in the 1990s, there are at least 82 countries presently operating wind power plants. As for manufacturing these devices, production was relocated from Europe to Asia, with the largest shares in China, India, and South Korea, which additionally become more devoted to renewable energy applications. China not only is the “production yard” of the world for conventional items including consumer electronics but also made, in 2009, 40% of PV devices (from solar cells to complete panels with control and DC/AC converter units), 30% of the world’s wind power plants, and even 77% of solar-driven water-heating collectors. All over Latin America, countries like Argentina, Brazil, Colombia, Ecuador, and Peru increase their production, including exports of biofuels, additionally investing in other kinds and technologies of renewable energies. More than 20 countries in the “Solar Belt” of the Middle, Northern, and sub-Saharan Africa are involved in renewable energy markets.

Large economic gains and considerable further technological changes are to be achieved from this state of affairs and developments beyond the principal players of the highly developed world such as the European Union (EU), the United States, Australia, Canada, or Japan. With renewable energies gaining a truly global footage and application area, there is internationally growing confidence into renewable energies being less susceptible towards perturbations by either political turnovers or market changes such as financial crises than "classical" energy carriers (which readily, like other raw materials, become an item of speculation undergoing tremendous price level oscillations in such conditions).

Another impetus to renewable energy development—as to every other kind of large-scale technical innovation—is its inherent potential to create entire new industries (or at least create vast uncharted fields of opportunities for existing technical branches) and thus millions of new workplaces. In photovoltaics, much larger amounts of semiconductor-grade to medium-purity silicon, SiH_x, and GaAs are used than in integrated (chip) solid-state microelectronics (e.g., GaAs chips are applied in cellphones), whereas technologies required to obtain control of a large rotating propeller in changing conditions, originally developed for helicopters, now increase reliability and output of wind power systems. Creating new workplaces is, by the way, the positive, friendly side of the "dual-use" problem: both engineers and blue-collar technicians who were mainly concerned with development and production of arms systems had to strive to save and "humanize" their workplaces after the Cold War came to an unanticipatedly happy end around 1990; these highly skilled metal workers and engineers at Kiel harbor (Germany) then started to focus on wind power plants (with rotor techniques borrowed from military helicopters) and integrated current-heat support systems (using diesel engines originally designed for tank propulsion to consume and convert plant and waste oils to produce some 1 MW each of electricity and heat). Hence, although not creating environmentally benign branches from scratch, this "Arbeitskreis für Alternative Produktion" (workgroup planning alternative production) and similar endeavors in U.S., British, Australian, and French arms enterprises effectively increased the array of possible customers much beyond the state (i.e., department of defense and sometimes police forces) and in the same turn reduced political-military dependences even though most of the respective employers did not really like the idea of employees considering what should (better) be produced on their own.

This is part of the ethical issues and bonuses associated with alternative energy production: there is a considerable bonus in terms of both workplace safety and numbers of workers required to install and run 1 GWel of alternative energies as compared to fossil and nuclear types; besides, there are fewer risks associated with making PV devices than with coal or uranium mining for the miners themselves, counting and comparing, for example, mine accidents and cancer fatalities per MWy. In 2009, there were an estimated three million workforce directly related and devoted to renewable energies, about half of them concerned with biofuels, and many more than this in branches indirectly connected with renewables (Table 1).

What are the recent performances of renewable energies as of 2009 (Figure 9, source: REN 21[28])? For the second consecutive year, in 2009, in both the United States and the EU, the newly installed renewable energy capacities exceeded those of combined conventional fossil energies and nuclear power. Renewables accounted for 60% of newly installed power capacity in Europe in 2009, and nearly 20% of annual power production.

Christopher Flavin (Worldwatch Institute) pointed out in his entry "Renewable Energy at the Tipping Point" within the REN 21[28] report that China's recent leader role in producing wind rotors and photovoltaic devices just gives proof of the political prerogatives in favor of renewable energy exploitation, including both laws and funding, to be successful. Although there were initial problems, the important reforms in China starting with the national legislation on renewable energies of 2005 caused fast and efficient development there. China since then increased its efforts to become a leading innovative power as well as key producer of renewable energy technologies.

Table 2 shows the five countries that are the most important players concerning renewable energies, as of 2009.

TABLE 1 Jobs Worldwide from Renewable Energy (REN 21[28] with Added Information from the United Nations Environment Programme Report 2010)

Industry	Estimated Jobs Worldwide	Selected National Estimates
Biofuels	>1,500,000	Brazil, 730,000 for sugar cane and ethanol production
Wind power	>500,000	Germany, 100,000; United States, 85,000; Spain, 42,000; Denmark, 22,000; India, 10,000
Solar hot water	~300,000	China, 250,000
Solar PV	~300,000	Germany, 70,000; Spain, 26,000; United States, 7,000
Biomass power	–	Germany, 110,000; Unites States, 66,000; Spain, 5,000
Hydropower	–	Europe, 20,000; United States, 8,000; Spain, 7,000
Geothermal	–	Germany, 9,000; United States, 9,000
Solar thermal power	~2,000	Spain, 1,000; United States, 1,000
Total	>3,000,000	

Note: Further information about the evaluation of the data are reported in REN 21,[28] p. 75, note 226. The table is incomplete.

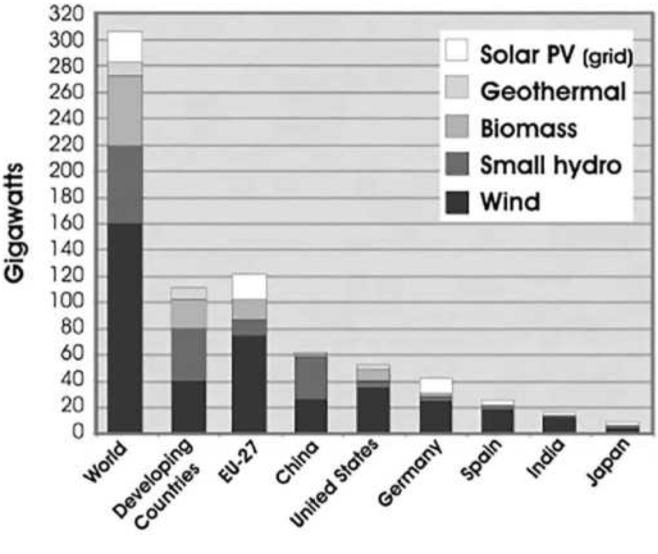


FIGURE 9 Renewable power capacities in 2009 without inclusion of large-scale hydropower: Developing World, European Union, and Top Six Countries (REN 21 2010).[28]
Note: Only includes small hydropower <10MW.

Owing to the well-known variabilities of renewable energy output, which are due to weather, time of day, and longer-period (tidal power plants) periodic or aperiodic changes, there are much larger theoretical (peak power output) renewable energy technical potentials than average yields. The present (as of end of 2009) capacity of a global 1.23 TW (1230 GW) that now constitutes just more than 25% of total electric generating capacity worldwide thus is considerably larger than the actual share/contribution of produced electricity.

What about Ocean-Related Energy [Waves, Ocean Currents, Tidal Power Plants, Osmotic Energy Conversion, Ocean Thermal Energy Conversion (OTEC)]?

The power associated with flowing water is impressive and has motivated people to use it many centuries ago in mills located at running creeks and rivers. It was an obvious idea to extend this technique to tapping ocean currents, like the Gulf current, as well as tidal water flows[31] that can reach speeds much above those seen in most rivers [e.g., some 6 m/sec (11 knots/hr) along the Welsh coast of the Atlantic

TABLE 2 The Five Countries That Are the Most Important Players concerning Renewable Energies [as of 2009[28]]

	#1	#2	#3	#4	#5
Annual amounts in 2009					
New capacity investment	Germany	China	United States	Italy	Spain
Wind power added	China	United States	Spain	Germany	India
Solar PV added (grid connected)	Germany	Italy	Japan	United States	Czech Republic
Ethanol production	United States	Brazil	China	Canada	France
Biodiesel production	France/Germany		United States	Brazil	Argentina
Existing capacity as of end of 2009					
Renewables' power capacity (including only small hydro)	China	United States	Germany	Spain	India
Renewables' power capacity (including all hydro)	China	United States	Canada	Brazil	Japan
Wind power	United States	China	Germany	Spain	India
Biomass power	United States	Brazil	Germany	China	Sweden
Geothermal power	United States	Philippines	Indonesia	Mexico	Italy
Solar PV (grid connected)	Germany	Spain	Japan	United States	Italy

Ocean (rather than perpendicular to it)], providing concentrated energy as flow speeds are close to those in air (wind), with water being 800 times as dense; thus, a rotor of equal diameter exposed to a water flow of equal speed delivers 800 times as much power [or the same power at $\sqrt[3]{(1/800)}$. this speed, i.e., some 10.8%]. Basins that are filled with water at maximum level differences of a few meters are commonplace in electric storage (in German: Pumpspeicherwerk), with the same being offered by tidal changes of ocean and estuary water levels [the largest tides are seen in river mounds, e.g., River Severn (Wales) or River Rance (France, Normandy) at some 10 m], and not just after pumping water into them by electrical power obtained otherwise, but for free twice a day (actually even 4 times daily using differences of either levels).

Concerning periodical filling and uncharging of storage basins connected to flowing water turbines, there is but one large power plant (240 MW_{el}) in the world still connected to the French grid back in 1967. It is located at La Rance in the mounding of River Rance, next to the famous island Mont St. Michel (Figure 10).

FIGURE 10 Aerial photograph of the world's biggest tidal power station. The Rance Tidal Power Station is located at La Rance (NW France) in the mounding of River Rance, next to the famous island Mont St. Michel.
Source: Photograph: Wikipedia.

Here, the tidal water level differences are far larger than the global average of about 3.5 m (isolated ocean basins, such as the Black Sea and the Baltic Sea, and even the Mediterranean Sea, tend to lack any significant tides, e.g., average tides in the Baltic Sea are some 20 cm); the minimum tidal heights required for a meaningful operation of a tidal power plant are estimated to be some 5 m.[32] Obviously, the gain of energy from a basin of given size interacting with the tide flows increases by the square of tidal water level changes: the amount of water flowing in and out is proportional to tidal height and so is the energy gain from a given mass of water flowing through the turbines. Now being operated for more than 40 years (it was connected to the French grid in 1967), effects from this plant and its overall performance can be well evaluated:

- Local tidal height decreased from some 14 m (!) to <8 m due to deposition and relocation of sediments.
- Corrosion is an issue, requiring both to avoid combination of different alloys in the construction to exclude galvanic effects in seawater, and active electrochemical protection measures.
- There were some ecological side effects concerning distributions of limnetic and marine fishes in and on either side of the basin, with the power plant shifting the limits of marine and limnetic populations of fishes somewhat downstream into the estuary. The turbines rotate so slowly that both fishes and squids can pass through them without being hurt, much unlike classical running-water power plants.
- Actual average output to the grid is some 540 GWh/yr, which is an average of 62 MW, about 25% of nominal power production.

Nevertheless, the La Rance plant remained unique, being the only large-scale plant existing, although sites having even higher tidal heights such as Bay of Fundy (Canada, up to 19 m) were used for yet smaller installations.

Waves[33] may be destructive, but their energy can also be technically exploited, e.g., with transducers based on either bending of some part of the device [hydraulically or by piezoelectric (piezoelectricity means electricity resulting from applying pressure, squeezing) devices]. Other devices employ either systems operating on periodically compressing and expanding air by by-passing volumes in a volume (some hollow metal or concrete bunker) under which the waves pass through and pass this air over a rotating turbine or using the alternating flow directions of water or directly converting the water level oscillation into electricity by electromagnetic induction.

It is estimated that one can gain some 15 kW_{el} (130,000 kWh/yr) from a single meter of coastline around the North Sea, which is comparable to the gain earned by covering about 1 km of land behind this shoreline by photovoltaics in the same climate (this sounds stunning if not paradoxical but consider the width of the seas where the waves could gain energy before from wind rather efficiently converting solar energy). So far, wave energy converters were constructed at Western ocean coasts in the Northern Hemisphere, including SW Norway (Toftestallen, near Bergen), the northernmost Scotch coast, open to the Atlantic Ocean and Oregon. A most straightforward way to convert mechanical movements into electricity, besides piezoelectric techniques, is by magnetic induction. A buoy is towed to the seafloor, with a magnetic shaft tethered to it, while the floating part of the buoy contains some induction coil that hence moves up and down around the shaft within the magnetic field produced by the latter, directly producing an alternating current even though its voltage and frequency still have to be adjusted.

Of course, such interferences with natural water flows may alter and affect the ecological situation in estuaries especially; in the best cases, using both tidal and wave energy reduces coastal erosion.

Less obvious sources of energy associated with ocean and shoreline,[34] respectively, include the temperature difference between surface waters and the deep sea (particularly in the tropics, except for the Red Sea and its annexes like the Gulf of Aqaba/Eilat), some 25–30°C there and the gradient in salt content and, thus, the osmotic pressure difference of some 26 bar between the ocean and freshwater from rivers and creeks running into it (the latter being identical to have this water falling through a Pelton turbine from about 260 m of height).

At a surface temperature of some 300–305°K in tropical waters, the Carnot (most efficient heat engine) maximum efficiency would be about 8%–10%. Usually, in OTEC plants, ammonia is used as a heat-transferring working medium; actual efficiencies are about 3% due to long hoses required to exchange hot and cold waters. Rather than steep cliffs selected earlier, OTEC plants are now located on either ships or floating constructions much like those employed in crude oil production. Given the problems associated with coral bleaching due to surface waters exceeding 30°C, some larger use of the OTEC technology should even provide a real ecological bonus while mixing nutrients between deep and surface waters might cause problems of eutrophication.

The osmotic pressure of some solution is about the pressure that would be exerted if the same concentration of dissolved entities (i.e., ions from, say, $MgCl_2$ solutions, to be counted individually, hence producing 3 times the osmotic pressure in water or acetonitrile than if dissolved in non-ionizing solvent like a long-chain alcohol); 1 M/kg of solvent (here water) hence is equivalent to a gas compressed to 1 M/L, which is tantamount to a pressure of some 24 bar at 20°C. A unimolar $MgCl_2$ solution in water would thus produce an osmotic pressure of more than 70 bar, while the corresponding value for seawater—which is 0.55 M NaCl solution to a first approximation—is about 26 bar [the actual value for Norway is a little smaller because of 1) the water being colder and 2) some dilution by the very freshwater discharges into the Fjord site used, which is rather remote from the open ocean].

Before dreaming of what the latter source could deliver when passing but substantial parts of the freshwater flows of giant rivers like River Amazon, Mississippi, Yangtzekiang (China's largest river), Kongo, or the Siberian rivers like Yenisei through such devices, be reminded of some problems that are not yet solved. The only demonstration plant existing so far is located at a creek mounding into a fjord in Norway, producing some 3–4 kW of electric power since November 2009. There are three quite different kinds of converting osmotic potentials into electricity:

- Using a water-permeable yet pressure-proof membrane (this is done in Norway). The saltwater content of this closed membrane volume is going to absorb pure water from the freshwater passed along outside (i.e., in the mounding of the river). Due to the osmotic pressure, the water level inside will increase considerably, allowing to pass it over a falling-water turbine and produce electricity eventually. This process is a somewhat periodical one: as the saltwater gets diluted during the process, you need to discard it into the ocean sooner or later and refill the chamber with "pure" (3.5% salt) ocean water. By now, electric power output is about $1\,W/m^2$ of membrane interface area; the aim is to achieve $5\,W/m^2$ soon.
- Electrochemical settings using the diffusion potential: concentration differences leveled off by diffusion create electric potentials even though no redox reactions are involved in charge transfer. Likewise, second-type electrochemical cells draw upon concentration potentials: an electrode made of combined silver and AgCl (both solid and mixed among each other) will adjust its potential according to the concentration level of chloride ions. This can be used both in analytical chemistry and for producing electric currents from chloride solutions that differ in concentration (freshwater typical values being 1–2 mM/L as opposed to 0.55 M/L in seawater, giving an open-cell voltage of 150–170 mV).
- Theoretically speaking, the zeta (electrokinetic) potential could also be used in osmotic energy conversion: by osmotic pressure differences forcing water through some membrane, or a column filled with a packed solid, it will produce an electrical potential difference between either side of the interface. This potential is due to selective adsorption of cations or anions onto a typically charged particle, charging being caused by oxide/ hydroxide particles (say, wet alumina) behaving as an acid (adsorbing hydroxide) or base (adsorbing protons) depending on local pH, co-sorbents, and material (point of zero zeta potential). Then, the other, non-adsorbed ions will be passed through along the solvent, and a potential of typically half a V forms. The effect is reversible (electrokinetic water pumping).

The problems are with membrane stability and, more generally, also affecting electrochemical systems, clogging of the interfaces by biomass (mainly phytoplankton) or even mineral concretions. The Norwegian success terminated a history of decades of failed experiments on osmotic power production.

Things became a little different—and better, rather more advanced—because of OTEC. The first plant ever of this kind was constructed by French Georges Claude (1870–1960) at the coast of Cuba back in the late 19th century. OTEC devices are simple thermal power plants using rather small T differences, more like a steam engine compared to turbines, coal-fired plants, or internal combustion engines.

The global distribution of chances for this way of harvesting is solar energy indirectly (having hot water from insolation in the tropics while cooling is provided by Arctic or Antarctic undercurrents at some 1000–2000 m of depth).

Eventually, there can be integrated offshore energy parks making use of almost all the energy sources discussed above combined on, or beneath, a tethered floating island.

Geothermal Energy

For many decades, *geothermal* energy[35,36] has been well established for heating purposes in countries like Iceland, New Zealand, and several developing countries in Central America such as El Salvador. Although electrical uses, with geothermal (fumarole) vapors directly run through a steam turbine, were first tried in Italy more than a century ago (at Larderello in 1904, delivering about 200 W), corrosion and clogging problems remain severe until this day. The obvious reason is the "contamination" of fumaroles with both clogging agents like boric acid and hydrolyzable volatile metal chlorides, besides the large shares of corrosive gaseous acid precursors like SO_2, HCl, and HF. Thus, one has to create a primary heat exchange cycle directly exposed to these corrosive items, as well as a secondary one linked to the heat/mechanical/electrical conversion systems, much like in NPPs and mainly for the same reasons (if not even worse here), and worse, due to the rather small heat difference, this decreases total efficiency of conversion. When obtaining the vapor from underground wells drilled several kilometers into the Earth's crust rather than operating close to active volcanoes, clogging of drilling holes or rock fractures required to circulate some operating medium also remains critical. Hence, it is safe to predict that geothermics will remain more concerned with heating (houses, swimming pools, etc.,) than with electricity production in the near future as well.

Wind Power

Wind power now is an established source of energy, with average production costs per energy unit (e.g., cent per kilowatt-hour) coming close to those by conventional (fossil) energy sources.[37–40] In certain countries, the share of wind power in total electricity production exceeds 20% (Denmark, Mongolia), and there is a broad international consensus that state subsidies are no longer required nor given to enhance the rate of implementation. Rather, as with all kinds of renewable energies that are subject to considerable periodic (sunlight) or non-periodic changes of supply, the optimum strategy of storage becomes imminent. Hence, electrolysis of water (and possibly secondary production of methane or methanol) by "excess" wind power (excesses being produced by mismatches with the grid also), storage of H_2 or CH_3OH, and use of the latter energy carriers in either vehicles or stationary fuel cells connected to the electric grid are gaining importance.

The size of individual plants is limited by the necessity to erect them and thereby place 100-ton items more than 100 m above the ground within millimeter precision, that is, by size of the available cranes. The largest wind power systems thus now have 200 m rotor diameter and deliver some 8 MW while arrays of them ("wind parks") can produce outputs in the size of classical power plants and NPPs both on- and offshore. However, in either case, interactions with local fauna may become significant.

Solar

Solar energy is the key source of almost all the biological and meteorological processes operating on Earth.[41,42] Semiconductor solid-state devices allow for a remarkably efficient exploitation of this source, with an additional role for thermal processes that latter rely on focusing and, thus, on non-scattered sunlight, that is, on clear skies. These thermal processes include production of fuels by cycles involving zinc or cerium oxides as well as metallurgical transformations.

Solar thermal power plants, augmented by natural gas or biogas combustion during night and other dark times, are now realized in a scale of hundreds of megawatts, while photovoltaics in 2011 first yielded an all-year average of more than 2000 MW (2 GW) electrical output in Germany (some 3% of total current). The price breakdown in production and processing of semiconductors (which need not be that pure or advanced than with electronic microdevices) supports the "boom" furthermore, regardless of fast cuts in state subsidies paid for supplying PV current to the public grid in all the EU member countries now.

When considering very large plants such as in the Desertec initiative, the increase of radiation absorption however becomes likely to influence the performance of solar parks by itself: large volumes of heated air will rise right above the plants, causing an increased dust advection to the panels as well as clouds to form on their top (both sailplane pilots and birds of prey look for typical kinds of clouds to spot regions of upwind over hotter surface areas!). High-yield photovoltaics by thin-layer solar cells [a few micrometers' total thickness, unlike 0.3–0.5 mm with polycrystalline SiH_x ("blue silicon")] demands rare (e.g., In) and/or highly toxic (Cd, As, Se, Te) elements, causing problems in all mining, processing them and eventually abandoning old PV devices when their performance sharply decreases after several decades.

Biomass

Biomass can be used in a variety of ways as a source of energy, with combustion of wood and vegetable or animal oils for both heating and illumination purposes dating back as far as the Stone Age.[43,44] More recently, plant oils were introduced into internal combustion engines (first with peanut oil; diesel, 1900), while other engines were fed with either wood distillates (containing mainly CH_3OH and acetone besides H_2 and CO) or ethanol produced by microbial activity, much after steam engines had been powered by either wood, wood-processing residues (e.g., sawdust), or peat to replace hard coal (which is biogenic in itself, of course).

Conversion methods of brown coal—lignite—by hydrogenation, gasification (steam gas process), and liquefaction can also be readily applied to biomass, including less obvious representatives of biomass such as sewage sludge (chiefly containing heterotrophic bacteria), and then mostly even take less vigorous conditions in terms of all temperature, H_2 pressure (3–10 bar rather than hundreds of bars), and needed catalysts. Finally, motivated by the fact that biomasses, especially scrap biomasses, became an item of fuel production (and waste treatment/compaction) once again, the very former coal liquefaction plants in, e.g., South Africa are now used for this purpose. Regionally, in Germany, success and economic performance were poorer, however.

Using scrap or digestible waste fractions relieves an ethical problem from the competition among food/fodder and "energy plants" for the same agrarian areas, but even after avoiding this, one must bear in mind the poor area productivity of photosynthesis—the only economically viable source of biomass energy carriers in a large scale—which typically is 0.5% or a few kilograms of reduced $C/m^{2}*a$, i.e., far short of photovoltaics. While there will be a role in waste processing, a large-scale use of biomass for energy purposes such as in Brazil poses a lot of difficult problems, including ecological ones associated with monocultures, possible fertilization/eutrophication, and high water requirements.

Renewable Energy in Germany and the Planned Nuclear Exit

Concerning its gross domestic product (GDP), Germany is the largest national economy in Europe and No. 4 in the world. In 2009, it was second in export and third in import values. Like with the GDP, Germany ranks No. 4 in energy consumption [measured in fossil fuel (hard coal) mass equivalents (BTU: British thermal unit)] but just No. 21 among the energy producers in the world).[46] The intention of the Federal Government of Germany in fulfilment of Kyoto Protocol obligations is a reduction of greenhouse gas production of about 40% by 2020 and up to about 80%–95% reduction by 2050. Quite recently, the greenhouse gas issue and the risk of NPPs (disaster of Fukushima in 2011) were aggravated by the decision to abandon nuclear energy use in Germany in the early 2020s. Figure 11 shows the fuel mix in 2010 and the aimed fuel mix in 2050 in Germany, which mainly will be supplied by wind and solar power.

By steadily replacing fossil energy sources with renewables, the share of the latter will increase, even allowing for a slight increase in energy consumption (which, in Germany, like most other highly developed countries, is rather constant for decades now, notwithstanding a slight decrease in population happening soon). Hydropower is fully established now except for reactivation of very small local plants, many of which had been in operation since the early 20th century. Hence, the present 3% share will remain almost the same; 35% from photovoltaics corresponds to an average output of some 23 GW, which, in our climates, is tantamount to an introduced peak power of 130–150 GW, more than twice that what now is funneled into the entire grid by all kinds of power plants. Dealing with this excess energy on sunny summer afternoons, possibly by chemical storage (water electrolysis, then linked to fuel cells), remains to be figured out. The total area required to produce this amount of PV electricity is about 1000 km^2 [<0.3% of Germany's total area (356,000 km^2) and <10% of the fields on which "energy plants" are now grown (>12,000 km^2 are covered with rape alone)], even assuming no further improvement in today's Si hydride polycrystallinic or $CuIn(S; Se)_2$ thin-layer solar cells (some 13% efficiency).

Growth and Booming Region Ems-Axis, Lower Saxony (NW Germany)

The previous entries dealt with the global-scale relevance of renewable energies. Besides the environmental issue, there are both economic and social surpluses produced by creating novel workplaces.

This can turn a formerly "just" agrarian region into some diversified boom area as will be shown by the example of the so-called River Ems-Axis (Lower Saxony, Northwestern Germany). For more than a decade now, the region keeps increasing its workforce by 3% per year—no "Mc jobs" but fully qualified jobs that produce social security and modest earnings, with >10,000 enterprises that keep expanding and creating new jobs one year after another. The regional motto reads: "Powerful, innovative and ready to achieve by unconventional solutions—these are our region's benchmarks."

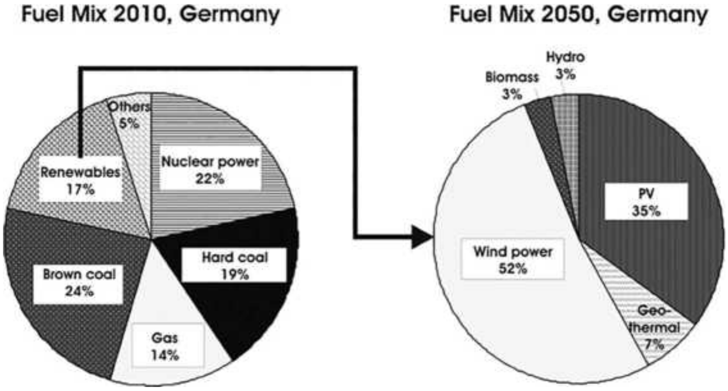


FIGURE 11 Energy mix contributions in Germany in 2010 and the probable future in 2050.[45]

As the Ems-Axis is located next to the North Sea, maritime-related activities are prominent by locating shipyards, shipping companies, and wind power plant producers, among suppliers of other renewable energies. This model region was created through a combination of prudent political support, improvement of infrastructures, synergy among regionally active enterprises, and finally the support of the public. It is located near the Dutch border in Central Europe, making use of already existing East-West connections, and, in addition, links the North Sea shores to the German megalopolis Ruhr district, which is the most populated part of the most populated and economically prolific *Bundesland* of Germany. The Ems-Axis includes the counties Wittmund, Aurich, Leer, Emsland, Grafschaft Bentheim, and Emden City with its large harbor (see Figure 12).

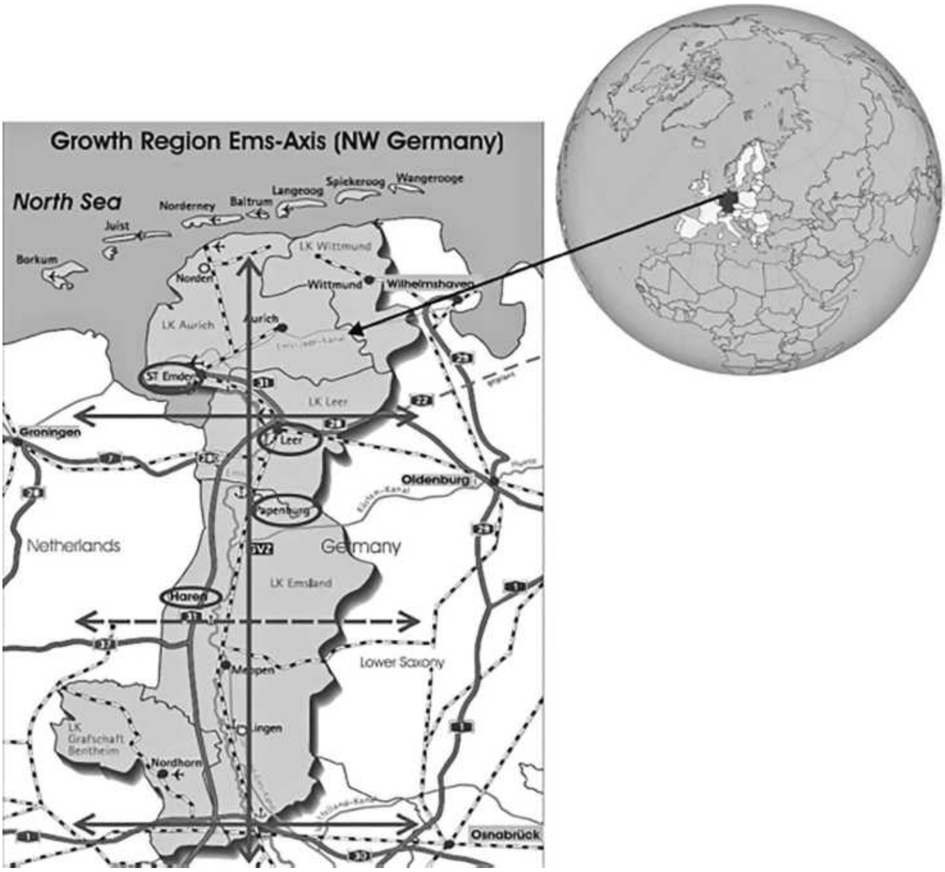


FIGURE 12 The growth region Ems-Axis in Northwestern Germany. This figure shows the excellent infrastructure that will imminently cause new enterprises to settle and expand here. The East-West and South-North highways (motorways) are marked with Nos. 7, 28, 30, 31, and 37, while rivers and channels for ship travel are marked medium gray. These are the River Ems, which is deep and wide enough to permit economically meaningful transportation by ship, and the Dortmund–Ems channel, which extends almost parallel to it. Railway tracks are outlined in black and white. Framed: the cities of Leer and Haren/Ems are among the most important locations for ship owners all over Germany. In Papenburg, there is the Meyer shipyards, among the largest in Europe and moreover the one producing the biggest ships (passenger and cruise ships). Other notable shipyards located next to the shore at Emden recently rather switched to producing wind power plants. The Ems-Axis is distinguished by intense economic activities covering all energy supply, integrated maritime activities, agriculture, and tourism; processing plastics and metals; building vehicles and machines; and providing logistical infrastructure.

Source: The main figure of the Ems-Axis is modified after http://www.emsachse.de. The figure on the right is from Wikipedia, TUBS.

There are six permanent workgroups concerned with energy, integrated maritime economy, tourism, production of plastics items, vehicles and machines, and finally logistics to initiate and run projects. It is the aim of these workgroups and the economic region to make Ems-Axis an independent axis along which economic, travel, and transport activities will organize. This implies strengthening economy-related infrastructure and creating networks for regional economy.

The cities of Haren and Leer combine to be the second-largest shipowner's site in Germany. A total of 750 ships are run from here, making these two special and significant players in running ship travel and dockyards and providing additional maritime goods and items, together with Papenburg. The existing travel infrastructure allows to process materials inshore and, using local logistics, build huge ocean liners such as that for Disney Cruise Line (340 m long and 37 m wide, 128,000 tons, can accommodate 2500 passengers) at Papenburg's Meyer dockyards (Figure 13). The latter commands the world's most advanced instrumentation and facilities for building ships, its workforce being about 2500.

One should mention that there is minimal required bureaucracy used to acquire these infrastructures. This enabled Motorway 31—a crucial North-South connection—to be completed years ahead of planning, with the region providing the required funds itself by joint and coordinated action. Another ambitious project was Euroharbour Emsland at Haren (operated jointly with nearby Meppen city), the construction of which began in 2007. In August 2011, construction of the plant of ENERCON wind power devices began here. ENERCON is the manufacturer of the most advanced wind rotors (the actual propellers), producing blades that are aimed to deliver 3 MW per unit at the Emsland Euroharbour site. The principal administrative person *(Landrat)* of the largest of the involved counties and cities of Ems-Axis, Emsland itself, uses to call this a "pro-climate climate," stressing that currently, an impressive 82% of the energy consumed in the county are derived from renewable energy. The location of Euroharbour, the 24,000-population town of Haren, even boasts a 100% renewable electric current production. Among the renewables, wind is most important for the Ems-Axis region. With the shore nearby and little terrain roughness, it is most suited to create onshore wind plants; thus, NW Germany outcompetes the southern parts of the country in this respect.

Yet, there are also offshore wind power parks in the region now. In 2010, the first one in German domestic waters, "Alpha Ventus," was erected and connected to the grid. As for crucial parts of wind power technology, BARD Energy at Emden both produces rotor blades specifically designed for offshore application (there are special criteria to withstand salt corrosion, impact of water drops on the fastly moving blades, etc.,) and likewise constructs entire power plants at offshore sites (Figure 14).

Suffice this to show features of the booming Ems-Axis economic region, which additionally sports, for example, the Transrapid (maglev) testbed at Lathen, and notably a big plant at Werlte, which will be the first in the world to convert excess wind power energy via hydrogen and hydrogenation of CO_2 into methane for energy storage purposes (to be combined with natural gas CH_4 and biogas). With regard to issues of energy use efficiency and extending the amount of renewable energy supply, the Ems-Axis consortium stated in May 2011:

FIGURE 13 Meyer shipyards at Papenburg (NW Germany). The largest dock is incredibly 504 m long. Right figure: Norwegian Jewel in front of the 70 m tall Meyerwerft Hall.
Source: Pictures courtesy of Wikipedia: left, C. Walther; right, satermedia.de, C. Brinkmann.

FIGURE 14 BARD Emden Energy GmbH & Co. KG produces rotor parts, etc., for offshore wind power plants, then mounting them at sea also. Located at Emden, it belongs to the economic region Ems-Axis. Photos: (Left) German special crane ship for the setups of offshore wind farms called Wind Lift I (BARD) in the harbor of Emden. **Source:** Wikipedia, photographer Carschten. (Right) BARD offshore 1 (Mai, 2011). Courtesy of the BARD Group.

"Partners in growing region Ems-Axis consider big chances for local and regional economy to be obtained from making energy supply a cornerstone of economical politics. Simultaneously they respond to their environmental responsibility by making energy use more efficient and increasing the share of renewable energy sources.

Growing region Ems-Axis is capable of becoming a model (blueprint) energy supply region for the future. Concerning Germany, this region both has the largest concentration of wind power plants and is the site of globally active producers of wind power devices. In addition, renewable energy is earned here from all biomass, sun and geothermal resources. So there is a bandwidth of competence in energy supply which yields new impetus to the region by enhanced cooperation and thus advantages in competition which in turn once more improves the economic performance of the local enterprises".

More pieces of information on the Ems-Axis region, including pertinent enterprises, can be obtained via http://www.emsachse.de.

Conclusion

The present mix of renewable resources used in both thermal and electrical energy delivery represents a superposition of both technical problems still to be overcome (the less so) and political decisions, many of which are made in favor of protecting the respective domestic industries for both producing energies and the very power plants required to obtain and convert them: this partly is a quite reasonable and, to some extent, even responsible industrial policy. Now, there are "old" energy sources, exploitation of which has become so costly that it is worthwhile only in certain most simple conditions, including hard coal and, in another way, oil sands. This statement refers to all economic costs of exploitation, ecological side effects (as well as cultural ones such as destruction of villages and first-nation settlements in favor of open pits), and risk production causes to the workers. The renewables make it to the market step by step with their increasing ability to compete economically and the perspective to relieve old dependences, in addition to avoiding the above risks by offering genuine technical alternatives.

Of course, this might produce problems for countries that have virtually nothing else to offer to today's global markets than their fossil energy carriers, including uranium, but not to some of the "big shots" in fossil fuel mining—highly industrialized countries such as the Unites States, Canada, Australia, and Russia. Apparently, however, there is no convincing perspective of sustainable development by which the common population might benefit from exploitation of fossil energy carriers alone for countries such as Niger in Western Africa (uranium) or Yemen in the Middle East (oil). Other large uranium suppliers like Gabon (West Central Africa) or Kazakhstan (Central Asia/ Eastern Europe) have a more diversified supply portfolio. Several of the Arab oil-producing countries are very aware of what might happen to them, their regimes, their population, and their common welfare (which is often truly restricted to some indigenous minorities) when oil continues to get scarce, and there are cautionary economic examples of countries, societies, and national economies running out of the single, principal minable resource the entire economy was based on, such as the tiny South Pacific Republic of Nauru (phosphate) and Bolivia in Central South America (tin, silver).

Nevertheless, the exchange of our joint economic basis for energy production appears feasible globally within some 50 years from now. It remains to be seen whether this is fast enough both to control climate effects from fossil combustion within acceptable limits and to reorganize completely our strategies of personal transportation while avoiding yet more catastrophes like those in Chernobyl or Fukushima [as well as the failure of a hydropower plant in Longarone (Friaul, NE Italy) which took some 2000 lives in 1963]. Besides this, nuclear power plants—like other technical systems—can run into operation states where they almost or entirely escape control. If a catastrophic accident then can be avoided due to self-regulation or simply luck, it is by no means satisfying or consoling that, e.g., nuclear reactors arrived at states that were not even known to their own operators for extended periods of time (like in Forsmark, Sweden, in 2006), let alone these people would be able to influence it anymore.

The future awaits us but is notoriously hard to predict, but we should take chances, even severe ones, if we decide either way, and we should be aware that doing nothing is tantamount not only to taking chances but also pursuing ways that we know for sure to be not sustainable, not even in the shorter term.

Acknowledgments

We are deeply thankful to the thousands of comments given by colleagues, students, friends, and especially opposite thinking people during attractive and highly motivating discussions during the past decades. For intensive support during the preparation of this entry, we would like to thank Prof. Michael Tomaschek, University of Applied Sciences Emden/Leer.

This entry corresponds in parts with chapter 4.4 (Energy—One of the Biggest Challenges of the 21st Century) of the textbook by Fraenzle, S., Markert, B., and Wuen-schmann, S. (2012) on an "Introduction to Environmental Engineering" published by Wiley/VCH, Weinheim.

References

1. Shafiee, S.; Topal, E. When will fossil fuel reserves be diminished? Energy Policy **2009**, *37* (1), 181–189.
2. Kutz, M. *Environmentally Conscious Fossil Energy Production;* John Wiley and Sons, Hoboken, NJ, 2010.
3. Kreith, F.; Goswami, D. *Principles of Sustainable Energy;* CRC Press, Boca Raton, FL, 2011.
4. Fraenzle, S.; Markert, B.; Wuenschmann, S. *Introduction to Environmental Engineering—Innovative Technologies for Soil, Air and (Ground)water Remediation and Pollution Control;* Wiley-VCH: Weinheim, 2012.
5. Federal Institute for Geosciences and Raw Material Research. *Reserven und Verfügbarkeit von Energierohstoffen (Kurzstudie);* Bundesanstalt für Geowissenschaften und Rohstoffe: Hannover, 2009.
6. Federal Ministry of Economics and Technology. *Energie in Deutschland (Energy in Germany);* BMWI: Berlin, 2010.
7. U.S. Geological Survey. *Circum-Arctic Resource Appraisal: Estimates of Undiscovered Oil and Gas North of the Arctic Circle;* Washington, DC, 2008.
8. Campbell, C. The Rimini Protocol: An oil depletion protocol. Heading off economic chaos and political conflict during the second half of the age of oil. Energy Policy **2006**, *34* (12), 1319–1325.

9. Blum, A. Die finale Ölkrise—fossile Brennstoffe, vor allem Erdöl, sind endlich und eine Verknappung ist absehbar, 2005, available at http://www.raize.ch/Geologie/erdoel/oil.html (accessed on April 11, 2005).
10. Kondratyev, K.; Krapivin, V.; Varostos, C. *Global Carbon Cycle and Climate Change;* Springer: Berlin, 2003.
11. Leroux, M.; Comby, J. *Global Warming. Myth or Reality;* Springer: Berlin, 2005.
12. Hoel, M.; Kverndokk, S. Depletion of fossil fuels and the impacts of global warming. Resour. Energy Econ. **1996**, *18* (2), 115–136.
13. Ostwald, W. *Die Mühle des Lebens;* Theod. Thomas: Leipzig, 1911.
14. Voiland, A. *2009: Second Warmest Year on Record; End of Warmest Decade;* NASA Goddard Institute for Space Studies (accessed January 22, 2010).
15. Reschetilowski, W. Vom energetischen Imperativ zur nachhaltigen Chemie. Nachr. Chem. **2012**, 134–136.
16. United Nations. United Nations Framework Convention on Climate Change: FCCC/CP/1997/INF. 5, List of participants (COP 3), available at http://en.wikipedia.org/wiki/United_Nations_Framework_Convention_on_Climate_ Change (accessed January 21, 2011).
17. Romberg, B. Reaktorneubau: Amerika wagt Renaissance der Kernenergie. Süddeutsche Zeitung, Febr. 13th, 2012.
18. Breban, D.C.; Moreno, J.; Mocanu, N. Activities of Pu radionuclides and ^{241}Am in soil samples from an alpine pasture in Romania. J. Radioanal. Nucl. Chem. **2003**, *258,* 613–617.
19. Gill, V. Chernobyl zone shows decline in biodiversity. BBC News online, Science and Environment, 2010, available at http://www.bbc.co.uk/news/science-environment-10819027 (accessed January 20, 2010).
20. Etspüler, M. *Transmutation: Die zauberhafte Entschärfung des Atommülls*; Frankfurter Allgemeine Zeitung: Frankfurt, 2011.
21. De Bruyn, D. European fast neutron transmutation reactor projects (MYRRHA/XT-ADS). IAEA Review Paper, 2009, available at http://www-pub.iaea.org/MTCD/publications/PDF/P1433_CD/datasets/summaries/Sum_SM-ADS.pdf (accessed November 2011).
22. Hoffert, M.; Caldeira, K.; Benford, G.; Criswell, D.; Green, C.; Herzog, H.; Jain, A.; Kheshgi, H.; Lackner, K.; Lewis, J.; Lightfoot, D.; Manheimer, W.; Mankins, J.; Mauel, M.; Perkins, J.; Schlesinger, M.; Volk, T.; Wigley, T. Advanced technology paths to global climate stability: Energy for a greenhouse planet. Science **2002**, *298* (5595), 981–987.
23. Turner, J. A realizable renewable energy future. Science **1999**, *285* (5428), 687–689.
24. Kaltschmitt, M.; Streicher, W.; Wiese, A. *Renewable Energy: Technology, Economics and Environment;* Springer: Berlin Heidelberg, 2010.
25. Pimentel, D., Ed. *Biofuels, Solar and Wind as Renewable Energy System. Benefits and Risks;* Springer, Science+Business Media B.V., Dordrecht, 2008.
26. Olah, G.; Goeppert, A.; Prakash, C.K. *Beyond Oil and Gas: The Methanol Economy*; Wiley-VCH: Weinheim, 2006.
27. REN 21. Renewables 2010 Global Status Report (Paris: REN 21 Secretariat). Deutsche Gesellschaft für Technische
28. Zusammenarbeit (GTZ) GmbH, 2010, http://www.ren21. net (accessed January 22, 2010).
29. REN 21. Renewables 2007 Global Status Report (Paris: REN 21 Secretariat and Washington, DC: Worldwatch Institute). Deutsche Gesellschaft für Technische Zusammenarbeit (GTZ) GmbH, 2008; 21, available at http://www. ren21.net (accessed January 22, 2010).
30. Markert, B.; Wuenschmann, S.; Fraenzle, S.; Figueiredo, A.; Ribeiro, A.P.; Wang, M. Bioindication of trace metals—With special reference to megacities. Environ. Pollut. **2011**, *159,* 1991–1995.
31. Charlier, R.; Finkl, C. *Ocean Energy: Tide and Tidal Power;* Springer: Berlin, Heidelberg, 2009.
32. Hoffmann, V. Energie aus Sonne, Wind und Meer. Harri Deutsch Thun (SUI): Frankfurt/Main, 1990.

33. Cruz, J. *Ocean Wave Energy. Current Status and Future Perspectives;* Springer: Heidelberg, 2008.
34. Multon, B. *Marine Renewable Handbook*; Wiley and Sons, Weinheim, 2011.
35. Glassley, W. *Geothermal Energy. Renewable Energy and the Environment*; CRC Press, Taylor and Francis Group, Boca Raton, FL, 2010.
36. Ghosh, T.; Prelas, M. Geothermal energy. In *Energy Resources and Systems;* Gosh, T., Prelas, M., Eds.; Springer, Weinheim, 2011; 217–266.
37. European Wind Energy Association. *Wind Energy—The Facts: A Guide to the Technology, Economics and Future of Wind Power;* Taylor and Francis Ltd., Oxford, 2009.
38. Nelson, V. *Wind Energy: Renewable Energy and the Environment;* CRC Press, Boca Raton, Florida, 2009.
39. Hau, E. *Wind Turbines: Fundamentals, Technologies, Application,* Economics, 3rd Ed.; Springer: Berlin, 2012.
40. Maki, K.; Sbragio, R.; Vlahopoulos, N. System design of a wind turbine using a multi-level optimization approach. Renewable Energy, **2012**, *43,* 101–110.
41. Goetzberger, A.; Hoffmann, V. *Photovoltaic Solar Energy Generation;* Springer: Berlin, Heidelberg, 2005.
42. Foster, R.; Ghassemi, M.; Cota, A. *Solar Energy: Renewable Energy and the Environment*; CRC Press, Taylor and Francis Group, Boca Raton, FL, 2009.
43. Goldemberg, J.; Coelho, S. Renewable energy—Traditional biomass vs. modern biomass. Energy Policy, **2004**, 32 (6), 711–714.
44. Fraenzle, S.; Markert, B. Metals in biomass: From the biological system of elements to reasons of fractionation and element use. Environ. Sci. Pollut. Res. **2007**, (6), 404–413.
45. Burkhardt, M.; Weigand, T. Der Strom-Mix in Deutschland (the fuel mix in Germany). German television ZDF, 2011, available at http://www.heute.de/ZDFheute/inhalt/24/ 0,3672,8233016,00.html (accessed January 21, 2011).
46. U.S. Energy Information Administration. Total Primary Energy. Statistics on Germany, available at http://www.eia.gov/countries/country-data.cfm?fips=GM (accessed January 22, 2010).
47. Flavin, C. Last word: Renewable energy at the tipping point. In *Renewable Energy Policy Network for the 21st Century (REN 21) (2010) Renewables 2010 Global Status Report (Paris: REN 21 Secretariat)*; Deutsche Gesellschaft für Technische Zusammenarbeit (GTZ) GmbH, 2010; 52–53.

6

Energy and Environmental Security

Muhammad Asif

Introduction

The economies of all countries, and particularly of the developed countries, are dependent on secure supplies of energy. Energy security means consistent availability of sufficient energy in various forms at affordable prices. These conditions must prevail over the long term if energy is to contribute to sustainable development. Owing to the pivotal role of energy in the modern age, energy security is at the heart of national and international energy policies across the world.

The global environmental scene has changed dramatically over the last century. Global warming and its consequent climatic changes driven by human activities, in particular the production of greenhouse gases (GHGs), directly impact the environment. The environmental security to a certain extent is linked with the energy security as there is an intimate relationship between energy and environment. The production and use of all energy sources results in undesirable environmental effects, which vary based on the health of the existing ecosystem, the size and health of the human population, energy production and consumption technology, and chemical properties of the energy source or conversion device. A shorthand equation for the environmental impacts of energy production and use has been provided by Solomon:[1]

$$I = PAT$$

where I is the environmental impact, P is the size of the human population, A is the affluence of the population (e.g., per capita income and/or energy use), and T is the technology (e.g., energy efficiency, emission rate of air and water pollution). A comparison of different types of energy systems in terms of CO_2 emissions is provided in Table 1.[2]

The energy resources, presently being consumed in the world, can be broadly classified into three groups: fossil fuels, nuclear power, and renewable energy. Fossil fuels contribute to more than 80%

TABLE 1 Comparison of CO_2 Emissions from Different Energy Systems

Type of Power Plant	Fuel/Type of Energy	CO_2 (kg/kWh)
Steam power plant	Lignite	1.04–1.16
Steam power plant	Hard coal	0.83
Gas power plant	Pit coal	0.79
Thermal power plant	Fuel oil (heavy)	0.76
Gas turbine power plant	Natural gas	0.58
Nuclear power plant (pressurized water)	Uranium	0.025
Thermal power plant	Natural gas	0.45
Solar thermal power plant	Solar energy	0.1–0.15
Photovoltaic power plant	Solar energy	0.1–0.2
Wind power plant	Solar/wind energy	0.02
Hydroelectric power plant	Hydropower	0.004

of world's total energy supplies and are mainly classified into the following three types: coal, oil, and natural gas. It is the set of fossil fuels that primarily bond the environmental security with energy security. Global warming—the predominant threat to environmental security—to a great extent is a consequence of fossil fuel consumption. As the two issues go side by side, their potential solutions would also be closely related.

Energy and Sustainable Development

Energy is at the heart of the existence of present-day societies. The accomplishments of civilization have largely been achieved through the increasingly efficient and extensive harnessing of various forms of energy to extend human capabilities and ingenuity. Providing adequate and affordable energy is essential for eradicating poverty, improving human welfare, and raising living standards worldwide. The per capita energy consumption is an index used to measure the socioeconomic prosperity in any society—the human development index (HDI) of a country has strong relationship with its energy prosperity.[3] A direct correlation between the access to electricity and the economic well-being in a range of countries, for example, is indicated in Figure 1.[4] Throughout the course of history, with the evolution of civilizations, the human demand for energy has continuously swallowed. Of present, key factors driving the growth in energy demand include increasing human population, modernization, and urbanization.

Poverty, hunger, disease, illiteracy, and environmental degradation are among the most important challenges facing the world. Poor and inadequate access to secure and affordable energy is one of the crucial factors behind these issues. Electricity, for example, is vital for providing basic social services such as education and health, water supply and purification, sanitation, and refrigeration of essential medicines. Electricity can also be helpful in supporting a wide range of income-generating opportunities. Access to electricity remains to be a serious issue in developing countries as around 13% of the world's population—do not have access to it yet.[5] As of 2017, around 3 billion people rely on traditional biomass, including wood, agricultural residues, and dung, for cooking and heating. More than 99% of people without electricity live in developing regions, and four out of five live in rural areas of South Asia and sub-Saharan Africa.[6,7]

There is a global consensus that the provision of secure, affordable, and socially acceptable energy services is a prerequisite for eradicating poverty in order to achieve the Millennium Development Goals (MDGs). The Earth Summit 2002 strongly urged the nations to "Take joint actions and improve efforts to work together at all levels to improve access to reliable and affordable energy services for sustainable development sufficient to facilitate the achievement of the MDGs, including the goal of halving the proportion of people in poverty by 2015, and as a means to generate other important services that mitigate poverty, bearing in mind that access to energy facilitates the eradication of poverty." United

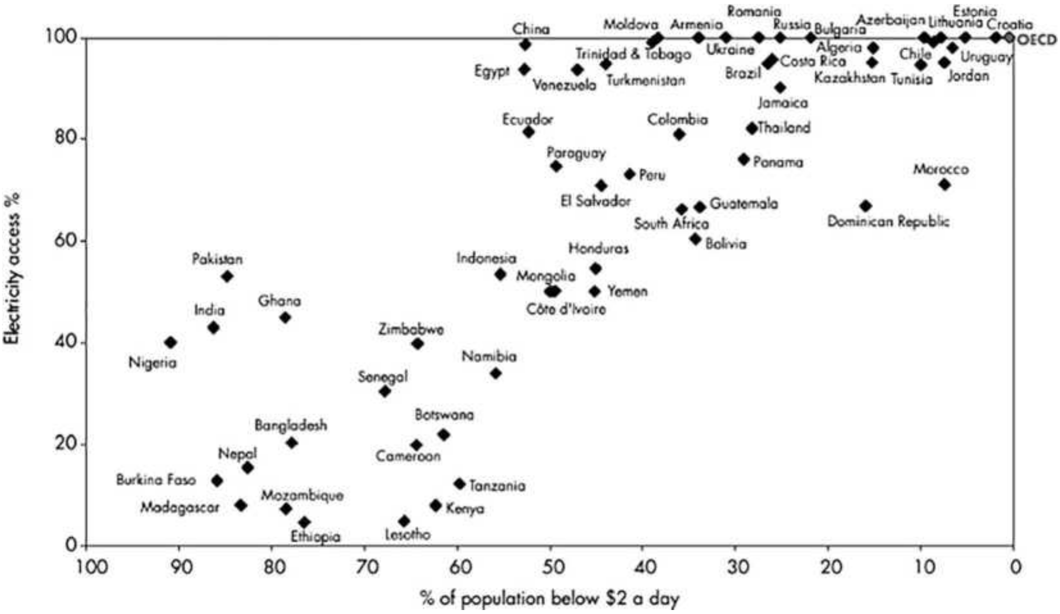


FIGURE 1 Relationship between economic prosperity and availability of electricity.

Nations also acknowledged that "without increased investment in the energy sector, the MDGs will not be achieved in the poorest countries."

With the growing world population and people's innate aspirations for improved life, a central and collective global issue in the new century is to sustain socioeconomic growth within the constraints of the Earth's limited natural resources while at the same time preserving the environment. This target—sustainable development—can only be met by ensuring energy sustainability.

Energy Security

Energy has attained the status of an indispensable strategic commodity, and ensuring its availability is one of the important responsibilities of governments across the world. Failure to ensure robust provision of interruption-free and affordable energy may result into serious financial and social problems. Breach of energy security is also prone to lead to national instability. A great number of countries in the world, particularly the industrialized mature economies, have placed energy security on top of their national policies. The high degree of multidisciplinary interdependency among nations makes energy security vital from both national and international perspectives. German Foreign Minister Frank-Walter Steinmeier emphasizes this fact by stating that maintaining global security in the 21st century will "inseparably also be linked to energy security."[8] At their summit in Brussels in March 2007, the European Union (EU) leaders adopted a road map, the highlight of which was "Energy is what makes Europe tick. It is essential, then, for the European Union to address the major energy challenges facing us today and in future." Similarly, the 2006 G8 summit held in St. Petersburg had energy security on top of its agenda ahead of issues such as education, health, trade, environment, and terrorism.[9] The seventh sustainable development goal (SDG) by the United Nations—ensure access to affordable, reliable, sustainable, and modern energy—is all about energy security.[5]

Owing to a great degree of dependency upon oil and gas, the international energy market generally gauges the energy security in terms of secure supplies of these two commodities. The balance of evidence suggests that the present international energy market is vulnerable to different types of security risks. Energy security risks can be broadly classified into two categories: first, man-made— international geopolitical and geostrategic conflicts resulting into military attacks and wars, and sabotage and terrorist activities on relatively smaller scale; and second, natural— depletion of fossil fuel reserves, floods, fires, and earthquakes.

Risk Assessment

As with the supply of any commodity, the security of energy supply depends on the availability, size, and location of the energy reserves. The availability and the size of the reserves and their location then determine the supply line: the exploitation, transportation, and utilization arrangements. Conversely, the security of the exploitation, transportation, and utilization arrangements then determines the security of energy supplies. Using a very simple model, the major components of the risk can be identified:

- The number of potential disruptive events
- The probability that, given a potential disruptive event, the supply line will be successfully disrupted
- The potential consequences of such a disruption.

The contingency strategy could have three possible prongs. In order to minimize the risk, the number of potential disruptive events must be made as low as possible. Similarly, the probability that such an event will result in disruption must also be minimized. Finally, contingencies must be in place to successfully survive such a disruption.

The effects of human errors can be minimized by proper education and training of the operators, and by ergonomically suitable design. Energy companies have always strived to achieve this but failed spectacularly on a number of occasions, such as Windscale in the U.K., Chernobyl in the Ukraine, or Three Mile Island in the United States, and most recently Fukushima in Japan. However, based on past experience, the probability of human error can be estimated with relative accuracy. Determining the probability of sabotage or a military attack is difficult, since it depends upon various political and economic considerations. Nevertheless, it can be observed that such events are more frequent in unstable countries with many unresolved internal and external conflicts. Finally, the probability and severity of the natural events can be, once again, estimated for various locations, based on geographical, geological, and historical data.

In order to minimize the subsequent disruptions, or the effect of the potentially disruptive events, defense systems or protection and safeguards systems are employed. These systems have been well developed, and in order to maximize their effectiveness, they use the principles of defense in depth, redundancy, and diversity. All energy companies are geared to deal with human errors and natural events, but problems can be encountered when dealing with sabotage or military attacks. To minimize the impact of sabotage and military attacks, security services must be called for assistance. These services are generally able to deal with these problems domestically in countries without serious internal and external conflicts. However, in countries with serious internal and external conflicts, either the internal security forces can be ineffective, particularly in remote locations, or their loyalty can be divided. In these cases, the governments sometimes require security support from other countries, which, while improving the defenses, may make the internal conflicts worse and lead to an increasing frequency of attacks. Hence, stable countries, where the probability of an initial attack is already low, can, generally, protect their installations, but countries, with significant internal and external conflicts, where the probability of an initial attack is already high, cannot, generally, protect their installations.

All developed countries have contingency plans in place to deal with the effects of supply disruptions in the form of strategic sock reserves. These plans used to be based on the diversity of the energy supplies and the availability of strategic stocks. Due to the reliance on oil and natural gas by the developed countries, the principle of diversity of energy supplies holds less and less. All developed countries now deal with the potential problems primarily by having strategic reserves of oil supplies. By definition, the strategic reserves would be only available for certain activities and a certain period of time. This selectivity and time limit would provide a serious blow to the economies of the developed countries, if the reserves had to be used for a significant period of time.

Many of the world's leading oil-producing countries such as Iraq and Nigeria are politically unstable. The Middle East region as a whole has quite a volatile geopolitical situation and has experienced frequent conflicts over the last 100 years. There are serious reservations regarding security of oil; production and supply channels of many countries are regarded as the legitimate targets of radical elements because of various internal and external conflicts.

Getting oil from the well to the refinery and from there to the service station involves a complex transportation and storage system. Millions of barrels of oil are transported every day in tankers, pipelines, and trucks. This transportation system has always been a possible weakness of the oil industry, but it has become even more so in the present volatile geopolitical situation, especially in the Middle East region. The threats of global terrorism have made the equation more complex. Tankers and pipelines are quite vulnerable targets. There are approximately 4000 tankers employed, and each of them can be attacked in the high seas and more seriously while passing through narrow straits in hazardous areas. Pipelines, through which about 40% of the world's oil flows, are no less vulnerable, and due to their length, they are very difficult to protect. This makes pipelines potential targets for terrorists. In recent years, there have been an increased number of pipeline sabotages in different countries particularly Nigeria and Iraq, sending shockwaves in international energy markets.

Depleting Oil Reserves

A combination of constrained production capacity and growing fears of a rapid depletion of oil reserves in the world is also an important factor that has been playing its behind-the-scene role in pushing oil prices. In recent years, global oil infrastructure particularly with regard to extraction and refining capacity has been stretched to its limits. The production capacity of various oil-rich countries in the world, such as Iraq and Venezuela, has also been curtailed. The aspect of depleting oil reserves, despite its critical role, for various reasons, is not being publicly accepted by the market forces.

The world's ultimate conventional oil reserves are estimated at 2,000 billion barrels. This is the amount of production that would have been produced when production eventually ceases. The demand for oil has grown rapidly over the last few decades as shown in Table 2.[10] The surging demand for oil has already stretched the production to its limits—in mid-2002, there were more than 6 million barrels per day of excess production capacity, but by mid-2003, the daily excess capacity was below 2 million barrels, which further skewed to less than 1 million barrels by 2006.[11]

Different countries are at different stages of their reserve depletion curves. Some, such as the United States, are past their midpoint and are in terminal decline, whereas others are close to midpoint such as U.K. and Norway. However, the five major Gulf producers—Saudi Arabia, Iraq, Iran, Kuwait, and United Arab Emirates—are at an early stage of depletion and can exert a swing role, making up the difference between world demand and what others can supply.

The expert consensus is that the world's midpoint of reserve depletion will be reached when 1,000 billion barrels of oil have been produced—that is to say, half the ultimate reserves of 2,000 billion barrels. It is estimated that around 1,000 billion barrels have already been consumed and 1,000 billion barrels of proven oil reserves are left in the world.[12] According to BP statistical review of oil reserves in 2017, the reserve-to-production ratio for North America, South and Central America, Europe, Asia Pacific, Middle East, and Africa was 12, 48, 13, 32, 120, and 62, respectively. The reserve-to-production ratio for the whole world is reported to be equal to 52 years.[13] In the backdrop of a continuous growth in oil demand—as according to the U.S. Department of Energy, by 2025 the global daily demand could be as much as 110 barrels/day—the global reserves are actually going to run out much quicker than 42 years.[14]

A growing number of opinions among energy experts suggest that global oil production will probably peak sometime during this decade as indicated in Table 3.[15] A siren call regarding energy

TABLE 2 Growth in World Oil Demand

Year	World Population (million)	Average Daily Oil Demand (million barrels/day)	World Average per Capita Consumption (barrels/yr)
1965	3310	31.23	3.65
1968	3520	39.04	4.05
1971	3750	51.76	5.04
1974	3990	59.39	5.44
1977	4200	63.66	5.53
1980	4410	64.14	5.31
1983	4650	58.05	4.56
1986	4890	61.76	4.60
1989	5150	65.88	4.67
1992	5400	66.95	4.52
1995	5610	69.88	4.54
1998	5870	72.92	4.51
2001	6140	75.99	4.53
2004	6400	82.35	4.67
2007	6610	86.30	4.76
2010	6958	86.41	4.53
2013	7213	91.82	4.64
2016	7466	96.15	4.70

TABLE 3 Various Projections of the Global Oil Reserves and the Peak Year (billions of barrels)

Author	Affiliation	Year	Estimated Ultimate Reserves	Peak Year
Hubert	Shell	1969	2100	2000
Bookout	Shell	1989	2000	2010
Mackenzie	Researcher	1996	2600	2007–2019
Appleby	BP	1996		2010
Ivanhoe	Consultant	1996		2010
Edwards	University of Colorado	1997	2836	2020
Campbell	Consultant	1997	1800–2000	2010
Bernabe	ENI	1998		2005
Schollenberger	Amoco	1988		2015–2035
IEA	OECD	1998	2800	2010–2020
EIA	DOE	1998	4700	2030
Laherrere	Consultant	1999	2700	2010
USGS	International Department	2000	3270	
Salameh	Consultant	2000	2000	2004–2005
Deffeyes	Princeton University	2001	1800–2100	2004

security was also raised in a recent speech made by the British Ambassador to the United States, Sir David Manning. He eloquently puts forward the case thus: “The International Energy Agency predicts that, if we do nothing, the global oil demand will reach 121 million barrels per day by 2030, up from 85 million barrels today. That will require increasing production by 37 million barrels per day over the next 25 years, of which 25 million barrels per day has yet to be discovered. That is, we’ll have to find four petroleum systems that are each the size of the North Sea. Production from existing fields is dropping at about 5% per year.” Only one barrel of oil is now being discovered for every

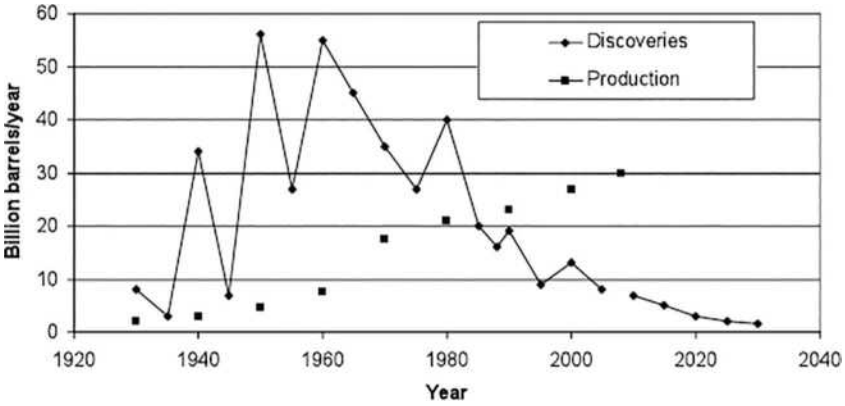


FIGURE 2 An overview of global oil discoveries and production.

four consumed as highlighted in Figure 2. Globally, the discovery rate of untapped oil peaked in the late 1960s. Over the past decade, oil production has been falling in 33 of the world's 48 largest oil-producing countries, including 6 of the 13 members of OPEC (Organization of the Petroleum Exporting Countries).[16]

Concerns over Security of Supplies from the Middle East

Oil and gas are jointly contributing to around 56% of the present global energy requirements. In terms of existence, both oil and gas are extremely localized by their very nature. Vast majority of their resources are found in relatively confined regions. For example, almost 88% of the world oil reserves exist within ten countries. Middle East is the oil headquarters of the world, holding almost 48% of the global oil as highlighted in Figure 3.[13] The situation with natural gas is even more intense. Statistics indicate that of the remaining known gas reserves in the world, equivalent to 6260 trillion cubic feet (tcf), around 57% are shared by only three countries: Russia, Iran, and Qatar, as shown in Figure 4.[17]

Middle East has been under the limelight for its vast oil reserves for nearly a century now. By the 1950s, it had established its standing as the oil-rich region of the world—its oil reserves were equal to the rest of the world's combined and double the U.S. reserves.[18] For the last six decades, Middle East

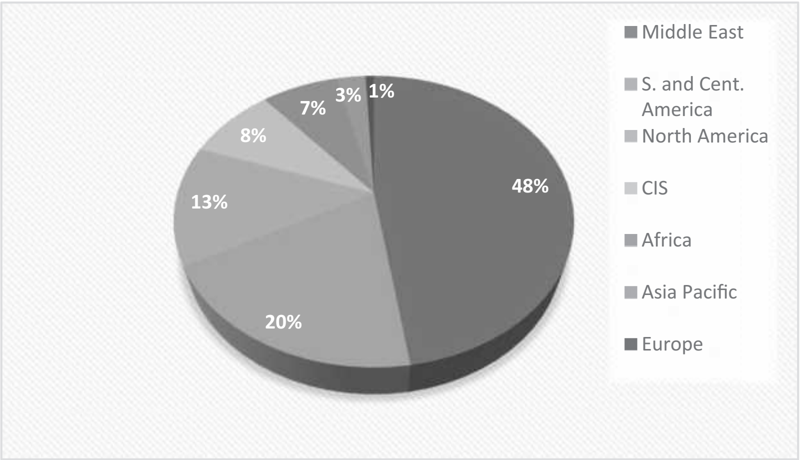


FIGURE 3 Remaining oil reserves in the world, 2017.

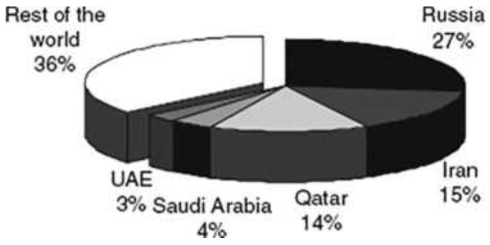


FIGURE 4 Remaining natural gas reserves in the world, 2009.

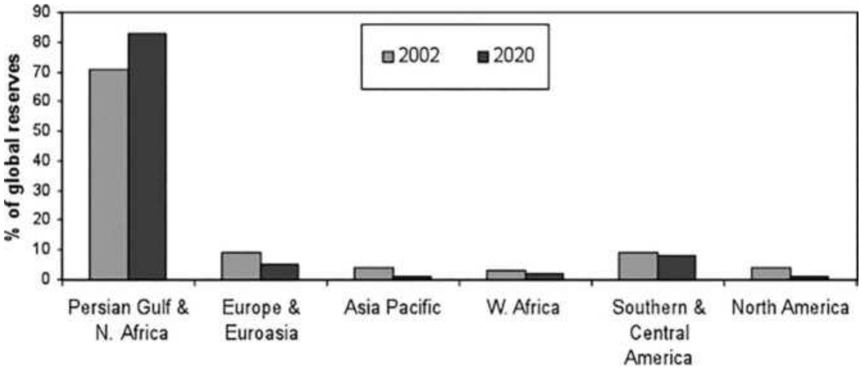


FIGURE 5 Growing share of Middle Eastern oil reserves.

has been the source of abundant and cheap oil that propelled the economic growth of the world over this period. In 2006, it produced about 28% of the world's total oil. During the same year, the region exported 18.2 million barrels of oil per day. Besides oil, the region also has huge reserves (2509 tcf) of natural gas, accounting for 41% of total proven world gas reserves.[19]

The importance of the Arabian Gulf countries is several-fold. The oil and gas reserves in non-Middle East countries are being depleted more rapidly than those of Middle East producers. If production continues at the present rate, many of the largest, non-Middle Eastern producers in 2002, such as Russia, Mexico, United States, Norway, and Brazil, will cease to be relevant players in the oil market in less than two decades. At that point, the Middle East will be the only major reservoir of abundant crude oil—within 20 years or so, about four-fifths of oil reserves could be in the hands of the Middle Eastern countries as shown in Figure 5.[11] The role of the region in future oil supplies of the world is thus going to become even more prominent.

Global Warming

Since the industrial revolution, human activities have resulted in a sharp rise in the concentration of GHGs in the Earth's atmosphere. The GHGs such as carbon dioxide (CO_2), nitrous oxide (N_2O), methane (CH_4), ozone (O_3), sulfur hexafluoride (SF_6), hydrofluorocarbons (HFCs), and perfluorocarbons (PFCs) are widely considered to be the root cause of a rapid increase in the atmospheric temperature, a phenomenon recognized as global warming. As a consequence of global warming, the climatic patterns of the earth are experiencing wide-ranging changes that are having numerous and far-reaching implications for the planet. The concentration levels of GHGs in the atmosphere are usually described in terms of carbon dioxide equivalent (CO_2e). Estimates suggest that since the advent of the industrial revolution, the level of GHGs in the atmosphere has grown from 280 parts per million (ppm) CO_2e to 430 ppm CO_2e.

Scientific circles believe that in order to avoid catastrophic implications, the level should be limited to 450–550 ppm CO_2e.[20] The growing concentration of GHGs in the atmosphere has increased the global average temperature. Some estimates suggest that the global temperature has increased by 0.8°C since the industrial revolution. The rise in the atmospheric temperature leads to the melting of glaciers and ice sheets, which, in turn, increases the sea level. According to the Intergovernmental Panel on Climate Change (IPCC), over the last century, the global sea level has increased at a rate of around 2 mm/yr. Estimates also indicate that the sea level rose twice as much between 1993 and 2003 as in the previous three decades.[21]

Global warming and climate change lead to a pattern of more frequent and more intense weather events such as floods, storms, droughts, heat waves, diseases, and loss of habitat. These events affect people across the world and result into huge casualties. According to the World Health Organization (WHO), as many as 160,000 people die each year from the side effects of climate change, and the numbers could almost double by 2020. These side effects range from malaria to malnutrition and diarrhea that follow in the wake of floods, droughts, and warmer temperatures.[22] Evidence suggests that the weather-related disasters have quadrupled over the last two decades. According to Oxfam, from an average of 120 disasters a year in the early 1980s, there are now as many as 500, with the rise to be attributed to unpredictable weather conditions caused by global warming. The year 2007 saw floods in South Asia, across the breadth of Africa and Mexico, which affected more than 250 million people. Devastating floods in 2010 affected a similar magnitude of people in China and Pakistan alone.[23] Statistics also suggest that over the last 25 years, the number of people affected by disasters has risen by 68%, from an average of 174 million a year from 1985 to 1994 to 254 million a year from 1995 to 2004.[24]

Climate change is responsible for huge economic consequences. Between the 1960s and the 1990s, the number of significant natural catastrophes such as floods and storms rose nine-fold, and the associated economic losses rose by a factor of nine. Figures indicate that the economic losses as a direct result of natural catastrophes over 5 years between 1954 and 1959 were $35 billion, while between 1995 and 1999, these losses were around $340 billion.[25] Natural catastrophes associated with global warming killed more than 190,000 people in 2004, twice as many as in 2003, with an economic cost of $145 billion. The August 2005 Hurricane Katrina was responsible for taking more than 1,000 human lives. Hurricane Katrina caused at least $125 billion in economic damage and could cost the insurance industry up to $60 billion in claims. That is significantly higher than the previous record-setting storm, Hurricane Andrew in 1992, which caused nearly $21 billion in insured losses in today's dollars.[26] Katrina shut down large portions of oil and gas production in the Gulf of Mexico at a time when worldwide energy output was already stretched thin. While the storm's impact was most acute in the United States, it also sent fuel costs higher around the globe, squeezing consumers in Europe and Asia.[27] It has been reported that since the advent of the 20th century, natural disasters such as floods, storms, earthquakes, and bushfires have resulted in an estimated loss of nearly 8 million lives and over $7 trillion of economic loss.[27]

The scale and intensity of global warming are set to increase in the 21st century. Global mean temperature is forecast to rise by between 1°C and 4.5°C by 2100, with best estimates somewhere between 2°C and 3°C. All projections produce rates of warming that are greater than those experienced in the last 10,000 years. Sea level is projected to rise by about 50 cm by 2100 (with a range of 20–90 cm).[11,28] Temperature and sea level changes will not be globally uniform. Land areas, particularly at high latitudes, will warm faster than the oceans, with a more vigorous hydrological cycle potentially affecting the rate and scale of various extreme events such as drought, flood, and rainfall. Impacts on natural and semi-natural ecosystems, agriculture, water resources, human infrastructure, and human health are subject to many uncertainties, but all will be subject to stresses that will exacerbate stresses from other sources such as land degradation, pollution, population growth and migration, and rising per capita exploitation of natural resources. Global warming is also set to have colossal ramifications on biodiversity—Stern Report suggests that a warming of even only 2°C could leave 15–40% species facing extinction.[20]

Global Warming Threats for Developing Countries

Although global warming is a threat to the whole planet, its intensity is not uniformly distributed. It is the low-lying and small island countries that are being hit harder. Most of them are actually poor and developing countries. It is feared that several hundred million people in densely populated coastal regions—particularly river deltas in Asia—are threatened by rising sea levels and the increasing risk of flooding. It is estimated that more than one-sixth of the world's population live in areas affected by water sources from glaciers and snow pack that will very likely disappear.[29] The number of people at risk of flooding by coastal storm surges is projected to increase from the current 75 million to 200 million by 2080, when sea levels may have risen by more than 1 ft. In the United Nation's list of countries under severe threat by global warming, Bangladesh is at the top. Being a low-lying and densely populated country, Bangladesh would be worst hit by any rise in the sea level. Coastal areas would experience erosion and inundation due to intensification of tidal action. A rise in seawater would enable saline water to intrude further inland during high tides. Destruction of agricultural land and loss of sweet water fauna and flora could also occur. The shoreline would retreat inland, causing changes in the coastal boundary and coastal configuration. The process will also shrink the land area of Bangladesh. Worst scenarios suggest that by the year 2050, one-third of the country could be under water, making more than 70 million people homeless.[30] For small island developing states (SIDSs), global warming poses an enormous set of challenges for their livelihood, safety, and security. Since most of the infrastructure in these countries is on the coast, the damage from consequent erosion and flooding is likely to be hugely burdensome for their already fragile economies. Owing to their smaller land area compared to other countries, they cannot afford to lose land due to surging sea level. For some, for example, Maldives, it threatens their very existence.

The health-related implications of global warming are also expected to be more formidable for developing countries. The WHO estimates that global warming is already causing about 5 million extra cases of severe illness a year. By 2030, however, the number of climate-related diseases is likely to more than double, with a dramatic increase in heat-related deaths caused by heart failure, respiratory disorders, the spread of infectious diseases, and malnutrition from crop failures. Countries with coastlines along the Indian and Pacific oceans and sub-Saharan Africa would suffer a disproportionate share of the extra health burden. According to WHO experts, many of the most important diseases in poor countries, such as diarrhea and malnutrition, are highly sensitive to climate. Also, that the health sector is already struggling to control these diseases and climate change threatens to undermine these efforts.[31]

One of the significant heartbreaks of global warming is that developed and industrialized countries are acutely responsible for the phenomenon, but the heavier price is to be paid by the poor and developing nations. For example, the average value of the per capita energy consumption—an index to measure the contribution towards global warming—in industrialized and developed countries is almost six times greater than that in developing countries. In 2007, the former U.S. Vice President Al Gore, in his famous work on global warming, heavily criticized the United States for turning a blind eye on the issue of global warming. According to him, the U.S. stance on the Kyoto Protocol is unfavorable despite the fact that it alone is responsible for more than 30% of the world GHG emissions. He concludes that a U.S. citizen emits nearly six times greater amount of carbon as compared to the world average emission as indicated in Figure 6, whereas compared to African or South Asian countries such as Pakistan, India, and Bangladesh, the per capita carbon emission in the United States is around 22 times higher.[32] Figure 7 provides a relationship between carbon dioxide emissions and gross national income (GNI) of a range of developing and developed countries. It can be seen that China has a substantially higher emission rate compared to other countries. Australia, Canada, and the United States also have a significantly higher emission rate in comparison with other developed countries. According to the World Wildlife Fund (WWF), in terms of eco-footprint, an index of sustainability, the excessive consumption of natural resources by the developed countries is imposing serious implications on the ecosystem of the planet.[33] These pieces of evidence indicate that the responsibility of the disturbance in the global ecosystem, also

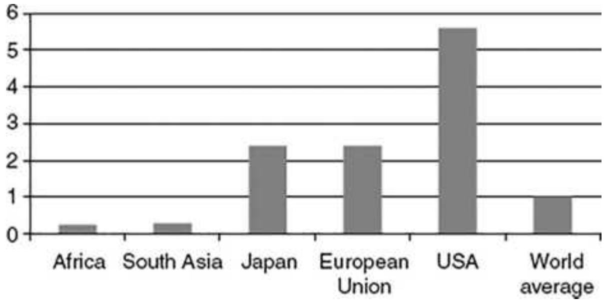


FIGURE 6 Annual per capita carbon emission in tons.

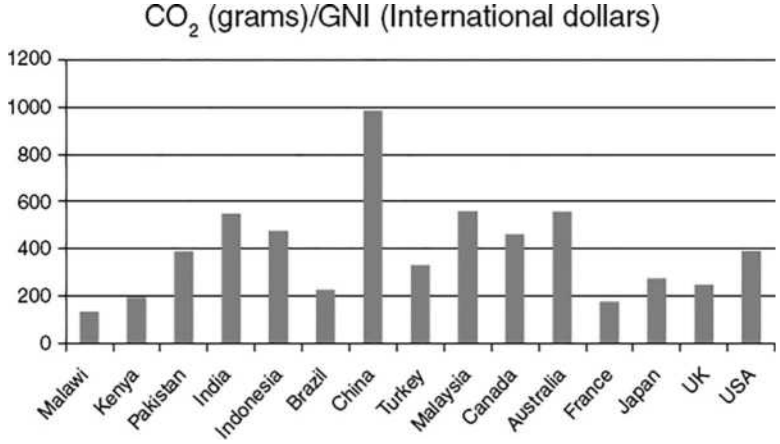


FIGURE 7 The per capita values of grams of carbon dioxide emitted for each international dollar of GNI in various countries.

leading to the phenomenon of global warming and climate change, to a great extent rests on the shoulders of the developed and industrialized countries of the world.

In the backdrop of the trajectory of the issue, the Kyoto Protocol urges the industrialized countries to reduce their collective emissions of GHGs by 5.2% compared to the year 1990. However, compared to the emission levels that would be expected by 2010 without the Protocol, this target represents a 29% cut. The goal is to lower overall emissions of six GHGs—carbon dioxide, methane, nitrous oxide, sulfur hexafluoride, HFCs, and PFCs—calculated as an average over the 5-year period of 2008–2012. National targets range from 8% reductions for the EU and some others to 7% for the United States and 6% for Japan. The developing countries of the world have been spared any emission reduction obligations not only in the Kyoto Protocol but also in its predecessor accords such as Agenda 21. According to Professor Jonathan Patz, it is incumbent on those countries bearing the greatest responsibility for climate change to show moral leadership. He goes on to say: "Those least able to cope and least responsible for the GHGs that cause global warming are most affected. Herein exists an enormous global ethical challenge."[34] On similar lines, Sir Nicholas Stern acknowledges that the developed countries should realize their responsibility towards addressing global warming. He concludes that climate change is the greatest and widest-ranging challenge mankind has ever faced. All countries will be affected by climate change, but the poorest countries will suffer earliest and most. Loss of biodiversity is another inevitable consequence—more than 40% of the species are likely to face extinction. He further urges that climate change should be fully integrated into development policy, and rich countries should honor pledges to increase support through overseas development assistance.[20]

To tackle global warming, a radical change in human attitude towards the environment and consumption of natural resources is required. A major shift in energy consumption practices—from the reliance on currently employed environmentally dangerous resources to the environmentally friendly ones—would be imperative to attain sustainable development. To safeguard the future of coming generations, the world thus has to move towards low-carbon energy systems.

Renewable Energy

Renewable energy as the name implies is the energy obtained from natural resources such as wind power, solar energy, hydropower, biomass energy, and geothermal energy. Renewable energy resources have also been important for humans since the beginning of civilization. Biomass, for example, has been used for heating, cooking, and steam production; wind has been used for moving ships; both hydropower and wind have been used for powering mills to grind grains. Renewable energy resources are abundant in nature and have the potential to provide energy services with zero or almost zero emissions. Renewable energy is acknowledged as a vital and plentiful source of energy. Technically, renewable energy resources have enormous potential and can meet many times the present world energy demand. However, due to their intermittent nature, they have to be used in conjunction with other energy resources. They can enhance diversity in energy supply markets, secure long-term sustainable energy supplies, and reduce local and global atmospheric emissions. They can also provide commercially attractive options to meet specific needs for energy services (particularly in developing countries and rural areas), create new employment opportunities, and offer possibilities for local manufacturing of equipment.

Renewable energy's favorable dimensions in terms of resources availability, reducing cost trends, and environmental friendliness are propelling its rapid growth.[35,36] Technological advancements and supportive policy frameworks have also played a vital role in the rapid growth of renewable energy application around the world.[37–39] In 2017, renewable energy technologies received a cumulative investment of $280 billion. At the end of the year, the global installed capacity of renewables stood at 2,195 GW. Renewables contributed to over 26% of the world's total electricity supplies.[40] Overall, renewables accounted for an estimated 70% of net additions to global power capacity in 2017. Solar photovoltaic (PV) made up led the renewables march, making up nearly 55% of the newly installed renewable power capacity in the year. Wind power and hydropower with respective shares of 29% and 11% account for most of the remaining capacity additions. Renewable energy technologies are becoming competitive with conventional power systems—in 2017, solar PV and wind power projects received bids as low as $30/MWh. Since 2010, the global weighted average levelized cost of electricity (LCOE) of utility-scale solar PV has dropped by over 73%.[35]

Conclusions

The world faces stringent energy and environmental challenges. Fossil fuels, contributing to more than 80% of global energy supplies, interlink the security of energy and environment. The present energy scenario has a number of concerns including depletion of fossil fuel reserves, surging energy prices, geopolitical conflicts in fossil-fuel-rich regions, and emission of GHGs. The emission of GHGs leads to environmental insecurity by contributing to global warming. The intensity of the global warming concerns can be gauged from the fact that the average atmospheric temperature is forecasted to rise by as much as 4.5°C during the 21st century compared to a 0.8°C increment since the industrial revolution. Global warming and the consequent sea level rise—which could be up to 50 cm over the same period—would have catastrophic implications for natural and seminatural ecosystems, agriculture, water resources, human infrastructure, and human health. Even the very existence of some of the small island countries is under threat. Energy and environmental security is thus absolutely vital for sustainable development. The world needs to shift to energy resources that are more reliable, affordable, secure, and environmentally friendly. Renewable energy is one possible solution that can significantly help address the energy and environmental security concern.

References

1. Solomon, B. Economic geography of energy. In Cleveland, C., *Encyclopedia of Energy*; Amsterdam: Elsevier, 2004; Vol. 2.
2. Asif, M. Energy Crisis in Pakistan: Origins, Challenges and Sustainable Solutions; Karachi: Oxford University Press, 2011.
3. BP Energy Outlook, 2019 Edition, British Petroleum, UK.
4. Weynand, G. Energy Sector Assessment for US Aid/Pakistan. United States Agency for International Development, June 2007.
5. Sustainable Development Goals, United Nations, 2019, https://www.un.org/sustainabledevelopment/energy/ (accessed on 25 March 2019).
6. Poverty, Energy and Society, Energy Forum. The Baker Institute, Rice University.
7. Energy Services for the Millennium Development Goals, UNDP/World Bank, 2006.
8. Muller-Kraenner, S. *Energy Security*; London: Earthscan, 2008.
9. Issues and Instruments, G8 Summit 2006, St Petersburg, 15–17 July 2006.
10. McKillop, A. Oil Shock and Energy Transition. POGEE Conference, Karachi, May 2008.
11. Asif, M.; Muneer, T. Energy Supply, its Demand and Security Issues for Developed and Emerging Economies. *Renewable and Sustainable Energy Reviews* 2007, 11 (7): 1388–1413.
12. Asif, M.; Currie, J.; Muneer, T. The Role of Renewable and Non-renewable Sources for Meeting Future U.K. Energy Needs. International Journal *of* Nuclear Governance, Economy *and* Ecology 2007, 1 (4): 372–383.
13. BP Statistical Review of World Energy 2018, British Petroleum, UK.
14. International Energy Outlook 2006.
15. Salameh, M. Oil Crises: Historic Perspective. In Cleveland, C., *Encyclopaedia of Energy*; Amsterdam: Elsevier, 2004; Vol. 4.
16. British Ambassador Sir David Manning Discusses Energy: A Burning Issue for Foreign Policy at Stanford University. British Embassy, Washington, DC, March 2006. https://fsi.stanford.edu/multimedia/energy-burning-issue-foreign-policy (accessed 15 May 2020).
17. International Energy Agency Statistics 2009, International Energy Agency, Washington DC, USA.
18. Chomsky, N. The Fateful Triangle: The United States Israel and The Palestinians; London: Pluto Press, 1999.
19. Persian Gulf Region, Energy Information Administration, EIA, 2008, available at http://www.eia.doe.gov/emeu/cabs/Persian_Gub7Background.html (accessed June 2011).
20. Stern Report: The Key Points. *The Guardian*, 30 Hilary Osborne, October 2006.
21. Dodds, F.; Higham, A.; Sherman, R. *Climate Change and Energy Insecurity*; UK: Earthscan, 2009.
22. Asif, M.; Muneer, T. Energy Supply, its Demand and Security Issues for Developed and Emerging Economies. Renewable and Sustainable Energy Reviews 2007, 11 (7): 1388–1413.
23. Available at http://en.wikipedia.org/wiki/2010_China_floods (accessed June 2011). China Flood, Wikipedia, Available at http://en.wikipedia.org/wiki/2010_China_floods (accessed 15 May 2020).
24. John Sinnott, Disasters quadruple over last 20 years: Oxfam, Reuters, 25 November 2007, https://www.reuters.com/article/us-britain-climate-oxfam-idUSL2518480220071125 (accessed: 15 May 2020)
25. Muneer, T.; Asif, M. Generation and Transmission Prospects for Solar Electricity: U.K. and Global Markets. *Energy Conversion and Management* 2003, 44: 35–52.
26. Available at http://www.usatoday.com/money/economy/2005-09-09-katrina-damage_x.htmS (accessed June 2011).
27. Khan, H.; Asif, M. Impact of Green Roof and Orientation on the Energy Performance of Buildings: A Case Study from Saudi Arabia. *Sustainability* 2017, 9(4): 640. doi:10.3390/su9040640
28. Intergovernmental Panel on Climate Change, IPCC Working Group I (scenario IS92a), available at http://www.ipcc.ch/ipccreports/sres/regional/140.htm (accessed June 2011).

29. IPCC,2007:Climate Change 2007: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, M.L. Parry O.F. Canziani, J.P. Palutikof, P.J. van der Linden and C.E. Hanson, Eds., Cambridge University Press, Cambridge, UK, 976.
30. UNDP. 2007/8. Human Development Report 2007/8: Fighting climate change - Human solidarity in a divided world. New York.
31. Connor, S. Climate Change Map Reveals Countries Most Under Threat, The Independent, 17 November 2005.
32. Guggenheim, D. An Inconvenient Truth, Documentary on Former U.S. Vice President Al Gore's Campaign on Environment, 2006.
33. Living Planet Report 2004, WWF. Living Planet Report 2004, WWF. http://wwf.panda.org/knowledge_hub/all_publications/living_planet_report_timeline/lpr_2004/ (accessed 15 May 2020).
34. Jonathan A. Patz, Diarmid Campbell-Lendrum, Tracey Holloway & Jonathan A. Foley, Impact of regional climate change on human health, Nature volume 438, 310–317, 2005.
35. Asif, M. Growth and Sustainability Trends in the GCC Countries with Particular Reference to KSA and UAE. *Renewable & Sustainable Energy Reviews Journal* 2016, 55: 1267–1273.
36. Dehwah, A.; Asif, M. Assessment of Net Energy Contribution to Buildings by Rooftop PV Systems in Hot-Humid Climates. *Renewable Energy* 2019, 131: 1288–1299, https://doi.org/10.1016/j.renene.2018.08.031
37. Asif, M.; Hassanain, M.; Nahiduzzaman, Kh.; Swalha, H. Techno-Economic Assessment of Application of Solar PV in Building Sector-A Case Study from Saudi Arabia. *Smart and Sustainable Built Environment* 2019, 8 (1): 34–52.
38. Dehwah, A.; Asif, M.; Tauhidurrahman, M. Prospects of PV Application in Unregulated Building Rooftops in Developing Countries: A Perspective from Saudi Arabia. *Energy and Buildings* 2018, 171: 76–87.
39. Khan, M.; Asif, M.; Stach, S. Rooftop PV Potential in the Residential Sector of the Kingdom of Saudi Arabia. *Buildings* 2017, 7(2): 46. doi:10.3390/buildings7020046.
40. Renewables 2018 Global Status Report, Renewable Energy Policy Network for the 21st Century, Paris.

7

Energy Commissioning: New Buildings

Janey Kaster

Introduction

This entry provides an overview of commissioning—the processes one employs to optimize the performance characteristics of a new facility being constructed. Commissioning is important to achieve customer satisfaction, optimal performance of building systems, cost containment, and energy efficiency, and it should be understood by contractors and owners.

After providing an overview of commissioning and its history and prevalence, this entry discusses what systems should be part of the commissioning process, the benefits of commissioning, how commissioning is conducted, and the individuals and teams critical for successful commissioning. Then the entry provides a detailed discussion of each of the different phases of a successful commissioning process, followed by a discussion of the common mistakes to avoid and how one can measure the success of a commissioning effort, together with a cost-benefit analysis tool.

The purpose of this entry will be realized if its readers decide that successful commissioning is one of the most important aspects of construction projects and that commissioning should be managed carefully and deliberately throughout any project, from predesign to postacceptance. As an introduction to those unfamiliar with the process and as a refresher for those who are, the following section provides an overview of commissioning, how it developed, and its current prevalence today.

Overview of Commissioning

Commissioning Defined

Commissioning is the methodology for bringing to light design errors, equipment malfunctions, and improper control strategies at the most cost-effective time to implement corrective action. Commissioning facilitates a thorough understanding of a facility's intended use and ensures that the design meets the intent through coordination, communication, and cooperation of the design and installation team. Commissioning ensures that individual components function as a cohesive system. For these reasons, commissioning is best when it begins in the predesign phase of a construction project and can in one sense be viewed as the most important form of quality assurance for construction projects.

Unfortunately, there are many misconceptions associated with commissioning, and perhaps for this reason, commissioning has been executed with varying degrees of success, depending on the level of understanding of what constitutes a "commissioned" project. American Society of Heating, Refrigerating and Air-Conditioning Engineers (AHSRAE) guidelines define commissioning as: the process of ensuring that systems are designed, installed, functionally tested, and capable of being operated and maintained to perform conformity with the design intent … [which] begins with planning and includes design, construction, startup, acceptance, and training, and is applied throughout the life of the building.[4] However, for many contractors and owners, this definition is simplified into the process of system startup and checkout or completing punch-list items.

Of course, a system startup and checkout process carried out by a qualified contractor is one important aspect of commissioning. Likewise, construction inspection and the generation and completion of punch-list items by a construction manager are other important aspects of commissioning. However, it takes much more than these standard installation activities to have a truly "commissioned" system. Commissioning is a comprehensive and methodical approach to the design and implementation of a cohesive system that culminates in the successful turnover of the facility to maintenance staff trained in the optimal operation of those systems.

Without commissioning, a contractor starts up the equipment but doesn't look beyond the startup to system operation. Assessing system operation requires the contractor to think about how the equipment will be used under different conditions. As one easily comprehended example, commissioning requires the contractor to think about how the equipment will operate as the seasons change. Analysis of the equipment and building systems under different load conditions due to seasonal conditions at the time of system startup will almost certainly result in some adjustments to the installed equipment for all but the most benign climates. However, addressing this common requirement of varying load due to seasonal changes most likely will not occur without commissioning. Instead, the maintenance staff is simply handed a building with minimal training and left to figure out how to achieve optimal operation on their own. In this seasonal example, one can just imagine how pleased the maintenance staff would be with the contractor when a varying load leads to equipment or system failure—often under very hot or very cold conditions!

Thus, the primary goal of commissioning is to achieve optimal building systems performance. For heating, ventilation, and air-conditioning (HVAC) systems, optimal performance can be measured by thermal comfort, indoor air quality, and energy savings. Energy savings, however, can result simply from successful commissioning targeted at achieving thermal comfort and excellent indoor air quality. Proper commissioning will prevent HVAC system malfunction—such as simultaneous heating and cooling, and overheating or overcooling—and successful malfunction prevention translates directly into energy savings. Accordingly, energy savings rise with increasing comprehensiveness of the commissioning plan. Commissioning enhances energy performance (savings) by ensuring and maximizing the performance of specific energy efficiency measures and correcting problems causing excessive energy use.[3] Commissioning, then, is the most cost-effective means of improving energy efficiency in commercial buildings. In the next section, the two main types of commissioning in use today—acceptance-based and processed-based—are compared and contrasted.

Acceptance-Based vs. Process-Based Commissioning

Given the varied nature of construction projects, contractors, owners, buildings, and the needs of the diverse participants in any building projection, commissioning can of course take a variety of forms. Generally, however, there are two types of commissioning: acceptance-based and process-based. Process-based commissioning is a comprehensive process that begins in the predesign phase and continues through postacceptance, while acceptance-based commissioning, which is perceived to be the cheaper method, basically examines whether an installation is compliant with the design and accordingly achieves more limited results.

Acceptance-based commissioning is the most prevalent type due to budget constraints and the lack of hard cost/benefit data to justify the more extensive process-based commissioning. Acceptance-based commissioning does not involve the contractor in the design process but simply constitutes a process to ensure that the installation matches the design. In acceptance-based commissioning, confrontational relationships are more likely to develop between the commissioning agent and the contractor because the commissioning agent and the contractor, having been excluded from the design phase, have not "bought in" to the design and thus may be more likely to disagree in their interpretation of the design intent.

Because the acceptance-based commissioning process simply validates that the installation matches the design, installation issues are identified later in the cycle. Construction inspection and regular commissioning meetings do not occur until late in the construction/installation phase with acceptance-based commissioning. As a result, there is no early opportunity to spot errors and omissions in the design, when remedial measures are less costly to undertake and less likely to cause embarrassment to the designer and additional costs to the contractor. As most contractors will readily agree, addressing issues spotted in the design or submittal stages of construction is typically much less costly than addressing them after installation, when correction often means tearing out work completed and typically delays the completion date.

Acceptance-based commissioning is cheaper, however, at least on its face, being approximately 80% of the cost of process-based commissioning.[2] If only the initial cost of commissioning services is considered, many owners will conclude that this is the most cost-effective commissioning approach. However, this 20% cost differential does not take into account the cost of correcting defects after the fact that process-based commissioning could have identified and corrected at earlier stages of the project. One need encounter only a single, expensive-to-correct project to become a devotee of process-based commissioning.

Process-based commissioning involves the commissioning agent in the predesign through the construction, functional testing, and owner training. The main purpose is quality assurance—assurance that the design intent is properly defined and followed through in all phases of the facility life cycle. It includes ensuring that the budget matches the standards that have been set forth for the project so that last-minute "value engineering" does not undermine the design intent, that the products furnished and installed meet the performance requirements and expectation compliant with the design intent, and that the training and documentation provided to the facility staff equip them to maintain facility systems true to the design intent.

As the reader will no doubt already appreciate, the author believes that process-based commissioning is far more valuable to contractors and owners than acceptance-based commissioning. Accordingly, the remainder of this entry will focus on process-based commissioning, after a brief review of the history of commissioning from inception to date, which demonstrates that our current, actively evolving construction market demands contractors and contracting professionals intimately familiar with and expert in conducting process-based commissioning.

History of Commissioning

Commissioning originated in the early 1980s in response to a large increase in construction litigation. Owners were dissatisfied with the results of their construction projects and had recourse only to the courts and litigation to resolve disputes that could not be resolved by meeting directly with their contractors. While litigation attorneys no doubt found this satisfactory approach to resolving construction project issues, owners did not, and they actively began looking for other means to gain assurance that they were receiving systems compliant with the design intent and with the performance characteristics and quality specified. Commissioning was the result.

While commissioning enjoyed early favor and wide acceptance, the recession of the mid-1980s placed increasing market pressure on costs, and by the mid-to late 1980s it forced building professionals to reduce fees and streamline services. As a result, acceptance-based commissioning became the norm, and process-based commissioning became very rare. This situation exists in most markets today; however, the increasing cost of energy, the growing awareness of the global threat of climate change and the need to reduce CO2 emissions as a result, and the legal and regulatory changes resulting from both are creating a completely new market in which process-based commissioning will become ever more important, as discussed in the following section.

Prevalence of Commissioning Today

There are varying degrees of market acceptance of commissioning from state to state. Commissioning is in wide use in California and Texas, for example, but it is much less widely used in many other states. The factors that impact the level of market acceptance depend upon

- The availability of commissioning service providers
- State codes and regulations
- Tax credits
- Strength of the state's economy[1]

State and federal policies with regard to commissioning are changing rapidly to increase the demand for commissioning. Also, technical assistance and funding are increasingly available for projects that can serve as demonstration projects for energy advocacy groups. The owner should investigate how each of these factors could benefit the decision to adopt commissioning in future construction projects.

Some of the major initiatives driving the growing market acceptance of commissioning are:

- Federal government's U.S. Energy Policy Act of 1992 and Executive Order 12902, mandating that federal agencies develop commissioning plans
- Portland Energy Conservation, Inc.; National Strategy for Building Commissioning; and their annual conferences
- ASHRAE HVAC Commissioning Guidelines (1989)
- Utilities establishing commissioning incentive programs
- Energy Star building program
- Leadership in Energy Environmental Design (LEED) certification for new construction
- Building codes
- State energy commission research programs

Currently, the LEED is having the largest impact in broadening the acceptance of commissioning. The Green Building Council is the sponsor of LEED and is focused on sustainable design—design and construction practices that significantly reduce or eliminate the cradle-to-grave negative impacts of buildings on the environment and building occupants. Leadership in energy efficient design encourages sustainable site planning, conservation of water and water efficiency, energy efficiency and renewable energy, conservation of materials and resources, and indoor environmental quality.

With this background on commissioning, the various components of the commissioning process can be explored, beginning with an evaluation of what building systems should be subject to the commissioning process.[5]

Commissioning Process

Systems to Include in the Commissioning Process

The general rule for including a system in the commissioning process is: the more complicated the system is the more compelling is the need to include it in the commissioning process. Systems that are required to integrate or interact with other systems should be included. Systems that require specialized trades working independently to create a cohesive system should be included, as well as systems that are critical to the operation of the building. Without a commissioning plan on the design and construction of these systems, installation deficiencies are likely to create improper interaction and operation of system components.

For example, in designing a lab, the doors should be included in the commissioning process because determining the amount of leakage through the doorways could prove critical to the ability to maintain critical room pressures to ensure proper containment of hazardous material. Another common example is an energy retrofit project. Such projects generally incorporate commissioning as part of the measurement and verification plan to ensure that energy savings result from the retrofit process.

For any project, the owner must be able to answer the question of why commissioning is important.

Why Commissioning?

A strong commissioning plan provides quality assurance, prevents disputes, and ensures contract compliance to deliver the intended system performance. Commissioning is especially important for HVAC systems that are present in virtually all buildings because commissioned HVAC systems are more energy efficient.

The infusion of electronics into almost every aspect of modern building systems creates increasingly complex systems requiring many specialty contractors. Commissioning ensures that these complex subsystems will interact as a cohesive system.

Commissioning identifies design or construction issues and, if done correctly, identifies them at the earliest stage in which they can be addressed most cost effectively. The number of deficiencies in new construction exceeds existing building retrofit by a factor of 3.[3] Common issues that can be identified by commissioning that might otherwise be overlooked in the construction and acceptance phase are: air distribution problems (these occur frequently in new buildings due to design capacities, change of space utilization, or improper installation), energy problems, and moisture problems.

Despite the advantages of commissioning, the current marketplace still exhibits many barriers to adopting commissioning in its most comprehensive and valuable forms.

Barriers to Commissioning

The general misperception that creates a barrier to the adoption of commissioning is that it adds extra, unjustified costs to a construction project. Until recently, this has been a difficult perception to combat because there are no energy-use baselines for assessing the efficiency of a new building. As the cost of energy continues to rise, however, it becomes increasingly less difficult to convince owners that commissioning is cost effective. Likewise, many owners and contractors do not appreciate that commissioning can reduce the number and cost of change orders through early problem identification. However, once the contractor and owner have a basis on which to compare the benefit of resolving a construction issue earlier as opposed to later, in the construction process, commissioning becomes easier to sell as a win–win proposal.

Finding qualified commissioning service providers can also be a barrier, especially in states where commissioning is not prevalent today. The references cited in this entry provide a variety of sources for identifying associations promulgating commissioning that can provide referrals to qualified commissioning agents.

For any owner adopting commissioning, it is critical to ensure acceptance of commissioning by all of the design construction team members. Enthusiastic acceptance of commissioning by the design team will have a very positive influence on the cost and success of your project. An objective of this entry is to provide a source of information to help gain such acceptance by design construction team members and the participants in the construction market.

Selecting the Commissioning Agent

Contracting an independent agent to act on behalf of the owner to perform the commissioning process is the best way to ensure successful commissioning. Most equipment vendors are not qualified and are likely to be biased against discovering design and installation problems—a critical function of the commissioning agent—with potentially costly remedies. Likewise, systems integrators have the background in control systems and data exchange required for commissioning but may not be strong in mechanical design, which is an important skill for the commissioning agent. Fortunately, most large mechanical consulting firms offer comprehensive commissioning services, although the desire to be competitive in the selection processes sometimes forces these firms to streamline their scope on commissioning.

Owners need to look closely at the commissioning scope being offered. An owner may want to solicit commissioning services independently from the selection of the architect/mechanical/electrical/plumbing design team or, minimally, to request specific details on the design team's approach to commissioning. If an owner chooses the same mechanical, electrical, and plumbing (MEP) firm for design and commissioning, the owner should ensure that there is physical separation between the designer and commissioner to ensure that objectivity is maintained in the design review stages. An owner should consider taking on the role of the commissioning agent directly, especially if qualified personnel exist in-house. This approach can be very cost effective. The largest obstacles to success with an in-house commissioning agent are the required qualifications and the need to dedicate a valuable resource to the commissioning effort. Many times, other priorities may interfere with the execution of the commissioning process by an in-house owner's agent.

There are three basic approaches to selecting the commissioning agent:

Negotiated—best approach for ensuring a true partnership
Selective bid list—preapproved list of bidders
Competitive—open bid list

Regardless of the approach, the owner should clearly define the responsibilities of the commissioning agent at the start of the selection process. Fixed-cost budgets should be provided by the commissioning agent to the owner for the predesign and design phases of the project, with not-to-exceed budgets submitted for the construction and acceptance phases. Firm service fees should be agreed upon as the design is finalized.

Skills of a Qualified Commissioning Agent

A commissioning agent needs to be a good communicator, both in writing and verbally. Writing skills are important because documentation is critical to the success of the commissioning plan. Likewise, oral communication skills are important because communicating issues uncovered in a factual and nonaccusatory manner is most likely to resolve those issues efficiently and effectively. The commissioning agent should have practical field experience in MEP controls design and startup to be able to identify potential issues early. The commissioning agent likewise needs a thorough understanding of how building structural design impacts building systems. The commissioning agent must be an effective facilitator and must be able to decrease the stress in stressful situations. In sum, the commissioning agent is the cornerstone of the commissioning team and the primary determinant of success in the commissioning process.

At least ten organizations offer certifications for commissioning agents. However, there currently is no industry standard for certifying a commissioning agent. Regardless of certification, the owner should carefully evaluate the individuals to be performing the work from the commissioning firm selected. Individual experience and reputation should be investigated. References for the lead commissioning agent are far more valuable than references for the executive members of a commissioning firm in evaluating potential commissioning agents. The commissioning agent selected will, however, only be one member of a commissioning team, and the membership of the commissioning team is critical to successful commissioning.

Commissioning Team

The commissioning team is composed of representatives from all members of the project delivery team: the commissioning agent, representatives of the owner's maintenance team, the architect, the MEP designer, the construction manager, and systems contractors. Each team member is responsible for a particular area of expertise, and one important function of the commissioning agent is to act as a facilitator of intrateam communication.

The maintenance team representatives bring to the commissioning team the knowledge of current operations, and they should be involved in the commissioning process at the earliest stage, defining the design intent in the predesign phase, as described below. Early involvement of maintenance team representatives ensures a smooth transition from construction to a fully operational facility, and aids in the acceptance and full use of the technologies and strategies that have been developed during the commissioning process. Involvement of the maintenance team representatives also shortens the building turnover transition period.

The other members of the commissioning team have defined and important functions. The architect leads the development of the design intent document (DID). The MEP designer's responsibilities are to develop the mechanical systems that support the design intent of the facility and comply with the owner's current operating standards. The MEP schematic design is the basis for the systems installed and is discussed further below. The construction manager ensures that the project installation meets the criteria defined in the specifications, the budget requirements, and the predefined schedule. The systems contractors' responsibilities are to furnish and install a fully functional system that meets the design specifications. There are generally several contractors whose work must be coordinated to ensure that the end product is a cohesive system.

Once the commissioning team is in place, commissioning can take place, and it occurs in defined and delineated phases—the subject of the following section.

Commissioning Phases

The commissioning process occurs over a variety of clearly delineated phases. The commission plan is the set of documents and events that defines the commissioning process over all phases. The commissioning plan needs to reflect a systematic, proactive approach that facilitates communication and cooperation of the entire design and construction team.

The phases of the commissioning process are:

Predesign
Design
Construction/installation
Acceptance
Postacceptance

These phases and the commissioning activities associated with them are described in the following sections.

Predesign Phase

The predesign phase is the phase in which the design intent is established in the form of the DID. In this phase of a construction project, the role of commissioning in the project is established if process-based commissioning is followed. Initiation of the commissioning process in the predesign phase increases acceptance of the commissioning process by all design team members. Predesign discussions about commissioning allow all team members involved in the project to assess and accept the importance of commissioning to a successful project. In addition, these discussions give team members more time to assimilate the impact of commissioning on their individual roles and responsibilities in the project. A successful project is more likely to result when the predesign phase is built around the concept of commissioning instead of commissioning's being imposed on a project after it has been designed.

Once an owner has decided to adopt commissioning as an integral part of the design and construction of a project, the owner should be urged to follow the LEED certification process, as discussed above. The commissioning agent can assist in the documentation preparation required for the LEED certification, which occurs in the postacceptance phase.

The predesign phase is the ideal time for an owner to select and retain the commissioning agent. The design team member should, if possible, be involved in the selection of the commissioning agent because that member's involvement will typically ensure a more cohesive commissioning team. Once the commissioning agent is selected and retained, the commissioning-approach outline is developed. The commissioning-approach outline defines the scope and depth of the commissioning process to be employed for the project. Critical commissioning questions are addressed in this outline. The outline will include, for most projects, answers to the following questions:

What equipment is to be included?
What procedures are to be followed?
What is the budget for the process?

As the above questions suggest, the commissioning budget is developed from the choices made in this phase. Also, if the owner has a commissioning policy, it needs to be applied to the specifics of the particular project in this phase.

The key event in the predesign phase is the creation of the DID, which defines the technical criteria for meeting the requirements of the intended use of the facilities. The DID document is often created based in part upon the information received from interviews with the intended building occupants and maintenance staff. Critical information—such as the hours of operation, occupancy levels, special environmental considerations (such as pressure and humidity), applicable codes, and budgetary considerations and limitations—is identified in this document. The owner's preference, if any, for certain equipment or contactors should also be identified at this time. Together, the answers to the critical questions above and the information in the DID are used to develop the commissioning approach outline. A thorough review of the DID by the commissioning agent ensures that the commissioning-approach outline will be aligned with the design intent.

With the commissioning agent selected, the DID document created, and the commissioning approach outline in place, the design phase is ready to commence.

Design Phase

The design phase is the phase in which the schematics and specifications for all components of a project are prepared. One key schematic and set of specifications relevant to the commissioning plan is the MEP schematic design, which specifies installation requirements for the MEP systems. As noted, the DID is the basis for creating the commissioning approach outline in the predesign phase. The DID also serves as the basis for creating the MEP schematic design in the design phase. The DID provides the MEP designer with the key concepts from which the MEP schematic design is developed.

The completed MEP schematic design is reviewed by the commissioning agent for completeness and conformance to the DID. At this stage, the commissioning agent and the other design team members should consider what current technologies, particularly those for energy efficiency, could be profitably included in the design. Many of the design enhancements currently incorporated into existing buildings during energy retrofitting for operational optimization are often not considered in new building construction. This can result in significant lost opportunity, so these design enhancements should be reviewed for incorporation into the base design during this phase of the commissioning process. This point illustrates the important principle that technologies important to retrocommissioning should be applied to new building construction—a point that is surprisingly often overlooked in the industry today.

For example, the following design improvements and technologies should always be considered for applicability to a particular project:

- Variable-speed fan and pumps installed
- Chilled water cooling (instead of DX cooling)
- Utility meters for gas, electric, hot water, chilled water, and steam at both the building and system level
- CO_2 implementation for minimum indoor air requirements

This list of design improvements is not exhaustive; the skilled commissioning agent will create and expand personalized lists as experience warrants and as the demands of particular projects suggest.

In addition to assisting in the evaluation of potential design improvements, the commissioning agent further inspects the MEP schematic design for:

- Proper sizing of equipment capacities
- Clearly defined and optimized operating sequences
- Equipment accessibility for ease of servicing

Once the commissioning agent's review is complete, the feedback is discussed with the design team to determine whether its incorporation into the MEP schematic design is warranted. The agreed-upon changes or enhancements are incorporated, thus completing the MEP schematic design.

The completed MEP schematic design serves as the basis on which the commissioning agent will transform the commissioning-approach outline into the commissioning specification.

The commissioning specification is the mechanism for binding contractually the contractors to the commissioning process. Expectations are clearly defined, including:

- Responsibilities of each contractor
- Site meeting requirements
- List of the equipment, systems, and interfaces
- Preliminary verification checklists
- Preliminary functional-performance testing checklists
- Training requirements and who is to participate
- Documentation requirements

- Postconstruction documentation requirements
- Commissioning schedule
- Definition for system acceptance
- Impact of failed results

Completion of the commissioning specification is required to select the systems contractor in a competitive solicitation. Alternatively, however, owners with strong, preexisting relationships with systems contractors may enter into a negotiated bid with those contractors, who can then be instrumental in finalizing the commissioning specification.

Owners frequently select systems contractors early in the design cycle to ensure that the contractors are involved in the design process. As noted above, if there are strong, preexisting relationships with systems contractors, early selection without a competitive selection process (described in the following paragraph) can be very beneficial. However, if there is no competitive selection process, steps should be taken to ensure that the owner gets the best value. For example, unit pricing should be negotiated in advance to ensure that the owner is getting fair and reasonable pricing. The commissioning agent and the MEP designer can be good sources for validating the unit pricing. The final contract price should be justified with the unit pricing information.

If the system selection process is competitive, technical proposals should be requested with the submission of the bid price. The systems contractors need to demonstrate a complete understanding of the project requirements to ensure that major components have not been overlooked. Information such as the project schedule and manpower loading for the project provide a good basis from which to measure the contractor's level of understanding. If the solicitation does not have a preselected list of contractors, the technical proposal should include the contractor's financial information, capabilities, and reference lists. As in the negotiated process described above, unit pricing should be requested to ensure the proper pricing of project additions and deletions. The review of the technical proposals should be included in the commissioning agent's scope of work.

A mandatory prebid conference should be held to walk the potential contractors through the requirements and to reinforce expectations. This conference should be held regardless of the approach—negotiated or competitive— used for contractor selection. The contractor who is to bear the financial burden for failed verification tests and subsequent functional-performance tests should be reminded of these responsibilities to reinforce their importance in the prebid meeting. The prebid conference sets the tone of the project and emphasizes the importance of the commissioning process to a successful project.

Once the MEP schematic design and commissioning specification are complete, and the systems contractors have been selected, the construction/installation phase begins.

Construction/Installation Phase

Coordination, communication, and cooperation are the keys to success in the construction and installation phase. The commissioning agent is the catalyst for ensuring that these critical activities occur throughout the construction and installation phase.

Frequently, value engineering options are proposed by the contractors prior to commencing the installation. The commissioning agent should be actively involved in the assessment of any options proposed. Many times, what appears to be a good idea in construction can have a disastrous effect on a facility's long-term operation. For example, automatic controls are often value engineered out of the design, yet the cost of their inclusion is incurred many times over in the labor required to perform their function manually over the life of the building. The commissioning agent can ensure that the design intent is preserved, the life-cycle costs are considered, and the impact on all systems of any value engineering modification proposed is thoroughly evaluated.

Once the design aspects are complete and value engineering ideas have been incorporated or rejected, the submittals, including verification checklists, need to be finalized. The submittals documentation is prepared by the systems contractors and reviewed by the commissioning agent. There are two types of submittals: technical submittals and commissioning submittals. Both types of submittals are discussed below.

Technical submittals are provided to document the systems contractors' interpretation of the design documents. The commissioning agent reviews the technical submittals for compliance and completeness. It is in this submittal review process that potential issues are identified prior to installation, reducing the need for rework and minimizing schedule delays. The technical submittals should include:

- Detailed schematics
- Equipment data sheets
- Sequence of operation
- Bill of material

A key technical submittal is the testing, adjusting, and balancing submittal (TAB). The TAB should include:

- TAB procedures
- Instrumentation
- Format for results
- Data sheets with equipment design parameters
- Operational readiness requirements
- Schedule

In addition to the TAB, other technical submittals, such as building automation control submittals, will be obtained from the systems contractors and reviewed by the commissioning agent.

The commissioning submittal generally follows the technical submittal in time and includes:

- Verification checklists
- Startup requirements
- Test and balance plan
- Training plan

The commissioning information in the commissioning submittal is customized for each element of the system.

These submittals, together with the commissioning specification, are incorporated into the commissioning plan, which becomes a living document codifying the results of the construction commissioning activities. This plan should be inspected in regular site meetings. Emphasis on the documentation aspect of the commissioning process early in the construction phase increases the contractors' awareness of the importance of commissioning to a successful project.

In addition to the submittals, the contractors are responsible for updating the design documents with submitted and approved equipment data and field changes on an ongoing basis. This update design document should be utilized during the testing and acceptance phase.

The commissioning agent also performs periodic site visits during the installation to observe the quality of workmanship and compliance with the specifications. Observed deficiencies should be discussed with the contractor and documented to ensure future compliance. Further inspections should be conducted to ensure that appropriate corrective action has been taken.

The best way to ensure that the items discussed above are addressed in a timely manner is to hold regularly scheduled commissioning meetings that require the participation of all systems contractors. This is the mechanism for ensuring that communication occurs. Meeting minutes prepared by the commissioning agent document the discussions and decisions reached. Commissioning meetings should be coordinated with the regular project meetings because many participants in a construction project need to attend both meetings.

Typical elements of a commissioning meeting include:

- Discussing field installation issues to facilitate rapid response to field questions
- Updating design documents with field changes
- Reviewing the commissioning agent's field observations
- Reviewing progress against schedule
- Coordinating multicontractor activities

Once familiar with the meeting process, an agenda will be helpful but not necessary. Meeting minutes should be kept and distributed to all participants.

With approved technical and commissioning plan submittals, as installation progresses, the contractor is ready to begin the system verification testing. The systems contractor generally executes the system verification independently of the commissioning agent. Contractor system verification includes:

- Point-to-point wiring checked out
- Sensor accuracy validated
- Control loops exercised

Each of the activities should be documented for each control or system element, and signed and dated by the verification technician.

The documentation expected from these activities should be clearly defined in the commissioning specification to ensure its availability to the commissioning agent for inspection of the verification process. The commissioning agent's role in the system verification testing is to ensure that the tests are completed and that the results reflect that the system is ready for the functional-performance tests. Because the commissioning agent is typically not present during the verification testing, the documentation controls how successfully the commissioning agent performs this aspect of commissioning.

In addition to system verification testing, equipment startup is an important activity during this phase. Equipment startup occurs at different time frames relative to the system verification testing, depending on the equipment and system involved. There may be instances when the system verification needs to occur prior to equipment startup to prevent a catastrophic event that could lead to equipment failure. The commissioning agent reviews the startup procedures prior to the startup to ensure that equipment startup is coordinated properly with the system verification. Unlike in verification testing, the commissioning agent should be present during HVAC equipment startup to document the results. These results are memorialized in the final commissioning report, so their documentation ultimately is the responsibility of the commissioning agent.

Once system verification testing and equipment startup have been completed, the acceptance phase begins.

Acceptance Phase

The acceptance phase of the project is the phase in which the owner accepts the project as complete and delivered in accordance with the specifications, and concludes with acceptance of the project in its entirety. An effective commissioning process during the installation phase should reduce the time and labor associated with the functional-performance tests of the acceptance phase.

Statistical sampling is often used instead of 100% functional-performance testing to make the process more efficient. A 20% random sample with a failure rate less than 1% indicates that the entire system was properly installed. If the failure rate exceeds 1%, a complete testing of every system may need to be completed to correct inadequacies in the initial checkout and verification testing. This random sampling statistical approach holds the contractor accountable for the initial checkout and test, with the performance testing serving only to confirm the quality and thoroughness of the installation. This approach saves time and money for all involved. It is critical, however, that the ramifications of not meeting the desired results of the random tests are clearly defined in the commissioning specifications.

The commissioning agent witnesses and documents the results of the functional-performance tests, using specific forms and procedures developed for the system being tested. These forms are created with the input of the contractor in the installation phase. Involvement of the maintenance staff in the functional-performance testing is important. The maintenance team is often not included in the design process, so they may not fully understand the design intent. The functional-performance testing can provide the maintenance team an opportunity to learn and appreciate the design intent. If the design intent is to be preserved, the maintenance team must fully understand the design intent. This involvement of the maintenance team increases their knowledge of the system going into the training and will increase the effectiveness of the training.

Training of the maintenance team is critical to a successful operational handover once a facility is ready for occupancy. This training should include:

- Operations and maintenance (O&M) manual overview
- Hardware component review
- Software component review
- Operations review
- Interdependencies discussion
- Limitations discussion
- Maintenance review
- Troubleshooting procedures review
- Emergency shutdown procedures review

The support level purchased from the systems contractor determines the areas of most importance in the training and therefore should be determined prior to the training process. Training should be videotaped for later use by new maintenance team members and in refresher courses, and for general reference by the existing maintenance team. Using the O&M manuals as a training manual increases the maintenance team's awareness of the information contained in them, making the O&M manuals more likely to be referenced when appropriate in the future.

The O&M manuals should be prepared by the contractor in an organized and easy-to-use manner. The commissioning agent is sometimes engaged to organize them all into an easily referenced set of documents. The manuals should be provided in both hard-copy and electronic formats, and should include:

- System diagrams
- Input/output lists
- Sequence of operations
- Alarm points list
- Trend points list
- Testing documentation
- Emergency procedures

These services—including functional-performance testing, training, and preparing O&M manuals—should be included in the commissioning plan to ensure the project's successful acceptance. The long-term success of the project, however, is determined by the activities that occur in the postacceptance phase.

Postacceptance Phase

The postacceptance phase is the phase in which the owner takes beneficial occupancy and forms an opinion about future work with the design team, contractors, and the commissioning agent who completed the project. This is also the phase in which LEED certification, if adopted, is completed. Activities that usually occur in the acceptance phase should instead occur in the postacceptance phase. This is due to constraints that are not controllable by the contractor or owner. For example, seasonal changes may make functional-performance testing of some HVAC systems impractical during the acceptance phase for certain load conditions. This generally means that in locations that experience significant seasonal climate change, some of the functional-performance testing is deferred until suitable weather conditions exist. The commissioning agent determines which functional-performance tests need to be deferred and hence carried out in the postacceptance phase.

During the postacceptance phase, the commissioning agent prepares a final commissioning report that is provided to the owner and design team. The executive summary of this report provides an overall assessment of the design intent conformance. The report details whether the commissioned equipment and systems meet the commissioning requirements. Problems encountered and corrective actions taken are documented in this report. The report also includes the signed and dated startup and functional-performance testing checklists.

The final commissioning report can be used profitably as the basis of a "lessons learned" meeting involving the design team so that the commissioning process can be continuously improved and adaptations can be made to the owner's commissioning policy for future projects. The owner should use the experience of the first commissioned project to develop the protocols and standards for future projects. The documentation of this experience is the owner's commissioning policy. Providing this policy and the information it contains to the design and construction team for the next project can help the owner reduce budget overruns by eliminating any need to reinvent protocols and standards and by setting the right expectations earlier in the process.

Commissioning therefore should not be viewed as a onetime event but should instead be viewed as an operational philosophy. A recommissioning or continuous commissioning plan should be adopted for any building to sustain the benefits delivered from a commissioning plan. The commissioning agent can add great value to the creation of the recommissioning plan and can do so most effectively in the postacceptance phase of the project.

Figure 1 depicts the information development that occurs in the evolution of a commissioning plan and summarizes the information presented in the preceding sections by outlining the various phases of the commissioning process.

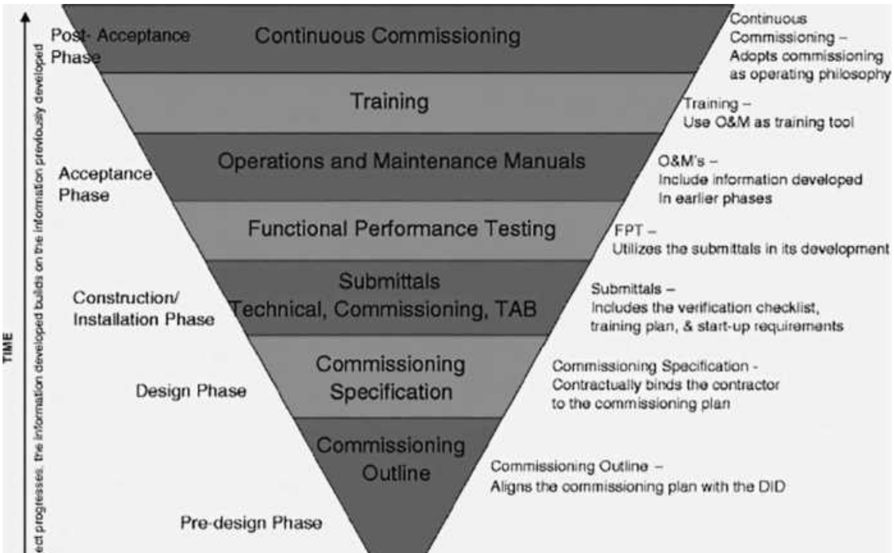


FIGURE 1 The commissioning plan.

With this background, the reader is better positioned for success in future commissioning projects and better prepared to learn the key success factors in commissioning and how to avoid common mistakes in the commissioning process.

Commissioning Success Factors

Ultimately, the owner will be the sole judge of whether a commissioning process has been successful. Thus, second only to the need for a competent, professional commissioning agent, keeping the owner or the owner's senior representative actively involved in and informed at all steps of the commissioning process is a key success factor. The commissioning agent should report directly to the owner or the owner's most senior representative on the project, not only to ensure that this involvement and information transfer occur, but also to ensure the objective implementation of the commissioning plan—a third key success factor.

Another key success factor is an owner appreciation— which can be enhanced by the commissioning agent—that commissioning must be an ongoing process to get full benefit. For example, major systems should undergo periodic modified functional testing to ensure that the original design intent is being maintained or to make system modification if the design intent has changed. If an owner appreciates that commissioning is a continuous process that lasts for the entire life of the facility, the commissioning process will be a success.

Most owners will agree that the commissioning process is successful if success can be measured in a cost/benefit analysis. Cost/benefit or return on equity is the most widely used approach to judge the success of any project. Unfortunately, the misapplication of cost/benefit analyses has been the single largest barrier to the widespread adoption of commissioning. For example, because new construction does not have an energy baseline from which to judge energy savings, improper application of a cost/benefit analysis can lead to failure to include energy savings technologies—technologies the commissioning agent can identify—in the construction process. Similarly, unless one can appreciate how commissioning can prevent schedule delays and rework by spotting issues and resolving them early in the construction process, one cannot properly offset the costs of commissioning with the benefits.

Fortunately, there are now extensive studies analyzing the cost/benefit of commissioning that justify its application. A study performed jointly by Lawrence Berkeley National Laboratory; Portland Energy Conservation, Inc.; and the Energy Systems Laboratory at Texas A&M University provides compelling analytical data on the cost/benefit of commissioning. The study defines the median commissioning cost for new construction as $1 per square foot or 0.6% of the total construction cost. The median simple payback for new construction projects utilizing commissioning is 4.8 years. This simple payback calculation does not take into account the quantified nonenergy impacts, such as the reduction in the cost and frequency of change orders or premature equipment failure due to improper installation practices. The study quantifies the median nonenergy benefits for new construction at $1.24 per square foot per year.[3]

While the primary cost component of assessing the cost/benefit of commissioning lies in whether there was a successful negotiation of the cost of services with the commissioning service provider, the more important aspect of the analysis relates to the outcomes of the process. For example, after a commissioning process is complete, what are the answers to these questions?

Are the systems functioning to the design intent?
Has the owner's staff been trained to operate the facility?
How many of the systems are operated manually a year after installation?

Positive answers to these and similar questions will ensure that any cost/benefit analysis will demonstrate the value of commissioning.

To ensure that a commissioning process is successful, one must avoid common mistakes. A commissioning plan is a customized approach to ensuring that all the systems operate in the most effective and efficient manner. A poor commissioning plan will deliver poor results. A common mistake is to use an existing commissioning plan and simply insert it into a specification to address commissioning. Each commissioning plan should be specifically tailored to the project to be commissioned.

Also, perhaps due to ill-conceived budget constraints, commissioning is implemented only in the construction phase. Such constraints are ill conceived because the cost of early involvement of the commissioning agent in the design phases is insignificant compared with the cost of correcting design defects in the construction phase. Significant cost savings can arise from identifying design issues prior to construction. Studies have shown that 80% of the cost of commissioning occurs in the construction phase.[2] Also, the later the commissioning process starts, the more confrontational commissioning becomes, making it more expensive to implement later in the process.[2] Therefore, adopting commissioning early in the project is a key success factor.

Value engineering often results in ill-informed, last-minute design changes that have an adverse and unintended impact on the overall building performance and energy use.[3] By ensuring that the commissioning process includes careful evaluation of all value engineering proposals, the commissioning agent and owner can avoid such costly mistakes.

Finally, the commissioning agent's incentive structure should not be tied to the number of issues brought to light during the commissioning process, as this can create an antagonistic environment that may create more problems than it solves. Instead, the incentive structure should be outcome based and the questions outlined above regarding compliance with design intent, training results, and postacceptance performance provide excellent bases for a positive incentive structure.

Conclusion

Commissioning should be performed on all but the most simplistic of new construction projects. The benefits of commissioning include:

- Optimization of building performance
  —Enhanced operation of building systems
  —Better-prepared maintenance staff
  —Comprehensive documentation of systems
  —Increased energy efficiency
  —Improved quality of construction
- Reduced facility life-cycle cost
  —Reduced impact of design changes
  —Fewer change orders
  —Fewer project delays
  —Less rework or postconstruction corrective work
  —Reduced energy costs
- Increased occupant satisfaction
  —Shortened turnover transition period
  —Improved system operation
  —Improved system reliability

With these benefits, owners and contractors alike should adopt the commissioning process as the best way to ensure cost-efficient construction and the surest way to a successful construction project.

References

1. Quantum Consulting, Inc. *Commissioning in Public Buildings: Market Progress Evaluation Report,* June 2005. Available at http://www.nwalliance.org/resources/reports/141.pdf.
2. ACG AABC Commissioning Group. ACG Commissioning Guidelines 2005. Available at http://www.commissioning.org/commissioningguideline/ACGCommissioningGuide-line.pdf.

3. Mills, E.; Friedman, H.; Powell, T. et al. *The Cost-Effectiveness of Commercial-Buildings Commissioning*, December 2004. Available at http://eetd.lbl.gov/Emills/PUBS/Cx-Costs-Benefits.html.
4. American Society of Heating. *Refrigerating and Air-Conditioning Engineers,* ASHRAE Guidelines 1–1996.
5. Green Building Council Web site. Available at http://www.usgbc.org/leed (accessed October 2005).

Bibliography

1. SBW Consulting, Inc. *Cost-Benefit Analysis for the Commissioning in Public Buildings Project*, May 2004. Available at http://www.nwalliance.org/resources/documents/CPBRe-port.pdf (accessed).
2. Turner, W.C. *Energy Management Handbook*, 5th Ed.; Fairmont Press: Lilburn, GA, 2005.
3. Research News Berkeley Lab. *Berkeley Lab Will Develop Energy-Efficient Building Operation Curriculum for Community Colleges,* December 2004. Available at http://www.lbl.gov/Science-Articles/Archive/EETD-college-curriculum.html.
4. Interview with Richard Holman. *Director of Commissioning and Field Services for Affiliated Engineers, Inc.* Walnut Creek, CA.

8

Energy Sources: Renewable versus Non-Renewable

Marc A. Rosen

Introduction

Energy can exist in many forms and be converted from form to form by energy-conversion technologies. Society and the people within it use energy carriers (often simply referred to as energy), which are produced from energy sources, in all aspects of living.

A basic understanding of renewable energy and non-renewable energy is provided in this entry. Renewable energy includes the energy received directly and indirectly from the sun as well as energy derived from other natural forces. Non-renewable energy includes non-renewable energy resources as well as energy forms that do not exist naturally but are produced by people. The main types of renewable energy are listed in Table 1, and the main types of non-renewable energy are listed in Table 2. All of these have been studied extensively.[1–10]

From the breakdown of renewable and non-renewable energy in Tables 1 and 2, it is clear that energy resources are often categorized into two groups: (1) those generally acknowledged to be finite and non-renewable and therefore not sustainable over the long term (e.g., fossil fuels, peat, uranium) and (2) those generally considered renewable and therefore sustainable over the relatively longer term (e.g., sunlight, wind, tides, falling water). Wastes (convertible to useful energy forms through, for example, waste-to-energy incineration facilities) and biomass fuels are also sometimes viewed as renewable energy sources.

In this entry, energy forms, sources, and carriers are explained. Renewable energy and non-renewable energy are discussed, and energy-conversion technologies are described. Then, energy use and factors

TABLE 1 Types of Renewable Energy

- Direct solar radiation
- Solar-related energy*
  - Water based
    - Hydraulic energy (falling and running water, including large and small hydro)
    - Wave energy
    - Ocean thermal energy (from temperature difference between surface and deep waters of the ocean)
  - Air based
    - Wind energy
  - Land based
    - Biomass (where the rate of use does not exceed the rate of replenishment)
    - Geothermal energy (ambient)
  - Non-solar-related energy
  - Geothermal energy (internal heat of the earth)
  - Tidal energy (from gravitational forces of the sun and moon and the rotation of the earth)

*As explained in the text, fossil fuels are originally solar energy, but with a very long time lag for their transformation. So technically, they are solar-related energies. But for practical purposes, fossil fuels are non-renewable, so they are not listed in this table.

TABLE 2 Types of Non-Renewable Energy

- Energy sources
  - Fossil fuels
    - Conventional
      - Coal
      - Oil
      - Natural gas
    - Alternative
      - Oil shales
      - Tar sands
      - Peat
  - Non-fossil fuels
    - Uranium
    - Fusion material (e.g., deuterium)
    - Wastes (which can be used as energy forms or converted to more useful energy forms)
- Energy currencies
  - Work
  - Electricity
  - Thermal energy
    - Heat (or a heated medium such as hot air, steam, exhaust gases)
    - Cold (or a cooled medium such as cold brine, ice)
  - Secondary chemical fuels
    - Conventional
      - Oil products (e.g., gasoline, diesel fuel, naphtha)
      - Synthetic gaseous fuels (e.g., from coal gasification)
      - Coal products (e.g., coke)
    - Non-conventional
      - Methanol
      - Ammonia
      - Hydrogen

in energy selection are discussed. Finally, efficiencies for energy use are presented, along with measures to improve energy efficiency.

Forms, Sources, and Carriers of Energy

Energy comes in a variety of forms, including fossil fuels (e.g., coal, oil, natural gas), fossil fuel-based products (e.g., gasoline, diesel fuel), uranium, electricity, work (e.g., mechanical energy in a rotating engine shaft), heat, heated substances (e.g., steam, hot air), and light and other electromagnetic radiation.

Energy sources (sometimes called primary energy forms) are found in the natural environment. Some are available in finite quantities (e.g., fossil fuels, fossil fuel-containing substances such as oil sands, peat, and uranium). Some energy resources are renewable (or relatively renewable), including sunlight (or solar energy), falling water, wind, tides, geothermal heat, wood, and other biomass fuels (provided the growth rate exceeds or meets the rate of use). Energy sources are often processed from their raw forms prior to use.

Energy carriers (sometimes called energy currencies) are the energy forms that we transport and use, and include some energy sources (e.g., fossil fuels) and processed (or secondary) energy forms (e.g., gasoline, electricity, work, heat). The processed energy forms are not found in the environment.

The distinction between energy carriers and sources is important. Energy carriers can exist in a variety of forms and can be converted from one form to another, while energy sources are the original resource from which an energy carrier is produced. Misunderstanding sometimes results between energy sources and carriers because some energy sources are also energy carriers. For example, hydrogen is an energy carrier, not an energy source, and can be produced from a wide range of resources using various energy-conversion processes (e.g., water electrolysis, reforming of natural gas, coal gasification). In this way, hydrogen energy is analogous to electricity. Nevertheless, hydrogen is often erroneously referred to as an energy source, especially in discussions of its potential future role as a chemical energy carrier to replace fossil fuels.

Renewable Energy

Renewable energy includes the solar radiation incident on the earth and the energy forms that directly result from that radiation. Renewable energy also includes the energy supplied by other natural forces, such as gravitation and the rotation of the earth. The types of renewable energy are summarized in Table 1. It is this energy that makes possible the existence of ecosystems, human civilizations, and life itself.

Solar Energy

Direct solar radiation is the main type of renewable energy.[11] The daily energy output of the sun is 8.33×10^{25} kWh, of which the earth receives 4.14×10^{15} kWh. At any instant, the rate of solar energy reaching the earth is 1.75×10^{17} W, which is about 20,000 times greater than the total energy-use rate of the world. Solar radiation can be converted directly to electricity in photovoltaic devices. Also, solar energy can be collected as heat and used for thermal processes such as space and water heating or concentrated for use for in high-temperature heating and thermal electricity generation.

Most of the energy that enters the system of the earth and its atmosphere eventually exits to space. This concept can be demonstrated by considering the earth–sun energy balance (see Figure 1). A general energy balance

$$\text{Energy input} - \text{Energy output} = \text{Energy accumulation}$$

can be applied to the earth when

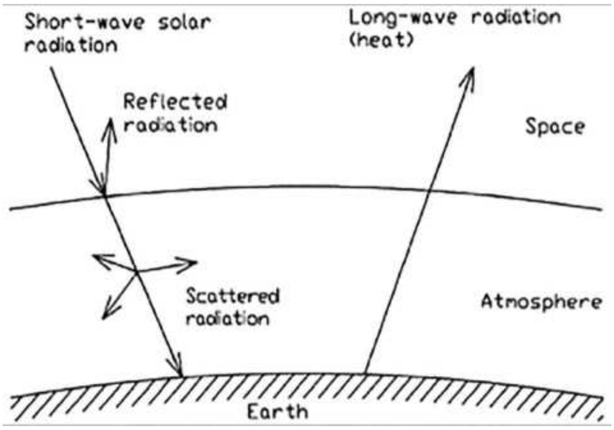


FIGURE 1 Earth–sun energy balance.

- The energy input is the short-wave solar radiation entering the atmosphere.
- The energy output is the long-wave radiation exiting the atmosphere to space.
- The energy accumulation term is the increase in energy of the earth and its atmosphere.

The main implication of this global energy balance is that since the average temperature of the earth is relatively constant, the energy accumulation term is zero. Therefore, the energy output is equal to the energy input for the planet.

Global warming is caused by a disruption in the earth–sun energy balance. The main cause of global warming is increased releases of atmospheric "greenhouse gases" that absorb radiation in the 8–20 μm region. When greenhouse gas concentrations increase in the atmosphere, energy output from the earth and its atmosphere (Figure 1) is reduced while energy input remains constant. Thus, the energy accumulation term becomes positive, leading to an increase in the average temperature of the earth. Eventually, if concentrations of greenhouse gases in the atmosphere stabilize at new levels, the energy balance is re-established but at some higher average planetary temperature.

Solar-Related Energy

Several types of renewable energy are a consequence of solar radiation. These can be loosely arranged into water, air, and land groupings. Some forms of solar-derived energy are more common, for reasons of practicality, economics, and environmental impact.

Water-Based Solar-Related Energy. The most common solar-related energy is hydraulic energy, which includes falling and running water in natural settings such as rivers and waterfalls. Large-scale hydroelectric generating installations are common. Most economically utilizable hydraulic resources have already been developed. Recently, interest in the potential uses of small-scale hydro has grown. Other forms of solar-related energy involve water. Wave energy systems that take advantage of the motion of waves have been proposed, although the potential contribution from wave energy is relatively small. Ocean thermal energy arises from the temperature difference between surface and deep waters of the ocean. This temperature difference can be utilized to drive a heat engine, and several ocean thermal energy-conversion (OTEC) devices have been tested.

Land-Based Solar-Related Energy. Biomass energy includes wood and other forms of plants and organic matter. Biomass can act as a fuel itself or can be converted into more desirable fuels. Several fast-growing trees have been identified as good candidates for biomass energy production. Biomass energy is only a renewable resource when the rate at which it is used does not exceed the rate at which it is replenished. Another form of solar-related energy that involves land is ambient geothermal energy. This is the energy at ambient temperatures in the ground near the surface, which is maintained predominantly by solar radiation and ambient air temperatures. This ambient energy can be used for heating and/or cooling, especially in conjunction with technologies like heat pumps.

Air-Based Solar-Related Energy. Wind energy is used extensively in some countries (e.g., Denmark) for electricity generation. Its use is growing in recent years, making it more widespread than even a decade ago.[12]

Non-Solar-Related Energy

The main types of renewable energy in this category are geothermal energy,[13] which exists as a consequence of the internal heat of the earth, and tidal energy, which is attributable to the gravitational forces of the sun and moon and the rotation of the earth. Both of these energy sources have been used in limited ways.

Non-Renewable Energy

Non-renewable energy, often called secondary energy, includes both energy resources that are available in limited quantities and not renewable and energy forms produced by humankind.[2–7,13] Different types of non-renewable energy are summarized in Table 2. Non-renewable energy is related to the level of technological development of a society.

Two main categories exist for non-renewable energy:

- Energy resources that are not renewable. The most common of these are fossil fuels, which are the basis of the economy for most industrialized countries. In addition to conventional fossil fuels, there exist alternative fossil fuels such as oil shales, tar sands, and peat. Other non-renewable energy resources include uranium and fusion material (e.g., deuterium).
- Energy currencies not existing naturally. This category of non-renewable energy encompasses energy currencies that do not exist naturally. They include such basic energy forms as work, electricity, and thermal energy. The latter can be either heat or a heated medium like steam or cold water.

At this point, a note on the non-renewability of fossil fuels is necessary for clarity and comprehensiveness. Fossil fuels are originally solar energy, but with a very long time lag for their transformation. Thus, considering the origin of fossil fuels, they could in theory be listed in Table 1 along with other solar-related energies. The non-renewability of fossil fuels deals with the lengthy time lag as well as the unique conditions on Earth that permit the creation of biomass at a much faster rate than its decomposition. Thus, fossil fuels are for practical purposes non-renewable, so they are not listed in Table 1.

Thermal energy in the form of heat or cold can be transported to users over long distances in district heating and/or cooling systems. District heating systems use centralized heating facilities to produce a heated medium that is transported to many users connected along a district heating network. For example, buildings in the cores of many cities are often connected by pipes through which hot water or steam flows to provide space and water heating. Similarly, district cooling involves the central production of a cold medium, which is transported to users through a piping network to provide cooling. Many cities and industrial parks utilize such district energy systems.

Wastes, which include recovered materials and energy that would otherwise be discarded, are also sometimes considered as an energy source in the category of non-renewable energy. Wastes can be used directly as energy forms or converted to more useful forms. Waste materials and waste heat can be recovered for utilization both within a facility and in other facilities where they are needed. For example, waste heat from hot gases (e.g., stack gases) and liquids (e.g., cooling-water discharges) can sometimes be recovered. Also, material wastes can be used in waste-to-energy incineration facilities, which burn garbage to provide heat and to generate electricity. Utilizing such wastes offsets the need for further supplies of external energy.

Non-renewable energy also includes secondary chemical fuels. Some conventional ones include oil-derived products such as gasoline, diesel fuel, and naphtha, as well as synthetic gaseous fuels (e.g., from coal gasification) and coal products (e.g., coke). The types of non-conventional chemical fuels proposed are numerous and include methanol, ammonia, and hydrogen.

Life Cycle Considerations

As many types of non-renewable energy can be produced from energy resources or converted from other types of energy, it is important to understand and consider all the steps in the entire life cycle of an energy product. The following life stages are usually included in assessments:

- Extraction or collection of raw energy resources
- Manufacturing and processing of the desired energy form(s)
- Transportation and distribution of the energy to users
- Energy storage
- Use of the energy to provide services and tasks
- Recovery and reuse of output energy that would otherwise be wasted (e.g., waste heat recovery)
- Recycling of wastes from any of the above steps
- Disposal of final wastes (e.g., materials such as stack gases and solid wastes including ash).

For example, the life cycle of a general energy form may involve the following chain of events:

Raw resource → Finished resource → Energy product → Waste → Waste disposal

Two or more types of energy can be simultaneously produced in some systems. For example, cogeneration is a process that usually refers to the combined generation of electricity (or work) and heat (or a heated medium). Trigeneration refers to an extended cogeneration process in which cooling is provided as a third product.

Some examples of the different life cycles for electricity generation methods from a range of energy sources are presented in Table 3. In that table, methods based on fossil fuels and non-fossil resources are considered. In addition, electricity generation from different energy sources via a less conventional technology, fuel cells, is considered in Table 4. All processes in Tables 3 and 4 include equipment disposal as a final step in the life cycle, although that step is not shown.

TABLE 3 Life Cycles for Selected Energy Sources and Methods for Electricity Generation

Method and Energy Source	Energy-conversion Processes in the Life Cycle
Fossil fuel-based methods	
Natural gas	Extraction via gas well → Transport → Processing → Natural gas-fired power plant
Coal	Coal mining → Transport → Processing → Coal-fired power plant
Oil	Extraction via oil well → Transport (tanker or pipeline) → Refining → Oil-fired power plant
Non-fossil-based methods	
Nuclear	Uranium mining → Transport → Processing → Nuclear power plant
Hydro	Hydraulic turbine
Wind	Wind generator
Solar (thermal)	Solar energy collection → Thermal power plant
Solar (non-thermal)	Solar photovoltaic panels
Geothermal	Well to geothermal source → Geothermal power plant
Ocean (thermal)	OTEC power plant

TABLE 4 Life Cycles for Electricity Generation from Various Energy Sources Using Fuel Cells

Energy Source	Energy-conversion Processes in the Life Cycle
Fossil fuels	Fossil-based hydrogen production → Fuel cell
Non-fossil energy sources	Hydrogen production → Fuel cell
High-temperature heat (e.g., solar)	Hydrogen production via thermochemical cycle → Fuel cell

Energy Use

Energy use involves the production of useful energy forms through energy-conversion processes, as well as energy transportation, distribution and storage, and the utilization of energy resources and processed forms of energy to provide services and perform tasks.

Energy-Conversion Technologies

Desired energy carriers are produced from energy sources, or converted from one form to another, using energy-conversion technologies. The energy-conversion technology appropriate in a given instance depends on the initial (or source) and final (or desired) energy forms. Conventional energy-conversion technologies include hydroelectric, fossil fuel, and nuclear generating stations; oil refineries; engines and motors; and heaters.[14–16] Some examples of less conventional energy-conversion technologies include fuel cells, solar photovoltaics, and high-efficiency and/or clean technologies for fossil fuels (e.g., combined-cycle systems).[4,17,18] Some energy-conversion technologies yield more than one product (e.g., cogeneration). Various technologies for electricity generation are listed in Tables 3 and 4.

A major energy-conversion process is combustion, which drives furnaces, engines, transportation vehicles, etc., and which can be expressed for a hydrocarbon fuel as

$$\text{Hydrocarbon fuel} + \text{Air}\left(\text{mainly } O_2 \text{ and } N_2\right) \rightarrow CO_2 + H_2O + N_2 + \text{Other substances}$$

The other substances are leftover reactants and other reaction products. The reaction for stoichiometric combustion (i.e., complete combustion with a balanced amount of reactants and no reactions that yield other products) of a general hydrocarbon (C_mH_n) in air (treated as only nitrogen and oxygen) can be written as

$$C_mH_n + (m + n/4)(O_2 + 3.76N_2) \rightarrow m\,CO_2 + n/2H_2O + 3.76\,(m + n/4)N_2$$

Here, m and n are variables that take on different values for different hydrocarbons (e.g., in approximate terms, $m = 1$ and $n = 4$ for natural gas, which is mainly methane, $m = 1$ and $n = 2$ for oil, $m = 1$ and $n = 1$ for typical coal, and $m = 1$ and $n = 0$ for pure carbon). Several important points stem from these expressions. First, carbon dioxide is an inherent product of the combustion of any carbon-containing fuel. Generally, the only way to avoid carbon dioxide emissions is to eliminate the use of carbon-based fuels (e.g., by using hydrogen as a fuel). Second, if insufficient oxygen is available (or if mixing is poor or reaction time short), the carbon will often only partially combust, yielding CO. Third, excess air (over stoichiometric) is usually used to improve fuel burnup, often lowering combustor efficiencies since some of the fuel energy must go towards heating the excess air. Fourth, often the fuel is not a pure hydrocarbon and contains other substances, e.g., sulfur. Sulfur combusts, yielding heat, and reacts to sulfur dioxide.

Energy Selection

Energy selection involves choosing both energy carriers and sources. The selection often depends on the energy service to be provided and the energy-conversion technologies available. Some energy selections involve energy or fuel substitution (e.g., heating with natural gas rather than electricity).

To prevent environmental impact, energy sources are usually preferred, which are renewable and cause relatively lower environmental impacts.[19–23] By extension, preference is also placed on energy sources and carriers that can be used with higher efficiency and more environmentally benign energy-conversion technologies (e.g., boilers having low emissions of nitrogen oxides, NO_x).[9,24,25]

Renewable energy sources (Table 1), usually being derived from sunlight or solar-derived sources (e.g., wind, waves), are sustainable. A major barrier to renewable energy sources is that they are usually more costly to use than non-renewable energy sources such as fossil fuels, although this observation is not true in some niche applications.

Environmental Considerations

The use of fossil fuels leads to combustion emissions. Problematic pollutants are generally lower for fuels having higher hydrogen-to-carbon atomic ratios, so natural gas is more benign than oil, which is more benign than coal. There are no emissions during normal operation of nuclear power facilities except the spent fuel, which remains radioactive for many years. Although falling and running water, derived from solar energy, drive hydroelectric generation, flooding of lands sometimes occurs. The use of renewable energy is relatively benign, but resources are required to build the relevant energy-conversion technologies and large land tracts are often needed (e.g., for solar collectors or wind generators). In all assessments of environmental impact, adoption of a life cycle perspective is critical.

Energy Efficiency and Other Measures of Merit for Energy Use

Many efficiencies can be defined for energy use. Other measures are also often used to assess the merit of energy use relative to other criteria. One important measure is energy intensity, which reflects the energy use per some unit of output, e.g., energy consumption per monetary unit of gross domestic product for a country or region. Since they deal with raw materials, primary processing industries (e.g., petroleum/coal production, chemicals production, primary metals) generally have higher energy intensities than intermediate and secondary processing industries, which, although they come in a wide variety of forms, tend to deal with more finished products.

Numerous methods and technologies exist for improving energy efficiency:

Use of High-Efficiency Devices. Energy efficiency can generally be increased via high-efficiency versions of energy intensive devices (e.g., home appliances, furnaces, air conditioners, motors, boilers). New lighting fixtures (including bulbs, reflectors, diffusers, and ballasts), for instance, have significantly higher efficiencies and longer lives than older equipment. Lighting efficiency can also be increased by task lighting (i.e., directing light where it is needed), reducing lighting to levels adequate for the human eye and the nature of the facility, and using timers, dimmers, and occupancy sensors.

Energy Leak and Loss Prevention. Energy losses are associated with energy leaks. Approaches to avoid such energy leaks include (1) applying leak-prevention technologies and methods (e.g., sealing leaks in storage vessels and pipelines and insulating to reduce undesired heat flows) and (2) inspecting periodically to detect leaks and initiate appropriate actions.

Application of Advanced and Integrated Energy Systems. Many advanced energy systems feature increased efficiencies compared to conventional technologies. Efficiency can also be increased by advantageously integrating energy systems so that wastes become inputs to other processes. Waste-recovery and waste-to-energy plants normally increase efficiency by offsetting some of the need for external energy. Cogeneration and trigeneration are often more efficient than separate processes for the individual products. Compared to the alternative of each facility having its own heating and/ or cooling plant, district heating and/or cooling often provide increased efficiency and reduced environmental impact because many efficiency-improvement and environmental control measures are possible in centralized, large-scale facilities. Attempts have been made to optimize regional heat markets and systems.[26]

Improved Monitoring, Control, and Maintenance. Energy efficiency can be improved through better monitoring, control, and maintenance of operations, so that performance degradation from design specifications is avoided. Computer systems allow automated acquisition of frequent and widespread readings and diagnostic evaluations. The life span of energy equipment can be increased through diligent maintenance (e.g., cleaning, lubricating, calibrating, tuning, regular testing, periodic overhauling, replacing consumable items regularly).

Improved Matching of Energy Supplies and Demands. Rather than supply an energy form of a level that greatly exceeds that required for a specific energy demand, it is usually more efficient to supply an energy form of a better matched level. For instance, better matching can often be achieved for heat-transfer flow temperatures (e.g., for space heating at about 22°C, a heat supply of perhaps 40°C could be sufficient, rather than furnace combustion gases at hundreds of degrees). Thus, industrial waste heat and low-temperature cogenerated heat can satisfy some heating needs.

Energy Storage. Sometimes, available energy supply exceeds demand, and the energy is not utilized. [27,28] At times, the supply is not controllable (e.g., sunlight may be available for heating during the day whether or not it is needed), while in other cases, the supply may not be easily alterable (e.g., the waste heat from a factory). Energy storage can improve system efficiency by storing energy between times when it is available and needed.

Improved Building Envelopes. The energy efficiency of a building can be improved by increasing insulation to reduce heat infiltration in summer and heat loss in winter, applying weather stripping and caulking to reduce air leakages, and utilizing advanced and high-efficiency windows. The latter reduce heat losses by using multiple glazings that are sometimes separated by insulating gases or a vacuum, electronic and photosensitive windows that automatically reflect or absorb excessive sunlight, low-emissivity window coatings that increase a window's resistance to heat loss, and efficient window shades that adjust themselves using photosensitive sensors to block excessive sunlight while permitting natural room lighting.

Use of Passive Strategies to Reduce Energy Use. Passive, as opposed to active, methods can be used to reduce energy use. The following are some examples: adjusting design temperatures such as those used for space heating in winter and space cooling in summer; shutting energy devices during periods of non-use and using timers to control operating hours; using daylight harvesting to reduce artificial lighting needs; applying "free cooling" by using cool outdoor air when it is available to cool warm indoor spaces, rather than active air conditioning; using solar radiation to heat buildings by exploiting the thermal storage capacity of buildings; and locating trees, windows, and window shades so as to keep buildings cool during summer.

Use of Exergy Analysis. Exergy analysis is used to improve and optimize the efficiency and performance of energy and other systems and processes.[1,13,29–31]

Energy Sustainability

Achieving appropriate energy technologies, resources, and efficiencies can help in efforts to make energy systems more sustainable, in terms of environmental protection as well as economic and social development.[32] Energy sustainability requires appropriate energy sources and carriers, efficiencies, and environmental protection, in addition to other needs.[33,34] Efforts to achieve energy sustainability greatly support efforts to achieve overall sustainability,[35] and contribute significantly to attaining the United Nations Sustainable Development Goals for 2015–2030.[36]

Example

Many examples can illustrate the methods and technologies described in the section "Energy Efficiency and Other Measures of Merit for Energy Use" for improving energy efficiency. Here, we focus on the use of high-efficiency lighting. New fixtures have significantly higher efficiencies and longer lives than older equipment. Several operating parameters for various types of lighting are shown in Table 5. For example, the efficiencies for some light sources (in lumens of light delivered per watt of electricity consumed) are as follows: incandescent (10–30), mercury (20–55), fluorescent (20–60), high-pressure sodium (50–130), and low-pressure sodium (80–155). Thus, the same amount of light can be delivered with high-efficiency lighting using much less than 20% of the electricity required for an incandescent bulb. Lighting using

TABLE 5 Operating Parameters for Various Types of Lighting

	Normalized Light Output (lumens/W)			Normalized Electricity Use for Lighting (W/lumens)			Increase in Normalized Light Output (%)[a]	Decrease in Normalized Electrical Use (%)[a]
Lighting Type	Mean	Lower Range	Upper Range	Mean	Upper Range	Lower Range		
Incandescent	20	10	30	0.050	0.100	0.033	0	0
Mercury	38	20	55	0.026	0.050	0.018	90	48
Fluorescent	40	20	60	0.025	0.050	0.017	100	50
High-pressure sodium	90	50	130	0.011	0.020	0.0080	350	78
Low-pressure sodium	118	80	155	0.0085	0.013	0.0065	490	83

[a] Changes are relative to incandescent lighting and based on mean values.

light-emitting diode (LED) technology is capable today of over 100 lumens per watt, and some prototypes have demonstrated the ability to achieve 200 and even 300 lumens per watt.

On a regional or national level, the savings can be significant. Since lighting is a significant energy consumer, accounting for about 20% of U.S. electrical energy use, the use of high-efficiency lighting across the United States could reduce that country's electricity consumption by 18% (i.e., 90% of the 20% of electricity use attributable to lighting). Of course, cost and other considerations limit the implementation of high-efficiency lighting.

Note that other measures to increase lighting efficiency could reduce the electricity used for lighting further. If, for instance, electricity use for lighting is reduced by an additional 5% through application of (1) timers, dimmers, and occupancy sensors to turn lights off when rooms are unoccupied; (2) reduced lighting intensities; and (3) task lighting, then electricity use in the United States could be reduced by 19% (i.e., 95% of the 20% of electricity use attributable to lighting).

Conclusion

Categorizing energy forms as renewable energy and non-renewable energy allows for improved understanding of energy systems. Renewable energy and non-renewable energy are important in societies, particularly since they affect many facets of life and living standards. These energy categories and their applications are described in this entry, as are energy forms, sources, carriers, and technologies; energy use; energy selection and efficiencies; and efficiency-improvement measures.

Glossary

m	number of carbon atoms in a hydrocarbon
n	number of hydrogen atoms in a hydrocarbon

Acknowledgments

The author is grateful for the support provided for this work by the Natural Sciences and Engineering Research Council of Canada.

References

1. Dincer, I., Rosen, M.A.; Ahmadi, P. *Optimization of Energy Systems*; Wiley: London, 2017.
2. Organization for Economic Cooperation and Development. *Energy: The Next Fifty Years*; OECD Publishing: Washington, DC, 1999.
3. Organization for Economic Cooperation and Development. *World Energy Outlook 2010*; OECD Publishing: Washington, DC, 2010.

4. Asdrubali, F.; Desideri, U., Eds. *Handbook of Energy Efficiency in Buildings: A Life Cycle Approach*; Elsevier: Oxford, 2019.
5. Niele, F. *Energy: Engine of Evolution*; Elsevier: Oxford, 2005.
6. Roosa, S.A.; Doty, S.; Turner, W.C. *Energy Management Handbook*, 9th Ed.; Fairmont Press: Lilburn, GA, 2018.
7. Goldemberg, J.; Johansson, T.B.; Reddy, A.K.N.; Williams, R.H. *Energy for a Sustainable World*; Wiley: New York, 1988.
8. Asif, M.; Muneer, T. Energy supply, its demand and security issues for developed and emerging economies. *Renewable and Sustainable Energy Reviews* **2007**, *11*, 1388–1413.
9. Ramanathan, R. A multi-factor efficiency perspective to the relationships among world GDP, energy consumption and carbon dioxide emissions. *Technological Forecasting and Social Change* **2006**, *73*, 483–494.
10. Birol, F. World energy prospects and challenges. *Asia Pacific Business Review* **2007**, *14*, 1–12.
11. Moharamian, A.; Soltani, S.; Rosen, M.A.; Mahmoudi, S.M.S.; Bhattacharya, T. Modified exergy and modified exergoeconomic analyses of a solar based biomass co-fired cycle with hydrogen production. *Energy* **2019**, *167*, 715–729.
12. Ehyaei, M.A.; Ahmadi, A.; Rosen, M.A. Energy, exergy, economic and advanced and extended exergy analyses of a wind turbine. *Energy Conversion and Management* **2019**, *183*, 369–381.
13. Rosen, M.A.; Koohi-Fayegh, S. *Geothermal Energy: Sustainable Heating and Cooling Using the Ground*; Wiley: London, 2017.
14. International Energy Agency. *Energy Technology Transitions for Industry: Strategies for the Next Industrial Revolution*; International Energy Agency: Paris, 2009.
15. International Energy Agency. *Transport, Energy and CO_2: Moving Toward Sustainability*; International Energy Agency: Paris, 2009.
16. Organisation for Economic Co-cooperation and Development. *Nuclear Energy Outlook 2008*; Organisation for Economic Co-operation and Development Nuclear Energy Agency, OECD Publishing: Paris, 2008.
17. Gnanapragasam, N.V.; Reddy, B.V.; Rosen, M.A. Feasibility of an energy conversion system in Canada involving large-scale integrated hydrogen production using solid fuels. *The International Journal of Hydrogen Energy* **2010**, *35*, 4788–4807.
18. Energy Technology Perspectives 2010: Scenarios & Strategies to 2030; International Energy Agency: Paris, 2010.
19. Shogren, J. (Ed.) *Encyclopedia of Energy, Natural Resource, and Environmental Economics*; Elsevier: Oxford, 2013.
20. Singh, D.P.; Kothari, R.; Tyagi, V.V., Eds. *Emerging Energy Alternatives for Sustainable Environment*; CRC Press: Boca Raton, FL, 2019.
21. Michaelides, E.E. *Energy, the Environment, and Sustainability*; CRC Press: Boca Raton, FL, 2018.
22. Lund, H.; Mathiesen, B.V. Energy system analysis of 100% renewable energy systems—The case of Denmark in years 2030 and 2050. *Energy* **2009**, *34*, 524–531.
23. Rosen, M.A. Combating global warming via non-fossil fuel energy options. *The International Journal of Global Warming* **2009**, *1*, 2–28.
24. Kreith, F; Goswami, D.Y., Eds. *Energy Management and Conservation Handbook*, 2nd Ed.; CRC Press: Boca Raton, FL, 2016.
25. Herring, H. Energy efficiency—A critical view. *Energy* **2006**, *31*, 10–20.
26. Karlsson, M.; Gebremedhin, A.; Klugman, S.; Henning, D.; Moshfegh, B. Regional energy system optimization—Potential for a regional heat market. *Applied Energy* **2009**, *86*, 441–451.
27. Ibrahim, H.; Ilinca, A.; Perron, J. Energy storage systems—Characteristics and comparisons. *Renewable and Sustainable Energy Reviews* **2008**, *12*, 1221–1250.
28. Dincer, I.; Rosen, M.A. *Thermal Energy Storage: Systems and Applications*, 2nd Ed.; Wiley: London, 2011.

29. Kotas, T.J. *The Exergy Method of Thermal Plant Analysis*, Reprint Ed.; Krieger: Melbourne, FL, 1995.
30. Dincer, I.; Rosen, M.A. *Exergy: Energy, Environment and Sustainable Development*, 2nd Ed.; Elsevier: Oxford, 2013.
31. Rosen, M.A. *Economics and Exergy: An Enhanced Approach to Energy Economics*; Nova Science Publishers: Hauppauge, NY, 2011.
32. Krewitt, W.; Simon, S.; Graus, W.; Teske, S.; Zervos, A.; Schafer, O. The 2°C scenario—A sustainable world energy perspective. *Energy Policy* **2007**, *35*, 4969–4980.
33. Rosen, M.A. Energy sustainability: A pragmatic approach and illustrations. *Sustainability* **2009**, *1*, 55–80.
34. Rosen, M.A. Towards energy sustainability: A quest of global proportions. *Forum on Public Policy Online: A Journal of the Oxford Round Table*, **2008**, *2008* (2), 1–20. Available at http://forumonpublicpolicy.com/Summer08papers/archivesummer08/rosen.pdf (accessed May 5, 2012).
35. Rosen, M.A. Issues, concepts and applications for sustainability. *Glocalism: Journal of Culture, Politics and Innovation* **2018**, *3*, 1–21.
36. Rosen, M.A. How can we achieve the UN sustainable development goals? *European Journal of Sustainable Development Research* **2017**, *1*(2), 6.

9

Energy: Physics

Milivoje M. Kostic

Introduction: From Work to Heat to General Energy Concept

Energy is a fundamental concept indivisible from matter and space, and energy exchanges or transfers are associated with all processes (or changes), thus indivisible from time. Actually, energy is "the building block" and fundamental property of matter and space, thus a fundamental property of existence. Energy transfer is needed to produce a process to change other system properties. Also, among all properties, energy is the only one that is directly related to mass and vice versa: $E = mc^2$ (known in some literature as *mass energy,* or mass-energy equivalence; the c is the speed of light in a vacuum); thus, mass and energy are interrelated. Mass and energy are a manifestation of each other and are equivalent; they have a holistic meaning of *mass* energy.[1,2]

Energy moves cars and trains, and boats and planes. It bakes food and keeps it frozen for storage. Energy lights our homes and plays our music. Energy makes our bodies alive and grow, and allows our minds to think. Through centuries, people have learned how to use energy in different forms in order to do work more easily and live more comfortably. No wonder that energy is often defined as *ability to perform work,* i.e., as a potential for energy transfer in a specific direction (displacement in force direction), thus achieving a purposeful process, as opposed to dissipative (less-purposeful) energy transfer in the form of heat. The above definition could be generalized as: "energy is providing existence, and if exchanged, it has the ability to perform change."[3,4]

Any and every material system in nature possesses energy. The structure of any matter and field is "energetic," meaning active; i.e., photon waves are traveling in space, electrons are orbiting around an atom nucleus or flowing through a conductor, atoms and molecules are in constant rotation, vibration,

TABLE 1 Material System Structure and Related Forces and Energies

Particles	Forces	Energies
Photons	Electromagnetic	Electromagnetic
Atom nucleus	Strong and weak	Nuclear
Electron shell, or electron flow	Electromagnetic	Electrical, magnetic, electromagnetic
Atoms/Molecules	Interatomic/molecular	Chemical
Molecules	Inertial kinetic due to random collision and potential Intermolecular	Sensible thermal
Molecules	Potential Intermolecular	Latent thermal
Molecules	Potential Intermolecular	Mechanical elastic
System mass	Inertial kinetic and gravitational	Mechanical kinetic and gravitational potential

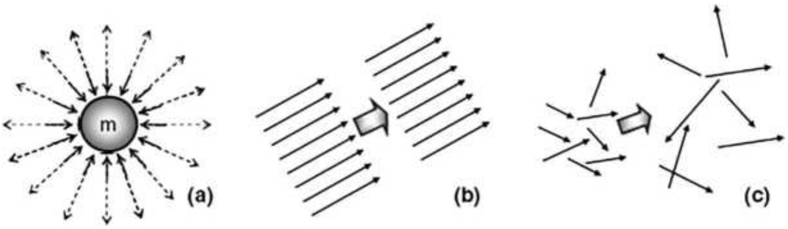


FIGURE 1 Different types of energy. (a) Potential gravitational and electromagnetic radiation. (b) Organized energy as work transfer. (c) Disorganized thermal energy as heat transfer.

or random thermal motion, etc., (see Table 1 and Figure 1). Thus, energy is a property of a material system (simply referred here as *system*) and, together with other properties, defines the system equilibrium state or existence in space and time.

Energy in transfer (E_{transfer}) is manifested as work (W) or heat (Q) when it is exchanged or transferred from one system to another, as explained next (see Figure 2).

Work

Work is a mode of energy transfer from one acting body (or system) to another resisting body (or system), with an acting force (or its component) in the direction of motion, along a path or displacement. A body that is acting (forcing) in motion-direction in time is doing work on another body that is resisting the motion (displacement rate) by an equal resistance force, including inertial force, in opposite direction of action. Energy transfer is a forced-displacement interaction between inertial material systems, in effect accelerating the reacting mass-energy sink system (i.e., its structure) at the expense of decelerating

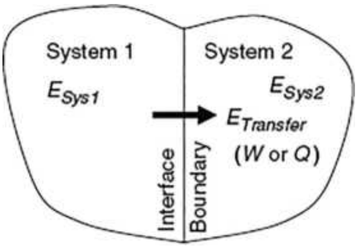


FIGURE 2 Energy as material system property and energy transfer from one system to another.

an acting-source system structure. The acting body (energy source) is imparting (transferring away) its energy to the resisting body (energy sink), and the amount of energy transfer is the work done by the acting onto the resisting body, equal to the product of the force component in the motion direction multiplied with the corresponding displacement, or vice versa (force multiplied by displacement component in the force direction, see Figure 3). If the force ($\vec{F}$) and displacement vectors ($d\vec{s} = d\vec{r}$) are not constant, then integration of differential work transfers from initial (1) to final state (2), defined by the corresponding position vectors $\vec{r}$, will be necessary (see Figure 4).

Work is a directional energy transfer. However, it is a scalar quantity like energy and is distinctive from another energy transfer in the form of *heat,* which is due to random motion (chaotic or random in all directions) and collisions of system molecules and their structural components.

Work transfer cannot occur without existence of a resisting body or system, or without finite displacement in the force direction. This may not always be obvious. For example, if we are holding a heavy weight or pushing hard against a stationary wall, there will be no work done against the weight or wall (neglecting their small deformations). However, we will be doing work internally due to the contraction and expansion of our muscles (thus force with displacement), thus converting (spending) a lot of chemical energy via muscle work and then dissipating it into thermal energy and heat transfer (sweating and getting tired).

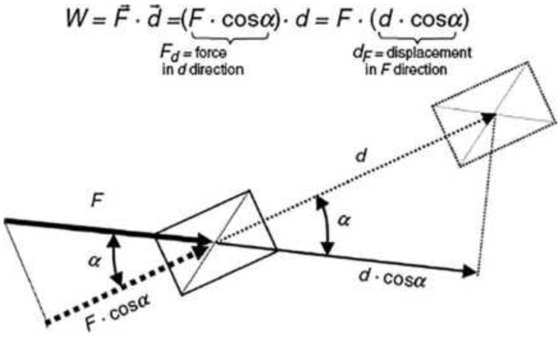


FIGURE 3 Work, force, and displacement.

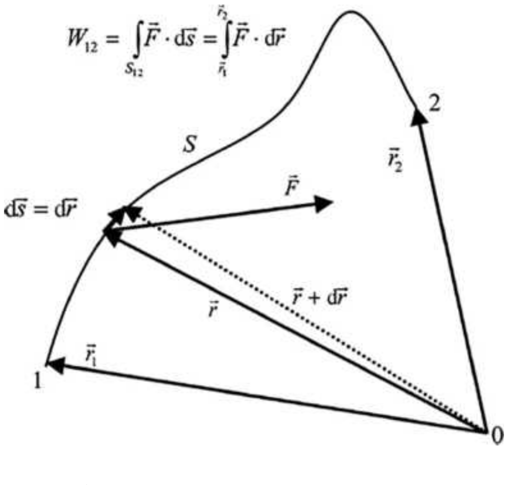


FIGURE 4 Work along arbitrary path.

Heat

Heat is another mode of energy transfer from one system at a higher temperature to another at a lower temperature due to their temperature difference. Fire was civilization's first great invention, long before people could read and write, and wood was the main heat source for cooking and heating for ages. However, true physical understanding of the nature of heat was discovered rather recently, in the middle of the 19th century, thanks to the development of the kinetic theory of gases. Thermal energy and its transfer as heat are defined as the energy associated with the random motion of atoms and molecules. The prior concept of heat was based on the *caloric* theory proposed by the French chemist Antoine Lavoisier in 1789. The caloric theory defines heat as a massless, colorless, odorless, and tasteless, fluid-like substance called the *caloric* that can be transferred or "poured" from one body into another. When caloric was added to a body, its temperature increased and vice versa. The caloric theory came under attack soon after its introduction. It maintained that heat is a substance that could not be created or destroyed. Yet it was known that heat can be generated indefinitely by rubbing hands together or from mechanical energy during friction, like mixing or similar. Finally, careful experiments by James P. Joule (1843) quantified correlation between mechanical work and heat, and thus put the caloric theory to rest by convincing the skeptics that heat was not the caloric substance after all. Although the caloric theory was totally abandoned in the middle of the 19th century, it contributed greatly to the development of thermodynamics and heat transfer. Actually, if conversion of all other energy types to heat (thermal energy "generation" from all types of "phlogistons") is included, i.e., conservation of energy in general, then caloric theory will be valid in general as stated by Clausius.[5]

Heat may be transferred by three distinctive mechanisms: conduction, convection, and thermal radiation (see Figure 5).[6] Heat conduction is the transfer of thermal energy due to interaction between the more energetic particles of a substance, like atoms and molecules (thus at higher temperature), and the adjacent, less energetic ones (thus at lower temperature). Heat convection is the transfer of thermal energy between a solid surface and the adjacent moving fluid, and it involves the combined effects of conduction and fluid motion. Thermal radiation is the transfer of thermal energy due to the emission of electromagnetic waves (or photons), which are products of random interactions between energetic electron shells of substance particles (thus related to the temperature). During those interactions, the electron shell energy level is changed, thus causing emission of photons, i.e., electromagnetic thermal radiation.

Joule's experiments of establishing equivalency between work and heat paved the way to establish the concept of internal thermal energy, to generalize the concept of energy, and to formulate the general law of energy conservation. The total internal energy includes all other possible but mechanical energy types or forms, including thermal, chemical, and nuclear energies. This allowed extension of

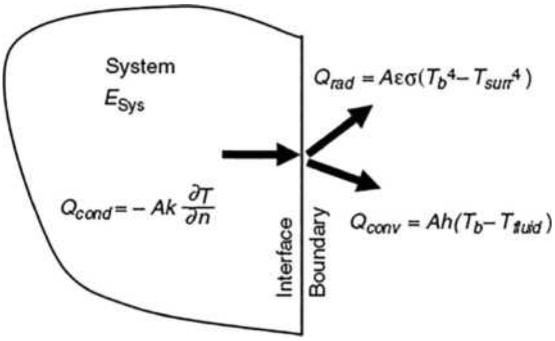


FIGURE 5 Heat as energy transfer by (a) conduction, (b) convection, and (c) radiation is due to difference in temperature.

the well-established law of mechanical energy conservation to the general law of energy conservation, known as the *first law of thermodynamics,* which includes all possible energy forms that a system could possess, and heat and all types of work as all possible energy transfers between the systems. The *law of energy conservation* will be elaborated later.

Energy

Energy is fundamental property of a physical system and refers to its potential to maintain a system identity or structure and to influence changes with other systems (via forced interaction) by imparting work (forced directional displacement) or heat (forced chaotic displacement/motion of a system's molecular or related structures). Energy exists in many forms: electromagnetic (including light), electrical, magnetic, nuclear, chemical, thermal, and mechanical (including kinetic, elastic, gravitational, and sound), where, for example, electromechanical energy may be kinetic or potential, while thermal energy represents overall potential and chaotic motion energy of molecules and/or related microstructure.[3,7,8]

Energy, Work, and Heat Units

Energy is manifested via work and heat transfer, with corresponding Force × Length dimension for work (N·m, kg_f·m, and lb_f·ft, in SI, metric, and English system of units, respectively), and the caloric units, in kilocalorie (kcal) or British thermal unit (Btu), the last two defined as heat needed to increase a unit mass of water (at reference state: specified pressure and temperature) for one degree of temperature in their respective units. Therefore, the water specific heat at reference state is 1 kcal/(kg °C) = 1 Btu/(lb °F) by definition, in metric and English system of units, respectively. It was demonstrated by Joule that 4187 Nm of work, when dissipated in heat, is equivalent to 1 kcal. In his honor, 1 N·m of work is named after him as 1 joule, or 1 J, the SI energy unit, also equal to electrical work of 1 W·s = 1 V·A·s. The SI unit for power, or work rate, is watt, i.e., 1 J/s = 1 W, and also corresponding units in other system of units, like Btu/hr, etc. Horsepower is defined as 1 HP = 550 lb_f·ft/sec = 745.7 *W*. Other common units for energy, work, and heat are given in Table 2.

Energy Forms and Classifications: Energy Transfer versus Energy Property

Any and all changes or processes (happening in space and time) are caused by energy exchanges or transfers from one substance *(system* or *subsystem)* to another (see Figure 2). A part of a system may be considered as a subsystem if energy transfer within a system is taking place, and inversely, a group of interacting systems may be considered as a larger isolated system, if they do not interact with the rest of the surroundings. Energy transfer may be in organized form (different types of work transfer due to different force actions) or in chaotic disorganized form (heat transfer due to temperature difference).

TABLE 2 Typical Energy Units with Conversion Factors

Energy Units	J	kWh	Btu
1 joule (J)	1	2.78×10^{-7}	9.49×10^{-4}
1 kilowatt hour (kWh)	3.6×10^{6}	1	3.412×10^{3}
1 kilocalorie (kcal = Cal = 1000 cal)	4187	1.19×10^{-3}	3.968
1 British thermal unit (Btu)	1055	2.93×10^{-4}	1
1 pound-force foot ($lb_f \cdot ft$)	1.36	3.78×10^{-7}	1.29×10^{-3}
1 electron volt (eV)	1.6×10^{-19}	4.45×10^{-26}	1.52×10^{-22}
1 horsepower × second (HP · s)	745.7	2.071×10^{-4}	0.707

Energy transfer into a system builds up energy potential or generalized force (called simply potential for short, like pressure, temperature, voltage, etc.) over energy displacement (like volume, entropy, charge, etc.). Conversely, if energy is transferred from a system, its energy potential is decreased. That is why net energy is transferred from a higher to a lower energy potential until the two equalize, due to virtually infinite probability of equipartition of energy over system microstructure, causing system equilibrium; otherwise, virtually impossible singularity of energy potential at infinite magnitude would result.[4]

There are many forms and classifications of energy (see Table 3), all of which could be classified as *microscopic* (or internal within a system microstructure) and/or *macroscopic* (or external as related to the system mass as a whole with reference to other systems or fields). Furthermore, energy may be "quasi-potential" (associated with a system equilibrium state and structure, i.e., system property) or "quasi-kinetic" (energy in transfer from one system or one structure to another, in the form of *work* or *heat*).

Every material system state is an equilibrium potential "held" by forces (space force fields); i.e., the forces "define" the potential and state—action and reaction; otherwise, a system will undergo dynamic change (in time) or quasi-kinetic energy exchange towards another stable equilibrium. Atoms (mass) are held by atomic and electromagnetic forces in small scale and by gravity in large scale (see Figure 1a); otherwise, mass would disintegrate ("evaporate" or radiate into energy) like partly in nuclear reactions—*nuclear* energy and/or electromagnetic radiation. Molecules are held by electrochemical bonding (e.g., valence) forces (chemical reactions—*chemical* energy). Liquids are held by latent intermolecular forces (gas condensation, when kinetic energy is reduced by cooling—*latent thermal* energy). Solids are held by "firm" intermolecular forces (freezing/fusing/solidification when energy is further reduced by cooling—*latent thermal* energy again). *Sensible thermal* energy represents intermolecular potential energy and energy of random molecular motion and is related to the temperature of a system. "Holding" potential forces may be "broken" by energy transfer (e.g., radiation, heating, high-energy particle interactions, etc.). States and potentials are often "hooked" (i.e., stable) and thus need to be "unhooked" (or to "be broken") to overcome existing "threshold" or equilibrium, like in igniting combustion, starting nuclear reaction, etc.

As stated above, energy transfer can be directional (purposeful or organized) and chaotic (dissipative or disorganized). For example, mass in motion (*mechanical kinetic* energy) and electron in motion (*electrical kinetic* energy) are organized kinetic energies (Figure 1b), while *thermal* energy is disorganized chaotic energy of motion of molecules and atoms (Figure 1c). System energy may be defined with reference to position in a force field, like *elastic potential* (stress) energy, *gravitational potential* energy, or *electromagnetic* field energy. There are many different energy forms and types (see Table 3). We are usually not interested in (absolute) energy level, but in the change of energy (during a process) from an initial state (*i*) to a final state (*f*), and thus zero reference values for different energy forms are irrelevant, and often taken arbitrarily for convenience. The following are basic correlations for energy changes of several typical energy forms, often encountered in practice: motion kinetic energy ($E_K = KE$) as a function of system velocity (V); potential elastic deformational, e.g., pressure or spring elastic energy ($E_{\text{Pdeff}} = E_{\text{Pp}} = E_{\text{Ps}}$), as a function of spring deformation displacement (x); gravitational potential energy ($E_{\text{Pg}} = PE_{\text{g}}$) as a function of gravitational elevation (z); and sensible thermal energy ($E_{\text{U}} = U$) as a function of system temperature (T):

$$\Delta E_k = \frac{1}{2} m\left(V_{\text{f}}^2 - V_{\text{i}}^2\right) \qquad \Delta E_{\text{Ps}} = \frac{1}{2} k\left(x_{\text{f}}^2 - x_{\text{i}}^2\right) \tag{1}$$

$$\Delta E_{\text{Pg}} = mg\left(z_{\text{f}} - z_{\text{i}}\right) \qquad \Delta E_{\text{U}} = mc_{\text{v}}\left(T_{\text{f}} - T_{\text{i}}\right)$$

If the reference energy values are taken to be zero when the above initial (i) variables are zero, then the above equations will represent the energy values for the final values (f) of the corresponding variables. If the corresponding parameters, spring constant k, gravity g, or constant-volume specific heat c_v, are not constant, then integration of differential energy changes from initial to final state will be necessary.

TABLE 3 Energy Forms and Classifications

Energy										
Scale							Type		Transfer (release)	
							Kinetic (Motion)		Kinetic (Motion)	
MACRO/External (mass Based)	MICRO/Internal (structure Based)	Energy Form (energy Storage)		Energy Process	Potential (state of field)	Directional[a]	Chaotic dissipative	Work[b] Directional[a]	Heat Dissipative	
		Mechanical								
x		—Kinetic	$mV^2/2$	Acceleration			X		X	
x		—Gravitational[c]	mgz	Elevation		X			X	
	x	—Elastic	$kx^2/2$ or $PV = mP/\rho$	Deformation		X			X	
		Thermal	U_{th}							
	x	—Sensible	$U_{th} = mc_{avg}T$	Heating				X		X
	x	—Latent	$U_{th} = H_{latent}$	Melting Evaporation		X				X
	x	Chemical	U_{ch}	Chemical reaction		X				X
	x	Nuclear	U_{nucl}	Nuclear reaction		X				X
		Electrical	E_{el}							
	x	—Electro kinetic	V(It) or $LI^2/2$	Electro-current flow			X		X	
	x	—Electrostatic	$(It)^2/(2C)$	Electro-charging		X			X	
		Magnetic	E_{magn}	Magnetization		X			X	
	x[d]	Electromagnetic	E_{el_mag}	Radiation			X[d]	X[e]	X	

[a] Electromechanical kinetic energy type (directional/organized, the highest energy quality) is preferable since it may be converted to any other energy form/type with high efficiency.
[b] All processes (involve energy transfer) are to some degree irreversible (i.e., dissipative or chaotic/disorganized).
[c] Due to mass position in a gravitational field.
[d] Electromagnetic form of energy is the smallest known scale of energy.
[e] Thermal radiation.

Energy transfer via work *W* (net-out), and heat transfer *Q* (net-in), may be expressed for reversible processes as product of related energy potentials (pressure *P*, or temperature *T*) and corresponding energy displacements (change of volume *V* and entropy *S*, respectively), i.e.:

$$W_{12} = \vec{F} \cdot \vec{d} = \left[\underbrace{\left(P \cdot A\vec{n} \right) . \vec{d}}_{\Delta V} \right] = P \cdot V_{12} = \Bigg| \underbrace{\int_{V1}^{V2} P \cdot dV}_{P \neq \text{const}} \tag{2}$$

$$Q_{12} = T \cdot \Delta S_{12|T \neq const} = \int_{S1}^{S2} T \cdot dS \tag{3}$$

Note, in Eq. 2, that force cannot act at a point but is distributed as pressure (P) over some area A (with orthogonal unit vector $\vec{n}$), which when displaced will cause the volume change ΔV. Also, note that it is custom in some references to denote heat transfer "in" and work transfer "out" as positive (as they appear in the above equations and a heat engine). In general, "in" (means "net-in") is negative "out" (means "net-out") and vice versa.

In general, energy transfer between systems is taking place at the system boundary interface and is equal to the product of energy potential or generalized force and the corresponding generalized displacement:[7–9]

$$\begin{aligned}\delta E_{\text{Transfer}} &= \delta Q_{\text{netIN}} - \left[\sum \delta W_{\text{netOUT}}\right] \\ &= \delta Q_{\text{netIN}} + \left[\sum \delta W_{\text{netIN}}\right] \\ &= T\text{d}s + \left[\underbrace{-PdV}_{\text{COMPR.}} + \underbrace{\sigma dA}_{\text{STRETCHING.}} + \underbrace{\tau d(A \cdot s)}_{\text{SHEARING.}}\right] \\ &\quad + \underbrace{Vdq}_{\text{CHARGING}} + \underbrace{\vec{E} \cdot d(V\vec{P})}_{\text{POLARIZATION}} \\ &\quad + \underbrace{\mu_\circ \vec{H} \cdot d(V\vec{P})}_{\text{MAGNETIZATION}} + \underbrace{\cdots}_{\text{ETC.}}\end{aligned} \tag{4}$$

where the quantities after the last equal sign are temperature and entropy; pressure and volume; surface tension and area; tangential stress and area with tangential displacement, voltage, and electrical charge; electric field strength and volume-electric dipole moment per unit volume product; and permeability of free space, magnetic field strength, and volume-magnetic dipole moment per unit volume product, respectively.

The total system energy stored within the system, as energy property, is:

$$E_{\text{Sys}} = \underbrace{E_K + E_{Pg} + E_{P\text{deff.}}}_{E_{\text{Mechanical}}} + \underbrace{\underbrace{E_{\text{Uth}}}_{\text{Thermal}} + E_{\text{Ch}} + E_{\text{Nucl}} + E_{\text{EI}} + E_{\text{Magn}} + \underbrace{\cdots}_{\text{Etc.}}}_{\text{Internal(total)}} \tag{5}$$

where the quantities after the equal sign are kinetic, potential gravitational, potential elastic deformational, thermal, chemical, nuclear, electric, magnetic energies, etc.

First Law of Energy Conservation: Work-Heat-Energy Principle

Newton formulated the general theory of motion of objects due to applied forces (1687). This provided for concepts of mechanical work, kinetic and potential energies, and development of solid-body mechanics. Furthermore, in absence of non-mechanical energy interactions, excluding friction and other dissipation effects, it is straightforward to derive (and thus prove) energy conservation, i.e.:

$$\underbrace{\int_{s1}^{s2} F_s \text{d}s}_{W_{\text{Fs}}} = \int_{s1}^{s2} \left(\underbrace{\frac{\text{d}V_s}{\text{d}t}}_{a_s} \right) \text{d}s = \int_{s1}^{s2} \left(\underbrace{m \frac{\text{d}V_s}{\text{d}s} \left\{ \frac{\text{d}s}{\text{d}t} \right\}}_{V_s} \right) \text{d}s$$

$$= \int_{V_1}^{V_2} mV_s \mathrm{d}V_s = \underbrace{\frac{1}{2}m\left(V_2^2 - V_1^2\right)}_{KE_2-KE_1}$$

$$\underbrace{\left(E_{\text{Transfer}} = W_{Fs}\right)}_{\text{work-energy transfer}} = \underbrace{\left(KE_2 - KE_1 = E_2 - E_1\right)}_{\text{energy transfer}} \tag{6}$$

The above correlation is known as the *work-energy principle*. The work-energy principle could be easily expended to include work of gravity force and gravitational potential energy as well as elastic spring force and potential elastic spring energy.

During a free gravity fall (or a free bounce) without air friction, for example, the potential energy is being converted to kinetic energy of the falling body (or vice versa for free bounce) and at any time the total mechanical energy (sum of kinetic and potential mechanical energies) is conserved, i.e., stays the same (see Figure 6). The mechanical energy is also conserved if a mass freely vibrates on an ideally elastic spring, or if a pendulum oscillates around its pivot, both in the absence of dissipative effects, like friction or non-elastic deformation. In general, for work of conservative forces only, the mechanical energy, E_{mech}, for N isolated systems is conserved since there is no dissipative conversion into thermal energy and thus no heat transfer, i.e.:

$$E_{\text{mech}} = E_K + E_{pg} + E_{Ps} = \sum_{i=1}^{N}\left(\frac{1}{2}mv^2 + mgz + \frac{1}{2}kx^2\right) = \text{const} \tag{7}$$

The mechanical work-energy concept could also be expended to fluid motion by inclusion of elastic-pressure force and potential elastic-pressure energy (referred to in some references as flow work; however, note that elastic- pressure energy is a system property while flow work is related energy transfer; see the *Bernoulli* equation below). For flowing or stationary fluid without frictional effects, the mechanical energy is conserved, including fluid elastic- pressure energy, $PV = mP/\rho$ (where V is volume, whereas v is used for velocity here), as expressed by the Bernoulli or hydrostatic equations below (see also Figure 7).[3]

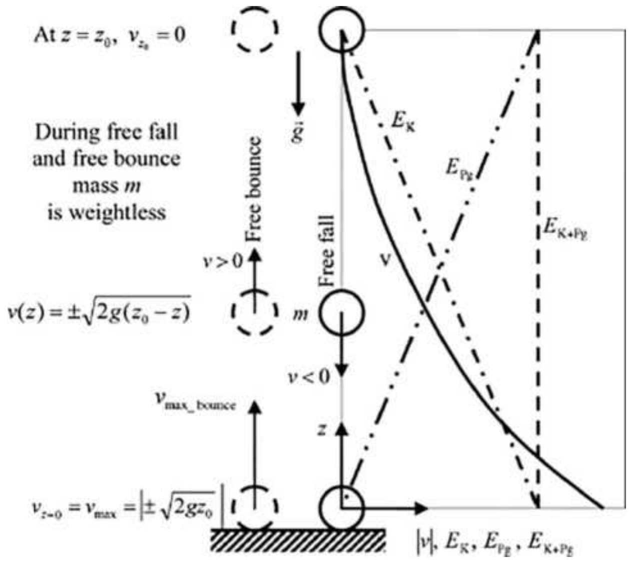


FIGURE 6 Energy and work due to conservative gravity force.

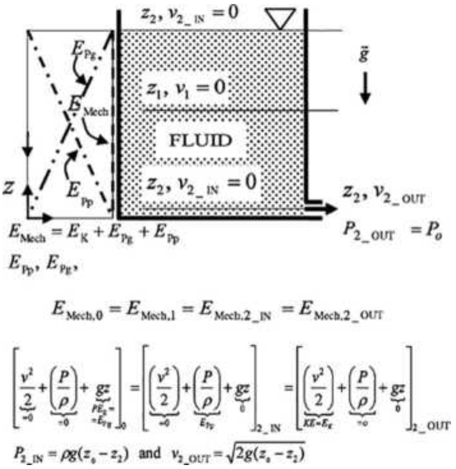


FIGURE 7 Conservation of fluid mechanical energy: Bernoulli equation, hydrostatic equation, and Torricelli's orifice velocity.

$$\frac{E_{\text{Mech}}}{m} = \frac{1}{m}\left(E_K + E_{Ps} + E_{Pg}\right) = \underbrace{\frac{v^2}{2} + \frac{P}{\rho} + gz}_{\substack{\text{Bernoulli}\\ \text{equation}}}\Bigg|_{v=0} = \underbrace{\frac{P}{\rho} + gz}_{\substack{\text{Hydrostatic}\\ \text{equation } (v=0)}} = \text{const.} \tag{8}$$

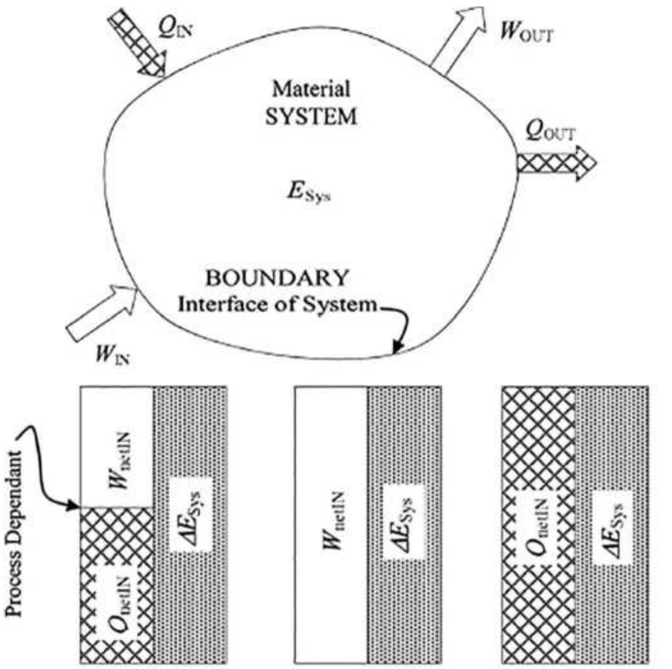
Work against inertial and/or conservative forces (also known as internal, or volumetric, or space potential field) is path independent, and during such a process, the mechanical energy is conserved. However, work of nonconservative, dissipative forces is process path dependent and part of the mechanical energy is converted (dissipated) to thermal energy (see Figure 8a).

When work of non-conservative (dissipative) forces, W_{nc}, is exchanged between N isolated systems, from an initial (i) to final state (f), then the total mechanical energy of all systems is reduced by that work amount, i.e.:

$$W_{nc,\, i\to f} = \left(\sum_{i=1}^{N} E_{\text{mech},\, i}\right)_i - \left(\sum_{i=1}^{N} E_{\text{mech},\, i}\right)_f \tag{9}$$

Regardless of the traveled path (or displacement), the work against conservative forces (like gravity or elastic spring in above cases) in absence of any dissipative forces will depend on the final and initial position (or state) only. However, the work of non-conservative, dissipative forces (W_{nc}) depends on the traveled path since the energy is dissipated during the force displacement, and mechanical energy will not be conserved but in part converted (via dissipation and heat transfer) into thermal energy. This should not be misunderstood with the total energy conservation, which is always the case, and it must include both work and heat transfer.

As already stated, there are many different types of work transfer into (or out of) a system that will change the corresponding energy form stored in the system (system properties). In addition to work, energy may be transferred as heat caused by temperature difference and associated with change of the thermal energy of a system. Furthermore, different forms of stored energy are often coupled so that one type of energy transfer may change more than one form of stored energy, particularly due to inevitable dissipative conversion of work to heat, i.e., to internal thermal energy. Conversely, heat and thermal



(a) Both Q and W
Work-Heat interaction
(process dependent)

(b) W only (No Q)
Work interaction
(adiabatic process)

(c) Q only (No W)
Heat interaction
(caloric process)

FIGURE 8 System energy and energy boundary interactions (transfers) for (a) arbitrary, (b) adiabatic, and (c) caloric processes.

energy may be converted into other energy forms. In the absence of a powerful nuclear reaction (conversion of mass into energy, $E = mc^2$, is virtually negligible), we then may assume that the mass and energy are conserved separately for an isolated system, a group of isolated systems, or for the Universe. Since the material system structure is of particulate form, then systems' interactions (collisions at different scale sizes) will exchange energy during the forced displacement— and similarly to the mechanical energy conservation—the totality of all forms of energy will be conserved[7,8] (also see Figure 8), which could be expressed as:

$$\underbrace{\sum_{\text{All } i\text{'s}} E_{i,\text{ Trans.}}}_{\text{BOUNDARY}} = \underbrace{\Delta E_{\text{Change}}}_{\text{SYSTEM}} \text{ or}$$

$$\underbrace{\sum_{\text{All } j\text{'s}} W_{j,\text{ netIN}} = \sum_{\text{All } k\text{'s}} Q_{k,\text{ netIN}}}_{\text{BOUNDARY}} \tag{10}$$

$$= \underbrace{\Delta E_{\text{netIncrease}}}_{\text{SYSTEM}} = \Delta E_{\text{Sys}}$$

Energy interactions or transfers across a system boundary, in the form of work, $W_{\text{netIN}} = \sum W_{\text{IN}} - \sum W_{\text{OUT}} = -\left(\sum W_{\text{OUT}} - \sum W_{\text{IN}}\right) = -\sum W_{\text{netOUT}}$, and heat, $Q_{\text{netIN}} = \sum Q_{\text{IN}} - \sum Q_{\text{OUT}}$, will change the system total energy, $\Delta E_{\text{Sys}} = E_{\text{Sys},\,2} - E_{\text{Sys},\,1}$. Different energy-type transfers are process (or process path) dependent, for the same ΔE_{Sys} change, except for special two cases: for adiabatic processes with work interaction only (no heat transfer, but fully irreversible, all work to heat generation) or for caloric processes with heat interaction only (no work transfer) (see Figure 8a–c). For the latter caloric processes without work interactions (no volumetric expansion or other mechanical energy changes), the internal thermal energy is conserved by being transferred from one system to another via heat transfer only, known as *caloric*. This demonstrates the value of the caloric theory of heat that was established by Lavoisier and Laplace (1789), the great minds of the 18th century. Ironically, the caloric theory was creatively used by Sadi Carnot (1824) to develop the concept of reversible cycles for conversion of caloric heat to mechanical work as it "flows" from high- to low-temperature reservoirs that later helped in dismantling the caloric theory. The caloric theory was discredited by establishing the "heat equivalent of work, e.g., mechanical equivalent of heat" by Mayer (1842) and experimentally confirmed by Joule (1843), which paved the way for establishing the *first law of energy conservation* and new science of *thermodynamics* (Clausius, Rankine, and Kelvin, 1850 and later). Prejudging the caloric theory now as a "failure" is unfair and unjustified since it made great contributions in calorimetry and heat transfer, and it is "valid" for caloric processes (without work interactions).[5] The coupling of work-heat interactions and conversion between thermal and mechanical energy are outside of the caloric theory domain and are further developed within the first and second laws of thermodynamics.

The first law of energy conservation for the control volume (CV, with boundary surface BS) flow process (see Figure 9) is:

$$\underbrace{\frac{d}{dt}E_{\text{CV}}}_{\substack{\text{RATE OF ENERGY}\\ \text{CHANGE IN CV}}} = \underbrace{\sum_{\text{BS}} \dot{Q}_{\text{netIN},\,i}}_{\substack{\text{BS TRANSFER}\\ \text{RATE OF HEAT}}} - \underbrace{\sum_{\text{BS}} \dot{W}_{\text{netOUT},\,i}}_{\substack{\text{BS TRANSFER}\\ \text{RATE OF WORK}}} + \underbrace{\sum_{\text{IN}} \dot{m}_j\left(e+Pv\right)_j}_{\substack{\text{ENERGY TRANSPORT}\\ \text{RATE WITH MASS IN}}} - \underbrace{\sum_{\text{OUT}} \dot{m}_K\left(e+Pv\right)_K}_{\substack{\text{ENERGY TRANSPORT}\\ \text{RATE WITH MASS OUT}}} \tag{11}$$

The first law of energy conservation equation for a material system in a differential volume per unit of volume around a point (x, y, z) in a flowing fluid is:[8]

$$\underbrace{\rho\frac{De}{Dt}}_{\text{energy change in time}} = \underbrace{-\vec{V}.(\nabla P)}_{\substack{\text{work rate}\\ \text{of normal pressure stresses}}} + \underbrace{\nabla.\left(\vec{V}.\tau_{ij}\right)}_{\substack{\text{work rate}\\ \text{of shearing stresses}}} + \underbrace{\nabla\cdot\left(k\nabla T\right)}_{\substack{\text{heat rate}\\ \text{via thermal conduction}}} \tag{12}$$

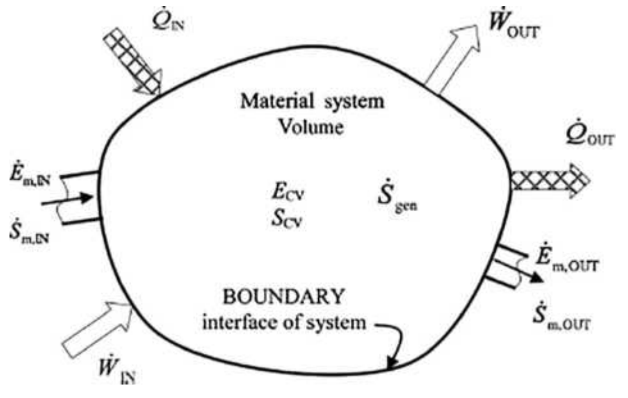


FIGURE 9 Control volume (CV) energy and entropy, and energy and entropy flows through the boundary interface of the CV.

Where $e = \hat{u} + \frac{\vec{V}^2}{2} + gz$.

Eq. 12 reduces after substitution, $\nabla \cdot \left(\vec{V} \cdot \tau_{ij}\right) = \vec{V}\left(\nabla \cdot \tau_{ij}\right) + \Phi$, and using the momentum equation, to:

$$\rho \frac{\mathrm{D}\hat{u}}{\mathrm{D}t} = -p\left(\nabla . \vec{V}\right) + \Phi_k + \Phi + \nabla \cdot \left(k \nabla T\right) \tag{13}$$

where $\Phi_k = k\left(\nabla \cdot \vec{V}\right)^2$ is the bulk-viscosity dissipation and Φ is the shear-viscosity dissipation function, which is the rate of mechanical work conversion to internal thermal energy for a differential volume per unit of volume, with [W/m^3] unit, given for Newtonian fluid as:

$$\begin{aligned} \Phi = \frac{\mathrm{d}\dot{W}_\Phi}{\mathrm{d}V} = &\left[\left(\frac{\partial u}{\partial x}\right)^2 \left(\frac{\partial v}{\partial y}\right)^2 \left(\frac{\partial w}{\partial z}\right)^2\right] \\ &+\mu\left[\left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\right)^2 + \left(\frac{\partial w}{\partial y} + \frac{\partial v}{\partial z}\right)^2\right. \\ &\left.+\left(\frac{\partial u}{\partial z} + \frac{\partial w}{\partial x}\right)^2\right] - \frac{2}{3}\mu\left(\nabla \cdot \vec{V}\right)^2 \end{aligned} \tag{14}$$

The power or work rate of viscous dissipation (irreversible conversion of mechanical to thermal energy) in a control volume V is:

$$\dot{W}_\Phi = \int_V \Phi dV \tag{15}$$

Second Law of Energy Degradation: Entropy and Exergy

Every organized or directional kinetic energy will, in part or in whole (and ultimately in whole), disorganize/dissipate in all directions within the microstructure of a system (in time over its mass and space) into disorganized thermal energy. Entropy represents the measure of energy disorganization or random energy redistribution to smaller-scale structure and/or space, per absolute temperature level. Contrary to energy and mass, which are conserved in the universe, entropy is continuously generated (increased) due to continuous redistribution and disorganization of energy in transfer and thus degradation of the quality of energy ("spreading" of energy towards and over lower potentials in time until equipartitioned over system structure and space). Often, we want to extract energy from one system in order to purposefully change another system, thus to transfer energy in organized form (as work, the ultimate energy quality). No wonder that energy is defined as *ability to perform work,* and a special quantity, *exergy,* is defined as the maximum possible work that may be obtained from a system by bringing it to the equilibrium in a process with reference surroundings (called reference *dead state*). The maximum possible work will be obtained if we prevent energy disorganization, thus with limiting *reversible* processes where the existing non-equilibrium is conserved (but rearranged and could be reversed) within interacting systems. Since the energy is conserved during any process, only in ideal reversible processes will entropy (measure of energy disorganization or degradation) and exergy (maximum possible work with regard to the reference surroundings) be conserved, while in real irreversible processes, entropy will be generated and exergy will be partly (or even fully) destroyed or permanently lost. Therefore, conversion of other energy types in thermal energy or heat generation and transfer are universal manifestations of all natural and artificial (man-made) processes, where all organized potential and/or quasi-kinetic energies are disorganized or dissipated in the form of thermal energy, in irreversible and spontaneous processes.

Reversibility and Irreversibility: Energy Transfer and Disorganization, and Entropy Generation

Energy transfer (when energy moves from one system or subsystem to another) through a system boundary and in time is of kinetic nature and may be directionally organized as work or directionally chaotic and disorganized as heat. However, the net-energy transfer is in one direction only, from a higher to a lower energy potential, and the process cannot be reversed. Thus, *all processes are irreversible* in the direction of decreasing energy potential (like pressure and temperature) and increasing energy displacement (like volume and entropy) as a consequence of energy and mass conservation in the universe. This implies that the universe (as isolated and unrestricted system) is expending in space with entropy generation (or increase) as a measure of continuous energy degradation, i.e., energy redistribution and disorganization per unit of absolute temperature. It is possible in the limit to have an energy transfer process with infinitesimal potential difference (still from higher to infinitesimally lower potential, P). Then, if infinitesimal change of potential difference direction is reversed ($P + \mathrm{d}P \rightarrow P - \mathrm{d}P$, with infinitesimally small external energy, since $\mathrm{d}P \rightarrow 0$), the process will be reversed too, which is characterized by infinitesimal entropy generation, thus in the limit, without energy degradation (no further energy disorganization) and no entropy generation—thus achieving a limiting, ideal *reversible process*. Such processes at infinitesimal potential differences, called quasi-equilibrium processes, maintain the system space equilibrium at any instant but with incremental changes in time. The quasi-equilibrium processes are reversible and vice versa. In effect, the quasi-equilibrium reversible processes are infinitely slow processes at infinitesimally small potential differences, but they could be reversed to any previous state and forwarded to any future equilibrium state, without any "permanent" change to the rest of the surroundings. Therefore, the changes are "fully" reversible and, along with their rate of change and time, completely irrelevant, as if nothing is effectively changing—time is irrelevant as if it does not exist since it could be reversed (no permanent change and relativity of time). Since the real time cannot be reversed, it is a measure of permanent changes, i.e., irreversibility, which is, in turn, measured by entropy generation. In this regard, time and irreversible entropy generation are related.[2]

Entropy and the Second Law of Thermodynamics

Entropy is also a system property, which, together with energy, defines its equilibrium state and actually represents the system energy displacement or random energy disorganization (dissipation) per absolute temperature level over its mass and the space it occupies. Therefore, entropy as property of a system, for a given system state, is the same regardless whether it is reached by reversible heat transfer (Eq. 16) or irreversible heat or irreversible work transfer (caloric or adiabatic processes, Eq. 17). For example, an ideal gas system entropy increase will be the same during a reversible isothermal heat transfer and reversible expansion to a lower pressure (heat-in equal to expansion work-out), as during an irreversible adiabatic unrestricted expansion (no heat transfer and no expansion work) to the same pressure and volume, as illustrated in Fig. 10a and b, respectively.

If heat or work at higher potential (temperature or pressure) than necessary is transferred to a system, the energy at excess potential will dissipate spontaneously to a lower potential (if left alone) before a new equilibrium state is reached, with entropy generation, i.e., increase of entropy (energy degradation per absolute temperature level). A system will "accept" energy if it is transferred at minimum necessary (infinitesimally higher) or higher potential with regard to the system energy potential. Furthermore, the higher potential energy will dissipate and entropy increase will be the same as with minimum necessary potential, like in reversible heating process, i.e.

$$dS = \frac{\delta Q}{T} \text{ or } S = \int \frac{\delta Q}{T} + S_{\text{ref}} \tag{16}$$

However, source entropy will decrease to a smaller extent over higher potential, thus resulting in overall entropy generation for the two (or all) interacting systems, which may be considered as a combined isolated system (no energy exchange with the rest of the surroundings). The same is true for energy exchange between different system parts (could be considered as subsystems) at different energy potentials (non-uniform, not at equilibrium at a given time). Energy at higher potential (say, close to boundary within a system) will dissipate ("mix") to parts at lower energy potential with larger entropy increase than decrease at higher potential, resulting in internal irreversibility and entropy generation, i.e., energy "expansion" from higher potential over more mass and/or space with lower potential. Entropy is not only displacement for heat as often stated but also displacement for any energy dissipation (energy disorganization) and the measure of irreversibility. Examples are spontaneous heat transfer or throttling (unrestricted expansion) where "work potential" is lost ($W_{\text{pot, LOSS}}$), i.e., spontaneously converted (dissipated) in thermal energy, known as heat generation ($Q_{\text{gen}} = W_{\text{pot, LOSS}}$), resulting in entropy generation:

$$S_{\text{ref}} = \frac{Q_{\text{gen}}}{T} \geq \text{where } Q_{\text{gen}} = W_{\text{pot, LOSS}} \tag{17}$$

Eq. 17 may refer to the entropy generation rate at any instant within infinitesimal system or for the entropy generated during a larger process duration for an integral system. Therefore, entropy generation is the fundamental measure of irreversibility or "permanent change" that cannot be reversed, and therefore cannot be negative. This is true in general for closed and open processes and without any exception, thus representing the most general statement of the second law of energy degradation.[10]

However, there are two classical statements of the second law (both negative, about impossibility; see Figure 11).[7–9] One is the *Kelvin-Planck statement,* which expresses the never-violated fact that it is impossible to obtain work in a continuous cyclic process *(perpetual mobile)* from a single thermal reservoir (100% efficiency impossible, since it is not possible to spontaneously create non-equilibrium within the single reservoir in equilibrium), and the second is the *Clausius statement,* which refers to the direction of heat transfer, expressing the never-violated fact that it is impossible for heat transfer to take place spontaneously by itself (without any work input) from a lower- to a higher-temperature thermal reservoir: it is impossible to spontaneously create non-equilibrium. Actually, the two statements imply each other and thus are the same, as well as imply that all reversible cycles between the two temperature reservoirs (or all reversible processes between the two states) are the most (and equally) efficient with regard to extracting the maximum work out of a system or requiring the minimum possible work into the system (thus conserving the existing non-equilibrium).

Even though directionally organized energy transfer as work does not possess or generate any entropy (no energy disorganization, Figure 1b), it is possible to obtain work from the equal amount of disorganized thermal energy or heat if such process is reversible. There are two typical reversible processes where disorganized heat or thermal energy could be entirely converted into organized work, and vice versa. Namely, they are the following:[2]

1. Reversible expansion at constant thermal energy, e.g., isothermal-heating ideal-gas expansion ($\delta W = \delta Q$) (Figure 10a)
2. Reversible adiabatic expansion ($\delta W = -\mathrm{d}U$)

During a reversible isothermal heat transfer and expansion of an ideal-gas system (S) for example (Figure 10a), the heat transferred from a thermal reservoir (R) will reduce its entropy for ΔS_R

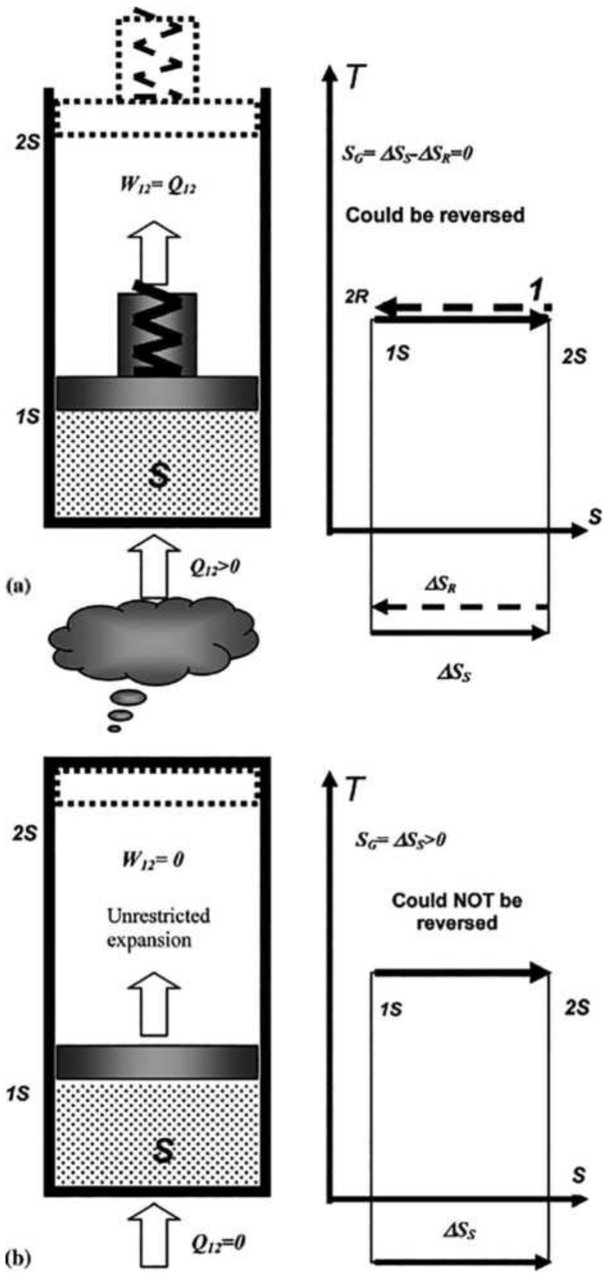


FIGURE 10 (a) Isothermal reversible heat transfer and restricted reversible expansion. (b) Adiabatic unrestricted irreversible expansion of the same initial system to the same final state (thus the same system entropy change).

magnitude, while ideal gas expansion in space (larger volume and lower pressure) will further disorganize its internal thermal energy and increase the gas entropy for ΔS_S, and in the process, an organized expansion work, equivalent to the heat transferred, will be obtained ($W_{12} = Q_{12}$). The process could be reversed, and thus, it is a reversible process with zero total entropy generation ($S_G = \Delta S_S - \Delta S_r = 0$). On the other hand, if the same initial system (ideal gas) is adiabatically expanded without any restriction (Figure 10b, thus zero expansion work) to the same final state, but without heat transfer, the system internal energy will remain the same but more disorganized over the larger volume, resulting in the same entropy increase as during the reversible isothermal heating and restricted expansion. However, this process cannot be autonomously reversed, since no work was obtained to compress back the system, and indeed the system entropy increase represents the total entropy generation ($S_G = \Delta S_S > 0$). Similarly, during reversible adiabatic expansion, the system internal thermal energy will be reduced and transferred in organized expansion work with no change of system entropy (isentropic process), since the reduction of disorganized internal energy and potential reduction of entropy will be compensated with equal increase of disorder and entropy in expanded volume. The process could be reversed back and forth (like elastic compression-expansion oscillations of a system) without energy degradation and entropy change, thus a reversible isentropic process. In reversible processes, energy is exchanged at minimum needed, not higher-than-needed potential, and isolated interacting systems as a whole do not undergo any energy-potential- related degradation/disorganization and with total conservation of entropy. The total non-equilibrium is conserved by reversible energy transfer within interacting systems, i.e., during reversible processes.

Heat Engines

Heat engines are devices undergoing thermomechanical cycles (converting thermal into mechanical energy), similar to the one in Figure 12, with mechanical expansion and compression network ($W = Q_h - Q_c$), obtained as the difference between the heat transferred to the engine from the

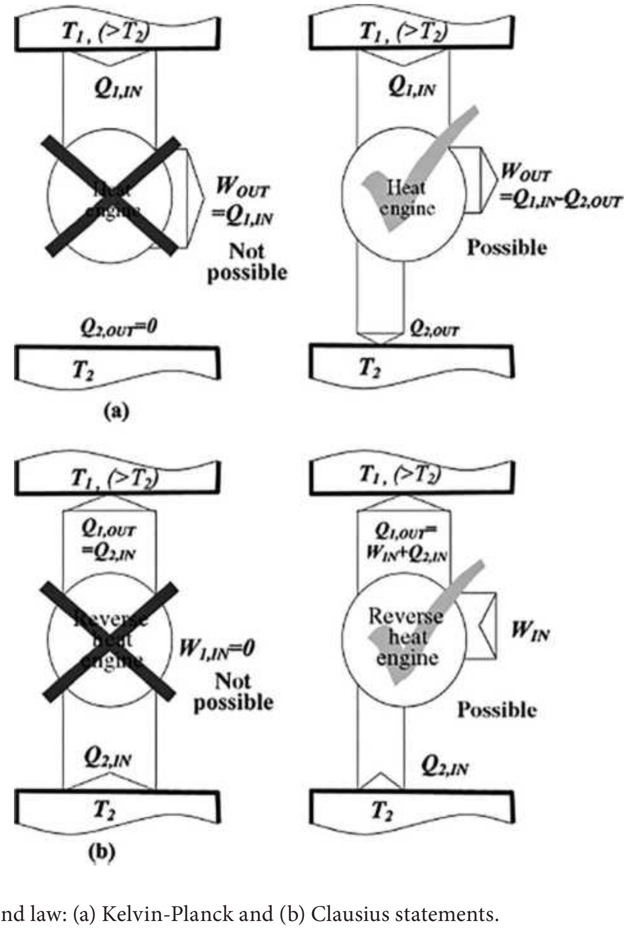


FIGURE 11 The second law: (a) Kelvin-Planck and (b) Clausius statements.

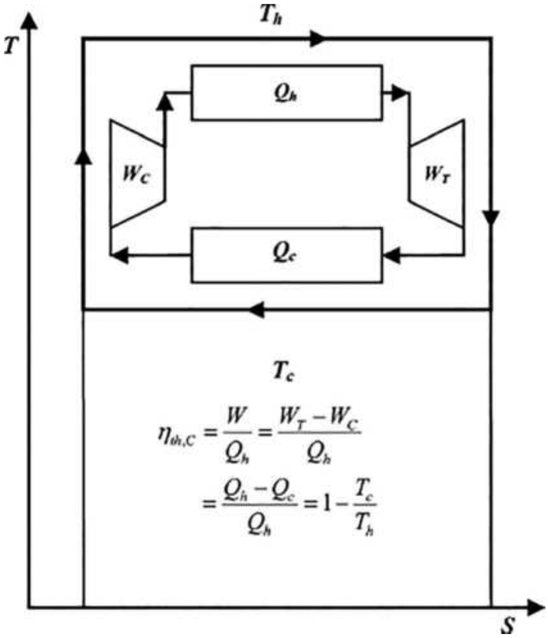


FIGURE 12 Heat engine ideal Carnot cycle.

high-temperature heat reservoir (at T_h) and rejected to a low-temperature (cold) heat reservoir (at T_c), thus converting part of the thermal energy into mechanical work. In a close-system cycle, the net-work-out is due to net-work of thermal expansion and thermal compression. Therefore, heat engine cycle cannot be accomplished without two thermal reservoirs at different temperatures, one at higher temperature to accomplish thermal expansion and work out, and another at lower temperature to accomplish thermal compression to initial volume and thus complete the cycle.[10]

The combustion process itself is an irreversible one, where chemical energy (electrochemical energy binding atoms in reactants' molecules) is chaotically released during combustion, i.e., converted in random thermal energy of products' molecules and cannot be fully converted into directional work energy. The *second law of thermodynamics* limits the maximum amount of work that could be obtained from thermal energy between any two thermal reservoirs at different temperatures, hot T_h and cold T_c, by using the ideal, reversible *Carnot cycle* (see Figure 12), with thermal efficiency given by Eq. 18. As an example, consider $T_c = 293$ K and $T_h = 2273$ K, then:

$$\eta_{\text{th, C\_ad}} = \frac{W}{Q_h} = \frac{Q_h - Q_c}{Q_h} = 1 - \frac{T_c}{T_h}\bigg|_{T_h = T_{\text{ad}} = 2273K,\ T_c = 293K} = 1 - \frac{293}{2273} = 87.1\% \tag{18}$$

where $W = W_T - W_C$ is the net-work of expansion, usually turbine (W_T) and compression (W_C) difference. The maximum efficiency is achieved if heat is supplied at the highest possible temperature, T_h, and released at the lowest possible temperature, T_c. However, both temperatures are limited by the fact that fuel combustion is performed using oxygen with ambient air, resulting in the maximum so-called adiabatic, stoichiometric combustion temperature, T_{ad}, which is for most fuels about 2000°C or 2273 K. A part of the heat supplied at high temperature, T_h, must be released to the surroundings at a low temperature of about $T_c = 20°\text{C} = 293$ K, which results in a Carnot efficiency of 87.1% (see Eq. 18 and Figure 12). However, the fuel heating value energy, $Q_{\text{HV}} = Q_{\text{ad\_var}}$, is not all available at the adiabatic temperature of the products but is distributed over their variable temperature range from an initial surrounding temperature before combustion, T_c, to a final adiabatic temperature, T_{ad}. For such a variable heat reservoir, a large number (infinite in the limit) of ideal Carnot engines operating at different temperatures (with $\mathrm{d}W = \mathrm{d}Q$) must be employed to achieve a reversible cycle, resulting in the variable- temperature Carnot cycle with the maximum possible combustion-products-to-work conversion efficiency:[1]

$$\eta_{\mathrm{th,\ Cvar\,max}} = \left.\frac{\ln\left(T_{\mathrm{ad}}/T_c\right)}{\left(T_{\mathrm{ad}}/T_c\right)-1}\right|_{T_{\mathrm{ad}}=2273K,\ T_c=293K} = 69.7\% \tag{19}$$

Eq. 19 is valid if the cyclic medium has constant specific heat; otherwise, integration will be required. Due to engine material property limitations and other unavoidable irreversibilities, it is impossible to reach the ideal Carnot efficiency. Different actual heat engines undergo similar but different cycles, depending on the system design. For example, internal combustion engines undergo the *Otto* cycle with gasoline fuel and the *Diesel* cycle with diesel fuel, while the steam and gas turbine power plants undergo the *Rankine* and *Brayton* cycles, respectively.[7,8] However, with improvements in material properties, effective component cooling and regeneration, and combining gas and steam turbine systems, the efficiencies above 50% are being achieved, which is a substantial improvement over the usual 30%–35% efficiency. The ideal Carnot cycle is an important reference to guide researchers and engineers to better understand limits and possibilities for new concepts and practical performance improvements of real heat engines.

The reversible Carnot cycle efficiency (Eq. 18) between the two thermal reservoirs does not depend on the cycling system, but only on the ratio of the absolute temperatures of the two reservoirs ($T = T_h$ and $T_{\mathrm{ref}} = T_{\mathrm{c}}$), known as Carnot cycle efficiency. Then, we may deduce the following ratio equality, where $Q = Q_h$ and $Q_{\mathrm{ref}} = Q_c$, i.e.:

$$\left.\frac{T}{T_{\mathrm{ref}}} = \frac{Q}{Q_{\mathrm{ref}}}\right|_{\mathrm{rev\ cycle}} \tag{20}$$

The above correlation, also known as the Carnot ratio equality, is much more important than what it appears at first. Actually, it is probably the most important equation in thermodynamics and among the most important equations in natural sciences.[10] Not only does Eq. 20 define the thermodynamic (absolute) temperature scale, which is independent of the substance of a thermometer, but it was also used by Clausius to deduce a fundamental property of matter, the entropy S, as well as the related process-equilibrium functions and correlations, and to quantify irreversibility (loss of work potential) and the far-reaching second law of thermodynamics.

The entropy balance equation for the control-volume flow process, complementing the related energy balance (Eq. 11, see Figure 9), is:[7–9]

$$\underbrace{\frac{\mathrm{d}}{\mathrm{d}t}S_{\mathrm{CV}}}_{\substack{\text{RATE OF ENTROPY}\\ \text{CHANGE IN CV}}} = \underbrace{\sum_{\mathrm{CS}}\frac{\dot{Q}_i}{T_i}}_{\substack{\text{CS TRANSFER}\\ \text{RATE WITH Q}}} + \underbrace{\sum_{\mathrm{IN}}\dot{m}_j S_j}_{\substack{\text{TRANSPORT}\\ \text{RATE WITH}\\ \text{MASS IN}}} - \underbrace{\sum_{\mathrm{OUT}}\dot{m}_k S_k}_{\substack{\text{TRANSPORT}\\ \text{RATE WITH}\\ \text{MASS OUT}}} + \underbrace{\dot{S}_{\mathrm{gen}}}_{\substack{\text{IRREVERSIBLE}\\ \text{GENERATION}\\ \text{RATE}}} \tag{21}$$

Exergy and the Second-Law Efficiency

Exergy, E_{X}, as described above, is defined as the maximum possible work (i.e., the work potential) that may be obtained from a system by bringing it to the equilibrium in a process with reference surroundings (called reference *dead state*). The maximum possible work will be obtained if we prevent energy disorganization, thus with limiting *reversible* processes where the existing non-equilibrium is conserved (but rearranged and could be reversed) within interacting systems. The specific exergy per unit of system mass, e_X, with reference to the surroundings with T_o, P_o, $V_o = 0$, and $z_o = 0$, is:[7]

$$e_x = \left(\breve{u} - \breve{u}_o\right) + P_o\left(\breve{v} - \breve{v}_o\right) - T_o\left(s - s_o\right) + \frac{1}{2}V^2 + gz \tag{22}$$

The specific exergy of a flowing fluid, used for control- volume analysis, in addition to Eq. 22 includes the so- called flow exergy, $(P - P_o)$, $\overset{\frown}{v}$ resulting in:

$$e_{xf} = \left(\breve{h} - \breve{h}_o\right) - T_o\left(s - s_o\right) + \frac{1}{2}V^2 + gz \tag{23}$$

where system state is defined by its specific properties (internal energy $\breve{u}$, enthalpy $\breve{h}$, specific volume $\breve{v}$, entropy s, velocity V, and elevation z, while 0th subscripted properties refer to the corresponding system properties in equilibrium with the surroundings (at T_o and P_o).

The exergy balance equation for the control-volume flow process, complementing the related energy and entropy balance (Eqs. 11 and 21, see Figure 9), is:[7,–9]

$$\frac{d}{dt}E_x = \sum\left(1 - \frac{T_o}{T_i}\right)\dot{Q}_{\mathrm{IN},\,i} - \left(\dot{W}_{\mathrm{out}} - P_o\frac{d}{dt}\dot{V}_{\mathrm{CV}}\right) + \sum_{\mathrm{in}}\dot{m}e_{xf} - \sum_{\mathrm{out}}\dot{m}e_{\mathrm{xf}} - \dot{E}_{\mathrm{X,\,LOSS}} \tag{24}$$

The destroyed or lost exergy ($E_{\mathrm{X,\,LOSS}}$), or irreversibility (I), is a measure of energy degradation causing entropy generation (S_{gen}), or vice versa (see Eq. 17), and represents the lost work potential ($W_{\mathrm{pot,\,LOSS}}$) with regard to reference or "dead" state (surroundings at T_o absolute temperature), as expressed by the following correlation:

$$I = W_{\mathrm{pot,\,LOSS}} = E_{\mathrm{X,\,LOSS}} = T_o S_{\mathrm{gen}} \geq 0 \tag{25}$$

The second-law efficiency is defined by comparing the real irreversible processes or cycles with the corresponding ideal reversible processes or cycles:

$$\eta_{\mathrm{II}} = \frac{E_{\mathrm{X,\,out}}}{E_{\mathrm{X,\,in}}} = 1 - \frac{E_{\mathrm{X,\,LOSS}}}{E_{\mathrm{X,\,in}}}, \text{ where } E_{\mathrm{X,\,LOSS}} = E_{\mathrm{X,\,in}} - E_{\mathrm{X,\,out}} \tag{26}$$

The second-law efficiency is true efficiency related to the "goodness" or quality of energy processes. For example, the ideal Carnot cycle has the second-law efficiency 100% as the best possible heat-engine conversion process, and the efficiency (quality) of the rejected heat to the surrounding is 0%; i.e., the rejected heat in an ideal cycle has no value at all; therefore, it is not a "waste energy" as usually presented but is a useful and necessary one for accomplishing the heat engine cycle. Many other misconceptions based on the first-law analysis are clarified using the second-law or exergy analysis.

Concluding Remarks: Energy Provides Existence and Is Cause for Change

Energy is the fundamental property of a physical system and refers to its potential to maintain a system identity or structure and to influence changes (via forced-displacement interactions) with other systems by imparting work (forced directional displacement) or heat (forced chaotic displacement/motion of a system's molecular or related structures). Energy exists in many forms: electromagnetic (including light), electrical, magnetic, nuclear, chemical, thermal, and mechanical (including kinetic, elastic, gravitational, and sound). Note that if all energy is literally expelled from a confined space, then nothing will be left (empty space). As long as any matter is left, it will contain energy—even at zero absolute temperature, the electrons will be orbiting around very energetic nucleus. Matter is and must be energetic, $E = mc^2$; thus, literally, "energy is everything"—no energy, nothing in the space. Mass and energy are a manifestation of each other and are equivalent; they have a holistic meaning of *mass energy*.

The philosophical and practical aspects of energy and entropy, including reversibility and irreversibility, could be summarized as follows:

- Energy is a fundamental concept indivisible from matter and space, and energy exchanges or transfers are associated with all processes (or changes), thus indivisible from time.
- Energy is "the building block" and fundamental property of matter and space, thus a fundamental property of existence. Any and every material system in nature possesses energy.
- For a given system state (structure and phase), addition of energy will spontaneously tend to randomly redistribute (disorganize, degrade, dissipate) over the system's finer microstructure and the space it occupies, called thermal energy, equalizing the thermal energy potential (temperature) and increasing the energy randomization per absolute temperature (entropy), and vice versa.
- Entropy is an integral measure of (random) thermal energy redistribution (due to heat transfer and/or irreversible heat generation due to energy degradation) within a system mass and/or space (during system expansion), per absolute temperature level. Entropy is increasing from a perfectly ordered (singular and unique) crystalline structure at zero absolute temperature (zero reference) during reversible heating (entropy transfer) and entropy generation during irreversible energy conversion (loss of work-potential to degraded thermal energy), i.e., energy degradation or random equipartition within system material structure and space per absolute temperature level. Energy and mass are conserved within interacting systems (all of which may be considered as a combined isolated system not interacting with its surrounding systems), and energy transfer (in time) is irreversible (in one direction) from a higher to a lower energy potential only, which then results in continuous generation (increase) of energy displacement, called entropy generation, which is a fundamental measure of irreversibility or permanent changes.
- Exergy refers to the maximum system work potential if it is reversibly brought to the equilibrium with reference surroundings; i.e., exergy is a measure of a system's non-equilibrium with regard to the reference system. Exergy is not conserved but is destroyed or lost during real irreversible processes commensurate to the related entropy generation.
- Reversible energy transfer is only possible as a limiting case of irreversible energy transfer at infinitesimally small energy potential differences, thus in quasiequilibrium processes, with conservation of entropy (no entropy generation). Since such changes are reversible, they are not permanent and, along with time, irrelevant.
- Non-equilibrium, i.e., non-uniform distribution of mass energy in space, tends in time to spontaneously and irreversibly redistribute over space towards a common equilibrium; thus, non-equilibrium cannot be spontaneously created. Therefore, all natural spontaneous or autonomous processes (proceeding by itself and without interaction with the rest of the surroundings) between systems in non-equilibrium have irreversible tendency towards a common equilibrium—and thus irreversible loss of the original work potential (measure of nonequilibrium), by converting other energy forms into the thermal energy accompanied with increase of entropy (randomized equipartition of energy per absolute temperature level). The spontaneous forced tendency of mass-energy transfer is due to a difference or nonequilibrium in space of the mass-energy space density or mass-energy potential. As mass energy is transferred from a higher to a lower potential, and thus conserved, the lower mass-energy potential is increased on the expense of the higher potential until the two equalize, i.e., until a lasting equilibrium is established. This explains a process tendency towards the common equilibrium and impossibility of otherwise (impossibility of spontaneous creation of non-equilibrium).[10]

In summary, energy provides existence and is cause for change. Energy is possessed (thus equilibrium property) by material systems and redistributed (transferred) between and within systems, due to systems' non-equilibrium, via forced- displacement interactions (process) towards the equilibrium (equipartition of energy over mass and space); thus, energy is conserved (the first law) but degraded (the second law). Finally, we may unify the two laws of thermodynamics as one *grand law of nature:* The universe consists of local material (mass energy) structures in forced equilibrium and their interactions via force fields. The forces are balanced at any time (including inertial—process rate forces), thus conserving momentum, while charges/mass and energy are transferred and conserved during forced displacement in space all the time, but energy is degraded (dissipated) as it is redistributed (transferred) from a higher to a lower non-equilibrium potential towards equilibrium (equipartition of energy).

Glossary

Energy: It is a fundamental property of a physical system and refers to its potential to maintain a material system identity or structure (force field in space) and to influence changes (via forced-displacement interactions, i.e., systems' restructuring) with other systems by imparting work (forced directional displacement) or heat (forced chaotic displacement/motion of a system's molecular or related structures). Energy exists in many forms: electromagnetic (including light), electrical, magnetic, nuclear, chemical, thermal, and mechanical (including kinetic, elastic, gravitational, and sound).

Energy conservation: It may refer to the fundamental law of nature that energy and mass are conserved, i.e., cannot be created or destroyed but only transferred from one form or one system to another. Another meaning of energy conservation is improvement of efficiency of energy processes so that they could be accomplished with minimal use of energy sources and minimal impact on the environment.

Energy conversion: A process of transformation of one form of energy to another, like conversion of chemical to thermal energy during combustion of fuels, or thermal to mechanical energy in heat engines, etc.

Energy efficiency: Ratio between useful (or minimally necessary) energy to complete a process and actual energy used to accomplish that process. Efficiency may also be defined as the ratio between energy used in an ideal energy-consuming process versus energy used in the corresponding real process, or vice versa, for an energy- producing process. Energy, as per the conservation law, cannot be lost (destroyed), but part of energy input that is not converted into "useful energy" is customarily referred to as "energy loss."

Entropy: It is an integral measure of (random) thermal energy redistribution (due to heat transfer or irreversible heat generation) within a system mass and/or space (during system expansion), per absolute temperature level. Entropy is increasing from a perfectly ordered (singular and unique) crystalline structure at zero absolute temperature (zero reference) during reversible heating (entropy transfer) and entropy generation during irreversible energy conversion (loss of work potential to thermal energy), i.e., energy degradation or random equipartition within system material structure and space per absolute temperature level.

Exergy: It is the maximum system work potential if it is reversibly brought to the equilibrium with reference surroundings; i.e., exergy is a measure of a system's nonequilibrium with regard to the reference system.

Heat: It is inevitable (spontaneous) energy transfer due to temperature differences (from a higher to a lower level), to a larger or smaller degree without control (dissipative) via chaotic (in all directions, non-purposeful) displacement/motion of system molecules and related microstructure, as opposed to controlled (purposeful and directional) energy transfer referred to as *work*.

Heat engine: It is a device undergoing thermomechanical cycle that partially converts thermal energy into mechanical work and is limited by the ideal *Carnot cycle* efficiency. The cycle mechanical expansion and compression net-work is obtained due to difference between heat transferred to the engine from a high-temperature heat reservoir and rejected to a low temperature reservoir, thus converting part of thermal energy into mechanical work.

Mechanical energy: It is defined as the energy associated with ordered motion of moving matter at a large scale (kinetic) and ordered elastic potential energy within the material structure (potential elastic), as well as potential energy in gravitational field (potential gravitational).

Power: It is the energy rate per unit of time and is related to work or heat transfer processes (different work power or heating power).

System (also *Particle* or *Body* or *Object*) refers to any arbitrarily chosen but fixed physical or material system in space (from a single particle to a system of particles), often called "closed system," which is subject to observation and analysis. System occupies a so-called system volume within its own enclosure interface or system boundary and thus separates itself from its surroundings, i.e., other surrounding systems.

Temperature: It refers to the average kinetic energy during thermal interaction of disordered microscopic motion of molecules and atoms. The concept of temperature is complicated by the particle internal degrees of freedom like molecular rotation and vibration and by the existence of internal interactions in solid materials, which can include the so-called collective molecular or atomic behavior. All of these phenomena could contribute to the kinetic energy during particle (thermal) interactions. When two objects are in thermal contact (i.e., interaction of random motion of their particles), the one that tends to spontaneously give away (lose) energy is at the higher temperature. In general, temperature is a measure of the tendency of an object to spontaneously exchange thermal energy with another object until their temperatures equalize, that is, until they achieve interacting particle kinetic energy equipartition (statistically equalize).

Thermal energy: It refers to the energy associated with the random, disordered motion of molecules and potential energy of intermolecular forces, as opposed to the macroscopic ordered energy associated with ordered "bulk" motion of system structure at a large scale, and excluding internal binding energy within atoms (nuclear) and within molecules (chemical).

Total internal energy: It refers to the energy associated with the random, disordered motion of molecules and intermolecular potential energy (thermal), potential energy associated with chemical molecular structure (chemical), and atomic nuclear structure (nuclear), as well as with other structural potentials in force fields (electrical, magnetic, elastic, etc.). It refers to the "invisible" microscopic energy on the subatomic, atomic, and molecular scale as opposed to "visible" mechanical, bulk energy.

Work: It is a type of controlled energy transfer when one system is exerting force in a specific direction and thus making a purposeful change (forced displacement) in the other systems. It is inevitably (spontaneously) accompanied, to a larger or smaller degree (negligible in ideal processes), by dissipative (without control) energy transfer referred to as *heat* (see above).

References

1. Kostic, M. Work, power, and energy. In *Encyclopedia of Energy*; Cleveland, C.J., Ed.; Elsevier Inc.: Oxford, U.K., 2004; Vol. 6, 527–538.
2. Kostic, M. Irreversibility and reversible heat transfer: The quest and nature of energy and entropy, IMECE2004. In ASME Proceedings, ASME: New York, 2004; 6 pp.
3. Kostic, M. Treatise with Reasoning Proof of the First Law of Energy Conservation, Copyrighted Manuscript, Northern Illinois University: DeKalb, IL, 2006.
4. Kostic, M. Treatise with Reasoning Proof of the Second Law of Energy Degradation, Copyrighted Manuscript, Northern Illinois University: DeKalb, IL, 2006.

5. Clausius, R. Über die bewegende Kraft der Warme. Annalen der Physik **1850**, *79*, 368–397, 500–524. English Translation: "On the moving force of heat, and the laws regarding the nature of heat itself which are deducible therefrom." Phil. Mag. **1957**, *2* (1–21), 102–119.
6. Cengel, Y.A.; Ghajar, A.J. *Heat and Mass Transfer, Fundamentals and Applications,* 4th Ed.; McGraw-Hill: New York, 2011.
7. Cengel, Y.A.; Boles, M.A. *Thermodynamics, An Engineering Approach*, 5th Ed.; WCB McGraw-Hill: Boston, 2006.
8. Bejan, A. *Advanced Engineering Thermodynamics*; Wiley: New York, 1988.
9. Zemansky, M.W.; Dittman, R.H. *Heat and Thermodynamics*; McGraw-Hill: New York, 1982.
10. Kostic, M.; Revisiting the 2nd Law of Energy Degradation and Entropy Generation: From Sadi Carnot's Ingenious Reasoning to Holistic Generalization. AIP Conf. Proc. 2011, 1411, 327, doi: 10.1063/1.3665247.

10

Energy: Renewable

John O. Blackburn

Introduction

Renewable energy is a category of energy supply that is receiving increasing attention in the United States. The term refers, of course, to energy sources that can be fully utilized without diminishing future supplies. The obvious contrast is with fossil fuels which, once used, are no longer available.

Renewable energy is certainly nothing new. Mankind lived on it almost exclusively until the industrial revolution in the 1700s. Falling water and wind provided mechanical energy for tasks like pumping water or grinding grain, while ships were propelled by wind. Wood fires provided heat for dwellings or for tasks like smelting metals.

This entry examines the various renewable energy resources in the United States and the status of conversion technologies. Because the largest two sources are also intermittent, integration into reliable supply systems is also examined. This turns out to be a less formidable task than it first appears to be.

Present energy end uses are examined in order to see how renewable sources might be deployed in a highly renewable energy future.

Renewable Energy Forms

Renewable energy comes in many forms, though most sources come directly or indirectly from solar radiation. Solar energy may be used directly as a source of heat as, for example, in heating water. It may also produce electricity, either directly in photovoltaic cells or by using concentrating collectors to produce steam for turbines. Wind energy, which is derived from solar energy, is now used almost exclusively to produce electricity, as is falling water. Winds over ocean waters produce waves, which are beginning to be harnessed as a source of electricity. Ocean currents or temperature differences between surface and deep waters also offer opportunities to extract useful energy.

Trees, crops, and other plant matter (referred to as "biomass" in the renewable energy literature) likewise represent stored solar energy. Half the world still uses wood as a fuel, while forest product industries in the industrial world use wood waste to produce process heat, steam, and electricity. Energy crops, trees, or even grasses could be grown. Fuel alcohols are now produced mostly from sugarcane and corn, and they are beginning to be produced from cheaper, more abundant cellulosic biomass materials, such as wheat straw or other waste products. Other biomass sources include animal manures and sewage, which can produce methane through anaerobic digestion. There are a few renewable sources of energy that do not derive from solar sources.

The moon's gravitational pull on the earth creates tidal movements of ocean waters. In some locations, the rise and fall of water levels is sufficiently great to generate electricity.

Heat generated from nuclear processes deep in the earth sometimes comes close enough to the surface to be captured as useful energy. When underground water is present as well, steam or hot water may be available for electricity generation.

The principal difference between age-old uses of renewable energy and the present possibilities is the availability of modern technologies with conversion efficiencies much higher than those previously available. Most importantly, all forms of renewable energy can be converted into electricity—the most versatile form of energy for human purposes.

Potential Renewable Resources in the United States

Solar radiation falling on the lower forty-eight states amounts to an estimated 44,000 quads (quadrillion Btu) annually—440 times the annual energy use for the nation.[1] Few other resources are found in such abundance. Windpower, the next most abundant renewable energy source, has an estimated electricity-generating potential of 10,000 billion kilowatt-hours in the United States, almost three times present annual electricity consumption.[2] This estimate is for on-shore wind turbines only; the additional offshore potential is large, but has not been as carefully estimated. Moreover, recent studies find the U.S. potential to be underestimated in quantity and in the number of areas with economic wind-generating potential.[3]

Hydroelectric facilities already generate 7%–8% of the nation's electricity, with another 3%–4% that might still be developed.[4] Geothermal resources that are suitable for electricity generation are found only in the western third of the United States, but they amount to 25,000 megawatts— enough to provide 25%–30% of that area's present electricity use. Geothermal heat is available to more of the nation, with a much larger potential.[5] The wave-generating potential has not yet been fully assessed in the United States. It appears to be substantial along the Pacific coast and the North Atlantic coast. Tidal power possibilities in the United States are limited to Alaska, Maine, and Washington.

As for biomass sources, U.S. forests and farms now produce materials with an energy content of some 14 quads. With energy crops, this could increase fourfold, according to estimates of the National Renewable Energy Laboratory.[6] Other recent estimates give a near-term potential of 19–20 quads over and above materials taken for food, feed, and forest products, given relatively minor modifications in farm and forestry practices.[7] The author gives a figure of 1 cent per kwh in Japan, but this calculation does not include amortization of the equipment or annual maintenance costs.

There is clearly no shortage of renewable energy. It comes in many forms and, with respect to solar radiation and windpower, in great abundance.

In order to be useful in human activities, renewable energy must be captured and converted into forms which are consistent with energy demands. Time gaps between availability and need must also be bridged. For these tasks, equipment that is reliable, durable, and affordable must be available.

Conversion Efficiencies

Conversion efficiencies for various forms of renewable energy vary considerably. The growth of trees and other plant life generally convert only 0.5%–1% of solar radiation into stored biomass energy, which can be used commercially. The fastest growing species in areas with long growing seasons (sugarcane, for example) convert 2%–3% of solar energy into recoverable energy.[8]

Other renewable technologies, fortunately, have much higher conversion ratios. Solar panels for heating water convert 40%–55% of the solar energy that falls on their surfaces to useful energy.[9] Crystalline silicon photovoltaic cells now on the market convert 11%–14% of solar radiation into electricity, lab models reach 33%, and nanotechnologies hold some promise of yet-higher figures.[10] It is these conversion efficiencies that facilitate the use of large quantities of renewable energy collected in a relatively small portion of a nation's total area.

State of Technologies: Costs

Technologies Now Fully Developed and Cost Competitive

Hydroelectricity has been competitive since the earliest days of grid electricity. Expansion in the United States is constrained by environmental considerations, though output can be expanded by repowering existing facilities or installing generators at small existing dams. Hydro capacity becomes extremely valuable in grids with high components of intermittent power from wind and solar sources because it enables them to operate over a considerable range without further backup.

Wind energy is competitive with conventional fuels in areas with good wind resources and its costs continue to fall.

Geothermal Electricity

This technology is now well established in suitable sites around the world. Initial capital costs are high, both for drilling into steam or hot water sources and for generating equipment. California has 2200 megawatts in operation, and new facilities are planned or under construction in several Western states.[11]

Biomass

About three quads are already used annually for energy in the United States. The forest products industries use 1.8 quads of wood and wood wastes for process heat and cogenerated electricity. Firewood heats three million homes and supplies some occasional heat to another 20 million. These uses are fully competitive with fossil fuels, as is methane collected from landfills.[12] Other biomass possibilities are listed in the following section.

Ocean Sources

A small number of tidal plants have been installed around the world, but none have been installed as of yet in the United States. Other technologies for using ocean energy, except for waves, are not under active development at this time.[13]

Technologies That Are Well Developed but Not Yet Fully Cost Competitive

Solar Water Heating

This is a relatively simple technology consisting basically of water pipes in a box covered with glass and painted black. Today's versions are, to be sure, much more sophisticated, with freeze-protection features to make the devices useable nationwide. Making costs competitive with present energy sources will be possible upon mass production and mass installation of the equipment, a feat long since accomplished in Israel.

Photovoltaic Electricity

Fifty years of research and development efforts have brought higher efficiencies and greatly reduced costs. Costs are still too high, by a factor of two or three, for photovoltaic electricity to be fully competitive on customer facilities in the United States. They are already competitive for users remote from a power grid or for utilities at peak summer hours.

Cost reductions in photovoltaic systems manufacturing are proceeding more slowly than anticipated in the 1980s and 1990s, while conventional generation, with which these systems compete, has shown declining inflation- adjusted cost, as well. Authoritative analyses in the 1980s projected cost reductions of photovoltaic modules and other system costs to about one dollar per peak watt (two dollars in 2005 prices), which implied Photovoltaic (PV) electricity costs of six to seven cents per kilowatt-hour.[14] Module prices have indeed declined to about $3.50 for volume purchases in 2005 (in current dollars).[15] Costs are not likely to fall significantly until some alternatives to the (now) standard silicon wafer are developed for volume production. Nonetheless, manufacturing costs have fallen below three dollars for modules.[16] With rapid market and manufacturing development, Japan can produce photovoltaic electricity at about 15 cents per kwh, competitive with (expensive) Japanese-delivered grid electricity.[17]

Photovoltaic equipment benefits from its potential location at the point of consumption, thus avoiding most of the costs of transmission and distribution. PV electricity, therefore, can be priced higher than electricity from central generating plants. Because there is already enough rooftop space to accommodate nearly any conceivable volume of photovoltaic power, site costs are avoided as well.

Other Biomass Sources

Subsidized fuel ethanol is made mostly from corn, but more abundant cellulose sources are beginning to be utilized. Scale economies and cost decreases with production experience should make this technology competitive in the next few years. Iogen, the Canadian producer of ethanol from wheat straw, estimates a price of $1.30 per gallon from a proposed plant in Idaho.[18] Economically feasible processes and equipment to produce methane from animal wastes are already available, but the technology is still not widely used.

Technologies for Which Further Development Is Required

Solar Industrial Process Heat

Development of equipment for these tasks received some attention in the late 1970s, but was abandoned after that as fuel prices declined. Industrial process heat requirements and insolation overlap considerably in the United States, so their potential remains to be exploited.[19] Further development awaits higher fossil fuel costs and some renewed research attention.

Wave Energy

Several small demonstration facilities now operate. The first full-scale commercial generation facilities are about to be installed in Portugal by British and Portuguese firms using British designs.[20]

Ocean Currents and Thermal Differences between Surface and Deep Ocean Waters

There is very little activity at present.

In summary, hydroelectric, wind, and geothermal electricity are fully cost-competitive with fossil and nuclear-generated electricity in areas with good resources. These areas are, respectively, three-quarters of the states for wind, all but three or four states for hydroelectricity, and the western third of the nation for geothermal. Biomass-based heat and electricity are likewise well-established sources and they are available in every state. Widespread market penetration of solar hot water systems awaits economies of mass production and mass installation. Biomass-based liquid fuel production is set to expand rapidly. Sales of photovoltaic equipment are rising rapidly, but still with subsidies and incentives. Cost reduction proceeds apace at about 4%–5% per year, while several promising new and potentially much cheaper versions are moving into production.

Solar industrial process heat is receiving little attention, though it probably will as oil and gas prices trend upward in coming decades.

Dealing with Intermittent Sources

Solar radiation and wind are the most abundant renewable energy sources, but they are also intermittent in nature. This feature raises issues of complementarity and energy storage. Utilities think of generating capacity in terms of baseload, intermediate, and peaking capacities. Because intermittent sources do not fit into any of these categories, they are likely to be relegated to marginal roles.

Inquiries that pose the question of usefulness in another way have produced much more favorable conclusions. Utilities are already accustomed to demand patterns that are subject to seasonal, daily, weather-related, and random fluctuations. Their supply sources, in contrast, can be turned on or off as they wish. The more fruitful approach in considering intermittent sources is to regard them as "negative demand"—the side which is already subject to variations. Models that start with actual hour-by-hour demand loads through the year and then subtract out wind or solar contributions, hour-by-hour, consistently show sizeable potential contributions from intermittent renewables without further backup. Needless to say, the presence of hydroelectric facilities or pumped storage units supports the ability to accommodate intermittent sources.

A simulation of the British electricity system showed that windpower, though not steadily available and not correlated with demand patterns, could still meet 25%–45% of system demand without additional backup, provided that the conventional components could be reconfigured to accommodate wind patterns.[21] A U.S. simulation found that in a system with demand patterns similar to those of most utilities, a grid with a substantial hydroelectric component could accommodate 50% of its outputs from intermittent renewable sources (i.e., wind and solar) and at costs no higher than those for conventional generation.[22] These are certainly counter-intuitive results, and they would need to be supplemented by modeling studies of other utilities. The findings, though, are consistent with experiences in Denmark and North Germany. Denmark now obtains 21% of its electricity from wind turbines, without evident problems of integrating this output into the national grid. German reports of strains on the grid seem to have more to do with transmission bottlenecks than generation sources.

Solar electricity has the advantage, for most utilities, of coinciding in availability with summer peaks in air-conditioning use. A recent New Jersey study found that photovoltaic generation would not only supply peak power (and thereby displace fuel generation) but 40%–70% of PV capacity would displace the capital costs of conventional generation, fractions which rose quickly in the presence of grid storage capacity.[23]

These results are so much at variance with traditional thinking in utility managements that time, experience, and external pressures will be required to bring solar and wind technologies into widespread use, even when they are cost-competitive. Therein lies the case for incentives and regulatory requirements, such as the renewable portfolio standards now found in 18 states. Expanding markets enable the solar and wind equipment manufacturers to scale up and further reduce costs, while the utilities gain actual operating experience with intermittent sources.

Moving toward an All-Renewable Energy Future

A consideration of energy futures in which renewable sources contribute most or all energy needs begins with an assessment of the quantity of energy that is likely to be needed, and its appropriate forms. As a starting point, one examines present energy sources and end uses in the economy, and then allows for increasing efficiencies. One must show that renewable resources are sufficient in quantity, quality and cost to provide for each of the end uses.

A word of caution is in order. It is much too easy to focus on energy supply, while neglecting the rich possibilities for improving energy efficiency in all end uses. One thus can miss the most economical opportunities for meeting energy demands because efficiency measures usually are less expensive than any new supplies, renewable or conventional. The reader is referred to the many entries in this encyclopedia that deal with the significant potential of increasing the efficiency with which energy is used.

TABLE 1 Fuel and Energy Use by Sector, United States, 2003 (Quadrillion Btus)

Sector	Fuel	Electricity	Total End Energy Use	Electricity Losses	Total Primary Energy	Fuel End Uses in Sector
Residential	7.2	4.4	11.5	8.7	21.3	Space heat
Commercial	4.3	4.1	8.4	9.2	17.6	Water heat Space heat
Transportation Industry Manufacturing	26.8	—a	26.8	0.1	26.9	Water heat Liquid fuel
Agriculture, Mining	15.8	3.4	19.2	7.7	26.9	Process heat
Construction	3.0	—a	3.0		3.0	
Feed stocks	3.0	—	3.0		3.0	
Total industry	21.8	3.4	25.2	7.7	32.9	
Total all sectors	60.1	11.9	72.0	26.7	98.7	

[a] Less than 0.1 quad.

Source: Adapted from Energy Information Administration (additional detail for industry from International Energy Agency, IEA Energy Statistics, 2003, Energy Balance Table for the United States).[25]

Present U.S. Energy Use

U.S. primary energy use is now running at about 100 quads per year and growing at about the same rate as population, unlike the growth pattern prior to 1973. Data for 2003 are shown in Table 1.

As the table indicates, primary energy use in the United States in 2003 was 98.7 quads. About 40% of this primary energy was used to generate electricity, and two-thirds (26.7 quads) of that was lost as waste heat in generation and transmission. End-use energy was thus 72 quads. Transportation and industry each take about one third of end-use energy, with the rest going to the residential and commercial sectors. In U.S. energy statistics, electricity losses are allocated to each of the sectors in order to see the primary energy demand occasioned by the activities of each sector.

Table 1 also lists the kinds of energy end-uses provided by fuels because these would have to be supplied from renewable sources in the absence of fuels. (Sector end- uses for electricity are not shown, since electricity is readily produced from renewable sources.) In the residential sector, for example, fuels are used mostly to heat space and water or, put differently, to provide low temperature (under 100 degrees centigrade) heat. A much smaller use of fuel is for cooking, which utilizes temperatures in the 100–250 degree centigrade range. Fuel use in the commercial sector is applied for similar purposes. Eighty percent of industrial energy use is in manufacturing, while 20% goes to mining, agriculture, and construction or is used for fuel-based feedstocks. Manufacturing end-use energy is 19 quads of which three to four quads is for refining petroleum.

Energy Efficiency Gains

All of these uses can be reduced with gains in energy efficiency, and some of them can be reduced significantly. Indeed, if the best practices already existing in building design and construction, lights, appliances, heating and cooling systems, vehicle mileage, and industrial processes were uniformly applied in the whole economy, end-use energy demands use would be 50%–70% lower, or in the 22–36 quad range—not the 72 quads shown in Table 1.[24] These potential reductions in energy use make renewable energy futures much more readily conceivable for high-income, high-output nations. Any energy supply system in a world of diminishing oil and gas output would be hard pressed to meet the energy demands of our present inefficient economy.

Adequacy of Renewable Resources

The questions for renewable energies then become:

1. In the United States, can renewable energy sources provide 22–36 quads, of which 6 to 9 or so would meet efficiency-reduced electricity demands, 3 to 5 quads would be used for heat in buildings, 7 to 10 quads or so would power the transportation sector, and 6 to 10 quads would be composed of fuels for industry?
2. If that can be shown, then can these sources provide for any continuing growth in the economy?
3. What technological developments are needed for this potential to be realized?

The adequacy of renewable energy sources, considered in gross, is easy to establish. Wind and solar resources alone provide these amounts of energy many times over. In addition, the "nonintermittent" renewable sources (biomass, hydroelectric, and geothermal resources) sum to another 17 quads at a minimum.

As to end-uses, 6 to 9 quads of electricity fall well under available sources, with storage capability being required only in areas lacking hydroelectric resources.

Space and water heat can be provided in part from direct solar sources, with the rest coming from biomass or perhaps electricity.

For those who envision a hydrogen-fueled transportation system, intermittent sources of renewable electricity are tailor-made. Hydrogen can be produced by electrolysis whenever solar or wind sources exceed grid demands. Liquid fuels from biomass are likely to be available, as well, especially with emerging technologies to produce liquid fuels from cellulosic materials.

Industrial process heat would come primarily from biomass sources, with some assistance from solar equipment or perhaps hydrogen produced from renewable electricity.

As to growth over time in energy demands over time, one can observe that economic growth in the United States, at least in per capita terms, is directed toward sectors that generally do not have high energy contents (services vs. basic commodities). If the United States is ever to reach a measure of sustainability with respect to its physical resource demands, its population must level off, as must that of any other nation seeking sustainability.

There are two major technological developments that remain to be achieved in order to make an all-renewable future attainable. The first is cost reductions in the manufacture and installation of photovoltaic systems and the second is the rapid commercialization and cost reduction in processes which make fuel ethanol from cellulosic materials, rather than foods. These developments are already well underway.

Space does not permit an extended analysis of these matters; this brief discussion is intended to establish, in general terms, that a largely or wholly renewable energy future is possible, given several decades to adjust. This view is not now widely accepted, yet it is certainly the most hopeful one for mankind in an age of diminishing fossil fuel resources.

References

1. Kendall, H.W.; Nadis, S.J. Eds.; Energy strategies: toward a solar future. In *A Report of the Union of Concerned Scientists.* Ballenger: Cambridge, MA, 1979.
2. Grubb, M.J.; Meyer, N.I. Wind energy: resources, systems, and regional strategies. In *Renewable Energy: Sources for Fuel and Electricity*; Johansson, T.B., Kelly, H., Reddy, A.K.N. et al., Eds.; Island Press: Washington, DC, 1992.
3. Archer, C.L.; Jacobson, M.Z. Spatial and temporal distributions of U.S. winds and windpower at 80 m derived from measurements. J. Geophys. Res. Atmos. **2003**, *108* (D9), 4289.
4. National Hydropower Association, Hydrofacts. http://www.hydro.org/hydrofacts/forecast/asp (accessed May 24, 2005).
5. Wright, P.M. Status and potential of the U. S. geothermal industry. AAPG Bull. **1993**, *77* (8).

6. National Renewable Energy Laboratory, Biomass Energy, Potential, http://www.nrel.gov/documents/biomass_energy.html (accessed May 21, 2005).
7. Perlack, R.D.; Wright, L.L.; Turhollow, A.; Graham, R.L. *Biomass as Feedstock for a Bioenergy and Bioproducts Industry: The Technical Feasibility of a Billion-Ton Annual Supply;* Oak Ridge National Laboratory: Oak Ridge, TN, April, 2005.
8. Hall, D.O.; Rosillo-Calle, F.; Williams, R.H.; Woods, J. Biomass for energy: supply prospects. In *Renewable Energy: Sources for Fuels and Electricity*; Johansson, T.B., Kelly, H., Reddy, A.K.N., Williams, R.H., Eds.; Island Press: Washington, DC, 1992.
9. Florida Solar Energy Center. http://www.fsec.ucf.edu/solar/testcert/collectr/tpndhw.htm (accessed May 23, 2005).
10. http://www.oja-services.nl/iea-pvps/pv/materials.htm (accessed May 24, 2005).
11. Renewable Energy Policy Project http://www.repp.org./geothermal_brief_power_technologyand-generation.html (accessed May 24, 2005).
12. http://www.nrel.gov/documents/biomass_energy.html (viewed May 21, 2005).
13. http://www.poemsinc.org/FAQtidal.html (accessed May 25, 2005).
14. Maycock, P.D. *Photovoltaics: Sunlight to Electricity in One Step*; Brick House Publishing Co.: Andover, MA, 1981.
15. http://www.solarbuzz.com Price Survey, May 2005. (accessed May 20, 2005).
16. http://www.nrel.gov/docs/fy04osti/32578.pdf (viewed May 21, 2005).
17. Sawin, J. L. *Mainstreaming renewable energy in the 21st century.* Worldwatch Paper 169, Worldwatch Institute: Washington, DC, 2004.
18. Governors seek wider ethanol use. Wall Street Journal, April 12, 2005.
19. *A New Prosperity; Building a Sustainable Energy Future. The SERI Solar/Conservation Study.* Brick House Publishing: Andover, MA, 1981.
20. http://www.renewableenergyaccess.com (accessed May 20, 2005).
21. Grubb, M.J.; Meyer, N.I. Wind energy: resources, systems, and regional strategies. In *Renewable Energy: Sources for Fuels and Electricity*; Johanssen, T.B., Kelly, H., Reddy, A.K.N., Williams, R.H. Eds.; Island Press: Washington, DC, 1992.
22. Kelly, H.; Weinberg, C. Utility strategies for using renewables. In *Renewable Energy: Sources for Fuels and Electricity;* Johanssen, T.B., Kelly, H., Reddy, A.K.N., Williams, R.H. Eds.; Island Press: Washington, DC, 1992.
23. Perez, R. Determination of Photovoltaic Effective Capacity for New Jersey. http://cleanpower.com/research/capacityvaluation/ELCC_New_Jersey.pdf (accessed May 24, 2005).
24. Blackburn, J.O. Possible efficiency gains identified for the U.S. as of the mid-1980's are examined. *The Renewable Energy Alternative*; Duke University Press: Durham, NC, 1987.
25. Energy Information Administration, Monthly Energy Review, March 2005.

11

Energy: Storage

Rudolf Marloth

Introduction

Although specific types of energy storage are introduced below, note that this encyclopedia has separate and detailed entries on batteries, compressed air storage, and pumped hydroelectricity. First, some definitions are given:

- Primary storage methods are designed for a single use. A regular flashlight cell is used once and discarded (NOT into landfill).
- Secondary storage methods allow energy to be stored and recovered repeatedly. Rechargeable flashlight cells are available.
- Renewable energy is endlessly available in nature. Wind and solar power are currently in the spotlight.

The greater part of the entry concerns secondary storage. Renewables and primary storage are discussed briefly. A glossary is included at the end of the entry.

Need for Storage

Depending upon the use, energy is stored for varying periods. The energy stored in the gas tank of an automobile is used over a period of about a week. Part of that energy is stored in the car's battery, to be used two or several times in a day.

Power companies use large-scale energy storage[1] to defer investment in new generation, transmission, and distribution facilities. It is useful for integration with renewables, for voltage and frequency regulation, for demand leveling, and in lieu of spinning reserves. Long-term energy storage might depend upon availability and the cost of transporting it, such as the pile of coal at a generating plant.

Rate payers can use energy storage for load leveling and to protect themselves from the vagaries of the grid. Large quantities of electric energy generated at night when utility rates are cheap are stored in order to be released in daytime when rates are high. Smaller quantities are stored indefinitely to keep critical or sensitive equipment running during power failures or surges, sometimes only for seconds or minutes, until backup generators can come on line. A small and isolated user might want to store irregular energy supplied by wind or sun rather than build a road, pipeline, or power line for access to conventional sources.

Example 1. Consider electric rates that are \$10.00/kW-mo and \$0.10/kWh during weekdays from noon to 6:00 P.M. and \$5.00/kW-mo and \$0.05/kWh during weeknights from midnight to 6:00 A.M. and on weekends. Making reasonable assumptions about on-peak and off-peak hours, calculate a benefit of this use of energy storage.

Solution. Using an average of 22 weekdays and 8 weekend days per month gives 132 hr/mo on-peak hours and 324 hr/mo off-peak hours. Storing energy over all the offpeak hours will reduce the demand contribution to the stored energy cost. Assuming 75% round-trip efficiency apportioned evenly between storage and recovery (86.6% each way) and spreading the demand cost cover each hour of use, the blended cost of stored energy is

$$\frac{\$5.00/\text{kW}-\text{mo}/324\text{hr/mo}+\$0.05/\text{kWh}}{0.866}=\$0.076/\text{kWh} \tag{1}$$

while the blended cost of recovered energy is

$$\frac{\$10.00/\text{kW}-\text{mo}/132\text{hr/mo}+\$0.10/\text{kWh}}{0.866}=\$0.20/\text{kWh} \tag{2}$$

Valuation Criteria

Energy storage methods can be evaluated by many different criteria, with importance depending upon the application. Possible criteria are first cost in \$/kWh, lifetime in cycles or hours, maintenance cost, energy density in kWh/lb, allowable depth of discharge, footprint in kWh/ft^2 and area in ft^2, reliability, and safety.

Storing Electricity

A journal entry[2] whose title implies the possibility of storing electric energy seems to allow in its content that, technically, one cannot. The methods cited require conversion and reconversion, with storage in any of several non-electric forms. For storage in batteries, electricity is converted to chemical energy. For flywheel storage, electricity is converted to rotational kinetic energy. In pumped hydroelectric storage, electricity is converted to gravitational potential energy. Compressed air and energy stored in springs are mechanical potential energy. Superconducting magnetic energy storage (SMES) is magnetic energy. The energy stored in a capacitor is in an electric field, so at least this can be called stored electric energy.

Renewable Energy

A distinction is commonly made between stored energy and renewable (natural) energy. Some might say that energy is "stored" in the wind or ocean waves where it is available to be converted into a more useful form. Others contend that at least some processing is required for energy to qualify as stored, for instance, oil from a well must be refined, transported, and stored in a tank. For another example of primary stored energy, sulfur, charcoal, and saltpeter are combined to make gunpowder. Other examples of renewable energy are waterfalls, sunlight, trees and other biomass (saw grass, bagasse, corn), and geothermal energy. The case of water behind a dam is not so simple.

However, once some of these resources are harvested, processed, transported, and deposited at the consuming site, they can be considered as primary stored energy. Yet another category of natural energy becomes available at unpredictable times, such as that in volcanoes, hurricanes and tornadoes, and falling trees. A fascinating example of this type[3] is the energy in rocks tumbling down the San Gabriel Mountains into the backyards of suburban Los Angeles.

Hydrocarbons (coal, oil, and gas) are not considered renewable because they require eons to form, and indeed, they are being used up much faster than they form.

Primary Storage

Primary storage involves extraction or manufacture and a storage site or vessel, for example: 1) coal mined, transported, and stored at a power plant; 2) gasoline in the fuel tank of an automobile; and 3) uranium ore mined and processed into pellets of fuel. Examples of manufactured products containing primary stored energy are gunpowder, candles, lamp oil, and, indeed, non-rechargeable flashlight batteries (cells). Taking this one step further, the energy used in manufactured durable goods is stored when the goods are inventoried. This has the advantage of a very low dissipation rate but the disadvantage of frozen capital.

Secondary Methods

Storing on the Grid

References to storage on the electric grid conjure visions of electrons running about on transmission lines, looking for a place to light. In fact, the term merely means storage facilities connected to the grid, and it is in vogue because the lack of such storage is an impediment to the development of renewable power.[4] The problem is the need for a stable supply of power combined with the erratic availability of renewables. The solution is threefold: 1) an extended grid from the sources to the users; 2) increased storage facilities to receive energy when available and to supply it when needed; and 3) improved controls to gain access to stored energy and/or to limit the demand by users.

Capacitors

The energy, E, stored in a capacitor is equal to

$$E = \frac{1}{2}CV^2 \tag{3}$$

where the capacitance, C, of a plate capacitor is equal to

$$C = \frac{\varepsilon A}{d} \tag{4}$$

where ε is the permittivity of the dielectric between the plates, V is the voltage across the plates, A is the area of each plate, and d is the distance between the plates.

Evidently then, a high-energy capacitor requires a high voltage, a high permittivity, and a large area-to-separation ratio.

Conventional electrostatic capacitors have two metal plates separated by a dielectric. Electrolytic capacitors[5] are made up of two aluminum foils. The anode foil is coated with aluminum oxide, Al_2O_3, and the cathode is the other foil with a spacer saturated with electrolyte facing the oxide layer. Many types of both aqueous and non-aqueous electrolytes have been used.

Electric double-layer capacitors (EDLCs), also known hyperbolically as supercapacitors or ultracapacitors, have, between the electrodes, two layers of a sponge-like material with a separator that takes the place of plates. The material is typically activated carbon, although carbon nanotubes are envisioned for the purpose. These materials have thousands as much area per volume as conventional dielectrics and the distance between charges is submicroscopic, thus making their capacitance and energy density that much greater than that of electrolytic capacitors. Because of the small spacing, the voltage is only 2–3 V.

A major use of EDLCs is in combination with batteries. While batteries store more energy, EDLCs can charge and discharge much faster. Thus the EDLC is used in cameras to supply the flash and in electric cars to assist in hill climbing and acceleration or to absorb braking energy.

Chemical Batteries

The lead–acid battery has several disadvantages. It is heavy, toxic, maintenance-intensive, and short-lived. However, the technology is well-developed, and the materials are inexpensive, and this trumps all the disadvantages. The positive electrode (cathode) is PbO_2, lead dioxide, the negative electrode (anode) is Pb, lead, and the electrolyte is sulfuric acid, H_2SO_4. The overall reaction is

$$PbO_2 + Pb + 2H_2SO_4 \leftrightarrow 2PbSO_4 + 2H_2O \tag{5}$$

moving to the right during discharge and to the left during charging. Thus both electrodes and the electrolyte are used up during discharge, forming lead sulfate and water. The reaction can be reversed by application of a reverse current at the electrodes. A half-reaction occurs at each electrode, with electrons being produced at the anode and absorbed at the cathode.

A large lead–acid battery test facility[6] was built by Southern California Edison in Chino, California, and went into operation in 1998. Four strings of series-wired cells were housed in each of two large buildings, with a smaller building in between to house the converter and controls. Each string consisted of 1032 two-volt cells tied in series to provide a nominal voltage of 2000 Vdc. The facility was programed for an 80% depth of discharge to provide 20.8 kAh, or 40 MWh plus the energy to run auxiliaries, primarily the air compressor and air handlers. Typical discharge cycles were 5 MW for 8 hr or 10 MW for 4 hr.

Compare these figures to those for an automotive SLI (starting, lighting, ignition) battery that has six 2V cells to provide 40 Ah or 480 Wh. The SLI battery is designed for high current for a short time and for shallow discharge, whereas a traction battery, e.g., a 6, 8, or 12V golf cart battery, is designed for deep discharge and power over long periods.

Compressed Air

In the compressed air energy storage (CAES) method,[7] air is stored at high pressure in an underground cavern for use in a gas turbine. Since a conventional gas turbine consumes about two-thirds of its fuel to compress air at the time of generation, the CAES method benefits by using electricity from the grid at off-peak times for use at on-peak times in lieu of the gas turbine compressor. One improvement to the basic method is a surface pond connected to the cavern that serves to keep the pressure in the cavern constant. When air is stored, water is displaced, and when air is withdrawn, water reenters. Second, between the compressor and the cavern is a thermal storage device to store the heat of compression. When the air is withdrawn, it is reheated.

Example 2. Compute the specific work required to compress air at 1 atm and 20°C to 50 atm in an adiabatic compressor having an isentropic efficiency of 80%. Assuming the air is stored at 20°C in a cavern of 200,000 m^3, what total work is required?

Solution. The specific work of compression is

$$w = \left(\frac{1}{\eta}\right)\left(\frac{k}{k-1}\right)RT_1\left[\left(\frac{P_2}{P_1}\right)^{\frac{k-1}{k}} - 1\right] = \left(\frac{1}{0.80}\right)\left(\frac{1.4}{0.4}\right)(0.287\text{kJ/kg}-\text{K})\left[(20+273)\text{K}\right]$$

$$\left[\left(\frac{50}{1}\right)^{\frac{1.4-1}{1.4}} - 1\right] = 758\text{kJ/kg} \tag{6}$$

$$m = \frac{P_2V_2}{RT_2} = \frac{50\left(101.3\ \text{kPa}\right)\left(2\times10^5\ \text{m}^3\right)}{\left(0.287\ \text{kPa}-\text{m}^3/\text{kg}-\text{K}\right)293\ \text{K}} = 1.20\times10^7\ \text{kg} \tag{7}$$

$$W = \frac{mw}{c_1} = \frac{\left(1.20\times10^7\ \text{kg}\right)\left(758\text{kJ/kg}\right)}{3.6\times10^6\ \text{kJ/MWh}} = 2530\ \text{MWh} \tag{8}$$

Flow Batteries

These batteries are so named because the electrolyte flows between the cell and a storage tank for each electrode. In the best-known flow battery the electrolyte is a solution of vanadium and sulfuric acid and the reaction is an oxidation-reduction reaction, hence the name vanadium redox battery[8] (VRB). The cathode electrolyte carries V^{5+} and V^{4+} ions, while the anode electrolyte carries V^{2+} and V^{3+} ions. The reactions during discharge are oxidation at the cathode

$$V^{4+} + V^{5+} + e^- \tag{9}$$

and reduction at the anode

$$V^{3+} + e^- \rightarrow V^{2+} \tag{10}$$

These reactions are reversed during charging. The electrodes are separated by a permeable ionic membrane to allow hydrogen ions to pass through to complete the circuit.

The electrodes do not take part in the reactions, so they do not deteriorate as in a lead–acid battery. The electrolyte can be recycled indefinitely, eliminating the problem of disposal. The battery can be fully discharged without damage, it can be discharged at high voltage, and it could be "instantly" recharged by replacing the electrolyte.

Flywheels

Originally, flywheels were used to regulate the speed of rotating machinery. By virtue of its large rotational inertia, the flywheel tended to oppose any tendency in the machine to slow down or speed up. Now flywheels are being used as energy storage devices.

Modern energy storage flywheels are massive rotating cylinders supported on stators by magnetically levitated bearings to eliminate bearing wear. The flywheel is operated in a low vacuum environment to reduce drag. Some of the key features of flywheels are low maintenance, long life, and inert material.

The choice of solid steel vs. composite rims is based on the system cost, weight, and size. Steel is denser (heavier) but composites can withstand much higher tip speeds. Actual delivered energy depends on the speed range of the flywheel as it cannot deliver its rated power at very low speeds.

The energy stored in a flywheel can be calculated as follows: The energy of translation of a particle of mass m and velocity v is $1/2mv^2$, so the energy of rotation of a particle of mass m and rotational speed ω rotating about a center at distance r is $1/2m(r\omega)^2$. For a group of particles all at distance R and totaling mass M, the rotational energy $1/2MR^2\omega^2$, is or $1/2I\omega^2$ where $I = MR^2$ is the rotational inertia (or moment of inertia). Therefore to maximize the stored energy, a flywheel should be heavy (large M), big (large R), fast (large ω), and have the mass concentrated at the outer rim (mass at R).

Flywheels are used to provide power to uninterruptible power supplies.[9] They pair well with batteries in this application, since flywheels are tolerant to frequent cycling (which batteries are not), and batteries can provide power for longer periods. Comparative advantages of flywheels over batteries are longer life, smaller footprint, less maintenance, and less stringent housing conditions. The main disadvantage of a flywheel is the short period in which backup power can be supplied because of the continuously decreasing speed of rotation. This can be overcome by introducing special circuitry, but this adds cost and complexity.

Example 3. Consider a modest-sized flywheel of 40 kg mass and 0.3 m diameter rotating at 5000 rpm. Find the stored energy. Assume all the mass to be located at the extremity.

Solution. The rotational speed is (5000 rev/min) (2π)/(60 sec/min) = 523.6 rad/sec. The kinetic energy in the wheel is (1/2) (40 kg) $(0.3\ \text{m})^2$ $(523.6\ \text{rad/sec})^2 = 493{,}500$ (kg-m/sec^2)-m (= N-m = J = W-sec). If only 10% of this energy is usable until the flywheel slows to an unacceptable speed, 49,350 W-sec is available, or about 5 kW for 10 sec until a longer-term backup can be brought on line.

Fuel Cells

Fuel cells are not properly an energy storage device, but an energy conversion device, since they take a continuous flow of materials and combine them to produce energy and heat. However, they have a certain analogy with chemical batteries. Fuel cells convert hydrogen and oxygen into electricity and heat. A fuel cell, like a battery, is made up of an anode, a cathode, and an electrolyte. A current is produced when hydrogen ionizes at the anode with the aid of a catalyst. The hydrogen ions pass through the electrolyte to the cathode, where they combine with oxygen to produce water. The electrons released at the anode travel through an electric circuit to the cathode. The overall reaction is

$$2H_2 + O_2 \rightarrow 2H_2O \tag{11}$$

Notice that the reaction only goes to the right, i.e., it is not reversible as in an energy storage device. As in a lead-acid battery, the half-reactions produce electrons at the anode and absorb them at the cathode.

Fuel cells use a variety of electrolytes, including phosphoric acid, molten carbonate, solid oxide, and a proton exchange membrane.

Pumped Hydroelectricity

In this method, water is pumped into an elevated reservoir, with the energy recovered when the water flows back down to a lower reservoir through a turbine generator. The reservoirs may be man-made or natural bodies of water. Reversible turbine/pumps reduce the equipment required. Pumped hydro is particularly advantageous in conjunction with gas- or coal-fired plants, first by demand leveling, allowing the thermal plant to run at its most efficient load, and second by responding quickly to fluctuations in demand, where a thermal plant cannot. Pumped hydro also has been found to pair well with wind power and its variations in availability. Of course pumped hydro can be used on a regular schedule of pumping at night when rates are cheap and recovering by day when rates are high. The energy stored, E, is the potential energy of position, $E = mgh$, where m is the mass of water, g is the acceleration due to gravity, and h is the difference in height between reservoirs. Conversion efficiency is good, with losses depending on efficiency of the motor/generator and pump/turbine, pipe friction, and evaporation/seepage. Dam failures notwithstanding, this method is generally safe. The essential requirement is a site where high and low reservoirs are naturally available or can be constructed at a reasonable cost.

Springs

The force, F, required to stretch or compress a linear spring is $F = kx$, where k is the spring constant and x is the change in length. The energy, E, stored is the integral of force over the change in length, $E = 1/2kx^2$. The torque, τ, required to twist a linear torsion spring is $\tau = k\theta$, where k is the spring constant and θ is the change in angle. The energy, E, stored is the integral of torque over the change in angle, $E = 1/2k\theta^2$.

Torsion springs are typically used in instruments in spiral form. A torsion pendulum is a heavy disk that rotates at the end of a wire in torsion.

It is interesting to note that a bolt is a spring, and the clamping force (i.e., the stored energy) is the stretch in the bolt times its spring constant.

Superconducting Magnets

SMES is the storage of a current, I, in a cryogenically cooled inductor with energy E equal to

$$E = \frac{1}{2} L I^2 \tag{12}$$

where the inductance of a coil, L, is equal to

$$L = \frac{\mu N^2 A}{X} \tag{13}$$

where μ is the permeability of the core, N is the number of turns in the coil, A is the cross-sectional area of the coil, and X is the length of the coil.

Evidently then, a high-energy SMES requires high current, a highly permeable core, many turns, and a large cross-sectional area-to-length ratio.

The energy is maintained without dissipation by cryo-genically cooling the coil.[10] Helium is used as a coolant because its freezing temperature is 4.2°C above absolute zero. The salient characteristic of SMES is fast response with extremely high power. Thus a prominent use would be to stabilize short-term faults in grid power. Advantages cited for this method of energy storage are high conversion efficiency because of low resistance losses, and low maintenance because of few moving parts.

Thermal Storage

A wag has suggested that the thermos bottle is the greatest invention of the human mind because it keeps hot things hot and cold things cold. How does it know? Three examples of thermal storage are discussed here: chilled water, regenerators, and Trombe walls.

The energy consumed by a chiller can be stored in the form of chilled water or ice in an insulated tank, usually as a play on varying electric rates. Water is an excellent medium for this since it is inexpensive but also because it has a high specific heat, giving a good energy-to-volume ratio. For instance, an energy-efficient chiller will operate at 0.5 kW/T, (T = ton) so that over 6 hr, a 100T chiller will make

$$(100\text{T})(12{,}000 \text{ Btu/hr} - \text{T})(6\text{hr}) = 7.2 \text{ MMBtu} \tag{14}$$

of cooling for

$$(0.5 \text{ kW/T})(100 \text{ T})(6 \text{ hr.}) = 300 \text{ kWh of electrity}$$

This energy can be stored by chilling water from 70°F to 35°F, which will require a volume of water equal to

$$\frac{7.2 \text{ MM Btu}}{(1 \text{ Btu/1 bm} - {}^\circ\text{F})(70 - 35{}^\circ\text{F})(62.41\text{bm/ft}^3)} = 3300\text{ft}^3 \tag{15}$$

which is equivalent to a cubical tank about 15 ft on a side. The energy would typically be recovered by running the chilled water through an air handler to provide air conditioning. Good insulation, the small temperature difference between the medium and the surroundings of the tank, and the relatively short storage time would keep the storage loss low.

A regenerator is a regenerative heat exchanger, in which energy is stored in the material of the heat exchanger itself. A regenerator cycles such that half of it is heated by hot flue or exhaust gases while cool combustion air is drawn in and heated in the other half. In the second half of the cycle, cool air enters through the previously heated material of the regenerator. For example, older industrial furnaces have two chimneys, where combustion air enters in one while flue gases heat the other. Then the paths are switched so that new air is heated as it descends the first, heated chimney. A newer type of regenerator is a slowly rotating heat wheel, with half of it being heated by exiting flue gas and the other half heating the incoming air. The paths of the intake and exhaust remain constant while the heat-storing element moves.

A Trombe wall[11] (after the engineer Felix Trombe who used it extensively) is a form of passive solar energy storage. It is most commonly used in small buildings in combination with other passive energy-saving methods. The wall is a south-facing thick construction of brick or concrete with a layer of glass on the outside and an air space between. The greenhouse effect allows solar radiation to heat the wall but blocks the longer-wavelength, lower-temperature radiation from the wall. The wall stores heat during the day and releases it into the living space at night. Vents with one-way flaps pierce the wall at the top and bottom, whereby cool air from the room enters the air space at the bottom, is heated, rises to the top of the space, and reenters the room through the top vent. The flaps block reverse flow.

Conclusion

Stored energy is considered here as involving some human activity, ruling out natural energy contained in the winds or the tides or the Earth's crust. Primary storage is advantageous when the energy source is cheap or available. A disadvantage of acquiring in advance of need is that energy leaks from the medium.

Most secondary methods involve converting from and to electrical energy, so the electric grid becomes a primary factor. The practice of "buy low, sell high" leads to storing energy during off-peak hours and recovering it during on-peak hours. The availability and location of renewables causes that energy to be transported from the source to the point of use and stored until most valuable. Finally, the grid is not completely reliable, so energy is stored to carry critical services through outages.

Since energy is lost in every round trip to and from a storage medium, ongoing efforts to improve conversion efficiencies are of great economic importance.

Glossary

Bagasse	the solid residue of a plant, e.g., sugarcane, after the juice has been extracted
Battery	a collection of chemical cell
Btu	British Thermal Unit, 1055 J or 778 ft-lbf
CAES	See Compressed Air Energy Storage
Chemical Cell	a unit that stores chemical energy and converts it from and electric energy
Chemical Energy	energy stored or released in a chemical reaction
Chiller	a mechanical device that cools a refrigerant in a thermodynamic cycle
Combustion	a chemical reaction that oxidizes fuel and releases thermal energy
Compressed Air Energy storage	large-scale secondary storage, used in gas turbines
Demand Leveling	what the power company does to reduce the peak demand
EDLC	electric double-layer capacitor

Energy	the capacity to do work, usual units Btu or kWh
Ethane	C_2H_6, a minor component of natural gas
Ethanol	C_2H_6O, ethyl alcohol, sometimes used as a component of automotive fuel
Ethanol	ethyl alcohol, grain alcohol, C_2H_5OH
Flow battery	a battery in which the electrolyte flows between the cell and a storage tank
Flywheel	a device that spins to store kinetic energy
Fuel Cell	an energy conversion device that generates energy from hydrogen and oxygen
Gasoline	primarily octane
Geothermal Energy	energy associated with heat in the Earth's crust
Heat Energy	more properly called thermal energy
Kinetic	energy associated with motion
Latent Energy	energy associated with a change in phase, also called latent heat
Load Leveling	what the rate payer does to reduce the peak demand
Mechanical	Energy kinetic or potential energy
Methane	CH_4, the main component of natural gas, 1000 Btu/ft3
MMBtu	million Btu
MMBtu/hr	million Btu per hour
Natural Gas	a mixture of hydrocarbon gases, primarily methane
Natural Energy	renewable energy
Octane	C_8H_{18}, the main component of gasoline
Potential Energy	energy associated with the position of a system or its component parts
Regenerator	an energy-storing heat exchanger
Renewable Energy	energy endlessly available in nature
Saw Grass	a sedge, with leaves having edges set with small sharp teeth, a source of ethanol
Sensible Energy	energy associated with a change in temperature, also called sensible heat
SMES	See superconducting magnetic energy storage
Superconducting Magnetic Energy Storage	energy stored by energizing (inducing a magnetic current in) a magnet
T	ton of cooling, 12,000 Btu/hr
Thermal Energy	sensible plus latent internal energy
Trombe Wall	a building component that stores heat from the sun

References

1. Akhil, A.; Swaminathan, S.; Sen, R.K. Cost Analysis of Energy Storage Systems for Electric Utility Applications, available at http://infoserve.sandia.gov/cgi-bin/techlib/access-control.pl/1997/970443.pdf (accessed January 2009).
2. Zink, J.C. Who says you cant [sic] store electricity? *Power Engineering,* March 1997, available at http://pepei.pennnet.com/articles/article_display.cfm?article_id=44573 (accessed October 2008).
3. McPhee, J. *The Control of Nature,* Farrar, Strauss, and Giroux: New York, 1989.
4. Talbot, D. Lifeline for renewable power, *Technology Review*, January/February 2009.
5. Schindall, J. The Charge of the Ultra-Capacitors, available at http://www.spectrum.ieee.org/nov07/5636 (accessed February 2009).
6. Rodriguez, G.D. Operating Experience with the Chino 10 MW/40 MWh Battery Energy Storage Facility, available at http://ieeexplore.ieee.org/ielx5/852/2490/00074691.pdf? arnumber=74691 (accessed February 2009).
7. El-Wakil, M.M. *Powerplant Technology*; McGraw–Hill Primus: New York, 2002.
8. Miyake, S.; Tokuda, N. Vanadium Redox Flow Battery (VRB) for a Variety of Applications, available at http://www.electricitystorage.org/pubs/2001/IEEE_PES_Sum-mer2001/Miyake.pdf (accessed December 2008).

9. Flywheel Energy Storage, available at http://www1.eere.energy.gov/femp/pdfs/fta_flywheel.pdf (accessed February 2009).
10. Boyes, J.D. Technologies for Energy Storage, available at http://www.osti.gov/bridge/servlets/purl/756082-Q5CBAv/webviewable/756082.pdf (accessed August 2009).
11. Cengel, Y.A. *Heat and Mass Transfer,* 3rd Ed.; McGraw Hill: New York, 2007.

12

Fossil Fuel Combustion: Air Pollution and Global Warming

Dan Golomb

Introduction

Currently, fossil fuels supply about 86% of global primary energy consumption (39% oil, 24% coal, and 23% natural gas), providing energy for transportation, electricity generation, and industrial, commercial and residential uses. The air emissions of fossil fuel combustion are transported by winds and dispersed by atmospheric turbulence, eventually falling or migrating to the surface of the earth or ocean at various rates. While in the atmosphere, pollutants cause considerable harmful effects on human health, animals, vegetation, structures, and aesthetics.

In recent decades, it has become evident that rising levels of atmospheric carbon dioxide (and other greenhouse gases) have already warmed the earth's surface slightly. It is predicted that, with continuous and increasing use of fossil fuels, the warming trend will increase. Most of the increase in carbon dioxide levels is a direct consequence of fossil fuel combustion.

The goal of this entry is to describe the characteristics of fossil fuel-generated air pollutants, including carbon dioxide and other greenhouse gases, their transport and fate in the atmosphere, and their effects on human health, the environment, and global climate change.

Ambient and Emission Standards of Air Pollutants

Air pollutants, when they exceed certain concentrations, can cause acute or chronic diseases in humans, animals, and plants. They can impair visibility, cause climatic changes, and damage materials and structures. Regulatory agencies in most developed countries prescribe ambient standards (concentrations) for air pollutants. The standards are set at a level below which it is estimated that no harmful effect will ensue to human health and the environment. On the other hand, if the standards are exceeded, increased human mortality and morbidity, as well as environmental degradation, is expected. For example, in the United States, the Environmental Protection Agency (EPA) sets National Ambient Air Quality Standards (NAAQS) for six pollutants, the so-called criteria pollutants: carbon monoxide (CO), nitrogen dioxide (NO_2), ozone (O_3), sulfur dioxide (SO_2), particulate matter (PM), and lead (Pb). The standards are listed in Table 1. In the past, particulate matter was regulated regardless of size (the so-called Total Suspended Particles, TSP). Starting in 1978, particles were regulated with an aerodynamic diameter less than 10μm. This is called PM_{10}. In 1997, a new standard was introduced for particles with an aerodynamic diameter less than 2.5μm. This is called $PM_{2.5}$. The reason for regulating only small particles is that these particles can be lodged deep in the alveoli of the lungs, and hence are detrimental to our health, whereas the larger particles are filtered out in the upper respiratory tract. While there is a standard for lead concentrations in the air, this pollutant is no longer routinely monitored. With the phasing out of leaded gasoline in the 1970s, lead concentrations in the air steadily declined, and allegedly, in the United States, airborne lead no longer poses a health hazard. (Of course, lead in its other forms, such as in paint, pipes, groundwater, solder, and ores, is still a hazard.)

In addition, the EPA sets emission standards for stationary and mobile sources. Stationary sources include power plants, incinerators, steel, cement, paper and pulp factories, chemical manufacturers, refineries, and others. Mobile sources include automobiles, trucks, locomotives, ships, and aircraft. As an example, Table 2 lists the U.S. emission standards (New Source Performance Standards) for vfossil-fueled steam generators, which include large fossil-fueled power plants and industrial boilers that were constructed after 1970. The emission standards are set in terms of mass emitted per fuel heat input (g/GJ). However, in 1978, new regulations were implemented. Instead of numerical emission standards, EPA prescribed *emission control technologies,* the so-called best available control technologies (BACT), purporting to reduce emissions to a minimum with practical and economic pollution abatement devices. For example, current BACT for sulfur oxides is a "scrubber," usually employing a slurry of pulverized limestone ($CaCO_3$) counter-flowing to the flue gas that contains the sulfur oxides. For control of nitric oxides (NO_x), the current BACT is the low-NO_x-burner (LNB). Because the LNB reduces

TABLE 1 U.S. 2000 National Ambient Air Quality Standards (NAAQS)

Pollutant	Standard	Averaging Time
Carbon monoxide	9 ppm (10 mg/m^3)	8-h[a]
	35 ppm (40 mg/m^3)	1-h[a]
Lead	1.5 μg/m^3	Quarterly Average
Nitrogen dioxide	0.053 ppm (100 μg/m^3)	Annual (arithmetic mean)
Particulate matter (PM10)	50 μg/m^3	Annual[b] (arithmetic mean)
	150 μg/m^3	24-h[a]
Particulate matter (PM2.5)	15.0 pg/m^3	Annual[c] (arithmetic mean)
	65 μg/m^3	24-h[d]
Ozone	0.08 ppm	8-h[e]
Sulfur oxides	0.03 ppm	Annual (arithmetic mean)
	0.14 ppm	24-h[a]
	–	3-h[a]

[a] Not to be exceeded more than once per year.

[b] To attain this standard, the 3-year average of the weighted annual mean PM_{10} concentration at each monitor within an area must not exceed 50 μg/m^3.

[c] To attain this standard, the 3-year average of the weighted annual mean PM2.5 concentrations from single or multiple community-oriented monitors must not exceed 15.0 μg/m^3.

[d] To attain this standard, the 3-year average of the 98th percentile of 24-h concentrations at each population-oriented monitor within an area must not exceed 65 μg/m^3.

[e] To attain this standard, the 3-year average of the fourth-highest daily maximum 8-h average ozone concentrations measured at each monitor within an area over each year must not exceed 0.08 ppm.

TABLE 2 U.S. New Source Performance Standards (NSPS) for Fossil Fuel Steam Generators with Heat Input >73 MW

Pollutant	Fuel	Heat Input (g/GJ)
SO_2	Coal	516
	Oil	86
NO_2	Gas	86
	Coal (bituminous)	260
PM[a]	Coal (subbituminous)	210
	Oil	130
	Gas	86
	All	13

[a] PM, particulate matter. For PM emissions an opacity standard also applies, which allows a maximum obscuration of the background sky by 20% for a 6-min period.

Source: Data from EPA. Standards of performance for new stationary sources. Electric steam generating units. *Federal Register,* 45, February 1980, 8210–8213.

NO_x emissions only up to 50%, new sources may be required to employ more efficient NO_x-reducing technologies, such as selective catalytic reduction (SCR) or selective non-catalytic reduction (SNCR), based on ammonia or urea injection into the flue gas. For control of particulate matter, the current prescribed technology is the electrostatic precipitator (ESP). Because the ESP is not efficient in removing sub-micron and 1–2 pm particles, and in view of the new $PM_{2.5}$ standard, new sources may be required to install instead of an ESP a Fabric Filter (FF), also called a bag house. The FF consists of a porous fabric or fiberglass membrane that filters out efficiently the smallest particles, albeit at an increased cost and energy penalty compared to the ESP. The detailed description of the workings of these emission control technologies is beyond the scope of this entry; the reader is referred to the excellent handbooks on the subject.[1,2]

Table 3 lists the U.S. emission standards in units of mass emitted per length traveled (g/km) for passenger cars and light trucks for the different model years from 1968 to 1994. At present, the 1994 standards are still in effect. Notice that light truck standards are not as strict as those for light duty vehicles. This dichotomy is very controversial, because the light truck category includes sport utility vehicles (SUV), minivans, and pick-up trucks, which are mostly used for personal and not for cargo transport. Therefore, their emission standards ought to be equal to those for passenger cars. Until now, these vehicles captured the majority of sales in the United States. (With rising gasoline prices, this trend may be reversed.) The achievement of the emission standards relies on emission control technologies. For unleaded gasoline-fueled vehicles, the prevailing control technology is the three-way catalytic converter, which simultaneously reduces carbon monoxide (CO), nitric oxides (NO_x), and fragmentary hydrocarbon (HC) emissions. Unfortunately, the catalytic converter would not work on diesel-fueled vehicles, because the relatively high sulfur and particle emissions from diesel engines would poison the catalyst. Automobile and truck manufacturers are intensively investigating possible technologies that would reduce emissions from diesel-fueled autos and trucks.

In addition to the aforementioned criteria pollutants, one finds in the air a host of other gaseous and particulate pollutants, generally designated as hazardous air pollutants (HAP), or simply *air toxics.* The EPA has identified 189 HAPs. Of course, not all HAPs are related to fossil fuel usage. Examples of fossil fuel HAPs are products of incomplete combustion (PIC), volatile organic compounds (VOC), polycyclic aromatic hydrocarbons (PAH), toxic metals (e.g., mercury, cadmium, selenium, arsenic, vanadium, etc.) Many of the fossil fuel–related HAPs are found as condensed matter on particles emitted by stationary sources (e.g., fly ash from power plants) or mobile sources (e.g., exhaust smoke from trucks). While HAPs may be more harmful to our health than the criteria pollutants (some of them are carcinogens), it is difficult to establish a dose–response relationship. Therefore, instead of setting HAP emission standards,

TABLE 3 U.S. Federal Vehicle Emission Standards

	Light-duty Vehicles (Auto)				Light-duty Trucks (Gasoline)			
Model Year	HC (g/km)	CO (g/km)	NO (g/km)	PM (g/km)	HC (g/km)	CO (g/km)	NOx (g/km)	PM (g/km)
1968	2.00	20.50						
1971[a]	2.90	29.20	2.49					
1974	2.11	24.20	1.86					
1977	0.93	9.32	1.24					
1978	0.93	9.32	1.24		1.24	12.40	1.93	
1979	0.93	9.32	1.24		1.06	11.20	1.43	
1980	0.25	4.35	1.24		1.06	11.20	1.43	
1981	0.25	2.11	0.62		1.06	11.20	1.43	
1982	0.25	2.11	0.62	0.37	1.06	11.20	1.43	
1985	0.25	2.11	0.62	0.37	0.50	6.21	1.43	0.99
1987	0.25	2.11	0.62	0.12	0.50	6.21	1.43	1.62
1988	0.25	2.11	0.62	0.37	0.50	6.21	0.75	1.62
1994	0.25	2.11	0.25	0.05	0.50	3.42	0.60	0.06

[a] Test method changed in 1971.

the EPA mandates that specific control technologies be installed on major emitting sources. These are called maximum achievable control technologies (MACT). For example, for toxic volatile gases, the most often used control technologies are physical adsorption on porous adsorbents (usually activated carbon), or chemical absorption in solvents, including water, that have a large absorption capacity for these gases. When the concentration of toxic gases in the effluent is high, secondary incineration of the effluent may be warranted. For detailed descriptions of MACTs, the reader is also referred to the appropriate handbooks.[1,2]

Pollutant Transport and Dispersion

When air pollutants exit a smoke stack or exhaust pipe (called the *sources*), they are transported by winds and dispersed by turbulent diffusion. Winds blow from high pressure toward low-pressure cells at speeds that depend on the pressure gradient. Because of the Coriolis force, wind trajectories are curvilinear in reference to fixed earth coordinates, although within a relatively short (few to tens of km) distance, wind trajectories can be approximated as linear. Winds have a horizontal and vertical component. Over flat terrain, the horizontal component predominates; in mountainous and urban areas with tall buildings, the vertical component can be significant, as well at the land/sea interface.

In the bottom layer of the atmosphere, called the troposphere, the temperature usually declines with altitude. For a dry atmosphere, the temperature gradient is −9.6°C/km. This is called the *dry adiabatic lapse rate.* In a moist atmosphere, the temperature gradient is less. Under certain conditions (e.g., nocturnal radiative cooling of the earth surface), the temperature gradient may even become positive. This is called *inversion.* The global and temporal average temperature gradient in the troposphere is −6.5°C/km. When a negative temperature gradient exists, upper parcels of the air are denser than lower ones, so they tend to descend, while lower parcels buoy upward, giving rise to eddy or turbulent mixing. The more negative the temperature gradient, the stronger the turbulence, and the faster the dispersion of a pollutant introduced into the troposphere. Wind shears also cause turbulence. A wind shear exists when, in adjacent layers of the atmosphere, winds blow in different directions and speeds.

FIGURE 1 Los Angeles smog. Note the low inversion beneath which the smog accumulates. Above the inversion, the air is relatively clear. Photo by the South Coast Air Quality Management District.

During an inversion, there is little turbulence, and a pollutant will disperse very slowly. Inversions may also occur aloft, that is, a negative temperature gradient exists at the bottom, followed by a positive gradient above. The layer up to the altitude at which the inversion occurs is called the *mixing layer* and the altitude at the inversion is called the *mixing height*. The shallower the mixing layer, the greater chances for air pollution episodes to develop, because pollutants emitted near the ground are confined to the shallow mixing layer. This often occurs in cities such as Los Angeles, Houston, Atlanta, Salt Lake City, Denver, Mexico City, Sao Paulo, Athens, Madrid, Rome, and Istanbul. Some of these cities are surrounded by mountain chains. In the basin of the mountain chain, or in the valley, the mixing layer is shallow, and the winds in the layer are usually weak, leading to poor ventilation. During the morning rush hour, pollutants are emitted into the shallow mixing layer, where they are concentrated because of the “lid” imposed by the inversion aloft. Later in the day, when solar radiation breaks up the inversion, the pollutants disperse to higher altitude, and the pollutant concentration becomes more diluted. The photo in Figure 1 shows a pollution episode in Los Angeles. The mixing height extends only to the middle of the tall building. The lower part is obscured by the smog, the upper part, which is above the mixing layer, is in relatively clear air.

Air Quality Modeling

The estimation of ambient pollutant concentrations in space and time due to emissions of single or multiple sources is called air quality modeling (AQM), or source–receptor modeling (SRM). The basic ingredients of AQM are the emission strengths of the sources, meteorological conditions, and solar irradiation. Air quality models are of the trajectory-type, where the coordinate system moves together with the plume of pollutants, or grid-type, where the coordinate system is fixed over an area, and the emission strengths and meteorological variables are inserted in each grid of the model domain. Some models consider only conservative pollutants, where the emitted pollutant does not change en route to the receptor. Other models consider chemical transformation processes and “sinks,” e.g., dry and wet deposition to the ground. Air quality models where transformation and deposition processes need to be considered are acid deposition, regional haze, and photo-oxidants.

Acid Deposition

While commonly called acid rain, acid deposition is a better term, because deposition can occur both in the wet and dry form. The ingredients of acid deposition are sulfuric and nitric acids. The primary pollutants are SO_2 and NO_x, which is the sum of NO and NO_2 molecules. Both SO_2 and NO_x result from fossil fuel combustion. Sulfur is a ubiquitous ingredient of coal and petroleum. When these fuels are burned in air, SO_2 is emitted from a smoke stack or the exhaust pipe of a vehicle. Coal and petroleum also contain nitrogen in their molecular make-up, resulting in NO_x emissions when these fuels are combusted in air. In addition, some NO_x is formed from the recombination of air oxygen and nitrogen at the high flame temperatures. So, even the combustion of natural gas, which has no nitrogen in its molecular make-up, produces some NO_x

Primary emitted SO_2 and NO_x are transformed in the atmosphere to sulfuric and nitric acids. The resulting acids can be either deposited in the dry phase on land or water, a process called dry deposition, or scavenged by falling rain drops or snow flakes, resulting in wet deposition.[3] Acid deposition modeling is successful in predicting the amount of acid deposition given the emission strength of the precursors SO_2 and NO_x.[4] From these models, it was concluded that acid deposition within a geographic domain is approximately linearly dependent on SO_2 and NO_x emission strength in that domain, so a certain percentage of precursor emission reduction results in a proportional deposition reduction of sulfuric or nitric acid. Indeed, in countries and continents where serious curtailments of precursor emissions have been made, a proportional reduction of acid deposition occurred. In the United States, as a result of reducing emissions of SO_2 by approximately one-half since the enactment of the Clean Air Act of 1990, sulfuric acid deposition has declined by approximately one-half.[5] Nitric acid deposition has not fallen appreciably, because the control of NO_x is much more difficult to accomplish than the control of SO_2, especially from dispersed sources such as commercial and residential boilers and furnaces, automobiles, and diesel trucks.

Regional Haze

Small particles (also called fine particles or aerosols), less than 1–2 μm in diameter, settle very slowly on the ground. Small particles can be either in the solid or liquid phase. For example, fly ash or smoke particles are solid; mist is liquid. Small particles can be of natural origin (e.g., volcanic dust, wind blown soil dust, forest and brush fires) or of anthropogenic origin (e.g., fly ash, diesel truck smoke). They can be emitted as primary particles or formed by transformation and condensation from primary emitted gases. They can travel hundreds to thousands of kilometers from their emitting sources. The particles can envelope vast areas, such as the northeastern or southwestern United States, southeastern Canada, western and central Europe, and southeastern Asia. Satellite photos often show continental areas covered with a blanket of haze. The haze may extend far out over the ocean. This phenomenon is called regional haze. Small particles are efficient scatterers of sunlight. Light scattering prevents distant objects from being seen. This is called visibility impairment. Figure 2 shows an encroaching haze episode on a mountain chain in Vermont. As the haze thickens, distant mountain peaks are no longer visible, and eventually neighboring peaks disappear. Increasing concentration of particles in urbanized parts of continents causes the loss of visibility of the starlit nocturnal sky. These days, small stars, less than the fifth order of magnitude, rarely can be seen from populated areas of the world.

The composition of fine particles varies from region to region, depending on the precursor emissions. In the northeastern United States, central Europe, and southeastern Asia, more than half of the composition is made up of sulfate particles, due to the combustion of high sulfur coal and oil. The rest is made up of nitrate particles, carbonaceous material (elemental and organic carbon), and crustal matter (fugitive particles from soil, clay, and rock erosion).

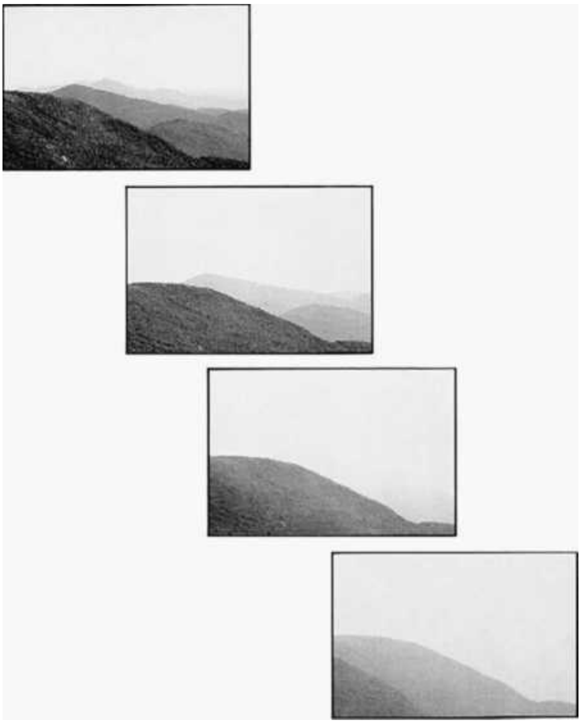


FIGURE 2 Regional haze in the Vermont green mountains. Note that as the pollution episode progresses, adjacent peaks are no longer visible. Photo by the Vermont agency of environmental conservation.

Photo-Oxidants

The family of photo-oxidants includes tropospheric ozone, O_3 (the bad ozone), ketones, aldehydes, and nitrated oxidants, such as peroxyacetylnitrate (PAN) and peroxybenzoylnitrate (PBN). The modeling of photo-oxidants is more complicated than that of acid deposition.[6] Here, the primary precursor is NO_x, which as mentioned previously, is emitted because of fossil fuel combustion. A part of NO_x is the NO_2 molecule, which splits (photo-dissociates) by solar ultraviolet and blue photons into NO and atomic oxygen. The photo-dissociation rate is dependent on solar irradiation, which, in turn, is dependent on latitude, season, time of day, and cloudiness. Atomic oxygen combines with molecular oxygen to form O_3. The NO that is formed in the photo-dissociation is quickly re-oxidized into NO_2 by peroxy radicals, RO_2, present in the polluted atmosphere. The peroxy radicals are formed from VOCs that are emitted as a consequence of incomplete combustion of fossil fuels, or from evaporation and leakage of liquid fossil fuels and solvents. The VOCs are oxidized in a complicated sequence of photo-chemical reactions to the peroxy radicals. The oxidation rates of VOCs are also dependent on solar irradiation and on the specific VOC molecule. Long-and branch-chained hydrocarbons (e.g., *n*-octane and *iso*octane) are more reactive then short-and straight-chained ones (e.g., methane and ethane); unsaturated hydrocarbons (e.g., ethene) are more reactive than saturated ones (e.g., ethane). Aromatic hydrocarbons (e.g., benzene) are more reactive than aliphatic ones (e.g., hexane), and so on.

Thus, photo-oxidants have two kinds of precursors, NO_x and VOCs, which make abatement of these secondary pollutants, as well as their modeling, so complicated. First of all, as mentioned previously, complete NO_x emission control is difficult to accomplish, because in addition to coming from large stationary sources, NO_x is emitted from a myriad of dispersed sources, such as home and commercial furnaces and boilers, automobiles, trucks, off-road vehicles, aircraft, locomotives, and ships. In principle, anthropogenic VOCs could be substantially controlled, for example, by ensuring complete combustion of the fossil fuel, or with the catalytic converter on automobiles. But, not all VOCs are of anthropogenic origin. Trees and vegetation emit copious quantities of VOCs, such as terpenes and pinenes, which are pleasant smelling, but they do participate in photo-chemical reactions that produce photo-oxidants. Even though great effort and expenses are being made in many developed countries to control precursors, photo-oxidant concentrations over urban-industrial continents have improved only slightly, if at all. In less developed countries that do not have the means of controlling photo-oxidant precursors, their concentrations are on a steady increase.

Global Warming

Of all environmental effects of fossil fuel usage, global warming, including its concomitant climate change, is the most perplexing, potentially most threatening, and arguably most intractable problem. It is caused by the ever-increasing accumulation in the atmosphere of carbon dioxide (CO_2) and other gases, such as methane (CH_4), nitrous oxide (N_2O), and chloro-fluoro-carbons (CFC), collectively called greenhouse gases (GHG). Atmospheric aerosols, natural as well as anthropogenic, also may contribute to global warming.

The term greenhouse effect is derived by analogy to a garden greenhouse. There, a glass-covered structure lets in the sun's radiation, warming the soil and plants that grow in it, while the glass cover restricts the escape of heat into the ambient surroundings by convection and radiation. Similarly, the earth atmosphere lets through most of the sun's radiation, which warms the earth surface, but the GHGs and some aerosols trap outgoing terrestrial infrared (IR) radiation, keeping the earth's surface warmer than if the GHGs and aerosols were absent from the atmosphere.

A schematic of the greenhouse effect is represented in Figure 3. To the left of the schematic is the atmospheric temperature structure, which defines the various spheres. In the troposphere, the temperature decreases on average by 6.5°C/km. This sphere is thoroughly mixed by thermal and mechanical turbulence, and this sphere contains most of the air pollutants mentioned in the previous sections. In the stratosphere, the temperature increases steadily. Because of the positive temperature gradient, this sphere is very stable with very little turbulence and mixing. At about 50–60 km height, the temperature declines again with altitude, giving rise to the mesosphere. Finally, above 90–100 km, the temperature gradient reverses itself, and becomes positive. The highest sphere is called the thermosphere. Its temperature can reach hundreds to one thousand degrees, depending on solar radiation intensity that directly heats this sphere.

The average solar radiation that impinges on the top of the atmosphere is about 343 W/m^2. This is the annual, diurnal, and spatial average irradiation. Of this irradiation, currently about 30% (103 W/m^2) is immediately reflected into space by land, ocean, icecaps, and clouds, and scattered into space by atmospheric molecules. The reflected and scattered sunlight is called albedo. The albedo may not remain constant over time. With increased melting of the ice caps, and increased cloud cover, in part due to anthropogenic influences, the albedo may change over time. The remaining 70% of solar irradiation heats the earth's surface, land, and oceans. Currently, the global average surface temperature is 288 K (about 15°C). A body (the so-called black body) that is heated to 288 K radiates 390 W/m^2. The earth's surface radiation occurs in the far IR. This is called earth shine. A part of the earth shine is reflected back to the earth's surface by clouds and aerosols; another part is first absorbed by certain gaseous molecules, and then re-radiated back to the surface. The absorption/re-radiation occurs by poly-atomic molecules, including water vapor and the GHGs: CO_2, CH_4, N_2O, O_3, CFCs, and others. The reflection and re-radiation to the earth's surface of the outgoing terrestrial IR radiation is causing the earth's surface to become warmer than it would be merely by solar irradiation. This is the greenhouse effect. With increasing concentrations of anthropogenic GHGs and aerosols, the earth's surface temperature will

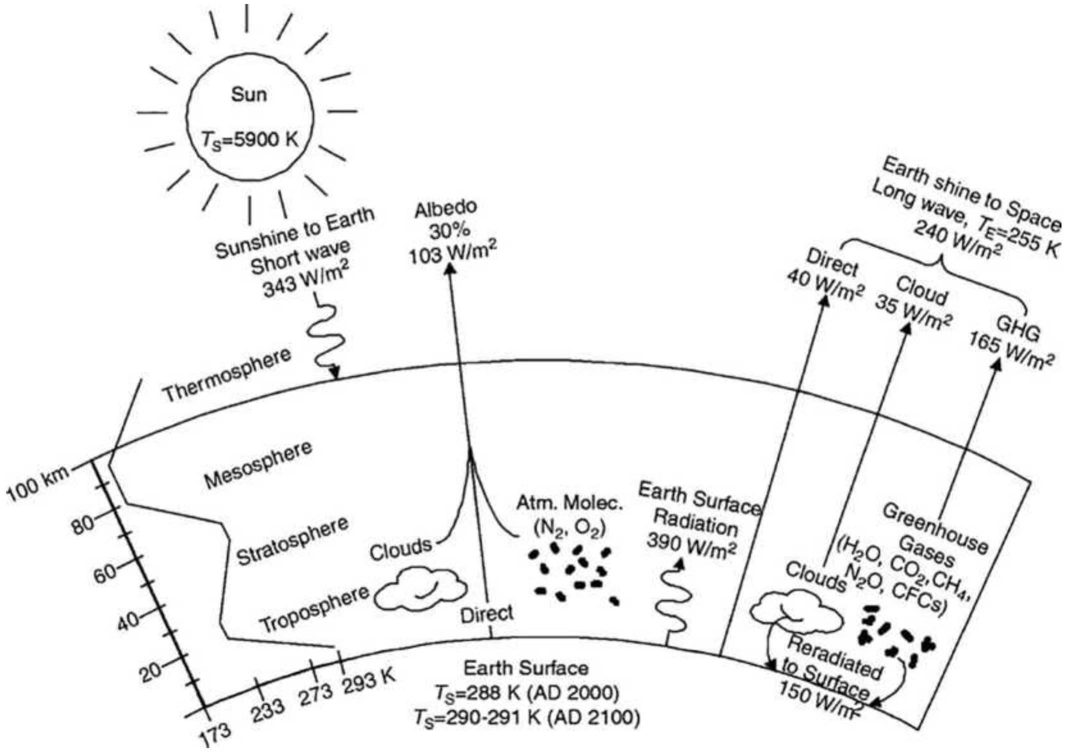


FIGURE 3 Schematic of the greenhouse effect.

increase. It should be emphasized that the greenhouse effect is not due to trapping of the incoming solar radiation, but outgoing terrestrial IR radiation. In fact, because of the trapping of earth shine by GHGs, the upper layers of the atmosphere (primarily the stratosphere) will become colder, not warmer.[7]

The extent of global warming can be predicted by radiative transfer models. These models include the radiative properties of GHGs and their distribution in the earth's atmosphere, as well as the temperature and pressure gradients in the atmosphere. There is general agreement among the models as to the extent of surface warming due to GHG absorption/re-radiation, called radiative forcing. Based on the models, it is predicted that the average earth surface temperature will increase as shown in Figure 4. The middle, "best," estimate predicts a rise of the earth's surface temperature by the end of the 21st century of about 2°C; the "optimistic" estimate predicts about 1°C, and the "pessimistic" estimate predicts about a 3°C rise. The optimistic estimate relies on the slowing of CO_2 and other GHG emissions; the pessimistic estimate relies on "business-as-usual," i.e., on the continuing rate of growth of CO_2 and other GHG emissions, and the "best" estimate is somewhere in between.[8] If the GHG concentrations increase in the atmosphere at their current rate, by the year 2100, CO_2 will contribute about 65% to global warming, CH_4 15%, N_2O 10%, and CFC about 5%–10%. (By international conventions, CFCs are being phased out entirely. But because of their very long lifetime in the atmosphere, they still will contribute to global warming by the year 2100.)

In addition to radiative forcing, global warming may be enhanced by the so-called feedback effects. For example, water vapor is a natural GHG. When the temperature of the ocean surface increases, the evaporation rate will increase. As a consequence, the average water vapor content of the atmosphere will increase. This causes more absorption of the outgoing infrared radiation and more global warming. Furthermore, increased evaporation may cause more cloud formation. Clouds and aerosols also can

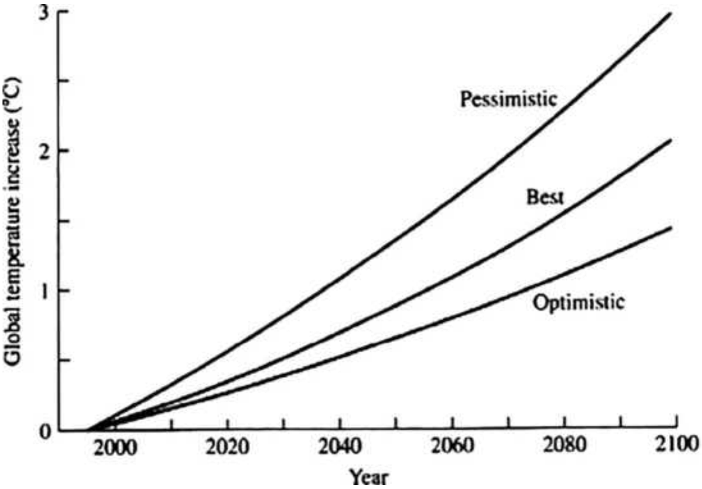


FIGURE 4 Projected trend of the earth surface temperature increase. Upper curve: pessimistic scenario with no emission curtailment; lower curve: optimistic scenario with significant emission curtailment; middle curve: in-between scenario.
Source: Cambridge University Press (see Intergovernmental Panel on Climate Change).[8]

trap outgoing terrestrial radiation, further increasing global warming. Melting ice caps and glaciers decrease the reflection of incoming solar radiation (reduced albedo), which also increases global warming. The prediction of the feedback effects is more un-certain than the prediction of radiative forcing, but generally, it is assumed that the feedback effects may double the surface temperature increases because of radiative forcing alone.

Has the surface temperature already increased due to anthropogenic activities? Figure 5 plots the global average surface temperature over the last century and a half. Even though there are large annual fluctuations, the smoothed curve through the data points indicates an upward trend of the temperature. From 1850, the start of the Industrial Revolution, to date, the global average surface temperature increased by about 0.5°C–1°C. This is in accordance with radiative models that predict such a trend, considering the increase of GHG concentrations over that period.

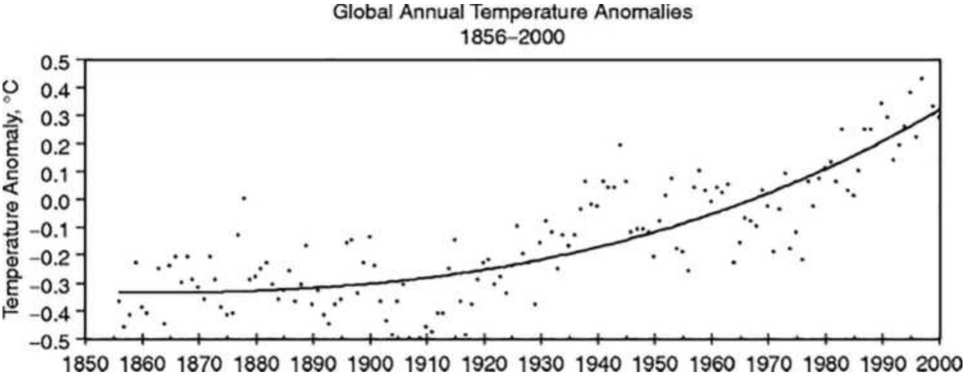


FIGURE 5 Global average surface temperature trend 1850–2000.
Source: Oak Ridge National Laboratory (see CO_2 Information Center[9]).

Other Effects of Global Warming

Because of increased GHG concentrations in the atmosphere, the earth's surface temperature may rise, as discussed in the previous section. The surface temperature rise may cause several ancillary effects on global climate and hydrogeology, which in turn will affect human habitat, welfare, and ecology.

Sea Level Rise

With increasing surface temperatures, the average sea level will rise because of three factors: melting of polar ice caps, receding of glaciers, and thermal expansion of the ocean surface waters. Combining all three factors, it is estimated that by the end of the next century, the average sea level may be 30–50 cm higher than it is today. This can seriously affect low-lying coastal areas, such as the Netherlands in Europe, Bangladesh in Asia, and low-lying islands in the Pacific and other oceans.[8]

Climate Changes

Predicting global and regional climatic changes because of average surface temperature rise is extremely difficult and fraught with uncertainties. It is expected that regional temperatures, prevailing winds, and storm and precipitation patterns will change, but where and when, and to what extent changes will occur, is a subject of intensive investigation and modeling on the largest available computers, the so-called supercomputers. Climate is not only influenced by surface temperature changes, but also by biological and hydrological processes, and by the response of ocean circulation, which are all coupled to temperature changes. It is expected that temperate climates will extend to higher latitudes, probably enabling the cultivation of grain crops further toward the north than at present. But crops need water. On the average, the global evaporation and precipitation balance will not change much, although at any instant, more water vapor (humidity) may be locked up in the atmosphere. However, precipitation patterns may alter, and the amount of rainfall in any episode may be larger than it is now. Consequently, the runoff (and soil erosion) may be enhanced, and areas of flooded watersheds may increase. Hurricanes and typhoons spawn in waters that are warmer than 27°C, in a band from 5 to 20° north and south latitude. As the surface waters become warmer, and the latitude band expands, it is very likely that the frequency and intensity of tropical storms will increase.

The sea level and climatic changes may cause a redistribution of agricultural and forestry resources, a considerable shift in population centers, and incalculable investments in habitat and property protection.

Greenhouse Gas Concentrations Trends

Currently, about 6.8 Gt/y of carbon (25 Gt/y CO_2) are emitted into the atmosphere by fossil fuel combustion. Another 1.5 ± 1 Gt/y are emitted due to deforestation and land use changes, mainly artificial burning of rain forests in the tropics, and logging of mature trees, which disrupts photosynthesis. Figure 6 plots the trend of atmospheric concentrations of CO_2, measured consistently at Mauna Loa, Hawaii since 1958 to date. At present, the average CO_2 concentration is about 375 parts per million by volume (ppmv). The plot shows seasonal variations due mainly to assimilation/respiration of CO_2 by plants, but there is a steady increase of the average concentration at a rate of approximately 0.4%/y. If that rate were to continue into the future, a doubling of the current CO_2 concentration would occur in about 175 years. However, if no measures are taken to reduce CO_2 emissions, then due to the population increase, and the concomitant enhancement of fossil fuel use, the rate of growth of CO_2 concentration will increase more than 0.4%/y, and the doubling time will be achieved sooner.

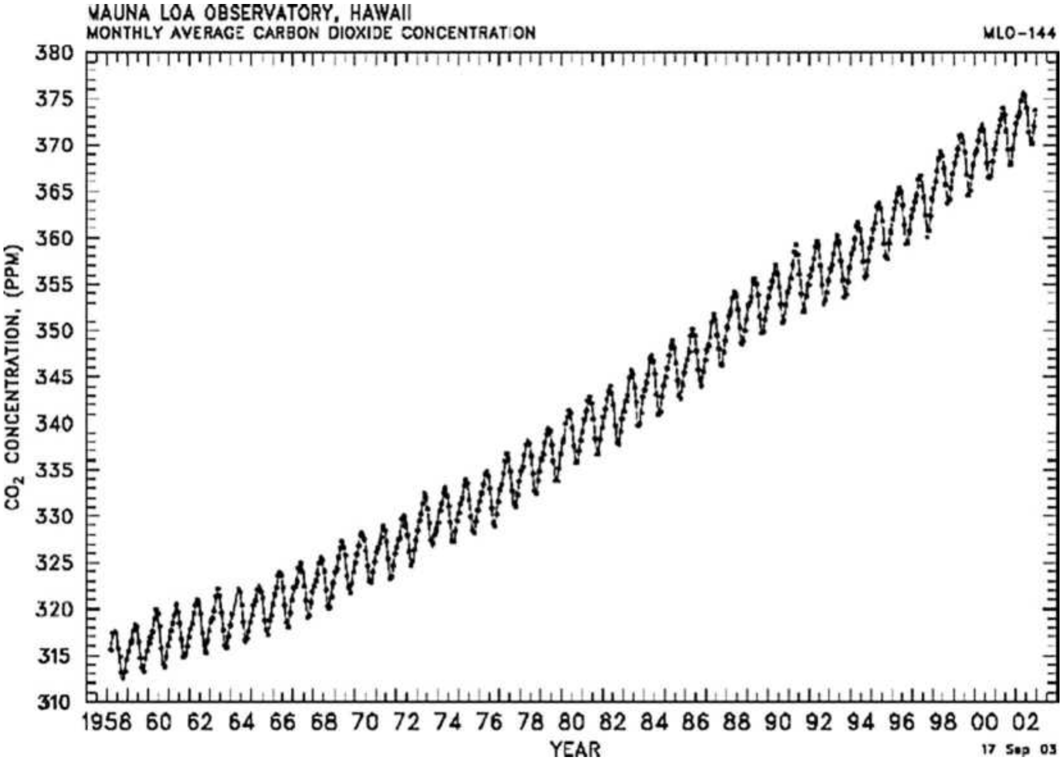


FIGURE 6 Carbon dioxide concentration trend 1958–2002.
Source: Oak Ridge National Laboratory (see CO_2 Information Center).[9]

Methane emissions are in part due to fossil energy usage, because CH_4 leaks from gas pipes, storage tanks, tankers, and coal mine shafts. Anthropogenic emissions of CH_4 from fossil fuel usage amount to about 100 Mt/y. However, CH_4 is also emitted from municipal waste landfills, sewage treatment, biomass burning, cultivated rice paddies, enteric fermentation of cattle, and other anthropogenic activities, so that the total amount of CH_4 emissions is about 400 Mt/y. Currently, the average atmospheric concentration of CH_4 is about 1.7 ppmv, growing at about 0.6%/y. Nitrous oxide (N_2O) is a minor product of combustion of fossil fuels. Currently, the concentration of N_2O is about 0.3 ppmv, growing by about 0.25%/y. CFCs are not directly associated with fossil fuel usage; however, they are emitted inadvertently from energy using devices, such as refrigerators, air conditioners, chillers, and heat pumps. Current concentrations of the various CFCs are about 0.5 parts per billion by volume (ppbv), and their concentrations are slowly declining due to the phase-out of production of CFCs.

What Can Be Done about Global Warming

Global warming could be lessened by reducing significantly the emissions of GHGs into the atmosphere. Most GHG emissions are a consequence of fossil fuel use. While it is important to reduce the emissions of all GHGs, the greatest preventative measure would come from reducing CO_2 emissions. CO_2 emission reductions can be accomplished by a combination of several of the following approaches.

Demand-Side Conservation and Efficiency Improvements

This includes less space heating and better insulation, less air conditioning, use of more fluorescent instead of incandescent lighting, more energy efficient appliances, process modification in industry, and very importantly, more fuel efficient automobiles. Some measures may even incur a negative cost, i.e., consumer savings by using less energy, or at least a rapid payback period for the investment in energy saving devices.

Supply-Side Efficiency Measures

Here we mean primarily increasing the efficiency of electricity production. Natural gas combined cycle power plants (NGCC) emit less CO_2 than single cycle coal fired power plants. First, because NG emits about one-half the amount of CO_2 per fuel heating value than coal, and second, because the thermal efficiency of combined cycle power plants is in the 45%–50% range vs. the 35%–38% range of single cycle plants. In the future, integrated coal gasification combined cycle power plants (IGCC) may come on-line with a thermal efficiency in the 40%–45% range reckoned on the basis of the coal heating value. Furthermore, IGCC may enable the capture of CO_2 at the gasification stage, with subsequent sequestration of the captured CO_2 in geologic and deep ocean repositories.

Shift to Non-Fossil Energy Sources

The choices here are agonizing, because the largest impact could be made by shifting to nuclear and hydro electricity, both presently very unpopular and fraught with environmental and health hazards. The shift to solar, wind, geothermal, and ocean energy is popular, but because of their limited availability and intermittency, and because of their larger cost compared to fossil energy, a substantial shift to these energy sources cannot be expected in the near future.

None of these options can prevent global warming by itself. They have to be taken in combination and on an incremental basis, starting with the least expensive ones and progressing to the more expensive ones. Even if the predictions of global climate change were to turn out exaggerated, the fact that fossil fuel usage entails many other environmental and health effects, and the certainty that fossil fuel resources are finite, makes it imperative that we curtail fossil energy usage as much as possible.

Summary and Conclusions

The use of fossil fuels to supply energy for the use of the world's population has resulted in the release to the atmosphere of troublesome chemical byproducts that present harm to humans and other natural species. These effects can be localized (near the emission source), can extend to large regional areas (involving sub-continents), and can even cover the globe, from pole to pole. A large portion of the human population is exposed to one or more of these environmental effects. The scientific understanding of how fossil fuel use causes these effects is well advanced, providing quantitative means for explaining what is currently observed and predicting what changes will occur in the future from projected future fuel consumption. These projections provide a basis for modifying the amount and character of future energy supply so as to lessen harmful environmental consequences. The technological systems that employ fossil fuel energy have been developed to lessen the amounts of harmful emissions, albeit at significant energy and economic cost. Further improvements can be expected, but at increasing marginal cost. The most severe emission control problem, in terms of economic and energy cost, is CO_2, a major contributor to global warming. The implementation of policies by national governments and international bodies to curtail the use of fossil energy and the concomitant emissions of CO_2 will become a growing task for humankind in this century.

References

1. Wark, K.; Warner, C.F.; Davis, W.T. *Air Pollution: Its Origin and Control,* 3rd Ed.; Addison-Wesley: Menlo Park, CA, 1998.
2. Cooper, C.D.; Alley, F.C. *Air Pollution Control: A Design Approach,* 3rd Ed.; Waveland Press, Prospects Heights: IL, 2002.
3. Seinfeld, J.H.; Pandis, S.N. *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*; Wiley: New York, 1998.
4. Fay, J.A.; Golomb, D.; Kumar, S. Source apportionment of wet sulfate deposition in Eastern North America. Atm. Environ. **1985**, 19, 1773–1782.
5. Butler, T.J.; Likens, G.E.; Stunder, B.J.B. Regional-scale impacts of phase I of the clean air act amendments in the U.S.A.: the relation between emissions and concentrations, both wet and dry. Atmos. Environ. **2000**, 35, 1015–1028.
6. National Research Council. *Rethinking the Ozone Problem in Urban and Regional Air Pollution*; National Academy Press: Washington, DC, 1991.
7. Mitchell, J.F.B. The greenhouse effect and climate change. Rev. Geophys. **1989**, 27, 115–139.
8. Intergovernmental Panel on Climate Change. *Climate Change: The Scientific Basis*; Cambridge University Press: Cambridge, U.K., 2001.
9. CO_2 Information Center. *Trends Online*; Oak Ridge National Laboratory: Oak Ridge, TN, 2000.

13

Geothermal Energy Resources

Ibrahim Dincer
and Arif Hepbasli

Introduction

With increasing awareness of the detrimental effects of the burning of fossil fuels on the environment, there has been an increasing interest worldwide in the use of clean and renewable energy sources. Although geothermal energy is categorized in international energy tables as one of the "new renewables," it is not a new energy source at all. People have used hot springs for bathing and washing clothes since the dawn of civilization in many parts of the world.[1] Heat is a form of energy, and *geothermal energy* is, literally, the heat contained within the Earth that generates geological phenomena on a planetary scale. "Geothermal energy" is often used nowadays, however, to indicate that part of the Earth's heat that can, or could, be recovered and exploited by man, and it is in this sense that we will use the term from now on.[2] Geothermal energy utilization is basically divided into two categories, i.e., electric energy production and direct uses. Direct use of geothermal energy is one of the oldest, most versatile, and also the most common form of utilization of geothermal energy.[3] The early history of geothermal direct use has been well documented for over 25 countries in the *Stories from a Heat Earth—Our Geothermal Heritage*[4] that documents geothermal use for over 2000 years.[5]

Geothermal energy has until recently had a considerable economic potential only in areas where thermal water or steam is found concentrated at depths less than 3 km in restricted volumes, analogous to oil in commercial oil reservoirs. The use of ground-source heat pumps has changed this. In this case, the earth is the heat source for the heating and/or the heat sink for cooling, depending on the season. This has made it possible for people in all countries to use the heat of the earth for heating and/or cooling, as appropriate. It should be stressed that the heat pumps can be used basically everywhere.[1]

Based on country update papers submitted for the World Geothermal Congress 2010 (WGC-2010), covering the period 2005 to 2009, an estimate of the installed thermal power for direct utilization at the end of 2009 is 50,583 MW_t, almost a 79% increase over the 2005 data, growing at a compound rate of 12.3% annually with a capacity factor of 0.27. The thermal energy used is 438,071 TJ/yr (121,696 GWh/yr), about a 60% increase over 2005, growing at a compound rate of 9.9% annually. The distribution of thermal energy used by category is approximately 49.0% for ground-source heat pumps, 24.9% for bathing and swimming (including balneology), 14.4% for space heating (of which 85% is for district heating), 5.3% for greenhouses and open ground heating, 2.7% for industrial process heating, 2.6% for aquaculture pond and raceway heating, 0.4% for agricultural drying, 0.5% for snow melting and cooling, and 0.2% for other uses. Energy savings amounted to 307.8 million barrels (46.2 million tons) of equivalent oil annually, preventing 46.6 million tons of carbon and 148.2 million tons of CO_2[5] being released to the atmosphere, which includes savings in geothermal heat pump cooling (compared to using fuel oil to generate electricity).

The main objective of this entry is to provide a comprehensive picture about geothermal energy resources and their use for various purposes, ranging from power production to hydrogen production. It first gives a brief historical development of geothermal energy resources and geothermal energy applications. The following sections delve the nature of geothermal resources, classification of geothermal resources by energy and exergy, Lindal diagram, and some energetic and exergetic parameters for practical applications.

Historical Background

Early humans probably used geothermal water that occurred in natural pools and hot springs for cooking, bathing, and hot water and heating purposes. Archeological evidence indicates that the Indians of the Americas occupied sites around these geothermal resources for over 10,000 years to recuperate from battle and take refuge. Many of their oral legends describe these places and other volcanic phenomena. Recorded history shows uses by Romans, Japanese, Turks, Icelanders, Central Europeans, and the Maori of New Zealand for bathing, cooking, hot water, and space heating. Baths in the Roman Empire, the middle kingdom of the Chinese, and the Turkish baths of the Ottomans were some of the early uses of balneology, where body health, hygiene, and discussions were the social custom of the day. This custom has been extended to geothermal spas in Japan, Germany, Iceland, and countries of the former Austro-Hungarian Empire, the Americas, and New Zealand. Early industrial applications include chemical extraction from the natural manifestations of steam, pools, and mineral deposits in the Larderello region of Italy, with boric acid being extracted commercially starting in the early 1800s. At Chaudes-Aigues in the heart of France, the world's first geothermal district heating system was started in the 14th century and is still in use. The oldest geothermal district heating project in the United States is on Warm Springs Avenue in Boise, Idaho, which came on line in 1892 and continues to provide space heating for up to 450 homes.[2,4]

As outlined in various literature (e.g., Dickson and Fanelli[2] and Cataldi[4]), a brief summary is given as follows. The first use of geothermal energy for electric power production was carried out in Italy with some experimental tests by Prince Gionori Conti between 1904 and 1905. The first commercial power plant with a capacity of 250 kW_e was commissioned in 1913 at Larderello, Italy. An experimental plant was installed in the Geysers in 1932 to power a local resort. Such initiatives were followed in Wairakei (New Zealand) in 1958, with an experimental plant in Pathe (Mexico) in 1959, and the first commercial plant at The Geysers in the United States in 1960. Japan followed with a capacity of 23 MW_e in Matsukawa in 1966. All of these early plants utilized steam directly from the earth (mainly dry steam fields), except the one in New Zealand, which was considered the first to use flashed or separated steam for running the turbines. The former USSR produced power from the first true binary power plant with a capacity of 680 kW_e using 81°C water in Paratunka on the Kamchatka Peninsula—the lowest temperature at that time. Iceland first produced power at Namafjall in the Northern part of the country by using a 3 MWe non-condensing turbine. These were later followed by various plants in El Salvador, China, Indonesia, Kenya, Turkey, Philippines, Portugal (Azores), Greece, and Nicaragua in the 1970s and 1980s. Some recent plants have been installed in Thailand, Argentina, Taiwan, Australia, Costa Rica, Austria, Guatemala, and Ethiopia, with the latest installations in Germany and Papua New Guinea.

More recently, renewable energy resources have received increasing interest for sustainable hydrogen production. In this regard, Balta et al.[6,7] have recently studied on geothermal based hydrogen production methods comprehensively and defined various potential methods using thermochemical and hybrid cycles driven essentially by heat, as considered appropriate for geothermal sources. This has further widened the scope of geothermal energy resources. Geothermal based hydrogen production offers potential advantages over other sources for a growing hydrogen economy. These advantages do not ensure that geothermal based hydrogen will be the only option, but it will compete with other hydrogen production options. In some countries, geothermal energy is considered a primary energy source of producing hydrogen since it provides reliable energy supply in an environmentally benign manner. Hydrogen can be produced from geothermal energy using five potential methods such as 1) direct production of hydrogen from geothermal steam; 2) water electrolysis using the electricity generated from a power plant; 3) high- temperature steam electrolysis; 4) thermochemical cycles driven by geothermal heat; and 5) hybrid cycles driven by geothermal heat and electricity. A comprehensive discussion of these methods is available elsewhere.[6,7]

Nature of Geothermal Resources

The *geothermal gradient* expresses the increase in temperature with depth in the Earth's crust. Down to the depths accessible by drilling with modern technology, i.e., over 10,000 m, the average geothermal gradient is about 2.5–3°C/100 m. For example, if the temperature within the first few meters below ground level, which on average corresponds to the mean annual temperature of the external air, is 15°C, then we can reasonably assume that the temperature becomes about 65–75°C at 2000 m depth, 90–105°C at 3000 m, and so on for a further few thousand meters. There are, however, vast areas in which the geothermal gradient is far from the average value. In areas in which the deep rock basement has undergone rapid sinking, and the basin is filled with geologically "very young" sediments, the geothermal gradient may be lower than 1°C/100 m. On the other hand, in some "geothermal areas," the gradient is more than 10 times the average value (further details are given in Dickson and Fanelli[2]).

Geothermal energy essentially comes from the natural heat of the earth as a result of the decay of the naturally radioactive isotopes of uranium, thorium, and potassium. Because of the internal heat, the Earth's surface heat flow averages 82 mW/m^2 with an equivalent total heat loss rate (the mean terrestrial heat flow of continents and oceans is 65 and 101 mW/m^2) of about 42 million MW.[2,8,9] The estimated total thermal energy above mean surface temperature to a depth of 10 km is 1.3×10^{27} J, which is equivalent to burning 3.0×10^{17} barrels of oil. Since the global energy consumptions for all types of energy are equivalent to use of about 100 million barrels of oil per day, the Earth's energy to a depth of 10 km could theoretically supply all of mankind's energy needs for about 6 million years (as further discussed in Wright[8] and Lund[9]).

A geothermal system essentially consists of three elements, namely, 1) a heat source; 2) a reservoir; and 3) a heat-carrying fluid (geothermal liquid) as illustrated in Figure 1. The heat source can be either at very high temperature greater than 600°C with magmatic intrusion at relatively shallow depths (5–10 km) or at relatively low temperatures. As mentioned above, the Earth's temperature increases with depth. The reservoir is a volume of hot permeable rocks from which the circulating fluids extract heat and connected to a surfacial recharge area through which the meteoric waters can replace or partly replace the fluids that escape from the reservoir through springs or are extracted by boreholes. The geothermal fluid is water (in the majority of cases, meteoric water) either in liquid or in vapor phase,

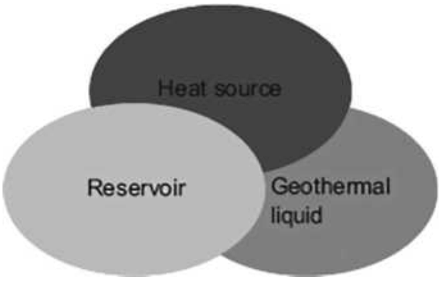


FIGURE 1 Geothermal system.

depending on its temperature and pressure. The geothermal water often carries with it some chemicals and gases such as CO_2 and H_2S.

Geothermal Energy and Environmental Impact

Problems with energy supply and use are related not only to global warming but also to such environmental concerns as air pollution to climate change, forest destruction, and emission of radioactive substances. These issues must be taken into consideration simultaneously if humanity is to achieve a bright energy future with minimal environmental impacts. Much evidence exists, which suggests that the future will be negatively impacted if humans keep degrading the environment.

One solution to both energy and environmental problems is to make much more use of renewable energy resources, particularly geothermal. This cause is sometimes espoused with a fervor that leads to extravagant and impossible claims being made. Engineering practicality, reliability, applicability, economy, scarcity of supply, and public acceptability should all be considered accordingly.

In order to achieve the energetic, economic, and environmental benefits that geothermal offers, the following integrated set of activities should be acted accordingly:

- *Research and Development:* Research and development priorities should be set in close consultation with industry to reflect their needs. Most research is conducted through cost-shared agreements and falls within the short-to- medium term. Partners in these activities should include a variety of stakeholders in the energy industry, such as private-sector firms, utilities across the country, provincial governments, and other federal departments.
- *Technology Assessment:* Appropriate technical data should be gathered in the laboratory and through field trials on factors such as cost benefit, reliability, environmental impact, safety, and opportunities for improvement. These data should also assist the preparation of technology status overviews and strategic plans for further research and development.
- *Standards Development:* The development of technical and safety standards is needed to encourage the acceptance of proven technologies in the marketplace. Standards development should be conducted in cooperation with national and international standards writing organizations, as well as other national and provincial regulatory bodies.
- *Technology Transfer:* Research and development results should be transferred through sponsorship of technical workshops, seminars, and conferences, as well as through the development of training manuals and design tools, Web tools, and the publication of technical reports.

Such activities will also encourage potential users to consider the benefits of adopting geothermal energy technologies. In support of developing near-term markets, a key technology transfer area is to accelerate the use of geothermal energy technologies, particularly for district heating applications and power generation.

Geothermal Energy and Sustainability

Sustainable development requires a sustainable supply of clean and affordable energy resources that do not cause negative societal impacts. Supplies of such energy resources as fossil fuels and uranium are finite. Energy resources such as geothermal are generally considered renewable and therefore sustainable over the relatively long term.

Sustainability often leads local and national authorities to incorporate environmental considerations into energy planning. The need to satisfy basic human needs and aspirations, combined with increasing world population, will make the need for successful implementation of sustainable development increasingly apparent. Various criteria that are essential to achieving sustainable development in a society follow:[10,11]

- Information about and public awareness of the benefits of sustainability investments
- Environmental education and training
- Appropriate energy and exergy strategies
- The availability of renewable energy sources and cleaner technologies
- A reasonable supply of financing
- Monitoring and evaluation tools

Geothermal as a renewable energy has an important role to play in meeting future energy needs in both rural and urban areas. The development and utilization of geothermal energy applications should be given a high priority, especially in the light of increased awareness of the adverse environmental impacts of fossil-based generation. The need for sustainable energy development is increasing rapidly in the world. Widespread use of geothermal energy is important for achieving sustainability in the energy sectors in both developing and industrialized countries.

Geothermal energy can be a key component for sustainable development for three main reasons:

1. They generally cause less environmental impact than other energy resources. The variety of geothermal energy applications provides a flexible array of options for their use.
2. They cannot be depleted. If used carefully in appropriate applications, renewable energy resources can provide a reliable and sustainable supply of energy almost indefinitely. In contrast, fossil fuel and uranium resources are diminished by extraction and consumption.
3. They favor both system centralization and decentralization and local solutions that are somewhat independent of the national network, thus enhancing the flexibility of the system and providing economic benefits to small isolated populations. Also, the small scale of the equipment often reduces the time required from initial design to operation, providing greater adaptability in responding to unpredictable growth and/or changes in energy demand.

Not all renewable energy resources are inherently clean in that they cause no burden on the environment in terms of waste emissions, resource extraction, or other environmental disruptions. Nevertheless, use of geothermal almost certainly can provide a cleaner and more sustainable energy system than increased controls on conventional energy systems.

To seize the opportunities, it is essential to establish a geothermal energy market and gradually build up the experience with the cutting-edge technologies. The barriers and constraints to the diffusion of geothermal energy use should be removed. The legal, administrative, and financing infrastructure should be established to facilitate planning and application of geothermal energy projects. Government could/should play a useful role in promoting geothermal energy technologies through fundings and incentives to encourage research and development as well as commercialization and implementation in both urban and rural areas.

Classification of Geothermal Resources

There is no standard international terminology in use throughout the geothermal community, which is unfortunate, as this would facilitate mutual comprehension.[2] In the following, some of the most common classifications in this discipline are presented in terms of energetic and exergetic aspects.

Classification of Geothermal Resources by Energy

The most common classification of geothermal resources is done based on the enthalpy values of the geothermal fluids, acting as the energy carrier for transporting the heat from deep hot rocks to the Earth's surface. In this regard, enthalpy and temperature levels are accordingly considered as given in Table 1. The temperatures of these geothermal resources range from the mean annual temperature of around 20°C to over 300°C. In general, resources above 150°C are used for electric power generation, although power has recently been generated at Chena Hot Springs Resort in Alaska using a 74°C geothermal resource.[12] Resources below 150°C are usually used in direct-use projects for individual and district heating and cooling applications. Furthermore, the ambient temperatures in the range of 5–30°C can be used with geothermal (ground source) heat pumps to provide both heating and cooling. The resources may also be categorized into three enthalpy levels (low, medium, and high) according to criteria that are generally based on the energy content of the fluids and their potential forms of utilization. Furthermore, Table 2 presents a summary of the types of geothermal resources, reservoirs temperatures, reservoirs fluids, applications, and systems.

Although numerous criteria can be utilized for classification, the reservoir temperature still remains as the primary criterion. Table 3 lists a possible scheme in which geothermal resources are classified into

TABLE 1 Geothermal Resource Types and Temperature Ranges

Resource Type		Temperature Range (°C)
Convective hydrothermal resources	Vapor dominated	≈240
	Hot water dominated	20–350
Other hydrothermal resources	Sedimentary basin	20–150
	Geopressured	90–200
	Radiogenic	30–150
Hot rock resources	Solidified (hot dry rock)	90–650
	Part still molten magma	>600

Source: Adapted from Lund[12] and Dickson and Fanelli.[13]

TABLE 2 Geothermal Resource Types, Applications, and System Types

Reservoir Temperature	Reservoir Fluid	Applications	Systems
High temperature (>220°C)	Water or steam	Power generation Direct use	• Flash steam • Combined (flash and binary) cycle • Direct fluid use • Heat exchangers • Heat pumps
Medium temperature (100–220°C)	Water	Power generation Direct use	• Binary cycle • Direct fluid use • Heat exchangers • Heat pumps
Low temperature (50–150°C)	Water	Direct use	• Direct fluid use • Heat exchangers • Heat pumps

Source: Adapted from World Bank, Geothermal Energy.[14]

TABLE 3 Classification of Different Schemes for Geothermal Resources

Class of Resource	Reservoir Temperature (°C)	Mobile Fluid Phase in Reservoir	Production Mechanism	Fluid State at Well Head	Well Productivity and Controlling Factors Other than Temperature	Applicable Power Conversion Technology	Unusual Development or Operational Problems
1. Non-electrical	<100	Liquid water	Artesian self-flowing wells; pumped wells	Liquid water	Well productivity dependent on reservoir flow capacity and static water level	Direct use	
2. Very low temperature	100 to <150	Liquid water	Pumped wells	Liquid water (for pumped wells); steam–water mixture (for self-flowing wells)	Typical well capacity 2 to 4 MW_e; dependent on reservoir flow capacity and gas content in water; well productivity often limited by pump capacity	Binary	
3. Low temperature	150 to <190	Liquid water	Pumped wells; self-flowing wells (only at the higher temperature end of the range)	Liquid water (for pumped wells); steam–water mixture (for self-flowing wells)	Typical well capacity 3 to 5 MW_e; dependent on reservoir pressures, reservoir flow capacity and gas content in water; productivity of pumped wells typically limited by pump capacity and pump parasitic power need; productivity of self-flowing wells strongly dependent on reservoir flow capacity	Binary; two-stage flash; hybrid	Calcite scaling in production wells and stibnite scaling in binary plant are occasional problems
4. Moderate temperature	190 to <230	Liquid water	Self-flowing wells	Steam–water mixture (enthalpy equal to that of saturated liquid at reservoir temperature)	Well productivity highly variable (3 to 12 MW_e); strongly dependent on reservoir flow capacity	Single-stage flash; two-stage flash; hybrid	Calcite scaling in production wells occasional problem; alumino-silicate scale in injection system a rare problem

(*Continued*)

TABLE 3 (*Continued*) Classification of Different Schemes for Geothermal Resources

Class of Resource	Reservoir Temperature (°C)	Mobile Fluid Phase in Reservoir	Production Mechanism	Fluid State at Well Head	Well Productivity and Controlling Factors Other than Temperature	Applicable Power Conversion Technology	Unusual Development or Operational Problems
5. High temperature	230 to <300	Liquid water; liquid-dominated two-phase	Self-flowing wells	Steam–water mixture (enthalpy equal to or higher than that of saturated liquid at reservoir temperature); saturated steam	Well productivity highly variable (up to 25 MW_e); dependent on reservoir flow capacity and steam saturation		
6. Ultrahigh temperature	300+	Liquid-dominated two-phase	Self-flowing wells	Steam–water mixture (enthalpy equal to or higher than that of saturated liquid at reservoir condition); saturated steam; superheated steam	Well productivity extremely variable (up to 50 MW_e); dependent on reservoir flow capacity and steam saturation		
7. Steam field	240 (33.5 bar-a pressure; 2800 kJ/kg enthalpy)	Steam	Self-flowing wells	Saturated or superheated steam	Well productivity extremely variable (up to 50 MW_e); dependent on reservoir flow capacity		

Source: Adapted from Sanyal.[15]

seven categories based on temperature, while they are explained in more detail in Sanyal.[15] This classification includes non-electrical grade (<100°C), very low temperature (100°C to <150°C), low temperature (150°C to 190°C), moderate temperature (190°C to <230°C), high temperature (230°C to <300°C), ultrahigh temperature (>300°C), and steam fields (approximately 240°C with steam as the only mobile phase).[16] In the first four classes of reservoirs, liquid water is the mobile phase; in the "high" and "ultrahigh" temperature reservoirs, the mobile fluid phase is either liquid or a liquid–vapor mixture. Thus, this makes steam fraction a second criterion. Unfortunately, there is no readily available procedure to classify enhanced geothermal systems or hot dry rock resources other than perhaps considering temperature-based classification as discussed above.

Classification of Geothermal Resources by Exergy

Although many agree on the classification of geothermal resources based on temperature and enthalpy levels as part of the energy aspects under the first law of thermodynamics, there are still some ambiguities with respect to their quality that address the second law of thermodynamics through exergy. Therefore, geothermal resources should be classified by their exergy, a measure of their ability to do work[17,18]).

Lee[17] proposed a new parameter, the so-called specific exergy index (SExI) for better classification and evaluation as follows:

$$\text{SExI} = \frac{h_{\text{brine}} - 273.16\ S_{\text{brine}}}{1192} \tag{1}$$

which is a straight line on an $h\!-\!s$ plot of the Mollier diagram. Straight lines of SExI = 0.5 and SExI = 0.05 can therefore be drawn in this diagram and used as a map for classifying geothermal resources by taking into account the following criteria:

- SExI < 0.05 for low-quality geothermal resources
- 0.05 ≤ SExI < 0.5 for medium-quality geothermal resources
- SExI ≥ 0.5 for high-quality geothermal resources

Here, the demarcation limits for these indexes are exergies of saturated water and dry saturated steam at 1 bar absolute.

In order to map any geothermal field on the Mollier diagram as well as to determine the energy and exergy values of the geothermal brine, the average values for the enthalpy and entropy are then calculated from the following equations:[18]

$$h_{\text{brine}} = \frac{\sum_{i}^{n} \dot{m}_{w,i} h_{w,i}}{\sum_{i}^{n} \dot{m}_{w,i}} \tag{2}$$

$$S_{\text{brine}} = \frac{\sum_{i}^{n} \dot{m}_{w,i} S_{w,i}}{\sum_{i}^{n} \dot{m}_{w,i}} \tag{3}$$

The above-given exergetic classification has been used by some investigators in their studies. From these, Quijano[19] performed an exergy analysis of the Ahuachapán and Berlin geothermal fields. After plotting the thermodynamic states of both fields on the Mollier diagram and calculating the values of SExI, he classified these geothermal fields as the medium high exergy zone. Ozgener et al.[20] and Baba et al.[21] utilized this index (SExI) for the Balcova geothermal field located in the western part of Turkey, endowed with considerably rich geothermal resources. The number of the operating wells in this field varies depending on the heating days and operating strategies. Taking into account the eight operating wells and using Eqs. 1–3, the SExI is found to be 0.07 since it falls into the medium-quality geothermal resources according to the classification of Lee.[17] However, given the classification of the Balcova geothermal field with respect to its average reservoir as 113.9°C, the Balcova geothermal field becomes an intermediate-enthalpy resource. This clearly indicates that the classification with respect to the SExI is more meaningful as there is no need for a temperature range classification. Ozgener et al.[22,23] also determined the SExI values for the Gonen and Salihli geothermal fields in Turkey as 0.025 and 0.049, respectively. They concluded that both these fields fall into the low-quality geothermal resources according to the classification of Lee. Etemoglu and Can[24] studied geothermal resources in Turkey based on SExI. In their analysis, the triple point of water was selected as the dead-state conditions since the enthalpy and entropy values were zero. Geothermal resources with SExI values of 0.56 were classified as high-exergy resources; those whose values were <0.05 were low-exergy resources, and those whose values are between these two were medium-exergy resources.

The Lindal Diagram

There is another way of classifying the geothermal fields developed by Lindal[25] as illustrated in Figure 2. This diagram has frequently appeared in books, reports, and articles on direct uses of geothermal energy and shows the temperature required for both current and potential applications of geothermal energy. In other words, the minimum production temperatures in a geothermal field are generally required for the different types of use as shown in the diagram. This is now widely used by the geothermal community to depict temperature as the yardstick of direct applications. The boundaries, however, serve only as guidelines. Conventional electric power production is limited to fluid temperatures about 150°C, but considerably lower temperatures can be used with the application of binary fluids (outlet temperatures commonly at 100°C, while there are also some applications up to 74°C). The ideal inlet temperature into houses for space heating using radiators is about 80°C, but by using radiators of floor heating, or by applying heat pumps or auxiliary boilers, thermal waters with temperatures only a few degrees above the ambient can be used beneficially. It is a common misconception that direct use of geothermal is confined to low-temperature resources. High-temperature resources can, of course, also be used for heating and drying purposes even if the process is at a very low temperature. Refrigeration is, in fact, only possible with temperatures above about 120°C.[26]

Some Energetic and Exergetic Parameters

Coskun et al.[27] have introduced some geothermal energy systems-related parameters, namely, energetic renewability ratio, exergetic renewability ratio, energetic reinjection ratio, and exergetic reinjection ratio:

- *Energetic renewability ratio:* This is defined as the ratio of useful renewable energy obtained from the system to the total energy input (all renewable and non-renewable together) to the system.

$$R_{\mathrm{Ren}_E} = \frac{\dot{E}_{\mathrm{usf}}}{\dot{E}_{\mathrm{total}}} \tag{4}$$

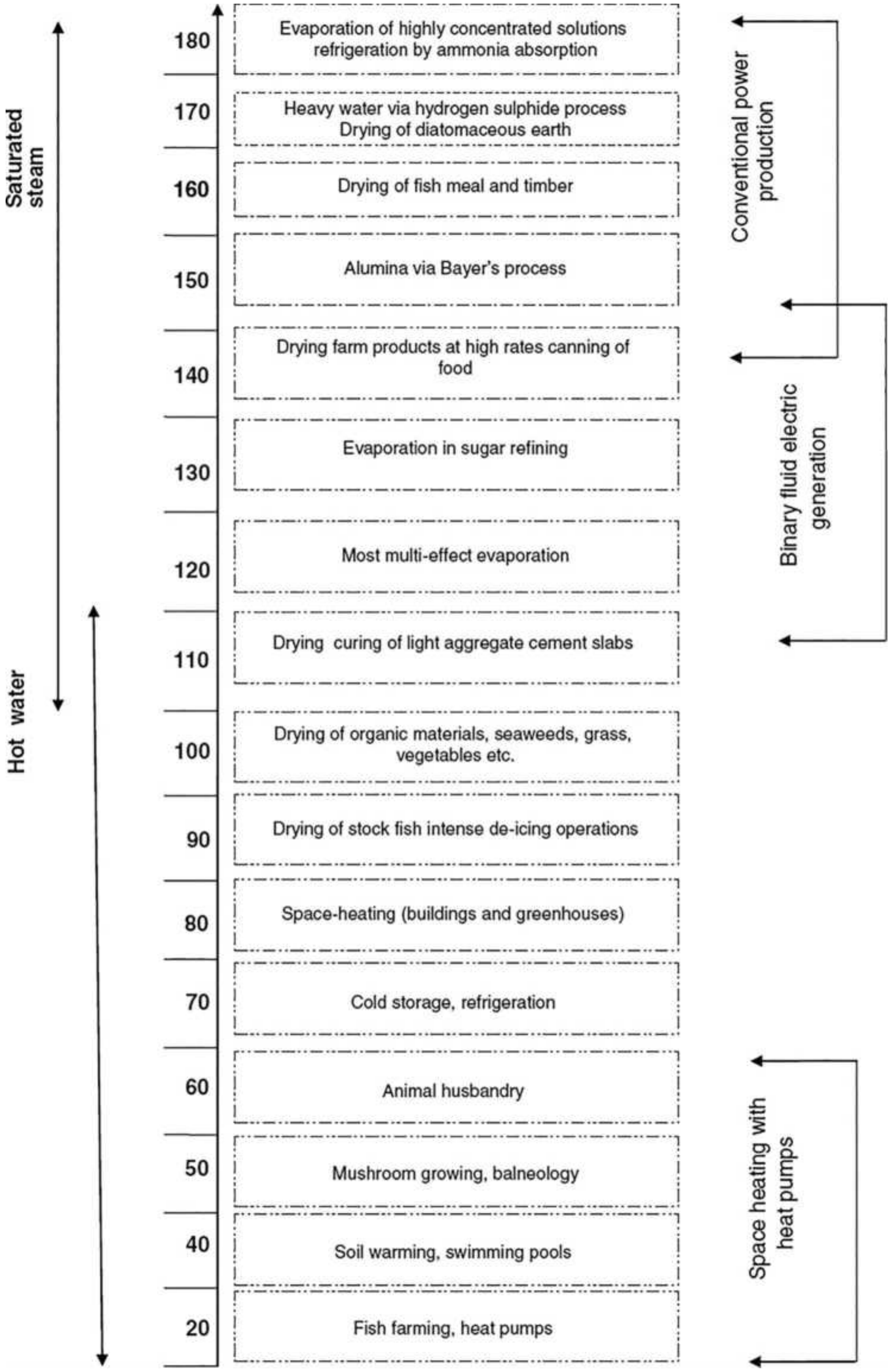


FIGURE 2 Lindal diagram indicating the applications for geothermal resources depending on the temperature. **Source:** Adapted from Hepbasli.[28]

- *Exergetic renewability ratio:* This is defined as the ratio of useful renewable exergy obtained from the system to the total exergy input (all renewable and non-renewable together) to the system:

$$R_{\mathrm{Re\ in}_{\mathrm{Ex}}} = \frac{\dot{E}x_{\mathrm{usf}}}{\dot{E}x_{\mathrm{total}}} \tag{5}$$

Here, total energy and exergy input terms include the wellhead geothermal water energy only in energy and exergy efficiencies. However, in the above equations for energetic and exergetic renewability ratios, all renewable and non-renewable energy and exergy inputs are considered for input.

- *Energetic reinjection ratio:* This is defined as the ratio of renewable energy discharged to environment or reinjected to the well from the system to the total geothermal energy supplied to the system:

$$R_{\mathrm{Re\ in}_{\mathrm{E}}} = \frac{\dot{E}_{\mathrm{nd}}}{\dot{E}_{\mathrm{gw}}} \tag{6}$$

- *Exergetic reinjection ratio:* This is defined as the ratio of renewable exergy discharged to environment or reinjected to the well from the system to the total geothermal exergy supplied to the system:

$$R_{\mathrm{Re\ in}_{\mathrm{Ex}}} = \frac{\dot{E}x_{\mathrm{nd}}}{\dot{E}x_{\mathrm{gw}}} \tag{7}$$

Here, the monthly variations of the above-listed parameters as studied by Coskun et al.[27] for the Edremit Geothermal District Heating System (GDHS) are illustrated in Figure 3. As seen in Figure 3a and b, both energetic and exergetic renewability ratios decrease with increasing temperature in the summer season. Here, one can extract the following:

- The energetic renewability ratio varies between 0.32 and 0.33 for the Edremit system during summer when heating is not required.
- The exergetic renewability ratio varies between 0.25 and 0.35 for the Edremit system during the summer when heating is not required.

It was observed that the outdoor reference ambient temperatures do not have a direct effect on both mass flow rate and energy/exergy losses as a result of no heating requirement for the summer season. Domestic hot water requirement is one important parameter for the summer season. Domestic hot water requirement increases for hot days. Because of the increase in domestic hot water requirement, pump energy consumption in total energy input increases. Moreover, calculations show that total exergy loss/destruction tended to increase in the summer season. For instant, pump exergy destruction increases in the summer season as a result of the increase in the required domestic mass flow rate. In the heating season, total "winter" heat demand is the sum of the domestic hot water+heating proper. The mass flow rates for the network water and geothermal water vary during the season dependent on the average useful energy demand. The demand is the highest in January, declines until May, and begins increasing until the end of December. It is obvious that based on the demand, the mass flow rates proportionally change. The energetic and exergetic renewability ratios become the lowest in January, increasing until May, and begin decreasing until the end of December. We have understood from these results that heat demand and renewability ratio values are opposite characteristic properties. The renewability values are given for the heating season as below:

- The energetic renewability ratio varies between 0.33 and 0.34 for the Edremit system.
- The exergetic renewability ratio varies between 0.53 and 0.63 for the Edremit system.

In addition, both energetic and exergetic reinjection ratios are calculated using Eqs. 6 and 7 for the Edremit GDHS annually as given in Figure 3c and d, respectively. These figures can be explained for summer and winter seasons as follows. The energetic reinjection ratio increases in higher temperatures with the effect of pump energy consumption. They vary between 0.59 and 0.62 for the Edremit system during the summer season when heating is not required (Figure 3c). Variation of the exergetic reinjection ratio depends on outdoor temperature and has similar characteristic with average annual temperature (see Figure 3d). Used geothermal water exergy rate decreases as a result of increase in reference ambient temperature in the summer season. Used geothermal exergy rate decreases through the middle of summer, at the highest temperature. The exergetic reinjection ratio varies between 0.10 and 0.20 for the Edremit system during the summer season (see Figure 3d).

Furthermore, Hepbasli[28] has recently proposed the following new exergetic parameters, based on the system given in Ozgener et al.[29] while their energetic approaches may be also considered using their exergetic relations:

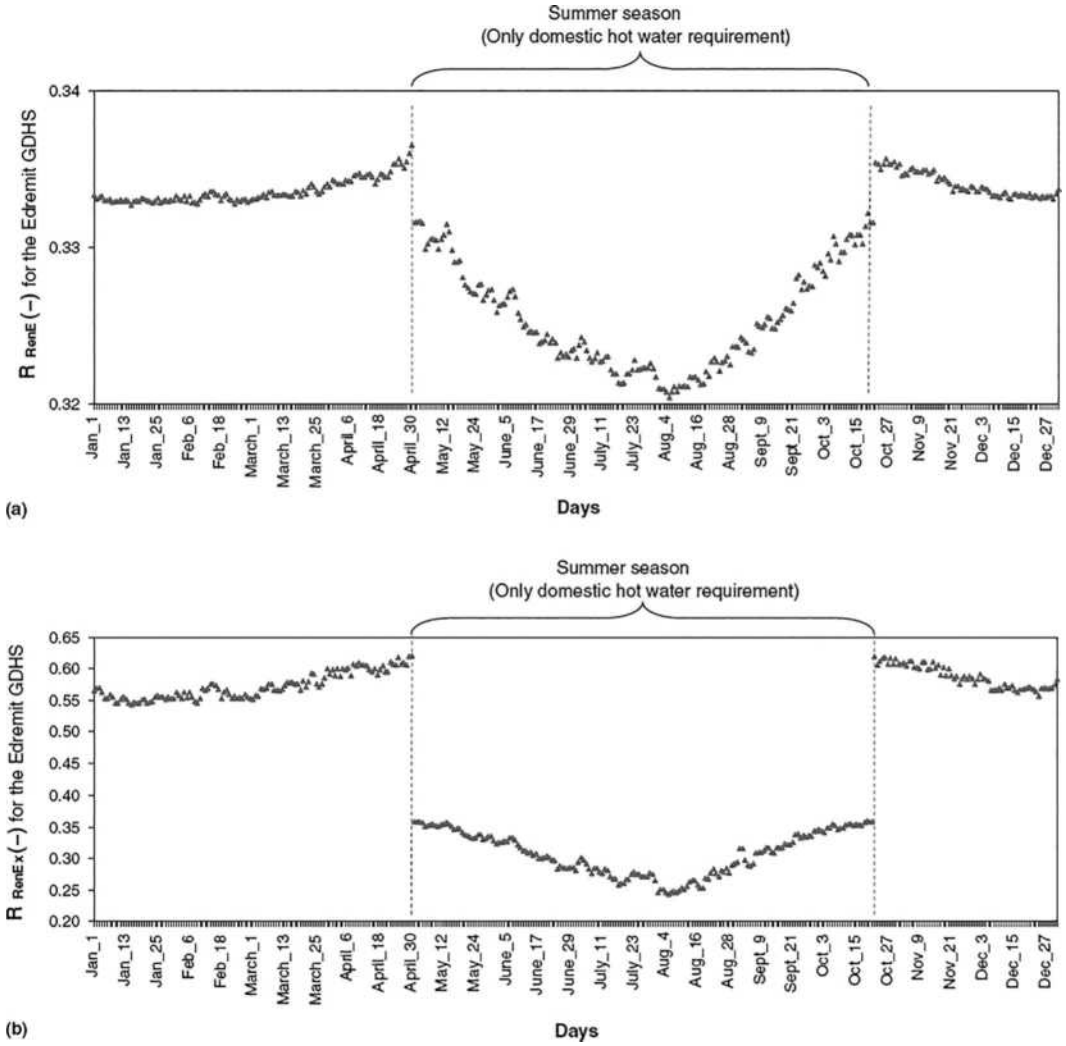


FIGURE 3 (a) Change of energetic renewability ratio for the Edremit GDHS annually. (b) Change of exergetic renewability ratio for the Edremit GDHS annually (modified from Coskun et al.[27]).

(*Continued*)

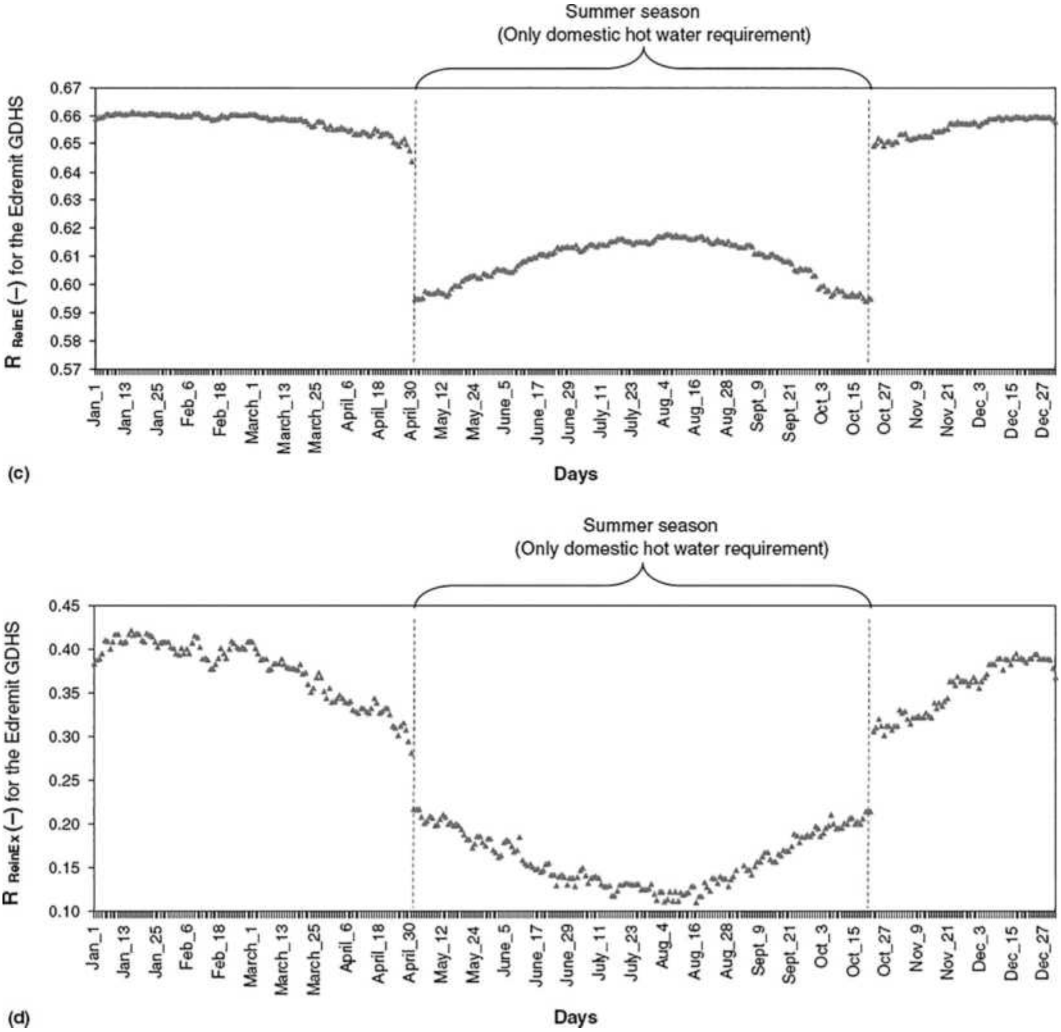


FIGURE 3 (CONTINUED) (c) Change of energetic reinjection ratio for Edremit GDHS annually. (d) Change of exergetic reinjection ratio for Edremit GDHS annually (modified from Coskun et al.[27])**S**

1. *Distribution cycle exergetic ratio* (R_{DCEx}): This is defined as the ratio of the exergetic capacity of the district heating water circulating through the district heating distribution network (MW) to the exergetic capacity of the geothermal reservoir in the field considered (MW):

$$R_{\mathrm{DCEx}} = \frac{\dot{E}x_{\mathrm{dhw}}}{\dot{E}x_{\mathrm{gr}}} \tag{8}$$

2. *Energy consumptions circuit exergetic ratio* (R_{ECCEx}): This is defined as the ratio of the exergetic capacity of the fluid circulating through the energy consumptions circuit (building substations) district heating distribution network (MW) to the exergetic capacity of the geothermal reservoir in the field considered (MW):

$$R_{\mathrm{ECCEx}} = \frac{\dot{E}x_{\mathrm{dhw}}}{\dot{E}x_{\mathrm{gr}}} \tag{9}$$

The average values for the enthalpy and entropy to be used in the calculations may be calculated from Eqs. 2 and 3.

3. *Reservoir specific exergy utilization index* (RSExUI): This is defined as the ratio of the exergetic capacity of the geothermal reservoir in the field considered (MJ) to the total heating surface area of all the buildings (m^2) and given by

$$\mathrm{RSExUI}\left(\mathrm{MJ}/m^2\right) = \frac{\dot{E}x_{gr}\,(\mathrm{MJ})}{A_{\mathrm{building}}\left(m^2\right)} \tag{10}$$

4. *Geothermal brine specific exergy utilization index* (GBSExUI): This is defined as the ratio of the geothermal brine exergy input value (MJ) to the total heating surface area of all the buildings (m^2) and given by

$$\mathrm{GBSxUI}\left(\mathrm{MJ}/m^2\right) = \frac{\dot{E}x_{gb}\,(\mathrm{MJ})}{A_{\mathrm{building}}\left(m^2\right)} \tag{11}$$

Some more details on the use of exergy for design and analysis of geothermal district heating systems are available elsewhere.[30]

Performance Evaluation

From the thermodynamics point of view, exergy is defined as the maximum amount of work that can be produced by a system or a flow of matter or energy as it comes to equilibrium with a reference environment. Exergy is a measure of the potential of the system or flow to cause change, as a consequence of not being completely in stable equilibrium relative to the reference environment. Unlike energy, exergy is not subject to a conservation law (except for ideal, or reversible, processes). Rather, exergy is consumed or destroyed, due to irreversibilities in any real process. The exergy consumption during a process is proportional to the entropy created due to irreversibilities associated with the process.

Exergy analysis is a technique that uses the conservation of mass and conservation of energy principles together with the second law of thermodynamics for the analysis, design, and improvement of geothermal energy systems as well as others. It is also useful for improving the efficiency of energy-resource use, for it quantifies the locations, types, and magnitudes of wastes and losses. In general, more meaningful efficiencies are evaluated with exergy analysis rather than energy analysis, since exergy efficiencies are always a measure of how nearly the efficiency of a process approaches the ideal. Therefore, exergy analysis identifies accurately the margin available to design more efficient energy systems by reducing inefficiencies. We can suggest that thermodynamic performance is best evaluated using exergy analysis because it provides more insights and is more useful in efficiency-improvement efforts than energy analysis. For exergy analysis, the characteristics of a reference environment must be specified. This is commonly done by specifying the temperature, pressure, and chemical composition of the reference environment. The results of exergy analyses, consequently, are relative to the specified reference environment, which, in most applications, is modeled after the actual local environment. The exergy of a system is zero when it is in equilibrium with the reference environment.

Energy and exergy balances for an unsteady-flow process in a system during a finite time interval can be written as:

$$\text{Energy input} - \text{Energy output} = \text{Energy accumulation} \tag{12}$$

$$\text{Exergy input} - \text{Exergy output} - \text{Exergy consumption} = \text{Energy accumulation} \tag{13}$$

For a general steady-state, steady-flow process, it is assumed that the accumulation terms in the above equations become zero. Therefore, input terms become equal to output terms.

For a steady-state, steady-flow process, the mass balance equation can be expressed as

$$\sum_{\dot{m}_{\text{in}}} = \sum_{\dot{m}_{\text{out}}} \tag{14}$$

For a general steady-state, steady-flow process, the general energy and exergy balance equations can also be written more explicitly as

$$\dot{Q} + \sum_{\dot{m}_{\text{in}}} h_{\text{in}} = \dot{W} + \sum_{\dot{m}_{\text{out}}} h_{\text{out}} \tag{15}$$

And

$$\sum \left(1 - \frac{T_0}{T_s}\right)\dot{Q}_s + \sum \dot{m}_{\text{in}}\left(h_{\text{in}} - T_0 s_{\text{in}}\right) - \dot{E}x_d \tag{16}$$

$$= \dot{W} + \sum m_{\text{out}}\left(h_{\text{out}} - T_0 s_{\text{out}}\right)$$

We can generally define the energy and exergy efficiencies as follows:

$$\eta = \frac{\dot{E}_{\text{net}}}{\dot{E}_{\text{tot}}} \tag{17}$$

And

$$\psi = \frac{\dot{E}x_{\text{net}}}{\dot{E}x_{\text{tot}}} \tag{18}$$

For further details of both energy and exergy analyses and applications to various energy systems, see Dincer and Rosen.[31]

Note that in the calculations, the temperature T_0 and pressure P_0 of the environment are often taken as standard-state values, such as 1 atm and 25°C. However, these properties may be specified differently depending on the application. T_0 and P_0 might be taken as the average ambient temperature and pressure, respectively, for the location at which the system under consideration operates. Or, if the system uses atmospheric air, T_0 might be specified as the average air temperature. If both air and water from the natural surroundings are used, T_0 would be specified as the lower of the average temperatures for air and water.

Detailed energy and exergy as well as exergoeconomic analysis methodologies for geothermal energy systems are given elsewhere.[22,23]

Case Study

Here, we present a case study analysis of the Salihli Geothermal District Heating System (SGDHS), which has a maximal yield of 87 L/s at an average reservoir temperature of 95°C, with a minimal capacity of 838 MW.[30] The SGDHS was originally designed for 20,000 residences equivalence. Of these, 2400 residences equivalence are heated by geothermal energy as of February 2004. The outdoor and indoor

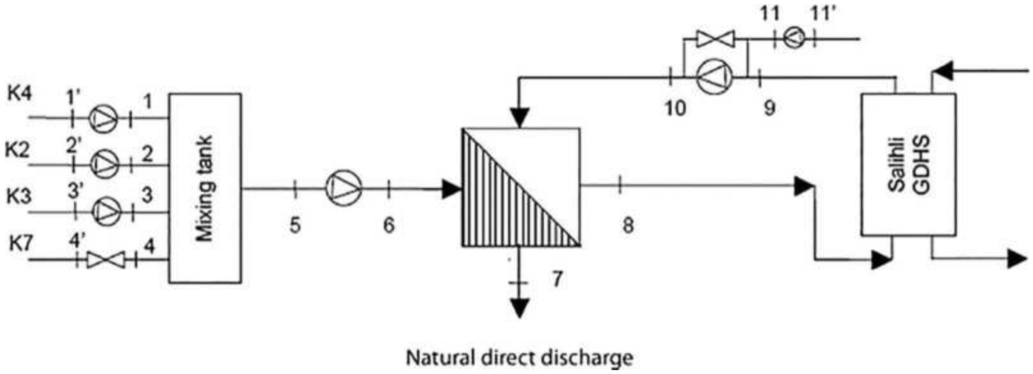


FIGURE 4 A schematic of the SGDHS.
Source: Adapted from Dincer et al.[30]

design temperatures for the system are 4°C and 22°C, respectively. Figure 4 illustrates a schematic of the SGDHS where two hospitals and official buildings heated by geothermal energy were also included. The SGDHS consists mainly of three cycles, such as 1) energy production cycle (geothermal well loop and geothermal heating center loop); 2) energy distribution cycle (district heating distribution network); and 3) energy consumption cycle (building substations). At the beginning of 2004, there were seven wells ranging in depth from 70 to 262 m in the SGDHS. Of these, five wells were in operation at the date studied and two wells (K5 and K6) were out of operation. Four wells (designated as K2, K3, K4, and K7) and one well (K1) are production and balneology wells, respectively. The well head temperatures of the production wells vary from 56°C to 115°C, while the volumetric flow rates of the wells range from 2 to 20 L/s. The geothermal fluid is basically sent to two primary plate-type heat exchangers and is cooled to about 45°C, as its heat is transferred to secondary fluid. The geothermal fluid[30] is discharged via natural direct discharge, with no recharge to Salihli geothermal field production, but reinjection studies are expected to be completed in the near future. The temperatures obtained during the operation of the SGDHS are, on average, 98°C/45°C for the district heating distribution network and 62°C/42°C for the building circuit. By using the control valves for flow rate and temperature at the building main station, the needed amount of water is sent to each housing unit and the heat balance of the system is achieved. Geothermal fluid, collected from the four production wells at an average well heat temperature of 95.5°C, is pumped to the inlet of the heat exchanger mixing tank later through a main collector (from four production wells) with a total mass flow rate of about 47.62 kg/s. Geothermal fluid of intermingling molecules of different species through molecular diffusion was neglected in this study. As a result, not only irreversibility of the mixing tank was assumed equal to zero but also heat losses from the tank and main collector pipeline through the mixing process were neglected. In the SGDHS investigated, the thermal data on pressures, temperatures, and flow rates of the fluids were measured using a number of devices and employed in the models.[22,23]

Results and Discussion

The thermodynamic data used in the case study are based on the actual data taken from the SGDHS on February 1, 2004. The number of the wells in operation in the Salihli geothermal field may normally vary depending on the heating days as well as operating strategy. Taking into account the four productive wells when this study was conducted, the specific exergy index (SExI) is found to be 0.049, which is very close to the limit of the medium-quality geothermal resources. This shows

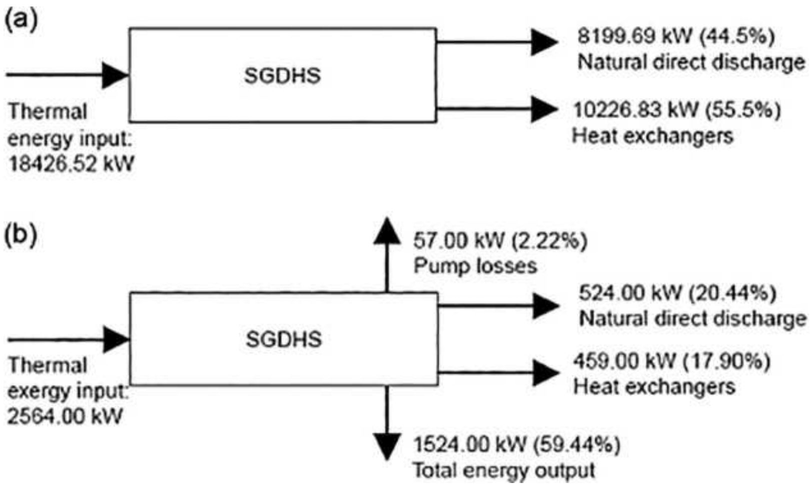


FIGURE 5 System flow diagrams for (a) energy and (b) exergy.
(**Source:** Adapted from Dincer et al.[30])

that the Salihli geothermal field falls into the low-quality geothermal resources according to the classification of Lee.[18]

The total mass flow rates of the geothermal fluid at the inlet of the heat exchanger (the total mass flow rate of the production wells) were measured to be 47.62 kg/s at an average temperature of 44.2°C on February 1, 2004. If hot water losses measured in the district heating distribution network are above 5 m^3/h, the water is generally added via a pump (using pressurized water tanks) to the network in order to cover the leaks taking place. However, the hot water losses were neglected in this study as these losses were below 5 m^3/h on the day of the study.

We present our calculation results, based on the actual data taken, given elsewhere.[30] These include all energetic and exergetic as well as other thermodynamic parameters along with the efficiencies. The energy and exergy efficiencies of the SGDHS are determined to be 55.5% and 59.4%, respectively. Here, the exergy efficiency is higher due to the recirculation and heat recovery processes.

We also present our energy and exergy flow diagrams in Figure 5. As illustrated in Figure 5a, the natural direct discharge of the system accounts for 44.5% of the total energy input. An investigation of the exergy flow diagram given in Figure 5b shows that 40.56% (accounting for about 1040 kW) of the exergy entering the system is lost, while the remaining 59.44% is utilized. The highest exergy loss of 20.44% occurs from the natural direct discharge in this study. The second largest exergy destruction occurs from the heat exchanger with 17.90% (about 459 kW) of the total exergy input. This is followed by the total exergy destruction associated with the pumps amounting to some 57 kW, which accounts for 2.22% of the exergy input to the system.

Conclusions

This entry has presented some historical background on geothermal energy resources and geothermal energy applications and classified geothermal resources by energy and exergy in detail along with some practical cases studies. It has discussed energetic, environmental, and sustainability aspects of geothermal energy and systems and proposed some energetic and exergetic parameters. Also, a procedure for performance assessment of geothermal energy systems is presented. A case study on how both energy and exergy analyses are necessary for better performance of the systems is likewise presented.

Nomenclature

$\dot{E}$	energy rate (kW)
$\dot{E}x$	exergy rate (kW)
$\dot{F}$	exergy rate of the fuel (kW)
h	specific enthalpy (kJ/kg)
IP	improvement potential rate (kW)
$\dot{m}$	mass flow rate (kg/s)
$\dot{P}$	product exergy rate (kW)
$\dot{Q}$	heat transfer rate (kW)
S	specific entropy (kJ/kg/K)
SExI	specific exergy index (dimensionless)
T	temperature (°C or K)
$\dot{W}$	work rate, power (kW)

Greek letters

β	exergetic factor (%)
η	energy efficiency (%)
ψ	exergy efficiency (%)
δ	fuel depletion rate (%)
ξ	productivity lack (%)
χ	relative irreversibility (%)

Subscripts

0	reference environment, dead state
d	natural direct discharge; destruction
HE	heat exchanger
in	inlet
s	source
out	outlet
tot	total

References

1. Fridleifsson, I.B. Status of geothermal energy amongst the world's energy sources. Geothermics **2003**, *32,* 379–388.
2. Dickson, M.H.; Fanelli, M. *What is geothermal energy?* Prepared on February 2004. Available at http://www. geothermal-energy.org/314,what_is_geothermal_energy.html (accessed June 2010).
3. Dickson, M.H.; Fanelli, M., Eds.; *Geothermal Energy: Utilization and Technology;* UNESCO Renewable Energy Series, UNESCO Publishing: Paris, 2003; 205 pp.
4. Cataldi, R.; Hodgson, S.F.; Lund, J.W., Eds.; *Stories from a Heated Earth—Our Geothermal Heritage;* Geothermal Resources Council and International Geothermal Association: Davis, CA, 1999; 569 pp.
5. Lund, J.W.; Freeston, D.H.; Boyd, T.L. Direct utilization of geothermal energy 2010 worldwide review. Geothermal Congress 2010, Bali, Indonesia, 25–29 April, 2010; 23 pp.

6. Balta, M.T.; Dincer, I.; Hepbasli, A. Geothermal-based hydrogen production using thermochemical and hybrid cycles: A review and analysis. Int. J. Energy Res., **2010**, *34,* 757–775.
7. Balta, M.T.; Dincer, I.; Hepbasli, A. Thermodynamic assessment of geothermal energy use in hydrogen production. Int. J. Hydrogen Energy, **2009**, *34* (7), 2925–2939.
8. Wright, M. Nature of geothermal resources. In *Geothermal Direct-Use Engineering and Design Guidebook;* Lund, J.W., Ed.; Geo-Heat Center: Klamath Falls, OR, 1998; 27–69.
9. Lund, J.W.; Bjelm, L.; Bloomquist, G.; Mortensen, A.K. Characteristics, development and utilization of geothermal resources—A Nordic perspective. Episodes **2008**, *31* (1), 140–147.
10. Dincer, I. Renewable energy and sustainable development: A crucial review. Renewable Sustainable Energy Rev. **2000**, *4* (2), 157–175.
11. Dincer, I.; Rosen, M.A. Thermodynamic aspects of renewables and sustainable development. Renewable Sustainable Energy Rev. **2005**, *9* (2), 169–189.
12. Lund, J.W. Chena Hot Springs. Geo-Heat Center Quarterly Bulletin **2006**, *27* (3) (September), Klamath Falls, OR, 2–4.
13. Dickson, M.H.; Fanelli, M. Geothermal energy and its utilization. In *Small Geothermal Resources*; Dickson, M.H., Fanelli, M., Eds.; UNITAR/UNDP Centre for Small Energy Resources: Rome, Italy, 1990; 1–29.
14. World Bank, Geothermal Energy, available at http://www.worldbank.org/html/fpd/energy/geothermal/technology.htm (accessed October 23, 2005).
15. Sanyal, S.K. Classification of geothermal systems—A possible scheme. Proceedings of Thirtieth Workshop on Geothermal Reservoir Engineering, SGP-TR-176, Stanford University: Stanford, California, January 31–February 2, 2005, 8 pp.
16. Sanyal, S.K.; Butler, S.J. National Assessment of U.S. Enhanced Geothermal Resources Base—A Perspective, Trans.; Geothermal Resources Council: Palm Springs, California, August-September, 2004; Vol. 28, 233–238.
17. Lee, K.C. Classification of geothermal resources by exergy. Geothermics **2001**, *30,* 431–442.
18. Lee, K.C. Classification of geothermal resources: An engineering approach. Proceedings, Twenty-First Workshop on Geothermal Reservoir Engineering; Stanford University, 1996; 5 pp.
19. Quijano J. Exergy analysis for the Ahuachapan and Berlin Geothermal fields, El Salvador. Proceedings World Geothermal Congress, May 28–June 10; Kyushu-Tohoku, Japan, 2000.
20. Ozgener, L.; Hepbasli, A.; Dincer, I. Thermo-mechanical exergy analysis of Balcova geothermal district heating system in Izmir, Turkey. ASME J. Energy Resour. Technol. **2004**, *126,* 293–301.
21. Baba, A.; Ozgener, L.; Hepbasli, A. Environmental and exergetic aspects of geothermal energy. Energy Sources **2006**, *28* (7), 597–609.
22. Ozgener, L.; Hepbasli, A.; Dincer, I. Energy and exergy analysis of the Gonen geothermal district heating system, Turkey. Geothermics **2005**, *34,* 632–645.
23. Ozgener, L.; Hepbasli, A.; Dincer, I. Exergy analysis of Salihli geothermal district heating system in Manisa, Turkey. Int. J. Energy Res. **2005**, *29* (5), 398–408.
24. Etemoglu, A.B.; Can, M. Classification of geothermal resources in Turkey by exergy analysis. Renewable Sustainable Energy Rev. **2007**, *11*, 1596–1606.
25. Lindal, B. Industrial and other applications of geothermal energy. In *Geothermal Energy*; Armstead, H.C.H., Ed.; Earth Science 12; UNESCO: Paris, 1973; 135–148.
26. Gudmundsson, J.S.; Freeston, D.H.; Lienau, P.J. The Lindal diagram. Trans. Geotherm. Resour. Counc. **1985**, *9*, Part 1, August.
27. Coskun, C.; Oktay, Z.; Dincer, I. New energy and exergy parameters for geothermal district heating systems. Appl. Therm. Eng. **2009**, *29* (11–12), 2235–2242.
28. Hepbasli, A. A review on energetic, exergetic and exergoeconomic aspects of geothermal district heating systems (GDHSs). Energy Convers. Manage. **2010**, *51* (10), 2041–2061.

29. Ozgener, L.; Hepbasli, A.; Dincer, I. A key review on performance improvement aspects of geothermal district heating systems and applications. Renewable Sustainable Energy Rev. **2007**, *11* (8), 1675–1697.
30. Dincer, I.; Hepbasli, A.; Ozgener, L. Geothermal energy resources. In *Encyclopedia of Energy Engineering and Technology;* Capehart, B.L., Ed.; CRC Press, Taylor and Francis Group, 2007; Vol. 1, 744–752.
31. Dincer, I.; Rosen, M.A., Exergy, Elsevier: New York, 2007.

14

Green Energy

Ibrahim Dincer
and Adnan Midili

Introduction

Green energy can be considered a catalyst for energy security, sustainable development, and social, technological, industrial, and governmental innovations in a country. An increase in the green energy consumption of a country provides a positive impact on the economic as well as the social development of the country. Moreover, the supply and utilization of low-priced green fuel are particularly significant for global stability and peace because energy plays a vital role in industrial and technological developments around the world.

Critical energy issues in the 21st century will likely include energy security for almost 7 billion people, the expected global population by the middle of the 21st century, and global warming, mainly caused by CO_{2s} emissions generated from the combustion of fossil fuels.[1,2] Fossil fuels have caused some major problems for human health and human welfare due to their extensive use in various industrial and nonindustrial sectors. In reality, the main source of these problems is the extensive use of fossil-based technologies and strategies used by humans to govern throughout the centuries. In recent decades, the world has struggled with shortages of fossil fuels, pollution, and increased global energy requirements due to fast population growth, fast technological developments, and higher living standards. These factors have led to world population transition, migration, hunger, environmental (especially air and water pollution) problems, deteriorating health and disease, terrorism, energy and natural resources concerns, and wars. Also, problems with energy supply and use are related not only to global unrest, but also to such environmental concerns as air pollution, acid precipitation, ozone depletion, forest destruction, and the emission of radioactive substances. These issues must be taken into consideration simultaneously if humanity is to achieve a bright energy future with minimal environmental impact.

Other environmental considerations have been given increasing attention by energy industries and the public. The concept that consumers share responsibility for pollution and its cost has been increasingly accepted. In some jurisdictions, the prices of many energy resources have increased over the last one to two decades, in part to account for environmental costs. Global demand for energy services is expected to increase by as much as an order of magnitude by 2050, while primary-energy demands are expected to increase by 1.5 ± 0.3 times.[2] Accordingly, humans have reached energy shortage and pollution levels that are not tolerable anymore. The urgent need in this regard is to develop green energy–based permanent solutions for a sustainable future without any negative environmental and societal impacts. As a consequence, investigations for green energy should be encouraged particularly for green energy–based sustainability and future world stability.

In this book contribution, the authors present key information on green energy–based sustainability and global stability in accordance with major considerations that are presented in the following subtitles. In addition, this presents some key parameters like the green energy impact ratio and the green energy–based sustainability ratio. Moreover, anticipated patterns of future energy use and consequent environmental impacts (focusing on acid precipitation, stratospheric ozone depletion, and the greenhouse effect) are comprehensively discussed. Also, potential solutions to current environmental problems are identified along with green energy technologies. The relationships between green energy and sustainable development are described. Throughout the entry, several issues relating to green energy, the environment, and sustainable development are examined from both current and future perspectives. The specific objectives of this entry can be enumerated as follows:

- To help understand the main concepts and issues about green energy use and sustainability aspects
- To develop relationships between green energy use and sustainability development
- To encourage the strategic use and conservation of green energy sources
- To provide methods for energy security, implementation, and development
- To increase motivation on the implementation of green energy strategies for better energy supply
- To provide solutions to reduce negative environmental impacts by considering the possible green energy strategies
- To discuss possible green energy strategies for sectoral use

Green Energy

The most important property of green energy sources is environmental compatibility. In line with this characteristic, green energy sources will likely become the most attractive energy sources in the near future and will be the most promising energy sources for the technological and environmental perspectives of the 21st century, particularly in the context of sustainable development for the future. Green energy is one of the main factors that must be considered in discussions of sustainable development and future world stability. Several definitions of sustainable development have been put forth, including the following common one: "development that meets the needs of the present without compromising the ability of future generations to meet their own needs."[3,4] There are many factors that can contribute to achieving sustainable development. One of them is the requirement for a supply of energy resources that is fully sustainable. A secure supply of green energy resources is generally agreed to be a necessary but not sufficient requirement for development within a society.

Furthermore, sustainable development within a society demands a sustainable supply of green energy sources and an effective and efficient utilization of green energy technologies. Green energy can be defined as an energy source that has zero or minimum environmental impact, is more environmentally benign and more sustainable, and is produced from solar, hydro, biomass, wind, and geothermal, etc., energies. These types of green energy reduce the negative effects of fossil energy resources and the overall emissions from electricity generation and decrease greenhouse gases. They give an opportunity to take an active role in improving the environment and meet clean energy demand for both industrial and nonindustrial applications. Considering the benefits of green energy, it can be said that the sustainability of green energy supply and progress is assumed to be a key element in the interactions between nature and society. In addition, sustainable development requires a supply of energy resources that is sustainably available at reasonable cost and causes no or minimal negative societal impacts. Clearly, energy resources such as fossil fuels are finite and thus lack the characteristics needed for sustainability while others such as green energy sources are sustainable over the relatively long term.[5] Particularly, low-priced green energy is the most essential means for increasing the sustainable technological development and industrial productivity as well as people's living standards in a society. Therefore, achieving solutions to the energy shortages and environmental problems presented today requires long-term potential actions for sustainable development. In this regard, green energy sources and technologies appear to be the one of the most efficient and effective solutions. It can be said that one solution to the impending energy shortage and environmental problems is to make much more use of green energy sources and technologies. Another important solution is to develop permanent and effective sustainable green energy strategies for the increase of sustainable global stability.[1,6] For these purposes, engineering practicality, reliability, applicability, economy, scarcity of supply, and public acceptability should also be considered.

Green Energy and Environmental Consequences

Some of the global problems affecting world stability are presented in Figure 1. Fossil fuel utilization effects such as global climate change, world energy conflicts, and energy source shortages have increasingly threatened world stability. These negative effects may be observed locally, regionally, and globally.

This concern arises due to world population growth, fast technological development, and increasing energy demand. In the past, fossil energy sources could be used to solve world energy problems. However, in this century, fossil fuels cannot continue indefinitely as the principal energy sources due to the rapid increase of world energy demand and energy consumption. Due to world population growth and the advance of technologies that depend on fossil fuels, the reserves of fossil fuels eventually will

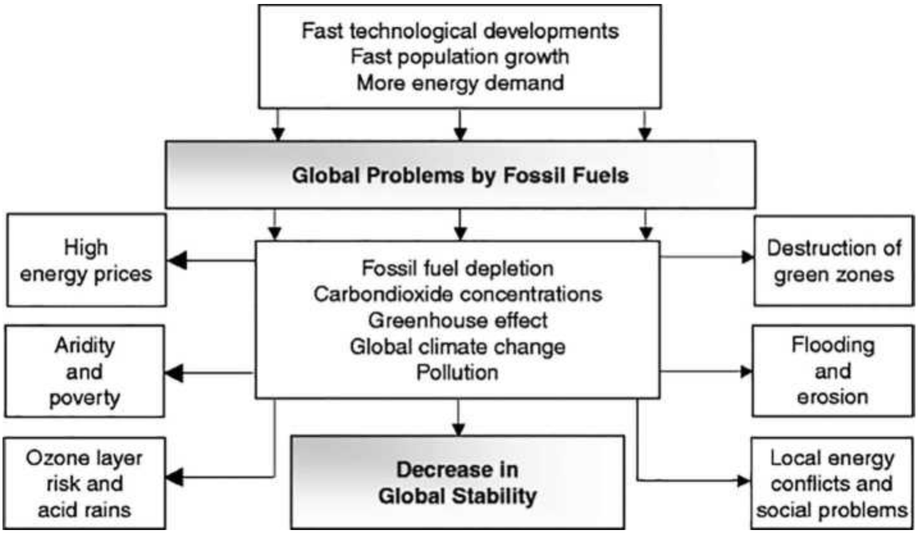


FIGURE 1 Problems affecting global stability.

not be able to meet world energy demand. Energy experts point out the fact that reserves account for less than 40 years for petroleum, 60 years for natural gas, and 250 years for coal.[1] The increase of the energy consumption and energy demand indicates our dependence on the fossil fuels. If the increase of fossil fuel utilization continues in this manner, it is likely that the world will be affected by many negative problems due to fossil fuels. As one outcome, global stability will probably decrease. This effect is presented as flow chart in Figure 2. For instance, shortages of fossil fuel reserves and global environmental problems will likely lead to global unrest throughout the world. As a result, local, regional, and world conflicts may appear across the world.

To further support these arguments, the observed and predicted consumptions of world primary energy, fossil fuels, and green energy from 1965 to 2050 are displayed in Figure 3 based on the data taken from literature.[7,8]

According to Figure 3, the quantities of world primary energy consumption are expected to reach 16502 Mtoe (Million tons of oil equivalent) by the year 2050. World population is now over 6 billion, double the population of 40 years ago, and it is likely to double again by the middle of the 21st century. The world's population is expected to rise to about 7.0 billion by 2010. Even if birth rates fall so that the world population becomes stable by 2050, the population will still be about 10 billion. Because the population is expected to increase drastically, conventional energy resource shortages are likely due to insufficient fossil fuel resource supplies. In the near future, therefore, green energy will become increasingly important to compensate for shortages of conventional resources. The correlations that have been applied here are as follows:

For world primary energy consumption:

$$\mathrm{WPEC} = 143.57 \times Y - 277808 \left(R^2 = 0.9902\right) \tag{1}$$

For world fossil fuel consumption:

$$\mathrm{WPEC} = 113.61 \times Y - 219092 \left(R^2 = 0.9968\right) \tag{2}$$

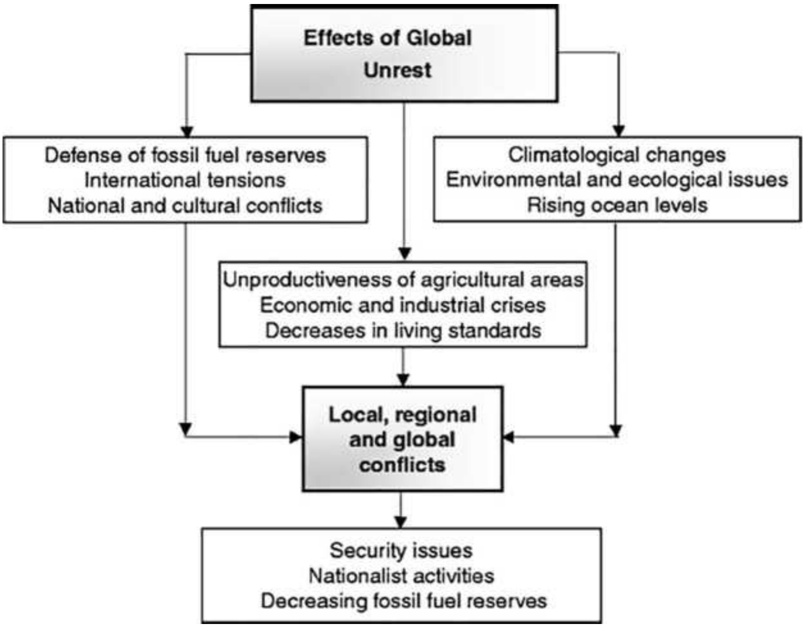


FIGURE 2 Some possible effects and results of the global unrest.

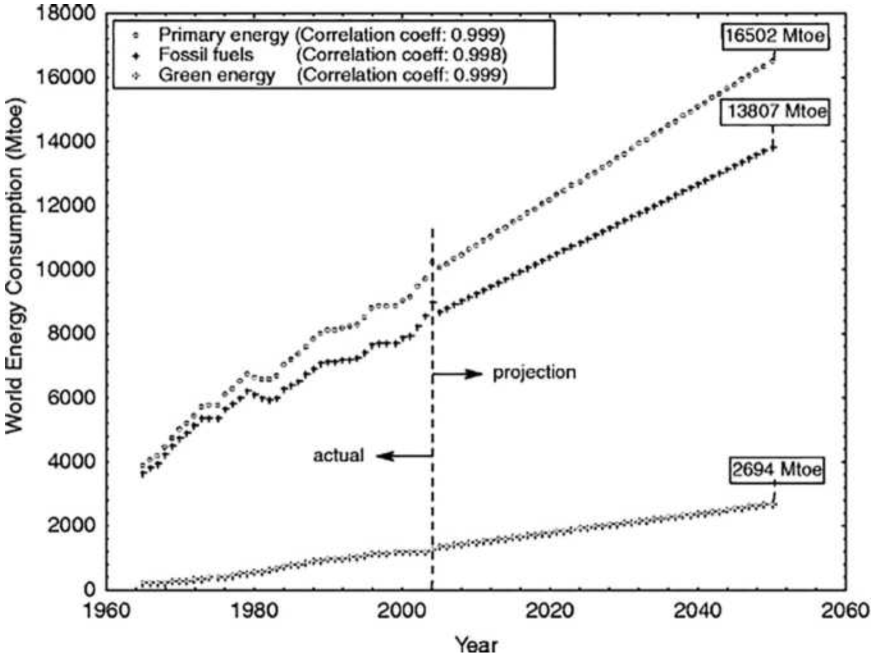


FIGURE 3 Observed and projected quantities of world primary energy, fossil fuels, and green energy consumptions (Mtoe: million tons of oil equivalent).

For world green energy consumption:

$$\text{WPEC} = 29.956 \times Y - 58715 \left(R^2 = 0.9985\right) \tag{3}$$

Here, WPEC is world primary energy consumption (Mtoe); WFFC, world fossil fuel consumption (Mtoe); WGEC, world green energy consumption; *Y* is time (years); and *R* is the correlation coefficient.

Energy shortages will accelerate the fluctuations of energy prices and economic recessions and decrease living standards and increase the unrest among countries. Decreased available fossil fuel reserves and increased fuel costs since the middle to late 1900s have led to variations in lifestyles and standards of life. These effects have, in some regions, decreased living standards of entire societies. Countries that need more energy resources have been purchasing cheaper energy sources. Countries that look after the future welfare of their societies have received the attention of countries that possess fossil fuel supplies, posing the potential for world conflict. Problems are often attributed to decreases in fossil fuel energy reserves. Those who seek a clean world must find appropriate alternatives to fossil fuels. Why not green energy? Green energy that is abundantly available all over the world can help:

- Reduce or stop conflicts among countries regarding energy reserves
- Facilitate or necessitate the development of new technologies
- Reduce air, water, and land pollution and the loss of forests
- Reduce illnesses and deaths caused by direct or indirect use of energy

Therefore, green energy and related technologies are needed to improve global stability by reducing the harmful effects of fossil-based energy consumption. Thus, the importance of green energy in reducing world problems and achieving a sustainable energy system should be emphasized and a transition to the green economy should be encouraged.

Energy, Environment, and Sustainability

Environmental concerns are an important factor in sustainable development. For a variety of reasons, activities that continually degrade the environment are not sustainable over time, e.g., the cumulative impact on the environment of such activities often leads over time to a variety of health, ecological, and other problems. A large portion of the environmental impact in a society is associated with its utilization of energy resources. Ideally, a society seeking sustainable development utilizes only energy resources that release no or minimal emissions to the environment and thus cause no or little environmental impact. However, because all energy resources lead to some environmental impact, it is reasonable to suggest that some (not all) of the concerns regarding the limitations imposed on sustainable development by environmental emissions and their negative impacts can be in part overcome through increased energy efficiency. Clearly, a strong relation exists between energy efficiency and environmental impact because for the same services or products less resource utilization and pollution is normally associated with increased energy efficiency. While not all green energy resources are inherently clean, there is such a diversity of choices that a shift to renewables carried out in the context of sustainable development could provide a far cleaner system than would be feasible by tightening controls on conventional energy.[9] Furthermore, being by nature site-specific, they favor power system decentralization and locally applicable solutions more or less independent of the national network. It enables citizens to perceive positive and negative externalities of energy consumption. Consequently, the small scale of the equipment often makes the time required from initial design to operation short, providing greater adaptability in responding to unpredictable growth and changes in energy demand. In this regard, sustainability often leads local and national authorities to incorporate environmental considerations into energy planning. The need to satisfy basic human needs and aspirations combined with the increasing world population will make the need for successful implementation of sustainable development increasingly apparent. Various criteria that are essential to achieving sustainable development in a society are as follows:[3,4,10]

- Information about and public awareness of the benefits of sustainability investments, environmental education, and training
- Appropriate energy strategies
- The availability of green energy resources and technologies
- A reasonable supply of financing
- Monitoring and evaluation tools

Environmental concerns are significantly linked to sustainable development. Activities which continually degrade the environment are not sustainable. For example, the cumulative impact on the environment of such activities often leads over time to a variety of health, ecological, and other problems. Clearly, a strong relation exists between efficiency and environmental impact because for the same services or products, less resource utilization and pollution is normally associated with increased efficiency.[3] Improved energy efficiency leads to reduced energy losses. Most efficiency improvements produce direct environmental benefits in two ways. First, operating energy input requirements are reduced per unit of output and pollutants generated are correspondingly reduced. Second, consideration of the entire life cycle for energy resources and technologies suggests that improved efficiency reduces environmental impact during most stages of the life cycle. In recent years, the increased acknowledgment of humankind's interdependence with the environment has been embraced in the concept of sustainable development. With energy constituting a basic necessity for maintaining and improving standards of living throughout the world, the widespread use of fossil fuels may have impacted the planet in ways far more significant than first thought. In addition to the manageable impacts of mining and drilling for fossil fuels and discharging wastes from processing and refining operations, the greenhouse gases created by burning these fuels is regarded as a major contributor to a global warming threat.

Global warming and large-scale climate change have implications for food chain disruption, flooding, and severe weather events.

Therefore, use of green energy sources can help reduce environmental damage and achieve sustainability. The attributes of green energy technologies (e.g., modularity, flexibility, low operating costs) differ considerably from those for traditional, fossil-fuel-based energy technologies (e.g., large capital investments, long implementation lead times, operating cost uncertainties regarding future fuel costs). Green energy technologies can provide cost-effective and environmentally beneficial alternatives to conventional energy systems. Some of the benefits of the green energy- based systems are as follow:[3,10]

- They are relatively independent of the cost of oil and other fossil fuels, which are projected to rise significantly over time. Thus, cost estimates can be made reliably for green energy systems and they can help reduce the depletion of the world's nongreen energy resources.
- Implementation is relatively straightforward.
- They normally cause minimum environmental degradation and so they can help resolve major environmental problems. Widespread use of green energy systems would certainly reduce pollution levels.
- They are often advantageous in developing countries. In fact, the market demand for renewable energy technologies in developing nations will likely grow as they seek a better standard of living.

Under these considerations, green energy resources have some characteristics that lead to problematic but often solvable technical and economic challenges:

- Generally diffuse
- Not fully accessible
- Sometimes intermittent
- Regionally variable

The overall benefits of green energy technologies are often not well understood, leading to such technologies often being assessed as less cost-effective than traditional technologies. For green energy technologies to be assessed comprehensively, all of their benefits must be considered. For example, many green energy technologies can provide, with short lead times, small incremental capacity additions to existing energy systems. Such power generation units usually provide more flexibility in incremental supply than large devices like nuclear power stations.

Green Energy and Sustainable Development

Sustainability has been called a key to the solution of current ecological, economic, and developmental problems by Dincer and Rosen[10] A secure supply of energy resources is generally agreed to be a necessary but not sufficient requirement for development within a society. Furthermore, sustainable development demands a sustainable supply of energy resources that, in the long term, is readily and sustainably available at reasonable cost and can be utilized for all required tasks without causing negative societal impacts. Supplies of such energy resources as fossil fuels (coal, oil, and natural gas) and uranium are generally acknowledged to be finite; other energy sources such as sunlight, wind and falling water are generally considered renewable and therefore sustainable over the relatively long term. Wastes (convertible to useful energy forms through, for example, waste-to-energy incineration facilities) and biomass fuels are also usually viewed as sustainable energy sources.

In general, the implications of these statements are numerous and depend on how sustainable is defined.[3] For sustainable development, green energy can play an important role for meeting energy requirements in both industrial and local applications. Therefore, development and utilization of green energy sources and technologies should be given a high priority for sustainable development in a country. The need for sustainable energy development is increasing rapidly in the world. Widespread use of green energy sources and technologies is important for achieving sustainability in the energy sectors in both developing and industrialized countries. Thus, green energy resources and technologies are a key component of sustainable development for three main reasons:

- They generally cause less environmental impact than other energy sources. The variety of green energy resources provides a flexible array of options for their use.
- They cannot be depleted. If used carefully in appropriate applications, green energy resources can provide a reliable and sustainable supply of energy almost indefinitely.
- They favor system decentralization and local solutions that are somewhat independent of the national network, thus enhancing the flexibility of the system and providing economic benefits to small isolated populations. Also, the small scale of the equipment often reduces the time required from initial design to operation, providing greater adaptability in responding to unpredictable growth and changes in energy demand.
- Major considerations involved in the development of green energy technologies for sustainable development as modified from[10] are presented in Figure 4.

As a consequence, it can be said that green energy and technologies are definitely needed for sustainable development that ensures a minimization of global unrest. The relation between green energy and sustainability is of great significance to the developed countries as well as developing or less developed countries. Moreover, examining the relations between green energy sources and sustainability makes it clear that green technology is directly related to sustainable development. Therefore, attaining

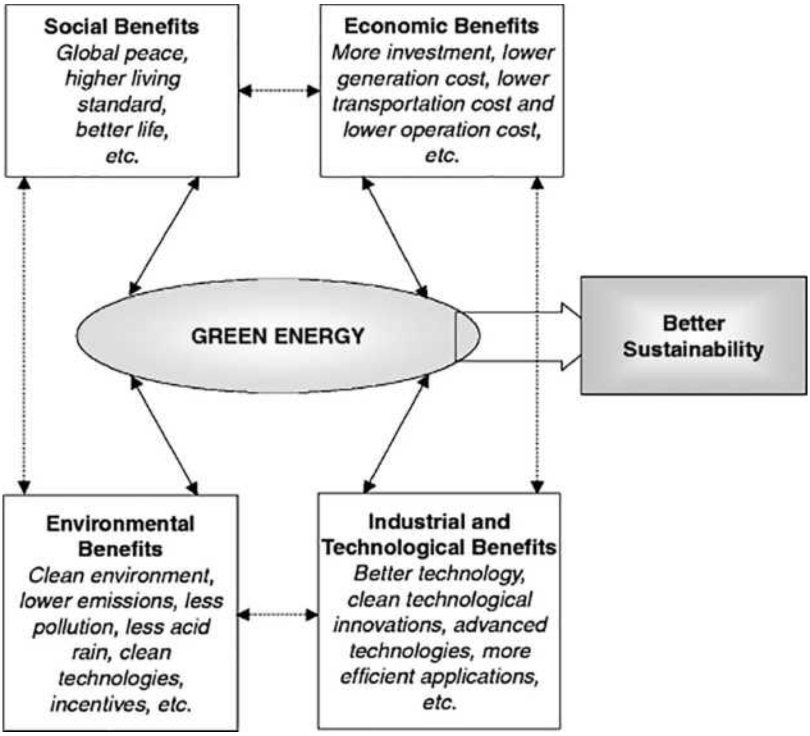


FIGURE 4 Major considerations involved in the development of green energy technologies for sustainable development.

sustainable development requires that green energy resources be also used, and is assisted if resources are used efficiently.[4,10] Thus, if sustainable green energy technologies are effectively put into practice, the countries may maximize the benefits from the green energy sources and technologies while minimizing the global unrest[1] associated with the use of fossil fuel energy sources. It is expected that this initiative can contribute to development over a longer period of time, i.e., to make development more sustainable.

Green Energy Resources and Technologies

Since the oil crises in the early 1970s, there has been active worldwide research and development in the field of green energy resources and technologies. During this time, energy conversion systems that were based on green energy technologies appeared to be most attractive because of facts such as the projected high cost of oil and the cost effectiveness estimates and easy implementation of green energy technologies. Furthermore, in more recent times, it has been realized that green energy sources and technologies can have a beneficial impact on essential technical, environmental, economic, and political issues of the world. As pointed out by Hartley,[3,9] green energy technologies produce marketable energy by converting natural phenomena into useful energy forms. These technologies use the energy inherent in sunlight and its direct and indirect impacts on the earth (photons, wind, falling water, heating effects, and plant growth), gravitational forces (the tides), and the heat of the earth's core (geothermal) as the resources from which they produce energy. These resources represent a massive energy potential which dwarfs that of equivalent fossil resources. Therefore, the magnitude of these is not a key constraint on energy production.

However, they are generally diffused and not fully accessible, some are intermittent, and all have distinct regional variabilities. Such aspects of their nature give rise to difficult, but solvable, technical, institutional, and economical challenges inherent in development and the use of green energy resources. Despite having such difficulties and challenges, the research and development on green energy resources and technologies has been expanded during the past two decades because of the facts listed above. Recently, significant progress has been made by (1) improving the collection and conversion efficiencies, (2) lowering the initial and maintenance costs, (3) increasing the reliability and applicability, and (4) understanding the phenomena of green energy technologies.

Green energy technologies become important as environmental concerns increase, utility (hydro) costs climb, and labor costs escalate.[3,9] The uncertain global economy is an additional factor. The situation may be turned around with an increase in research and development in the advanced technologies fields, some of which are closely associated with green energy technologies. This may lead to innovative products and job creation supported by the governments. The progress in other technologies, especially in advanced technologies, has induced some innovative ideas in green energy system designs. The ubiquitous computer has provided means for optimizing system performance, costs/benefits, and environmental impacts even before the engineering design was off the drawing board. The operating and financial attributes of green energy technologies, which include modularity and flexibility and low operating costs (suggesting relative cost certainty) are considerably different than those for traditional, fossil-based technologies, whose attributes include large capital investments, long implementation lead times, and operating cost uncertainties regarding future fuel costs. The overall benefits of green energy technologies are often not well understood and consequently they are often evaluated to be not as cost effective as traditional technologies. In order to comprehensively assess green energy technologies, however, some of their benefits that are often not considered must be accounted for. Green energy technologies, in general, are sometimes seen as direct substitutes for existing technologies so that their benefits and costs are conceived in terms of assessment methods developed for the existing technologies. Many government organizations and universities recognize the opportunity and support the efforts to exploit some commercial potential by:

- Analyzing opportunities for green energy and working in consultation with industry to identify research, development, and market strategies to meet technological goals.
- Conducting research and development in cooperation with industry to develop and commercialize technologies.
- Encouraging the application of green energy technologies to potential users, including utilities.
- Providing technical support and advice to industry associations and government programs that are encouraging the increased use of green energy. In order to realize the energy and the economic and environmental benefits that green energy sources offer, the following integrated set of activities should be acted on accordingly.[3,9,10]
- Conducting research and development. The priorities should be set in close consultation with industry to reflect their needs. Most research is conducted through cost- shared agreements and falls within the short-to-medium term. Partners in research and development should include a variety of stakeholders in the energy industry such as private sector firms, utilities across the country, provincial governments, and other federal departments.
- Assessing technology. Data should be gathered in the lab and through field trials on factors such as cost benefit, reliability, environmental impact, safety, and opportunities for improvement. This data should also assist the preparation of technology status overviews and strategic plans for research and development.
- Developing standards. The development of technical and safety standards is needed to encourage the acceptance of proven technologies in the marketplace. Standards development should be conducted in cooperation with national and international standards writing organizations as well as other national and provincial regulatory bodies.
- Transferring technology. Research and development results should be transferred through the sponsorship of technical workshops, seminars, and conferences, as well as through the development of training manuals and design tools and the publication of technical reports. Such activities will also encourage potential users to consider the benefits of adopting green energy technologies. In support of developing near-term markets, a key technology transfer area is to accelerate the use of green energy technologies in a country's remote communities.

Such activities will also encourage potential users to consider the benefits of adopting renewable energy technologies. In support of developing near-term markets, a key technology transfer area is to accelerate the use of green energy technologies in a country's remote communities.

Essential Factors for Sustainable Green Energy Technologies

There are various essential parameters, as outlined and detailed in Figure 5. These factors can help in identifying and achieving required green energy strategies and technologies for sustainable development. As shown in Figure 5, green energy technologies are largely shaped by broad and powerful trends that are rooted in basic human needs. In conjunction with this, the increasing world population requires the definition and successful implementation of green energy technologies. Briefly, the important parameters and their interrelations, as outlined in Figure 5, are definitely required to carry out the best green energy program and to select the most appropriate green energy technologies for sustainable development.

Exergetic Aspects of Green Energy Technologies

The impact of energy resource utilization on the environment and the achievement of increased resource-utilization efficiency are best addressed by considering exergy. The exergy of an energy form or a substance is a measure of its usefulness, quality, or potential to cause change and provide the basis for an effective measure of the potential of a substance or energy form to impact the environment. In practice,

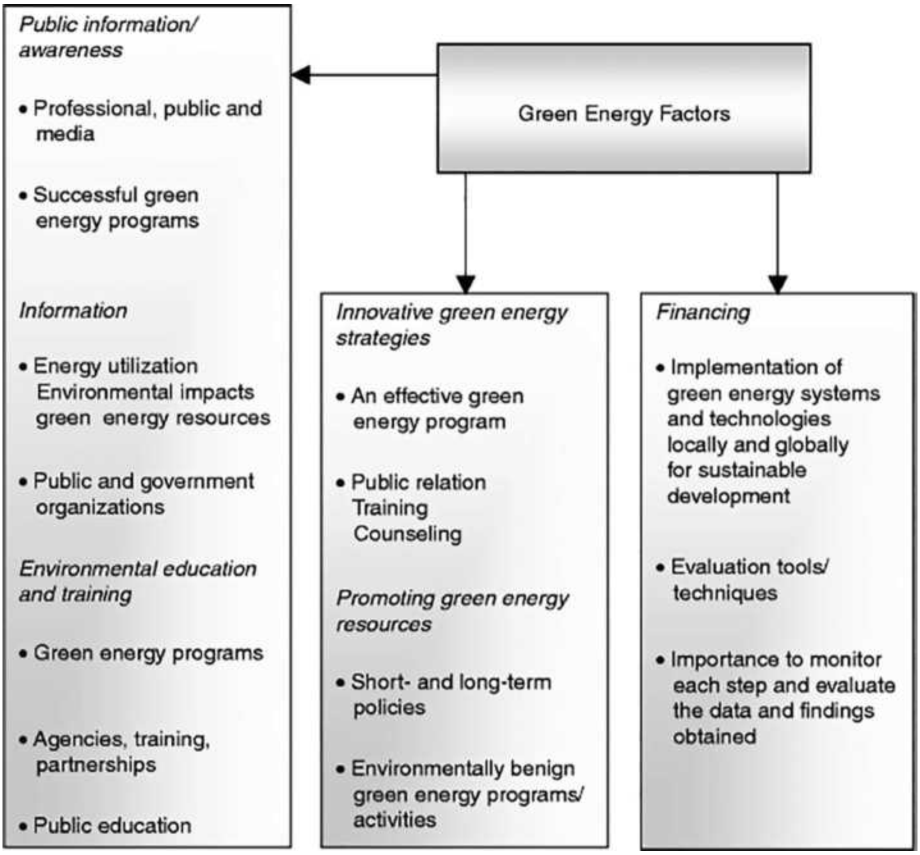


FIGURE 5 Essential factors for green energy technologies.

a thorough understanding of exergy and the insights it can provide into the efficiency, environmental impact, and sustainability of green energy technologies are helpful if not required for the engineer or scientist working in the area of green energy–based environmental systems.[4,11] For green energy technologies, applications of exergy methods can have numerous broad and comprehensive benefits:

- A better knowledge of the efficiencies and losses for the technologies and systems and how they behave and perform
- A clearer appreciation of the environmental impacts of green energy technologies as well as the mitigation of environmental that they can facilitate
- Better identification of the ways green energy technologies can contribute to sustainable development

Green Energy Applications

Green energy technologies are expected to play a key role in sustainable energy scenarios for the future. The foremost factor that will determine the specific role of green energy and technologies will likely be energy demand. Therefore, in order to compensate the energy requirement, it will be possible to produce green energy from renewable energy sources such as hydraulic, solar, wind, geothermal, wave and biomass, etc. If so, the green energy and technologies can be utilized for many application fields as shown in Figure 6.

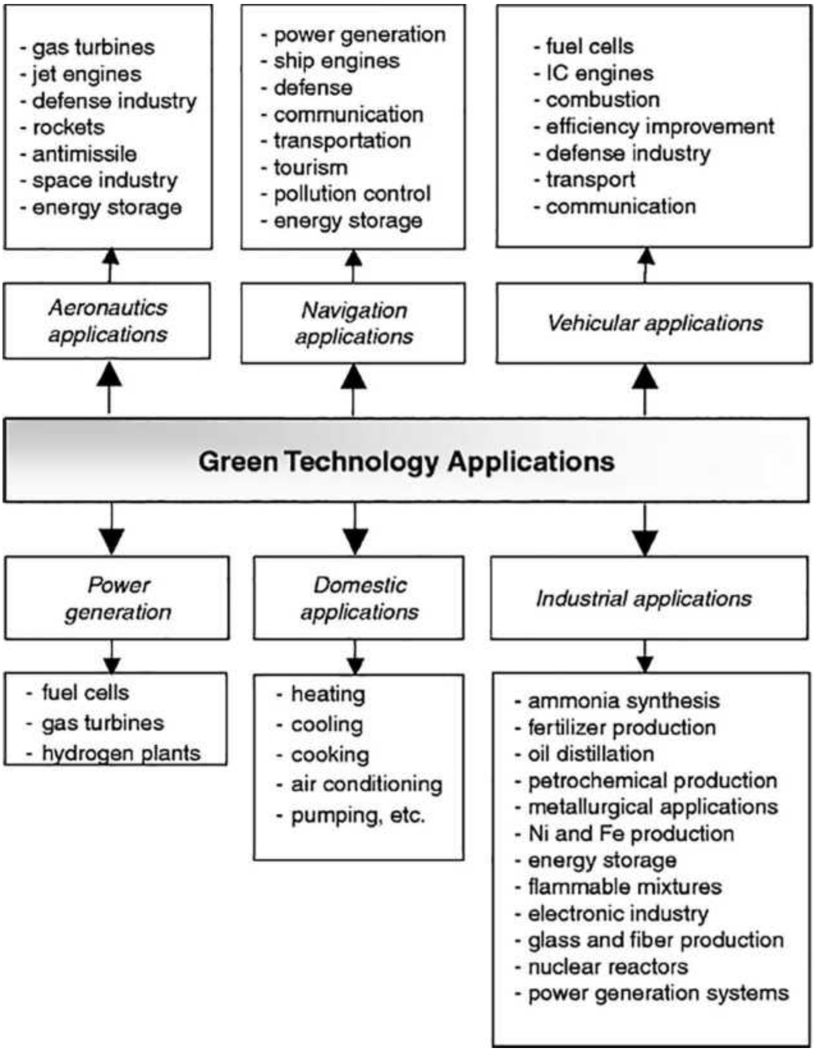


FIGURE 6 Possible application fields of green energy and technologies for sustainable

Thus, it can be said that green energy and technologies, which are abundantly available, can help:

- Provide more environmentally benign and more sustainable future
- Increase energy security
- Facilitate or necessitate the development of new, clean technologies
- Reduce air, water, and soil pollution and the loss of forests
- Reduce energy-related illnesses and deaths
- Reduce or stop conflicts among countries regarding energy reserves

Therefore, green energy and related technologies are needed to ensure global stability by reducing the harmful effects of fossil-based energy consumption. Thus, the importance of green energy in reducing the world problems and achieving a sustainable energy system should be emphasized considering the sustainable green energy strategies; and a transition to green energy economy should be encouraged and developed countries in particular should increase investments in green energy and technologies.

Green Energy Analysis

In light of the above major considerations, in order to accelerate the use of green energy sources and technologies and the implementation of green energy strategies, and in this regard, to describe the relationship between the global stability and sustainable development, some key steps are presented as follows:

- Key strategies
- Green energy based sustainability ratio
- Global unrest and peace

Green Energy-Based Sustainable Development Ratio

In order to discuss the key role of green energy for sustainable development and global stability, the general algebraic form of the equation that is the green energy-based sustainable development ratio should be presented based on the above strategies. For this purpose, the works that were early presented by Midilli et al.[1,2,6,12] and Dincer[3,10] were taken as the reference basis. The following important parameters, which are sectoral, technological, and practical application impact ratios, are taken into consideration to estimate the green-energy-based sustainable development ratio. The detailed derivations of these parameters are presented in the literature.[2] Briefly, the green energy–based sustainable development ratio is mainly based on the following parameters:

- *Sectoral impact ratio,* (R_{si}) (ranging from 1 to 1/3), is based on the provided financial support ($C_{p,\ si}$) of public, private, and media sectors for transition to green energy–based technologies, and depending on the total green energy financial budget (C_{geb}), as a reference parameter.[2]
- *Technological impact ratio,* (R_{ti}) (ranging from 1 to 1/3), is based on the provided financial support ($C_{p,\ ti}$) for research and development, security, and analysis of green energy–based technologies, and also depending on the total green energy financial budget (C_{geb}) as a reference parameter.[2]
- *Practical application impact ratio,* (R_{pai}) (ranging from 1 to 1/3), is based on the provided financial support ($C_{p,\ pai}$) for projection, production, conversion, marketing, distribution, management, and consumption of green fuel from green energy sources, and also depending on the total green energy financial budget (C_{geb}), as a reference parameter.[2] Here, it should be emphasized that these parameters can be defined as the ratio of the provided financial support to the total green energy financial budget in a country. In addition, it should be always remembered that it is assumed that the financial share of each parameter is equal to one-third of the total green energy financial budget in a country. The utilization ratio of green energy depends on the green energy utilization and the world primary energy quantity. Green energy–based global stability is a function of utilization ratios of coal, petroleum, and natural gas, and utilization ratios of green energy at certain utilization ratios of these fuels, and utilization ratio of green energy. Therefore, the green-energy-based sustainable development ratio is written in algebraic form as follows:[2]

$$R_{ges} = R_{gei} \times R_{geu} \tag{4}$$

where

$R_{gei} = \left(R_{si} + R_{ti} + R_{pai}\right)$ green energy impact ratio; $R_{si} = C_{p,\,si}/C_{geb}$ sectoral impact ratio, which is estimated based on the provided financial support of the public, private, and media sectors for transition to green energy ($C_{p,\ si}$) and also the green energy budget allocation of a country:(C_{geb});

$R_{ti} = C_{p,\,ti}/C_{geb}$ technological impact ratio, which is estimated based on the provided financial supports for research and development, security, and analysis of green- energy-based technologies ($C_{p,\ ti}$) and also the green energy budget allocation of a country (C_{geb});

$R_{\text{pai}} = C_{\text{p, pai}}/C_{\text{geb}}$ practical application impact ratio, which is estimated based on the provided financial supports for projection, production, conversion, marketing, distribution, management, and consumption of green fuel from green energy sources, and also the green energy budget allocation of a country (C_{geb});

$R_{\text{geu}} = 1 - Q_{\text{wffc}}/Q_{\text{wpec}}$ green energy utilization ratio, which is also defined as a function of fossil fuel utilization ratio (R_{ffu}).

Here Q_{wffc} explains world fossil fuel consumption (M) and Q_{wpec} world primary energy consumption (M).

Global Unrest and Peace

Fossil fuels such as petroleum, coal, and natural gas, which have been extensively utilized in industrial and domestic applications for a long time, have often been the cause of global destabilization and unrest. This problem is likely to increase in significance in the future and suggests the need for investigations of, among other factors, the role of hydrogen energy relative to future global unrest and global peace. In general, the global unrest arising from the use of fossil fuels is considered as a function of the usage ratios of these fossil fuels.[12] In order to estimate the level of global unrest, it is important to select a proper reference value. Therefore, it is assumed that the lowest value of global unrest and the highest value of global peace are equal to 1, which is a reference point to evaluate the interactions between global unrest and global peace.[12] Consequently, the general algebraic case form of global unrest expression is as follows[12]:

$$\text{GU} = \left\{ 1 - \left[\left(\frac{q_p}{q_{\text{wpec}}} + \frac{q_c}{q_{\text{wpec}}} + \frac{q_{\text{ng}}}{q_{\text{wpec}}} \right) - \left(r_{H_2-p} + r_{H_2-c} + r_{H_2-ng} \right) \right] \right\}^{-1} \tag{5}$$

where q_{p}, q_{c} and q_{ng} define the consumption quantities of petroleum, coal, and natural gas, respectively; q_{wpec} defines the quantity of world primary energy consumption; r_{H_2-p} is the utilization ratio of hydrogen from nonfossil fuels at a certain utilization ratio of petroleum; r_{H_2-c} is the utilization of hydrogen from nonfossil fuels at a certain utilization ratio of coal; $r_{H_2-\text{ng}}$ is the utilization of hydrogen from nonfossil fuels at a certain utilization ratio of natural gas. The values of q_{p}, q_{c}, q_{ng}, and q_{wpec} can be taken from the literature.[8] The values for r_{H_2-p}, r_{H_2-c}, and $r_{H_2-\text{ng}}$ can be taken depending on the utilization ratios of the petroleum, coal, and natural gas presented in the literature.

In order to estimate the level of global peace quantitatively, it is assumed that the highest value of global peace is equal to 1, which is a reference point to evaluate the interactions between global unrest and global peace.[12] Thus, the relationship between global unrest and global peace can be written as a function of the utilization ratio of hydrogen from nonfossil fuels. The general algebraic case form of global peace expression proposed is as follows:

$$\text{GP} = \frac{U_{H_2}}{Q_{\text{wpec}}} \times \frac{1}{(\text{GU})} \tag{6}$$

where U_{H_2} defines the utilization of hydrogen from non-fossil fuels, which can be taken from the literature.

Case Study

The first case study is given to determine the green-energy- based sustainability ratio depending on the three subcases that are carried out based on the sectoral impact ratio, the technological impact ratio, and the practical application impact ratio by using actual data from the literature.[8] All cases are presented in Table 1. The results of the first case study are presented in Figures. 7 and 8a–c.

TABLE 1 The Cases Depending on Green Energy Impact Ratio for the Green Energy–Based Sustainability Ratio

Effect of Variable Parameters (entity)			10%	30%	50%	70%	90%
Cases		Percent of Total Financial Budget for Green Energy	Green Energy Impact Ratio (Rgei)				
I	2 constant parameters	n/3	0.699	0.766	0.833	0.899	0.966
	1 variable parameter	k ×(1/3)×∈					
II	1 constant parameter	n/3	0.400	0.533	0.666	0.800	0.933
	2 variable parameters	k ×(1/3)× ∈					
III	3 variable parameters	k×(1/3)×∈	0.100	0.300	0.500	0.700	0.900

Here, n defines the number of constant parameters; k, the number of variable parameters.

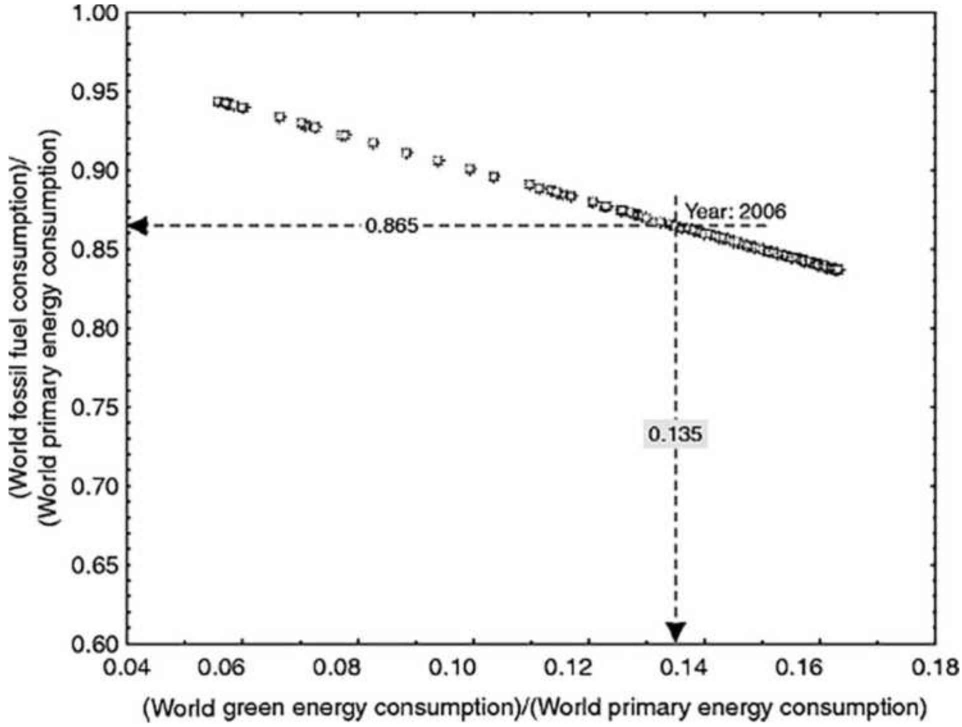


FIGURE 7 World fossil fuel consumption ratio as a function of green energy consumption ratio.

The second case study is performed to determine the global unrest and global peace level based on the predicted utilization ratios of hydrogen from nonfossil fuels. In this regard, two important empirical relations that describe the effects of fossil fuels on world peace and global unrest are taken into consideration for this case study. The results of this case study are presented in Figure 9.

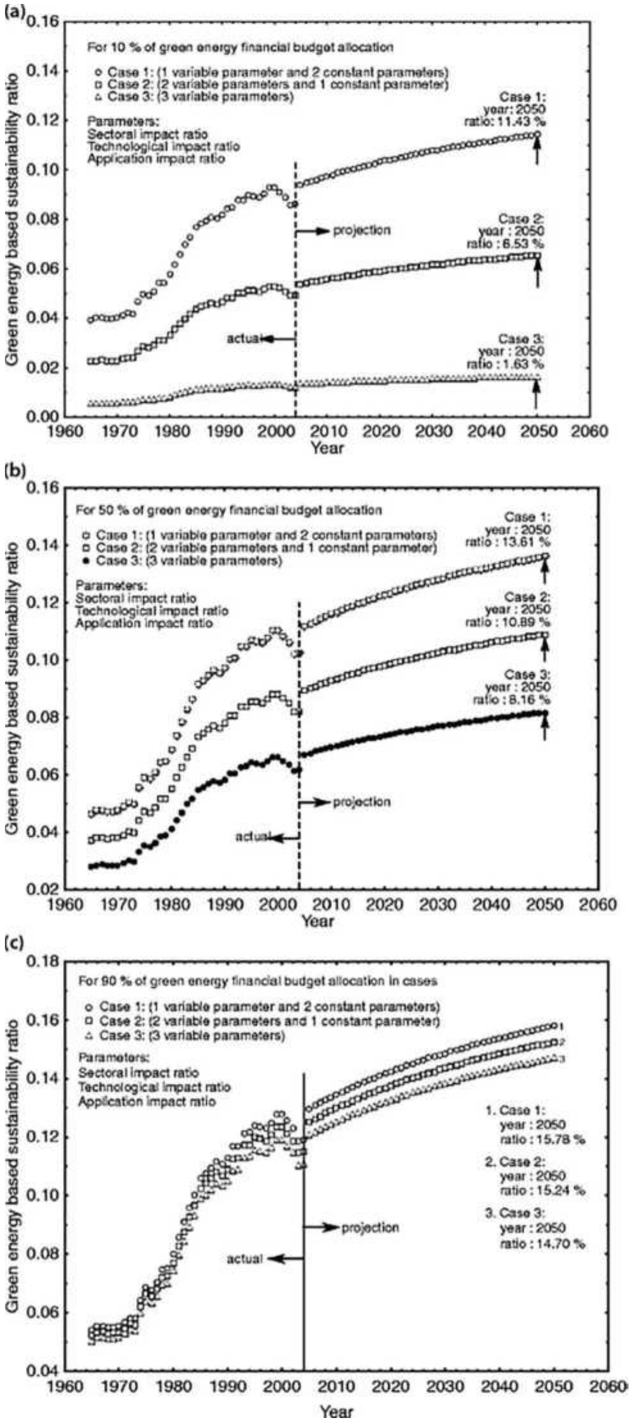


FIGURE 8 (a) Green energy–based sustainability ratio versus time depending on the actual and projected green energy consumption data for 10% of green energy financial budget allocation. (b) Green energy–based sustainability ratio versus time depending on the actual and projected green energy consumption data for 50% of green energy financial budget allocation. (c) Green energy–based sustainability ratio versus time depending on the actual and projected green energy consumption data for 90% of green energy financial budget allocation.

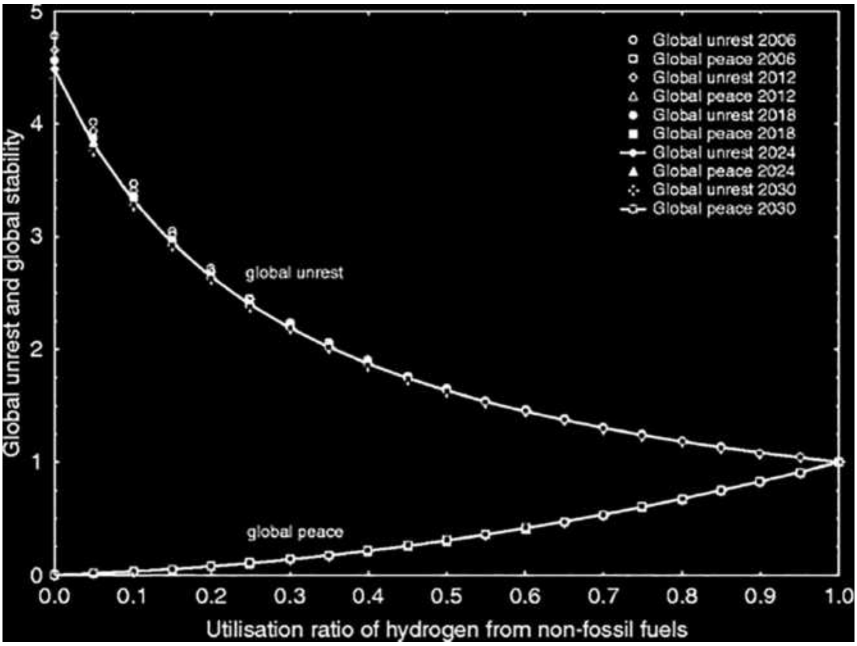


FIGURE 9 Comparison of levels of global unrest and global stability as a function of hydrogen utilization ratios from non-fossil fuels.

Results and Discussion

In accordance with the objective of this entry, the projection data is obtained using actual data of primary energy, fossil fuel, and green energy consumptions. Considering the technological impact ratio (max. value=1/3 of the green energy financial budget), the sectoral impact ratio (max. value=1/3 of the green energy financial budget allocation), and the practical application impact ratio (max. value=1/3 of the green energy financial budget allocation), three cases are analyzed and discussed in detail in this part.

Figure 7 shows the variations of the fossil fuel utilization ratio (world fossil fuel consumption/world primary energy consumption) as a function of the green energy utilization ratio (world green energy consumption/world primary energy consumption). It is found out from this figure that the fossil fuel utilization ratio decreases depending on the rise of the year while green energy utilization ratio increases. For example, the green energy utilization ratio was 5.71% in 1970, 8.25% in 1980, 11.67% in 1990, and 13.27% in 2000 while the fossil fuel utilization ratio was 94.29% in 1970, 91.75% in 1980, 88.32% in 1990, and 86.77% in 2000, based upon the actual data. However, it is observed that the green energy utilization ratio increased and reached 12.31% while fossil fuel utilization ratio decreased to 87.69% in 2004. Based on the projected data, it expected that green energy utilization ratio will reach almost 13.52% in 2006, 14.09% in 2012, 14.58% in 2018, 15.01% in 2024, and 15.38% in 2030. However, it is expected that the fossil fuel utilization ratio will decrease to almost 86.48% in 2006, 85.91% in 2012, 85.42% in 2018, 85.99% in 2024, and 84.62% in 2030. Thus, in order to increase the green energy utilization ratio and to reduce the harmful effects resulting from the fossil fuel consumption, the investments on green energy should be encouraged and the green energy strategies should be put into practice for sustainable development.

First, it should be stated that one or two of the parameters that are sectoral, technological, and practical application impact ratios can be selected as constant parameters. As shown in Table 1, a variable and two constant parameters are considered in Case 1, two variables and one constant parameter are considered in Case 2, and three variable parameters are considered in Case 3. When Case 1 is applied for green energy supply and progress, it is found that the green energy impact ratio changes between 0.699 and 0.966 depending on the percentages of the variable parameter. In the application of Case 2 it is obtained that the green energy impact ratio varies between 0.40 and 0.933 depending on the percentages of two variable parameters. In the application of Case 3, it is calculated that the green energy impact ratio changes from 0.1 to 0.9 depending on the percentages of three variable parameters. When the three Cases are compared to each other, it can be said that the highest values of green energy impact ratio are found by applying Case 1, and also that Case 3 gives the lowest green energy impact ratios. Thus, Case 1 should be selected to increase the green energy impact ratio and the green-energy-based sustainability ratio.

Figure 8a–c show a variation of the green energy–based sustainability ratio (R_{ges}) as a function of year by depending on the percentages of the green energy financial budget as 10%, 50%, and 90%, or the effect of the parameters in the Cases, respectively. The values of green-energy-based sustainability ratios were calculated using Figure 4.

As shown in these figures, the values of R_{ges} increase with time based on the cases. The highest values of R_{ges} are obtained when Case 1 is applied, as shown in Figure 8a–c. For example, the green-energy-based sustainability ratios are estimated to be 9.46% in 2006, 9.86% in 2012, 10.21% in 2018, 10.50% in 2024, and 10.76% in 2030 in case of 10% of green energy financial budget; 13.06% in 2006, 13.62% in 2012, 14.09% in 2018, 14.51% in 2024, and 14.86% in 2030 in case of 90% of green energy financial budget.

It is important to implement green energy strategies through green energy systems and applications for sustainable future. If so, the green energy–based sustainability ratio increases and green energy is more easily supplied, thus its technologies are more preferred and applied. Hence, as long as sustainable green energy strategies are increasingly applied and the green technologies are more utilized and encouraged, the negative effects stemming from the fossil fuel utilization will decrease and the green-energy-based sustainability ratio will increase.

Figure 9 compares the levels of global unrest and global peace as a function of the predicted utilization ratio of hydrogen from non-fossil fuels. Figure 4 indicates that there is an inversely proportional relationship between global peace and global unrest, depending on the utilization ratio of hydrogen from non-fossil fuels.

To better appreciate this figure, some of the key energy- related reasons for global unrest need to be understood. They include:

- Increases in fossil fuel prices
- Environmental effects of energy use, including pollution due to emissions, stratospheric ozone layer depletion, and global warming
- Decreases in the amount of fossil fuel available per capita and the associated decrease in living standards
- Increases in energy demand due to technological developments attributable to and based on fossil fuels
- Depletion of fossil fuel reserves
- Increases in conflicts for fossil fuel reserves throughout the world
- The lack of affordable and practical alternative energy sources to fossil fuels

It is found out from Figure 9 that an increase in hydrogen utilization accordingly decreases the reasons for global unrest, allowing the benefits of global peace to be realized over time; and the lowest levels of global unrest occur when hydrogen from non-fossil fuels is substituted completely for fossil fuels. In general, the level of global unrest is higher than 1 and the problems causing global unrest can be reduced by using hydrogen energy from non-fossil sources instead of fossil fuels. Figure 5 suggests that the utilization of hydrogen from non-fossil fuels at certain ratios of petroleum, coal, and natural gas decreases the amount of fossil fuel consumption and thus reduces the level of global unrest closer to 1. As shown in Figure 9, it is expected that, depending on the actual and projected fossil fuel consumption data, the levels of global unrest will be approximately 4.78 in 2006, 4.66 in 2012, 4.56 in 2018, 4.48 in 2024, and 4.41 in 2030 when the utilization ratio of hydrogen is zero. If the utilization ratio of hydrogen from non-fossil fuels is lower than 100%, the level of global peace is less than 1 and the reasons for global unrest increase. Therefore, it is beneficial to encourage the utilization of hydrogen from non-fossil fuels in place of fossil fuels. The highest level of global peace is attained when 100% of hydrogen from non-fossil fuels is used in place of fossil fuels. Some advantages of having the highest level of global peace are:

- Lifetimes of fossil fuel reserves are extended and real fossil fuel prices, consequently, can be held constant or reduced relative to present prices.
- Environmental effects from using fossil fuels are reduced or prevented because of the utilization of hydrogen from renewable energy sources and technologies.
- Technological developments based on hydrogen from nonfossil fuels increase and the requirement of technologies based on fossil fuels decrease.
- Living standards are probably higher than at present due in part to the increased consumption of the technologies related to hydrogen from sustainable green energy sources.
- Pressures to discover energy sources reduce because hydrogen can be abundantly produced and conflicts for energy supplies subside.

To increase global peace, the relationship of hydrogen to renewable energy sources needs to be understood, as does the importance of producing hydrogen from renewable energy sources.

Figure 10 describes routes using sustainable green energy sources for green power production. It is expected that the utilization of sustainable green energy sources will reduce the negative energy-related environmental effects such as global climate change and emissions of CO, CO_2, NO_x, SO_x, non-methane hydrocarbons, and particulate matter. In this regard, some potential solutions to decreasing the global unrest associated with the harmful pollutant emissions have evolved, including:[3]

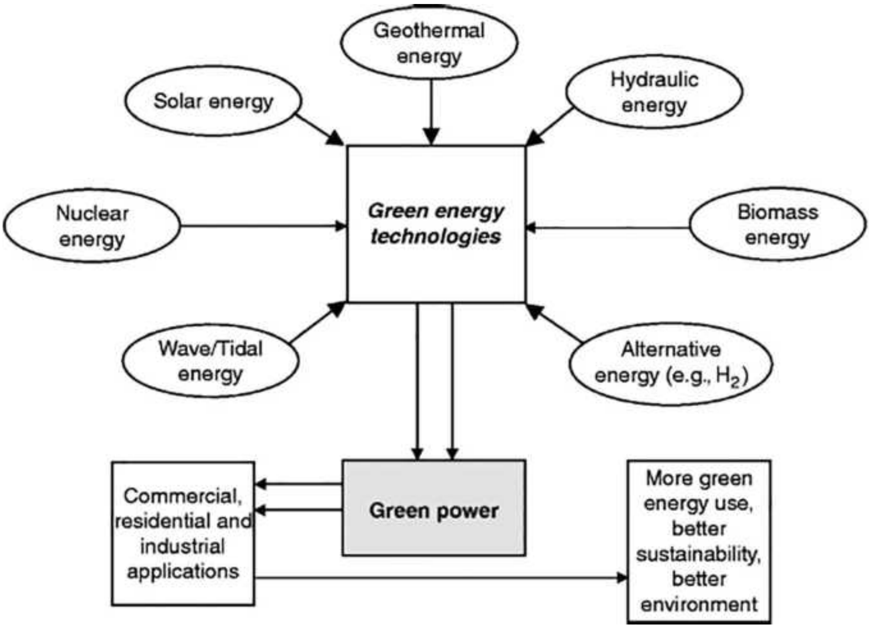


FIGURE 10 Routes for green power production from sustainable green energy sources.

- Education and training for a sustainable future
- Renewable energy technologies
- Energy conservation (efficient energy utilization)
- Cogeneration and district heating
- Energy storage technologies
- Alternative energy dimensions for transport
- Energy source switching from fossil fuels to environmentally benign energy forms
- Coal cleaning technologies
- Optimum monitoring and evaluation of energy indicators
- Policy integration
- Recycling
- Process change and sectoral shiftment
- Acceleration of forestation
- Carbon or fuel taxes
- Greener materials substitution
- Promoting public transport
- Changing life styles
- Increasing public awareness

Considering the above explanations, the following important remarks can be extracted. Fossil fuel consumption and green energy consumption are expected to reach 13807.2 and 2694.9 M, respectively, by the year 2050. This increase indicates that humans will still be dependent on fossil fuels. Based on the projected data, the green energy consumption ratio expects that the green energy utilization ratio will reach almost 16.33% and the fossil fuel utilization ratio will decrease to almost 83.67% in 2050. If the increase of fossil fuel consumption continues in this manner, it is likely that the world will be affected by many negative problems due to fossil fuels. More utilization of fossil fuels will harm world stability and increase local and global environmental problems, resulting in increasing global unrest. It is thus suggested that the utilization of fossil fuels should be reduced, and fossil-based technologies should be gradually converted to green-energy-based technologies.

Case 1 gives better results than Case 2 and Case 3. Therefore, for a higher green energy impact ratio in practice, Case 1 should be applied to increase the green energy sustainability ratio depending on the green energy strategies. Moreover, Case 1 gives the best results of the green energy–based sustainability ratio depending on the green energy impact ratio and the green energy utilization ratio.

The approximate quantified measures developed for level of global peace (ranging between 0 and 1) and level of global unrest (ranging between 1 and ∞)expressions can help understand and measure levels of global unrest and global peace. The highest level of global peace occurs when $GP=1$ and correspondingly, the lowest level of global unrest when $GU=1$, and efforts to increase global peace and stability should cause the values of GP and GU to shift towards these limiting cases. Hydrogen from non-fossil fuels can replace oil, coal, and natural gas to reduce the level of global unrest.

Sustainable green energy strategies are definitely required to ensure the global stability by reducing the harmful effects of fossil-based energy consumption. So, it is suggested that the importance of green energy and technologies that probably reduce world problems and achieve a sustainable energy system should be emphasized with consideration of the sustainable energy strategies. Moreover, a transition to a green energy–based economy should be encouraged and developed countries in particular should increase investments in green energy and technologies. Progress of green energy and technologies is based on sustainable green energy strategies for future green energy scenarios. The foremost factor that will determine the specific role of green energy and technologies will likely be energy demand. In order to balance the energy demand now and in the future, it is suggested that sustainable green energy sources and technologies be taken into consideration to increase the sustainable development in a country.

Conclusion

Green energy for sustainable development has been discussed and some key parameters to increase green energy–based sustainability and the global peace level have been presented. The effects of technological, sectoral, and practical application impact ratios on the green energy impact ratio and the green energy-based sustainability ratio are examined thoroughly. In addition, the key role of hydrogen—one of the most green energy carriers for the future—is discussed in terms of global unrest and peace. Accordingly, sustainable green energy and technologies are definitely required to ensure global stability by reducing the harmful effects of fossil-based energy consumption. The most important scenario that encourages the transition to green energy and technologies and promotes green energy–based technologies is to supply the required incentives and interactions among the countries, scientists, researchers, societies, and all. Therefore, the investments in green energy supply should be, for the future of world nations, encouraged by governments and other authoritative bodies who, for strategic reasons, wish to have a green alternative to fossil fuels.

References

1. Midilli, A.; Ay, M.; Dincer, I.; Rosen, M.A. On hydrogen and hydrogen energy strategies: I: current status and needs. Renewable Sustainable Energy Rev. **2005**, *9* (3), 255–271.
2. Midilli, A.; Dincer, I.; Ay, M. Green energy strategies for sustainable development. Energy Policy **2006**, *34* (18), 623–633.
3. Dincer, I. Renewable energy and sustainable development: a crucial review. Renewable Sustainable Energy Rev. **2000**, *4* (2), 157–175.
4. Dincer, I. The role of exergy in energy policy making. Energy Policy **2002**, *30,* 137–149.
5. Dincer, I.; Rosen, A.M. Exergy as a driver for Achieving Sustainability. Int. J. Green Energy **2004**, *1* (1), 1–19.
6. Midilli, A.; Ay, I.; Dincer, M.; Rosen, M. A.; *On Hydrogen Energy Strategy: The Key Role in This Century and Beyond,* Proceedings of The First Cappadocia International Mechanical Engineering Symposium (CMES-04), 1, 211–217, Cappadocia, Turkey, 2004.
7. World Energy Outlook, EO&W, Institute of Shipping Analysis, http://saiweb.sai.se, Available at http://www.world energyoutlook.org/weo/pubs/weo2002/WEO21.pdf, 2002.
8. Workbook, Statistical Review Full Report of World Energy 2005, Available at http://www.bp.com/centres/energy, 2005.
9. Dincer, I. Environmental impacts of energy. Energy Policy **1999**, *27*, 845–854.
10. Dincer, I.; Rosen, A.M. Thermodynamic aspects of renewables and sustainable development. Renewable Sustainable Energy Rev. **2005**, *9*, 169–189.
11. Rosen, M.A.; Scott, D.S. Entropy production and exergy destruction—part I: hierarchy of earth's major constituencies, and part II: illustrative technologies. Int. J. Hydrogen Energy **2003**, *28* (12), 1307–1323.
12. Midilli, A.; Ay, M.; Dincer, I.; Rosen, M.A. On hydrogen and hydrogen energy strategies II: future projections affecting global stability and unrest. Renewable Sustainable Energy Rev. **2005**, *9* (3), 273–287.

15

Ozone Layer

Luisa T. Molina

Introduction

The Earth's ozone layer shields all life from the sun's harmful ultraviolet (UV) radiation. It is mainly located in the lower stratosphere, between 12 and 30 km above Earth's surface. Ozone is a gas that is present naturally in the Earth's atmosphere; it is continuously being made by the action of solar radiation on molecular oxygen, predominantly in the upper stratosphere and at low latitudes; it is also continuously being destroyed throughout the atmosphere by a variety of chemical processes. The ozone abundances in the atmosphere are therefore determined by the balance between chemical production and destruction processes.

The average concentration of ozone in the atmosphere is about 300 parts per billion by volume (ppbv); most of it (~90%) is contained in the stratosphere. Even though it occurs in such small quantities, ozone plays a vital role in sustaining life on Earth by absorbing most of the biologically damaging UV sunlight. However, the fragile ozone layer is being depleted since the late 1970s as a consequence of the emission of human-made chemicals, chlorofluorocarbons (CFCs), to the atmosphere. The CFCs are industrial chemicals that have been used in the past as coolants for refrigerators and air conditioners, propellants for aerosol spray cans, foaming agents for plastics, and cleaning solvents for electronic components, among other uses. They are thought of as "miracle" compounds because they are non-flammable, noncorrosive, and unreactive with most other substances. Ironically, it is their chemical inertness that creates a global scale problem by enabling them to reach the stratosphere, where they decompose, releasing chlorine atoms that deplete the ozone layer.

Observations of the ozone layer itself showed that depletion was indeed occurring; the most dramatic loss was discovered over Antarctica—far from the emitted sources. In response to the likelihood of increasing depletion of ozone in the stratosphere, an international agreement, the Montreal Protocol on Substances that Protect the Ozone Layer, was signed by most national governments of the world calling for an orderly phase out of all ozone-depleting substances (ODSs). Thus, ozone, a trace constituent in the atmosphere, has become an issue of global prominence and the model for international cooperation to protect the environment from unintended consequences of human activities.

While the Montreal Protocol has made great strides in phasing out ODSs, new challenges are emerging. Changes in climate are expected to have an increasing influence on stratospheric ozone in the coming decades. International efforts to protect the ozone layer would require improved understanding of the complex linkages between stratospheric ozone and climate change. Moreover, some new ODSs replacements are extremely powerful global warming gases and represent a potential focal area within the overall climate change challenge. This entry describes the discovery of the ozone depletion phenomenon, the chemicals that cause ozone depletion in the stratosphere, polar ozone destruction processes, impacts of ozone depletion on human health and on ecosystems, international treaties that regulate the ODSs, and the linkage between stratospheric ozone and climate change.

Atmospheric Ozone

Ozone was discovered by the German chemist Christian Schoenbein in 1840 while observing an electrical discharge; he noted its distinctively pungent odor and named it "ozone, " which means "smell" in Greek. An ozone (O_3) molecule is made of three oxygen atoms, instead of two of the normal oxygen molecule (O_2), which makes up 21% of the air we breathe. Ozone is found mainly in two regions of the Earth's atmosphere (Figure 1):

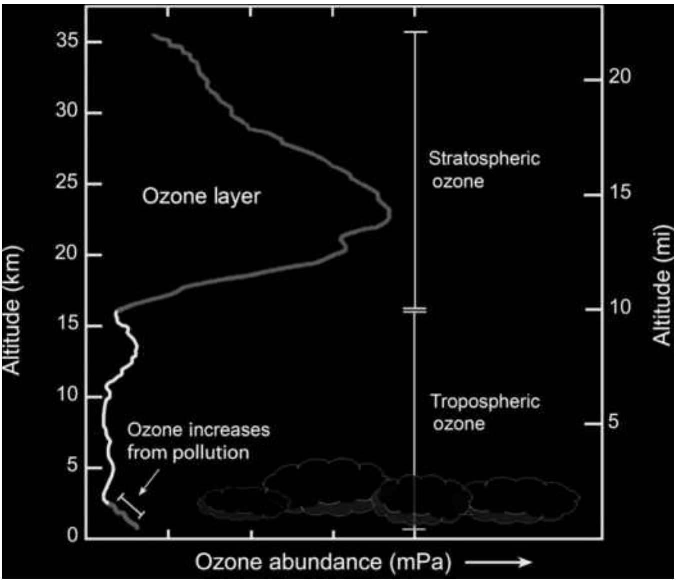


FIGURE 1 Typical atmospheric ozone profile. Ozone abundances are shown here as the pressure of ozone at each altitude using the unit "milli-Pascals" (mPa) (100 million mPa = atmospheric sea-level pressure).

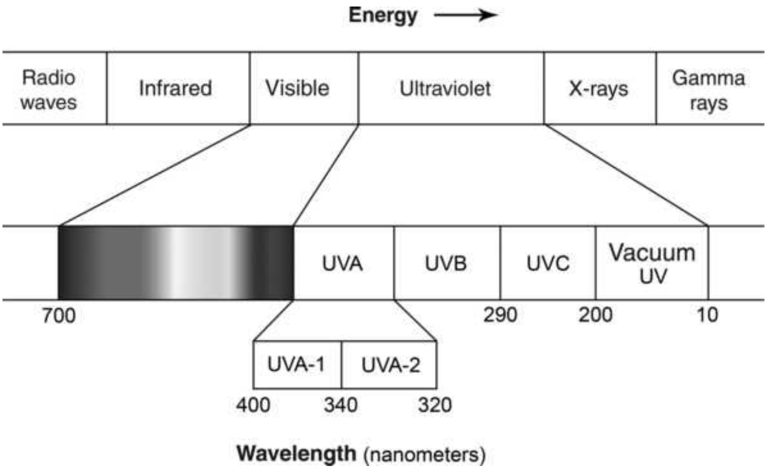


FIGURE 2 Spectrum of electromagnetic radiation.

1. Stratosphere, from about 10–16 km above Earth's surface up to 50 km. The ozone in this region is commonly known as the ozone layer.
2. Troposphere, the lowest region of the atmosphere, between Earth's surface and the stratosphere. Some of the tropospheric ozone is generated by atmospheric photochemical reactions (smog), and some is transported from the stratosphere.

The ozone molecules in the stratosphere and the troposphere are chemically identical; however, they have very different roles in the atmosphere and very different effects on humans and the biosphere. Stratospheric ozone—sometimes referred to as "good ozone"—plays a beneficial role by absorbing most of the biologically damaging UV sunlight (UV-B), allowing only a small amount to reach the Earth's surface (Figure 2). At the Earth's surface, ozone is a key component of urban smog and is harmful to human health, agriculture, and ecosystems; hence, it is often referred as "bad ozone." Furthermore, because ozone is a powerful greenhouse gas, increase in tropospheric ozone contributes to global warming.[1] Thus, depending on where ozone resides, it can have positive or negative impacts on human well-being and the environment.[2]

Origin of Ozone

The fundamental ozone formation–destruction mechanism consists of the following reactions, suggested initially by the British physicist Sidney Chapman[3] in the 1930s:

$$O_2 + \text{UV light} \rightarrow O + O \tag{1}$$

$$O + O_2 + M \rightarrow O_3 + M \tag{2}$$

$$O_3 + \text{UV light} \rightarrow O + O_2 \tag{3}$$

$$O + O_3 \rightarrow O_2 + O_2 \tag{4}$$

Molecular oxygen absorbs solar radiation at ~200 nm and releases oxygen atoms (Reaction 1), which rapidly combine with oxygen molecules to form ozone (Reaction 2). Ozone absorbs solar radiation very efficiently at wavelengths ~200 to 300 nm.[4] This absorption process leads to the decomposition of ozone, producing oxygen atoms (Reaction 3), which in turn regenerate the ozone molecule by Reaction 2. Thus, the net effect of Reactions 2 and 3 is the conversion of solar energy into heat, without the net loss of ozone. This process leads to an increase of temperature with altitude, which is the feature that gives rise to the stratosphere; the inverted temperature profile is responsible for its large stability toward vertical movements (Figure 3). In contrast, the troposphere is characterized by decreasing temperature with altitude because of the heating effects from the absorption of the sun's energy at the Earth's surface. Because hot air rises, this causes rapid vertical mixing so that chemical substances emitted on the ground can rise to the tropopause (transition zone between troposphere and stratosphere) in a matter of days; they are also dispersed horizontally throughout the troposphere on the time scale of weeks to months by winds and convection. However, once they reach the stratosphere, the transport time scales in the stratosphere are much slower and mixing in the stratosphere can take months to years. Movement of air between the troposphere and stratosphere is very slow compared to the movement of air within the troposphere itself. However, this small air exchange is an important source of ozone from the stratosphere to the troposphere.

Most of the time, oxygen atoms react with molecular oxygen to make ozone (Reaction 2), but occasionally they destroy ozone (Reaction 4). The overall amount of ozone is determined by a balance between the production and the removal processes. Models based only on the Chapman's mechanism were found to overpredict stratospheric ozone levels; thus, there are other reactions that contribute to the destruction of ozone.

In the early 1970s, Crutzen[5] suggested that trace amounts of nitrogen oxides ($NO_x = NO + NO_2$) formed in the stratosphere through the decomposition of nitrous oxide (N_2O), which originates from soil-borne microorganisms, control the ozone abundance through the following catalytic cycle:

$$NO + O_3 \rightarrow NO_2 + O_2 \tag{5}$$

$$NO_2 + O \rightarrow NO + O_2 \tag{6}$$

$$\frac{O_3 + \text{UV light} \rightarrow O + O_2 \left(\text{for Reaction 3}\right)}{\text{Net: } 2\,O_3 \rightarrow 3\,O_2}$$

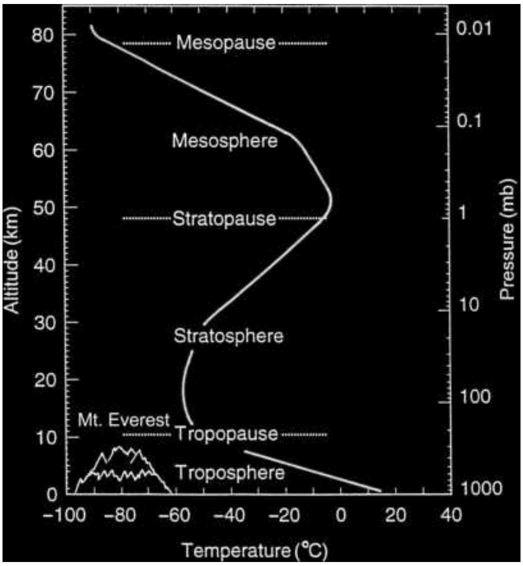


FIGURE 3 Typical variation of atmospheric temperatures and pressures with altitude.

The term “catalyst” refers to a compound that reacts with one or more reactants to form intermediates that subsequently give the final reaction product, in the process regenerating the catalyst. In the above cycle, the species NO (nitric oxide) and NO_2 (nitrogen dioxide) are still present after these three reactions have occurred, but two molecules of ozone have been destroyed. These species have an odd number of electrons; they are free radicals and are chemically very reactive. Although the concentration of NO and NO_2 is small (several ppbv), each radical pair can destroy thousands of ozone molecules before being temporarily removed, mainly by reaction with hydroxyl (OH) radical to form nitric acid:

$$OH + NO_2 \rightarrow HNO_3 \tag{7}$$

Independently, Johnston[6] suggested that the NO_x chain reaction can be initiated by the direct release of NO_x in the exhaust of supersonic transport (SST) aircraft and could disturb the delicate natural balance between ozone formation and destruction.

In the troposphere, NO_x, together with volatile organic compounds in the presence of sunlight, are the ingredients for the photochemical formation of ground-level ozone. Thus, NO_X plays a dual role, destroying or generating O_3 depending on the altitude.

Other free radicals that destroy stratospheric ozone are OH and HO_2, derived from the water molecule:

$$OH + O_3 \rightarrow HO_2 + O_2 \tag{8}$$

$$\begin{array}{l} HO_2 + O_3 \rightarrow OH + 2O_2 \\ \hline \text{Net: } 2\,O_3 \rightarrow 3\,O_2 \end{array} \tag{9}$$

Chlorine atoms are also very efficient catalysts for ozone destruction and may proceed in a similar cycle[7–9] as will be discussed below. Small amounts of chlorine compounds of natural origin exist in the stratosphere; the most important source is methyl chloride (CH_3Cl), which is emitted mainly from oceanic and terrestrial ecosystems. Most of the CH_3Cl is destroyed in the troposphere, but a few percent reaches the stratosphere. There are also large natural sources of inorganic chlorine compounds at the Earth’s surface, e.g., NaCl and HCl from the oceans; however, they are water soluble and are removed efficiently from the atmosphere by clouds and rainfall.

Measurements and Distribution of Stratospheric Ozone

In the 1920s, the British meteorologist G.M.B. Dobson developed the Dobson spectrophotometer that measures sunlight at two UV wavelengths: one that is strongly absorbed by ozone and one that is weakly absorbed. The difference in light intensity at the two wavelengths provides the total ozone above the location of the instrument.[10] In 1957, a global network of ground-based, total ozone observing stations was established as part of the International Geophysical Year; currently, there are about 100 sites distributed throughout the world. Ozone concentrations are also being routinely measured by a variety of instruments on balloons, aircraft, and satellites. One of the most commonly used units for measuring ozone concentration is called “Dobson unit” (DU), which is a measure of how much ozone is contained in a vertical column of air. The average amount of ozone in the atmosphere is about 300 DU, equivalent to a layer 3 mm thick. By comparison, if all of the air in a vertical column that extends from the ground up to space were collected and squeezed together at 0°C and 1 atm pressure, that column would be 8 km thick.

As shown in Figure 4, the distribution of total ozone over the globe varies with latitude, longitude, and season.[11] The variations are caused by large-scale movements of stratospheric air and the chemical production and destruction of ozone. In general, the total ozone is highest in the polar regions and lowest at the equator. In the Northern Hemisphere, the ozone layer is thicker during spring and thinner during autumn.

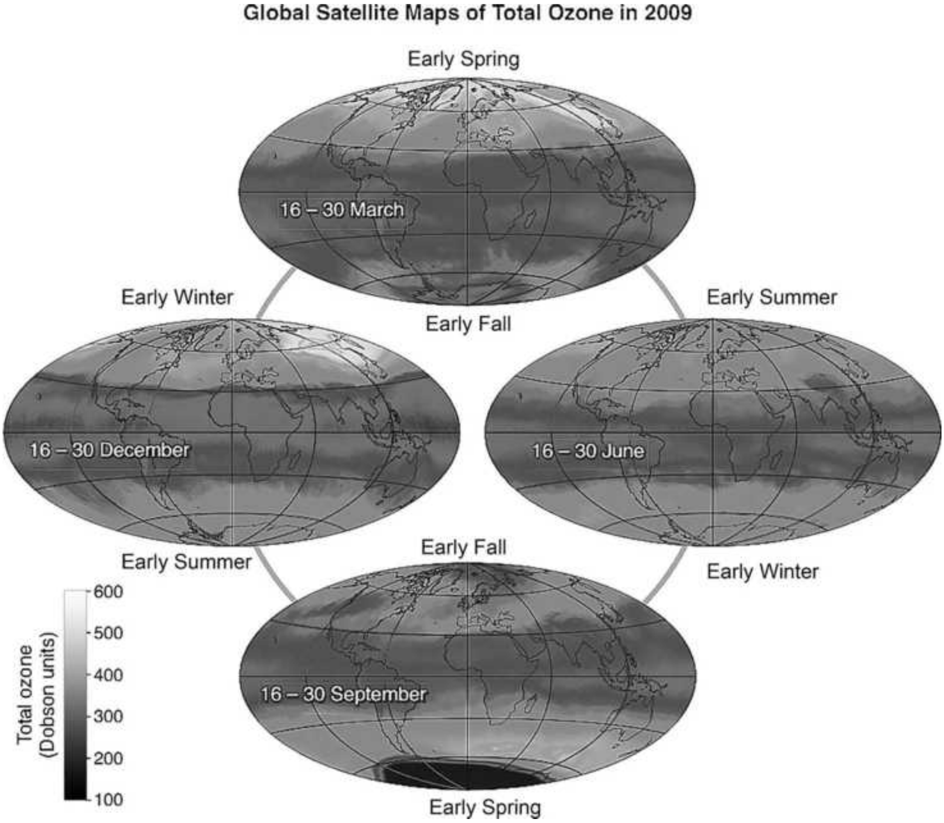


FIGURE 4 Global satellite maps of total ozone in 2009 showing the variation with latitude, longitude, and season. The variations are demonstrated here with 2-week averages of total ozone in 2009 as measured with a satellite instrument. Total ozone shows little variation in the tropics (20°N–20°S latitudes) over all seasons. Total ozone outside the tropics varies more strongly with time on a daily to seasonal basis as ozone-rich air is moved from the tropics and accumulates at higher latitudes. The low total ozone values over Antarctica in September constitute the "ozone hole" in 2009. Since the 1980s, the ozone hole in late winter and early spring represents the lowest values of total ozone that occur over all seasons and latitudes.
Source: Fahey and Hegglin).[11]

CFCS and the Ozone Layer

CFCs: The Miracle Compounds

In the 1930s, American mechanical engineer and chemist Thomas Midgley[12,13] invented the CFCs during a search for nontoxic and non-flammable substances that could be used as coolants in home refrigerators and air conditioners. The CFCs are compounds that contain only chlorine, fluorine, and carbon; they are also known under trademarks such as Freon (DuPont) and Genetron (Allied Signal).

The two important properties that make the CFCs commercially valuable are their volatility (they can be readily converted from a liquid to a vapor and vice versa) and their chemical inertness (they are stable, nontoxic, and nonflammable). The CFCs were thought of as miracle compounds and soon replaced the toxic ammonia and sulfur dioxide as the standard cooling fluids. Subsequently, the CFCs found uses as propellants for aerosol sprays, blowing agents for plastic foam, and cleaning agents for electronic components. All of these activities doubled the worldwide use of CFCs every 6 to 7 years and eventually reached over 700,000 metric tons annually by the early 1970s.

CFCs and the Destruction of Stratospheric Ozone

In the early 1970s, James Lovelock showed that trichlorofluoromethane (CCl_3F or CFC-11) was present in the air over Ireland using a newly developed electron capture detector.[14] Subsequently, Lovelock and co-workers[15] detected measurable levels of CCl_3F in the atmosphere over the South and North Atlantic and concluded that the CFCs were carried over by large-scale wind motions. Their interest in CFCs lies in the potential use of these compounds as inert tracers; they did not expect the CFCs to present any conceivable harm to the environment.

In 1973, Molina and Rowland[7,8,16–18] decided to investigate the ultimate fate of these new miracle compounds upon their release to the atmosphere. After carrying out a systematic search of chemical and physical processes that might destroy the CFCs in the lower atmosphere, they concluded that these compounds would break up in the middle stratosphere (~25–30 km) by solar UV radiation.

Because the CFCs are chemically inert and practically insoluble in water, they are not removed by the common cleansing mechanisms that operate in the lower atmosphere. Furthermore, the CFC molecules are transparent from 230 nm through the visible wavelengths; they are effectively protected below 25 km by the stratospheric ozone layer that shields the Earth's surface from UV light. Instead, they rise into the stratosphere, where they are eventually destroyed by the short-wavelength (~200 nm) solar UV radiation to yield radicals that can destroy stratospheric ozone through a catalytic process.[7] Because transport into the stratosphere is very slow, the atmospheric lifetime for the CFCs is about 50 to 100 years.

The destruction of CFCs by high-energy solar radiation leads to the release of chlorine atoms, as shown in Reaction 12 for CCl_3F:

$$CCI_3F + \text{UV light} \rightarrow CI + CCI_2F \tag{10}$$

The chlorine atoms attack ozone within a few seconds and are regenerated on a time scale of minutes; the net result is the conversion of one ozone molecule and one oxygen atom to two oxygen molecules.

$$CI + O_3 \rightarrow CIO + O_2 \tag{11}$$

$$\begin{array}{c} CIO + O \rightarrow CI + O_2 \\ \hline \text{Net: } O + O_3 \rightarrow O_2 + O_2 \end{array} \tag{12}$$

In the above cycle, chlorine acts as a catalyst for ozone destruction because Cl and ClO react and are reformed, and ozone is simply removed. Oxygen atoms are formed when solar UV radiation reacts with ozone and oxygen molecule (Reactions 1 and 3). It is estimated that one Cl atom can convert 100,000 molecules of ozone into oxygen molecules before that chlorine becomes part of a less reactive compound, e.g., by reaction of ClO with HO_2 or NO_2 to produce hypochlorous acid (HOCl) or chlorine nitrate ($ClONO_2$), respectively, or by reaction of the Cl atom with methane (CH_4) to produce the relatively stable hydrogen chloride (HCl):

$$CIO + HO_2 \rightarrow HOCI + O_2 \tag{13}$$

$$CIO + NO_2 \rightarrow CINO_2 \tag{14}$$

$$CI + CH_4 \rightarrow HCI + CH_3 \tag{15}$$

The chlorine-containing product species HCl, $ClONO_2$, and HOCl function as temporary inert reservoirs: they are not directly involved in ozone depletion, but they are eventually broken down by reaction with other free radicals or by absorption of solar radiation, thus returning chlorine to its catalytically active free radical form. At low latitudes and in the upper stratosphere, where the

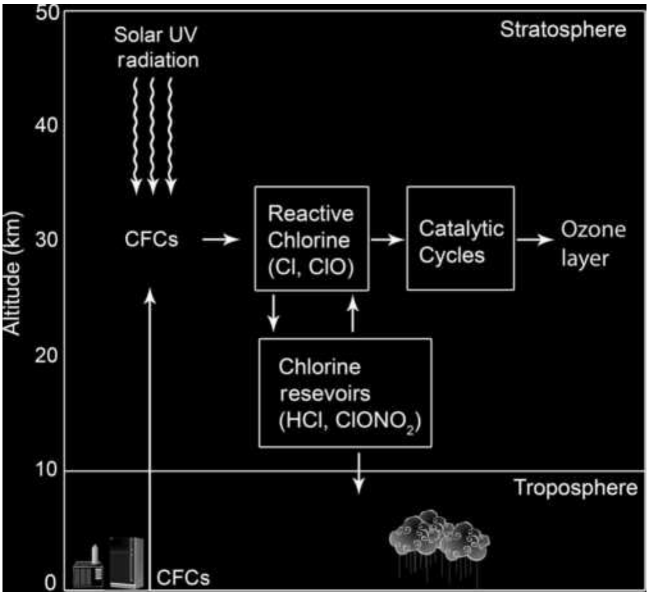


FIGURE 5 Schematic representation of the CFC-ozone depletion hypothesis.

formation of ozone is fastest, a few percent of the chlorine is in this active form; most of the chlorine is in the inert reservoir form, with HCl being the most abundant species. The temporary chlorine reservoirs remain in the stratosphere for several years before returning to the troposphere, where they are rapidly removed by rain or clouds. A schematic representation of these processes is presented in Figure 5.

Other ODSs

Besides CFCs, there are other chlorine substances from human activities that destroy ozone in the stratosphere. Carbon tetrachloride and methyl chloroform are important ODSs that are used as fire extinguisher, cleaning agents, and solvents.

Another category of ODSs contains bromine. There are industrial sources of brominated ODSs as well as natural ones; the most important are the halons and methyl bromide (CH_3Br). Halons are halocarbon gases containing carbon, bromine, fluorine, and (in some cases) chlorine; they are produced industrially as fire extinguishers. Methyl bromide is both natural and human-made; it is used as an agricultural fumigant. These sources release bromine to the stratosphere at pptv levels, compared with ppbv for chlorine. On the other hand, bromine atoms are about 60 times more efficient than chlorine atoms for ozone destruction on an atom-per-atom basis;[19–21] a large fraction of the bromine compounds is present as free radicals because the temporary reservoirs are less stable and are formed at considerably slower rates than the corresponding chlorine reservoirs.

Many of the ODSs contain fluorine atoms in addition to chlorine and bromine. However, in contrast to chlorine and bromine, fluorine atoms abstract hydrogen atoms very rapidly from methane and from water vapor, forming the stable hydrogen fluoride (HF), which serves as a permanent inert fluorine reservoir. Hence, fluorine free radicals are extremely scarce and the effect of fluorine on stratospheric ozone is negligible. Halogen source gases that contain fluorine and no other halogens are not classified as ODSs. An important category is the hydrofluorocarbons (HFCs) discussed in the "International Response: Montreal Protocol" section.

The publication of the 1974 Molina–Rowland article[7] stimulated numerous scientific studies, including laboratory studies, computer modeling, and field measurements, to understand the impacts of chlorine and bromine on stratospheric ozone.[20,21] The U.S. National Academy of Sciences issued two reports in 1976, verifying the Molina-Rowland findings.[22,23]

Stratospheric Ozone Depletion

Discovery of the Antarctic Ozone Hole

In 1985, Joseph Farman and coworkers[24] reported that the ozone concentrations recorded at the Halley Bay Observatory in Antarctica has dropped dramatically in the spring months starting in the early 1980s, compared to the data obtained since 1957, when the British Antarctic Survey began ozone measurements using a Dobson Spectrometer. The 1984 October monthly ozone averaged less than 200 DU, about 35% lower than the 300 DU levels recorded in 1957–1958 and on through the 1960s. Farman et al.'s findings were subsequently confirmed by satellite observation from the total ozone mapping spectrometer (TOMS).[25] Furthermore, satellite measurements confirmed that the bulk of the chlorine in the stratosphere is of human origin.[26] Additional measurements from ground-based Dobson instruments[27] and from satellites indicate that the extent of ozone depletion over Antarctica in the spring months continued to increase after 1985, with concentrations as low as 85 to 95 DU reported from some of the polar stations.

Measurements show that ozone was also being depleted in the Northern Hemisphere, particularly at high latitudes and in the winter and spring months. Examination of the ozone records shows that significant changes have also occurred in the lower stratosphere at mid- latitudes.[20,21] Figure 6 shows an acute drop in total atmospheric ozone during October in the early and mid-1980s measured by instruments from the ground and from a satellite.[28]

The discovery of the depletion of ozone over Antarctica—the ozone hole—was not predicted by the atmospheric scientists (see Figure 7). The large magnitude of the depletion suggests that the stratospheric

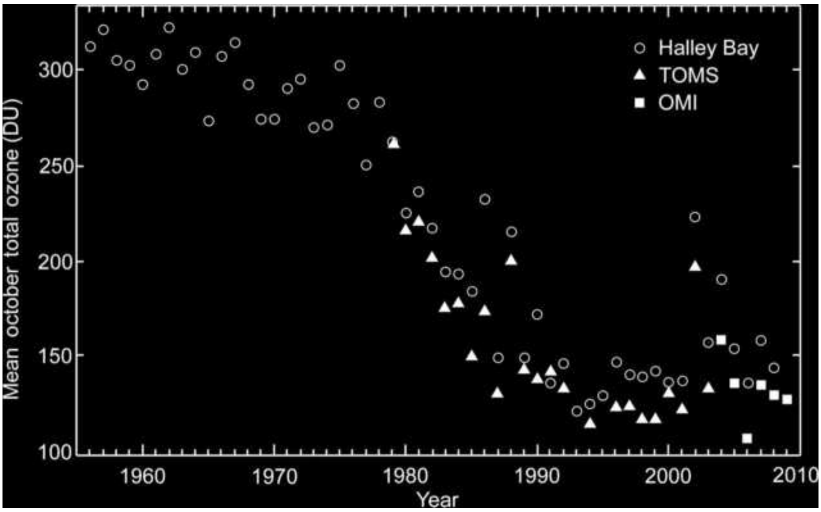


FIGURE 6 Average total amount of ozone measured in October over Antarctica. Instruments on the ground (at Halley Bay) and high above Antarctica (the TOMS and ozone monitoring instrument [OMI] measured an acute drop in total atmospheric ozone during October in the early and middle 1980s).
Source: Adapted from NASA Ozone Hole Watch.[28]

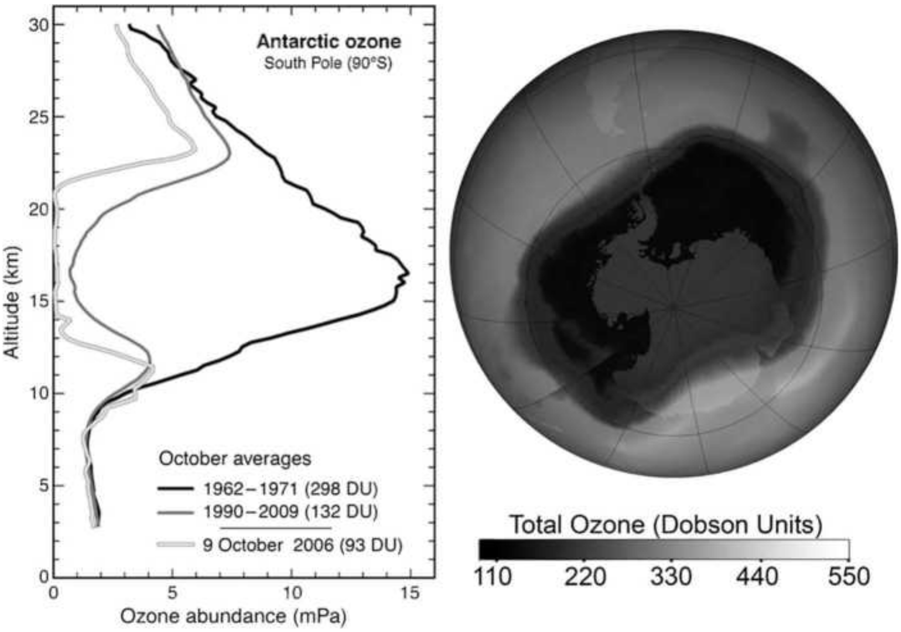


FIGURE 7 Antarctic ozone hole. Left panel: Vertical distributions of Antarctic ozone. A normal ozone layer was observed to be present between 1962 and 1971. In more recent years, as shown here for October 9, 2006, ozone is almost completely destroyed between 14 and 21 km in the Antarctic in spring (Source: Adapted from Fahey and Hegglin[11]). Right panel: The darker shaded regions over the Antarctic continent show the severe ozone depletion or ozone hole on October 9, 2006, measured by satellite instrument. The hole reached 26.2 million km^2, the greatest extent recorded in the Antarctic.
Source: NASA Ozone Hole Watch.[28]

ozone is influenced by processes that had not been considered previously. Researchers all over the world raced to develop plausible explanation; the cause of this depletion soon became very clear. Laboratory experiments, field measurements over Antarctica, and model calculations showed unambiguously that the ozone hole over Antarctica can indeed be traced to the industrial CFCs.[20,21]

Polar Ozone Chemistry

Characteristics of the Polar Regions

ODSs emitted at Earth's surface are transported over great distances to the stratosphere by atmospheric air motions and are present throughout the stratospheric ozone layer. Yet, the most dramatic ozone depletion was over Antarctica—the ozone hole—far from the emitted sources. A major reason why an Antarctic ozone hole of the observed extent could happen is because of the unique atmospheric and chemical conditions that exist there. The very low winter temperatures in the Antarctic stratosphere cause polar stratospheric clouds (PSCs) to form. Special reactions that occur on PSCs, combined with the relative isolation of polar stratospheric air, allow chlorine and bromine reactions to produce the ozone hole in the Antarctica when the sunlight returns in the springtime.

Both polar regions of the earth are cold, primarily because they receive far less solar radiation than the tropics and mid-latitudes do. Moreover, most of the sunlight that does shine on the polar regions is reflected by the bright white surface. Winter temperatures at the North Pole can range from about −45°C to −25°C and summer temperatures can average around the freezing point (0°C).

In comparison, the annual mean temperature at the South Pole is about −60°C in the winter and −28°C in the summer.

What makes the South Pole so much colder than the North Pole is that it sits on top of a very thick ice sheet, which itself sits on the continent of Antarctica. The surface of the ice sheet at the South Pole is more than 2700 m (9000 ft) above sea level; Antarctica is by far the highest continent on the earth. In contrast, the North Pole is at sea level in the middle of the Arctic Ocean, which also acts as an effective heat reservoir.

Stratospheric air in the polar regions is relatively isolated from other stratospheric regions for long periods in the winter months. The isolation comes about because of strong winds that encircle the poles, forming a polar vortex (or polar cyclone), which prevents substantial air masses into or out of the polar stratosphere. This cyclonic circulation strengthens in winter as stratospheric temperatures decrease.[11] All through the long, dark winter, chemical changes occur in polar regions from reactions on PSCs inside the vortex; the isolation preserves those changes until the spring sunlight strikes the stratosphere above the frozen continent in late August. The result is massive ozone destruction inside the vortex forming an ozone hole, as described below. The polar vortex diminishes when the continent and the air above it begin to warm up and ozone-rich air from outside the vortex flows in, replacing much of the ozone that was destroyed.

The Antarctic polar vortex is more pronounced and persistent than its Arctic counterpart with the result that the isolation of air inside the vortex is much more effective in the Antarctic than in the Arctic. In addition to being significantly warmer, the Northern Hemisphere also has numerous mountain ranges and a more active tropospheric meteorology, giving rise to enhanced planetary wave and a less stable Arctic vortex.

Polar Stratospheric Clouds

The conditions in the polar stratosphere are unique in several ways. Firstly, ozone is not generated there because the high-energy solar radiation that is absorbed by molecular oxygen is scarce over the poles. Secondly, the total ozone column abundance at high latitudes is large because ozone is transported toward the poles from higher altitudes and lower latitudes. Thirdly, the prevailing temperatures over the stratosphere above the poles in the winter and spring months are the lowest throughout the atmosphere, particularly over Antarctica. Typically, average daily minimum values are as low as −90°C in July and August over Antarctica and near −80°C in late December and January over the Arctic (Figure 8). Ozone is expected to be rather stable over the poles if one considers only gas-phase chemical and photochemical processes, because regeneration of ozone-destroying free radicals from the reservoir species would occur very slowly at those temperatures. However, another unique feature of the polar stratosphere is the seasonal presence of PSCs (Fig. 9). Different types of liquid and solid PSC particles form when stratospheric temperatures fall below −78°C. As a result, PSCs are often found over large areas of the winter polar regions and over significant altitude ranges. With a temperature threshold of −78°C, PSCs exist in larger areas and longer periods in the Antarctica than in the Arctic.

The stratosphere is normally very dry; water is present only at a level of a few ppmv, comparable to that of ozone itself. Over the poles, a somewhat larger amount of water is present, resulting from the oxidation of methane. Furthermore, the temperature can drop below −85°C over Antarctica in the winter and spring months, leading to the formation of ice clouds. The presence of trace amounts of nitric and sulfuric acids enables the formation of PSCs a few degrees above the frost point (the temperature at which ice can condense from the gas phase); these acids can form cloud particles consisting of crystalline hydrates.

Solomon et al.[29] first suggested that PSCs could play a major role in ozone depletion over Antarctica by promoting the release of photolytically active chlorine from its reservoir species (HCl, $ClONO_2$, and HOCl). This occurs mainly by the following reaction:

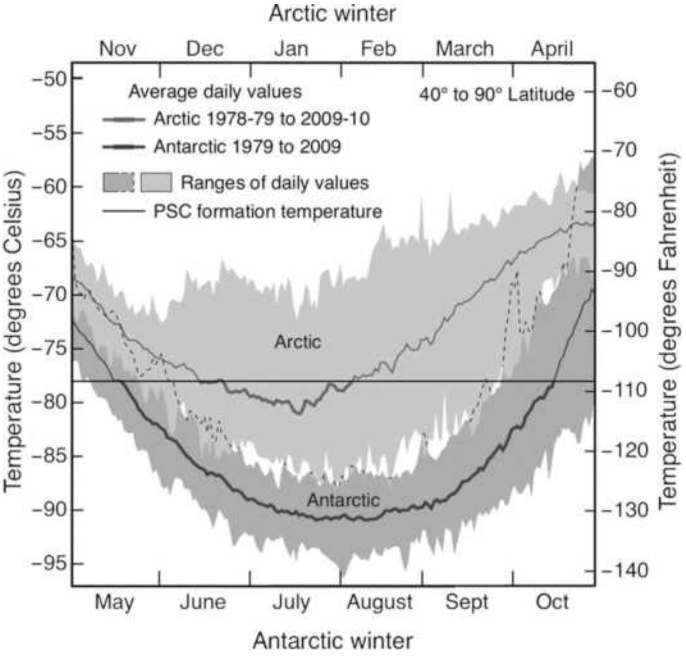


FIGURE 8 Arctic and Antarctic temperatures. Air temperatures in both polar regions reach minimum values in the lower stratosphere in the winter season. Typically, average daily minimum values are as low as −90°C in July and August in the Antarctic and near −80°C in late December and January in the Arctic. PSCs are formed in the polar ozone layer when winter minimum temperatures fall below the formation temperature of about −78°C. Note that the dashed black lines denote the upper limits of the Antarctic temperature range where they overlap with the Arctic temperature range.
Source: Fahey and Hegglin.[11]

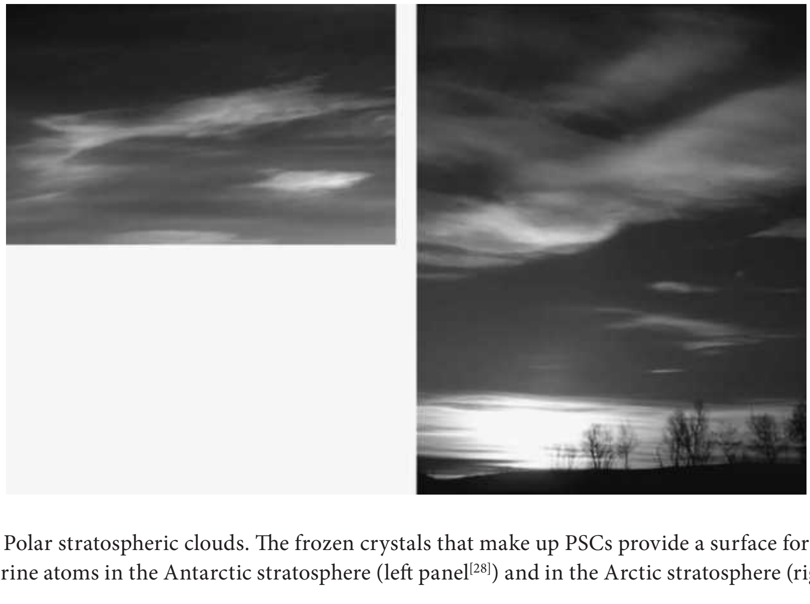


FIGURE 9 Polar stratospheric clouds. The frozen crystals that make up PSCs provide a surface for the reactions that free chlorine atoms in the Antarctic stratosphere (left panel[28]) and in the Arctic stratosphere (right panel[11]).

$$\mathrm{HCI + CIONO_2 \rightarrow CI_2 + HNO_3} \tag{16}$$

$$\mathrm{CI_2} + \text{sunlight} \rightarrow \mathrm{CI + CI} \tag{17}$$

Indeed, laboratory studies have shown that this reaction occurs very slowly in the gas phase;[30] however, it proceeds with remarkable efficiency in the presence of ice surfaces.[31] The product, Cl_2, is immediately released to the gas phase while the other product, HNO_3, remains in the condensed phase. Cl_2 decomposes readily to Cl atoms even with the faint amount of sunlight present over Antarctica in the early spring:

When average temperatures begin increasing by late winter, PSCs form less frequently and their surface conversion reactions produce less ClO. Without continued ClO production, ClO amounts decrease and other chemical reactions re-form the reactive reservoirs, $ClONO_2$ and HCl. The most intense period of ozone depletion will end when PSC temperatures no longer occur.

An important feature for Reaction 16 in the presence of PSCs is the removal of NO_x from the gas phase; the source for these free radicals in the polar stratosphere is nitric acid, which condenses in the cloud particles. NO_x normally interfere with the catalytic ozone loss reactions by scavenging ClO to form chlorine nitrate (Reaction 14). In the absence of NO_x, the fraction of chlorine compounds in the form of free radical is larger and makes it possible for the ozone depletion reaction to occur. These experimental results have been corroborated by other studies.[32] Further laboratory studies[33,34] and theoretical calculations[35] indicate that HCl solvates readily on the ice surface, forming hydrochloric acid. Therefore, chlorine activation reactions on the surfaces of ice crystals proceed through ionic mechanisms analogous to those in aqueous solutions.

Another natural source of ozone depletion is the sulfate aerosols that come from volcanic eruptions. The most recent large eruption was that of Mt. Pinatubo in 1991, which ejected large amounts of sulfur dioxide into the stratosphere, resulting in up to a 10-fold increase in the number of particles available for surface reactions. These particles increased the global ozone depletion by 1%–2% for several years following the eruption.[11]

Polar Ozone Destruction

While the presence of the PSCs explains how chlorine can be released from the inactive reservoir chemicals, it remains unanswered how a catalytic cycle might be maintained to account for the large ozone destruction observed. The ClO_x cycles such as Reactions 11 and 12 are not efficient in the polar stratosphere because they require the presence of free oxygen atoms, which are scarce at high latitudes.

Several catalytic cycles have been suggested as occurring over Antarctica during winter and spring. One of the dominant cycles involving ClO dimer or chlorine peroxide (ClOOCl) was proposed by Molina and Molina.[36] The cycle is initiated by the combination of two ClO radicals forming ClOOCl, which then photolyzes to release molecular oxygen and free chlorine atoms:

$$\mathrm{C1O + C1O + M \rightarrow C1OOC1 + M} \tag{18}$$

$$\mathrm{C1OOC1} + \text{sunlight} \rightarrow \mathrm{C1 + C1OO} \tag{19}$$

$$\mathrm{C1OO + M \rightarrow C1 + O_2 + M} \tag{20}$$

$$\frac{2\left[\mathrm{C1 + O_3 \rightarrow C1O + O_2}\right]\ (\text{for Reaction 11})}{\text{Net: } 2\ \mathrm{O_3} \rightarrow 3\ \mathrm{O_2}}$$

No free oxygen atoms are involved in this cycle. Visible sunlight is required to complete and maintain the cycle; thus, this cycle can occur only in late winter/early spring when sunlight returns to the polar region. Laboratory studies have shown that the photolysis products of ClOOCl are indeed Cl atoms.[37,38]

Another important cycle operating in the polar stratosphere involves the reaction of bromine monoxide (BrO) with ClO suggested by McElroy et al.:[39]

$$\mathrm{BrO+C1O+sunlight \rightarrow C1+Br+O_2} \tag{21}$$

$$\mathrm{Br+O_3 \rightarrow BrO+O_2} \tag{22}$$

$$\underline{\mathrm{C1+O_3 \rightarrow C1O+O_2}\ (\text{for Reaction } 11)}$$
$$\text{Net: } 2\ \mathrm{O_3} \rightarrow 3\ \mathrm{O_2}$$

Field Measurements of Atmospheric Trace Species

Ground-based and aircraft expeditions were launched in the years following the ozone hole discovery to measure trace species in the stratosphere over Antarctica.[40] The results provided strong evidence for the crucial role played by industrial chlorine in the ozone depletion. One of the most convincing evidence was provided by the NASA ER-2 aircraft measurements in 1987, which flew into the Antarctic vortex. The flight data (Figure 10) showed an anticorrelation between ClO measured by Anderson et al.[41] with in situ ozone measurements monitored by Proffitt et al.[42] The results show that the ClO + ClO cycle accounts for about three-quarters of the observed ozone loss, with the BrO + ClO cycle accounting for the rest. Furthermore, NO*x* levels were found to be very low and nitric acid was shown to be present in the cloud particles, as expected from the laboratory studies.

Recent laboratory measurements of the dissociation cross section of ClO dimer[43] and analyses of observation from aircraft and satellites have reaffirmed that polar springtime ozone depletion is caused primarily by the ClO + ClO catalytic ozone destruction cycle, with substantial contributions from the BrO + ClO cycle.[44]

Field measurements have also been conducted in the Arctic stratosphere,[45] indicating that a large fraction of the chlorine is also activated there. Nevertheless, ozone depletion is less severe over the Arctic and is not as localized because the atmosphere above the Arctic is warmer than above the Antarctic; the active chlorine does not remain in contact with ozone long enough and at low enough temperatures to destroy it before the stratospheric air over the Arctic mixes with warmer air from lower latitudes. This warmer air also contains NO_2, which deactivates the chlorine. On the other hand, cold winters can lead to significant ozone depletion—30% or more—over large areas, as described below.[46]

Depletion of the Global Ozone Layer

Global total ozone levels are influenced not only by the concentrations of ODSs but also by atmospheric transport (winds), incoming solar radiation, aerosols (fine particles suspended in the air), and other natural compounds. Global total ozone has decreased beginning in the 1980s, reaching a maximum of about 5% in the early 1990s (Figure 11a). The lowest global total ozone values occurred in the years following the eruption of Mt. Pinatubo in 1991, which resulted in up to a 10-fold increase in the number of sulfuric-acid-containing particles available for surface reaction in the stratosphere, thereby increased global ozone depletion for several years before they were removed from the stratosphere by natural processes. The depletion has lessened since then; the average global ozone for 2005–2009 is about 3.5% below the 1964–1980 average.

Observed total ozone loss varies significantly with latitude on the globe. Figure 11b shows how the 2005–2009 ozone depletion varies with latitude. The ozone loss is very small near the equator and increases with latitude towards the poles. The largest decreases have occurred at high latitudes in both

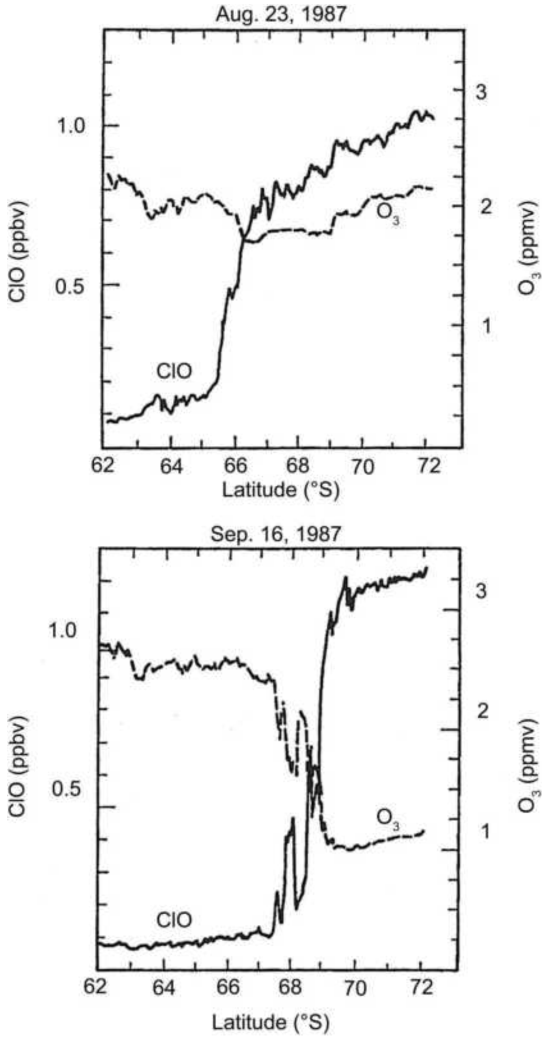


FIGURE 10 Aircraft measurements conducted on August 23 and September 16, 1987, of chlorine monoxide (ClO) by Anderson et al.[41] and of ozone (O_3) by Proffitt et al.[42]
Source: Adapted from Anderson.[41]

hemispheres because of the large winter/spring depletion in the polar regions; the losses are greater in the Southern Hemisphere because of the Antarctic ozone hole. Since the 1980s, Antarctic ozone loss in the springtime has been quite large, covering nearly the entire continent with virtually all of the ozone destroyed between 15 and 20 km. Ozone loss over the Arctic is smaller than its Antarctic counterpart; it is modulated strongly by variability in atmospheric dynamics, transport, and temperature. The degree of spring Arctic depletion is highly variable from year to year, but large Arctic depletion has also been observed recently with the most dramatic occurring in the spring of 2011. Observations over the Arctic region as well as from satellites show an unprecedented ozone column loss comparable to some Antarctic ozone holes. The formation of the "Arctic ozone hole" in 2011 was driven by an anomalously strong stratospheric polar vortex and an unusually long cold period, leading to persistent enhancement of active chlorine and severe ozone loss that exceeded 80% over 18–20 km altitude. This result raises the possibility of yet more severe depletion as lower stratospheric temperatures decrease. More acute Arctic

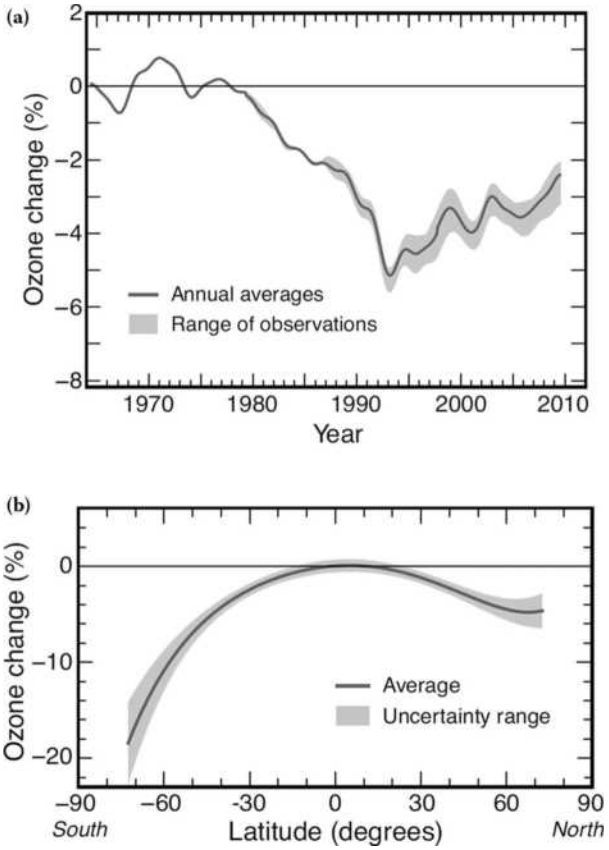


FIGURE 11 Global total ozone changes. Satellite observations show depletion of global total ozone beginning in the 1980s. Panel (a) compares annual averages of global ozone with the average from the period 1964 to 1980 before the ozone hole appeared. Seasonal and solar effects have been removed from the observational data set. On average, global ozone decreased each year between 1980 and 1990. The depletion worsened for a few years after 1991 following Mt. Pinatubo eruption. Panel (b) shows variation of the 2005–2009 depletion with latitude over the globe. The largest decreases have occurred at high latitudes in both hemispheres because of the large winter/spring depletion in polar regions; the losses are greater in the Southern Hemisphere because of the Antarctic ozone hole.

Source: Fahey and Hegglin.[11]

ozone loss could exacerbate biological risks from exposure to increased UV radiation, especially if the vortex shifted over densely populated mid-latitudes, as it did in April 2011.[46]

Ozone depletion is also observed at the mid-latitudes between equatorial and polar latitudes. Total ozone averaged for 2005–2009 is about 3.5% lower in northern mid-latitudes (35°N–60°N) and about 6% lower at southern mid-latitudes (35°S–60°S) compared with the 1964–1980 average (Figure 11b). Chemical destruction processes occurring at midlatitudes contributes to observed depletion in these regions, although it is much smaller than in the polar regions because the amounts of reactive halogen gases are lower and a seasonal increase of the most reactive halogen gases in Antarctic late winter does not occur in mid-latitude regions. Changes in mid-latitude ozone are also affected by changes occurring in the polar regions when the ozone-depleted air over both polar regions is dispersed away from the poles following the vortex breakdown, thus reducing average ozone concentrations at nonpolar latitudes.

Ozone Depletion and Biological Effects

Because ozone absorbs UV radiation from the sun, depletion of the stratospheric ozone layer is expected to lead to increases in the amount of solar UV radiation reaching the Earth's surface, predominantly in the wavelength range of 290 to 320 nanometers (UV-B radiation). UV-B radiation is also partially shielded by clouds, dust, and air pollutants. Large ozone losses in the Antarctica have produced a clear increase in surface UV radiation. Ground-based measurements show that the average spring erythemal (sunburning) irradiance for 1990–2006 is up to 85% greater than the modeled irradiance for 1963–1980, depending on site. The Antarctic spring erythemal irradiance is approximately twice that measured in the Arctic for the same season.

Analyses based on surface and satellite measurements show that erythemal UV irradiance over mid-latitudes has increased since the late 1970.[44] This is in qualitative agreement with the observed decrease in column ozone, although other factors (mainly clouds and aerosols) have influenced long-term changes in erythemal irradiance. Clear-sky UV observations from unpolluted sites in midlatitudes show that since the late 1990s, UV irradiance levels have been approximately constant, consistent with ozone column observations over this period.

The environmental and health effects of stratospheric ozone depletion have been summarized in several assessment reports.[47–49] UV-B radiation can induce acute skin damage in humans, such as sunburn, as well as eye diseases and infectious diseases.[50,51] Human epidemiological studies and animal experiments have established that UV-B radiation is a key risk factor for development of skin cancer, both melanoma and non-melanoma, especially in light-skinned population. Non-melanoma (squamous cell carcinoma and basal cell carcinoma) is the more common form of skin cancer and can be readily treated and is rarely fatal,[51] whereas melanoma is the most dangerous and the leading cause of death from skin cancer. Absorption of strong UV radiation damages the DNA molecule, eventually leading to faulty replication and mutation.[52] Numerous experiments have shown that the cornea and lens of the eye can also be damaged by UV-B radiation, and that chronic exposure to this radiation increased the likelihood of certain cataracts.[49] Studies in human subjects show that exposure to UV-B radiation can suppress proper functioning of the body's immune system. Animal experiments indicate that overexposure to UV-B radiation decreases the immune response to skin cancers and some infectious agents.[53,54]

Terrestrial plants can also be affected by UV-B radiation, although the response varies to a large extent among different species. In addition to plant growth, the changes induced by UV radiation can be indirect, for example, by affecting the timing of developmental phases or the allocation of biomass to the different parts of the plant.[55] Aquatic ecosystems can also be damaged by UV-B radiation; for example, there is evidence for impaired larval development and decreased reproductive capacity in some amphibians, shrimp, and fish.[56] There is also direct evidence of UV-B effects under the ozone hole on the productivity of natural phytoplankton communities in Antarctic waters.[57]

International Response: Montreal Protocol

Following the publication of the Molina–Rowland article,[7] the U.S. National Academy of Sciences issued two reports in 1976 stating that the atmospheric sequence outlined by Molina and Rowland was essentially correct.[22,23] The United States, Canada, Norway, and Sweden responded in late 1970s by banning the sale of aerosol spray cans containing CFCs; this caused a temporary halt in the growing demands for CFCs. However, worldwide use of the chemicals continued and the production rate began to rise again.

The discovery of the massive ozone losses in Antarctica in 1985 spurred a rush of scientific research activities, leading to improved understanding of stratospheric chemistry and the evolution of the ozone layer. These new scientific developments have provided the foundation for the critical policy decisions that followed.

In 1985, under the auspices of the United Nations Environment Programme (UNEP), 20 nations signed the Vienna Convention for the Protection of the Ozone Layer.[58] In September 1987, the recognition that CFC use was increasing and the mounting scientific evidence that this increase would cause large ozone depletions led to an international agreement limiting the production of CFCs.[58] This agreement, the Montreal Protocol on Substances that Deplete the Ozone Layer, initially called for a reduction of only 50% in the manufacture of CFCs by the end of the century. In view of the scientific evidence that emerged in the following years, the initial provisions were strengthened through the London (1990), Copenhagen (1992), Montreal (1997), and Beijing (1999) amendments as well as several adjustments by both controlling additional ODSs and by moving up the date by which already controlled substances must be phased out. At the 19th Meeting of the Parties to the Montreal Protocol in 2007, the Parties agreed to adjust their commitments related to the phase out of HCFCs.[58] Fig. 12 shows the projected abundance of ODSs in the stratosphere according to the provisions of the Montreal Protocol and subsequent amendments.[44]

The Montreal Protocol is now more than 20 years old and has been ratified by 196 countries, although not all the parties have ratified the subsequent amendments. It is widely regarded as one of the most effective multilateral environmental agreements in existence. The production of CFCs in industrialized countries was phased out at the end of 1995, and other compounds such as the halons, methyl bromide, carbon tetrachloride, and methyl chloroform (CH_3CCl_3) were also regulated. Developing countries were allowed to continue CFC production until 2010, to facilitate their transition to the newer CFC-free technologies. An important feature of the Montreal Protocol was the establishment of a funding mechanism to help these countries meet the costs of complying with the protocol and with its subsequent amendments.[58]

A significant fraction of the former CFC usage is being dealt with by conservation and recycling. Some of the former use of CFCs is being temporarily replaced by hydrochlorofluorocarbons (HCFCs)—compounds that have similar physical properties to the CFCs, but their molecules contain hydrogen atoms and are less stable in the atmosphere. A large fraction of the HCFCs released industrially reacts

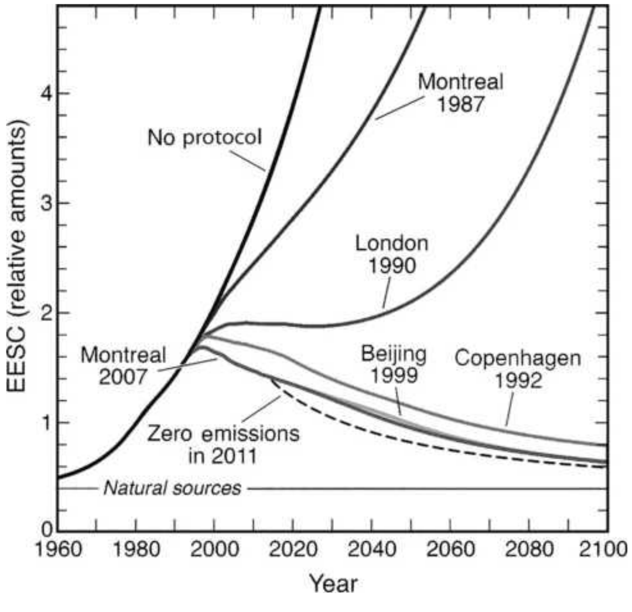


FIGURE 12 Measured and projected chlorine concentrations in the stratosphere according to the provisions of the Montreal Protocol and subsequent amendments. EESC is the equivalent effective stratospheric chlorine.
Source: Fahey and Hegglin.[11]

in the lower atmosphere with the OH radical before reaching the stratosphere, forming water and an organic free radical that rapidly photo-oxidizes to yield water-soluble products, which are then removed from the atmosphere mainly by rainfall. Although HCFCs are more ozone friendly than CFCs, they still destroy some ozone. They are now also regulated under the Montreal Protocol; the concentrations of HCFCs are projected to grow for another two decades before decreasing.

Some HFCs, which do not contain chlorine atoms, are now being used as CFC replacements; they are ozone friendly because fluorine forms stable compounds in the stratosphere. However, they have the potential to contribute to global warming. HFCs are now regulated under the Kyoto Protocol, an agreement under the United Nations Framework Convention on Climate Change.

About half of the CFC usage has been replaced by not-in-kind alternatives; for example, CFC-113—used extensively as a solvent to clean electronic components—has been phased out by CFC-free cleaning technologies such as soap and water or terpene-based solvents; there are also new technologies to manufacture clean electronic boards. Other examples include the use of stick or spray pump deodorants to replace CFC-12 aerosol deodorants and the use of mineral wool to replace CFC, HFC, or HCFC insulating foam.

Overall, the provisions of the Montreal Protocol have been successfully enforced. Atmospheric measurements indicate that the abundance of chlorine contained in the CFCs and other halocarbons has declined in response to the Montreal Protocol regulations.[44] On the other hand, because of the long lifetime of the CFCs in the atmosphere, relatively high chlorine levels in the stratosphere—with the consequent ozone depletion—are expected to continue well into the 21st century.

The total tropospheric abundance of chlorine from ODSs and methyl chloride had declined in 2008 to 3.4 parts per billion (ppb) from its peak of 3.7 ppb between 1992 and 1994. However, the rate of decline in total tropospheric chlorine was only two-thirds as fast as was expected because of the increase in the HCFC abundances. The rapid HCFC increases are coincident with increased production in the developing countries, particularly in East Asia. The rate of decline of total tropospheric bromine from controlled ODSs was close to that expected and was driven by changes in methyl bromide.[44]

By the middle of the 21st century, the amounts of halogens in the stratosphere are expected to be similar to those present in 1980 prior to the onset of the ozone hole. However, the influence of climate change could accelerate or delay ozone recovery.[44]

Stratospheric Ozone and Climate Change Linkage

Greenhouse Gases and Climate Change

The Earth absorbs energy from the Sun, and also radiates energy back into space. However, much of this energy going back to space is absorbed by greenhouse gases occurring naturally in the atmosphere, such as CO_2, CH_4, water vapor, N_2O, and ozone. Because the atmosphere then radiates most of this energy back to the Earth's surface, our planet is warmer than it would be if the atmosphere did not contain these gases. This is the natural "greenhouse effect," which maintains the surface of the Earth at a temperature that is suitable for life as we know it today. However, this natural greenhouse effect has intensified since the start of the industrial era as human activities emit more greenhouse gases to the atmosphere, including industrial compounds such as CFCs, HFCs, perfluorocarbons (PFCs), and sulfur hexafluoride (SF_6), resulting in a shift in the radiative balance of the Earth's atmosphere.

According to the Intergovernmental Panel on Climate Change, there is now visible and unequivocal evidence of climate change impacts, and there is a consensus that greenhouse gas emissions from human activities are the main drivers of change. The Earth's average temperature has been recorded to have increased by approximately 0.74°C over the past century.[1] Climate change is now recognized as a major global challenge that will have significant and long-lasting impacts on human well-being and development.[59]

One way of quantifying the contribution of greenhouse gases to climate change is through the standard metric known as radiative forcing, which is defined as positive if it results in a gain of energy for the Earth system (warming) and negative if it results in a loss (cooling).[1] The largest radiative forcing comes from CO_2, followed by CH_4, tropospheric ozone, halocarbons, and N_2O. All of these gases absorb infrared radiation emitted from the Earth's surface and re-emit it at a lower temperature, thus decreasing the outgoing radiation flux and producing a positive forcing, leading to warming. In contrast, stratospheric ozone depletion represents a small negative forcing, which leads to cooling of Earth's surface (Fig. 13); however, this contribution is expected to decrease as ODSs are gradually removed from the atmosphere.

Halocarbons and Climate Change

The understanding of the interaction between ozone depletion and climate change has been strengthened in recent years.[44,49] Stratospheric ozone and tropospheric ozone both absorb infrared radiation emitted by Earth's surface. Stratospheric ozone also significantly absorbs solar radiation. Therefore, changes in the stratospheric ozone and tropospheric ozone are directly linked to climate change. Stratospheric ozone and climate change are also indirectly linked because both ODSs and their replacements are greenhouse gases and represent an important contribution to the radiative forcing (see Fig. 13).

One approach of comparing the influence of individual halocarbons on ozone depletion and climate change is to use ozone depletion potentials (ODPs) and global warming potentials (GWPs). The ODP and GWP of a gas quantify its effectiveness in causing ozone depletion and climate forcing, respectively. GWP is defined as the total forcing attributed to a mass of emitted pollutant during a specified time after emissions (typically 100 years) as compared to the same mass of CO_2. ODP is the relative value that indicates the potential of a substance to destroy ozone as compared with the potential of CFC-11, which is assigned a reference value of 1.

Table 1 lists the atmospheric lifetime, global emissions, ODP, and GWP of some halogen source gases and the HFC replacement gases.[11,44] All ODSs and their substitutes shown here have a non-zero GWP,

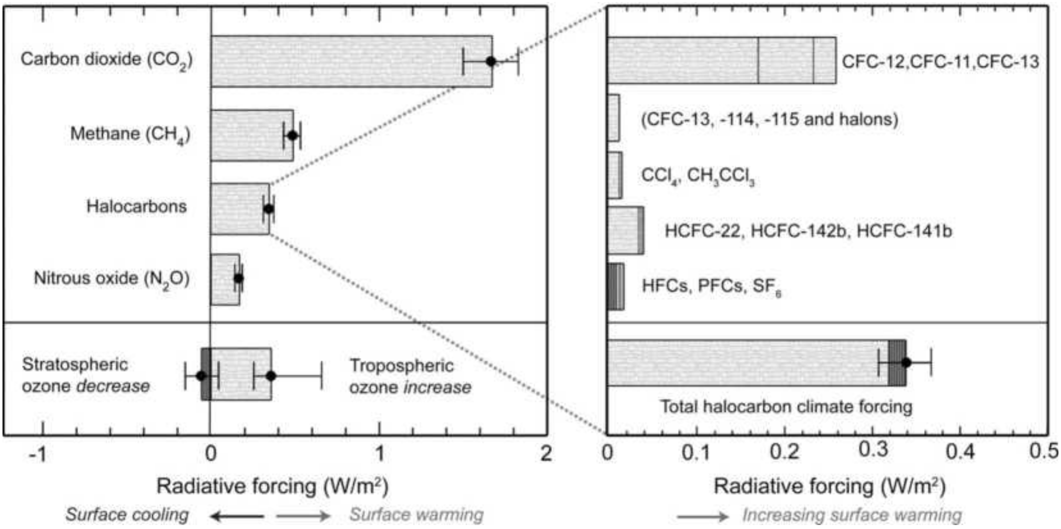


FIGURE 13 Radiative forcing of climate change. Left panel: radiative forcing of major greenhouse gases. Right panel: radiative forcing of halocarbons (darker shading designates Kyoto Protocol gases).
Source: Adapted from Fahey and Hegglin.[11]

TABLE 1 Atmospheric Lifetimes, Global Emissions, ODPs, and GWPs of Some Halogen Source Gases and HFC-Substituted Gases

Gas	Atmospheric Lifetime	Global Emissions (2008) (kt/yr)[a]	ODP	GWPb
Halogen source gases				
Chlorine gases				
CFC-11	45	52–91	1	4,750
CFC-12	100	41–99	0.82	10,900
CFC-113	85	3–8	0.85	6130
Carbon tetrachloride (CClq)	26	40–80	0.82	1400
HCFCs	1–17	385–481	0.01–0.12	77–2220
Methyl chloroform (CH_3CCl_3)	5	<10	0.16	146
Methyl chloride (CH_3Cl)	1	3600–4600	0.02	13
Bromine gases				
Halon-1301	65	1–3	15.9	7140
Halon-1211	16	4–7	7.9	1890
Methyl bromide (CH_3Br)	0.8	110–150	0.66	5
Very short-lived gases (e.g., $CHBr_3$)	<0.5	c	Very low[c]	Very low[c]
HFCs				
HFC-134	13.4	149±27	0	1370
HFC-23	222	12	0	14,200
HFC-143^a	47.1	17	0	4180
HFC-125	28.2	22	0	3420
HFC-152^a	1.5	50	0	133
HFC-32	5.2	8.9	0	716

[a] Includes both human activities (production and banks) and natural sources. Units are in kilotons per year.
[b] One-hundred-year GWPs. Values are calculated for emissions of an equal mass of each gas.
[c] Estimates are very uncertain for most species.
Source: Fahey and Hegglin[11] and WMO.[44]

with values spanning the wide range of 4 to 14,000; they are far more effective than equivalent amount of CO_2 in causing climate forcing. Therefore, the future selection of specific HFCs as replacement for ODS will have important consequences for climate change.

The CFCs, halons, and HCFCs are ODSs; they are controlled under the Montreal Protocol. Thus, the Montreal Protocol has provided collateral benefit of reducing the contributions of ODSs to climate change. In 2010, the decrease of annual ODS emissions under the Montreal Protocol is estimated to be about 10 gigatons of avoided CO_2-equivalent emissions per year, which is about 5 times larger than the annual emissions reduction target for the first commitment period (2008–2012) of the Kyoto Protocol.[44]

The HFCs, used as ODSs substitutes, do not destroy ozone (ODPs equal zero) and are considered ozone safe and have become the major replacement for CFCs and HCFCs. However, like the ODSs they replace, many HFCs have high GWP. HFCs, together with CO_2, CH_4, N_2O, PFCs, and SF_6, are controlled under the Kyoto Protocol. According to a new UNEP report, emissions of HFCs are growing at a rate of 8% per year due to growing demand in emerging economies and increasing populations. Without intervention, the increase in HFCs emissions is projected to offset much of the climate benefit achieved by the earlier reduction in ODS emissions. It is therefore important to select HFCs with low GWP potential and short lifetimes to minimize the climate impact while protecting the ozone layer.[60] This is one of the focal areas for the newly launched Climate and Clean Air Coalition on Short-lived Climate Pollutants.[61]

Impact of Climate Change on Ozone

The ODSs have declined as a result of the Montreal Protocol and its subsequent amendments; it is expected that this will lead to the recovery in the stratospheric ozone abundances. However, it is difficult to attribute ozone increases to the decreases in ODSs alone during the next few years because of natural variability, observational uncertainty, and confounding factors, such as changes in stratospheric temperature or water vapor. In contrast to the diminishing role of ODSs, changes in climate are expected to have an increasing influence on the recovery of ozone layer from the effects of ODSs.

Stratospheric ozone is influenced by changes in temperatures and winds in the stratosphere. For example, lower temperatures and stronger polar winds could both increase the extent and severity of winter polar ozone depletion. Observations show that the global-mean lower stratosphere has cooled by 1–2°C and the upper stratosphere cooled by 4–6°C between 1980 and 1995. There have been no significant long-term trends in global-mean lower stratospheric temperatures since about 1995. The two main reasons for the cooling of the stratosphere are depletion of stratospheric ozone and increase in atmospheric greenhouse gases. Ozone absorbs solar UV radiation, which heats the surrounding air in the stratosphere. Loss of ozone means that less UV light gets absorbed, resulting in cooling of the stratosphere. A significant portion of the observed stratospheric cooling is also due to human-emitted greenhouse gases. While the Earth's surface is expected to continue warming in response to the net positive radiative forcing from greenhouse gas increases, the stratosphere is expected to continue cooling.

A cooler stratosphere would extend the time period over which PSCs are present in winter and early spring, leading to increased polar ozone depletion. In the upper stratosphere at altitudes above PSC formation regions, a cooler stratosphere is expected to increase ozone amounts because lower temperatures decrease the effectiveness of ozone loss reactions.

Climate change may also affect the stratospheric circulation, which will significantly alter the distribution of ozone in the stratosphere. These changes tend to decrease total ozone in the tropics and increase total ozone at mid- and high latitudes. Changes in circulation induced by changes in ozone can also affect patterns of surface wind and rainfall. The projected changes in ozone and clouds may lead to large decreases in UV at high latitudes, where UV is already low, and to small increases at low latitudes, where it is already high. This could have important implications for human and ecosystem health. However, these projections depend strongly on changes in cloud cover, air pollutants, and aerosols, all of which are influenced by climate change. It is therefore important to improve our understanding of the processes involved and to continue monitoring ozone and surface UV spectral irradiances both from the surface and from satellites.[62]

Conclusions

The depletion of the stratospheric ozone layer exemplifies the global environmental challenges human face: it is an unintended consequence of human activity. Strong involvement and cooperation of stakeholders at all levels (scientists, technologists, economic and legal experts, environmentalists, and policy makers); strengthening of human and institutional capacities, coupled with suitable mechanisms for facilitating technological and financial flows; and changes in human behavior have been critical to the success in phasing out the ODSs.

The 1987 Montreal Protocol on Substances that Deplete the Ozone Layer is a landmark agreement that has successfully reduced the global production, consumption, and emissions of ODSs. By protecting the ozone layer from much higher levels of depletion, it has provided direct benefits to human health and agriculture, which in turn provide economic benefits by decreasing health costs and increasing crop production. Furthermore, because many ODSs are also potent greenhouse gases, the Montreal Protocol has provided substantial co-benefits to climate change.

On the other hand, demand for replacement substances such as HFCs has increased. Many of these substances are Kyoto gases. Additional climate benefits could be achieved by managing the emissions of replacement fluorocarbon gases and by implementing alternative gases with lower GWPs, as well as designing buildings that avoid the need for air conditioning.

Assuming full Montreal Protocol compliance, midlatitude ozone is expected to return to 1980 levels before mid-century. The recovery rate will be slower at high latitudes. Springtime ozone depletion is expected to continue to occur at polar latitudes, especially in Antarctica, in the next few decades. It is estimated that the ozone layer over the Antarctica will recover to pre-1980 levels between 2060 and 2075, and probably one or two decades earlier in the Arctic. However, effective control mechanisms for new chemicals threatening the ozone layer are essential; continued monitoring of the ozone layer is crucial to maintain momentum on recovering the ozone layer while simultaneously minimizing the influence on climate.

Changes in climate are expected to have an increasing influence on stratospheric ozone in the coming decades. An important scientific challenge is to project future ozone abundance based on an improved understanding of the complex linkages between stratospheric ozone and climate change.

Human activities will continue to change the composition of the atmosphere and new challenges that require international cooperation and collaboration will emerge. The ozone-hole phenomenon demonstrates the importance of long-term atmospheric monitoring and research, without which depletion of the ozone layer might not have been detected until more serious damage was evident. It is important for national and international agencies to continue their coordinated efforts on atmospheric monitoring, research, and assessment activities to provide sound scientific data needed to understand environmental changes on both regional and global scales.

Acknowledgments

The author gratefully acknowledges the use of the material and figures from “WMO Scientific Assessment of Ozone Depletion: 2010” and NASA Ozone Hole Watch presented in this entry.

References

1. IPCC (Intergovernmental Panel on Climate Change). Climate Change 2007: The Physical Science Basis, Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change; Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Avery, K.B., Tignor, M., Miller, H.L., Eds.; Cambridge University Press: Cambridge United Kingdom and New York, USA, 2007, 996 pp.
2. Molina M.J.; Molina L.T. Chlorofluorocarbons and destruction of the ozone layer. In *Environmental and Occupational Medicine,* 4th Ed.; Rom, W.N., Ed.; Lippincott, Williams and Wilkins: Philadelphia, 2007; 1605–1615.
3. Chapman, S. A theory of upper atmospheric ozone. Mem. R. Meteorol. Soc. **1930**, *3,* 103.
4. Molina, L.T.; Molina, M.J. Absolute absorption cross sections of ozone in the 185–350 nm wavelength range. J. Geophys. Res. **1986**, *91,* 14501–14508.
5. Crutzen, P.J. The influence of nitrogen oxides on atmosphere ozone content. Q. J. R. Meteorol. Soc. **1970**, *96,* 320–325.
6. Johnston H.S. Reduction of stratospheric ozone by nitrogen oxide catalysts from supersonic transport exhaust. Science **1971**, *173,* 517–522.
7. Molina, M.J.; Rowland, F.S. Stratospheric sink for chlorofluoromethanes: Chlorine-atom catalyzed destruction of ozone. Nature **1974**, *249,* 810–812.
8. Rowland, F.S.; Molina, M.J. Chlorofluoromethanes in the environment. Rev. Geophys. Space Phys. **1975**, *13,* 1–35.

9. Stolarski, R.S.; Cicerone, R. Stratospheric chlorine: A possible sink for ozone. Can. J. Chem. **1974**, *52,* 1610–1650.
10. Dobson, G.M.B.; Harrison, D.N. Measurement of the amount of ozone in the Earth's atmosphere and its relation to other geophysical conditions. Proc. R. Soc. London **1926**, *110,* 660–693.
11. Fahey, D.W.; Hegglin, M.I. Twenty Questions and Answers About the Ozone Layer: 2010 Update, Scientific Assessment of Ozone Depletion: 2010; World Meteorological Organization: Geneva, Switzerland, 2011; 72 pp.
12. Midgley, T. From the periodic table to production. Ind. Eng. Chem. **1937**, *29,* 241–244.
13. C&EN Special Issue **2008**, *86* (14), available at http://pubs.acs.org/cen/priestley/recipients/1941midgely.html (accessed May 29, 2011).
14. Lovelock, J.E. Atmospheric fluorine compounds as indicators of air movements. Nature **1971**, *230,* 379.
15. Lovelock, J.E.; Maggs, R.J.; Wade, R.J. Halogenated hydrocarbons in and over the Atlantic. Nature **1973**, *241,* 194–196.
16. Molina, M.J. Polar ozone depletion (Nobel lecture). Angew. Chem. Int. Ed. Engl. **1996**, *35,* 1778–1785.
17. Rowland, F.S. Stratospheric ozone depletion by chlorofluorocarbons (Nobel lecture). Angew. Chem. Int. Ed. Engl. **1996**, *35,* 1786–1798.
18. Rowland, F.S.; Molina, M.J. The CFC-Ozone Puzzle: Environmental Science in the Global Area. John H. Chafee Memorial Lecture on Science and the Environment. National Academy of Sciences, December 7, 2000.
19. DeMore, W.B.; Sander, S.D.; Golden, D.M. et al. Chemical kinetics and photochemical data for use in the stratospheric modeling. Evaluation no. 11, JPL publication no. *94–26.* NASA Jet Propulsion Laboratory: Pasadena, CA, 1994.
20. World Meteorological Organization. Scientific assessment of ozone depletion: 1994. WMO Global Ozone Research and Monitoring Project, report no. 37, Geneva: WMO, 1995.
21. World Meteorological Organization. Scientific assessment of ozone depletion: 1998. WMO Global Ozone Research and Monitoring Project, report no. 44, Geneva: WMO, 1999.
22. National Research Council. *Halocarbon: Environmental Effects of Chlorofluoromethane Release;* National Academy of Sciences: Washington, DC, 1976.
23. National Research Council. *Halocarbons: Effects on Stratospheric Ozone;* National Academy of Sciences: Washington, DC, 1976.
24. Farman, J.C.; Gardiner, B.G.; Shanklin, J.D. Large losses of total ozone in Antarctica reveal seasonal ClO_x/NO_x interactions. Nature **1985**, *315,* 207–210.
25. Stolarski, R.S.; Bloomfield, P.; McPeters, R.D.; Herman, J.R. Total ozone trends deduced from Nimbus 7 TOMS data. Geophys. Res. Lett. **1991**, *18,* 1015–1018.
26. Russell, J.M. III; Luo, M.; Cicerone, R.J.; Deaver, L.E. Satellite confirmation of the dominance of chlorofluorocarbons in the global stratospheric chlorine budget. Nature **1996**, *379*, 526–529.
27. Jones, A.E.; Shanklin, J.D. Continued decline of total ozone over Halley, Antarctica, since 1985. Nature **1995**, *376,* 409–411.
28. NASA Ozone Hole Watch, available at http://ozonewatch.gsfc.nasa.gov/facts/hole.html (accessed May 28, 2011).
29. Solomon S.; Garcia, R.R.; Rowland, F.S.; Wuebbles, D.J. On the depletion of Antarctic ozone. Nature **1986**, *321*, 755–758.
30. Molina, L.T.; Molina, M.J.; Stachnick, R.A.; Tom, R.D. An upper limit to the rate of the $HCl+ClONO_2$ reaction. J. Phys. Chem. **1985**, *89,* 3779–3781.
31. Molina, M.J.; Tso, T.-L.; Molina, L.T.; Wang, F.C.-Y. Antarctic stratospheric chemistry of chlorine nitrate, hydrogen chloride and ice. Release of active chlorine. Science **1987**, *238,* 1253–1260.
32. Tolbert, M.A.; Rossi, M.J.; Malhotra, R.; Golden, D.M. Reaction of chlorine nitrate with hydrogen chloride and water at Antarctic stratospheric temperatures. Science **1987**, *238,* 1258–1260.

33. Abbatt, J.P.D.; Beyer, K.D.; Fucaloro, A.F.; McMahon, J.R.; Wooldridge, P.J.; Zhang, R. Interaction of HCl vapor with water-ice: Implications for the stratosphere. J. Geophys. Res. **1992**, 97, 15819–15826.
34. Molina, M.J. The probable role of stratospheric 'ice' clouds: Heterogeneous chemistry of the ozone hole. In *Chemistry of the Atmosphere: The Impact of Global Change*; Calvert, J.G., Ed.; Blackwell Scientific: Oxford, U.K., 1994; 27–38.
35. Gertner, B.J.; Hynes, J.T. Molecular dynamics simulation of hydrochloric acid ionization at the surface of stratospheric ice. Science **1996**, *271,* 1563–1566.
36. Molina L.T.; Molina M.J. Production of Cl_2O_2 from the self-reaction of the ClO radical. J. Phys. Chem. **1987**, *91,* 433–436.
37. McElroy, M.B.; Salawitch, R.J.; Wofsy, S.C.; Logan, JA. Reduction of Antarctic ozone due to synergistic interactions of chlorine and bromine. Nature **1986**, *321,* 759–762.
38. Cox, R.A.; Hayman, G.D. The stability and photochemistry of dimers of the ClO radical and implications for Antarctic ozone depletion. Nature **1988**, *332,* 796–800.
39. Molina, M.J.; Colussi, A.J.; Molina, L.T.; Schindler, R.N.; Tso. T-L. Quantum yield of chlorine-atom formation in the photodissociation of chlorine peroxide (ClOOCl) at 308 nm. Chem. Phys. Lett. **1990**, *173,* 310–315.
40. Tuck, A.F.; Watson R.; Condon, E.P.; Margitan, J.J.; Toon, O.B. The planning and execution of ER-2 and DC-8 aircraft flights over Antarctica, August and September 1987. J. Geophys. Res. **1989**, *94,* 181–222.
41. Anderson, J.G., Toohey, D.W., Brune, W.H. Free radicals within the Antarctic Vortex: the role of CFCs in Antarctic ozone loss. Science **1991**, *251,* 39–46.
42. Proffitt, M.H.; Steinkamp, M.J.; Powell, J.A. et al. In situ ozone measurements within the 1987 Antarctic ozone hole from a high-altitude ER-2 aircraft. J. Geophys. Res. **1989**, 94, 547–555.
43. Chen, H.-Y.; Lien, C.Y.; Lin, W.-Y.; Lee, Y.T.; Lin, J.J. UV absorption cross sections of ClOOCl are consistent with ozone degradation models. Science **2009**, *324,* 781–784.
44. WMO (World Meteorological Organization). Scientific Assessment of Ozone Depletion: 2010, Global Ozone Research and Monitoring Project—Report No. 52, Geneva, Switzerland, 2011; 516 pp.
45. Turco, R.; Plumb, A.; Condon, E. The Airborne Arctic Stratospheric Expedition: Prologue. Geophys. Res. Lett. **1990**, *17,* 313–316.
46. Manney, G.L.; Santee, M.L.; Rex, M. et al. Unprecedented Arctic ozone loss in 2011. Nature **2011**, *478,* 469–475.
47. van der Leun, J.; Tang, X.; Tevini, M. Environmental effects of ozone depletion: 1994 assessment. Ambio **1995**, *24,* 138.
48. Biggs R.H.; Joyner, M.E.B., Eds. *Stratospheric Ozone Depletion/UV-B Radiation in the Biosphere;* Springer-Verlag: New York, 1994.
49. UNEP, Environmental Effects of Ozone Depletion and its Interaction with Climate Change: 2010 Assessment. United Nations Environment Programme, December 2010, Nairobi, Kenya, available at http://ozone.unep.org/Assessment_Panels/EEAP/eeap-report2010.pdf (accessed May 20, 2011).
50. Longstreth, J.D.; de Grujil, F.R.; Kripke, M.L.; Takizawa, Y.; van der Leun, J.C. Effects of solar radiation on human health. Ambio **1995**, *24,* 153–165.
51. Molina, M.J.; Molina L.T.; Fitzpatrick, T.B.; Nghiem, P.T. Ozone depletion and human health effects. In *Environmental Medicine;* Moller L., Ed.; Joint Industrial Safety Council Product 33: Sweden, 2000; 28–51.
52. Brash, D.E.; Rudolph, J.A.; Simon, J.A. et al. A role for sunlight in skin cancer: UV-induced p53 mutations in squamous cell carcinoma. Proc. Natl. Acad. Sci. U. S. A. **1991**, *88,* 124–128.
53. De Fabo, E.C.; Noonan, F.P. Mechanism of immune suppression by ultraviolet irradiation in vivo. I. Evidence for the existence of a unique photoreceptor in skin and its role in photo-immunology. J. Exp. Med. **1983**, *157,* 84–98.

54. Cooper, K.D.; Oberhelman, L.; Hamilton, T.A. et al. UV exposure reduces immunization rates and promotes tolerance to epicutaneous antigens in humans—Relationship to dose, CD1a-DR+ epidermal macrophage induction and Langerhans cell depletion. Proc. Natl. Acad. Sci. U. S. A. **1992**, *89,* 8497–8501.
55. Caldwell, M.M.; Teramura, A.H.; Tevini, M.; Bomman, J.F.; Björn, L.O.; Kulandaivelu, G. Effects of increased solar ultraviolet radiation on terrestrial plants. Ambio **1995**, *24,* 166–173.
56. Häder, D.P.; Worrest, R.C.; Kumar, H.D.; Smith, R.C. Effects of increased solar ultraviolet radiation on aquatic ecosystems. Ambio **199**5, *24,* 174–180.
57. Smith, R.C.; Prezelin, B.B.; Baker, K.S. et al. Ozone depletion: Ultraviolet radiation and phytoplankton biology in Antarctic waters. Science **1992**, 255, 952–959.
58. UNEP (United Nations Environment Program). *Handbook for the Montreal Protocol on Substances that Deplete the Ozone Layer,* 8th Ed.; UNEP Ozone Secretariat, 2009.
59. UNEP (United Nations Environment Program). *Global Environmental Outlook (GEO-4): Environment for Development;* Progress Press Ltd.: Malta, **2007**.
60. UNEP (United Nations Environment Program). HFCs: A Critical Link in Protecting Climate and the Ozone Layer, November 2011; 36 pp.
61. The Climate and Clean Air Coalition to Reduce Short-Lived Climate Pollutants, Fact Sheet. http://www.state.gov/rZpa/prs/ps/2012/02/184055.htm (accessed February 16, 2012)
62. McKenzie, R.L.; Aucamp, P.J.; Bais, A.F.; Björn, L.O.; Ilyas, M.; Madronich, S. Ozone depletion and climate change: impacts on UV Radiation. Photochem. Photobiol. Sci. **2011**, *10,* 182–198.

16

Thermodynamics

Ronald L. Klaus

Introduction

Thermodynamics is concerned with the interaction between matter and energy on a macroscopic level. It is therefore the basic science that underlies the engineering of the use of energy in all its forms. It is governed by two basic laws with far-reaching implications. The first concerns the conservation of energy, and the second puts limits on the amount of energy that can be converted into work over and above what the first law might be thought to imply.

The actual performance of practical process equipment can be predicted by application of these two laws. That, in turn, depends on the ability to estimate the thermodynamic properties of the streams that enter and leave this equipment. The exact equations needed to calculate these properties can be derived from the formalism of thermodynamics. Certain simplifications, such as the ideal gas law or the assumption of an equation of state, are usually required to make numerical estimates of these properties.

First Law of Thermodynamics

Formulation

The first law of thermodynamics is a statement of the principle of conservation of energy for a closed system—namely, one in which there is no transfer of mass across the system boundaries.

$$\Delta U = Q - W_{\text{total}} \tag{1}$$

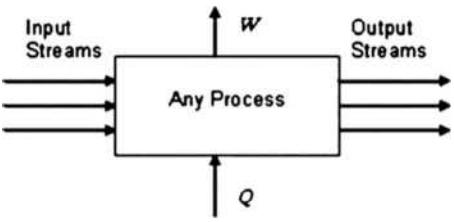


FIGURE 1 Schematic representation of any thermodynamic process.

This equation ignores certain effects such as differences in fluid velocities, which become important only at very high velocities, and differences in fluid elevations, which become important only when there are greatly varying elevations. U is a thermodynamic property called internal energy. It accounts for energy due to the motion of molecules in a fluid and the exchanges of energy between them. The Δs refer to differences in the property value between the final and initial states of the system. U, Q, and W all have units of energy. However, this equation can also have another interpretation for open flow systems at steady state. Such systems have input and outlet streams, and possibly an exchange of heat or work with the surroundings (Figure 1). However, the stream flows and property values throughout the system are all considered constant over time. In this case, the first law is a rate equation. ΔU is the difference between the sum of U times the mass flow rate for all the output streams minus that same sum for all the inlet streams. Q and W are rates of energy transfer. The arrows in Figure 1 indicate the direction of the flow of work, W, and heat, Q, when these quantities are positive.

When applying Eq. 1 to a flow process, W_{total} includes the work done in moving fluids into and out of the process. To eliminate these effects, another thermodynamic property, namely enthalpy, H, is defined,

$$H = U + PV \tag{2}$$

where P is pressure and V is volume.

For a flow process at steady state, this changes the first law to

$$\Delta H = Q - W \tag{3}$$

In Eq. 3, W is the work produced by the process apart from that involved in moving fluids into and out of the process. It is often called the shaft work.

In these equations, it is important to make a distinction between intrinsic and extrinsic properties. The former are properties per unit mass or per mole, whereas the latter are total values of the property for the mass flow of the given process stream. These and other thermodynamic relationships can be applied to either, provided that they are applied consistently.

U, H, P, and V are examples of thermodynamic state functions—that is, their intrinsic values depend only on the state of the fluid (usually its temperature, pressure, and composition) and not on the process used to bring the fluid to that state. On the other hand, quantities like Q and W are not state functions. When a fluid makes a transition from one state to another, the values of Q and W associated with that transition are dependent on the "path" over which that transition was made.

Enthalpies of Formation and Reaction

Enthalpy cannot be calculated in an absolute sense. It requires the definition of a reference state, at which it is taken to be zero. Enthalpies in all other states are referenced to this state. The most rigorous definition of a reference state is to take it as comprising the constituent atomic species in their normal molecular configurations at some temperature and pressure—typically 25°C and 1 atm. Normal molecular configuration means the normal configuration the atomic molecules take in nature, such as O_2, H_2, N_2, etc. Thus, these molecules would be taken to have zero enthalpy at 25°C and 1 atm. The phases of these elemental species also have to be stated. They are generally taken to be the normal phase of the element at the stated conditions. The reference states for gases are usually taken to be in their "ideal gas state" (ig). (The ideal gas state is difficult to describe briefly. It is a fluid with the same heat capacity as the real fluid's heat capacity at very low pressure. However, unlike the real fluid, it obeys the ideal gas law. For low-pressure calculations, the difference between the ideal gas state and the real fluid at low pressure is not important.)

With these ideas as a starting point, the enthalpies of compounds can be built up by considering real or imaginary reactions through which they are formed from their constituent elements. For example, the enthalpy of formation of methane would be the enthalpy change associated with the following reaction:

$$\mathrm{C(s)+2H_2(ig)=CH_4(ig)} \quad (25°\mathrm{C}, 1\ \mathrm{atm})$$

where s indicates solid and ig indicates the ideal gas state. Such chemical reactions are usually accompanied by the absorption or release of a certain amount of heat. If the reaction is carried out at constant temperature and pressure, and if no work is done, according to the first law, this heat is the enthalpy difference of the reaction and defines the standard enthalpy (or heat) of formation of the resulting compound.

Standard enthalpies of formation are widely tabulated. An excellent source is the collection by the National Institute of Standards and Technology (NIST).[1] Enthalpies of formation for some common species are given in Table 1. They are given both in the units that appear in the NIST collection and in nondimensionalized form. The latter permits their use in any system of units simply by multiplying by RT_0 in the desired units. To is taken to be 0°C (= 273.15°K=459.67°*R*), and *R* is the Universal Gas Constant. Note that °*R* represents degrees Rankine and °K represents degrees Kelvin.

Many different units are used in energy calculations. The values of some useful conversion factors and the values of the universal gas constant appear in Table 2.

These enthalpies of formation can be combined to produce enthalpy changes for any chemical reaction whose species have known enthalpies of formation. For example, in the combustion reaction

$$\mathrm{CH_4+2O_2=CO_2+2H_2O}$$

the enthalpy change may be calculated as follows:

$$\Delta H=\Delta H_\mathrm{f}(\mathrm{CO_2})+2\Delta H_\mathrm{f}(\mathrm{H_2O})-\Delta H_\mathrm{f}(\mathrm{CH_4})-2\Delta H_\mathrm{f}(\mathrm{O_2})$$
$$=(-393.51)+2(-285.830)-(-74.84)-0=-890.3\mathrm{kj/gmole}$$

TABLE 1 Enthalpies of Formation of Some Common Compounds

Chemical Specie	Formula	State	MW	ΔH_f (T=298.15°C)	ΔH_f (Non dim.)
Methane	CH_4	IG	16.04	−74.87	−32.968
Ethane	C_2H_6	IG	30.08	−83.8	−36.901
Propane	C_3H_8	IG	44.10	−104.7	−46.104
n-Octane	C_8H_{18}	IG	114.23	−208.4	−91.767
n-Octane	C_8H_{18}	L	114.23	−250.3	−110.217
Carbon dioxide	CO_2	IG	44.01	−393.51	−173.278
Carbon monoxide	CO	IG	28.01	−110.53	−48.671
Water	H_2O	IG	18.02	−241.826	−106.486
Water	H_2O	L	18.02	−285.830	−125.863

Note: Units of ΔH_f= kJ/gmole; Non-dimensionalization, $\Delta H_f/RT_0$ where T_0= 273.15°K.

TABLE 2 Some Useful Conversion Factors and Constants

Type	Value	Conversion
Length	1 m	3.28084 ft
		39.3701 in.
Mass	1 kg	2.20462 lbm
Force	1 N	1 kg m/s^2
		0.224809 lbf
Pressure	1 bar	10^5 N/m^2
		10^5 Pa
		0.986923 atm
		14.5038 psia
Pressure	1 atm	14.6960 psia
Volume	1 m^3	35.3147 ft^3
Density	1 g/cm^3	62.4278 lbm/ft^3
Energy	1 J	1 N m
		1 m^3 Pa
		10^{-5} m^3 bar
		10^{-3} kW s
		9.86923 cm^3 atm
		0.239006 cal
		5.12197×10^{-3} ft^3 psia
		0.737562 ft lbf
		9.47831×10^{-4} Btu
Power	1 kW	10^3 J/s
		239.006 cal/s
		737.562 ft lbf/s
		0.94783 Btu/s 1.34102 hp

Note: Values of the universal gas constant, R
8.314 J/gmole K = 8.314 m^3 Pa/(gmol K) = 83.14 cm^3 bar/(gmole K) 82.06 cm^3 atm/(gmole K)
1.987 cal/(gmole K) = 1.986 Btu/(lbmole R)
0.7302 ft^3 atm/(lbmole R) = 10.73 ft^3 psia/(lbmole R)
1545 ft lbf/(lbmole R)

If ΔH is positive, the reaction is called endothermic, whereas if it is negative, it is called exothermic. Because this particular reaction is a combustion reaction, the absolute value of its enthalpy change is also called the enthalpy change (or heat) of combustion.

Application to Combustion

The goal of a power cycle is to convert the chemical energy of a fuel into usable mechanical energy that can produce useful shaft work or drive a generator to produce electricity (Figure 2). What one hopes to do is to convert as much of the fuel's chemical energy into useful work. The hope for this ideal is reflected in what is called the heating value of the fuel. It is taken to be the negative of the enthalpy change of the combustion reaction through which a fuel is completely oxidized at the so-called ISO (International Organization for Standardization) conditions—namely, 15°C and 1 atm pressure. (ISO conditions for air also include a relative humidity of 60%.) Thus, it is the negative of the enthalpy change of the following kind of reaction:

$$\text{Fuel} + aO_2(\text{ig}) \rightarrow bCO_2(\text{ig}) + cH_2O(\text{ig or 1}) + \text{Any other products}(51°\text{C}, 1\text{ atm})$$

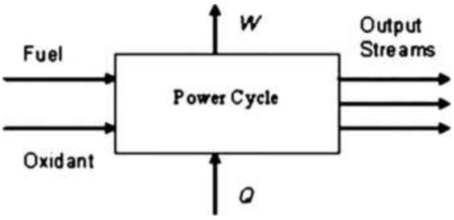


FIGURE 2 Schematic representation of a power cycle.

TABLE 3 Heating Values of Some Common Fuels

	HHV		LHV		LHV/
Fuel	Btu/lbm	Btu/gal	Btu/lbm	Btu/gal	HHV
No. 2 oil	19,580	142,031	18,421	133,623	0.9408
No. 4 oil	18,890	146,476	17,804	138,055	0.9425
No. 6 oil	18,270	150,808	17,312	142,901	0.9476
Diesel fuel	19,733		18,487		0.9368
Hydrogen	61,007		51,635		0.8464
Methane	23,876		21,518		0.9012
Typical natural gas	22,615		20,450		0.9019
Propane	21,653		19,922		0.9201
Butane	21,266		19,623		0.9227
Gasoline	19,657	121,808	18,434	114,235	0.9379
Reformulated gasoline	19,545	120,103	18,304	112,377	0.9365
Methanol	11,274	73,882	10,115	66,289	0.8972
Anthracite coal	14,661		14,317		0.9765
Bituminous coal	14,100		13,600		0.9645

Source: Based on calculations by the author and data from The Association of Energy Engineers.[2–4]

where *a, b,* and *c* are coefficients that depend on the particular fuel.

It is unfortunate that ISO conditions differ from the reference state at which enthalpies of formation are usually tabulated. Nevertheless, the latter can be adjusted to ISO conditions, and values for some common fuels appear in Table 3.

Heating Values, Heat Rates, and Power Cycle Efficiency

One area of confusion about heating values comes because of the choice of the phase of the water produced as a combustion product. If the product water is considered to be a liquid, its enthalpy will be lower than that if it is considered to be a vapor. The former choice leads to a higher heating value (HHV), whereas the latter leads to a lower heating value (LHV). If the chemical composition of the fuel is known, the heating value is the mole fraction average of the heating values of the individual constituents divided by the average molecular weight of the fuel. Where the composition of the fuel is not known, accurate heating values are determined experimentally.

The heat rate is one often-tabulated measure that is used to describe the effectiveness of a machine—such as an internal combustion engine or a gas turbine—used to convert the chemical energy of a fuel into usable mechanical or electrical energy. Several definitions go into it.

$$\text{Heat Energy Input(Btu/s)}$$

$$= \text{First Law Available Chemical Energy}$$

$$= \text{Heating value of a Fuel } (\text{Btu/lb}) \times \text{Fuel flow rate } (\text{lb/s})$$

Heat Rate (Btu/kWh)

$$= \text{Heat Energy Input } (\text{Btu/s}) \times 3600 (\text{Btu/s}) / \text{Power Produced } (\text{kW})$$

The heat rate is based on a particular choice of either HHV or LHV of the fuel. The higher the heat rate, the more fuel is being used per kW of power produced and therefore the less efficient the machine. The heat rate is directly related to the efficiency.

$$\text{Efficiency} = \frac{\text{Power Output}(\text{kW}) \times 9.47831 \times 10^{-4}\,\text{Btu/J}}{\text{Heat Energy Input } (\text{Btu/s})}$$

Comparison with the previous equations yields Efficiency

$$\text{Efficiency} = \text{Conversion Factor } (\text{Btu/kWh}) / \text{ Heat Rate} (\text{Btu/kWh})$$

The conversion factor's role is strictly to convert the units. In the units shown it is equal to

$$\text{Conversion Factor} = 9.47831 \times 10^{-4}\,\text{Btu/J} \times 1000\text{J/kW} - \text{s} \times 3600\text{s/h} = 3412.19\text{Btu/kWh}$$

When the power-generating equipment is used to produce electricity, the efficiency is often called the electrical efficiency.

The efficiency calculated by the above equations is based on the hope of converting the entire heating value of the fuel into useful work. It is unfortunate that this has come to be the standard because there is a further limitation imposed by the second law of thermodynamics. It tells us that nature will not allow conversion of this much chemical energy into useful work, even if the power cycle were constructed as perfectly as possible under the most ideal conditions.

The Second Law of Thermodynamics

Reversibility

Before stating the second law specifically, it is necessary to introduce the concept of reversibility. Most processes produce changes in the substances on which they operate. Reversibility has to do with how easy it would be to undo the changes that are produced in a substance through some process. Specifically, reversibility means that a change that a process makes on a substance could theoretically be undone (reversed) with only an infinitesimal change in the conditions in the process.

For example, suppose that heat is exchanged by allowing heat to flow from a hotter substance to a cooler one. For this change to be reversible, temperatures at every point where heat transfer takes place must differ only infinitesimally. Thus, if we raised the temperature profile of the cooler substance by just an infinitesimal amount, the heat could be transferred back to the hotter substance.

As another example, suppose that a gas is compressed in a cylinder by moving a piston in such a way that it reduces its volume. For the change to be reversible, the pressure the piston exerts must be only differentially greater than the pressure of the gas. In this case, if the piston pressure were reduced only infinitesimally, the gas could be expanded back to its original state, and all the work of compression would be recovered. On the other hand, if a pressure were applied that differed substantially from that of the gas in the piston, the change brought about would be irreversible, because a differential change in the applied pressure could not reverse the effects of the compression.

Formulation

The second law of thermodynamics has a certain mystique to it because unlike most physical laws, it is not a conservation law (e.g., conservation of mass, energy, momentum, etc.). Instead, it puts certain limits on what nature allows in spite of the great conservation laws. It is also not as intuitive as the other great laws. Therefore, much has been written to try to explain it. However, it can be applied rigorously based on the following three-part formulation. This formulation refers to a given substance of constant mass that experiences changes from an initial to a final state through the possible exchange of heat and work with its surroundings.

Part 1: There is a thermodynamic state function, called entropy, S, which is defined by the following equation:

$$\mathrm{d}S = \frac{\mathrm{d}Q_{\mathrm{rev}}}{T} \tag{4}$$

where $\mathrm{d}S$ is the differential change in entropy and $\mathrm{d}Q_{\mathrm{rev}}$ refers to a differential amount of heat transferred to or from a substance in a reversible manner. That is, to calculate finite entropy differences, this equation needs to be integrated along a reversible path.

Stating that entropy is a state function is not a trivial claim. What it means is that no matter what reversible path is chosen, the net change in entropy will be the same, provided that the initial and final states are the same.

Part 2: No matter how a change of state is brought about in a given substance

$$\Delta S_{\mathrm{substance}} + \Delta S_{\mathrm{surrounding}} \geq 0 \tag{5}$$

The differences implied by the Δs are differences between the end state and the initial state of both the substance and surroundings. Notice that there is such a thing as a change in the entropy of the surroundings as well as the substance. It too can be calculated by Eq. 4.

Part 3: The inequality in the previous equation becomes an equality if, and only if, the process used to bring about the change is completely internally reversible and if the process also exchanges heat with the surroundings in a completely reversible manner.

Thus, this third part becomes a criterion of reversibility for a process or piece of equipment. Furthermore, the greater the inequality, the greater the irreversibility.

In general, the surroundings are considered to be at a constant ambient temperature, T_0. Thus,

$$\Delta S_{\mathrm{surroundings}} = \frac{Q_{\mathrm{surroundings}}}{T_0} = \frac{Q_{\mathrm{substance}}}{T_0} \tag{6}$$

where the two Qs are heat transfers to the surroundings and substance, respectively, which are the same in magnitude but opposite in sign. This equation can be combined with Eq. 5 to give

$$Q \leq T_0 \Delta S \tag{7}$$

Along with the first law for flow processes, this gives

$$W \leq T_0 \Delta S - \Delta H \tag{8}$$

In these equations, the subscripts have been dropped and all quantities refer to the substance, not the surroundings. Furthermore, although these equations were derived for changes in a substance of constant mass (closed system), as with the first law, they can also be applied to a steady flow system. In this case, the Δs refer to differences in the fluxes of the thermodynamic properties between the output and inlet streams.

The implications of these two equations are very significant. They mean that in a flow process, if the inlet and outlet conditions of the various streams are fixed, there are an upper bound to the amount of heat that can be transferred to the process and an upper bound on the amount of work it can produce. The latter claim has an impact on power generation systems because it imposes a limit on the amount of power that can be produced from a given amount of fuel, regardless of its energy content. Furthermore, because the entropy change of such processes is usually negative, another implication of Eq. 7 is that a certain amount of heat must be rejected to the surroundings regardless of how well the process is designed internally.

Carnot Engine

Some of the implications of the second law can be illustrated through the so-called Carnot engine. This is an imaginary reversible engine in which heat, Q_H, is transferred from a high-temperature source at T_H into a reversible engine, which produces work, W, and discards heat, Q_C, into a cold sink—possibly the environment—at T_C (Figure 3). From the first law,

$$W = Q_H - Q_C$$

From the second law,

$$\frac{Q_C}{T_C} - \frac{Q_H}{T_H} \geq 0$$

If the Carnot efficiency is defined as the amount of work that can be produced from the high-temperature heat (i.e., W/Q_H),

$$\frac{W}{Q_H} = \frac{Q_H - Q_C}{Q_H} = \frac{T_H - T_C}{T_H} \tag{9}$$

The last term on the right is of particular interest. It means that the Carnot efficiency depends on only the temperatures of the two heat reservoirs. All the heat in a high-temperature source can never be converted to work, even if the apparatus is perfectly reversible. The higher the temperature of the hot source relative to the cold sink, the higher the Carnot efficiency.

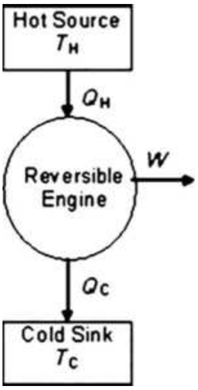


FIGURE 3 The Carnot engine.

Ideal Work, Lost Work

The second law gives a means to measure the irreversibility of a process or piece of equipment. That, in turn, is a measure of its departure from an ideal design—one which makes changes to substances in a way that most preserves their work-producing potential. This analysis begins by imagining a perfectly reversible process and defining W_{ideal} and Q_{ideal} to be the limiting values defined by Eqs. 7 and 8 (see Figure 4).

$$W_{\text{ideal}} = T_0\Delta S - \Delta H \tag{10}$$

$$Q_{\text{ideal}} = T_0\Delta S \tag{11}$$

These ideal values can be determined from the enthalpies and entropies of the various streams entering and leaving the process alone, without reference to any of the internal details of the process.

The level of irreversibility in a real process or in some part of it can be quantified through what is called lost work.[5] Because the actual work that can be obtained from any real process is always less than that what would have resulted from a completely reversible process, this leads to the following definition:

$$W_{\text{lost}} = W_{\text{ideal}} - W \tag{12}$$

By combining this with Eqs. 10 and 3, this becomes

$$W_{\text{lost}} = T_0\Delta S - Q \tag{13}$$

A real process—or any subprocess—can be represented as shown in Figure 5. The W_{lost} stream is not an actual work output. However, showing it in this manner is a reminder that in any real process a certain amount of the potential to do work has been lost due to the irreversibilities of the process. Furthermore, the lost work of all the subprocesses within the overall process sum up to the lost work of the process as a whole,

$$W_{\text{ideal}} = W + \sum W_{\text{lost}} \tag{14}$$

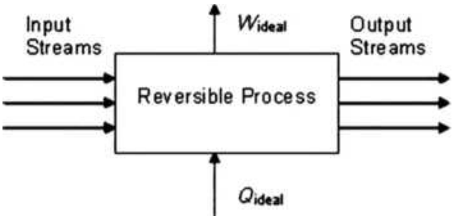


FIGURE 4 Schematic representation of a reversible thermodynamic process.

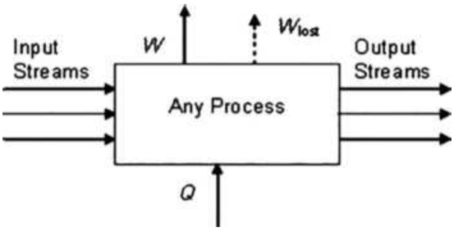


FIGURE 5 Schematic representation of any process showing lost work.

where the summation is carried out over each of the subprocesses. When analyzing a process, Eq. 13 can be applied to each piece of equipment. Then the W_{lost} for each of these can be summed according to Eq. 14. Such an analysis brings out where the greatest irreversibilities occur and therefore shows where the greatest opportunities are for improvement.

Exergy

Lost work is related to the modern concept of exergy. Lost work gives a quantitative measure of the irreversibility of a process or subprocesses within it. Furthermore, the term "lost work" seems quite descriptive. However, it is also possible to assign values—perhaps even monetary values—to streams as well as equipment. This can be done through the concept of exergy.

The exergy of a stream is the maximum amount of work that could theoretically be extracted from that stream if its temperature and pressure were reduced to that of the environment and if the concentrations of its chemical species were brought to that of the environment. It should be clear that from a power-production point of view, the economic value of a stream is related to its exergy because the production of work is usually more valuable than the production of heat alone.

Exergy analysis is based on the idea that it costs something in terms of equipment, power, adding costly streams, etc., to raise a stream's exergy—and hence its value. If what it costs to raise the value of a given stream is less than the increased value of the stream, it is a desirable thing to do.

Exergy is generally divided into four parts: physical, chemical, kinetic (having to do with the velocity of the fluid), and potential (having to do with its height above some datum level). The first two of these parts are of primary interest here. Physical exergy is the negative of the ideal amount of work that could be extracted from a stream if its temperature and pressure were reduced to ambient conditions. That is,

$$e^{\text{Ph}} = \Delta H - T_0 \Delta S \tag{15}$$

In Eq. 15 and the following equations, the Δs refer to differences between the present state of the fluid and its state at ambient conditions.

Chemical exergy, e^{Ch}, also has to be added because the exergy, say, of a fuel at ambient conditions ought to be higher than that of its eventual combustion products at those conditions because it has the potential of producing work in being transformed into its products. This can be accounted for in an equation similar to the one above, but this time involving properties of formation at ambient conditions.

$$e^{\text{Ch}} = \Delta H_{\text{f}} - T_0 \Delta S_{\text{f}} \tag{16}$$

Exergy is not quite a thermodynamic state function because it depends on environmental temperature. Also, unlike enthalpy or internal energy, it is not conserved in a process. Nevertheless, it is a useful measure of the power-producing value of a process stream. In any real processes, every process step results in a net loss of exergy.

The relationship between exergy and lost work is very straightforward—namely,

$$\text{Lost Work} = \sum \text{Exergies of incoming streams} - \sum \text{Exergies of outgoing streams} \tag{17}$$

where the summations are carried out over all incoming and outgoing streams, respectively.

Reversibility Revisited

The preceding analysis shows how the second law can be used to provide a measure of the irreversibility of a process or part of a process. Every irreversibility causes a loss in the potential to convert energy into work. In this era of diminishing supplies of cheap fuels and increasing demands for electric power, this is an important consideration. Furthermore, irreversibilities in one part of a process reduce the amount of work that can be produced—or else increase the amount of work that is needed—in the rest of the process, even though the source of the irreversibility may be far removed from the place in the process where work is produced or consumed.

The following is a list of some process steps that introduce irreversibility into processes, including power cycles:

- Heat transfer across a nonzero temperature difference
- Mixing of fluids of differing pressures, temperatures, or compositions
- Pressure drops that do not recover work
- Chemical reactions that do not take place at equilibrium conditions
- Flashing of liquids when pressure is reduced
- Two-phase contact and mass transfer between fluids whose various species have concentrations that differ from their equilibrium values
- Temperature and pressure shocks when fluids enter equipment at different temperatures or pressures from those that are present within the equipment

The most thermodynamically efficient designs—those that most preserve the work-producing potential of the fluids in the process—are those that strive for the greatest reversibilities in the various pieces of process equipment. However, there is another side to this question. The more reversible a piece of equipment is, the greater its capital cost is apt to be. Thus, in the practical design of equipment there is often a trade-off between reversibility and capital cost. Processes need to be evaluated not only for their thermodynamic efficiency, but also for their thermoeconomic effectiveness.

Calculation of Thermodynamic Properties

Rigorous Equations

Process equipment cannot be designed or evaluated apart from the ability to estimate the thermodynamic properties of the entering and exiting streams. The following are rigorous equations for the calculation of some of the important thermodynamic properties of pure or constant-composition substances (The equations can be found in many thermodynamics text books. A concise useful summary is found in Van Ness and Abott.)[6]:

$$dH = C_p dT + \left[V - T\left(\frac{\partial V}{\partial T_p} \right) \right] dP \tag{18}$$

$$dS = \left(\frac{C_P}{T} \right) dT - \left(\frac{\partial V}{\partial T} \right)_P dP \tag{19}$$

$$G = H - TS \tag{20}$$

C_p is the heat capacity at constant pressure, defined by

$$C_P = \left(\frac{\partial H}{\partial T} \right)_P \tag{21}$$

G is called the Gibbs free energy or Gibbs function and has importance in equilibrium calculations. Eq. 20 can be considered its definition. By differentiation and comparison with the other two, it can be shown that

$$dG = VdP - SdT \tag{22}$$

These are differential equations that emphasize that these properties need reference values from which differences can be calculated. To obtain values for the property differences for real substances, the pressure–volume–temperature (PVT) behavior of the fluid needs to be inserted into the equations, which would then be integrated to produce the final results.

Ideal Gas

One way of inserting PVT behavior is through equations of state. One of the simplest models for gases is that of the ideal gas, whose equation of state on a molar basis is

$$\mathrm{PV} = \mathrm{RT} \tag{23}$$

This equation reflects the fact that the molecules in the gas do not interact in any way, although they may have very complex energy interactions within a given molecule.

When this equation is inserted into the rigorous equations, one obtains

$$\mathrm{d}H^* = C_p^* \mathrm{d}T \tag{24}$$

$$\mathrm{d}S^* = \left(\frac{C_p^*}{T}\right)\mathrm{d}T + \left(\frac{R}{P}\right)\mathrm{d}P \tag{25}$$

$$\mathrm{d}G^* = RT\mathrm{d}\,\mathrm{In}\,P - S^*\mathrm{d}T \tag{26}$$

where the asterisk designates the properties as those of an ideal gas.

The integrated forms of Eqs. 24 and 25 are

$$\Delta H^*(P2,T2;P1,T1) = \int_{T1}^{T2} C_p^* \mathrm{d}T \tag{27}$$

$$\Delta S^*(P2,T2;P1,T1) = \int_{T1}^{T2} \left(\frac{C_p^*}{T}\right)\mathrm{d}T + R\ \mathrm{In}\left(\frac{P2}{P1}\right) \tag{28}$$

The heat capacity has been left behind the integral sign because even for most ideal gases, it varies with temperature. However, neither ideal gas heat capacities nor enthalpies vary with pressure, but ideal gas entropies do. Ideal gas heat capacity data is often presented in the form of an analytical equation. One common form used by NIST[7] is

$$C_p^* = A + BT + CT^2 + DT^2 + DT^3 + \frac{E}{T^2} \tag{29}$$

NIST tabulates the constants (*A*, *B*, etc.), for many compounds.

Real Fluids

Real fluids may be represented by the following general equation of state:

$$\mathrm{V} = \mathrm{Z}(\mathrm{P},\mathrm{T})\mathrm{RT}/\mathrm{P} \tag{30}$$

where *Z* is the compressibility factor, which can be a function of both *P* and *T*. It is common to report PVT data in the form of the compressibility factor. For an ideal gas, *Z*=1. Thus, the compressibility factor is a measure of the deviation from ideal gas behavior.

When Eq. 30 is substituted in Eqs. 18 and 19, one obtains

$$\mathrm{d}H = C_{\mathrm{P}}\mathrm{d}T - \left[\frac{RT^2}{P}\left(\frac{\partial Z}{\partial T}\right)_{\mathrm{P}}\right]\mathrm{d}P \tag{31}$$

$$\mathrm{dS} = \left(\frac{C_{\mathrm{P}}}{T}\right)\mathrm{d}T - R\left(Z + T\frac{\partial Z}{\partial T}\right)\frac{\mathrm{d}P}{P} \tag{32}$$

One consequence of these equations is that thermodynamic properties of real fluids can be calculated from their ideal gas–heat capacity (as a function of temperature) and their PVT behavior. This follows from the fact that to get from one state to another, these equations can first be integrated to zero pressure (which requires only PVT data), then integrated at zero pressure to the final temperature (which requires only ideal gas heat capacity data), and then integrated back to the end-point pressure.

Several approaches have been used to apply these equations for practical calculations. A first approach is to take known ideal gas-heat capacity and PVT data, and use them directly in these equations. Such data are not often available, and even when they are, this is a tedious process. Nevertheless, it has been done for a number of well-studied chemical species, most notably water. This is the basis of the steam tables, which have been put into analytical form for easy computer calculations.[8]

A second approach is through generalized correlations. An early attempt to describe the PVT behavior of many fluids was begun by Pitzer[9] through the use of his so- called acentric factor. His correlations have been extended so that they can be used to predict the more important thermodynamic properties of real no-polar fluids in both vapor and liquid phases.[10]

A third approach is to insert an equation of state into Eqs. 31 and 32 so that they may be integrated analytically. One excellent equation of state for low-density gases is the virial equation

$$Z = 1 + \frac{B(T)}{V} + \frac{C(T)}{V^2} + \frac{D(T)}{V^3} + \cdots \tag{33}$$

where B, C, D, etc., are functions of temperature only and are called the second, third, fourth, etc., virial coefficients. There is a great deal of data for the second virial coefficients.[11] Where data are not available, correlations have been developed to estimate them.[12,13] Third virial coefficients are usually not available, but prediction methods have been proposed.[14] Very little information is available for higher-order coefficients. Other equations of state have also enjoyed success in predicting properties of real fluids, both in the vapor and liquid phases. Among the more successful are the Soave–Redlich–Kwong[15] and Peng–Robinson[16] equations.

Solutions

The properties of solutions—whether they are gas or liquid mixtures—are not merely the mole fraction averages of the properties of the pure components. For any thermodynamic property, M, the exact equation for the properties of mixtures is

$$(P,T,\bar{x}) = \sum_i x_i \overline{M_i}(P,T,\bar{x}) \tag{34}$$

where $\overline{M}_i$ is called the "partial molal property of i" and $\bar{x}$ is a vector containing all of the compositions of the various constituents of the solution. The summation is carried out over all constituents. $\overline{M_i}$ is the property of constituent i at pressure, P, and temperature, T, as it exists in solution of composition $\bar{x}$. In general, this is different from its value as a pure substance.

The most general equation through which partial molal properties may be evaluated is

$$\overline{M_i} = \left(\frac{\partial (nM)}{\partial_{ni}} \right)_{P,T,\mathrm{nj}} \tag{35}$$

where n is the total number of moles and n_i is the number of moles of specie "i".

The partial derivative implies that the extensive property, *nM*, is differentiated with respect to the particular specie of interest, *i*, with pressure, temperature, and all of the other constituent amounts held constant. If properties of the various possible mixtures are known or can be approximated analytically, this equation provides a means to calculate their partial molal properties.

An important approximation for solutions is that of so- called "ideal solutions," whose properties are calculated as follows,

$$V^{\mathrm{id}} = \sum_i x^i V^i \tag{36}$$

$$H^{\mathrm{id}} = \sum_i x_i H_i \tag{37}$$

$$S^{\mathrm{id}} = \sum_i x_i S_i - R \sum_i x_i \operatorname{In} x_i \tag{38}$$

$$G^{\mathrm{id}} = \sum_i x_i G_i + RT \sum_i x_i \operatorname{In} x_i \tag{39}$$

where the superscript, "id," designates these as the properties for ideal solutions. All ideal gases are ideal solutions. The ideal-solution approximation is useful in some situations, but most liquids depart from it significantly. Notice also that properties involving entropy have a mixing effect (second terms on the right side of Eqs. 38 and 39) even if they are ideal solutions. This reflects the fact that the entropy of a solution is always higher than the combined entropies of its constituents.

The true properties of solutions can be calculated from their PVT behavior. For humid air, for example, the mixing effect is often neglected. However, one recent study done by M. Conde Company[17] includes virial coefficient information, including information for mixtures, sufficient to produce very accurate psychrometric properties of air-water mixtures.

Phase Equilibrium

Phase equilibrium can also be predicted from thermodynamic properties. The way this is done goes beyond the scope of this entry. However, several ideas can be stated here. One aspect of phase equilibria—namely pure-component vapor pressures—has been widely measured and can be used as a starting point for multicomponent phase equilibrium estimates. Vapor pressure data is often correlated by the Antoine equation,

$$\log_{10}\left(P^V\right) = A - \left(\frac{B}{T+C} \right) \tag{40}$$

where *A*, *B*, and *C* are constants. Such constants for many compounds also appear in the NIST collection.[18]

One calculations is the two-phase equilibrium between (1) pure special case of interest in energy liquid water and (2) a vapor phase one of whose components is water vapor. If the vapor phase is considered to be an ideal gas (and thus an ideal mixture), and if certain other assumptions are made, the composition of water in the vapor phase can be estimated by

$$yH_2O = \frac{P^V}{P} \tag{41}$$

where the vapor pressure can be calculated from the Antoine equation. This is a useful approximation that can be refined if PVT data for the vapor mixture are available.

Conclusions

This entry attempts to give an introduction to some of the basic principles of thermodynamics and how they relate to various energy engineering applications. The two main laws of thermodynamics govern the performance of any energy-conversion process. Their application requires the ability to estimate the thermodynamic properties of the various process streams. Such estimates are also based on the formalism that thermodynamics provides. Thus, thermodynamics provides the basic framework within which the design and evaluation of all energy-conversion equipment and processes must be performed.

Acknowledgments

Support for this work by VAST Power Systems, Elkhart, Indiana, is gratefully acknowledged.

Nomenclature

All thermodynamic properties have the units of energy or energy per unit time except where otherwise specified.

C_p	Heat capacity at constant pressure (energy/temperature)
e^{Ch}	Chemical exergy
e^{Ph}	Physical exergy
G	Gibbs free energy, or Gibbs function
H	Enthalpy
M	General thermodynamic property
M_i	Partial molal property of specie "i"
n	total number of moles
n_i	number of moles of specie "i"
P	Pressure (force/area)
P^V	Vapor pressure (force/area)
Q	Heat or heat flux added to a process (energy or energy per unit time)
Q_{ideal}	Heat or heat flux added to a completely reversible process (energy or energy per unit time)
R	Universal gas constant (energy/temperature)
S	Entropy (energy/temperature)
T	Temperature
T_0	Reference temperature or temperature of the surroundings
U	Internal energy
V	Volume (length^3)
W	Work or power produced by a process. Usually does not include work done by or on fluids entering or leaving the process (energy or energy per unit time)
W_{ideal}	Work or power produced by a perfectly reversible process (energy or energy per unit time)
W_{lost}	Lost work or power (energy or energy per unit time)
W_{total}	Total work or power produced by the process (energy or energy per unit time)
X_i	mole fraction of specie "i"
x	vector containing all mole fractions

yH_2O	Mole fraction of water vapor in a vapor phase
Z	Compressibility factor
Δ	Difference in the value between two states or differences between the properties of outlet and input streams
ΔH_f	Enthalpy of formation
ΔS_f	Entropy of formation (energy/temperature)

Subscripts

f	Formation
o	Base or reference state
C	Cold sink
H	Hot source
rev	Reversible

Superscripts

*	Ideal gas or ideal gas state
id	Ideal solution

References

1. Available at http://webbook.nist.gov/chemistry/form-ser. html (accessed July 2005).
2. Capehart, B.L.; Salas, C.E. *The Theory and Practice of Distributed Generation and On-Site CHP Systems*; The Association of Energy Engineers: Atlanta, GA, June 2003.
3. Bosul, Ulf. *Well-to-Wheel Studies, Heating Values and the Energy Conversation Principle,* available at http://www. efce.com/reports/E10.pdf (accessed July 2005).
4. Available at http://www.chpcentermw.org/pdfs/toolkit/7c_ rules_thumb.pdf (accessed July 2005).
5. Van Ness, H.C.; Abbott, M.M. Thermodynamics. In *Perry's Chemical Engineers' Handbook,* 7th Ed.; Perry, R.H., Green, D.W., Eds.; McGraw Hill: New York, 1997; 4-1–4–36; However, note that because of a different sign convention their formulas for ideal work have the opposite signs from those in this entry.
6. Van Ness, H.C.; Abbott, M.M. In *Perry's Chemical Engineers' Handbook,* 7th Ed.; Perry, R.H., Green, D.W., Eds.; McGraw Hill: New York, 1997; 4-1–4–36.
7. Available at http://webbook.nist.gov/chemistry/form-ser. html (accessed July 2005).
8. Available at http://www.cheresources.com/iapwsif971.pdf (accessed July 2005). A free downloadable Excel add-in which embodies the equations is also available on this Web site.
9. Pitzer, K.S. *Thermodynamics,* 3rd Ed.; Mc-Graw Hill: New York, 1995; App. 3.
10. Lee, B.I.; Kessler, M.G. A generalized thermodynamic correlation based on three-parameter corresponding states. AIChE J. **1975**, *21* (3), 510–527.
11. Daubert, T.E.; Danner, R.P. *Physical and Thermodynamic Properties of Pure Chemicals Data Compilation;* Taylor and Francis: Bristol, PA, 1994.
12. Tsonopoulos, C. An empirical correlation of second virial coefficients. AIChE J. **1974**, *20* (2), 263–272.
13. Liley, P.E.; Thomson, G.H.; Daubert, T.E.; Buck, E. Physical and chemical data. In *Perry's Chemical Engineers' Handbook,* 7th Ed.; Perry, R.H., Green, D.W., Eds.; McGraw Hill: New York, 1997; 2–355–2–358.
14. De Santis, R.; Grande, B. An equation for predicting third virial coefficients of nonpolar gases. AIChE J. **1979**, *25* (6), 931–938.
15. Soave, G. Equilibrium constants from a modified Redlich- Kwong equation of state. Chem. Eng. Sci. **1972**, *27,* 1197–1203.

16. Peng, D.-Y.; Robinson, B.B. A new two-constant equation of state. Ind. Eng. Chem. Fund. **1976**, *15*, 59–64.
17. Available at http://www.mrc-eng.com/Downloads/Moist% 20Air%20Props%20English.pdf (accessed July 2005).
18. Available at http://webbook.nist.gov/chemistry/form-ser. html (accessed July 2005).

III

CSS: Case Studies of Environmental Management

17

Energy Conversion: Coal, Animal Waste, and Biomass Fuel

Kalyan Annamalai,
Soyuz Priyadarsan,
Senthil Arumugam,
and John M.
Sweeten

Introduction and Objectives

The overall objective of this entry is to provide the basics of energy conversion processes and to present thermochemical data for coal and biomass fuels. Energy represents the capacity for doing work. It can be converted from one form to another as long as the total energy remains the same. Common fuels like natural gas, gasoline, and coal possess energy as chemical energy (or bond energy) between atoms in molecules. In a reaction of the carbon and hydrogen in the fuel with oxygen, called an oxidation reaction (or more commonly called combustion), carbon dioxide (CO_2) and water (H_2O) are produced, releasing energy as heat measured in units of kJ or Btu (see Table 1 for energy units). Combustion processes are used to deliver (i) work, using external combustion (EC) systems by generating hot gases and producing steam to drive electric generators as in coal fired power plans, or internal combustion (IC) engines by using the hot gases directly as in automobiles or gas turbines; and (ii) thermal energy, for applications to manufacturing processes in metallurgical and chemical industries or agricultural product processing.

TABLE 1 Energy Units and Terminology

The section on energy Units and Conversion factors in Energy is condensed from Chapter 01 of Combustion Engineering by Annamalai and Puri [2005] and Tables.

Energy Units

1 Btu (British thermal unit) = 778.14 ft lbf = 1.0551 kJ, 1 kJ = 0.94782 Btu = 25,037 $lbmft/s^2$

1 mBtu = 1k Btu = 1000 Btu, 1 mmBtu = 1000 k Btu = 106 Btu, 1 trillion Btu = 109 Btu or 1 giga Btu

1 quad = 1015 Btu or 1.05 × 1015 kJ or 2.93 × 10n kW h

1 Peta J = 1015 J = 1012 kJ»0.00095 Quads

1 kilowatt-hour of electricity = 3,412 BTU = 3.6 Mj

1 cal: 4.1868 J, One (food) calorie = 1000 cal or 1 Cal

1 kJ/kg = 0.43 Btu/lb, 1 Btu/lb = 2.326 kJ/kg

1 kg/GJ = 1 g/MJ = 2.326 lb/mmBtu; 1 lb/mmBtu = 0.430 kg/GJ = 0.430 g/MJ

1 Btu/SCF = 37 kJ/m^3

1 Therm = 105 Btu = 1.055 × 105 kJ

1 m^3/GJ = 37.2596 ft^3/mmBTU

1 hp = 0.7064 Btu/s = 0.7457 kW = 745.7 W = 550 lbf ft/s = 42.41 Btu/min

1 boiler HP = 33475 Btu/h, 1 Btu/h = 1.0551 kJ/h

1 barrel (42 gal) of crude oil = 5,800,000 Btu = 6120 MJ

1 gal of gasoline = 124,000 Btu = 131 MJ

1 gal of heating oil = 139,000 Btu = 146.7 MJ, 1 gal of diesel fuel = 139,000 Btu = 146.7 MJ

1 barrel of residual fuel oil = 6,287,000 Btu = 6633 MJ

1 cubic foot of natural gas = 1,026 Btu = 1.082 MJ, 1 Ton of Trash = 150 kWh

1 gal of propane = 91,000 Btu = 96 MJ, 1 short ton of coal = 20,681,000 Btu = 21821 MJ

Emission reporting for pollutants: (i) parts per million (ppm), (ii) normalized ppm, (iii) emission Index (EI) in g/kg fuel, (iv) g/GJ, v) mg/ m^3 of flue gas:

Conversions in emissions reporting: (ii) normalized ppm = ppm × (21-O_2% std)/(21-O_2% measured); (iii) EI of species k: C % by mass in fuel × mol Wt of k × ppm of species k × 10–3/{12.01(CO_2% + CO%)}, (iv) g/GJ = EI/ {HHV in GJ/kg}; (v) mg/m^3 = ppm of species k × Mol Wt of k/24.5

Volume of 1 kmol (SI) and 1 lb mole (English) of an ideal gas at STP conditions defined below: Pressure at 101.3 kPa (1 atm, 14.7 psia, 29.92 in.Hg, 760 Torr) fixed; T changes depending upon type of standard adopted

Scientific (or SATP, standard ambient T and P)	US standard (1976) or ISA (International standard atmosphere)	NTP (gas industry reference	Chemists-standard-atmosphere (CSA)
25°C (77°F)	15°C (60°F)	20°C(68°F), 101.3 kPa	0°C (32°F),
24.5 m^3/kmol (392 ft^3/lb mole); $\rho_{air,SATP} = 1.188$ kg/m^3 = 0.0698 lbm/ft^3	23.7 m^3/kmol (375.6 ft^3/lb mole); $\rho_{air,ISA} = 1.229$ kg/m^3 = 0.0767 lbm/ft^3	24.06 m^3/kmole or 385 ft^3/lb mole; $\rho_{air,NTP} = 1.208$ kg/m^3 = 0.0754 lbm/ft^3	22.4 m^3/kmol (359.2 ft^3/lb mole), $\rho_{air,CSA} = 1.297$ kg/m^3 = 0.0810 lbm/ft^3

Fuels can be naturally occurring (e.g., fossil fuels such as coal, oil, and gas, which are residues of ancient plant or animal deposits) or synthesized (e.g., synthetic fuels). Fuels are classified according to the phase or state in which they exist: as gaseous (e.g., natural gas), liquid (e.g., gasoline or ethanol), or solid (e.g., coal, wood, or plant residues). Gaseous fuels are used mainly in residential applications (such as water heaters, home heating, or kitchen ranges), in industrial furnaces, and in boilers. Liquid fuels are used in gas turbines, automotive engines, and oil burners. Solid fuels are used mainly in boilers and steelmaking furnaces.

During combustion of fossil fuels, nitrogen or sulfur in the fuel is released as NO, NO_2 (termed generally as NO_x) and SO_2 or SO_3 (termed as SO_x). They lead to acid rain (when SO_x or NO_x combine with H_2O and fall as rain) and ozone depletion. In addition, greenhouse gas emissions (CO_2, CH_4, N_2O,

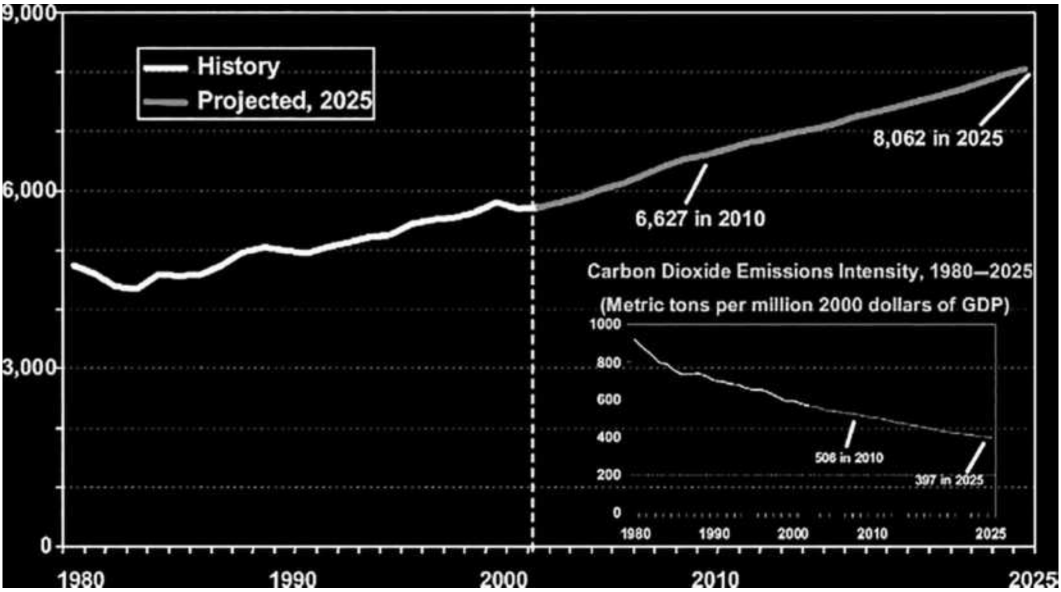


FIGURE 1 Total CO_2 emission in million metric tons per year: History and Projected 1980–2025.
Source: DOE-EIA.[1]

CFCs, SF_6, etc.,) are becoming a global concern due to warming of the atmosphere, as shown in Figure 1 for CO_2 emissions. Global surface temperature has increased by 0.6°C over the past 100 years. About 30%–40% of the world's CO_2 is from fossil fuels. The Kyoto protocol, signed by countries that account for 54% of the world's fossil based CO_2 emissions, calls for reduction of greenhouse gases by 5% from 1990 levels over the period from 2008 to 2012.

The total worldwide energy consumption is 421.5 quads of energy in 2003 and is projected to be 600 quads in 2020, while U.S. consumption in 2004 is about 100 quads and is projected to be 126 quads in 2020. The split is as follows: 40 quads for petroleum, 23 for natural gas, 23 for coal, 8 for nuclear power, and 6 for renewables (where energy is renewed or replaced using natural processes) and others sources. Currently, the United States relies on fossil fuels for 85% of its energy needs. Soon, the U.S. energy consumption rate which distributed as electrical power (40%), transportation (30%), and heat (30%), will outpace the growth in the energy production rate, increasing reliance on imported oil. The Hubbert peak theory (named after Marion King Hubbert, a geophysicist with Shell Research Lab in Houston, Texas) is based upon the rate of extraction and depletion of conventional fossil fuels, and predicts that fossil-based oil would peak at about 12.5 billion barrels per year worldwide some time around 2000. The power cost and percentage use of coal in various U.S. states varies from 10 cents (price per kWh) at 1% coal use for power generation in California to 48 cents at 94% use of coal in Utah.

Biomass is defined as "any organic material from living organisms that contains stored sunlight (solar energy) in the form of chemical energy."[1] These include agro-based materials (vegetation, trees, and plants); industrial wastes (sawdust, wood chips, and crop residues); municipal solid wastes (MSWs), which contain more than 70% biomass (including landfill gases, containing almost 50% CH_4); and animal waste. Biomass is a solid fuel in which hydrogen is locked with carbon atoms. Biomass production worldwide is 145 billion metric tons. Biomass now supplies 3% of U.S. energy, and it could be increased to as high as 20%. Renewable energy sources (RES) include biomass, wind, hydro, solar, flowing water or hydropower, anaerobic digestion, ocean thermal (20°C temperature difference), tidal energies, and geothermal (a nonsolar source of energy), and these supply 14% of the world demand. The RES constitute only 6%, while coal, petroleum, and natural gas account for 23%, 40%, and 24%, respectively. About 9%

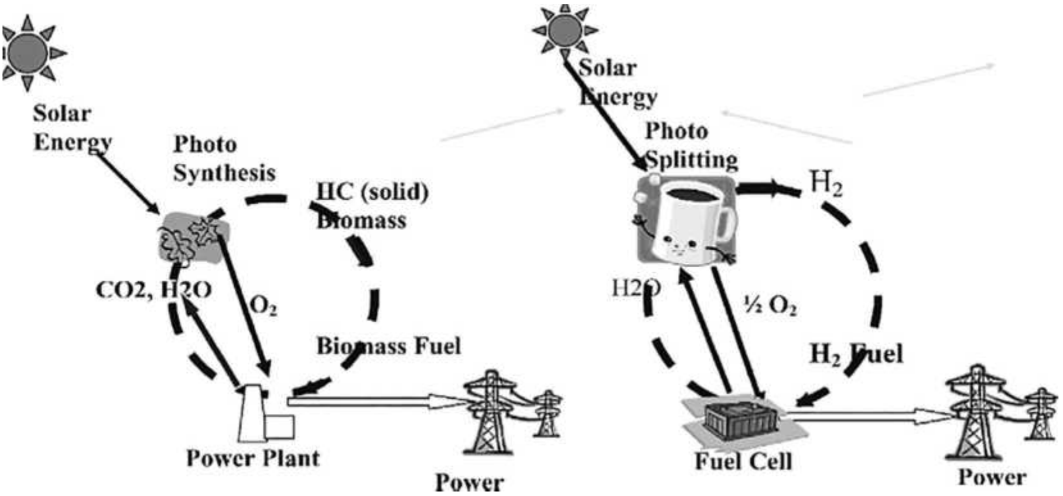


FIGURE 2 Comparison of biomass energy and H_2 energy cycles.

of the world's electricity is from RES, and 65% of the electricity contributed by biomass. About 97% of energy conversion from biomass is by combustion. Many U.S. states have encouraged the use of renewables by offering REC (Renewable Energy Credits). One REC = 1 MW/h = 3.412 mmBtu; hence the use of 1 REC is equivalent to replacing approximately 1500 lb of coal, reducing emission of NO_x and SO_x by 1.5 lb for every 1 REC, assuming that emissions of NO_x and SO_x are 0.45 lb per mmBtu generated by coal. Several emission-reporting methods and conversions are summarized in Table 1. Recently, H_2 is being promoted as a clean-burning, non-global-warming, and pollution-free fuel for both power generation and transportation.

Figure 2 shows a comparison between biomass and hydrogen energy cycles. In the biomass cycle, photosynthesis is used to split CO_2 into C and O_2, and H_2O into H_2 and O_2, producing Hydrocarbons (HC) fuel (e.g., leaves) and releasing O_2. The O_2 released is used to combust the HC and produce CO_2 and H_2O, which are returned to produce plant biomass (e.g., leaves) and O_2. On the other hand, in the hydrogen cycle, H_2O is disassociated using the photosplitting process to produce H_2 and O_2, which are then used for the combustion process. The hydrogen fuel can be used in fuel cells to obtain an efficient conversion. Photosynthesis is water intensive; most of the water supplied to plants evaporates through leaves into the atmosphere, where it re-enters the hydrology cycle.

This entry is organized in the following format: (i) coal and biosolid properties; (ii) coal and biosolid pyrolysis (a process of thermal decomposition in the absence of oxygen), combustion, and gasification; (iii) combustion by cofiring coal with biosolids; (iv) gasification of coal and biosolids (a process that includes pyrolysis, partial oxidation due to the presence of oxygen, and hydrogenation); and (v) reburn for NO_x reduction.

Fuel Properties

Fuel properties play a major role in the selection, design, and operation of energy conversion systems.

Solid Fuels

The primary solid fuel widely used in power plants is coal containing combustibles, moisture, and intrinsic mineral matter originating from dissolved salts in water. During the "coalification" process, lignite, the lowest rank of coal (low C/O ratio), is produced first from peat, followed by subbituminous (black lignite, typically low sulfur, noncaking), bituminous (soft coal that tends to stick when heated and is typically high in S), and finally anthracite (dense coal; has the highest carbon content, >90%, low volatile <15%) with a gradual increase in the coal C/O ratio. The older the coal, the higher its rank. Anthracite (almost carbon) is the highest-ranked coal, with a high heating value. To classify coals and ascertain the quality of coal, it is essential to perform proximate and ultimate analyses according to American Society of Testing Materials (ASTM) standards.

Proximate Analysis (ASTM D3172)

A solid fuel consists of combustibles, ash, and moisture. Combustibles together with ash are called the solid content of fuel. A proximate analysis provides the following information: surface moisture (SM) or dry loss (DL), i.e., moisture in air-dried coal; the inherent moisture in the coal (M); volatile matter (VM; produced by pyrolysis, a thermal decomposition process resulting in release of water, gases, oil and tar); fixed carbon (FC; skeletal matter left after release of volatiles); mineral matter (MM; inert collected with solid fuel); and heating value (HV). On combustion, the MM may be partially oxidized or reduced, and the material left after combustion of C and H in the fuel is called ash (CaO, $CaCO_3$, Fe_2O_3, FeO, etc.).

Table 2 shows comparative proximate analyses of coal, advanced feedlot biomass (FB, low-ash cattle manure; see "Coal and Bio-Solids Cofiring"), and litter biomass (LB, chicken manure).[2] Feedlot manure has higher moisture, nitrogen, chlorine, and ash content than coal. With aging or composting, the VM in manure decreases as a result of the gradual release of hydrocarbon gases or dehydrogenation, but fuel becomes more homogeneous.

Ultimate/Elemental Analysis (ASTM D3176)

Ultimate analysis is used to determine the chemical composition of fuels in terms of either the mass percent of their various elements or the number of atoms of each element. The elements of interest are C, H, N, O, S, Cl, P, and others. It can be expressed on an "as received" basis, on a dry basis (with the moisture in the solid fuel removed), or on a dry ash free (DAF) basis (also known as the moisture ash free basis MAF). Tables 3 and 4 show the ultimate analyses of various types of coal and biomass fuels.[3] While nitrogen is not normally present in natural gas, coal has 1%–1.5%; cattle manure and chicken waste contain high amounts of N (Table 2).

Heating Value (ASTM D3286)

The gross or higher heating value (HHV) of a fuel is the amount of heat released when a unit (mass or volume) of the fuel is burned. The HHV of solid fuel is determined using ASTM D3286 with an isothermal jacket bomb calorimeter. For rations fed to animals and animal waste fuels, the HHV for DAF roughly remains constant at about 19,500 kJ/kg (8400 Btu/lb),[4] irrespective of stage of decomposition of animal waste. The HHV can also be estimated using the ultimate analysis of the fuel and the following empirical relation from Boie[5]:

$$\mathrm{HHV}_{\mathrm{fuel}}\left(\mathrm{kJ/kg\ fuel}\right) = 35{,}160\ Y_c + 116{,}225\ Y_H - 11{,}090\ Y_O + 6280\ Y_N + 10465\ Y_S \tag{1}$$

$$\mathrm{HHV}_{\mathrm{fuel}}\left(\mathrm{BTU/lb\ fuel}\right) = 15{,}199\ Y_C + 49{,}965\ Y_H - 4768\ Y_O + 2700\ Y_N + 4499\ Y_S, \tag{2}$$

where Y denotes the mass fraction of an element C, H, O, N, or S in the fuel. The higher the oxygen content, the lower the HV, as seen in biomass fuels.

Annamalai et al. used the Boie equation for 62 kinds of biosolids with good agreement.[6] For most biomass fuels and alcohols, the HHV in kilojoules per unit mass of stoichiometric oxygen is constant at 14,360–14,730 kJ/kg of O_2 (6165–6320 Btu/lb of O_2).[7]

TABLE 2 Coal, Advanced Feedlot Biomass (FB) and Litter Biomass (LB)

Parameter	Wyoming Coal	Cattle Manure (FB)	Chicken Manure (LB)a	Advanced Feedlot Biomass (AFB)b	High-Ash Feedlot Biomass (HFB)b
Dry loss (DL)	22.8	6.8	7.5	10.88	7.57
Ash	5.4	42.3	43.8	14.83	43.88
FC	37.25	40.4	8.4	17.33	10.28
VM	34.5	10.5	40.3	56.97	38.2
C	54.1	23.9	39.1	50.08	49.27
H	3.4	3.6	6.7	5.98	6.13
N	0.81	2.3	4.7	38.49	38.7
O	13.1	20.3	48.3	4.58	4.76
S	0.39	0.9	1.2	0.87	0.99
Cl	<0.01%	1.2			
HHV-as received (kJ/kg)	21385	9560	9250	14983	9353
T adiab, Equilc	2200 K (3500°F)	2012 K (3161°F)			
DAF formula	$CH_{0.76}O_{0.18}N_{0.013}S_{0.0027}$	$CH_{1.78}O_{.64}N_{.083}S_{.014}$	$CH_{2.04}O_{0.93}N_{0.10}S_{0.012}$	$CH_{1.4184}O_{0.5764}N_{0.078}S_{0.0056}$	$CH_{1.4775}O_{0.5892}N_{0.083}S_{0.0076}$
HHV-DAF (kJ/kg)	29785	18785	18995	20168	19265
CO_2, g/GJ					
N, g/GJ					
S, g/GJ					

[a] Priyadarsan et al.

[b] Priyadarsan et al.[37]

[c] Equilibrium temperature for stoichiometric mixture from THERMOLAB Spreadsheet software for any given fuel of known composition (Annamalai and Puri.[36] website http://www.crcpress.com/e_products/downloads/download.asp?cat_no=2553)

TABLE 3 Coal Composition (DAF basis)

ASTM Rank	State (U.S.A.)	Ash, % (dry)	C	H	N	S*	O**	HHV Est kj/kg	CO$_2$ kg/GJ	N kg/GJ	S kg/GJ
Lignite	ND	11.6	63.3	4.7	0.48	0.98	30.5	24,469	94.8	0.196	0.401
Lignite	MT	7.7	70.7	4.9	0.8	4.9	22.3	28,643	90.4	0.279	1.711
Lignite	ND	8.2	71.2	5.3	0.56	0.46	22.5	28,782	90.7	0.195	0.160
Lignite	TX	9.4	71.7	5.2	1.3	0.72	21.1	29,070	90.4	0.447	0.248
Lignite	TX	10.3	74.3	5	0.37	0.51	19.8	29,816	91.3	0.124	0.171
Sbb. A	WY	8.4	74.3	5.8	1.2	1.1	17.7	31,092	87.6	0.386	0.354
Sbb. C	WY	6.1	74.8	5.1	0.89	0.3	18.9	30,218	90.7	0.295	0.099
HVB	IL	10.8	77.3	5.6	1.1	2.3	13.6	32,489	87.2	0.339	0.708
HVC	IL	10.1	78.8	5.8	1.6	1.8	12.1	33,394	86.5	0.479	0.539
HVB	IL	11.8	80.1	5.5	1.1	2.3	11.1	33,634	87.3	0.327	0.684
HVB	UT	4.8	80.4	6.1	1.3	0.38	11.9	34,160	86.2	0.381	0.111
HVA	wv	7.6	82.3	5.7	1.4	1.8	8.9	34,851	86.5	0.402	0.516
HVA	KY	2.1	83.8	5.8	1.6	0.66	8.2	35,465	86.6	0.451	0.186
MV	AL	7.1	87	4.8	1.5	0.81	5.9	35,693	89.3	0.420	0.227
LV	PA	9.8	88.2	4.8	1.2	0.62	5.2	36,153	89.4	0.332	0.171
Anthracite	PA	7.8	91.9	2.6	0.78	0.54	4.2	34,974	96.3	0.223	0.154
Anthracite	PA	4.3	93.5	2.7	0.24	0.64	2.9	35,773	95.8	0.067	0.179

Notes: HHV$_{est}$: Boie Equation. CO$_2$ in g/MJ or kg/GJ = C content in % ×36645/{HHV in kJ/kg}. CO$_2$ in lb per mmBtu = Multiply CO$_2$ in (g /MJ) or kg/GJ by 2.32. N in g/MJ or kg/GJ = N% × 10000/{HHV in kJ/kg}. For NO$_x$ estimation, multiply N content in g/MJ by 1.15 to get NO$_x$ in g/MJ which assumes 35% conversion of fuel N; For SO$_2$ estimation, multiply S content in g/MJ by 2 to get SO$_2$ in g/MJ assuming 100% conversion of fuel S (Multiply HHV in kJ/kg by 0.430 to get Btu/lb); *Organic sulfur; **by difference.

TABLE 4 Ultimate Analyses and Heating Values of Biomass Fuels

Biomass	C	H	O	N	S	Residue	Measured HHVm	[a]Estimated HHV	CO$_2$ g/MJ	N, g/MJ	S, g/MJ
Field crops											
Alfalfa seed straw	46.76	5.40	40.72	1.00	0.02	6.07	18.45	18.27	92.9	0.542	0.011
Bean straw	42.97	5.59	44.93	0.83	0.01	5.54	17.46	16.68	90.2	0.475	0.006
Corn cobs	46.58	5.87	45.46	0.47	0.01	1.40	18.77	18.19	90.9	0.250	0.005
Corn stover	43.65	5.56	43.31	0.61	0.01	6.26	17.65	17.05	90.6	0.346	0.006
Cotton stalks	39.47	5.07	39.14	1.20	0.02	15.10	15.83	15.51	91.4	0.758	0.013
Rice straw (fall)	41.78	4.63	36.57	0.70	0.08	15.90	16.28	16.07	94.0	0.430	0.049
Rice straw (weathered)	34.60	3.93	35.38	0.93	0.16	25.00	14.56	12.89	87.1	0.639	0.110
Wheat straw	43.20	5.00	39.40	0.61	0.11	11.40	17.51	16.68	90.4	0.348	0.063
Switchgrass[b]	42.02	6.30	46.10	0.77	0.18	4.61	15.99	15.97	96.3	0.482	0.113
Orchard prunings											
Almond prunings	51.30	5.29	40.90	0.66	0.01	1.80	20.01	19.69	93.9	0.330	0.005
Black Walnut	49.80	5.82	43.25	0.22	0.01	0.85	19.83	19.50	92.0	0.111	0.005
English Walnut	49.72	5.63	43.14	0.37	0.01	1.07	19.63	19.27	92.8	0.188	0.005
Vineyard prunings											
Cabernet Sauvignon	46.59	5.85	43.90	0.83	0.04	2.71	19.03	18.37	89.7	0.436	0.021

(*Continued*)

TABLE 4 (*Continued*) Ultimate Analyses and Heating Values of Biomass Fuels

Biomass	C	H	O	N	S	Residue	Measured HHVm	[a]Estimated HHV	CO_2 g/MJ	N, g/MJ	S, g/MJ
Chenin Blanc	48.02	5.89	41.93	0.86	0.07	3.13	19.13	19.14	92.0	0.450	0.037
Pinot Noir	47.14	5.82	43.03	0.86	0.01	3.01	19.05	18.62	90.7	0.451	0.005
Thompson seedless	47.35	5.77	43.32	0.77	0.01	2.71	19.35	18.60	89.7	0.398	0.005
Tokay	47.77	5.82	42.63	0.75	0.03	2.93	19.31	18.88	90.7	0.388	0.016
Energy Crops											
Eucalyptus											
Camaldulensis	49.00	5.87	43.97	0.30	0.01	0.72	19.42	19.19	92.5	0.154	0.005
Globulus	48.18	5.92	44.18	0.39	0.01	1.12	19.23	18.95	91.8	0.203	0.005
Grandis	48.33	5.89	45.13	0.15	0.01	0.41	19.35	18.84	91.5	0.078	0.005
Casuarina	48.61	5.83	43.36	0.59	0.02	1.43	19.44	19.10	91.6	0.303	0.010
Cattails	42.99	5.25	42.47	0.74	0.04	8.13	17.81	16.56	88.5	0.415	0.022
Popular	48.45	5.85	43.69	0.47	0.01	1.43	19.38	19.02	91.6	0.243	0.005
Sudan grass	44.58	5.35	39.18	1.21	0.08	9.47	17.39	17.63	93.9	0.696	0.046
Forest residue											
Black Locust	50.73	5.71	41.93	0.57	0.01	0.97	19.71	19.86	94.3	0.289	0.005
Chaparral	46.9	5.08	40.17	0.54	0.03	7.26	18.61	17.98	92.3	0.290	0.016
Madrone	48	5.96	44.95	0.06	0.02	1	19.41	18.82	90.6	0.031	0.010
Manzanita	48.18	5.94	44.68	0.17	0.02	1	19.3	18.9	91.5	0.088	0.010
Ponderosa Pine	49.25	5.99	44.36	0.06	0.03	0.3	20.02	19.37	90.1	0.030	0.015
Ten Oak	47.81	5.93	44.12	0.12	0.01	2	18.93	18.82	92.6	0.063	0.005
Redwood	50.64	5.98	42.88	0.05	0.03	0.4	20.72	20.01	89.6	0.024	0.014
White Fur	49	5.98	44.75	0.05	0.01	0.2	19.95	19.22	90.0	0.025	0.005
Food and fiber processing wastes											
Almond hulls	45.79	5.36	40.6	0.96	0.01	7.2	18.22	17.89	92.1	0.527	0.005
Almond shells	44.98	5.97	42.27	1.16	0.02	5.6	19.38	18.14	85.0	0.599	0.010
Babassu husks	50.31	5.37	42.29	0.26	0.04	1.73	19.92	19.26	92.5	0.131	0.020
Sugarcane bagasse	44.8	5.35	39.55	0.38	0.01	9.79	17.33	17.61	94.7	0.219	0.006
Coconut fiber dust	50.29	5.05	39.63	0.45	0.16	4.14	20.05	19.2	91.9	0.224	0.080
Cocoa hulls	48.23	5.23	33.09	2.98	0.12	10.25	19.04	19.56	92.8	1.565	0.063
Cotton gin trash	39.59	5.26	36.33	2.09		16.68	16.42	16.13	88.4	1.273	0.000
Macadamia shells	54.41	4.99	39.69	0.36	0.01	0.56	21.01	20.55	94.9	0.171	0.005
Olive pits	48.81	6.23	43.48	0.36	0.02	1.1	21.39	19.61	83.6	0.168	0.009
Peach pits	53	5.9	39.14	0.32	0.05	1.59	20.82	21.18	93.3	0.154	0.024
Peanut hulls	45.77	5.46	39.56	1.63	0.12	7.46	18.64	18.82	90.0	0.874	0.064
Pistachio shells	48.79	5.91	43.41	0.56	0.01	1.28	19.26	19.25	92.8	0.291	0.005
Rice hulls	40.96	4.3	35.86	0.4	0.02	18.34	16.14	15.45	93.0	0.248	0.012
Walnut shells	49.98	5.71	43.35	0.21	0.01	0.71	20.18	19.45	90.8	0.104	0.005
Wheat dust	41.38	5.1	35.19	3.04	0.19	15.1	16.2	16.78	93.6	1.877	0.117

[a] HHV based on Boie equation.

[b] Aerts et al[20]; [Adapted from Ebeling and Jenkins[3] and AnnamalaiJ[17] See foot note of Table 3.2 for conversions to English units and estimation of NO_x and SO_2 emissions.

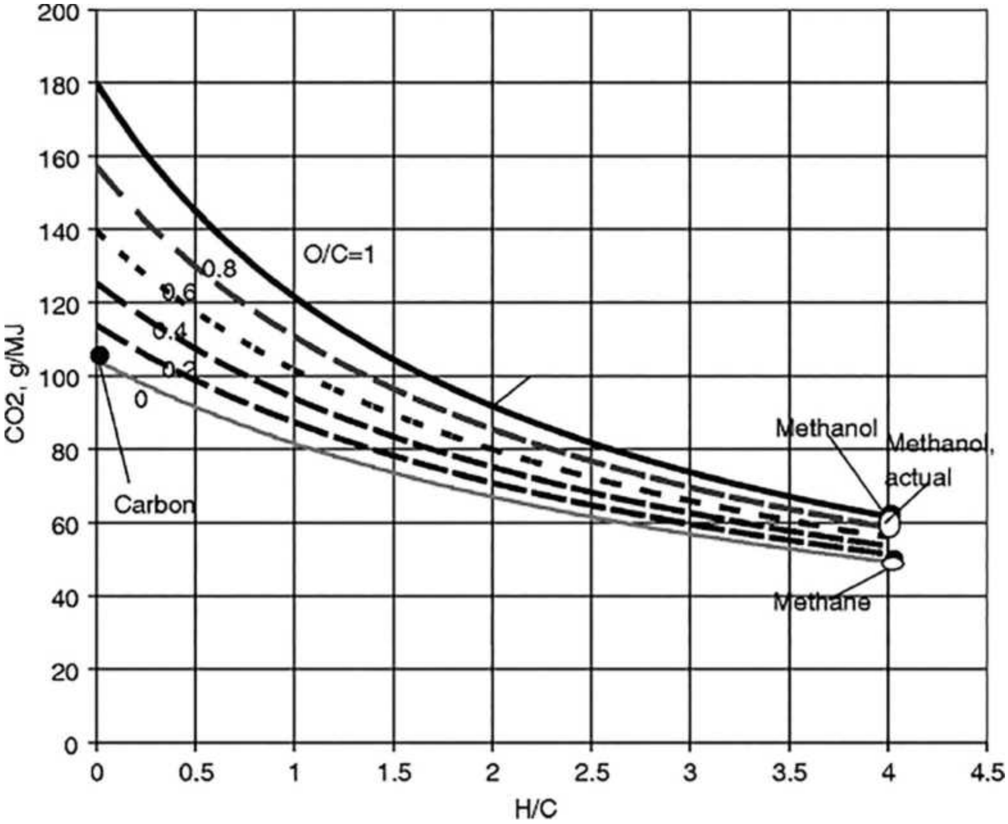


FIGURE 3 Emission of CO_2 as a function of H/C and O/C atom ratios in hydrocarbon fuels.
Source: Adapted from Annamalai and Puri.[8]

Estimate of CO_2 Emission

Using the Boie-based HVs for any fuel of known elemental composition, one can plot the CO_2 emission in g/MJ (Figure 3) as a function of H/C and O/C ratios.[8] Comparisons for selected fuels with known experimental HVs are also shown in the same figure. Coal, with H/C ratio ≈ 0.5, releases the highest CO_2, while natural gas (mainly CH_4) emits the lowest CO_2. Because the United States uses fossil fuels for 86% of its energy needs (100 quads), the estimated CO_2 emission is 6350 million ton/year, assuming that the average CO_2 emission from fossil fuels is 70 kg/GJ (methane: 50 kg/GJ vs coal: 90 kg/GJ). Figure 1 seems to confirm such estimation within a 10% error.

Flame Temperature

Figure 4 shows a plot of maximum possible flame temperature vs moisture percentage with combustion for biomass fuels. The result can be correlated as follows[4]:

$$T(\mathrm{K}) = 2290 - 1.89\mathrm{H_2O} + 5.06\ \mathrm{Ash} - 0.309\ \mathrm{H_2O}\,Ash - 0.180\ \mathrm{H_2O_2} - 0.108\ \mathrm{Ash}^2 \quad (3)$$

$$T(^{\circ}\mathrm{F}) = 3650 - 3.40\ \mathrm{H_2O} + 9.10\ \mathrm{Ash} - 0.556\ \mathrm{H_2O\ Ash} - 0.324\ \mathrm{H_2O_2} - 0.194\ \mathrm{Ash}^2 \quad (4)$$

The adiabatic flame temperature decreases if the ash and moisture contents increase.

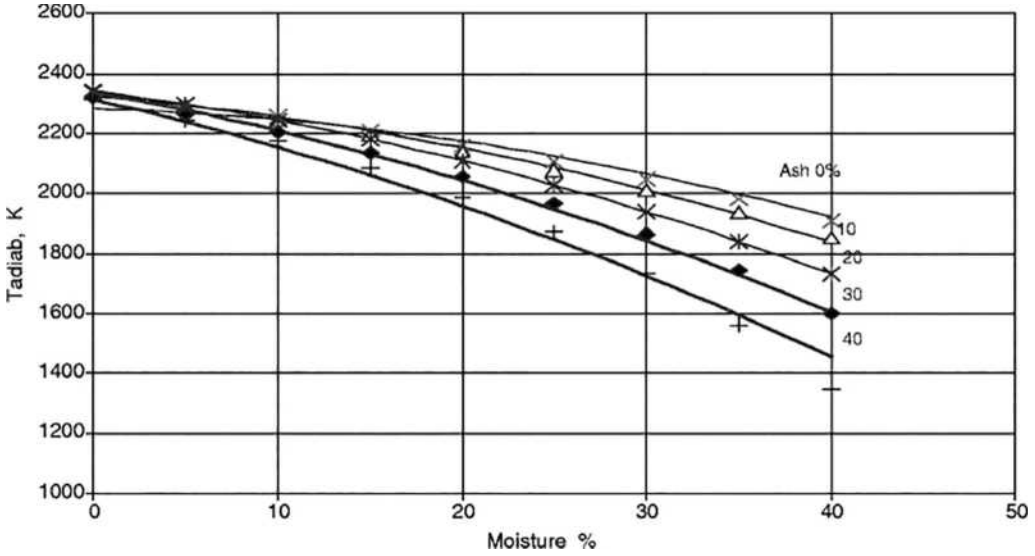


FIGURE 4 Correlation of adiabatic flame temperature with moisture and ash contents.

Flue Gas Volume

The flue gas volume for C–H–O is almost independent of O/C ratios. The fit at 6% O_2 in products gives the following empirical equation for flue gas volume (m^3/GJ) at SATP[8]:

$$\text{Flue gas}_{\text{vol}}\left(\text{m}^3/\text{GJ}\right)=4.96\left(\frac{\text{H}}{\text{C}}\right)^2-38.628\left(\frac{\text{H}}{\text{C}}\right)+389.72 \tag{5}$$

$$\text{Flue gas}_{\text{vol}}\left(\text{ft}^3/\text{mmBtu}\right)=184.68\left(\frac{\text{H}}{\text{C}}\right)^2-1439.28\left(\frac{\text{H}}{\text{C}}\right)+14520.95 \tag{6}$$

Liquid Fuels

Liquid fuels, used mainly in the transportation sector, are derived from crude oil, which occurs naturally as a free-flowing liquid with a density of $\rho \approx 780\,\text{kg/m}^3$–$1000\,\text{kg/m}^3$, containing 0.1% ash and 0.15%–0.5% nitrogen. Crude oil normally contains a mixture of hydrocarbons, and as such, the “boiling” temperature keeps increasing as the oil is distilled. Most fuel oils contain 83%–88% carbon and 6%–12% (by mass) hydrogen.

Gaseous Fuels

The gaseous fuels are cleaner-burning fuels than liquid and solid fuels. They are a mixture of HC but dominated by highly volatile CH_4 with very little S and N. Natural gas is transported as liquefied natural gas (LNG) and compressed natural gas (CNG), typically at 150–250 bars. Liquefied petroleum gas (LPG) is a byproduct of petroleum refining, and it consists mainly of 90% propane. A low-Btu gas contains 0–7400 kJ/SCM (Standard Cubic Meter, 0–200 Btu/SCF, standards defined in Table 1); a medium-Btu gas. 7400–14,800 kJ/SCM (200–400 Btu/SCF); and a high-Btu gas, above 14,800 kJ/SCM (more than 400 Btu/ SCF). Hydrogen is another gaseous fuel, with a heat value of 11,525 kJ/SCM (310 Btu/SCF). Because the fuel quality (heat value) may change when fuel is switched, the thermal output rate at a fixed gas-line pressure changes when fuels are changed.

Coal and Biomass Pyrolysis, Gasification, and Combustion

Typically, coal densities range from 1100 kg/m^3 for low-rank coals to 2330 kg/m^3 for high-density pyrolytic graphite, while for biomass, density ranges from 100 kg/ m^3 for straw to 500 kg/m^3 for forest wood.[9] The bulk density of cattle FB as harvested is 737 kg/m^3 (CF) for high ash (HA-FB) and 32 lbs/CF for low ash (LA-FB).[10] The processes during heating and combustion of coal are illustrated in Figure 5, and they are similar for biomass except for high VM. The process of release of gases from solid fuels in the absence of oxygen is called pyrolysis, while the combined process of pyrolysis and partial oxidation of fuel in the presence of oxygen is known as gasification. If all combustible gases and solid carbon are oxidized to CO_2 and H_2O, the process is known as combustion.

Pyrolysis

Solid fuels, like coal and biomass, can be pyrolyzed (thermally decomposed) in inert atmospheres to yield combustible gases or VM. While biomass typically releases about 70%–80% of its mass as VM (mainly from cellulose and hemicellulose) with the remainder being char, mainly from lignin content of biomass, coal releases 10%–50% of its mass as VM, depending upon its age or rank. Typically, a medium-rank coal consists of 40% VM and 60% FC, while a high-rank coal has about 10% VM. Bituminous coal pyrolyzes at about 700 K (with 1% mass loss for heating rates <100°C/s), as in the case of most plastics. Pyrolytic products range from lighter volatiles like CH_4, C_2H_4, C_2H_6, CO, CO_2, H_2, and H_2O to heavier molecular mass tars. Apart from volatiles, nitrogen is also evolved from the fuel during pyrolysis in the form of HCN, NH_3, and other compounds or, more generally, XN.

Sweeten et al. performed the thermogravimetric analysis (TGA) of feedlot manure.[4] The results are shown in Figure 6. In the case of manure, drying occurred between 50 and 100°C, pyrolysis was initiated around 185°C–200°C for a heating rate of 80°C/min, and the minimum ignition temperature was approximately 528°C. The gases produced during biomass pyrolysis can also be converted into transportation fuels like biodiesel, methanol, and ethanol, which may be used either alone or blended with gasoline.

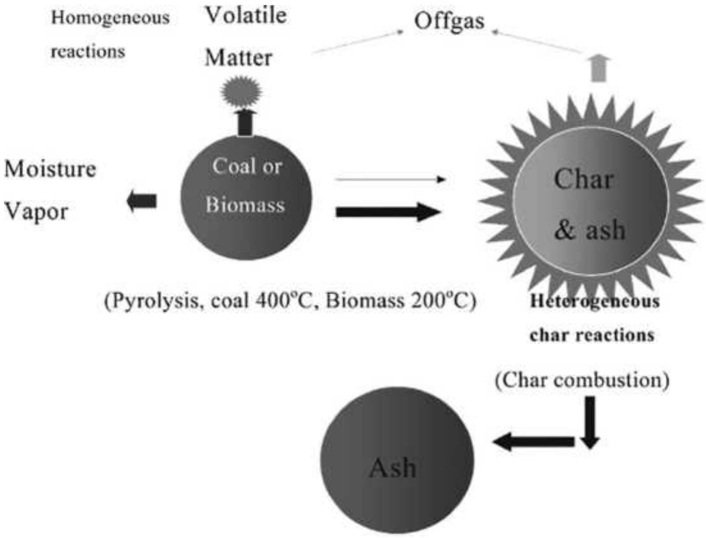


FIGURE 5 Processes during coal pyrolysis, gasification, and combustion.

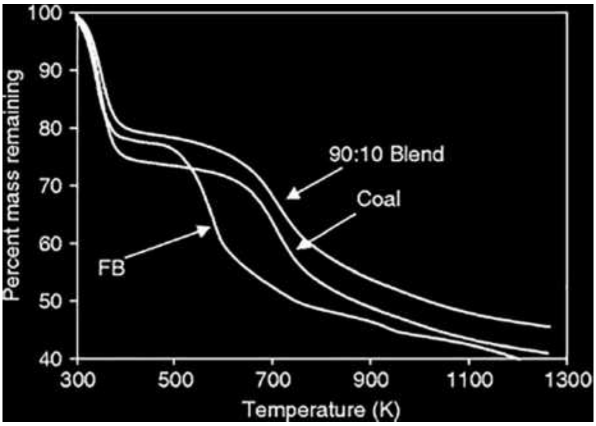


FIGURE 6 Thermo-gravimetric analyses of Feedlot Biomass (FB or cattle manure), coal, and 90:10 coal: FB blends.
Source: Sweeten et al.[4]

Volatile Oxidation

Once released, volatiles (HC, CO, H_2, etc.,) undergo oxidation within a thin gas film surrounding the solid fuel particle. The oxidation for each HC involves several steps. The enveloping flame, due to volatile combustion, acts like a shroud by preventing oxygen from reaching the particle surface for heterogeneous oxidation of char. Following Dryer,[11] the one-step global oxidation of a given species can be written as

$$\text{Fuel} + \upsilon_{CO_2} O_2 \overset{\text{oxidation}}{\rightarrow} \upsilon_{CO_2} CO_2 + \upsilon_{H_2O} H_2O \tag{7}$$

$$-\frac{d[\text{Fuel}]}{dt}, \frac{\text{kg}}{\text{m}^3\ \text{sec}} = \text{A}\ \exp\left(\frac{-E}{RT}\right)[Y_{\text{fuel}}]^a [Y_{O_2}]^b \tag{8}$$

where [ ] represents the concentration of species in kg/m^3, Y the mass fraction, A the pre-exponential factor, E the activation energy in kJ/kmole, and a and b the order of reaction; they are tabulated in Bartok and Sarofim for alkanes, ethanol, methanol, benzene and toluene.[11]

Char Reactions

The skeletal char, essentially FC, undergoes heterogeneous reactions with gaseous species. The heterogeneous combustion of carbon or char occurs primarily via one or more of the following reactions:

Reaction I. $C + \frac{1}{2} O_2 \rightarrow CO$
Reaction II. $C + O_2 \rightarrow CO_2$
Reaction III. $C + CO_2 \rightarrow 2\ CO$
Reaction IV. $C + H_2O \rightarrow CO + H_2$

Assuming a first-order reaction for scheme I, the oxygen consumption rate is given as

$$\dot{m}_{O2} \approx \pi d_P^2 B_I T^n \exp\left(-\frac{E}{R_V T_P}\right) \rho_\infty Y_{O2,w.} \tag{9}$$

The dominant oxygen transfer mechanism at high temperatures is via reaction I with an E/R (a ratio of activation energy to universal gas constant) of about 26,200 K, where $B_I = 2.3 \times 10^7$ m/s and $n = 0.5$ to 1. Reaction II has an E/R of 20,000 K, and $B_{II} = 1.6 \times 10^5$ m/s. Reaction III, the Boudouard reduction reaction, proceeds with an E/R of about 40,000 K. The reduction reactions, III and IV, may become significant, especially at high temperatures for combustion in boiler burners. Reaction with steam is found to be 50 times faster than CO_2 at temperatures up to 1800°C at 1 bar for 75–100 micron-sized Montana Rosebud char.[12] The combustible gases CO and H_2 undergo gas phase oxidation, producing CO_2 and H_2O.

Ignition and Combustion

Recently, Essenhigh et al. have reviewed the ignition of coal.[13] Volatiles from lignite are known to ignite at $T > 950$ K in fluidized beds. Coal may ignite homogeneously or heterogeneously depending upon size and volatile content.[14,15] A correlation for heterogeneous char ignition temperature is presented by Du and Annamalai, 1994.

Once ignited, the combustion of high volatile coal proceeds in two stages: combustion of VM and combustion of FC. Combustion of VM is similar to the combustion of vapors from an oil drop. The typical total combustion time of 100-micron solid coal particle is on the order of 1 s in boilers and is dominated by the time required for heterogeneous combustion of the residual char particle, while the pyrolysis time ($t_{pyr} = 10^6$ (s/m^2) d_p^2) is on the order of 1/10th–1/100th of the total burning time. Since biosolid contains 70%–80% VM (coal contains 10% VM), most of the combustion of volatiles occurs within a short time (about 0.10 s).

For liquid drops and plastics of density ρ_c, simple relations exist for evaluating the combustion rates and times. If the transfer number B is defined as

$$B = \frac{cp\{T_\infty - T_w\}}{L} + \frac{Y_{O_{2,\infty}}}{\nu_{O_2}} \frac{h_c}{L}, \tag{10}$$

where $T_w \approx$ TBP for liquid fuels; $T_w = T_g$, the temperature of gasification for plastics; L is the latent heat for liquid fuel and $L = q_g$, heat of gasification for plastics; $Y_{O_{2\infty}}$ is the stoichiometric oxygen mass per unit mass of fuel (typically 3.5 for liquid fuels); and h_c is the lower heating value of fuel; then the burn rate ($\dot{m}$) and time (t_b) for spherical condensates (liquid drops and spherical particles of diameter d_p and density ρ_c) are given by the following expressions:

$$\dot{m} \approx 2\pi \frac{\lambda}{c_p} d_p \ln(1+B) \tag{11}$$

$$t_b = \frac{d_0^2}{\alpha_c}, \tag{12}$$

Where

$$\alpha_c = 8 \frac{\lambda}{c_P} \frac{\ln(1+B)}{\rho_c} \tag{13}$$

and c_p and λ are the specific heat and thermal conductivity of gas mixture evaluated at a mean temperature (approximately 50% of the adiabatic flame temperature).

The higher the B value, the higher the mass loss rate, and the burn time will be lower. The value of B is about 1–2 for plastics (polymers), 2–3 for alcohols, and 6–8 for pentane to octane. The burn time of plastic waste particles will be about 3–4 times longer than single liquid drops of pentane to octane (≈ gasoline) of similar diameter.

Combustion in Practical Systems

The time scales for combustion are on the order of 1000, 10, and 1 ms for coal burnt in boilers and liquid fuels burnt in gas turbines and diesel engines. Coal is burnt on grates in lumped form (larger-sized particles, 2.5 cm or greater with a volumetric intensity on the order of 500 kW/m^3), medium-sized particles in fluidized beds (1 cm or less, 500 kW/m^3), or as suspensions or pulverized fuel (pf; 75 micron or less, 200 kW/m^3) in boilers.

Apart from pyrolysis, gasification, and combustion, another option for energy conversion (particularly if solid fuel is in slurry firm, such as flushed dairy manure), is the anaerobic digestion (in absence of oxygen) to CH_4 (60%) and CO_2 (40%) using psychrophilic (ambient temperature), mesophilic (95°F) and thermophilic (135°F) bacteria in digesters.[10] Typical options of energy conversion, indicated in Figure 7, include anaerobic digestion (path 1, the biological gasification process), thermal gasification with air to produce CO, HC, CO_2 (path 2A) or with steam to produce CO_2 and H_2 (path 2B), cofiring (path 3), reburn (path 4; see "Reburn with Bio-Solids"), and direct combustion (path 5).

Suspension Firing

In suspension-fired boilers, solid fuel is pulverized into smaller particles (d_p = 75 μm or less) so that more surface area per unit mass is exposed to the oxidant, resulting in improved ignition and combustion characteristics. Typical boiler burners use swirl burners for atomized oil and pulverized coal firing, while a gas turbine uses a swirl atomizer in highly swirling turbulent flow fields. A swirl burner for pf firing is shown in Figure 8. The air is divided into a primary air stream which transports the coal (10%–20% of the total air, heated to 70°C–100°C to prevent condensation of vapors and injected at about 20 m/s to prevent settling of the dust, loading dust and gas at a ratio of 1:2) and a secondary air stream (250°C at 60–80 m/s) which is sent through swirl vanes, supplying the remaining oxygen for combustion and imparting a tangential momentum to the air. In wall-fired boilers, burners are stacked above each other on the wall; while in tangential-fired boilers, the burners are mounted at the corners of rectangular furnaces.

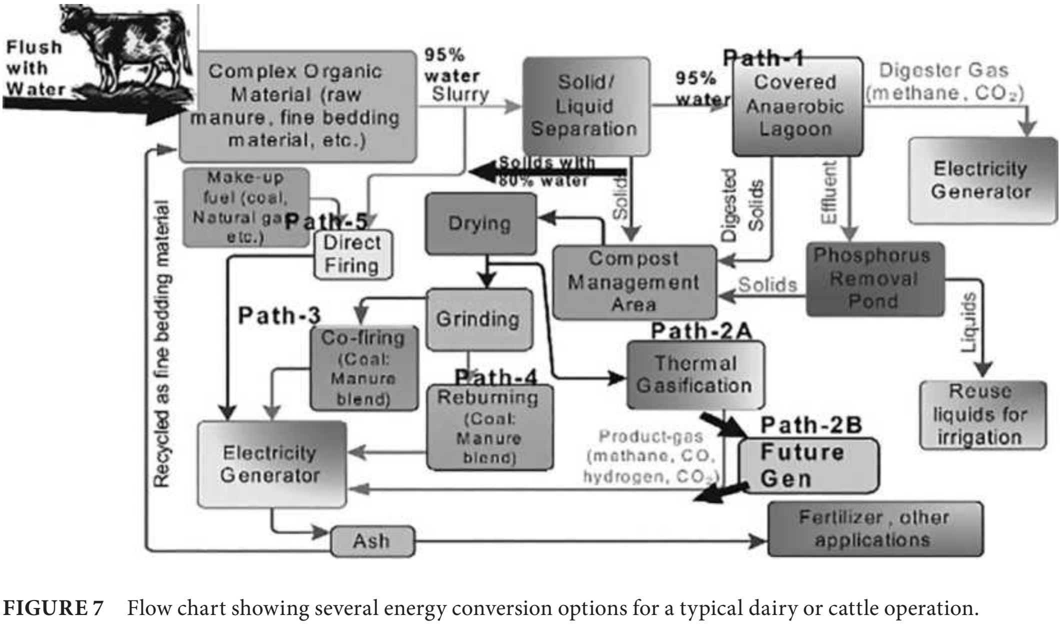


FIGURE 7 Flow chart showing several energy conversion options for a typical dairy or cattle operation.

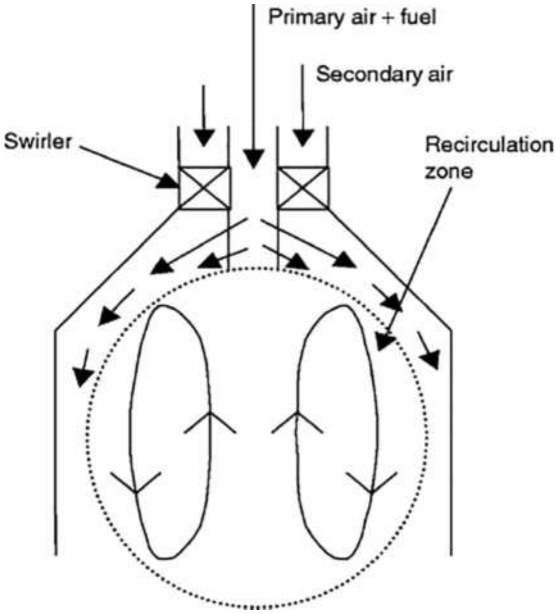


FIGURE 8 Pulverized Fuel (pf) fired swirl burner.

Stoker Firing

The uncrushed fuel [fusion temperature <1093°C (2000°F); volatile content >20%; sizes in equal proportions of 19 mm × 12.5 mm (3/4 in. × ½ in.), 6.3 mm × 3.2 mm (½ in. × ¼ in.), 3.2 mm × 3.2 mm (¼ in × ¼ in.)][16] is fed onto a traveling chain grate below to which primary air is supplied (Figure 9), which may be preheated to 177°C (350°F) if moisture exceeds 25%. The differential pressure is on the order of 5–8 mm (2–3 in.). The combustible gases are carried into an over-fire region into which secondary air (almost 35% of total air at three levels for low emissions) is fired to complete combustion. The over-fire region acts like a perfectly stirred reactor (PSR). It is apparent that solid fuels need not be ground to finer size.

Fixed Bed Combustor

The bed contains uncrushed solid fuels, inert materials (including ash), and processing materials (e.g., limestone to remove SO_2 from gases as sulfates). It is fed with air moving against gravity for complete combustion, but the velocity is low enough that materials are not entrained into the gas streams. Large solid particles can be used.

Fluidized Bed Combustor

When air velocity (V) in fixed bed combustor (FXBC) is increased gradually to a velocity called minimum fluidization velocity V_{mf}, the upward drag force is almost equal to the weight of the particle, so that solids float upward. The bed behaves like a fluid (like liquid water in a tank), i.e., it becomes fluidized. If $V > V_{mf}$, then air escapes as bubbles and is called a bubbling fluidized bed combustor (BFBC). The bed has two phases: the bubble phase, containing gases (mostly oxygen), and the emulsion phase (dense phase, oxygen deficient), containing particles and gas. Many times gas velocity is so high that gaseous combustibles produced within the bed burn above the bed (called free board region), while solids (e.g., char and carbon) burn within the bed. Fluidized Bed Combustor (FBC) is suitable for fuels which are difficult to combust in pf-fired boilers.

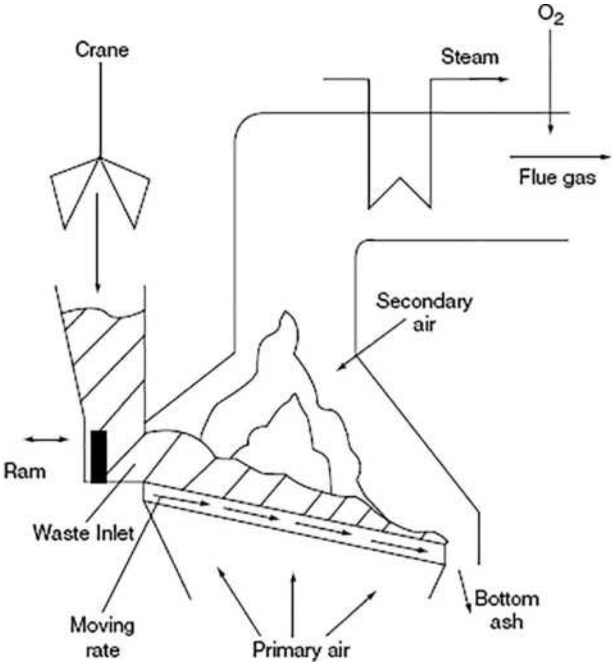


FIGURE 9 Schematic of stoker firing.
Source: Loo and Kessel.[39]

Circulating Fluidized Bed Combustor (CFBC)

When air velocity in FBC is increased at velocity $V >> V_{mf}$, particles are entrained into the gas stream. Since the residence time available to particles for combustion is shorter, unburned particles are captured using cyclones located downstream of the combustor and circulated back to the bed.

The residence time (t_{res}) varies from a low value for pf-fired burners to a long residence time for fixed-bed combustors. The reaction time (t_{reac}) should be shorter than t_{res} so that combustion is complete. The reaction time includes time to heat up to ignition temperature and combustion. The previous section on fuel properties and the homogenous (e.g., CH_4, CO oxidation) and heterogeneous (e.g., carbon oxidation) reaction kinetics can be used to predict t_{reac} or burn time t_b.

Coal and Bio-Solids Cofiring

General Schemes of Conversion

Most of the previously reviewed combustion systems typically use pure coal, oil, or gas. The same systems require redesign for use with pure biomass fuels. A few of the technologies, which utilize bio-solids as an energy source, are summarized in Annamalai et al.[17] These technologies include direct combustion (fluidized beds), circulating fluidized beds, liquefaction (mostly pyrolysis), onsite gasification for producing low to medium Btu gases, anaerobic digestion (bacterial conversion), and hydrolysis for fermentation to liquid fuels like ethanol.[18,19]

Cofiring

Although some bio-solids have been fired directly in industrial burners as sole-source fuels, limitations arose due to variable moisture and ash contents in bio-solid fuels, causing ignition and combustion problems for direct combustion technologies. To circumvent such problems, these fuels have been fired along with the primary fuels (cofiring) either by directly mixing with coal and firing (2%–15% of heat input basis) or by firing them in between coal-fired burners.[20–24]

Cofiring has the following advantages: improvement of flame stability problems, greater potential for commercialization, low capital costs, flexibility of adaptation of biomass fuels and cost effective power generation, mitigated NO_x emissions from coal-fired boilers, and reduced CO_2 emissions. However, a lower melting point of biomass ash could cause fouling and slagging problems.

Some of the bio-solid fuels used in cofiring with coal are cattle manure,[25,26] sawdust and sewage sludge,[21] switch grass,[20] wood chips,[24,27] straw,[22,28] and refuse-derived fuel (RDF).[21] See Sami et al. for a review of literature on cofiring.[7]

Coal and Agricultural Residues

Sampson et al. reported test burns of three different types of wood chips (20%, HHV from 8320 to 8420 Btu/lb) mixed with coal (10,600 Btu/lb) at a stoker (traveling grate) fired steam plant.[24] The particulate emission in grams per SCF ranged from 0.05 to 0.09. An economic study, conducted for the 125,000 lb/h steam power plant, concluded that energy derived from wood would be competitive with that from coal if more than 30,000 tons of wood chips were produced per year with hauling distances less than 60 mi. Aerts et al. carried out their experiments on cofiring switch grass with coal in a 50-MW, radiant, wall-fired, pulverized coal boiler with a capacity of 180 tons of steam at 85 bar and 510°C (Figure 10). The NO_x emissions decreased by 20%, since switchgrass contains lesser nitrogen (Table 4).[20] It is the author's hypothesis that a higher VM content of bio-solids results in a porous char, thus accelerating the char combustion process. This is validated by the data from Fahlstedt et al. on the cofiring of wood chips, olive pips and palm nut shells with coal at the ABB Carbon 1 MW Process Test Facility; they found that blend combustion has a slightly higher efficiency than coal-only combustion.[27]

Coal and RDF

Municipal solid waste includes residential, commercial, and industrial wastes which could be used as fuel for production of steam and electric power. MSW is inherently a blended fuel, and its major components are paper (43%); yard waste, including grass clippings (10%); food (10%); glass and ceramics (9%); ferrous materials (6%); and plastics and rubber (5%). Refer to Tables 5 and 6 for analyses. When

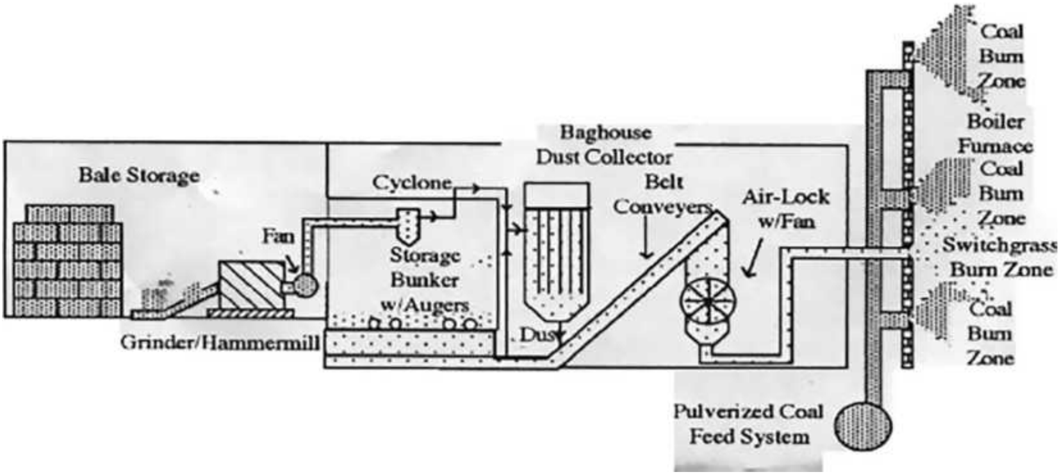


FIGURE 10 A cofiring scheme for coal and biomass (Alternate Fuel Handling Facility at Blount St. Generating Station).
Source: Aerts et al.[20]

TABLE 5 Chemical Composition of Solid Waste

	Percent	
Proximate Analysis	Range	Typical
Volatile matter (VM)	30–60	50
Fixed carbon (FC)	5–10	8
Moisture	10–45	25
Ash	10–30	25

	Percent by Mass (dry basis)					
Ultimate Analysis	C	H	O	N	S	Ash
Yard wastes	48	6	38	3	0.3	4.7
Wood	50	6	43	0.2	0.1	0.7
Food wastes	50	6	38	3	0.4	2.6
Paper	44	6	44	0.3	0.2	5.5
Cardboard	44	6	44	0.3	0.2	5.5
Plastics	60	7	23			10
Textiles	56	7	30	5	0.2	1.8
Rubber	76	10		2		12
Leather	60	9	12	10	0.4	8.6
Misc. organics	49	6	38	2	0.3	4.7
Dirt, ashes, etc.	25	3	1	0.5	0.2	70.3

TABLE 6 Heat of Combustions of Municipal Solid Waste Components

	Inerts (%)		Heating Values (kJ/kg)	
Component	Range	Typical	Range	Typical
Yard wastes	2–5	4	2,000–19,000	7,000
Wood	0.5–2	2	17,000–20,000	19,000
Food wastes	1–7	6	3,000–6,000	5,000
Paper	3–8	6	12,000–19,000	17,000
Cardboard	3–8	6	12,000–19,000	17,000
Plastics	5–20	10	30,000–37,000	33,000
Textiles	2–4	3	15,000–19,000	17,000
Rubber	5–20	10	20,000–28,000	23,000
Leather	8–20	10	15,000–20,000	17,000
Misc. organics	2–8	6	11,000–26,000	18,000
Glass	96–99	98	100–250	150
Tin cans	96–99	98	250–1,200	700
Nonferrous	90–99	96		
Ferrous metals	94–99	98	250–1,200	700
Dirt, ashes, etc.	60–80	70	2,000–11,600	7,000

raw waste is processed to remove non-combustibles like glass and metals, it is called RDF. MSW can decompose in two ways, aerobic and anaerobic. Aerobic decomposition (or composting) occurs when O_2 is present. The composting produces CO_2 and water, but no usable energy products. The anaerobic decomposition occurs in the absence of O_2. It produces landfill gas of 55% CH_4 and 45% CO_2.

Coal and Manure

Frazzitta et al. and Annamalai et al. evaluated the performance of a small-scale pf-fired boiler burner facility (100,000 Btu/h) while using coal and premixed coal manure blends with 20% manure. Three types of feedlot manure were used: raw, partially composted, and fully composted. The burnt fraction was recorded to be 97% for both coal and coal-manure blends.[25,26]

NO_x Emissions

During combustion, the nitrogen evolved from fuel undergoes oxidation to NO_x; and this is called fuel NO_x to distinguish it from thermal NO_x, which is produced by oxidation of atmospheric nitrogen. Unlike coal, most of the agricultural biomass being burned is very low in nitrogen content (i.e., wood or crops), but manure has a higher N content than fossil fuels. A less precise correlation exists between cofiring levels on a Btu basis and percent NO reduction under cofiring. The following Eq. (valid between 3 and 22% mass basis cofiring) describes NO_x reduction as a function of cofiring level on a heat input basis:

$$NO_x \text{ Reduction}(\%) = 0.0008(\text{COF}\%)^2 + 0.0006 \text{ COF}\% + 0.0752, \tag{14}$$

where COF% is the percentage of co-firing on a heat input basis. The mechanisms used to reduce NO_x emissions by cofiring vary between cyclone firing and PC firing.

Figure 11 shows the percentage reduction in NO with percentage cofiring of low-N agricultural biomass fuels. This relationship does not apply to high-N biofuels such as animal manure.

Fouling in Cofiring

Hansen et al. investigated the ash deposition problem in a multi-circulating fluidized bed combustor (MCFBC) fired with fuel blends of coal and wood straw.[25] The Na and K lower the melting point of ash. For ash fusion characteristics see Table 7. Rasmussen and Clausen evaluated the performance of an 80-MW co-generation power plant at Grenaa, Denmark, fired with hard coal and bio-solids (surplus straw from farming). Large amounts of Na and K in straw caused superheater corrosion and

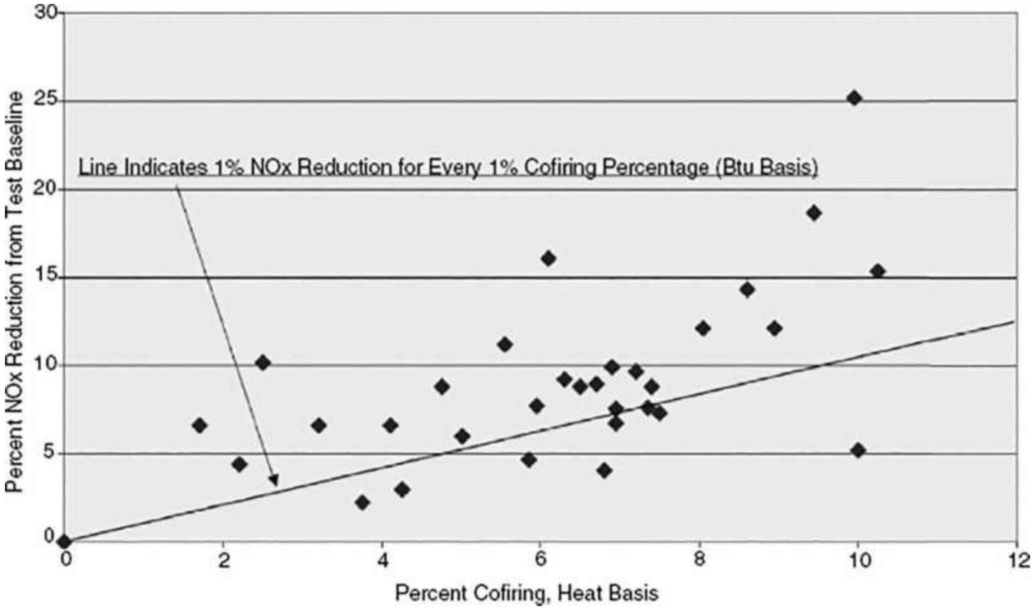


FIGURE 11 NO_x due to cofiring with low-N agricultural residues.
Source: Grabowski.[40]

TABLE 7 Ash Fusion Behavior and Ash Composition, Fusion Data: ASTM D-188

	FB	PRB Coal	Blend
Ash Fusion, (reducing) Initial deformation, IT, °C (°F)	1140 (2090)	1130 (2060)	NA
Softening, °C (°F)	1190 (2170)	1150(2110)	NA
Hemispherical, HT, °C (°F)	1210(2210)	1170(2130)	NA
Fluid, °C (°F)	1230 (2240)	1200 (2190)	NA
Ash fusion, (oxidizing) Initial deformation, IT, °C (°F)	1170(2130)	1190(2180)	NA
Softening, °C (°F)	1190 (2180)	1200 (2190)	NA
Hemispherical, HT, °C (°F)	1220(2230)	1210(2210)	NA
Fluid, °C (°F)	1240(2270)	1280 (2330)	NA
Slagging Index, Rs, °C (°F)	1160(2120)	1140 (2090)	
Slagging classification	High	Severe	
Ash composition (wt%) SiO_2	53.63	36.45	43.56
Al_2O_3	5.08	18.36	12.87
Fe_2O_3	1.86	6.43	4.54
TiO_2	0.29	1.29	0.88
CaO^+	14.60	19.37	17.40
MgO^+	3.05	3.63	3.39
Na_2O^+	3.84	1.37	2.39
K_2O^+	7.76	0.63	3.58
P_2O_5	4.94	0.98	2.62
SO_3	3.71	10.50	7.69
MnO_2	0.09	0.09	0.09
Sum	98.84	99.11	99.00
Volatile Oxides	30.77	28.25	
Basic oxides	32.73	35.51	
Silica ratio	0.73	0.53	
Na_2O+K_2O	11.60	2.00	5.97
Inherent Ca/S Ratio	6.71	1.86	2.48
kg alkali (Na_2O+K_2O)/GJ	5.37	0.06	0.29

combustor fouling.[29] Annamalai et al. evaluated fouling potential when feedlot manure biomass (FB) was cofired with coal under suspension firing.[30] The 90:10 Coal:FB blend resulted in almost twice the ash output compared to coal and ash deposits on heat exchanger tubes that were more difficult to remove than baseline coal ash deposits. The increased fouling behavior with blend is probably due to the higher ash loading and ash composition of FB.

Gasification of Coal and Bio-Solids

Gasification is a thermo-chemical process in which a solid fuel is converted into a gaseous fuel (primarily consisting of HC, H_2 and CO_2) with air or pure oxygen used for partial oxidation of FCs. The main products during gasification are CO and H_2, with some CO_2, N_2, CH_4, H_2O, char particles, and tar (heavy hydrocarbons). The oxidizers used for the gasification processes are oxygen, steam, or air. However, for air, the gasification yields a low-Btu gas, primarily caused by nitrogen dilution present in the supply air. Syngas ($CO + H_2$) is produced by reaction of biomass with steam. The combustible product, gas, can be used as fuel burned directly or with a gas turbine to produce electricity; or used to make chemical feedstock (petroleum refineries). However, gas needs to be cleaned to remove tar, NH_3,

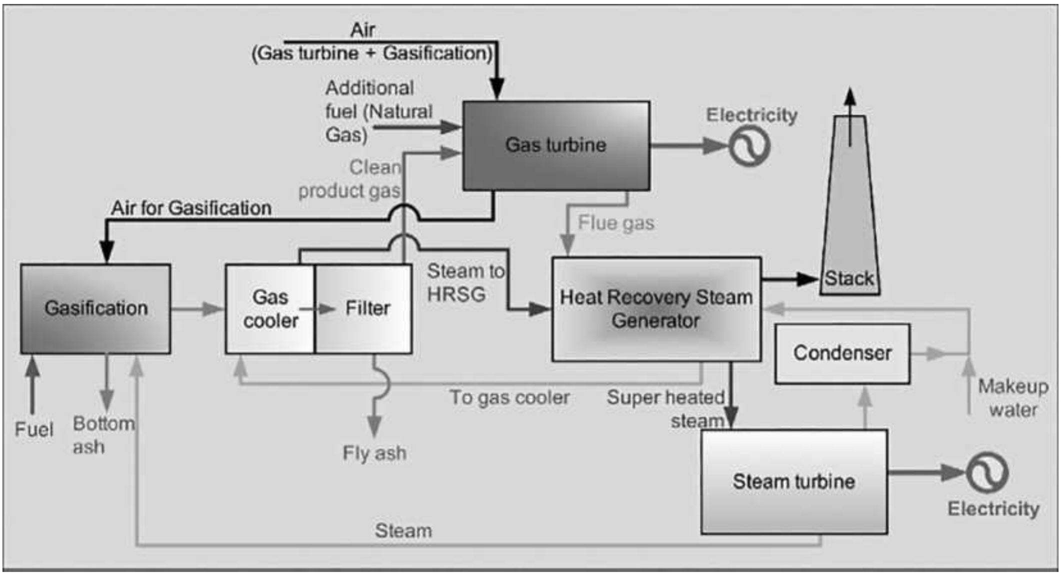


FIGURE 12 Fluidized-bed gasification for Integrated Gasification Combined Cycle (IGCC) Process.

and sulfur compounds. The integrated gasification combined cycle (IGCC) (Figure 12), for combined heat and power (CHP), and traditional boilers use combustible gases from gasifiers for generation of electric power.

Typically in combined cycles, gaseous or liquid fuel is burnt in gas turbine combustors. High-temperature products are expanded in a gas turbine for producing electrical power; a low-temperature (but still hot) exhaust is then used as heat input in a boiler to produce low-temperature steam, which then drives a steam turbine for electrical power. Therefore, one may use gas as a topping cycle medium, while steam is used as fluid for the bottoming cycle. The efficiency of a combined cycle is on the order of 60%, while a conventional gas turbine cycle has an efficiency of 42%.[31] Commercial operations include a 250-MW IGCC plant at Tampa, Florida, operating since 1996; a 340-MW plant at Negishi, Japan, since 2003; and a 1200-MW GE-Bechtel plant under construction in Ohio for American Electric Power, to start in 2010.[31]

There are three basic gasification reactor types: (i) fixed-bed gasifiers (Figure 13); (ii) fluidized-bed gasifiers, including circulating-bed (CFB) or bubbling-bed; and (iii) entrained-flow gasifiers. The principles of operation are similar to those of combustors except that the air supplied is much below stoichiometric amounts, and instead of a combination of steam, air and CO_2, air can also be used. The oxidant source could also include gases other than air, such as air combined with steam in Blasiak et al.[32]

Futuregen

FutureGen is a new U.S. initiative to build the world's first integrated CO_2 sequestration and H_2 production research power plant using coal as fuel. The technology shown in Figure 14 employs modern coal gasification technology using pure oxygen, resulting in CO, C_nH_m (a hydrocarbon), H_2, HCN, NH_3, N_2, H_2S, SO_2, and other combinations which are further reacted with steam (reforming reactions) to produce CO_2 and H_2. The bed materials capture most of the harmful N and S compounds followed by gas-cleaning systems; the CO_2 is then sequestered and H_2 is used as fuel, using either combined cycle or fuel cells for electricity generation or sold as clean transportation fuel. With partial oxidation of gasification products and char supplying heat for pyrolysis and other endothermic reactions (i.e., net zero external

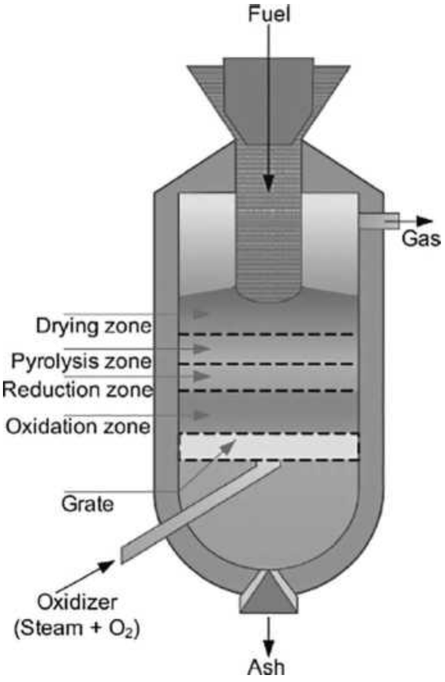


FIGURE 13 Updraft fixed-bed gasifier.

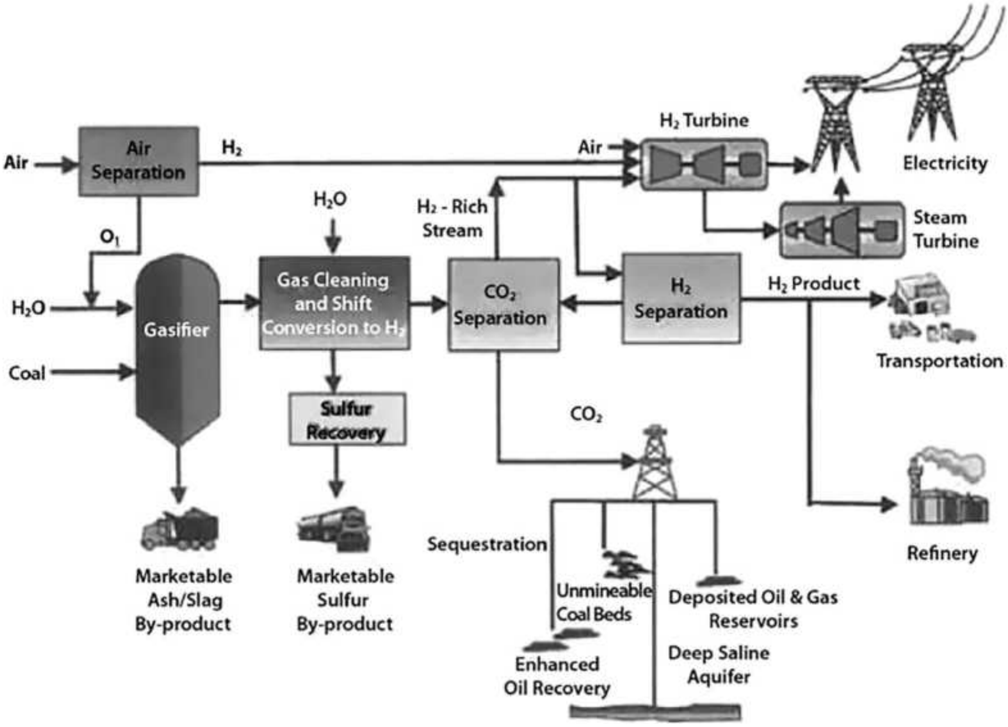


FIGURE 14 FutureGen layout. Source: http://www.fe.doe.gov.

heat supply in gasifier), the overall gasification reaction can be represented as follows for 100 kg of DAF Wyoming coal:

$$\text{Reaction V} \quad C_{6.3}H_{4.4}N_{0.0076}O_{1.15}S_{0.014} + 8.14\ H_2O(l) + 1.67\ O_2 \rightarrow 6.3\ CO_2 + 10.34\ H_2 + 0.038\ N_2 + 0.014\ SO_2$$

It is apparent that the FutureGen process results in enhanced production of H_2, using coal as an energy source to strip H_2 from water. For C–H–O fuels, it can be shown that theoretical H_2 production (N_{H2}) in moles for an empirical fuel CH_hO_o is given as $\{0.4115\,h - 0.6204\,o + 1.4776\}$ under the above conditions. For example, if glucose $C_6H_{12}O_6$ is the fuel, then empirical formulae is CH_2O; thus, with h = 2, o = 1, $N_{H_2} = 1.68$ kmol, using atom balance, $CH_2O + 0.68\ H_2O(\ell) + 0.16\ O_2\ CO_2 + 1.68\ H_2$.

Reburn with Bio-Solids

NO_x is produced when fuel is burned with air. The N in NO_x can come both from the nitrogen-containing fuel compounds (e.g., coal, biomass, plant residue, animal waste) and from the N in the air. The NO_x generated from fuel N is called fuel NO_x, and NO_x formed from the air is called thermal NO_x. Typically, 75% of NO_x in boiler burners is from fuel N. It is mandated that NO_x, a precursor of smog, be reduced to 0.40–0.46 lb/mmBtu for wall and tangentially fired units under the Clean Air Act Amendments (CAAA). The current technologies developed for reducing NO_x include combustion controls (e.g., staged combustion or low NO_x burners (LNB), reburn) and post-combustion controls (e.g., Selective Non-Catalytic Reduction, SNCR using urea).

In reburning, additional fuel (typically natural gas) is injected downstream from the primary combustion zone to create a fuel rich zone (optimum reburn stoichiometric ratio (SR), usually between SR 0.7 and 0.9), where NO_x is reduced up to 60% through reactions with hydrocarbons when reburn heat input with CH_4 is about 10%–20%. Downstream of the reburn zone, additional air is injected in the burnout zone to complete the combustion process. A diagram of the entire process with the different combustion zones is shown in Figure 15. There have been numerous studies on reburn technology found in literature, with experiments conducted, and the important results summarized elsewhere.[33] Table 8 shows the percentages of reduction and emission obtained with coal or gas reburn in coal-fired installations and demonstration units.

The low cost of biomass and its availability make it an ideal source of pyrolysis gas, which is a more effective reburn fuel than the main source fuel, which is typically coal. Recently, animal manure has been tested as a reburn fuel in laboratory scale experiments. A reduction of a maximum of 80% was achieved for pure biomass, while the coal experienced a reduction of between 10 and 40%, depending on the equivalence ratio.[34] It is believed that the greater effectiveness of the feedlot biomass is due to its greater volatile content on a DAF basis and its release of fuel nitrogen in the form of NH_3, instead of HCN.[35]

Acknowledgments

Most of this work was supported in part by the U.S. Department of Energy of Pittsburgh, PA, the DOE of Golden, CO, the USDA, the Texas Advanced Technology Program, and the Texas Commission on Environmental Quality (TCEQ) of Austin, TX, through the Texas Engineering Experiment Station (TEES), Texas A&M University, and College Station, TX.

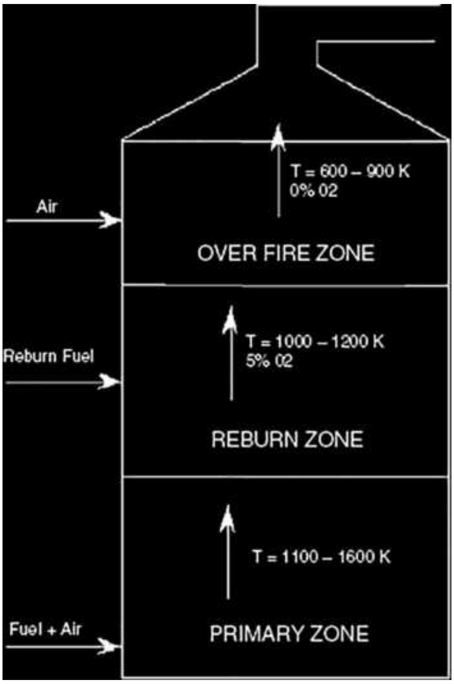


FIGURE 15 Schematic of reburn process.

TABLE 8 Percentage Reduction in NO_x: Demonstration and/or Operating Reburn Installations on Coal-Fired Boilers in the United States.

Type of Burner	% Reburn Heat in	% Reduction	NOx with Reburn lb/mmBtu[a]
Gas reburning			
Tangential	18	50–67	0.25
Cyclone	20–23	58–60	0.39–0.56
Wall without LNB	18	63	0.27
Coal reburn			
Cyclone (micronized)	30 (17)	52 (57)	0.39 (0.59)
Tangential (micron) with LNB	14	28	0.25

Note: LNB: Low NO_x Burners.

[a] *1 lb per mmBtu = 0.430 kg/GJ.*

Source: DOE.[38]

References

1. U.S. DOE, Energy Information Administration, available at http://www.eia.doe.gov/ (accessed April 12, 2005).
2. Priyadarsan, S.; Annamalai, K.; Sweeten, J.M.; Holtzapple, M.T.; Mukhtar, S. *Co-gasification of blended coal with feedlot and chicken litter biomass,* Proceedings of the 30th Symposium (International) on Combustion; The Combustion Institute: Pittsburgh, PA, 2005; Vol. 30, 2973–2980.
3. Ebeling, J.M.; Jenkins, B.M. Physical and chemical properties of biomass fuels. Trans. ASAE **1985**, *28* (3), 898–902.

4. Sweeten, J.M.; Annamalai, K.; Thien, B.; McDonald, L. Co-firing of coal and cattle feedlot biomass (FB) fuels, part I: feedlot biomass (cattle manure) fuel quality and characteristics. Fuel **2003**, *82* (10), 1167–1182.
5. Boie, W. Wiss z Tech. Hochsch. Dresden 1952/1953, 2, 687.
6. Annamalai, K.; Sweeten, J.M.; Ramalingam, S.C. Estimation of the gross HVs of biomass fuels. Trans. Soc. Agric. Eng. **1987**, *30,* 1205–1208.
7. Sami, M.; Annamalai, K.; Wooldridge, M. Co-firing of coal and biomass fuel blends. Prog. Energy Combust. Sci. **2001**, *27,* 171–214.
8. Annamalai, K.; Puri, I.K. 2006, *Combustion Science and Engineering,* Taylor and Francis: Orlando, FL, 2006.
9. Tillman, D.A.; Rossi, A.J.; Kitto, W.D. *Wood Combustion: Principles: Processes and Economics*; Academic Press: New York, 1981.
10. Sweeten, J.M.; Heflin, K.; Annamalai, K.; Auvermann, B.; Collum, Mc; Parker, D.B. *Combustion-Fuel Properties of Manure Compost from Paved vs Un-paved Cattle feedlots*; ASABE 06–4143: Portland, OR, July 9–12, 2006.
11. Bartok, W., Sarofim, A.F. 1991, *Fossil Fuel Combustion*, chapter 3: FL Dryer, John Wiley, and Hobokan, NJ. pp. 121–214.
12. Howard, J.B.; Sarofim, A.F. Gasification of coal char with CO_2 and steam at 1200–1800°C. *Energy Lab Report;* Chemical Engineering; MIT: MA, 1978.
13. Essenhigh, R.H.; Misra, M.K.; Shaw, D.W. Ignition of coal particles: a review. Combust. Flame **1989**, *77,* 3–30.
14. Annamalai, K.; Durbetaki, P. A theory on transition of ignition phase of coal particles. Combust. Flame **1977**, *29,* 193–208.
15. Du, X.; Annamalai, K. The transient ignition of isolated coal particles. Combust. Flame **1994**, *97,* 339–354.
16. Johnson, N. *Fundamentals of Stoker Fired Boiler Design and Operation*, CIBO Emission Controls Technology Conference, July 15–17, 2002.
17. Annamalai, K.; Ibrahim, Y.M.; Sweeten, J.M. Experimental studies on combustion of cattle manure in a fluidized bed combustor. Trans. ASME, J. Energy Res. Technol. **1987**, *109,* 49–57.
18. Walawender, W.P.; Fan, L.T.; Engler, C.R.; Erickson, L.E. Feedlot manure and other agricultural wastes as future material and energy resources: II. Process descriptions. *Contrib.30,* Deptartment of Chemical Engineering, Kansas Agricultural Experiment Station: Manhattan, KS, 30, 1973.
19. Raman, K.P.; Walawander, W.P.; Fan, L.T. Gasification of feedlot manure in a fluidized bed: effect of temperature. Ind. Eng. Chem. Proc. Des. Dev. **1980**, *10,* 623–629.
20. Aerts, D.J.; Bryden, K.M.; Hoerning, J.M.; Ragland, K.W. *Co-firing Switchgrass in a 50 MW Pulverized Coal Boiler,* Proceedings of the 59th Annual American Power Conference, Chicago, IL, **1997**, *50* (2), 1180–1185.
21. Abbas, T.; Costen, P.; Kandamby, N.H.; Lockwood, F.C.; Ou, J.J. The influence of burner injection mode on pulverized coal and biosolid co-fired flames. Combust. Flame **1994**, *99*, 617–625.
22. Siegel, V.; Schweitzer, B.; Spliethoff, H.; Hein, K.R.G. Preparation and co-combustion of cereals with hard coal in a 500 kW pulverized-fuel test unit. Biomass for energy and the environment, *Proceedings of the 9th European Bioenergy Conference,* Copenhagen, DK, June 24–27, 1996; 2, 1027–1032.
23. Hansen, L.A.; Michelsen, H.P.; Dam-Johansen, K. Alkali metals in a coal and biosolid fired CFBC—measurements and thermodynamic modeling, *Proceedings of the 13th International Conference on Fluidized Bed Combustion*, Orlando, FL, May 7–10, 1995; 1, 39–48.
24. Sampson, G.R.; Richmond, A.P.; Brewster, G.A.; Gasbarro, A.F. Co-firing of wood chips with coal in interior Alaska. Forest Prod. J. **1991**, *41* (5), 53–56.
25. Frazzitta, S.; Annamalai, K.; Sweeten, J. Performance of a burner with coal and coal: feedlot manure blends. J. Propulsion Power *1999*, *15* (2), 181–186.

26. Annamalai, K.; Thien, B.; Sweeten, J.M. Co-firing of coal and cattle feedlot biomass (FB) fuels part II: performance results from 100,000 Btu/h laboratory scale boiler burner. Fuel **2003**, *82* (10), 1183–1193.
27. Fahlstdedt, I.; Lindman, E.; Lindberg, T.; Anderson, J. Cofiring of biomass and coal in a pressurized fluidized bed combined cycle. Results of pilot plant studies, *Proceedings of the 14th International conference on Fluidized Bed Combustion,* Vancouver, Canada, May 11–14, 1997; 1, 295–299.
28. Van Doom, J., Bruyn, P., Vermeij, P. Combined combustion of biomass, fluidized sewage sludge and coal in an atmospheric Fluidized bed installation, Biomass for energy and the environment, *Proceedings of the 9th European Bioenergy Conference*, Copenhagen, DK, June 24–27, 1996; 1007–1012.
29. Rasmussen, I.; Clausen, J.C. ELSAM strategy of firing biosolid in CFB power plants, *Proceedings of the 13th International Conference on Fluidized Bed Combustion*, Orlando, FL, May 7–10, 1, 1995; 557–563.
30. Annamalai, K.; Sweeten, J.; Freeman, M.; Mathur, M.; O'Dowd, W.; Walbert, G.; Jones, S. Co-firing of coal and cattle feedlot biomass (FB) fuels, part III: fouling results from a 500,000 Btu/hr pilot plant scale boiler burner. Fuel **2003**, *82* (10), 1195–1200.
31. Langston, L.S. 2005 New Horizons. *Power and Energy* vol 2, June 2, 2005.
32. Blasiak, W.; Szewczyk, D.; Lucas, C.; Mochida, S. *Gasification of Biomass Wastes with High Temperature Air and Steam*, Twenty-First International Conference on Incineration and Thermal Treatment Technologies, New Orleans, LA, May 13–17, 2002.
33. Thien, B.; Annamalai, K. *National Combustion Conference,* Oakland, CA, March 25–27, 2001.
34. Arumugam, S.; Annamalai, K.; Thien, B.; Sweeten, J. Feedlot biomass co-firing: a renewable energy alternative for coal-fired utilities, Int. Natl J. Green Energy **2005**, *2* (4), 409–419.
35. Zhou, J.; Masutani, S.; Ishimura, D.; Turn, S.; Kinoshita, C. Release of fuel-bound nitrogen during biomass gasification. Ind. Eng. Chem. Res. **2000**, *39*, 626–634.
36. Annamalai, K.; Puri, I.K. *Advanced Thermodynamics;* CRC Press: Boca Raton, FL, 2001.
37. Priyadarsan, S.; Annamalai, K.; Sweeten, J.M.; Holtzapple, M.T.; Mukhtar, S. Waste to energy: fixed bed gasification of feedlot and chicken litter biomass. Trans. ASAE **2004**, *47* (5), 1689–1696.
38. DOE, Reburning technologies for the control of nitrogen oxides emissions from coal-fired boilers Topical Report, No 14, U.S. Department of Energy, May, 1999.
39. Loo, S.V.; Kessel, R., 2003, Optimization of biomass fired grate stroker systems, EPRI/IEA Bioenergy Workshop, Salt Lake City, Utah, February 19, 2003.
40. Grabowski, P. *Biomass Cofiring, Office of the Biomass Program*, Technical Advisory Committee, Washington, DC, March 11, 2004.

18

Energy Demand: From Individual Behavioral Changes to Climate Change Mitigation

Leila Niamir and
Felix Creutzig

Introduction

Decarbonization of the economy requires massive worldwide efforts and a strong involvement of regions, cities, businesses, and individuals in addition to commitments at national levels (1). Human consumption, in combination with a growing population, contributes to climate change by increasing the rate of green house gas (GHG) emissions (2,3). Over the last decade, instigated by the Paris Agreement, the efforts to limit global warming have been expanding. However, significant attention is being devoted to new energy technologies on both the production and consumption sides, while changes in individual behavior and management practices as part of the mitigation strategy are often neglected (4–8).

Although demand-side solutions are promising, they are not given the same level of attention as technological supply-side solutions in assessments, like the IPCC's AR5, and in integrated assessment models (IAMs) in general, or in the popular media, as mirrored in the observation that directed innovation perversely privileges energy-supply technologies over efficient end-use technologies (10). This potentially has two main reasons: (a) demand-side solutions are often embedded in a complex network of social institutions and practices (11), and thus less prone to quantitative analysis and clear-cut implementation; (b) demand-side solutions also often involve explicit normative positions, or values, making those solutions subject to value-laden discourses (12). Specifically, as demand-side solutions presuppose modified behavior, they are less compatible with the revealed-preference framework prevalent in economics. Both reasons apply less to supply-side solutions: people consume electricity via plug-in, and the original supply of energy for electricity is irrelevant for actual consumption.

Table 1 shows the potential contribution to climate change mitigation from three sectors: buildings, transport, and food under two main perspectives: (a) hard infrastructures, such as the built environment, and (b) soft infrastructures, such as habits and norms. On the one hand, hard infrastructures, epitomized by the urban built infrastructure, provide a physical setting for shaping preferences, practices, and opportunity spaces. On the other hand, changing norms, practices, and nudges modify the opportunity space and can induce direct behavioral change (13).

In building sector, the behavioral change potential can be as high as 50% over long periods of time (Table 1). There is a range in the energy savings achievable in buildings due to behavioral changes, depending on the type of end use. For example, savings from heating loads of 10–30% are possible for changes in the thermostat setting. Similarly, cooling savings of 50–67% are recorded with measures such as substituting air-conditioning with fans in moderately hot climates with tolerable, brief heat exposures. Meanwhile, subsequent literature has provided quantitative estimates of the importance of urban transport and behavioral options (14). Overall, behavioral and infrastructural measures in cities can potentially reduce GHG emissions from urban passenger transport by 20–50% until 2050 (15). This may be achieved via three routes, technological change, modal shift, and reduced travel demand, and

TABLE 1 Climate Mitigation Options and Their Quantitative Reduction Potential

End-User Sector	Mitigation Option	Reduction Potential	Reference(s)
Urban areas and spatial planning	Modifying the emerging urbanization	20–25% reduction of future urban energy use until 2050	(17)
Transportation	Optimal pricing	15–20% reduction in automobile transport	(17)
	Compactness	United States: 10% reduction in distance traveled	(18)
		Europe: 5% reduction in distance traveled	(19)
	Behavioral measures (marketing, information provision, and tailored services)	10% reduction in transport demand	(20,21)
Tourism	Transport modal shifts and increases in average length of stay	44% emission reductions worldwide until 2035	(22)
Buildings	Developed context: short-term behavioral change potential	>20%	(2)
	Developed context: long-term behavioral change potential	50%	(23,24)
	Developing context: behavioral change potential	Relevant but lower potential	(25,26)
	Heating: adjusting thermostat setting	10–30%	(27)
	Cooling: using substitutes for air-conditioning, increasing thermostat setting, changing dress codes	50–67%	(27,28)
	Clothes washing and drying behavior (e.g., by operation at full load versus one-third to one-half load)	10–100%	(29)
	Water-heating and cooking energy savings (e.g., by shorter showers, using different cooking practices)	50%	(29)
	Lighting energy savings (e.g., by turning off unneeded lights)	70%	(29)
	Refrigerator energy savings (e.g., by smaller refrigerators)	30–50%	(29)
	Dishwasher energy savings (e.g., by full-load operation)	75%	(29)
Agriculture and other land use	Technical potential of demand-side options for 2050	70%	(30,31)
	Healthy diet: meat, fish, and egg consumption of 90 g/(person/day)	36% greenhouse gas savings	(32)

Source: Ref. (13).

information technology plays a key role in all three areas of innovation (16). Behavioral options could be fostered by information campaigns that facilitate social learning, active choice, and frame alternative mobility choices.

Reshaping urban forms and the urban environment provides ample opportunity to make private car travel obsolete and save high amounts of energy in buildings. Both transport and buildings also offer significant opportunities for reduced energy demand by soft measures, such as providing targeted information on sustainable mobility for specific groups of people. In agriculture, demand-side action, in particular, dietary shift, could reduce emissions by more than 70% compared to the trend in 2055, thereby surpassing the potential of technological options.

Demand-Side Solutions

Demand-side solutions for mitigating climate change include strategies targeting technology choices, consumption, behavior, lifestyles, coupled production–consumption infrastructures and systems, service provision, and associated sociotechnical transitions. Disciplines vary in their approaches and in the research questions that they ask on demand-side issues (4). For example, psychologists and behavioral economists focus on emotional factors and cognitive biases in the decision-making process (33); economists elaborate on how, under rational decision-making, carbon pricing and other fiscal instruments can trigger change in demand (34); sociologists emphasize everyday practices, structural issues, and socioeconomic inequality (35); anthropologists address the role of culture in energy consumption (36); and studies in technological innovation consider sociotechnical transitions and the norms, rules, and pace of adoption that support dominant technologies (37).

Figure 1 proposed a demand-side assessment framework and discuss key topics that need to be addressed: the characterization of demand; policy instruments and how they would affect demand; techno-economic evaluation; implications for well-being; mitigation pathways; and the sustainable development context.

The starting point for a demand-side assessment seeks to characterize patterns of demand for energy, mobility, food and shelter, and the associated GHG emissions. Hence, the first question to ask is: what norms, values, preferences, and structural factors shape energy demand and GHG emissions (Figure 1a)? Policy instruments can spur demand-side solutions, in ways that depend on the specific energy service and socioeconomic context. Different disciplines have provided important pieces of this big jigsaw, but much remain to be done to put the assessment of policy instruments together in a truly interdisciplinary effort and address the questions posed (Figure 1b).

Beyond specific technologies, research should take a wider scope and ask for the efficient and reliable provision of end-use services, rather than efficient technology design alone (Figure 1c). Technological studies contribute to an understanding of dynamic systems, describing cost reductions and strategies to overcome barriers on the path from research and development of a technology to market-scale deployment and uptake. Another question is: how do demand-side mitigation measures affect well-being (Figure 1d)? Reducing energy use or GHG emissions need to be balanced with the goal of enhancing human well-being (39).

Sketched approaches such as transition theory, study of behavioral tipping points and social norms, and political economy insights into policy sequencing have the potential for laying out short-term and action-oriented mitigation pathways. Such approaches, together with bottom-up assessments from technological studies, can be soft-coupled and combined with IAMs and similar economic models that assess system-wide potentials, reflecting the interaction between sectors and mitigation options (Figure 1e). The linkage between sustainable development and climate change is also articulated in the 'nationally determined' language of the Paris Agreement, which promotes climate mitigation that coincides with nationally determined development outcomes. A demand-side assessment should also be able to inform sustainable development pathways (Figure 1f).

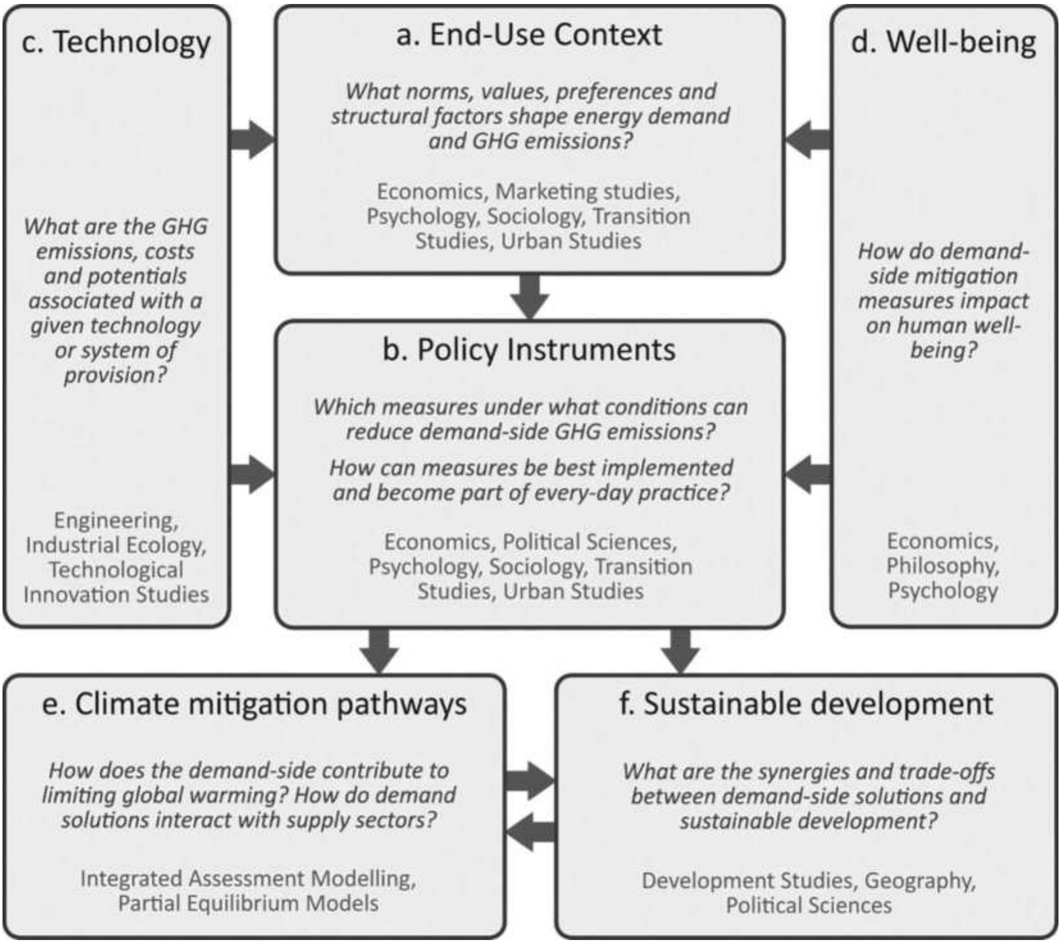


FIGURE 1 Key research questions and contributing disciplines for assessing demand-side solutions. (From Ref. 38.)

Modeling Individuals Energy Behavioral Changes

The understanding of how bottom-up processes can impact climate mitigation guides us to effective development and implementation of policies (9,40). In the last decade, a variety of macroeconomic models and assessment tools emerged (41–43) and were predominately used to support climate-energy policy decisions (44) range from the assessment of macroeconomic and cross-sectoral impacts (45–47), to detailed micro-simulation models for a specific technology (5,47,48). Much can be done to make the assumptions in macroeconomic and integrated assessment models more realistic concerning the scale and nature of damage (5). These models usually assume that economic agents form a representative group(s), have perfect access to information and adapt instantly and rationally to new situations, maximizing their long-run personal advantage (49–52).

Unlike other approaches, agent-based model (ABM) is not limited to perfectly rational agents or to abstract microdetails in aggregate system-level equations. Instead, ABM can represent the behavior of energy consumers – such as individual households – using a range of behavioral theories (9,50,53). In addition, ABM has the ability to examine how interactions of heterogeneous agents at micro-level give rise to the emergence of macro-outcomes, including those relevant for climate mitigation such as an adoption of low-carbon behavioral strategies and technologies over space and time (54). The ABM approach simulates complex and nonlinear behavior that is intractable in equilibrium models (52).

Individual energy behavior, especially when amplified through social context, shapes energy demand and, consequently, carbon emissions. By changing their behaviors, individuals can play an essential role in the transformation process towards a low-carbon society and global emissions reduction (9). However, explaining and affecting human behavior is a difficult task since human nature is complex and heterogeneous. As a result, quantitative tools to assess cumulative household emissions, given the diversity of behavior and a variety of psychological and social factors influencing it beyond purely economic considerations, are scarce (6,40,55).

BENCH Agent-Based Model

To assess the impact of individual behavior on carbon emissions, we went beyond classical economic models and the stylized representation of a perfectly informed optimizer. An agent-based simulation model was designed and developed to quantify the cumulative impacts of individual behavioral changes on regional dynamics of saved energy and CO_2 emissions (40). The BENCH (Behavioral change in ENergy Consumption of Households) model[1] builds up on the advances in agent-based modeling applied in the energy domain, and adds theoretically and empirically grounded individual behavioral rules that drive households' energy-related choices (40,52).

Driven by the empirical evidence from environmental behavioral studies (56–61), the *BENCH* model assumes that a decision regarding energy action is driven by psychological and social factors in addition to the standard economic drivers such as prices relative to incomes (40,52,55). Behavioral factors including personal norms and awareness may either amplify the economic logic behind a decision-making or impede it, serving either as a trigger or a barrier (Figure 2).

Household agents in *BENCH* are heterogeneous in socioeconomic characteristics, preferences, and awareness of environment and climate change, so they can pursue various energy choices and actions. Figure 3 illustrates our conceptual framework that represents household energy behavioral change as a dynamic process unfolding in stages. Knowledge and awareness can have an important role in triggering individual behavior change (4,15,62–67). If individuals have enough knowledge and awareness about climate, environment, and energy issues, a feeling of guilt develops and activates motivational factors, which may lead to energy-related behavior change. Motivation is enhanced by personal and social norms (58,67), which can lead to a feeling of responsibility and provoke an individual to change their behavior. When intentions for the latter are high, individuals do a formal feasibility assessment according their income, dwelling conditions, and own perceived behavioral control. Individuals compare their current energy-use habits with alternatives, and if things can be improved, the intention to pursue an alternative rises and may lead to a behavior change. This conceptual framework combines some behavioral constructs that are common between TPB (in red) and NAT (in green).

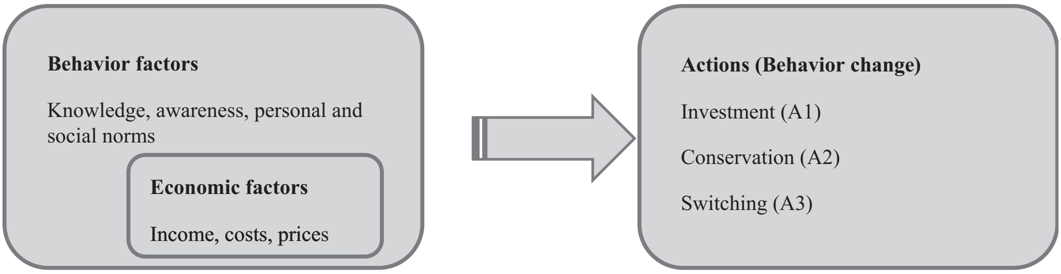


FIGURE 2 Factors affecting household decision-making regarding their energy use. (From Ref. 52.)

---

[1] **B**ehavioral change in **EN**ergy **C**onsumption of **H**ouseholds Model.

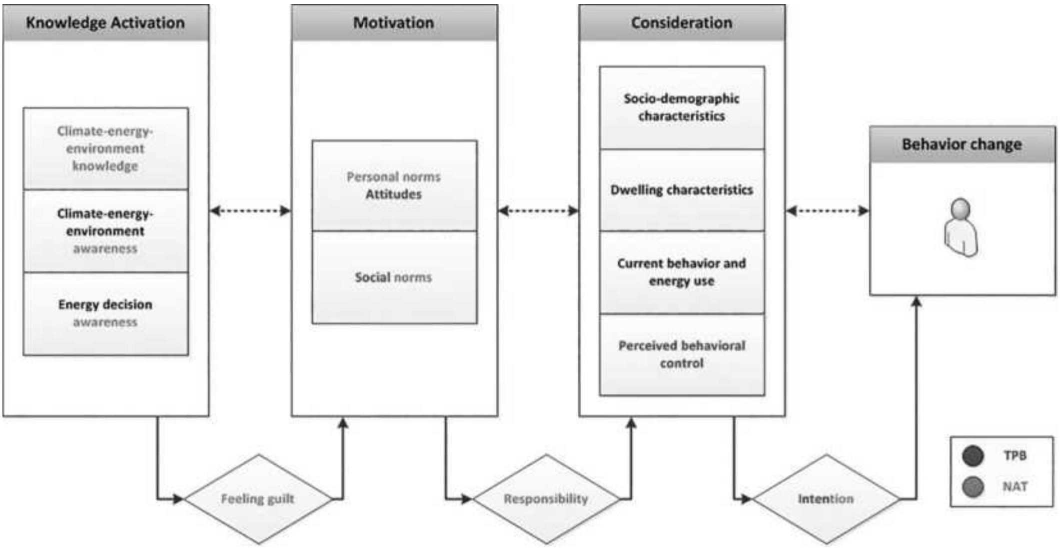


FIGURE 3 conceptual representation of multi-stage household behavioral change. (From Ref. 68.)

Empirical Data

Household agents in *BENCH* are heterogeneous in socioeconomic characteristics, preferences, and awareness of environment and climate change, so they can pursue various energy choices and actions. In order to quantify which factors – socioeconomic (e.g., income, age), behavioral (e.g., personal and social norms, knowledge and awareness about the environment, social influence), and structural (e.g., size and type of house) – trigger or attenuate a transition to a lower and greener energy footprint at the household level, the comprehensive survey (*N*=1,790 households) was designed and conducted in two provinces in Europe.

Our data analysis demonstrates that awareness and personal and social norms are equally as important as monetary factors when it comes to individual energy actions. Education and structural dwelling factors, e.g., size and type of dwelling appear to be very significant in bottom-up actions contributing to the reduction of the regional CO_2 footprint from the residential sector (68) (Figure 4).

End-User Scenarios

The *BENCH* calculated changes in electricity consumption annually and implied carbon emissions based on the primary source of energy by simulating individuals' behavior under different end-user behavioral and climate scenarios. The results indicate that accounting for demand-side heterogeneity provides a better insight into possible transition pathways to a low-carbon economy and climate change mitigation. Namely, the model including household heterogeneity, as represented by socio-demographic, dwelling, and behavioral factors, shows rich dynamics and provides a more realistic image of socioeconomics by simulating the economy through the social interactions of heterogeneous households.

Four end-user scenarios are designed and analyzed, which varied from the baseline scenario through the introduction of agent heterogeneity, the intensity of social interactions among households (slow or fast), and the lack or presence of carbon price (52). By comparing end-user scenarios, the relative impact of bottom-up drivers (social dynamics and learning on the diffusion of information) and top-down market policies (carbon price) on carbon emission reduction are estimated (Figure 5).

The impact of household attribute heterogeneity and social dynamics brought about a 5–9% CO_2 emission reduction by 2030. Adding carbon price cuts CO_2 emission down to 55% compared to the baseline scenario, which mimics the traditional economic setup of a representative rational, fully informed household making the optimal decision.

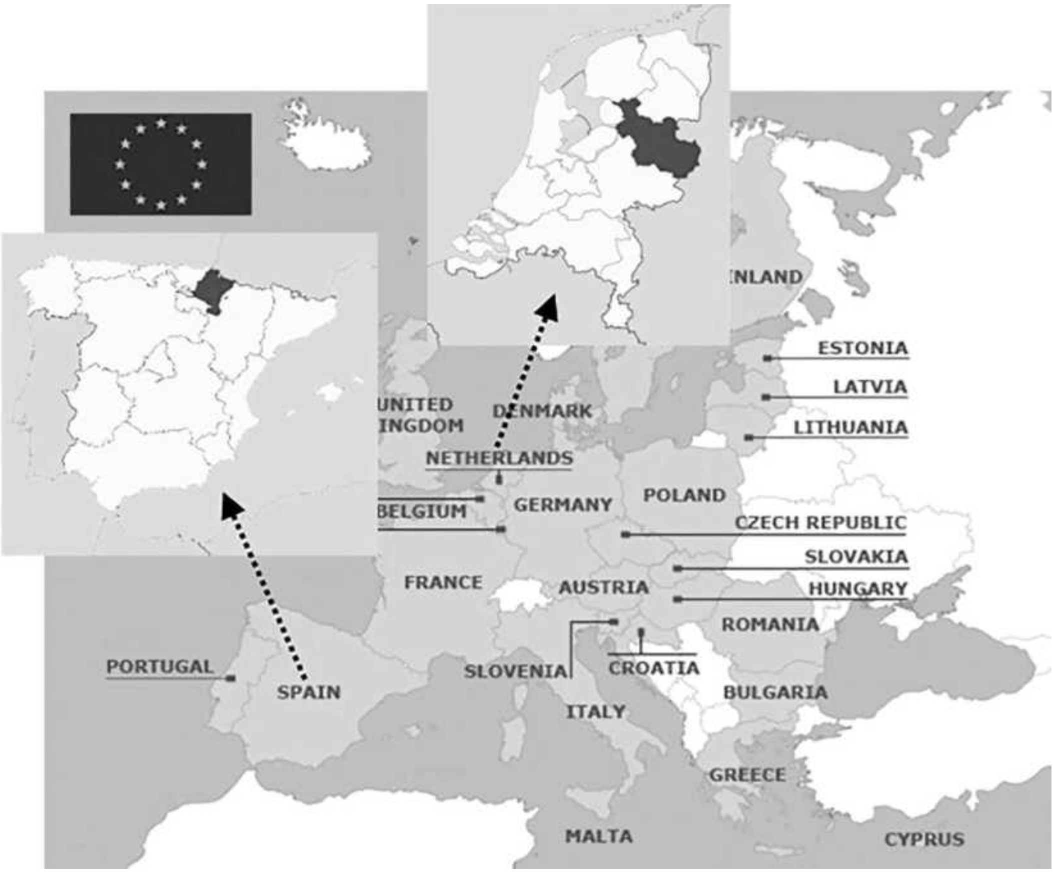


FIGURE 4 Survey case studies: the Overijssel province in the Netherlands and the Navarre province in Spain. (From Ref. 68.)

These insights have also implications for more standard policymaking. As preferences are malleable, and change with infrastructures and policy environments (13,69), policies themselves need to consider social norms and malleable preferences to avoid inefficiencies, or to even crowd out desirable outcomes (70,71).

Conclusion

Avoiding and shifting consumption plays a key role to achieve Sustainable Development Goals (SDGs) and for climate change mitigation. We can achieve the environmental targets (e.g., SDGs) through controlled consumption. Demand-side management should hence be better represented in energy systems scenarios. Demand-side management covers a broad range from technology interventions and diffusion to energy efficiency solutions and behavioral changes. Designing energy systems scenario requires an urgent understanding of which human activities are more culpable, what causes them, and how we can effectively change them. Since the nature of human is complex and heterogeneous, explaining and affecting human behavior is a difficult task. Therefore, this requires quantitative tools to assess cumulative individual emissions, given the diversity of behavior and variety of psychological and social factors influencing it beyond purely economics considerations.

The potential of individual behavioral changes in reducing carbon emissions attracts considerable attention as one of the climate change mitigation strategies (13,52). The theoretically and empirically

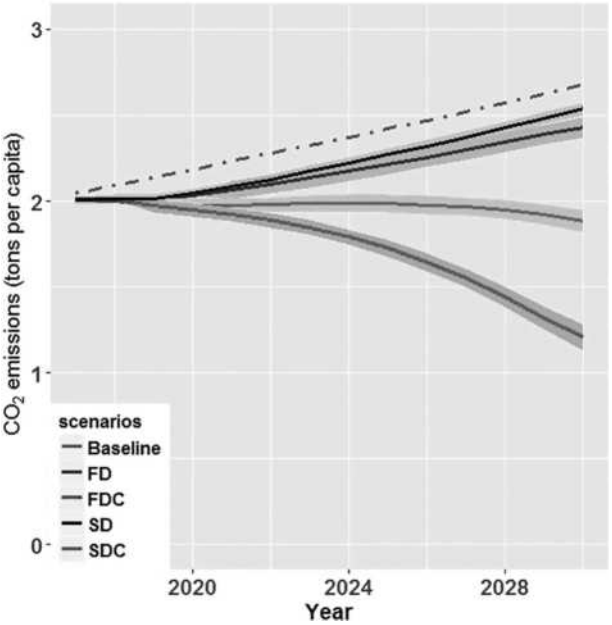


FIGURE 5 Macroimpact of heterogeneous households behavioral change combing (bottom-up) strategy and carbon price pressure (top-down) strategy on CO_2 emissions over time. Combining behavioral-climate scenarios: combination of carbon price and slow and fast social dynamics (SDC, FDC) (2017–2030). The shaded bounds around the curves indicate the uncertainty intervals across 100 runs. (From Ref. 52.)

grounded modeling tools such as the BENCH model can serve as useful instruments to quantify regional impacts of qualitative and untraceable individual behavioral aspects. The model can serve as a simulation platform to support the engagement of stakeholders. The results of this novel model indicate that accounting for the demand-side heterogeneity provides better insights into possible transition pathways to a low-carbon economy and into potential of behavioral changes as a climate change mitigation strategy (72). In order to facilitate this transition, the broader view on the social environment, cultural practices, public knowledge, producers technologies and services, and the facilities used by consumers are needed to design implementable and politically feasible policy options. It offers policymakers ways to explore various policy mixes combining price instruments (subsidies and taxes) with various targeted information policies to amplify the positive effect of individual behavioral changes regarding energy use. The reflected consideration have important implications for management practices and professionals. They point to the importance of seeing avoid and shift measures as part of management in business, administrations, and among professions.

The key message is the individuals are more than just consumers in climate change mitigation. Therefore, climate mitigation policies should go beyond economic cost-benefit incentives (e.g., subsidies and taxes). First, the social environment, cultural practices, public knowledge, producer technologies and services, and facilities used by consumers – all should be considered when designing implementable and politically feasible policy options. Second, various financial, social, and other instruments in the policy mix should be designed as a coherent set to reinforce each other, optimizing the joint effectiveness. In particular, policies such as the provision of targeted information, social advertisements, and power of celebrities for the broader public in combination with education can be used to create more knowledge and awareness in the longer run and could accompany and reinforce the effectiveness of other stimuli such as subsidies. These types of policies (soft policies) may prove to be more effective in promoting green energy solutions implemented by households compared to fiscal policy measures alone.

References

1. Grubler A, Wilson C, Bento N, Boza-Kiss B, Krey V, McCollum DL, et al. A low energy demand scenario for meeting the 1.5°C target and sustainable development goals without negative emission technologies. *Nat Energy*. 2018;3(6):515. Available from: https://www.nature.com/articles/s41560-018-0172-6
2. Dietz T, Gardner GT, Gilligan J, Stern PC, Vandenbergh MP. Household actions can provide a behavioral wedge to rapidly reduce US carbon emissions. *Proc Natl Acad Sci U S A*. 2009 Nov 3 [cited 2019 Jun 3];106(44):18452–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19858494
3. Dietz T, Rosa EA. Effects of population and affluence on CO_2 emissions. *Proc Natl Acad Sci U S A*. 1997 Jan 7 [cited 2019 Jun 3];94(1):175–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/8990181
4. Creutzig F, Roy J, Lamb WF, Azevedo IML, Bruine de Bruin W, Dalkmann H, et al. Towards demand-side solutions for mitigating climate change. *Nat Clim Chang*. 2018 Apr 3 [cited 2019 May 27];8(4):260–3. Available from: http://www.nature.com/articles/s41558-018-0121-1
5. Stern N. Economics: Current climate models are grossly misleading. *Nature*. 2016 Feb 24 [cited 2019 May 27];530(7591):407–9. Available from: http://www.nature.com/articles/530407a
6. Farmer JD, Foley D. The economy needs agent-based modelling. *Nature*. 2009 Aug 5;460(7256):685–6.
7. Niamir L. Behavioural mitigation from individual energy choices to demand-side potential. [cited 2019 May 10]. Available from: https://ris.utwente.nl/ws/portalfiles/portal/86376675/Dissertation_LeilaNiamir_BehaviouralClimateChangeMitigation.pdf
8. Niamir L, Filatova T. From climate change awareness to energy efficient behaviour. *Int Congr Environ Model Softw*. 2016 Jul 13 [cited 2019 Jun 3]. Available from: https://scholarsarchive.byu.edu/iemssconference/2016/Stream-A/74
9. Niamir L. *Behavioural Climate Change Mitigation*. Enschede, The Netherlands: University of Twente; 2019 [cited 2019 May 27]. Available from: http://purl.org/utwente/doi/10.3990/1.9789036547123
10. Wilson C, Grubler A, Gallagher KS, Nemet GF. Marginalization of end-use technologies in energy innovation for climate protection. *Nat Clim Chang*. 2012;2(11):780–8. Available from: http://www.nature.com/nclimate/journal/v2/n11/abs/nclimate1576.html
11. Spaargaren G. Theories of practices: Agency, technology, and culture: Exploring the relevance of practice theories for the governance of sustainable consumption practices in the new world-order. *Glob Environ Chang*. 2011 Aug 1 [cited 2019 Jun 12];21(3):813–22. Available from: https://www.sciencedirect.com/science/article/abs/pii/S0959378011000379?via%3Dihub
12. Tribe LH, Schelling CS, Voss J. *When Values Conflict: Essays on Environmental Analysis, Disco urse, and Decision*. Cambridge, MA: Ballinger Pub. Co; 1976 [cited 2019 Jun 12]. 178 p. Available from: https://www.tib.eu/en/search/id/TIBKAT%3A022211993/When-values-conflict-essays-on-environmental-analysis/
13. Creutzig F, Fernandez B, Haberl H, Khosla R, Mulugetta Y, Seto KC. Beyond Technology: Demand-Side Solutions for Climate Change Mitigation. 2016 [cited 2019 May 27]. Available from: www.annualreviews.org
14. Creutzig F, Jochem P, Edelenbosch OY, Mattauch L, van Vuuren DP, McCollum D, et al. Transport: A roadblock to climate change mitigation? *Science*. 2015;350(6263):911–2. Available from: http://www.sciencemag.org/content/350/6263/911
15. Creutzig F. Evolving narratives of low-carbon futures in transportation. *Transport Rev*. 2015 [cited 2019 Jun 3];36:341–60. Available from: http://dx.doi.org/10.1080/01441647.2015.1079277
16. Nykvist B, Whitmarsh L. A multi-level analysis of sustainable mobility transitions: Niche development in the UK and Sweden. *Technol Forecast Soc Chang*. 2008 [cited 2019 Jun 12];75:1373–87. Available from: http://www.matisse-project.net

17. Creutzig F, Baiocchi G, Bierkandt R, Pichler P-P, Seto KC. Global typology of urban energy use and potentials for an urbanization mitigation wedge. *Proc Natl Acad Sci*. 2015;112(20):6283–8. Available from: http://www.pnas.org/content/112/20/6283
18. Ewing R, Bartholomew K, Winkelman S, Walters J, Chen D. Growing cooler: the evidence on urban development and climate change. *Renew Resour J*. 2009;25(4):6–13.
19. Echenique MH, Hargreaves AJ, Mitchell G, Namdeo A. Growing cities sustainably. *J Am Plan Assoc*. 2012 Apr [cited 2019 Jun 19];78(2):121–37. Available from: http://www.tandfonline.com/doi/abs/10.1080/01944363.2012.666731
20. Salon D, Boarnet MG, Handy S, Spears S, Tal G. How do local actions affect VMT? A critical review of the empirical evidence. *Transp Res Part D Transp Environ*. 2012 Oct 1 [cited 2019 Jun 19];17(7):495–508. Available from: https://www.sciencedirect.com/science/article/pii/S136192091200051X
21. Cairns S, Sloman L, Newson C, Anable J, Kirkbride A, Goodwin P. Smarter choices: Assessing the potential to achieve traffic reduction using 'soft measures.' *Transp Rev*. 2008 Sep [cited 2019 Jun 19];28(5):593–618. Available from: http://www.tandfonline.com/doi/abs/10.1080/01441640801892504
22. UNWTO (UNWorld Tour.Organ.), UNEP (UNEnviron. Progr.). Climate Change and Tourism: Responding to Global Challenges. 2008 [cited 2019 Jun 19]. Available from: www.unep.fr
23. Fujino J, Hibino G, Ehara T, Matsuoka Y, Masui T, Kainuma M. Back-casting analysis for 70% emission reduction in Japan by 2050. *Clim Policy*. 2008 Jan [cited 2019 Jun 19];8(sup1):S108–24. Available from: http://www.tandfonline.com/doi/abs/10.3763/cpol.2007.0491
24. Skea J, Ekins P, Winskel M. Energy 2050: Making the transition to a secure low carbon energy system. *Earthscan*. 2011 [cited 2019 Jun 19]. 381 p. Available from: https://www.ingentaconnect.com/content/rout/324pkr/2011/00000001/00000001/art00010;jsessionid=6olit2pe2ffh0.x-ic-live-02
25. Wei Y-M, Liu L-C, Fan Y, Wu G. The impact of lifestyle on energy use and CO_2 emission: An empirical analysis of China's residents. *Energy Policy*. 2007 Jan 1 [cited 2019 Jun 19];35(1):247–57. Available from: https://www.sciencedirect.com/science/article/pii/S0301421505003022
26. Shukla PR, Dhar S, Mahapatra D. Low-carbon society scenarios for India. *Clim Policy*. 2008 Jan [cited 2019 Jun 19];8(sup1):S156–76. Available from: http://www.tandfonline.com/doi/abs/10.3763/cpol.2007.0498
27. Jaboyedoff P, Roulet C-A, Dorer V, Weber A, Pfeiffer A. Energy in air-handling units—results of the AIRLESS European Project. *Energy Build*. 2004 Apr 1 [cited 2019 Jun 19];36(4):391–9. Available from: https://www.sciencedirect.com/science/article/pii/S0378778804000477
28. Lin Z, Deng S. A study on the characteristics of nighttime bedroom cooling load in tropics and subtropics. *Build Environ*. 2004 Sep 1 [cited 2019 Jun 20];39(9):1101–14. Available from: https://www.sciencedirect.com/science/article/pii/S0360132304000472
29. Goswami DY, Kreith F. *Energy Efficiency and Renewable Energy Handbook*.
30. Smith P, Haberl H, Popp A, Erb K, Lauk C, Harper R, et al. How much land-based greenhouse gas mitigation can be achieved without compromising food security and environmental goals? *Glob Chang Biol*. 2013 Aug 1 [cited 2019 Jun 20];19(8):2285–302. Available from: http://doi.wiley.com/10.1111/gcb.12160
31. Edenhofer O, Pichs-Madruga R, Sokona Y, Farahani E, Kadner S, Seyboth K, et al., editors. *IPCC, 2014: Climate Change 2014: Mitigation of Climate Change. Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press.
32. Stehfest E, Bouwman L, van Vuuren DP, den Elzen MGJ, Eickhout B, Kabat P. Climate benefits of changing diet. *Clim Change*. 2009 Jul 4 [cited 2019 May 27];95(1–2):83–102. Available from: http://link.springer.com/10.1007/s10584-008-9534-6
33. Vandenbergh MP, Nielsen KS. From myths to action. *Nat Clim Chang*. 2019 Jan 17 [cited 2019 Jun 12];9(1):8–9. Available from: http://www.nature.com/articles/s41558-018-0357-9
34. Hendrickson CY. Global carbon pricing: The path to climate cooperation. *Glob Environ Polit*. 2018 Jun 28;18(3):159–61. Available from: https://doi.org/10.1162/glep_r_00475

35. Shove E, Watson M, Spurling N. Conceptualizing connections. *Eur J Soc Theory*. 2015 Aug 22 [cited 2019 Jun 12];18(3):274–87. Available from: http://journals.sagepub.com/doi/10.1177/1368431015579964
36. Crate SA, Nuttall M. *Anthropology and Climate Change: From Actions to Transformations*. Routledge; 2019 [cited 2019 Jun 12]. 450 p. Available from: https://www.routledge.com/Anthropology-and-Climate-Change-From-Actions-to-Transformations-2nd-Edition/Crate-Nuttall/p/book/9781629580012
37. Nemet GF. *How Solar Energy Became Cheap: A Model for Low-carbon Innovation*. Routledge; 2019 [cited 2019 Jun 12]. Available from: https://www.routledge.com/How-Solar-Energy-Became-Cheap-A-Model-for-Low-Carbon-Innovation/Nemet/p/book/9780367136598
38. Creutzig F, Roy J, Lamb WF, Azevedo IML, Bruine De Bruin W, Dalkmann H, et al. Towards demand-side solutions for mitigating climate change. *Nat Clim Change*. 2018 [cited 2019 May 27];8:260–3. Available from: www.nature.com/natureclimatechange
39. Steinberger JK, Roberts JT, Peters GP, Baiocchi G. Pathways of human development and carbon emissions embodied in trade. *Nat Clim Change*. 2012 Feb 22 [cited 2019 Jun 12];2(2):81–5. Available from: http://www.nature.com/articles/nclimate1371
40. Niamir L, Filatova T, Voinov A, Bressers H. Transition to low-carbon economy: Assessing cumulative impacts of individual behavioral changes. *Energy Policy*. 2018;118:325–45.
41. Riahi K, van Vuuren DP, Kriegler E, Edmonds J, O'Neill BC, Fujimori S, et al. The Shared Socioeconomic Pathways and their energy, land use, and greenhouse gas emissions implications: An overview. *Glob Environ Change*. 2017 Jan 1 [cited 2019 Jun 12];42:153–68. Available from: https://www.sciencedirect.com/science/article/pii/S0959378016300681
42. Kriegler E, Weyant JP, Blanford GJ, Krey V, Clarke L, Edmonds J, et al. The role of technology for achieving climate policy objectives: overview of the EMF 27 study on global technology and climate policy strategies. *Clim Change*. 2014 Apr 28 [cited 2019 Jun 12];123(3–4):353–67. Available from: http://link.springer.com/10.1007/s10584-013-0953-7
43. O'Neill BC, Kriegler E, Riahi K, Ebi KL, Hallegatte S, Carter TR, et al. A new scenario framework for climate change research: the concept of shared socioeconomic pathways. *Clim Change* [Internet]. 2014 Feb 15 [cited 2019 Jun 12];122(3):387–400. Available from: http://link.springer.com/10.1007/s10584-013-0905-2
44. Babatunde KA, Begum RA, Said FF. Application of computable general equilibrium (CGE) to climate change mitigation policy: A systematic review. *Renew Sustain Energy Rev*. 2017 Oct 1 [cited 2019 Jun 12];78:61–71. Available from: https://www.sciencedirect.com/science/article/pii/S1364032117305713
45. Kancs A. Predicting European enlargement impacts: A framework of interregional general equilibrium. *East Europ Econ*. 2001 Sep 8 [cited 2019 Jun 5];39(5):31–63. Available from: https://www.tandfonline.com/doi/full/10.1080/00128775.2001.11041001
46. Siagian U, Yuwono B, Fujimori S, Masui T, Siagian UWR, Yuwono BB, et al. Low-carbon energy development in indonesia in alignment with intended nationally determined contribution (INDC) by 2030. *Energies*. 2017 Jan 5 [cited 2019 Jun 5];10(1):52. Available from: http://www.mdpi.com/1996-1073/10/1/52
47. Hunt LC, Evans J. *International Handbook on the Economics of Energy*. Edward Elgar; 2009. 831 p.
48. Bhattacharyya SC. *Energy Economics: Concepts, Issues, Markets and Governance*. Springer; 2011. 721 p.
49. Niamir L, Filatova T. Linking agent-based energy market with computable general equilibrium model: An integrated approach to climate-economy-energy system. 2015 [cited 2019 Jun 3]; Available from: https://ris.utwente.nl/ws/portalfiles/portal/17707456/linking.pdf
50. Filatova T, Niamir L. Changing climate – changing behavior: empirical agent based computational models for climate change economics. In: Lamperti F, Monasterolo I, Roventini A, editors. *Complexity and Climate Change: An Evolutionary Political Economy Approach*. Taylor & Francis; 2019.

51. Lebaron B, Tesfatsion LS, Lebaron B, Tesfatsion L. Modeling macroeconomies as open-ended dynamic systems of interacting agents. *The American Economic Review* 2008 Aug 19 [cited 2019 Jun 12];98:246–250. Available from: https://econpapers.repec.org/paper/isugenres/12973.htm
52. Niamir L, Kiesewetter G, Wagner F, Schöpp W, Filatova T, Voinov A, et al. Assessing the macroeconomic impacts of individual behavioral changes on carbon emissions. *Clim Change*. 2019.
53. Rai V, Robinson SA. Agent-based modeling of energy technology adoption: Empirical integration of social, behavioral, economic, and environmental factors. *Environ Model Softw*. 2015 Aug 1 [cited 2019 Jun 12];70:163–77. Available from: https://www.sciencedirect.com/science/article/pii/S1364815215001231?via%3Dihub
54. Rai V, Henry AD. Agent-based modelling of consumer energy choices. *Nat Clim Chang*. 2016 Jun 9 [cited 2019 Jun 12];6(6):556–62. Available from: http://www.nature.com/articles/nclimate2967
55. Niamir L, Filatova T. Transition to low-carbon economy: Simulating nonlinearities in the electricity market, Navarre region-Spain. Vol. 528, Advances in Intelligent Systems and Computing. 2017.
56. Abrahamse W, Steg L. Factors related to household energy use and intention to reduce it: The role of psychological and socio-demographic variables. *Human Ecol Rev*. 2011 [cited 2019 Jun 12];18:30–40. Available from: https://www.jstor.org/stable/24707684
57. Bamberg S, Rees J, Seebauer S. Collective climate action: Determinants of participation intention in community-based pro-environmental initiatives. *J Environ Psychol*. 2015 Sep 1 [cited 2019 Jun 12];43:155–65. Available from: https://www.sciencedirect.com/science/article/abs/pii/S0272494415300190?via%3Dihub
58. *Bamberg* S, Hunecke M, Blöbaum A. Social context, personal norms and the use of public transportation: Two field studies. *J Environ Psychol*. 2007 Sep [cited 2019 Jun 12];27(3):190–203. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0272494407000357
59. *Steg* L. Values, norms, and intrinsic motivation to act proenvironmentally. *Annu Rev Environ Resour*. 2016 Nov 20 [cited 2019 Jun 12];41(1):277–92. Available from: http://www.annualreviews.org/doi/10.1146/annurev-environ-110615-085947
60. Steg L, Vlek C. Encouraging pro-environmental behaviour: An integrative review and research agenda. *J Environ Psychol*. 2009 Sep 1 [cited 2019 Jun 12];29(3):309–17. Available from: https://www.sciencedirect.com/science/article/abs/pii/S0272494408000959
61. Mills B, Schleich J. Residential energy-efficient technology adoption, energy conservation, knowledge, and attitudes: An analysis of European countries. *Energy Policy*. 2012 Oct 1 [cited 2019 Jun 12];49:616–28. Available from: https://www.sciencedirect.com/science/article/pii/S0301421512005897
62. Matsui K, Ochiai H, Yamagata Y. Feedback on electricity usage for home energy management: A social experiment in a local village of cold region. *Appl Energy*. 2014 May 1 [cited 2019 Jun 12];120:159–68. Available from: https://www.sciencedirect.com/science/article/pii/S0306261914000683
63. Long C, Mills BF, Schleich J. Characteristics or culture? Determinants of household energy use behavior in Germany and the USA. *Energy Effic*. 2018 Mar 8 [cited 2019 Jun 12];11(3):777–98. Available from: http://link.springer.com/10.1007/s12053-017-9596-2
64. Kobus CBA, Klaassen EAM, Mugge R, Schoormans JPL. A real-life assessment on the effect of smart appliances for shifting households' electricity demand. *Appl Energy*. 2015 Jun 1 [cited 2019 Jun 12];147:335–43. Available from: https://www.sciencedirect.com/science/article/pii/S0306261915001099
65. Donaldson SI, Grant-Vallone EJ. Understanding self-report bias in organizational behavior research. *J Bus Psychol*. 2002 [cited 2019 Jun 12];17(2):245–60. Available from: http://link.springer.com/10.1023/A:1019637632584
66. Ajzen I, Fishbein M. *Understanding Attitudes and Predicting Social Behavior*. Prentice-Hall; 1980. 278 p.

67. Abrahamse W, Steg L. How do socio-demographic and psychological factors relate to households' direct and indirect energy use and savings? *J Econ Psychol*. 2009 Oct 1 [cited 2019 Jun 12];30(5):711–20. Available from: https://www.sciencedirect.com/science/article/abs/pii/S0167487009000579
68. Niamir L, Ivanova O, Filatova T, Voinov A, Bressers H. Demand-side solutions for climate change mitigation: bottom-up drivers of household energy behavior changestle. *Energy Res Soc Sci*. 2019.
69. Mattauch L, Ridgway M, Creutzig F. Happy or liberal? Making sense of behavior in transport policy design. *Transp Res Part D Transp Environ*. 2016 Jun 1 [cited 2019 Jun 12];45:64–83. Available from: https://www.sciencedirect.com/science/article/pii/S136192091500111X
70. van den Bijgaart I. Too slow a change? Deep habits, consumption shifts and transitory tax policy. 2018 [cited 2019 Jun 12]. CESifo Working Paper Series No. 6958. Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3185965
71. Mattauch L, Hepburn C, Stern N. Pigou pushes preferences: Decarbonisation and endogenous values. 2018 [cited 2019 Jun 12]. Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3338758
72. Bressers H, Ligteringen J. Political-administrative policies for sustainable household behavior. 2001 [cited 2019 Jun 12]. Available from: https://research.utwente.nl/en/publications/political-administrative-policies-for-sustainable-household-behav

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

19

Wind Farms: Noise

Daniel Shepherd,
Chris Hanning,
and Bob Thorne

Introduction

Planning authorities, environmental agencies, and policy makers in many parts of the world are seeking information on possible links between wind turbine noise and health in order to legislate permissible noise levels or setback distances. Concurrently, larger and noisier wind turbines are emerging, and consent is being sought for progressively larger windfarms to be placed even closer to human habitats. While noise standards can effectively and fairly facilitate decision-making processes if developed properly, the current standards on offer suffer severe conceptual difficulties. Specifically, noise metrics considered by many in the industry as best practice may in fact relate little to health outcome variables such as annoyance or sleep disruption. In this entry, we describe the physical characteristics of wind turbine noise, review the impact of such noise on humans, and critique current approaches to mitigation.

Industrial Wind Turbines

Industrial wind turbines transform kinetic energy from the wind into electricity, a practice dating back over 100 years. Structurally, wind turbines can be decomposed into three key components (Figure 1). First, wind turbines possess a rotor, consisting of one or more blades designed to rotate when exposed to wind. The rotor can be thought of as a type of sail, catching wind in order to induce movement. Depending on the axis of blade rotation, wind turbines can be categorized as either horizontal-axis (the most common) or vertical-axis turbines. The second major component is the generator or "dynamo." The generator component includes a gearbox to regulate the speed of the dynamo and components to change blade pitch and plane of rotation with respect to wind direction. The dynamo can be used as a motor to maintain rotation at very low wind speeds. Third, there is a tower supporting the rotor and, typically, the generator. The size of a wind turbine can be specified either as a dimension (e.g., tower

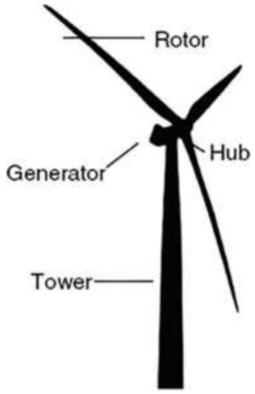


FIGURE 1 Components of a typical horizontal-axis wind turbine.

height measured from the ground to the top of a blade at its highest point) or as an electrical output (e.g., watts). Currently, turbines range from approximately 2 to 200 m high and from approximately 50 W to 6 MW.

Wind turbines can be erected in isolation or in sets and be located either onshore (i.e., terrestrial) or offshore (i.e., marine), though the latter is associated with higher construction costs. Industrial-scale wind energy generation, involving the saturation of an optimum number of wind turbines in a fixed area of land, gives rise to the concept of the "windfarm" or "wind park." Wind energy developers seek areas that have good consistent wind flow and close access to energy grids. The proliferation in the number of windfarms established globally in the past decade has been largely driven by environmental concerns such as climate change, renewableness and sustainability, and strategic energy considerations relating to the depletion of fossil fuels.[1] However, in the absence of large-scale electricity storage devices (i.e., batteries), the contribution of wind energy to a nation's electricity needs is likely to be peripheral. Another barrier is social acceptance, with reviewed social surveys indicating citizens supporting renewable energy in principle but opposed to having windfarms in their immediate vicinity due to visual impacts on the landscape, shadow flicker from the blades, and fears of noise-induced annoyance and sleep disruption.

Acoustic Profile of Wind Turbine Noise

The sound generated from a windfarm is qualitatively different from any sound source commonly met in the environment, can rapidly switch from being stationary to nonstationary, and can vary by as much as 20 dB within a single minute. When it interferes with human activities, wind turbine sound becomes a type of noise. Analysis of windfarm noise poses distinct challenges, including the identification of acoustic energy that can be directly attributed to the turbines and the detection of special audible characteristics, including distinct tonal complexes and modulation effects. Windfarm noise is often a broadband low-amplitude noise constantly shifting in character ("waves on beach," "rumble-thump," "plane never landing." etc.). In this respect, windfarm noise is not like, for example, traffic noise or the continuous hum from plant and machinery. When assessed, wind turbine noise is often related to either wind speed (m/s) or electrical output (watts) and typically increases with both.

When the wind reaches a blade, it flows both over and under the blade. The part of the airflow with momentum great enough to break away forms trailing vortexes and turbulence behind the blade, producing a set of sound sources. The power of each sound source depends on the strength of the turbulence, which in turn depends on the speed of airflow; the compressibility and viscosity of the air; the design and surface texture (roughness) of the blade; the wind speed; and the velocity of the blade at that point. The faster the blade rotates, the earlier the breakup in the boundary vortexes and the greater the interaction between the vortexes emanating by adjacent wind turbines. An amplification of potential noise occurs when two or more turbines are, or nearly are, synchronous, such that the blade passing pulses coincide and then go out of phase again.[2] With exact synchronicity, there is a fixed interference pattern; with near synchronicity, concurrent arrival of pulses will change over time and place.

Noise emissions from modern wind turbines are primarily due to turbulent flow and trailing edge sound, blade characteristics, blade/tower interaction, and to a lesser degree, mechanical processes. The most commonly used description of wind turbine noise is the A-weighted sound pressure level, which is expressed in decibels (notated dBA). The most commonly used noise compliance assessment methods for windfarms involve the "time average" sound level L_{Aeq} or the background sound level L_{A90}. These levels are quite different as the time-averaged ambient sound level includes all noises from near and far. The difference between these levels, and other common levels, is illustrated in Figure 2. The chart shows that sound levels change over time and that any derived sound level index is a summary of fluctuating levels in that time period. In a relatively short time period, such as 10 minutes, the unique noise events such as bangs or thuds from turbines shifting in the wind may be captured. If the time period is relatively long, for example, an hour, then evidence of unique short-term noise events is reduced because the sound energy is "averaged" over the whole hour, and the single-value A-weighted level will not represent short-term variations in sound character. If extraneous noise (e.g., insect noise) is included in the wind turbine measurement, its contribution to the overall level must be determined, though how this is undertaken remains a challenge.[3]

The A-frequency–weighted sound pressure level or "sound level" is the most common sound descriptor and is reputedly analogous to our hearing at medium sound levels. This is not strictly true, and the A-weighting has a significant restriction in that it does not permit measurement or assessment of low-frequency sound (i.e., 20 to 250 Hz). For more complex situations where dominant tonal components are significant (i.e., special audible characteristics), a procedure for determining tonal adjustment requiring one-third octave band frequency or narrow-band analysis is needed. These assessment procedures require the "C" weighting for low frequency or the unweighted (also known as "Z") response to measure both low-frequency and infrasonic sound. Whereas the dBC metric is able to include low-frequency sounds such as the audible rumble and thump from wind turbines, the dBZ response is more

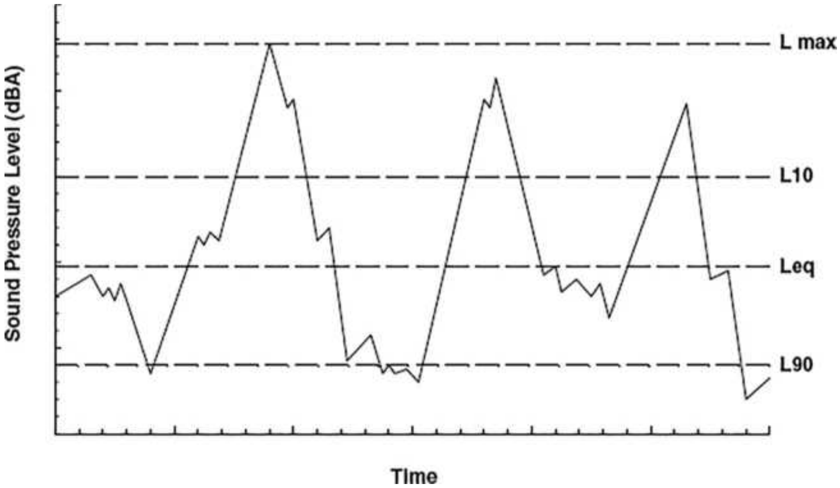


FIGURE 2 Chart illustrating different noise descriptors. L10 is the level exceeded 10% of the time, while L90 is the level exceeded 90% of the time. The time-average (equivalent continuous) sound pressure level, Leq, represents the average acoustic energy across a defined measurement epoch.

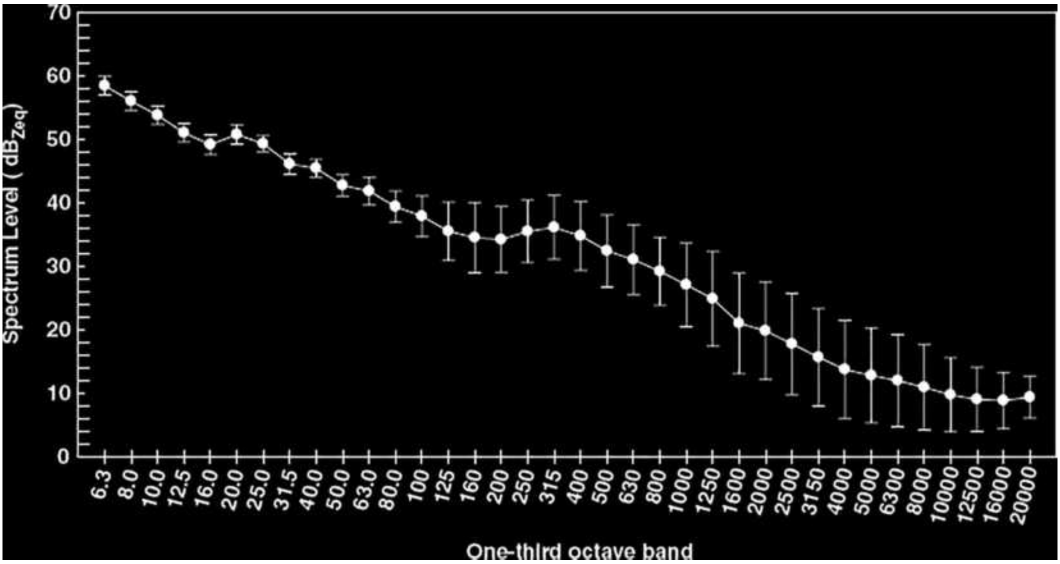


FIGURE 3 One-third octave band analysis of time-average unweighted sound pressure level (dB$_{Zeq}$) for wind turbine sound measured from 7:00 p.m. to 1:00 a.m. outside of a residence.

suitable for infrasound measurements (i.e., typically inaudible energy below 20 Hz). Figure 3 presents a third octave band analysis of outdoor wind turbine noise recorded over a 6-hr period. Other measures include assessments for tonality or low- frequency sound referenced to third octave bands and the "G" weighting for infrasound. Aside from physical measures of amplitude (e.g., dBA), wind turbine noise can be quantified with a variety of other acoustical and objective psychoacoustic measures, including amplitude modulation (for example, 100 msec samples of peak, time-average, or fast response), sound quality (including audibility, dissonance, roughness, fluctuation strength, sharpness, tonality), loudness (for steady, time-varying, and impulsive sounds), and unbiased annoyance.[4]

Certification of wind turbine noise is undertaken in accordance with the International Standard IEC 61400-11:2002.[5] Emission levels are to be reported as A- weighted time-averaged (L_{Aeq}) sound levels in one-third octave bands. Audibility is calculated by reference to tones. An informative entry in IEC 61400-11 states the following: "In addition to those characteristics of wind turbine noise described in the main text, this emission may also possess some, or all of the following: infrasound; low-frequency noise; impulsivity; low-frequency modulation of broad band or tonal noise; other, such as a whine, hiss, screech, or hum, etc., distinct pulses in the noise, such as bangs, clatters, clicks or thumps, etc." Unfortunately, many of these parameters are not reported by the turbine manufacturer and cannot be predicted with the simple calculation methods currently available. The prediction of windfarm sound levels is most often referenced to national or international standards that have been based on ISO 9613-2.[6] The propagation method is calculated with the receivers being downwind from the noise source(s). All prediction models have uncertainty to their accuracy of prediction. Table 5 of the ISO 9613-2 standard gives an estimated accuracy for broadband noise of ±3 dB at between 100 and 1000 m. This is due to the inherent nature of the calculation algorithms that go into the design of the model, the assumptions made in the implementation of the model, and the availability of good source sound power data. The ISO 9613-2 method holds for wind speeds of between approximately 1 and 5 m/s, measured at a height of 3 to 11 m above the ground. However, wind turbines are sound sources that operate at higher wind speeds than allowed for under the standard, and an accuracy of ±7 dB can be expected. [3] Ultimately, the received noise levels at residences will vary subject to varying meteorological conditions in the locality (e.g., wind speed and direction, wind shear, temperature, humidity, inversions),

TABLE 1 Factors Affecting the Prediction of Wind Farm Noise Levels at a Receiver[a]

- The true sound power level of the turbine(s) at the specified wind speed
- The reduction in sound level due to ground effects
- The increase or reduction in sound level due to atmospheric (meteorological) variations and wind direction
- The variation due to modulation effects from wind velocity gradient
- Increase and reduction in sound levels due to wake and turbulence modulation effects due to turbine placement and wind direction
- Increased sound levels due to synchronicity effects of turbines in phase due to turbine placement and wind direction
- Building resonance effects for residents inside a dwelling

[a] A conservative set of noise predictions should take all factors into account.

among other factors (see Table 1), all of which must be accounted for when measuring or modeling wind turbine noise levels.

Human Impacts of Wind Turbine Noise

A Psychological Description of Wind Turbine Noise

At the psychological level of description, wind turbine noise is most frequently characterized as a swishing or lashing sound or less commonly as thump/throb, low- frequency rumble, or a rustling sound.[7,8] Wind turbines produce noise with an impulsive character[9] and while the actual cause of the swishing or thumping has not yet been fully elucidated, it has been demonstrated that the swishing or thumping pattern is common with larger turbines[10] and may result from a fluctuating angle of attack between the trailing edge of the rotor blade and wind, or wind speed inequalities across the area being swept by the rotor blades.[11] It is thought that the swishing sound may be linked to activity in the 2000 to 4000 Hz band, with the pace of the rotor blades determining the degree of amplitude modulation.[12] Unfortunately, such amplitude-modulated sounds are generally attenuated poorly by background noise, especially so in rural areas.[13] Further, because human sensory systems behave as contrast analyzers, fluctuations in the incoming stimulus field tend to direct attention and so are more easily detected. Thus, amplitude-modulated sounds such as wind turbine noise are readily perceived and difficult to filter out, making them especially intrusive.[14] The loudness of a wind turbine depends on a number of factors, including wind speed, sound-attenuating materials between the turbines and the receiver, other masking sounds, the season, and time of day. The loudness of a modern 2 to 3 MW wind turbine can be compared to a car on a motorway, autobahn, or freeway,[15] with a sound power level of 94 to 104 dBA at a windspeed of 8 m/s.[16] Wind turbine noise is perceived louder at night and during the summer months and when the wind is blowing from the direction of the turbines toward the receiver.[7,8]

Quantifying the Health Impacts of Wind Turbine Noise

Elucidating a causal mechanism between an environmental event and health is a complicated undertaking, and noise effects are commonly “indirect” as opposed to “direct.” According to the biomedical model of health (Figure 4a), a direct health effect implies a direct pathological relationship between an environmental parameter (e.g., noise level) and a target organ. An alternative approach (Figure 4b) distinguishes between direct health effects and psychosomatic illness, the latter indicting that any physiological illness coinciding with the onset of wind turbine noise is caused by a negative psychological response to the noise and not the noise *per se.* Thus, anxiety or anger in the presence of wind turbine noise induces stress and strain that, if maintained, can eventually lead to adverse health effects. A counter argument to this approach is that some individuals are simply more susceptible to noise

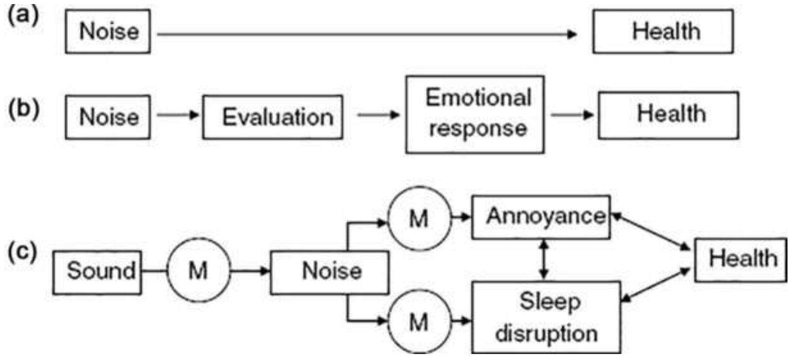


FIGURE 4 Three models representing the relationship between noise and health: the biomedical model (a) stipulating a direct causal relationship and indirect models (b and c) containing moderators and mediators.

than other individuals, which fits with the general concept of biological and physical variation. In the field of epidemiology, the differential susceptibilities of individuals are known as risk factors or vulnerabilities, with noise sensitivity being one risk factor related to negative responses to intrusive noise. A second challenge to the psychosomatic approach comes from documented instances of individuals who initially welcomed wind turbines into the community but who later campaigned to have them removed due to undesirable noise exposure.[17] Lastly, the veracity of psychosomatic arguments lessens in the face of feasible biological mechanisms describing the relationship between health and noise.[18]

An alternative and more accepted approach would be to adopt the World Health Organization's (WHO's) definition of health:[19] "A state of complete physical, mental and social well-being and not merely the absence of disease or infirmity." The forerunner of the biopsychosocial model, the WHO's definition states that optimal human functioning is determined by the interplay of biological, environmental, psychological, and social factors. Figure 4c displays a model consistent with the WHO's approach, in which the impact of noise is moderated by environmental, psychological, and social factors. A context-relevant model proposed by van den Berg and colleagues,[8] based on previous wind turbine literature, takes a similar shape to that presented in Figure 4c. They dichotomize moderators (denoted "M" in Figure 4c) into environmental moderators (e.g., degree of urbanization, house type, and ambient sound level) or psychological and demographic moderators (e.g., age, gender, education, employment status, attitudes to wind energy, noise sensitivity, and whether the individual receives a monetary return from the turbines). Other models linking wind turbine sound and health have been proposed[20] but can be considered extensions of that presented in Figure 4c.

As a new source of noise, the impact of wind turbine noise is understandably understudied relative to aviation and road traffic noise. Consequently, little data exist with which to assess the impacts of wind turbine noise on health, a state of affairs compounded by rapid development of wind turbine technology, in which data collected for smaller and less powerful turbines are not generalizable to larger, more modern turbines.[9,21] As of 2011, there have been two approaches to collecting wind turbine noise impact data, either epidemiological studies relying on masked surveys or direct clinical case studies.[22] Both approaches typically focus on the emotional impacts of noise (i.e., annoyance), upon sleep disruption, and/or the degradation of well-being and increases in stress that arise from sleep disturbance and annoyance. Irrespective of approach, however, case studies,[23–25] and epidemiological studies[7,8,20] have provided evidence that, like road traffic and aviation noise, wind turbine noise can be associated with negative health outcomes.

Wind Turbine Noise and Annoyance

People generally respond more negatively to man-made noise than to natural sounds,[26] and this generalization holds true for wind turbine noise.[16] From a psychological perspective, chronic exposure to community noise can impact health through information overload, overarousal, loss of coping strategies, loss of privacy, and loss of perceived control. These mechanisms give rise to a number of subjective responses to noise, of which the most common is annoyance. As a psychological stressor,[27] noise annoyance can express itself through malaise, fear, threat, uncertainty, restricted liberty, excitability, or defenseless- ness.[28] Furthermore, annoyance may be accompanied by intense anger, especially if one believes that they are being harmed unnecessarily. Thus, the term "annoyance" is often misinterpreted by the layperson as a feeling brought about by the presence of a minor irritant. The medical usage, in contrast, exists as a precise technical term and defines annoyance as a mental state capable of degrading health and well-being,[29,30] and it is classified as an adverse health effect by the WHO.[31]

There have been few studies estimating the health impacts of windfarms, with a series of studies undertaken in Scandinavia contributing the most to current knowledge. A seminal Swedish study undertaken by Pedersen and Persson Waye[7] sought to document the prevalence of wind turbine–induced annoyance and, further, to generate dose–response relationships between the two. Respondents were located between 150 and 1200 meters from the nearest wind turbine and were classified into noise exposure categories (see Figure 5). A significant relationship between dose (dBA) and annoyance was reported, but the variability in annoyance scores explained by noise level was small (adjusted $R^2 = 0.13$). Those reporting annoyance indicated a daily or nearly-everyday intrusion of windfarm noise. Those describing the noise as "swishing" were more likely to report noise annoyance, a finding replicated in a subsequent study reporting a high correlation (r = 0.664) between the swishing sound and annoyance.[14] Among those who noticed the noise, 11.2% reported being annoyed when indoors. A small but significant correlation was found between noise annoyance and noise sensitivity, with approximately 50% of the rural-dwelling respondents describing themselves as noise sensitive. Those making negative appraisals of the wind turbines, for example, as visually incongruent with the landscape, were at higher risk of an annoyance response. On the basis of their data, the authors undertook follow-up studies[14–16,22] supporting their conclusion that wind turbine noise maybe more potent than other categories of environmental noise (e.g., road or aviation) and appealed for further studies to determine why this might be. In a later report, Pedersen[22] suggests that coping strategy may moderate the relationship between wind turbine noise and stress.

Van den Berg et al.[8] analyzed data from 725 Dutch nationals residing within 2.1 km of a wind turbine and who were exposed to calculated outdoor noise levels between 24 and 54 dB(A). Approximately 60% of the sample could hear the turbines outdoors, while 33% reported that they could hear the wind turbines indoors. Of the 45% (n = 231) who noticed the sound of the rotor blades, 24.7% were not annoyed, 25.8% were slightly annoyed, 19.5% were rather annoyed, and 29.9% were very annoyed. The sound level explained approximately 25% of the variability in annoyance scores, and those who compared the noise to an amplitude modulation (i.e., swishing or lashing) were more likely to be annoyed, though this is not a novel finding.[14,32,33] Figure 5 plots the data from van den Berg et al., presenting proportions of detection and elicited annoyance as a function of noise level, for their entire dataset (Figure 5, circles) and for those receiving no economic benefit (Figure 5, squares). Note that, for those receiving no economic benefit, a monotonic relationship is evident, while a nonmonotonic function occurs when individuals benefiting financially from the turbines are included. Van den Berg[8] reports that this depreciation in annoyance of those benefiting economically can be explained by the control they have over the wind turbines, such that they can impede their operation if noise levels increase. Finally, it was reported that annoyance was positively correlated with stress scores, though a causal relationship could not be inferred.

It is accepted that both the physical parameters of the noise and the psychological characteristics of the listener combine to produce noise annoyance.[34] On the physical side, the relatively high annoyance

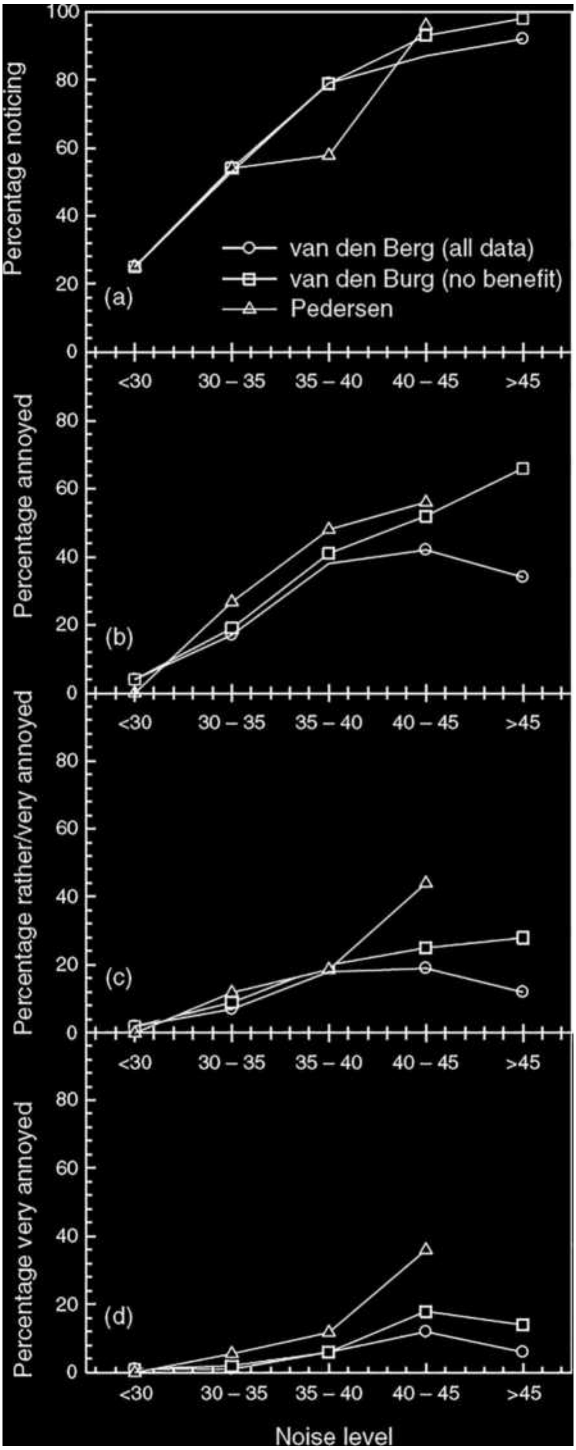


FIGURE 5 Perception of wind turbine noise as a function of noise level for three sets of data: Tables 7.25 (complete data set) and 7.26 (no economic benefit of turbines) from van den Berg et al.[8] and Pedersen and Persson Waye's[7] Table V. Plot (a) is percentage noticing the noise, while plots (b) to (d) are for annoyance. Plot (b) includes data from plots (c) and (d), and plot (c) includes data from plot (d).

levels elicited by wind turbine noises (e.g., swishing or thumping) may be explained by the increased fluctuation of the sound, up to 4 to 6 dB for a single turbine operating in a stable atmosphere.[11] Individuals are also highly sensitive to changes in frequency modulation variations of approximately 4 Hz or greater.[4] Noting that amplitude-modulated sound is known to be more annoying than unmodulated sound, Lee et al.,[34] in a laboratory setting, demonstrated that amplitude-modulated wind turbine noise was consistently judged to be more annoying than its unmodulated counterpart. Thus, the dominant acoustic driver of annoyance is likely to be noise dynamics rather than noise level. Other physical parameters linked to annoyance include terrain complexity, with rural terrain associated with greater annoyance than urban areas, possibility due to more complicated terrain exhibiting various focusing or defocusing effects and greater ground reflection.

While there is a strong correlation between the sound pressure level (i.e., amplitude) of a sound wave and the perceived loudness of a sound, there is no one-to-one mapping between sound pressure level and the psychological responses that individuals have to a sound.[35] Many non-acoustical factors determine how annoyed one will become toward a source of noise.[36–38] Thus, the response of the individual to the sound is just as important as the parameters of the acoustic wave, and the "people" side of noise should not be omitted from acoustical reports. Table 2 lists, in no particular order, non-acoustical factors found to influence levels of noise annoyance.[39] In relation to windfarms, the personal factors listed in Table 2 have been found to strongly influence how exposed individuals perceive the noise.[16] In addition, perceptions of amenity, individuals seeking refuge from urban noise, or the lower ambient sound levels typical of the rural environment may explain why annoyance responses are higher in rural as opposed to urban settings.[13,16]

When considering wind turbine noise and annoyance data emerging from the literature, a number of risk factors are evident, including an effect of age and educational status but not gender.[8] Employment status was also linked to wind turbine noise–induced annoyance in one study, possibly due to impeded restoration,[16] but to date, there are no data meaningfully comparing ethnicity or national groups (but see Pedersen et al.[40]). The general public view wind turbines as necessary but ugly,[14] and it is possible that the visual impact of a windfarm can interact with noise level to cause moderate annoyance. This amplification of annoyance is possibly due to a violation of the landscape soundscape continuum constructed by those who choose to live in areas that later contain windfarms,[41] or alternatively, multisensory engagement may enhance detection and identification of wind turbine noise.[42] The degree of influence of the visual aspects of windfarms has yet to be determined, with laboratory studies suggesting that it is wind turbine noise and not the visual impact that underlies the annoyance response,[41] while epidemiological studies suggest that the visual effects are nontrivial. [40]

TABLE 2 Non-Acoustical Factors Influencing the Degree of Annoyance to Noise

- Perceived predictability of the noise level changing
- Perceived control, either by the individual or others
- Trust and recognition of those managing the noise source
- Voice, the extent to which concerns are listened to
- General attitudes, fear of accidents, and awareness of benefits
- Personal benefits, how one benefits from the noise source
- Compensation, how one is compensated due to noise exposure
- Noise sensitivity
- Home ownership, concern about plummeting house values
- Accessibility to information relating to the noise source

Source: Flindell and Stallen.[39]

Wind Turbine Noise and Sleep

The deleterious effects of noise on sleep and the consequences of sleep loss are well documented and are a major concern for governments.[43] In comparison with road, rail, and aircraft noise, there is little research on the effects of wind turbine noise on sleep. However, there is no doubt that wind turbine noise can and does disturb the sleep of those living nearby. Sleep disruption is the predominant symptom in the thousands of anecdotal cases reported in the press and on the Internet and is confirmed by more structured surveys.[25] The quantity, consistency, and ubiquity of complaints has been taken as *prima facie* epidemiological evidence of a causal link between wind turbine noise, sleep disruption, and ill health.[44]

Early investigations into wind turbine noise and sleep are difficult to interpret as researchers used imprecise outcome measures, generally relying on recalled sleep disturbances such as difficulty in initiating or returning to sleep, which tends to underestimate the magnitude of the noise impact and its consequences.[45] One of the earliest studies ($n = 128$) reported that approximately 16% of respondents living at calculated outdoor turbine noise exposures exceeding 35 dB L_{Aeq} stated that wind turbine noise disturbed their sleep.[7] A New Zealand study of 604 households within 3.5 km of a windfarm found that 42 reported occasional and 26 frequent sleep disturbance.[46] The largest wind turbine noise study to date, "Project WINDFARM- perception,"[8] concluded that turbine noise was more of an annoyance at night and that interrupted sleep and difficulty in returning to sleep increased with both indoor and outdoor calculated noise levels. Even at the lowest noise levels, 20% of 725 respondents reported disturbed sleep at least one night per month. In a meta-analysis[40,47] of three European datasets ($n = 1764$),[7,8,16] there was a clear increase in levels of sleep disturbance with dB L_{Aeq} in two of the three studies. In one study, an increment in self-report sleep disturbance occurred between 35 and 40 dBA, while in the other, it occurred between 40 and 45 dBA.

More recent research into wind turbine noise and sleep includes two studies reported by Nissenbaum, Aramini, and Hanning.[48] In the first, a pilot study, a structured questionnaire was administered to 22 subjects living 370 to 1100 m from 28 1.5 mW turbines and a control group ($n = 28$) living at least 4.5 km from the nearest turbine. The study group had clinically and statistically worse sleep disturbance, headache, vestibular symptoms, and psychiatric symptomatology. The second study, using validated questionnaires, administered the Pittsburgh Sleep Quality Index (PSQI), Epworth Sleepiness Score (ESS), and Short- form health survey (SF36) to 79 subjects living between 375 and 6600 m from two windfarms. Those living within 375–1400 m reported worse sleep, were sleepier, and had worse SF36 mental summary scores than those between 3 and 6.6 km from a turbine. Psychiatric symptom scores (irritability, stress, anger, hopelessness, and anxiety) were significantly greater, as was a composite mental health score. They were also more likely to report headaches, nausea (31.6% vs. 12.2%), and a willingness to move away. Modeled dose–response curves of both sleep and health scores against distance from nearest turbine (Figs. 6–8) were significantly related after controlling for gender, age, and household clustering. There was a sharp increase in effects between 1 and 2 km. This study is the first to use appropriate sleep outcome measures[45] and to use a control group. While the sample size is modest ($n = 78$), it is convincing evidence that wind turbine noise adversely affects sleep and health for those living within 1.5 km of turbines.

Mechanisms explaining the effects of wind turbine noise on sleep have been considered, but would benefit from further empirical support.[45] Noise of any description can interfere with sleep by preventing the onset of sleep either at sleep initiation or at the return to sleep after a spontaneous or induced awakening. The amplitude, character, and associations of the noise are all important as is the noise sensitivity of the individual and the psychological response to the noise. In this respect, wind turbine noise seems to be particularly annoying, possessing an impulsive nature with short bursts of low-frequency sound, making it audible 10–15 dBA below background level.[38,49] Nocturnal atmospheric stability ensures that wind turbine noise is maintained while ground level ambient noise diminishes. Indoor noise levels for most noise sources can be reduced by closing windows; however, the low-frequency

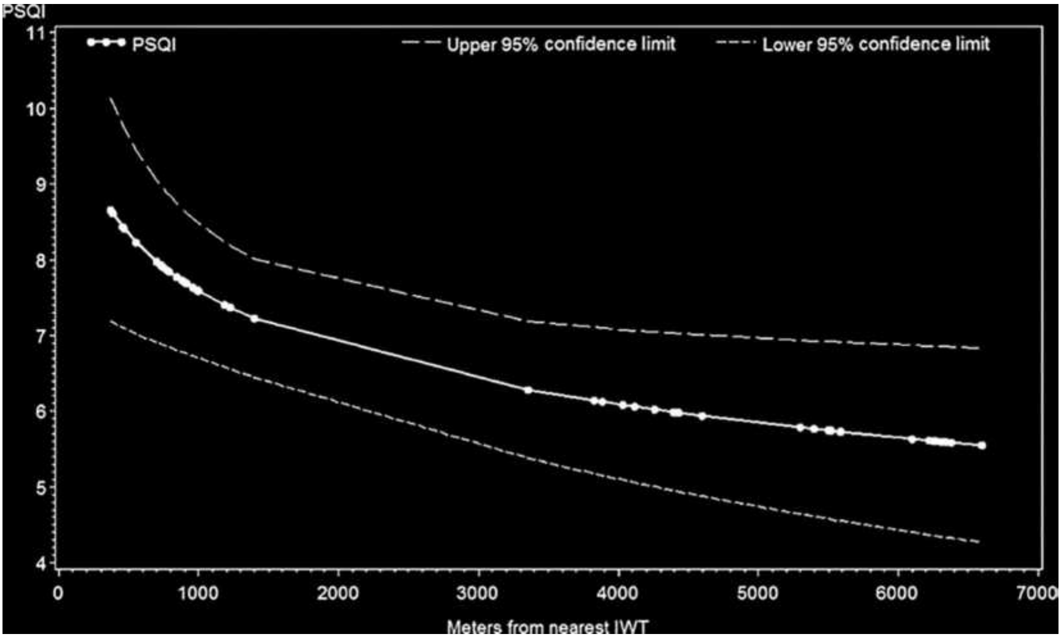


FIGURE 6 Mean Pittsburgh Sleep Quality Index (PSQI) scores as a function of setback distance. The dashed lines are 95% confidence intervals.
Source: Nissenbaum, Aramini, and Hanning.[48]

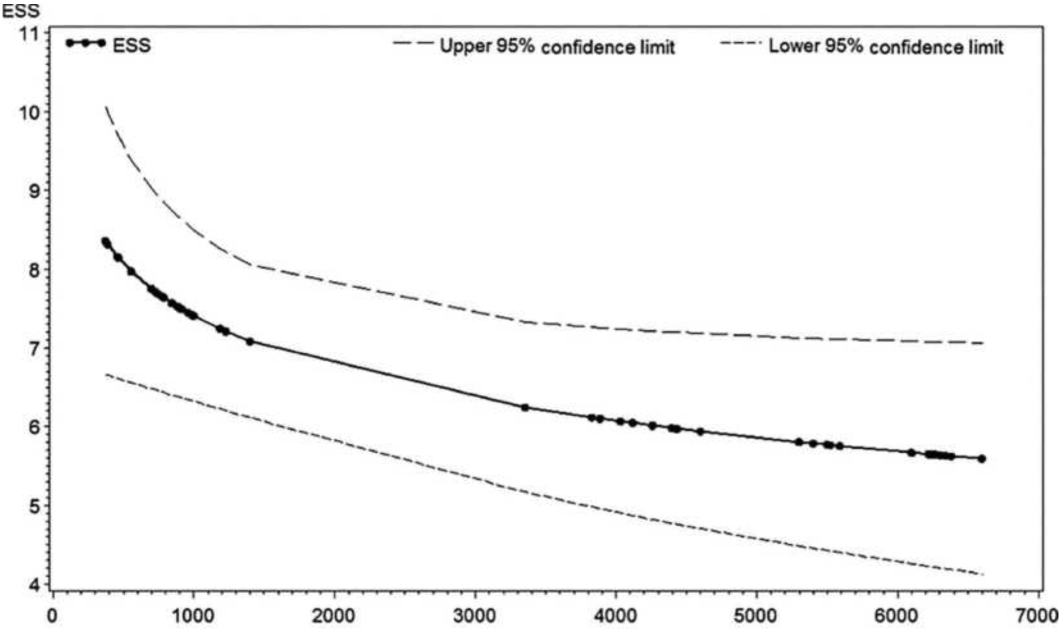


FIGURE 7 Mean Epworth Sleepiness Scale (ESS) scores as a function of setback distance. The dashed lines are 95% confidence intervals.
Source: Nissenbaum, Aramini, and Hanning.[48]

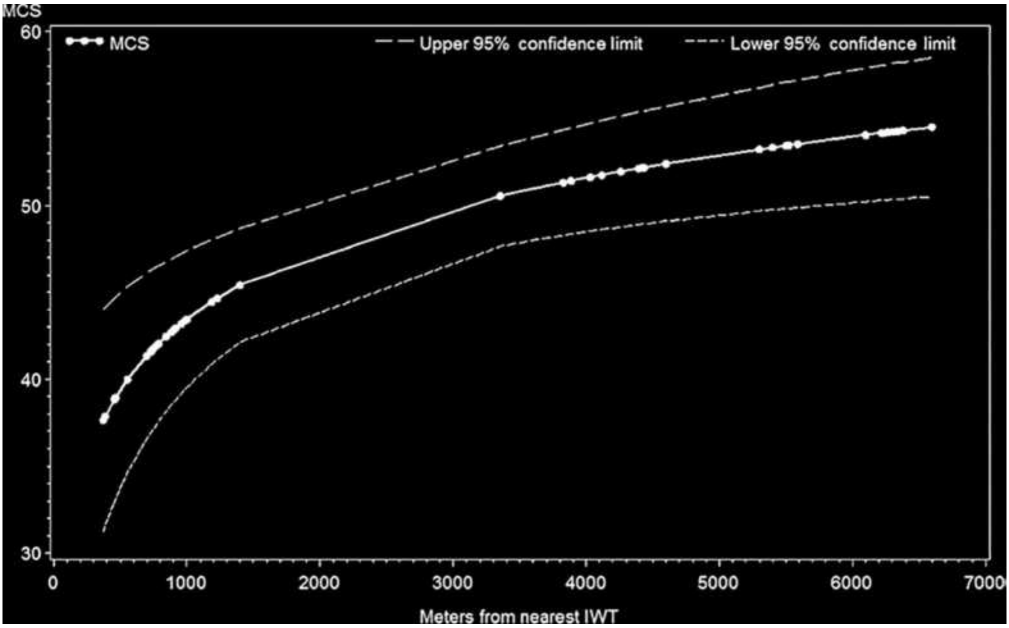


FIGURE 8 Mean SF36 mental component score (MCS) as a function of setback distance. The dashed lines are 95% confidence intervals.
Source: Nissenbaum, Aramini, and Hanning.[48]

content of wind turbine noise means that it may be more audible indoors than outdoors. Additionally, during warmer months, windows are more likely to stay open to control thermal parameters, whence the inability to control or modify wind turbine noise will contribute to the annoyance and, presumably, the effect on sleep onset.[16]

Noise may also cause awakenings and arousals. Arousal is a brief lightening of sleep that is not recalled. Sleep becomes fragmented and, if enough arousals occur, induces the same consequences as reduction of total sleep time. Awakenings are arousals of sufficient degree for wakefulness to be reached and long enough (greater than 10 sec) to be recalled. Arousals are more likely than awakenings, and thus, relying on reported awakenings underestimates the magnitude of the noise effects. The likelihood of an arousal depends upon the volume, character, and duration of the noise as well as the sleep stage and individual propensity (i.e., noise sensitivity). In an investigation into hospital noise, dose–response curves were created for different noises in different sleep stages.[50] Noises with characteristics designed to alert (e.g., telephone, alarms) were more likely to arouse. These noises tend to be impulsive in character, as does wind turbine noise. Noises that were classified as continuous broadband noises (e.g., traffic noise) were less likely to arouse. Another study[51] has shown that subjects with fewer sleep spindles (electrophysiological markers characteristic of stage II sleep) are more easily aroused by noise (Figure 9). Sleep spindles are taken as a marker of sleep stability and may provide a physiological marker of sleep quality.

To date, there are no electrophysiological studies of wind turbine noise on sleep. However, it is reasonable to expect that, in common with road, rail, and aircraft noise, it will induce arousals, fragmenting sleep, as well as preventing the onset of and return to sleep. The sleep measures used in the study by Nissenbaum, Aramini, and Hanning[48] (i.e., ESS and PSQI) are average scores, determining sleepiness and sleep quality, respectively, over a period of weeks. Thus, occasional sleep disturbance would not alter scores as the sleep loss would have been compensated quickly over one or two nights. The study results imply strongly that sleep was being disturbed to some degree on sufficient nights to prevent compensation occurring, thus leading to persistent daytime symptoms.

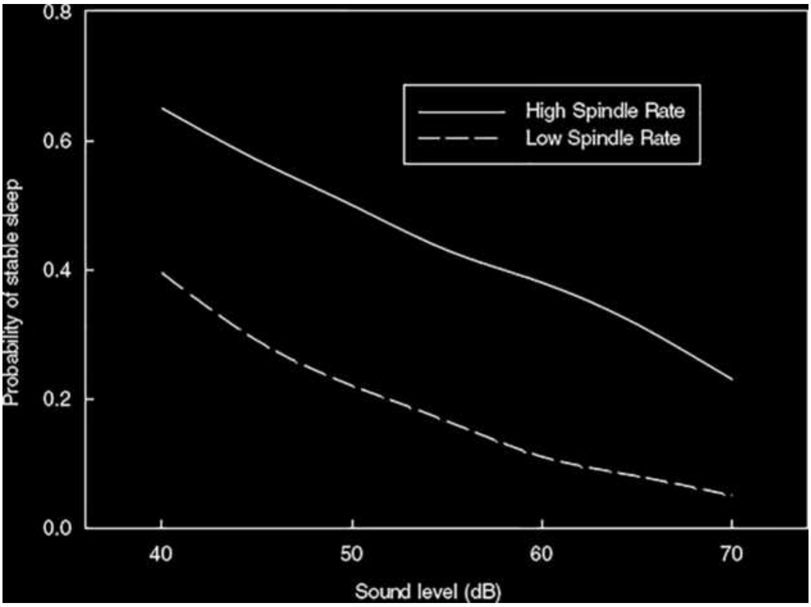


FIGURE 9 Sleep stability as a function of sound level for noise-resistant (high-spindle) and noise-sensitive (low-spindle) groupings.
Source: Estimated from Dang-Vu et al.[51]

Wind Turbine Syndrome

Wind turbine syndrome refers to a cluster of symptoms, which Pierpont,[24] who coined the phrase, claims are associated with exposure to wind turbine noise. Using direct clinical case studies, Pierpont describes the following symptoms to be characteristic of many individuals residing in close vicinity of wind turbines: insomnia, headaches, dizziness, unsteadiness, nausea, exhaustion, anxiety, anger, irritability, depression, memory loss, eye problems, problems with concentration and learning, and tinnitus. Pierpont hypothesizes that wind turbines may affect the vestibular system, that part of the inner ear that plays an important role in the maintenance of balance and stable visual perception. Wind turbines may compromise this system in two ways: first, by the visual disturbance of the moving blades and shadows (i.e., the flicker), and second, by direct vibration of the vestibular system. Such a model would explain why some residents in the close proximity of wind turbines (i.e., less than a kilometer) complain of vertigo, dizziness with nausea, and migraines. Wind turbine syndrome awaits further validation from the medical and scientific establishments, specifically the confirmation of a cause-and-effect relationship between wind turbine noise and vestibular function.

Wind Turbine Noise and Low-Frequency/Infrasound Components

Recent enquiry has focused on the impacts of low-frequency (20–200 Hz) and infrasonic frequencies (typically taken as below 20 Hz) being emitted by wind turbines. Infrasound is characterized by fluctuating pressure sensations at the eardrum, is atonal and countable, and is of a level proportional to wind speed.[21] Low-frequency acoustic waves emitted by wind turbines may be amplified by ground reflection and originate from varying lift forces as the rotors travel through spaces differing in wind speed and density.[21] Compared with medium (i.e., 250 to 4000 Hz) and high frequencies (above 4000 Hz), low-frequency energy decays slowly with distance, is less attenuated by conventionally designed structures, causes certain building materials to vibrate, and can sometimes resonate within rooms and undergo amplification. The effect of air absorption must also be taken into account, in which higher frequencies are attenuated at a greater rate as a function of distance, resulting in a shifting of the spectrum toward lower frequencies. The relationship between low-frequency wind turbine noise and building type creates an interesting proposition in which the low-frequency sound may be louder inside a dwelling than out,[21,52] and the assumption that walls and windows attenuate sound by 15 dB may not be applicable to frequencies below 200 Hz.

Research has shown that low-frequency noise increases cortisol levels in those who are sensitive to noise[12] and disturbs rest and sleep at levels below noise otherwise free from lower-frequency components.[31] Low-frequency noise and infrasound are known disturbers of sleep; however, the contribution, if any, of the low-frequency noise emissions of wind turbines to the sleep disturbances they induce remains to be scientifically determined. Beyond infrasound, the phenomenon of vibroacoustic disease is worthy of note. Humans chronically exposed to infrasound may exhibit elevated cortisol levels and generalized cell damage: a condition known as vibroacoustic disease.[53] A number of human and animal models explaining how infrasound can lead to cardiovascular and respiratory disease have been proposed[54] and applied to wind turbine noise.[55] The phenomenon of vibroacoustic disease is supported by correlational evidence coupled with a thoroughly detailed mechanism. However, further research is required to establish the veracity of this approach to human health within and beyond the wind turbine context.

Mitigation

There are multiple ways in which to reduce the impacts of audible and inaudible wind turbine noise. The first, and often the most effective, method is to control audible noise at the sound source. Thus, mechanical solutions invite technologies designed to attenuate wind turbine noise or to shift its spectral character in order to eliminate salient tonal characteristics. To safeguard health is more difficult, however, because wind turbine noise is largely aerodynamic in origin,[7] and it is not possible to obtain solutions that completely attenuate the noise at its source. Having minimized the noise through the implementation of technology, other approaches are often required, normally involving the application of noise standards to limit exposure levels or the determination of “safe” setback distances to mitigate noise impact. Still other approaches involve the positioning of wind turbines around preexisting noise generators,[15] in remote areas away from human habitations, or using social processes to determine wind turbine location.[27,56]

Regulating Permissible Noise Level

Permissible or safe exposure levels are often set in national noise standards, which may or may not be specific to wind turbine noise. These standards may serve one of two purposes, or sometimes both, with noise compliance guidelines naturally emerging from the two. The first purpose relates to methodologies for the physical quantification of the noise. This may involve standardized procedures for measuring noise from preexisting windfarms or detailing accepted mathematical models affording noise predictions of a planned windfarm. The second purpose is to determine what exposure levels can be considered safe and to clearly state criteria to this effect. However, there are a number of flaws inherent in wind turbine noise standards, including the metrics used to represent the noise, oversimplified modeling approaches that yield unrealistically low predictions of noise levels representing “best case” conditions,[5] or stimulus-oriented approaches that fail to account for human factors.[3,57]

There exists, in respect to levels-based noise standards, disagreement as to the relevance of physical measures such as dBA to human response,[58] not only for windfarm noise[14] but also for traffic and aviation noise. Of the few parametric studies that have been published,[7,8] only marginal dose–response relationships between wind turbine noise intensity and health measures have emerged. For example, Pedersen[22] noted that stress was not related to wind turbine noise level but rather noise annoyance.

Persson Waye and Öhtrsöm[12] reported that annoyance ratings varied for five distinct recordings of wind turbine noise, even though all five had equivalent noise levels. Others note that both laboratory and field studies have consistently found that the equivalent dBA measure fails to account for the relationship between wind turbine noise and annoyance.[14]

To some degree, then, it must be accepted that there is an uncoupling between wind turbine noise level and human response. A hitherto rarely measured characteristic of wind turbine noise is amplitude modulation, whereby noise levels fluctuate periodically as a function of blade passing frequency. Lee et al.[34] recommend that standardized metrics based on the modulation depth spectrum be developed and used in conjunction with sound levels. Other approaches to measuring amplitude modulation have existed for some time[4,59] but have yet to be seriously applied to the wind turbine noise context. However, the inability to account for amplitude modulation arises primarily due to the time-averaged dBA levels applied by noise standards, and arguably, smaller sampling epochs of around 100 msec should be adopted as best practice in order to record the amplitude modulation inherent in turbine noise.[60,61] The New Zealand Standard[62] applies a penalty for amplitude modulation, but does not describe an objective assessment. Furthermore, using aggregated metrics that average noise level over long periods underestimates the effect of peak levels and crest factors, important when considering sleep disturbance.

For the most part, the acceptable noise limits recommended by noise standards are derived from WHO guide-lines.[31,63,64] However, as Figure 10 demonstrates, using recommended noise levels from guidelines based on transport data risks exposing the population to unacceptable levels of noise. It follows that the Ldn (the “day–night” level in the United States) or Lden (the “day–evening– night” level in Europe) measures, derived from the measured LAeq sound level can be used in a wind farm context, but with caution.[65] Inspection of Figure 10 suggests that, relative to transport guidelines, at least a 10 dBA penalty should be placed on wind turbine noise. The differences in annoyance ratings between wind turbine noise and transport noise maybe accounted for by amplitude modulation, the typical location

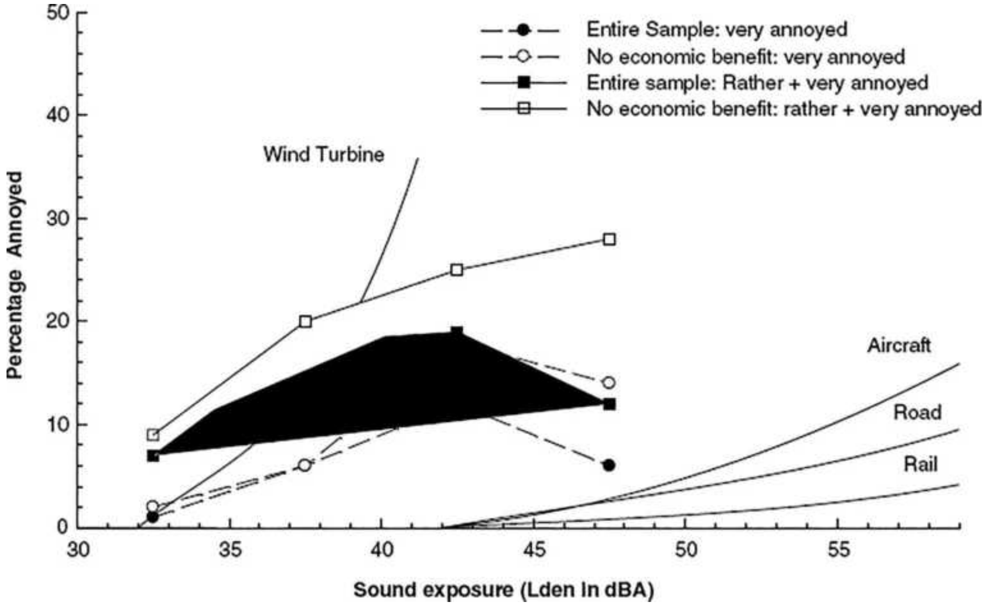


FIGURE 10 Annoyance plotted as a function of noise level for four theoretical models (rail, road, and air parameters: Miedema and Oudshoorm;[66] wind turbine parameters: Pedersen et al.;[7] and four sets of data obtained from Tables 7.24–7.26 of van den Berg et al.[8]). For the data, closed symbols are for the entire sample, while open symbols are for those who identified that they had no economic interest. Circles represent the percentage of “very annoyed” responses, while squares represent the sum of “very annoyed” and “rather annoyed” responses.

of windfarms (e.g., rural areas), or the over-representation of noise-sensitive individuals. A recent meta-analysis of three epidemiological studies revealed a consistent trend in wind turbine noise exposure and both annoyance and sleep disruption.[22] On the basis of her analysis, Pederson recommends that outdoor levels should not exceed 40 dBA, though this level could be more-or-less depending on situational factors, that is, ambient noise levels or the building's construction materials. When noise is continuous, the WHO[31] stipulates an indoor limit of 30 dBA, though for noises containing lower frequencies (e.g., wind turbine noise), a lower limit still is recommended. Thus, careful examination of the lower end of the frequency spectrum is important when judging appropriate exposure to wind turbine noise, and the use of dBC or spectral analysis in one-third octave bands or narrow bands is necessary.

In the comparison of global wind turbine noise level standards, there exist two chief methodologies, namely, sound levels not to be exceeded (usually in dBA) or a not- to-be-exceeded limit derived from the sum of the preconstruction ambient limit and a constant (e.g., L_{A90}+10 dBA). Critique of both these approaches can be found in Thorne.[3] The fact that noise limits differ between, and even within, a country is testament to the impoverished research database guiding their development or the political sensitivities around wind turbine placement. Examples of noise limits are presented in Table 3, and the variability in guidelines is evident. Based on the authors' collective experience, an interim guideline, providing a conservative noise limit capable of protecting the health of the public and susceptible individuals, would be a sound level of L_{Aeq} 35 dBA outside the residence and below the individual's threshold of hearing inside a residence. More specific guidelines are presented in Appendix A of this document.

Regulating Setback Distances

A setback distance is defined as the minimum distance between a dwelling and the closest wind turbine required to protect the health of the inhabitants. One difficulty is whether such setback distances can be standardized, as they will differ depending on a number of factors, including turbine type, terrain, and climate. Lee et al.[34] report that the perception of amplitude-modulated noise decreases with distances beyond a kilometer, though others claim that amplitude-modulated turbine noise can be heard up to 4km away from the source.[67] Setback distances maybe based on noise level, which, as

TABLE 3 A Comparison of Wind Turbine Noise Guidelines Taken from Nine Countries

Country	State	Limit (dBA)	Background Plus Constant
Australia	Victoria	LA90 35 or 40	LA90+5 dBA
	South Australia	LAeq 35 or 40	LA90+5 dBA
Australia	Queensland	LAeq 30 indoors	Health and well-being criteria
Canada	Ontario	LAeq 40 to 51	
Denmark		40	
France			Day: LA90+5 dBA Night: LA90+3 dBA
Netherlands		40	
New Zealand		LA90 35, 40	LA90+5 dBA
United Kingdom		Day: 40 Night: 43	LA90+5 dBA
United States	Illinois	Day: 50 Night: 46	
	Michigan	55	
	Oregon	35	

discussed in the preceding section, maybe an invalid approach. Instead, a better approach may be to link setbacks to turbine type. M0ller and Pedersen,[21] investigating the detection and annoyance of lower-frequency sound emitted from wind turbines, suggest that, for flat terrain, the minimum setback distance for modern turbines (2 to 3.6 MW) should be between 600 and 1200 metres. Other approaches rely on the establishment of dose–response curves relating a health outcome variable (e.g., annoyance or disturbed sleep) and distance (e.g., Figure 6). Medical professionals have proposed setback distances of 2.4 km[23,24] or 1.5 km.[45] Other research recommends a minimum of 2 km if wind turbines are sited in rough terrain.[3,20]

Conclusion

Windfarms have significant potential for sleep disruption and annoyance due to the intermittent nature and amplitude modulation of their sound emissions, even though exposure may be of low amplitude. The interactions between ambient levels, amplitude modulation, and the tonal character of windfarm noise overlaid within a soundscape are complex and difficult to measure and assess in terms of health and individual amenity. Additionally, currently employed sound level measurement and prediction approaches for complex noise sources of this nature are only partially relevant to environmental risk assessment. Aside from acoustic parameters, other factors such as noise sensitivity or amenity expectations may also predict the human response to wind turbine noise. Unfortunately then, for policymakers, there appears to be no proportional relationship between wind turbine noise levels and health, as these outcome factors will be influenced by characteristics associated with both the noise and the listener.[39]

As a relatively new source of intrusive noise, there is little research to draw upon when judging if a proposed windfarm constitutes a health threat to the exposed public. A liberal approach to assessing health impact will involve the application of previous knowledge obtained from other noise sources (e.g., road, aviation). A conservative approach, consistent with the precautionary principle, will consider wind turbine noise more potent than these other harmful noise sources. Thus, at this time, a constellation of acoustic and social metrics should be taken at preexisting wind farms in order to assess potential threat. Peak and crest noise levels, level metrics assessing low-frequency contributions (e.g., dBC), and amplitude modulation indices constitute the acoustic measures of importance. It should also be remembered that predicted levels derived from computer models represent estimates and not precise values, are constrained by numerous assumptions, contain substantial uncertainty, and as such should not constitute the sole criteria for wind turbine positioning. What form the social measures will take is yet to be elucidated, but research suggests that noise sensitivity[67] and procedural fairness[27] are the best approaches to minimize the health impacts and facilitate social acceptance of windfarms.

Acknowledgments

The authors thank Rick James of E-Coustic Solutions for sharing his acoustical expertise on the matter of wind turbine noise guidelines.

Appendix A

'Proposed Wind Turbine Siting Sound Limits', a revision by Thorne, R, of the Kamperman James criteria (2008) to include updates to ISO 1996-2 and U.K. Court of Appeal (Hulme re: Den Brook).

1. Audible Sound Limit
    a. No wind turbine or group of turbines shall be located so as to cause an exceedance of the preconstruction/operation background sound levels by more than 5 dBA. The background sound levels shall be the L_{A90} sound descriptor measured during a pre-construction noise study during the quietest time of evening or night. All data recording shall be a series of

contiguous ten (10) minute measurements. L_{A90} results are valid when L_{A10} results are no more than 15 dBA above L_{A90} for the same time period. Noise sensitive sites are to be selected based on wind development's predicted worst-case sound emissions in L_{Aeq} and L_{Ceq} which are to be provided by the developer.

- b. Test sites are to be located along the property line(s) of the receiving non-participating property(s).
- c. A 5 dB penalty is applied for tones as defined in IEC 61400-11 at the turbine and ISO1996-2 at any affected residence.
- d. A 5 dB penalty is applied for amplitude modulation as defined following. When noise from the wind farm has perceptible or audible characteristics that are perceived by the complainant as being cause for complaint, or greater than expected, the measured sound level of the source shall have a 5 dB penalty added. Audible characteristics include tonal character measured as amplitude or frequency modulation (or both); and tonality (where the tonal character/tonality of noise is described as noise with perceptible and definite pitch or tone). Amplitude modulation is the modulation of the level of broadband noise emitted by a turbine at blade passing frequency. Amplitude modulation will be deemed greater than expected if the following characteristics apply:
  - i. A change in the measured L_{Aeq}, 125 ms turbine noise level of more than 3 dB (represented as a rise and fall in sound energy levels each of more than 3 dB) occurring within a 2 second period.
  - ii. The change identified in (i) above shall not occur less than 5 times in any one minute period provided the L_{Aeq}, 1 minute turbine sound energy for that minute is not below 28 dB.
  - iii. The changes identified in (i) and (ii) above shall not occur for fewer than 6 minutes in any hour.

Noise emissions are measured outside a complainant's dwelling and shall be measured not further than 35 metres from the relevant building, and not closer than within 3.5 metres of any reflective building or surface, or within 1.2 metres of the ground.

2. Low Frequency Sound Limit
   - a. The L_{Ceq} and L_{C90} sound levels from the wind turbine at the receiving property shall not exceed the lower of either:
     - i. L_{Ceq} -L_{A90} greater than 20 dB outside any occupied structure, or
     - ii. A maximum not-to-exceed sound level of 50 dBC measured as the background sound level (L_{C90}) from the wind turbines without other ambient sounds for properties located at one mile or more from state highways or other major roads or measured as the background sound level (L_{C90}) for properties closer than one mile.
     - iii. These limits shall be assessed using the same night-time and wind/weather conditions required in 1(a). Turbine operating sound emissions (L_{Aeq} and L_{Ceq}) shall represent worst case sound emissions for stable night-time conditions with low winds at ground level and winds sufficient for full operating capacity at the hub.
3. General Clause
   - a. Sound levels from the activity of any wind turbine or combination of turbines shall not exceed L_{Aeq} 35 dB within 100 feet of any noise sensitive premises.
   - b. The monitoring shall include all the sound levels as required by these noise conditions and shall include monitoring for the characteristics described in Annex A of IEC 61400-11 including infrasound, low-frequency noise, impulsivity, low-frequency modulation of broad-band or tonal noise, and other audible characteristics. Wind speed and wind direction shall be measured at the same location as the noise monitoring location.
4. Requirements
   - a. All instruments must meet ANSI or IEC Class 1 integrating sound level meter performance specifications.
   - b. Procedures must meet ANSI S12.9, IEC61400-11 and ISO1996-2

   c. Procedures should meet ANSI, IEC and ISO standards applicable to the measurement of sound or its characteristics.
   d. Measurements must be made when ground level winds are 2m/s (4.5 mph) or less. Wind shear in the evening and night often results in low ground level wind speed and nominal operating wind speeds at wind turbine hub heights.
   e. IEC 61400-11 procedures are not suitable for enforcement of these requirements except for the presence of tones near the turbine.

5. Definitions

   ANSI S12.9 Quantities and Procedures for Description and Measurement of Environmental Sound, Parts 1 to 6.

   IEC 61400-11 Wind turbine generator systems—Part 11: Acoustic noise measurement techniques.

   ISO 1996-2 Acoustics—Description, measurement and assessment of environmental noise—Part 2: Determination of environmental noise levels.

   L_{A90}, L_{A10} Statistical measures calculated under ANSI S12.9.

   L_{Aeq}, L_{Ceq} Time average levels calculated under ANSI S12.9 or ISO 1996-2.

   Noise sensitive premises includes a residence, hotel, hostel or residential accommodation premises of any type.

6. References

   ANSI S12.9–2008, Quantities and Procedures for Description and Measurement of Environmental Sound—Part 6: Methods for Estimation of Awakenings Associated with Outdoor Noise Events Heard in Homes, 2008.

   Hulme V Secretary of State (2011). Hulme v Secretary of State for Communities and Local Government and Anor. Approved Judgment in the England and Wales Court of Appeal (Civil Division) on Appeal from the High Court of Justice, Case No: C1/2010/2166/ QBACF. Neutral Citation Number [2011] EWCA Civ 638; Available from: http://www.bailii.org/ew/cases/EWCA/Civ/2011/636.html.

   IEC 61400-11Wind Turbine Generator Systems—Part 11: Acoustic Noise Measurement Techniques, 2nd ed. (International Technical Commission, Geneva, 2002 plus Amendment 1 2006).

   International Standards Organization, (2007). ISO 1996-2 second edition, Acoustics—description, assessment and measurement of environmental noise- part 2: Determination of environmental noise levels. International Standards Organization.

   Kamperman, G. W.; James, R. R. In: Proceedings of the 30th Annual Conference of ASME's Noise Control and Acoustics Division, 28–31 July 2008, Dearborn, Michigan, USA.

References

1. Warren, C.R.; Lumsden, C.; O'Dowd, S.; Birnie, R.V. 'Green on green': Public perceptions of wind power in Scotland and Ireland. J. Environ. Plann. Manage. 2010, *48* (6), 853–875.
2. Van Den Berg, G.P. Wind-induced noise in a screened microphone. J. Acoust. Soc. Am. 2006, *119* (2), 824–833.
3. Thorne, R. The problems with "noise numbers" for wind farm noise assessment. Bull. Sci., Technol. Soc. 2011, *31,* 262–290.
4. Zwicker, E.; Fastl, H. *Psychoacoustics: Facts and Models*; 2nd Edition, Springer-Verlag, Berlin Heidelberg, New York, 1999.
5. International Electrotechnical Commission. *IEC 61400-11 Wind Turbine Generator Systems—Part 11: Acoustic Noise Measurement Techniques, 2nd Ed.*; International Technical Commission: Geneva, 2002 plus Amendment 1, 2006.
6. International Standards Organization. ISO 9613-2 Acoustics—Attenuation of Sound during Propagation Outdoors— Part 2: General Method of Calculation; International Organization for Standardization: Geneva, 1996.

7. Pedersen, E.; Persson Waye, K. Perception and annoyance due to wind turbine noise—A dose–response relationship. J. Acoust. Soc. Am. 2004, *116,* 3460–3470.
8. Van den Berg, G.P.; Pedersen, E.; Bouma, J.; Bakker, R. *Project WINDFARM perception. Visual and Acoustic Impact of Wind Turbine Farms on Residents.* FP6–2005- Science-and-Society-20. Specific Support Action Project no. 044628. Final report; University of Groningen: Holland, 2008.
9. Van Den Berg, G.P. Effects of the wind profile at night on wind turbine sound. J. Sound Vib. 2004, *277* (4–5), 955–970.
10. Stigwood, M. Large wind turbines—are they too big for ETSU-R-97? *Wind Turbine Noise*; Institute of Acoustics: Bristol, 2009.
11. Van Den Berg, G.P. The beat is getting stronger: The effect of atmospheric stability on low frequency modulated sound of wind turbines. J. Low Freq. Noise Vib. Active Control 2005, *24* (1), 1–24.
12. Persson Waye, K.; Öhtrsöm, E. Psycho-acoustic characters of relevance for annoyance of wind turbine noise. J. Sound Vib. 2002, *250* (1), 65–73.
13. Arlinger, S.; Gustafsson, H.A. How a Broadband Noise with Constant Sound Pressure Level Masks a Broadband with Periodically Varying Sound Pressure Levels; Department of Technical Audiology, Linkoping University: Sweden, 1988.
14. Pedersen, E.; Waye, K.P. Wind turbines—Low level noise sources interfering with restoration? Environ. Res. Lett. 2008, *3* (1), 1–5.
15. Pedersen, E.; van den Berg, F.; Bakker, R.; Bouma, J. Can road traffic mask sound from wind turbines? Response to wind turbine sound at different levels of road traffic sound. Energy Policy 2010, *38* (5), 2520–2527.
16. Pedersen, E.; Persson Waye, K. Wind turbine noise, annoyance and self-reported health and well-being in different living environments. Occup. Environ. Med. 2007, *64* (7), 480–486.
17. Martin, M. (2008). Statement of Evidence in Chief of Mr Murray Martin In the Matter of an Appeal Under Sections 120 and 121 of that Act Between Motorimu Wind Farm Ltd (Appellant) and Palmerston North City Council, Environment Court, ENV-2007-WLG-000098.
18. Lercher, P. Environmental noise and health: An integrated research perspective. Environ. Int. 1996, *22*, 117–129.
19. World Health Organization. Preamble to the Constitution of the World Health Organization as Adopted by the International Health Conference; WHO: New York, 1948.
20. Shepherd, D.; McBride, D.; Welch, D.; Dirks, K. N.; Hill, E. M. Evaluating the impact of wind turbine noise on health-related quality of life. Noise and Health, 2011, *13* (54), 333–339.
21. Møller, H.; Pedersen, C.S. Low-frequency noise from large wind turbines. J. Acoust. Soc. Am. 2011, *129* (6), 3727–3744.
22. Pedersen, E. Health aspects associated with wind turbine noise—Results from three field studies. Noise Control Eng. J. 2011, *59* (1), 47–53.
23. Harry, A. *Wind Turbines, Noise and Health*; 2007, available at http://www.flat-group.co.uk/pdf/wtnoise_health_2007_a_barry.pdf (accessed July 2011).
24. Pierpont N. *Wind Turbine Syndrome: A Report on a Natural Experiment*; K Selected Publications: Santa Fe, New Mexico, 2009.
25. Krogh, C.M.E.; Lorrie Gillis, L.; Kouwen, N.; Aramini, J. WindVOiCe, a self-reporting survey: Adverse health effects, industrial wind turbines, and the need for vigilance. Bull. Sci., Technol. Soc. 2011, *31,* 334–345.
26. Alvarsson, J.J.; Wiens, S.; Nilsson, M.E. Stress recovery during exposure to nature sound and environmental noise. Int. J. Environ. Res. Public Health 2010, *7* (3), 1036–1046.
27. Maris, E.; Stallen, P.J.; Vermunt, R.; Steensma, H. Noise within the social context: Annoyance reduction through fair procedures. J. Acoust. Soc. Am. 2007, *121* (4), 2000–2010.
28. Niemann, H.; Maschke, C. *WHO LARES: Report on Noise Effects and Morbidity*; World Health Organization: Geneva, 2004.

29. European Union. *Noise Policy of the European Union—Year* 2; Luxembourg, 2000, http://www.ec.europa.eu/environment/noise/pdf/noise_expert_network.pdf (accessed August 2011).
30. European Commission. *Relating to the Assessment and Management of Environmental Noise*, European Parliament and Council Directive 2002/49/EC; 2002, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2002:189:0012:0025:EN:PDF (accessed August 2011).
31. Berglund, B.; Lindvall, T.; Schwela, D.H. *Guidelines for Community Noise*; The World Health Organization: Geneva, 1999.
32. Hayes, M.; McKenzie, A. *The measurement of Low Frequency Noise at Three U.K. Windfarms;* Hayes Mckenzie Partnership Ltd, report to the Department of Trade and Industry: London, England, 2006.
33. Berglund, B.; Nilsson, M. Intrusiveness and dominant source identification for combined community noises. In *Fechner Day 1997,* Proceedings of the 13th Annual Meeting of the International Society for Psychophysics, Poznan, Poland, Aug 16–19, 1997; Preis, A., Hornowski, T., Eds.; The International Society for Psychophysics, 1997.
34. Lee, S.; Kim, K.; Choi, W.; Lee, S. Annoyance caused by amplitude modulation of wind turbine noise. Noise Control Eng. J. 2011, *59* (1), 38–46.
35. Shepherd, D. Wind turbine noise and health in the New Zealand context. In: *Sound, Noise, Flicker and the Human Perception of Wind Farm Activity;* Rapley, B., Bakker, H., Eds.; Atkinson and Rapley: Palmerston North, 2010; ISBN 978-0-473-16558-1.
36. Kuwano, S.; Seiichiro, N. Continuous judgment of loudness and annoyance. In *Fechner Day 90,* Proceedings of the Sixth Annual Meeting of the International Society of Psychophysics, Wurzburg, Germany, Aug 18–22, 1990; Muller, F., Ed.; International Society of Psychophysics.
37. Johansson, M.; Laike, T. Intention to respond to local wind turbines: The role of attitudes and visual perception. Wind Energy 2007, *10*, 435–45.
38. Nelson, D. 2007. Perceived loudness of wind turbine noise in the presence of ambient sound. In Proceedings of the Fourth International Meeting on Wind Turbine Noise, Rome, Italy, Apr12–14, 2011; INCE Europe; ISBN: 97888-88942-33–9.
39. Flindell, I.; Stallen, P.J. Non-acoustical factors in environmental noise. Noise Health 1999, *3*, 11–16.
40. Pedersen, E.; van den Berg, F.; Bakker, R.; Bouma, J. Response to noise from modern wind farms in the Netherlands. J. Acoust. Soc. America 2009, *126,* 634–643.
41. Pheasant, R.J.; Fisher, M.N.; Watts, G.R.; Whitaker, D.J.; Horoshenkov, K.V. The importance of auditory–visual interaction in the construction of 'tranquil space.' J. Environ. Psychol. 2010, *30,* 501–509.
42. Calvert, G.A. Crossmodal processing in the human brain: Insights from functional neuroimaging studies. Cereb. Cortex 2001, *11* (12), 1110–1123.
43. World Health Organization. *WHO Technical Meeting on Sleep and Health*; World Health Organization, Regional Office for Europe: Bonn, 2004.
44. Phillips, C.V. Properly interpreting the epidemiologic evidence about the health effects of industrial wind turbines on nearby residents. Bull. Sci., Technol. Soc. 2011, *31,* 303–315.
45. Hanning, C.D.; Nissenbaum, M. Selection of outcome measures in assessing sleep disturbance from wind turbine noise. In Proceedings of the Fourth International Meeting on Wind Turbine Noise, Rome, Italy, Apr 12–4, 2011; INCE Europe; ISBN: 978–88–88942-33–9.
46. Phipps, R. A.; Amati, M.; McCoard, S.; Fisher R.M. (2007). Visual and Noise Effects Reported by Residents Living Close to Manawatu Wind Farms: Preliminary Survey Results. New Zealand Planners Institute Conference, Palmerston North, March 27–30, 2007.
47. Pedersen, E. Effects of wind turbine noise on humans. In Proceedings of the Third International Meeting on Wind Turbine Noise, Aalborg, Denmark, June 17–19, 2009; INCE Europe.
48. Nissenbaum, M.; Aramini, J.; Hanning, C. Adverse health effects of industrial wind turbines: A preliminary report. In Proceedings of 10th International Congress on Noise as a Public Health Problem (ICBEN), London, July 24–28, 2011; Foxwoods, C.T., Ed; International Commission on Biological Effects of Noise.

49. Bolin, K. *Wind Turbine Noise and Natural Sounds-Masking, Propagation and Modelling*; Doctoral Thesis; Royal Institute of Technology: Stockholm, 2009.
50. Solet, J.M.; Buxton, O.M.; Ellenbogen, J.M.; Wang, W.; Carballiera, A. *Evidence-Based Design Meets Evidence- Based Medicine: The Sound Sleep Study*; the Center for Health Design: Concord, California, 2010.
51. Dang-Vu, T.T.; McKinney, S.M.; Buxton, O.M.; Solet, J.M.; Ellenbogen, J.M. Spontaneous brain rhythms predict sleep stability in the face of noise. Curr. Biol. 2010, *20,* R626–7.
52. Casella, S. *Low Frequency Noise Update*; DEFRA Noise Programme, Department of the Environment, Northern Ireland Scottish National Assembly for Wales, 2001; 1–11.
53. Castelo-Branco, N.A.A.; Alves-Pereira, M. The clinical stages of vibroacoustic disease. Aviat., Space Environ. Med. 1999, *70* (3), A32–9.
54. Castelo-Branco, N.A.A.; Alves-Pereira, M. Vibroacoustic disease. Noise Health 2004, *6* (23), 3–20.
55. Alves-Pereira, M.; Castelo-Branco, N.A.A. In-home wind turbine noise is conducive to vibroacoustic disease. In Proceedings of the Second International Meeting on Wind Turbine noise, Lyon, France, Sept 20–21, 2007.
56. Gross, C. Community perspectives of wind energy in Australia: The application of a justice and community fairness framework to increase social acceptance. Energy Policy 2007, *35,* 2727–2736.
57. Shepherd, D.; Billington, R. Mitigating the acoustic impacts of modern technologies: Acoustic, health and psychosocial factors informing wind farm placement. Bull. Sci., Technol. Soc. 2011, doi:10.1177/0270467611417841.
58. Fidell, S. The Schultz curve 25 years later: A research perspective. J. Acoust. Soc. of Am. 2003, *114* (6), 3007–3015.
59. Ando, Y.; Pompoli, R. Factors to be measured of environmental noise and its subjective responses based on the model of auditory-brain system. J. Temporal Des. Archit. Environ. 2002, *2* (1), 2–12.
60. Lundmark, G. Measurement and analysis of swish noise: A new method. In Proceedings of the Fourth International Meeting on Wind Turbine Noise, Rome, Italy, Apr 12–4, 2011; INCE Europe; ISBN: 978–88–88942-33–9.
61. Hulme. Approved Judgment in the Court of Appeal (Civil Division) on Appeal from the High Court of Justice, Case No: C1/2010/2166/QBACF, Neutral Citation Number [2011] EWCA Civ 638, http://www.bailii.org/ew/cases/EWCA/Civ/2011/636.html (accessed August 2011).
62. New Zealand Standards. *NZS6808:2010 Acoustics—Wind farm noise*; Standards New Zealand: Wellington, 2010.
63. World Health Organization, Regional Office for Europe. *Night Noise Guidelines for Europe*; World Health Organization: Geneva, Switzerland, 2009, http://www.euro.who.int/document/e92845.pdf (accessed August 2011).
64. World Health Organization. *Burden of Disease from Environmental Noise*; Bonn: World Health Organization: Geneva, 2011, http://www.euro.who.int/_data/assets/pdf_file/0008/136466/e94888.pdf (accessed August 2011).
65. Pedersen T.H, 2007, The “Genlyd” Noise Annoyance Model Dose-Response Relationships Modelled by Logistic Functions. Delta Acoustics and Electronics on behalf of the Ministry of Science, Technology and Innovation, Denmark.
66. Di Napoli, C. Long distance amplitude modulation of wind turbine noise. In Proceedings of the Fourth International Meeting on Wind Turbine Noise, Rome, Italy, Apr 1–14, 2011; INCE Europe; ISBN: 978–88–88942-33–9.
67. Miedema, H.M.E.; Oudshoorm, C.G.M. Annoyance from transportation noise: Relationship with exposure metrics DNL and DENL. Environ. Health Perspect. 2001, *109* (4), 409–416.

IV

DIA: Diagnostic Tools: Monitoring, Ecological Modeling, Ecological Indicators, and Ecological Services

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

20

Exergy: Analysis

Marc A. Rosen

Introduction

Energy analysis is based on the first law of thermodynamics, which embodies the principle of conservation of energy and is the traditional method used to assess the performance and efficiency of energy systems and processes.

Exergy analysis is a thermodynamic analysis technique for systems and processes that is based on the second law of thermodynamics. Exergy analysis has been increasingly applied over the last several decades, in large part because of its advantages over energy analysis:

- More meaningful efficiencies are evaluated with exergy analysis because exergy efficiencies are always a measure of the approach to the ideal.
- Inefficiencies in a process are better pinpointed with exergy analysis because the types, causes, and locations of the losses are identified and quantified.

In this entry, the role of exergy analysis in the assessment and improvement of energy systems is examined. First, exergy and its use as an analysis technique are briefly described. Second, the ranges of energy systems that have been assessed with exergy analysis are surveyed. Third, several example applications of exergy analysis are presented, ranging from simple devices to large and complex systems.

Exergy

Exergy can be regarded as a measure of the usefulness or quality of energy. Technically, exergy is defined as the maximum amount of work that can be produced by a stream of energy or matter, or from a system, as it is brought into equilibrium with a reference environment. Unlike energy, exergy is consumed during real processes due to irreversibilities and conserved during ideal processes. Exergy and related concepts have been recognized for more than a century.[1]

Exergy analysis is a methodology that uses the first and second laws of thermodynamics for the analysis, design, and improvement of energy and other systems.[2–14] The exergy method is useful for improving the efficiency of energy-resource use, for it quantifies the locations, types, and magnitudes of wastes and losses. In general, more meaningful efficiencies are evaluated with exergy analysis rather than energy analysis because exergy efficiencies are always a measure of the approach to the ideal. Therefore, exergy analysis accurately identifies the margin available to design more efficient energy systems by reducing inefficiencies.

In evaluating exergy, the characteristics of the reference environment must be specified,[2–15] usually by specifying the temperature, pressure, and chemical composition of the reference environment. The results of exergy analyses, consequently, are relative to the specified reference environment, which in most applications is modeled after the actual local environment. The exergy of a system is zero when it is in equilibrium with the reference environment. The tie between exergy and the environment has implications regarding environmental impact.[7,8,16,17]

The theory and the applications of exergy have been described in specialized books, e.g.,[2–8] general thermodynamics texts, e.g.,[9,10], and journal entries, e.g.,[11–14]. Many applications of exergy analysis have been reported in fields ranging from power generation,[18] hydrogen energy,[18] and cogeneration[9,19] to district energy,[19] thermal processes,[20,21] and thermal energy storage[21–23] and on to systems as large as countries[24] and the world.[25]

Exergy and the Reference Environment

Exergy quantities are evaluated with respect to a reference environment. The intensive properties of the reference environment in part determine the exergy of a stream or system. The reference environment is in stable equilibrium, with all parts at rest relative to one another and with no chemical reactions occurring between the environmental components. The reference environment acts as an infinite system and is a sink and source for heat and materials. It experiences only internally reversible processes in which its intensive state remains unaltered (i.e., its temperature T_0, pressure P_0, and the chemical potentials μ_{i00} for each of the i components present remain constant). The exergy of the reference environment is zero. More information on reference-environment models can be found in this encyclopedia in an entry by the present author entitled “Exergy: Environmental Impact Assessment.”

Exergy Balances

Energy and exergy balances can be written for a general process or system.

Since energy is conserved, an energy balance for a system may be written as

$$\text{Energy input} - \text{Energy output} = \text{Energy accumulation} \tag{1}$$

Energy input and output refer, respectively, to energy entering and exiting through system boundaries. Energy accumulation refers to build-up (either positive or negative) of the quantity within the system.

By contrast, an exergy balance can be written as

$$\text{Exergy input} - \text{Exergy output} - \text{Exergy accumulation} = \text{Exergy accumulation} \tag{2}$$

This expression can be obtained by combining the principles of energy conservation and entropy non-conservation, the latter of which states that entropy is created during a process due to irreversibilities. Exergy is consumed due to irreversibilities, with exergy consumption proportional to entropy creation.

Equations 1 and 2 demonstrate an important difference between energy and exergy—energy is conserved while exergy, a measure of energy quality or work potential, can be consumed.

Definitions

It is helpful to define some terms related to exergy for readers. The following are exergy quantities:

Exergy: A general term for the maximum work potential of a system, a stream of matter, or a heat interaction in relation to the reference environment (see definition below) as the datum state; or the maximum amount of shaft work obtainable when a steady stream of matter is brought from its initial state to the dead state (see definition below) by means of processes involving interactions only with the reference environment.

Physical exergy: The maximum amount of shaft work obtainable from a substance when it is brought from its initial state to the environmental state (see definition below) by means of physical processes involving interaction only with the environment.

Chemical exergy: The maximum work obtainable from a substance when it is brought from the environmental state to the dead state by means of processes involving interaction only with the environment.

Thermal exergy: The maximum amount of shaft work obtainable from a given heat interaction using the environment as a thermal energy reservoir.

Exergy consumption: The exergy consumed during a process due to irreversibilities within the system boundaries.

The following terms relate to the reference environment and its state:

Reference environment: An idealization of the natural environment, which is characterized by a perfect state of equilibrium, i.e., absence of any gradients or differences involving pressure, temperature, chemical potential, kinetic energy, and potential energy. The reference environment constitutes a natural reference medium with respect to which the exergy of different systems is evaluated.

Dead state: The state of a system when it is in thermal, mechanical, and chemical equilibrium with a conceptual reference environment, which is characterized by a fixed pressure, temperature, and chemical potential for each of the reference substances in their respective dead states.

Environmental state: The state of a system when it is in thermal and mechanical equilibrium with the reference environment, i.e., at the pressure and temperature of the reference environment.

Reference state: A state with respect to which values of exergy are evaluated. Several reference states are used, including environmental state, dead state, standard environmental state, and standard dead state.

Exergy Analysis

Exergy analysis involves the application of exergy concepts, balances, and efficiencies to evaluate and improve energy and other systems. Many engineers and scientists suggest that devices are best evaluated and improved upon using exergy analysis in addition to or in place of energy analysis.

A journal devoted to exergy matters entitled *The International Journal of Exergy* was established by Inderscience. Some extensive bibliographies have been compiled, including one by Goran Wall (see the website http://exergy.se).

A simple procedure for performing energy and exergy analyses involves the following steps:

- Subdivide the process under consideration into as many sections as desired, depending on the depth of detail and the understanding desired from the analysis.
- Perform conventional mass and energy balances on the process, and determine all basic quantities (e.g., work, heat) and properties (e.g., temperature, pressure).
- Based on the nature of the process, the acceptable degree of analysis complexity and accuracy, and the questions for which answers are sought, select a reference-environment model.

- Evaluate energy and exergy values relative to the selected reference-environment model.
- Perform exergy balances, including the determination of exergy consumptions.
- Select efficiency definitions depending on the measures of merit desired, and evaluate the efficiencies.
- Interpret the results and draw appropriate conclusions and recommendations relating to such issues as design changes and retrofit plant modifications.

Exergy Analysis and Efficiency

Increases in efficiency are subject to two constraints, which are often poorly understood:

- Theoretical limitations, which establish the maximum efficiency theoretically attainable for a process by virtue of the laws of thermodynamics
- Practical limitations, which further limit increases in efficiency.

First, consider practical limitations on efficiency. In practice, the goal when selecting energy sources and utilization processes is not to achieve maximum efficiency, but rather to achieve an optimal tradeoff between efficiency and such factors as economics, sustainability, environmental impact, safety, and societal and political acceptability. This optimum is dependent on many factors controllable by society. Furthermore, these factors can be altered to favor increased efficiency (e.g., governments can offer financial incentives that render high-efficiency technologies economically attractive or provide disincentives for low-efficiency alternatives through special taxes and regulations).

Next, consider theoretical limitations on efficiency, which must be clearly understood to assess the potential for increased efficiency. Lack of clarity on this issue in the past has often led to confusion, in part because energy efficiencies generally are not measures of how nearly the performance of a process or device approaches the theoretical ideal. The consequences of such confusion can be significant. For example, extensive resources have at times been directed towards increasing the energy efficiencies of devices that in reality were efficient and had little potential for improvement. Conversely, devices at other times have not been targeted for improved efficiency even though the difference between the actual and maximum theoretical efficiencies, which represents the potential for improvement, has been large.

The difficulties inherent in energy analysis are also attributed to the fact that it only considers quantities of energy and ignores energy quality, which is continually degraded during real processes. Exergy analysis overcomes many of the problems associated with energy analysis.

Overview of Exergy Analysis Applications

Exergy analysis has been applied to a wide range of processes and systems, including those that are mechanical, thermal, electrical, and chemical. The types of applications of exergy methods that have been reported over the last several decades include

- Electricity generation using both conventional devices such as fossil and nuclear power plants as well as alternative devices such as fuel cells and solar energy systems
- Energy storage systems such as batteries, pumped storages, and thermal energy storages
- Combustion technologies and systems and engines of various types
- Transportation systems for land, air, and water transport
- Heating and cooling systems for building systems and industrial applications
- Cogeneration systems for producing heating and electrical needs simultaneously
- Chemical processes such as sulfuric acid production, distillation, and water desalination, as well as petrochemical processing and synthetic fuels production
- Metallurgical processes such as lead smelting.

Examples of Exergy Analysis Applications

Three examples of differing complexity of applications of exergy analysis are presented:

- An electrical resistance space heater (a simple component)
- A thermal energy storage system (a simple system containing a number of components)
- A coal-fired electrical generating station (a complex system).

Electrical Resistance Space Heater

An electrical resistance space heater converts electricity to heat at a temperature suitable for room comfort and is illustrated in Figure 1a.

The energy efficiency of electric resistance space heating often exceeds 99%, implying that the maximum possible energy efficiency for electric resistance heating is 100%, corresponding to the most efficient device possible.

This understanding is erroneous; however, energy analysis ignores the fact that in this process, high-quality energy (electricity) is used to produce a relatively low-quality product (warm air). Exergy analysis recognizes this difference in energy qualities and indicates the exergy of the heat delivered to the room to be about 5% of the exergy entering the heater. Thus, the exergy efficiency of electric resistance space heating is about 5%.

The exergy results are useful. Since thermodynamically ideal space heating has an exergy efficiency of 100%, the same space heating can in theory be achieved using as little as 5% of the electricity used in conventional electric resistance space heating. In practical terms, one can achieve space heating with a greatly reduced electricity input using an electric heat pump (see Figure 1b), using 15% of the electricity that electric resistance heating would require, for a heat pump with a "coefficient of performance" of 7.

Thermal Storage System

A thermal energy storage system receives thermal energy and holds it until it is required. Thermal storages can store energy at temperatures above or below the environment temperature, and they come in many types (e.g., tanks, aquifers, ponds, caverns).

The evaluation of a thermal energy storage system requires a measure of performance which is rational, meaningful, and practical. The conventional energy storage efficiency is inadequate. A more perceptive basis is needed if the true usefulness of thermal storages is to be assessed and their economic benefit optimized, and exergy efficiencies provide such performance measures.

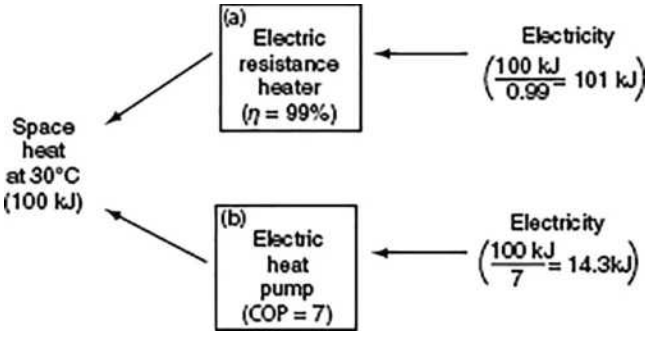


FIGURE 1 Comparison of the quantity of electricity required to provide 100 kJ of space heat using two different heating devices: (a) an electric resistance heater and (b) an electric heat pump. Here, η denotes the energy efficiency and COP the coefficient of performance.

The notion that energy efficiency is an inappropriate measure of thermal storage performance can be illustrated. Consider a perfectly insulated thermal storage containing 1,000 kg of water, initially at 40°C. The ambient temperature is 20°C, and the specific heat of water is taken to be constant at 4.2 kJ/kg K. A quantity of 4,200 kJ of heat is transferred to the storage through a heat exchanger from an external body of 100 kg of water cooling from 100°C to 90°C. This heat addition raises the storage temperature 1.0°C to a value of 41°C. After a period of storage, 4200 kJ of heat is recovered from the storage through a heat exchanger, which delivers it to an external body of 100 kg of water, raising the temperature of that water from 20°C to 30°C. The storage is returned to its initial state at 40°C.

For this storage cycle, the energy efficiency—the ratio of heat recovered from the storage to heat injected—is 4,200kJ/4,200 kJ = 1, or 100%. But the recovered heat is at only 30°C and is of little use, having been degraded even though the storage energy efficiency was 100%. The exergy recovered in this example is 70 kJ and the exergy supplied 856 kJ. Thus, the exergy efficiency, the ratio of the thermal exergy recovered from storage to that injected, is 70/856 = 0.082 or 8.2%, a much more meaningful expression of the achieved performance.

Consequently, a device which appears to be ideal on an energy basis is correctly shown to be far from ideal on an exergy basis, clearly demonstrating the benefits of using exergy analysis for evaluating thermal storage.

Coal-Fired Electrical Generating Station

Energy and exergy analyses are applied to the former Nanticoke coal-fired electrical generating station in Ontario, Canada, which has a net unit electrical output of approximately 500 MWe and is operated by the provincial electrical utility, Ontario Power Generation (formerly Ontario Hydro). This example illustrates how exergy analysis allows process inefficiencies to be better pinpointed than an energy analysis does and how efficiencies are to be more rationally evaluated.

A detailed flow diagram for a single unit of the station is shown in Figure 2. The symbols identifying the streams are described in a–c for material, thermal, and electrical flows, respectively, with corresponding data. Figure 2 has four main sections:

- *Steam Generation.* Eight pulverized-coal-fired natural circulation steam generators each produce 453.6 kg/s steam at 16.89 MPa and 538°C and 411.3 kg/s of reheat steam at 4.00 MPa and 538°C. Air is supplied to the furnace by two 1,080 kW 600-rpm motor-driven forced draft fans. Regenerative air preheaters are used. The flue gas passes through an electrostatic precipitator rated at 99.5% collection efficiency and exits the plant through two multiflued, 198 m high chimneys.
- *Power Production:* The steam passes through a series of turbine generators linked to a transformer. Extraction steam from several points on the turbines preheats feedwater in several low- and high-pressure heat exchangers and one spray-type open deaerating heat exchanger. The low-pressure turbines exhaust to the condenser at 5 kPa. Each station unit has a 3,600-rpm, tandem-compound, impulse-reaction turbine generator containing one single-flow high-pressure cylinder, one double-flow intermediate-pressure cylinder, and two double-flow low-pressure cylinders. Steam exhausted from the high-pressure cylinder is reheated in the combustor.
- *Condensation:* Cooling water from Lake Erie condenses the steam exhausted from the turbines. The cooling-water flow rate is adjusted to achieve a specified cooling-water temperature rise across the condenser.
- *Preheating:* The temperature and pressure of the feedwater are increased in a series of pumps and feedwater-heater heat exchangers.

The reference-environment model used here has a temperature of 15°C (the approximate mean temperature of the lake cooling water), a pressure of 1 atm, and a chemical composition consisting of air saturated with water vapor, and the following condensed phases at 15°C and 1 atm: water (H_2O),

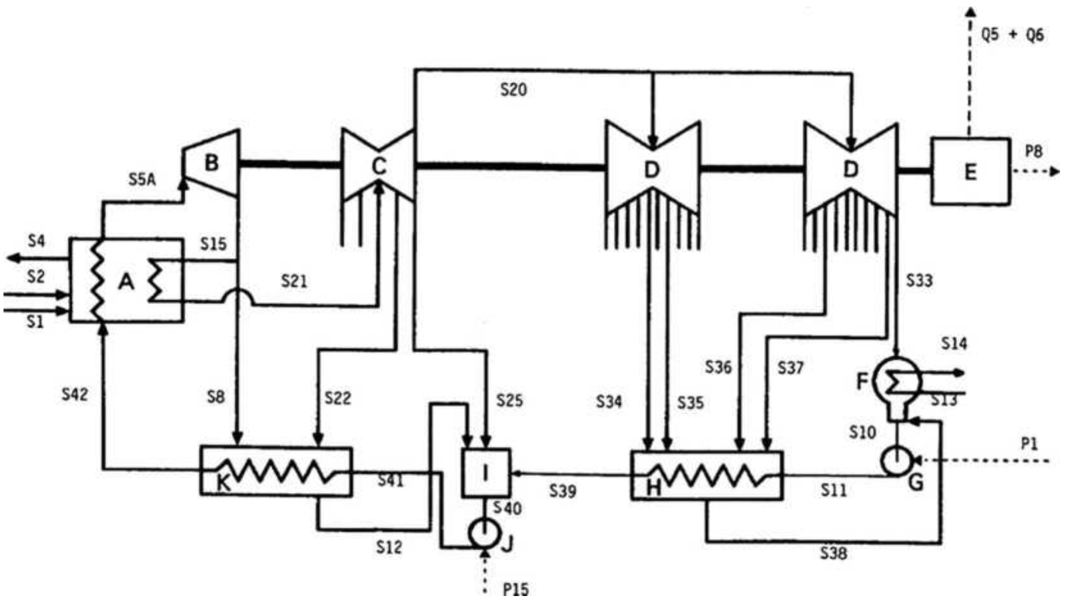


FIGURE 2 A unit of the coal-fired electrical generating station. Lines exiting the turbines represent extraction steam. The station has four main sections: *Steam Generation* (Device A), *Power Production* (B–E), *Condensation* (F), and *Preheating* (G–K). The external inputs for Device A are coal and air, and the output is stack gas and solid waste. The external outputs for Device E are electricity and waste heat. Electricity is input to Device G and Device J, and cooling water enters and exits Device F. A: steam generator and reheater, B: high- pressure turbine, C: intermediate-pressure turbine, D: low-pressure turbines, E: generator and transformer, F: condenser, G: hot well pump, H: low-pressure heat exchangers, I: open deaerating heat exchanger, J: boiler feed pump, and K: high-pressure heat exchangers.

gypsum ($CaSO_4 \bullet 2H_2O$), and limestone ($CaCO_3$). For simplicity, heat losses from external surfaces are assumed to occur at the reference-environment temperature of 15°C.

Energy and exergy values for the streams identified in Figure 2 are summarized in Tables 1a–1c. Exergy-consumption values for the devices are listed, according to process section, in Table 2. Figure 3a and b illustrates the net energy and exergy flows and exergy consumptions for the four main process sections.

Overall energy and exergy efficiencies are evaluated as

$$\text{Energy Efficiency} = (\text{Net energy output with electricity})/(\text{Energy input}) \tag{3}$$

and

$$\text{Exergy Efficiency} = (\text{Net exergy output with electricity})/(\text{Exergy input}) \tag{4}$$

Coal is the only input source of energy or exergy, and the energy and exergy efficiencies are 37% and 36%, respectively. The small difference in the efficiencies is due to the fact that the specific chemical exergy of coal is slightly greater than its energy. Although the station energy and exergy efficiencies are similar, these efficiencies differ markedly for many station sections.

In the *Steam Generation* section, exergy consumptions are substantial, accounting for 659 MW (or 72%) of the 916 MW station exergy loss. Of this 659, 444 MW is consumed with combustion and 215 MW with heat transfer. The energy and exergy efficiencies for the *Steam Generation* section, considering the increase in energy or exergy of the water as the product, are 95% and 49%, respectively.

TABLE 1A Data for Material Flows for a Unit of the Coal-Fired Electrical Generating Station

Stream	Mass Flow Rate (kg/s)[a]	Temperature (°C)	Pressure (N/m^2)	Vapor Frac.[b]	Energy Flow Rate (MW)	Exergy Flow Rate (MW)
S1	41.74	15.00	1.01×10^5	solid	1367.58	1426.73
S2	668.41	15.00	1.01×10^5	1.0	0.00	0.00
S3[c]	710.15	1673.59	1.01×10^5	1.0	1368.00	982.85
S4	710.15	119.44	1.01×10^5	1.0	74.39	62.27
S5A	453.59	538.00	1.62×10^7	1.0	1585.28	718.74
S8	42.84	323.36	3.65×10^6	1.0	135.44	51.81
S10	367.85	35.63	4.50×10^3	0.0	36.52	1.20
S11	367.85	35.73	1.00×10^6	0.0	37.09	1.70
S12	58.82	188.33	1.21×10^6	0.0	50.28	11.11
S13	18,636.00	15.00	1.01×10^5	0.0	0.00	0.00
S14	18,636.00	23.30	1.01×10^5	0.0	745.95	10.54
S15	410.75	323.36	3.65×10^6	1.0	1298.59	496.81
S20	367.85	360.50	1.03×10^6	1.0	1211.05	411.16
S21	410.75	538.00	4.00×10^6	1.0	1494.16	616.42
S22	15.98	423.23	1.72×10^6	1.0	54.54	20.02
S25	26.92	360.50	1.03×10^6	1.0	88.64	30.09
S33	309.62	35.63	4.50×10^3	0.93	774.70	54.07
S34	10.47	253.22	3.79×10^5	1.0	32.31	9.24
S35	23.88	209.93	2.41×10^5	1.0	71.73	18.82
S36	12.72	108.32	6.89×10^4	1.0	35.77	7.12
S37	11.16	60.47	3.45×10^4	1.0	30.40	5.03
S38	58.23	55.56	1.33×10^4	0.0	11.37	0.73
S39	367.85	124.86	1.00×10^6	0.0	195.94	30.41
S40	453.59	165.86	1.00×10^6	0.0	334.86	66.52
S41	453.59	169.28	1.62×10^7	0.0	347.05	77.57
S42	453.59	228.24	1.62×10^7	0.0	486.75	131.93

[a] The composition of all streams is 100% H_2O, except that, on a volume basis, the composition of S1 is 100% carbon, of S2 is 79% N_2 and 21% O_2, and of both S3 and S4 is 79% N_2, 6% O_2, and 15% CO_2.

[b] Vapor fraction is listed as 0.0 for liquids and 1.0 for superheated vapors.

[c] Stream S3 (not shown in Figure 2) represents the hot product gases for adiabatic combustion.

TABLE 1B Data for Principal Thermal Flows for a Unit of the Coal-Fired Electrical Generating Station

Stream	Energy Flow Rate (MW)	Exergy Flow Rate (MW)
Q5	5.34	0.00
Q6	5.29	0.00

TABLE 1C Data for Principal Electrical Flows for a Unit of the Coal-Fired Electrical Generating Station

Stream	Energy (and Exergy) Flow Rate (MW)
P1	0.57
P8	523.68
P15	12.19

TABLE 2 Breakdown of Exergy Consumption Rates for a Unit of the Coal-Fired Electrical Generating Station

Section/Device	Exergy Consumption Rate (MW)	
Steam generation		
Steam generator (including combustor)	659.0	
		659.0
Power production		
High-pressure turbine	26.4	
Intermediate-pressure turbine	22.3	
Low-pressure turbines	59.2	
Generator	5.3	
Transformer	5.3	
		118.5
Condensation		
Condenser	43.1	
		43.1
Preheat		
Low-pressure heat exchangers	10.7	
Deaerating heat exchanger	5.1	
High-pressure heat exchangers	6.4	
Hot well pumps	0.1	
Boiler feed pumps	1.1	
		23.4
Total		844.0

The *Steam Generation* section thus appears significantly more efficient on an energy basis than on an exergy basis. Physically, this discrepancy implies that although 95% of the input energy is transferred to the preheated water, the energy is degraded as it is transferred. Exergy analysis highlights this degradation.

In the condensers, a large quantity of energy enters (775 MW for each unit), of which close to 100% is rejected, while a small quantity of exergy enters (54 MW for each unit), of which about 25% is rejected and 75% is internally consumed. Thus, energy analysis leads to the erroneous conclusion that almost all losses in electricity-generation potential for the station are associated with the heat rejected by the condensers, while exergy analysis demonstrates quantitatively and directly that the condensers are responsible for little of these losses (see Figure 3b). This discrepancy arises because heat is rejected by the condensers at a temperature very near that of the environment.

In the *Power Production* and *Preheating* sections, energy losses are small (less than 10 MW) and exergy losses moderately small (118 MW in *Power Production* and 23 MW in *Preheating*). The exergy losses are almost completely associated with internal consumptions.

In assessing the thermodynamic characteristics of a coal-fired electrical generating station, several illuminating insights into performance are acquired:

- Although energy and exergy efficiencies are similar for the station, energy analysis does not identify the location and cause of process inefficiencies, while exergy analysis does. Energy losses are associated with emissions (mainly heat rejected by condensers), and exergy losses are primarily associated with consumptions (mainly in the combustors).
- Because devices with the largest thermodynamic losses have the largest margins for efficiency improvement, efforts to increase the efficiencies of coal-fired electrical generating stations should focus on the combustors. For instance, technologies capable of producing electricity without

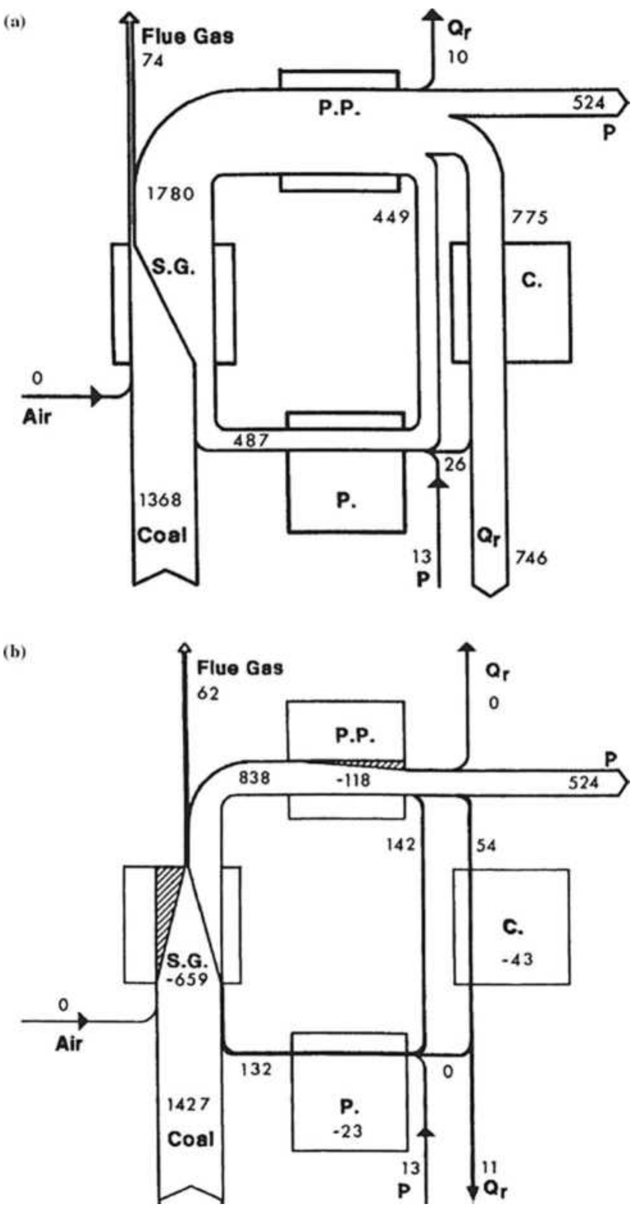


FIGURE 3 (a) Diagram for a coal-fired electrical generating station unit indicating net energy flow rates (MW) for streams. Stream widths are proportional to energy flow rates. Station sections shown are *Steam Generation* (S.G.), *Power Production* (P.P.), *Condensation* (C.), and *Preheating* (P.). Streams shown are electrical power (P), heat input (Q), and heat rejected (Q_r). (b) Diagram for a coal-fired electrical generating station unit indicating net exergy flow rates for streams and consumption rates (negative values) for devices (in MW). Stream widths are proportional to exergy flow rates and shaded regions to exergy consumption rates. Other details are as in (a).

combustion (e.g., fuel cells) or utilizing heat at high temperatures could increase efficiencies significantly. This suggestion is, of course, overly simplistic, as such decisions must also account for other technical and economic factors.

- The use of heat rejected by condensers only increases the exergy efficiencies by a few percent.

Applications beyond Thermodynamics

Exergy concepts can be applied beyond thermodynamics in such fields as environmental impact assessment,[7,8,16,17] economics,[5,17,20,26] and policy.[27,28]

Exergy and Environment

Many suggest that the impact of energy utilization on the environment is best addressed by considering exergy. Although the exergy of an energy form or a substance is a measure of its usefulness, exergy is also a measure of its potential to cause change. The latter point suggests that exergy may be or may provide the basis for an effective measure of the potential of a substance or energy form to impact the environment. The relation between exergy and the environment is discussed in this encyclopedia in an entry entitled "Exergy: Environmental Impact Assessment."

Exergy and Economics

Another area in which applications of exergy are increasing is that of economics. In the analysis and design of energy systems, techniques are often used that combine scientific disciplines like thermodynamics with economics to achieve optimum designs. For energy systems, costs are conventionally based on energy. Many researchers, however, have recommended that costs are better distributed among outputs based on exergy. Methods of performing exergy-based economic analyses have evolved (e.g., thermoeconomics, second-law costing, and exergoeconomics). These analysis techniques recognize that exergy, not energy, is the commodity of value in a system, and assign costs and prices to exergy-related variables. These techniques usually help in appropriately allocating economic resources so as to optimize the design and operation of a system and its economic feasibility and profitability (by obtaining actual costs of products and their appropriate prices).

Conclusion

Exergy analysis provides information that influences design, improvement, and application decisions, and it is likely to be increasingly applied. Exergy also provides insights into the "best" directions for research, where "best" is loosely considered most promising for significant efficiency gains. There are two main reasons for this conclusion:

- Unlike energy losses, exergy losses represent true losses of the potential to generate the desired product from the given driving input. Focusing on exergy losses permits research to aim at reducing the losses that degrade efficiency.
- Unlike energy efficiencies, exergy efficiencies always provide a measure of how closely the operation of a system approaches the ideal. By focusing research on plant sections or processes with the lowest exergy efficiencies, effort is directed to those areas that inherently have the largest margins for efficiency improvement. By focusing on energy efficiencies, on the other hand, research can inadvertently be expended on areas for which little margins for improvement exist, even theoretically.

Exergy analysis results typically suggest that improvement efforts should concentrate more on internal rather than external exergy losses based on thermodynamic considerations, with a higher priority for the processes that have larger exergy losses. Of course, effort should still be devoted to processes having low exergy losses when cost- effective ways to increase efficiency can be identified.

Energy-related decisions should not be based exclusively on the results of energy and exergy analyses even though these results provide useful information to assist in such decision-making. Other factors must also be considered, such as economics, environmental impact, safety, and social and political implications.

Glossary

COP	coefficient of performance
i	component
P	pressure
$\mathbf{P_0}$	reference-environment pressure
T	temperature
$\mathbf{T_0}$	reference-environment temperature
$\boldsymbol{\eta}$	energy efficiency
$\boldsymbol{\mu}$	chemical potential
μ00	reference-environment chemical potential

Acknowledgments

The author is grateful for the support provided for this work by the Natural Sciences and Engineering Research Council of Canada.

References

1. Rezac, P.; Metghalchi, H. A brief note on the historical evolution and present state of exergy analysis. *Int. J. Exergy* **2004**, *1*, 426–437.
2. Kotas, T.J. *The Exergy Method of Thermal Plant Analysis*; Krieger: Malabar, FL, Reprint ed.; 1995.
3. Moran, M.J. *Availability Analysis: A Guide to Efficient Energy Use*; American Society of Mechanical Engineers: New York, Revised ed.; 1989.
4. Brodyanski, V.M.; Sorin, M.V.; Le Goff, P. *The Efficiency of Industrial Processes: Exergy Analysis and Optimization*; Elsevier: London, 1994.
5. Szargut, J.; Morris, D.R.; Steward, F.R. *Exergy Analysis of Thermal, Chemical and Metallurgical Processes*; Hemisphere: New York, 1988.
6. Yantovskii, E.I. *Energy and Exergy Currents (An Introduction to Exergonomics)*; Nova Science Publishers: New York, 1994.
7. Szargut, J. *Exergy Method: Technical and Ecological Applications*; WIT Press: Southampton, 2005.
8. Dincer, I.; Rosen, M.A. *Exergy: Energy, Environment and Sustainable Development*, 2nd ed.; Elsevier: Oxford, 2013.
9. Moran, M.J.; Shapiro, H.N.; Boettner, D.D.; Bailey, M.B. *Fundamentals of Engineering Thermodynamics*, 9th ed.; Wiley: New York, 2018.
10. Sato, N. *Chemical Energy and Exergy: An Introduction to Chemical Thermodynamics for Engineers*; Elsevier: Oxford, 2005.
11. Moran, M.J.; Sciubba, E. Exergy analysis: principles and practice. *J. Eng. Gas Turbines Power* **1994**, *116*, 285–290.
12. Aghbashlo, M.; Rosen, M.A. Consolidating exergoeconomic and exergoenvironmental analyses using the emergy concept for better understanding energy conversion systems. *J. Clean. Prod.* **2018**, *172*, 696–708.
13. Milia, D.; Sciubba, E. Exergy-based lumped simulation of complex systems: an interactive analysis tool. *Energy* **2006**, *31*, 100–111.
14. Frangopoulos, C.A.; von Spakovsky, M.R.; Sciubba, E. A brief review of methods for the design and synthesis optimization of energy systems. *Int. J. Appl. Thermodyn.* **2002**, *5* (4), 151–160.
15. Rosen, M.A.; Dincer, I. Effect of varying dead-state properties on energy and exergy analyses of thermal systems. *Int. J. Therm. Sci.* **2004**, *43* (2), 121–133.
16. Jorgensen, S.E.; Svirezhev, Y.M. *Towards a Thermodynamic Theory for Ecological Systems*; Elsevier: Oxford, 2004.

17. Rosen, M.A. *Environment, Ecology and Exergy: Enhanced Approaches to Environmental and Ecological Management*; Nova Science Publishers: Hauppauge, NY, 2012.
18. Al-Zareer, M.; Dincer, I.; Rosen, M.A. Multi-objective optimization of an integrated gasification combined cycle for hydrogen and electricity production. *Comput. Chem. Eng.* **2018**, *117*, 256–267.
19. Rosen, M.A.; Koohi-Fayegh, S. *Cogeneration and District Energy Systems: Modelling, Analysis and Optimization*; Institution of Engineering and Technology: London, 2016.
20. Mahmoudi, S.M.S.; Salehi, S.; Yari, M.; Rosen, M.A. Exergoeconomic performance comparison and optimization of single-stage absorption heat transformers. *Energies* **2017**, *10*(4), 532.
21. Rad, F.M.; Fung, A.S.; Rosen, M.A. An integrated model for designing a solar community heating system with borehole thermal storage. *Energy Sustainable Dev.* **2017**, *36C*, 6–15.
22. Dincer, I.; Rosen, M.A. Energy and exergy analyses of thermal energy storage systems. In *Thermal Energy Storage: Systems and Applications*, 2nd ed.; Wiley: London, 2011; 233–334.
23. Rezaie, B.; Reddy, B.V.; Rosen, M.A. Thermodynamic analysis and the design of sensible thermal energy storages. *Int. J. Energy Res.* **2017**, *41*(1), 39–48.
24. Rosen, M.A.; Dincer, I. Sectoral energy and exergy modeling Turkey. *ASME J. Energy Resour. Technol.* **1997**, *119*, 200–204.
25. Rosen, M.A. Assessing global resource utilization efficiency in the industrial sector. *Sci. Total Environ.* **2013**, *461–462*, 804–807.
26. Jawad, H.; Jaber, M.Y.; Bonney, M.; Rosen, M.A. Deriving an exergetic economic production quantity model for better sustainability. *Appl. Math. Modell.* **2016**, *40*(11–12), 6026–6039.
27. Rosen, M.A. Benefits of exergy and needs for increased education and public understanding and applications in industry and policy. Part I: Benefits. *Int. J. Exergy* **2006**, *3*(2), 202–218.
28. Rosen, M.A. Benefits of exergy and needs for increased education and public understanding and applications in industry and policy. Part II: Needs. *Int. J. Exergy* **2006**, *3*(2), 219–229.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

V

ENT: Environmental Management Using Environmental Technologies

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

21

Air Pollution: Monitoring

Waldemar Wardencki

Introduction

Concern about air quality is not new. The first reports of air pollution problems appear to have been made by writers in ancient Rome who were aware of its adverse effects on human health. Air pollution and its consequences had originally been considered to be relatively local phenomena associated with urban and industrial centers. Complaints were recorded in the 13th century when coal was first used in London. Now, it has become apparent that pollutants may be transported long distances in the air, causing adverse effects in environments far removed from the source emission. Scientific research, conducted over 200 years, has evidently shown that polluted air has a negative influence on health and, in some cases, may lead to death. The World Health Organization (WHO) appraises that air pollution causes approximately 2 million premature deaths worldwide per year. The levels of pollutants, which have a negative influence on life, are nowadays well defined. Because current thresholds set by national or global air quality guideline values are frequently exceeded, further reductions of emissions are necessary.

The first essential step in controlling and mitigating air pollution is to quantify the emissions of air pollutants. Most countries entail controlling a range of key pollutants at their point of discharge. The most important tool in environmental protection is monitoring. Environmental monitoring is the general term for systematic observations of what is going on in the environment. In the broadest context, environmental monitoring is defined as a system of measurements, evaluations, and forecasts of environmental states, and the collecting, processing, and spreading of information on the environment.

Air pollution and its control are a global issue demanding international cooperation. Monitoring of air pollution is a very important source of data. However, measurement of air pollutant concentrations, in comparison to monitoring of other elements of the environment, is the most difficult. This is related to the dynamics of the atmosphere, making it the main route of pollutant transport between the remaining environmental compartments. Unlike the case of water and soil pollution, environmental pollution is not geographically restricted, as a result of which large human populations can be exposed to it. Another problem is low concentration of air pollutants and their interaction with other gases.

This entry reviews the issues in the field of air pollution monitoring. At the beginning, the general objectives of air monitoring, ambient air quality standards for so-called criteria pollutants, and their sources are presented. In the next part, both analytical methods and instruments for monitoring of ambient air and stack gases are briefly presented. Additionally, other approaches applied in air pollution monitoring, such as biomonitoring, geographical information system (GIS), or remote monitoring, are also briefly characterized.

Objectives of Air Monitoring

Collecting information on the presence and concentration of pollutants in the environment, both naturally occurring or from anthropogenic sources, may be achieved by measurements of such substances or phenomenon of interest. For realistic assessment, temporal and spatial variations of concentrations in the particular environmental compartment, repeated measurements rather than single ones, are made.

The general aim of monitoring is to provide information about the actual levels of harmful or potentially harmful pollutants to indicate areas in which the quality of air does not fulfill proper standards. The main objectives of air monitoring are as follows:

- To measure pollutant mixing ratios and their interactions, patterns, and fate in the environment.
- To carry out ecotoxicological studies and assessment of the effects of pollution on man and the environment, to identify possible cause-and-effect relationship between pollutant concentration and health effects.
- To assess emission sources and the need for legislative controls on emissions of pollutants and to ensure compliance with emission standards.
- To activate emergency procedures in areas prone to acute pollution episodes.
- To obtain a historical record of air quality to provide a database for future use.

The area of applications of air monitoring data is presented in Figure 1.

When the objectives of monitoring are clearly defined, several decisions should be made to generate suitable data for the intended use. Decisions on what to monitor, when and where to monitor, and how to monitor are usually undertaken at the beginning. More difficult are next decisions, e.g., establishing the number and location of sampling sites, the duration of the survey, and the time resolution of sampling. All the steps in the design of a monitoring program are presented in Figure 2.

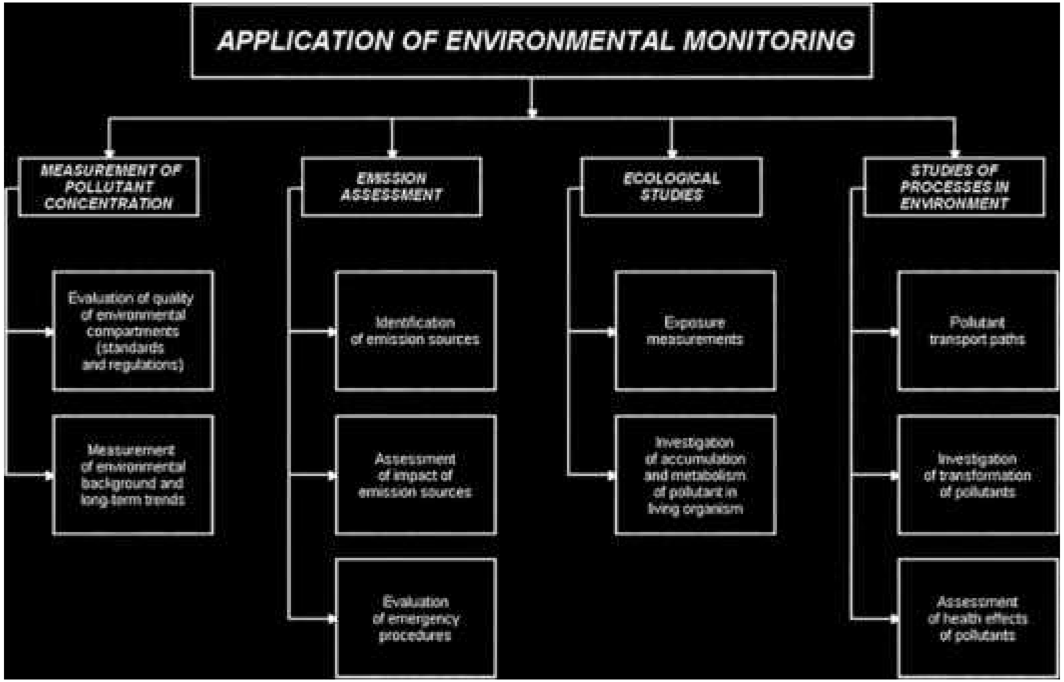


FIGURE 1 The detailed goals of activities in air monitoring.

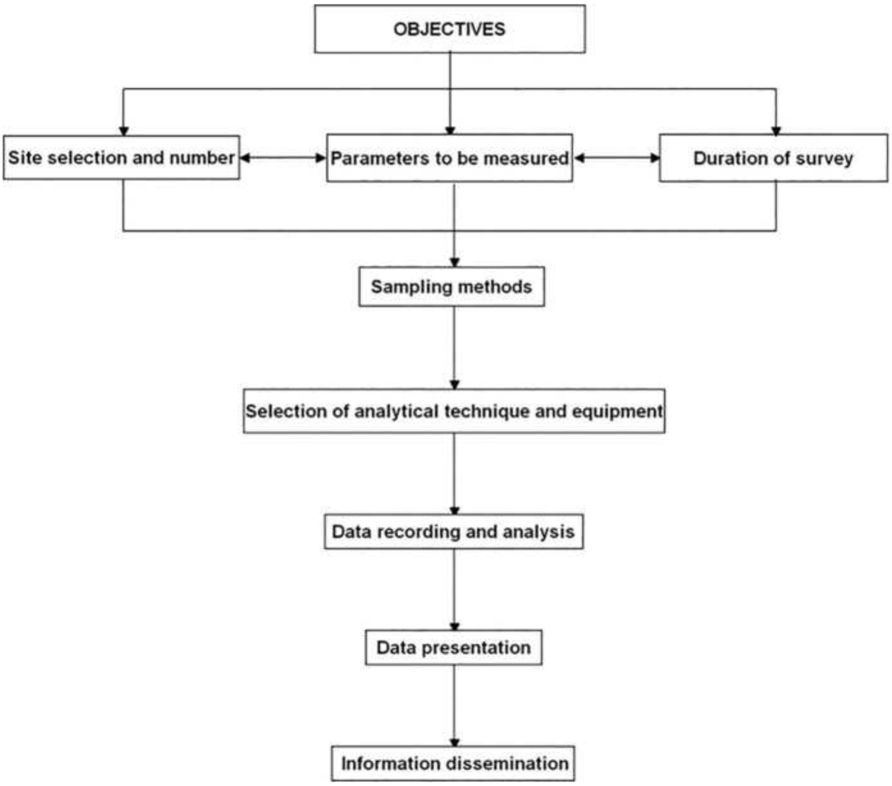


FIGURE 2 Steps in the design of a monitoring program.

History of Air Pollution Legislation

A growing concern over the influence of different air pollutants on human health was the main driving force to develop and implement air quality criteria and standards. Impetus was given to the development of air quality standards in 1958 when it was realized that photochemical problems could not be resolved without control of motor vehicle emission.

Air Quality Standards

Efforts to regulate air quality by law were discussed and undertaken in the 1960s. One of the first proposition was presented by Atkisson and Gaines[1] in 1970. The initial regulations were set in California in response to concerns over human health.[2] After that, similar air quality programs were soon adopted nationally. In 1967, the U.S. Congress enacted the Air Quality Acts, the first modern environmental law.

The Clean Air Act,[3] which was last amended in 1990, requires the United States Environmental Protection Agency (US EPA) to set National Ambient Air Quality Standards (NAAQS) for pollutants considered harmful to public health and the environment. The Clean Air Act established two types of national air quality standards. *Primary standards* set limits to protect public health, including the health of "sensitive" populations such as asthmatics, children, and the elderly. *Secondary standards* set limits to protect public welfare, including protection against decreased visibility and damage to animals, crops, vegetation, and buildings. EPA has established NAAQS for six principal pollutants, which are called criteria pollutants: sulfur dioxide, particulate matter (PM), nitrogen oxide, carbon monoxide, ozone, and lead. These standards are threshold concentrations based on a detailed review of scientific information related to effects.

In Europe, the first international air quality standards were introduced by the European Commission in 1980 for SO_2 and suspended particulates, mainly aimed at protecting human health. A few years later, the WHO, recognizing ecological damage as being relevant to human health, introduced air quality guidelines for Europe, which include the former as well as the latter revision in 2000.[4] The newest directive on ambient air quality and cleaner air of the European Union (EU) entered into force in June 2008.[5] It merges four earlier directives and one Council decision into a single directive on air quality. The new directive of the EU on air quality takes into account concerns from latest WHO air quality guidelines[6] on fine particles. Reflecting the latest WHO air quality guidelines that identify fine particles (PM2.5) as one of the most dangerous pollutants for human health, the new EU directive sets objectives and target dates for reducing population exposure to PM2.5. It also maintains limits for concentration of coarser particles known as PM10 and other main pollutants already subject to legislation.

Table 1 presents examples of air quality standards issued by the EPA, the WHO, and some states.

TABLE 1 Comparison of Limit Values (pm/m^3) for a Given Averaging Time and the Number of Exceedances per Year Issued by Different Countries and Organizations

Pollutant	Averaging Time	WHO	EPA	EU	UK	France	Germany	Poland
SO_2	10–15 min	500	–	–C	266 (not more than 35 times)	–	–	–
	30 min	–	–	–	–		–	–
	1 hr	–	–	350 (not more than 24 times)	350 (not more than 24 times)	350 (not more than 24 times)	350	350 (not more than 24 times)
	3 hr	–	1,300	–	–	–	–	–
	24 hr	125	365	125 (not more than 3 times)	125 (not more than 3 times)	125 (not more than 3 times)	125	125 (not more than 3 times)
	Year	50	80	20	20	20	20	20
NO_2	30 min	–	–	–	–	–	200	–
	1 hr	200	–	200 (not more than 18 times)	200 (not more than 18 times)	230 (not longer than 0.2 % of time)	–	200 (not more than 18 times)
	24 hr	–	–	–	–	–	100	–
	Year	40	100	40	40	46	–	40
PM10	30 min	–	–	–	–	–	–	–
	24 hr	20	150	50 (not more than 35 times)	50 (not more than 35 times)	50 (not more than 35 times)	50	50 (not more than 35 times)
	Year	50	–	40	40	40	40	40
CO	10–15 min	100,000	–	–	–	–	–	–
	30 min	60,000	–	–	–	–	–	–
	1 hr	30,000	4,000	–	–	–	–	–
	8 hr	10,000	1,000	1,000	1,000	1,000	1,000	1,000
	24 hr	–	–	–	–	–	–	–
	Year	–	–	–	–	–	–	–
O_3	30 min	–	–	–	–	–	–	–
	1 hr	–	235	–	–	–	–	–
	8 hr	100	157	120	100 (not more than 10 times)	120	120	120 (not more than 25 days)
Pb	24 hr	–	–	–	–	–	5	–
	3 months	–	1.5	–	0.5	–	–	–
	Year	0.5	–	0.5	–	0.5	0.5	0.5
Benzene	Year	–	–	5	16.25	8	–	5

Source: WHO.[6]

Air Quality Index

The Air Quality Index (AQI), also known as the Air Pollution Index (API) or Pollutant Standard Index (PSI), is a number used by different government agencies to characterize the quality of the air at a given location. The index aims to help the public easily understand air quality level and protect the health of people from air pollution. As the AQI increases, an increasingly large percentage of the population is likely to experience increasingly severe adverse health effects. Computing the AQI requires an air pollutant concentration from a monitor or model. The function used to convert air pollutant concentration to AQI varies by pollutant and is different among countries. AQI values expressed in different values (most frequently from 0 or 1 to 10, 100, or to 500) are divided into ranges (from 4 to 10), and each range is assigned a descriptor and a color code. Standardized public health advisories are associated with each AQI range. An agency might also encourage members of the public to take public transportation or work from home when AQI levels are high.

Not all air pollutants are characterized by AQI. Many countries monitor only some pollutants, e.g., ground-level ozone, sulfur dioxide, carbon monoxide, and nitrogen dioxide, and calculate AQIs for these pollutants.[7]

The EPA in the United States measures air quality in all parts of the country and publishes a daily AQI based on the data obtained. The following six priority pollutants are measured regularly in order to generate the AQI: carbon monoxide, nitrogen dioxide, particulates, sulfur dioxide, ozone, and lead. EPA has assigned a specific color to each AQI category to make it easier for people to understand quickly whether air pollution is reaching unhealthy levels in their communities. For example, the color orange means that conditions are "unhealthy for sensitive groups," while red means that conditions may be "unhealthy for everyone," and so on.

In Canada, API has values of 0 up to 100+ and is divided into four categories (from good to very poor).

CITEAIR (Common Information to European Air) has developed the first AQIs in Europe.[8] An important feature of the indices is that they differentiate between traffic and city background conditions. The Common Air Quality Index (CAQI) is designed to present and compare air quality in near real time on an hourly or daily basis. The CAQI has five levels, using a scale from 0 (very low) to >100 (very high) and the matching colors range from light green to dark red. The Year Average Common Air Quality Index (YACAQI) uses a different approach adopting the difference to target's principle. If the index is higher than 1.0, it means that for one or more pollutants, the limit values are not met. If the index is below 1, it means that on average the limit values are met. Both indices are practically implemented on a Common Operational Webpage (COW).[9] The project CITEAIR II will further develop the AQIs.[10]

The EU's Sixth Environment Action Programme (EAP), "Environment 2010: Our Future, Our Choice," includes Environment and Health as one of the four main target areas requiring greater effort—air pollution is one of the issues highlighted in this area. The Sixth EAP aims to achieve levels of air quality that do not result in unacceptable impacts on, and risks to, human health and the environment.

The EU is acting at many levels to reduce exposure to air pollution: through EC legislation, through work at the international level to reduce cross-border pollution, through cooperation with sectors responsible for air pollution, through national and regional authorities and NGOs, and through research. The Clean Air for Europe (CAFE) initiative has led to a thematic strategy setting out the objectives and measures for the next phase of European air quality policy.[11]

Regulated Air Pollutants

The contaminants in ambient air that are of concern are basically categorized as criteria and non-criteria pollutants.[12]

Criteria air pollutants are those air contaminants for which numerical concentration limits have been set as the dividing line between acceptable air quality and poor or unhealthy air quality. Criteria pollutants include five gases/vapors and two solids: nitrogen oxides (NO_x), sulfur dioxide (SO_2), carbon monoxide (CO), ozone (O_3), benzene, PM10, and lead (Pb). Non-criteria pollutants are those contaminants designated as toxic or hazardous by legislation or regulation. They fall in two further subcategories, depending on the legislation that defines them. In general, hazardous air pollutants may pose a variety of health effects, whereas toxic ones focus on one physiological response.

Main Sources of Air Pollutants

Air pollution may be defined as a situation in which substances change the qualitative composition of air in relation to average composition of troposphere sufficiently high above their normal ambient air levels to produce a measurable effect on humans, animals, vegetation, or material.[13–17]

Pollutants (both organic and inorganic) may be present in different forms such as gases, aerosols (liquid, solid), and sorbates and have a very broad range of concentration.

The concentrations of ambient air pollutants are expressed in terms of either a mass per unit volume ratio, such as $\mu g/m^3$, or a volumetric ratio (i.e., volumes of contaminant per million or billion volumes of air). The conversion between mass units and volumetric ratios at standard temperature or pressure is:[12]

$$\mu g / m^3 = \text{ppm} \times \text{MW}/0.02445 = \text{ppb} \times \text{MW}/24.45,$$

where:
$\mu g/m^3$—micrograms per cubic meter
ppm—parts per million by volume ($1{:}10^6$)
ppb—parts per billion by volume ($1{:}10^9$).
MW—molecular weight of the contaminant

Some of the most important atmospheric pollutants, their sources, and impacts on the environment and human health are presented in Table 2.

TABLE 2 Atmospheric Pollutants and Their Sources and Effects

Pollutant	Sources	Impact
Sulfur dioxide (SO_2)	Power generation, industry	Acid deposition, smog formation, threat to human health, smog formation
Nitrogen oxides (NO, N_2O, NO_2)	Transport, power generation, industry	Acid deposition, smog formation, O_3 precursor, threat to human health
Carbon dioxide (CO_2)	Combustion processes, power generation, transport, landfills	Global warming
Carbon monoxide (CO)	Combustion processes, power generation	Toxic to humans
Particulate matter (PM)	Power generation, industry, transport	Threat to human health, reduced visibility
Volatile organic compounds (VOCs)	Transport, industry	Photochemical smog, O_3 precursor, global warming
Ozone (O_3)	Photochemical reactions between VOCs and NO_x	Photochemical smogs, respiratory irritant, crop damage
Methane (CH_4)	Landfills, agriculture, gas industry	Global warming
Benzene, 1,3-butadiene	Transport industry	Carcinogenic
Ammonia (NH_3)	Industry, farming, refrigeration, power plant	Toxic to humans and wildlife
Heavy metals	Industry, transport	Toxic to humans and wildlife
Dioxins and furans	Incineration, electrical equipment	Toxic to humans

Source: Bogue.[18]

Urban traffic has become the most important cause of air pollution in the cities.[15] Road traffic is responsible for emission of several air pollutants; the most important of which are nitrogen oxides (NO and NO_2), sulfur dioxide (SO_2), PM, carbon monoxide (CO), and volatile organic compounds (VOCs), all of which can pose a health hazard.

Characteristics of Criteria Air Pollutants

Air pollutants arise from a wide variety of sources although they are mainly a result of the combustion process.[15] The largest sources include power generation, motor vehicles, and industries. The emissions of pollutants to the atmosphere badly influence vegetation, human and animal life, agriculture, and climate. Emissions of carbon monoxide (CO), nitrogen oxides (NO_x), and hydrocarbons are controlled by catalytic converters on new gasoline-driven cars. Emissions of sulfur oxides are being reduced through a lower sulfur content in gasoline. However, emissions of PM are not decreasing. Any successful strategies for controlling or countering these problems must be based on reliable air quality monitoring data for management, to make informed decisions on air pollution control.

Volatile organic compounds is a collective name for a very large number of different chemical species, including hydrocarbons, halocarbons, and oxygenates that have different physicochemical properties and are directly emitted from both anthropogenic and natural sources, and which can contribute to the formation of secondary pollutants with different efficiencies. For vehicular emissions, the list of compounds is long and variable depending on fuel, type of engine, and operating conditions. Hydrocarbons such as ethane, ethyne, higher aliphatic hydrocarbons, benzene, toluene, and xylenes are typical emissions in most cases. Each of these compounds can be released unreacted or can undergo oxidation reactions. One of them, benzene, is found in highest concentrations. Ambient concentrations are typically between 1 and 50 ppb, but close to major emissions can be as high as several hundred parts per billion. In the unreacted state, it has undesirable ecotoxicological properties. Besides causing annoying physiological reactions such as dizziness and membrane irritation, it is known to be a human carcinogen.

The two nitrogen oxides, NO and NO_2 (together called NO_x), from anthropogenic sources are present as a consequence of various combustion processes from both stationary sources, i.e., power generation (21%), and mobile sources, i.e., transport (44%). These species have very short atmospheric lifetimes, around 5 days, and have been ultimately converted to nitric acid and removed in rainfall. However, nitrogen oxide is important because it is a precursor to tropospheric ozone. Whereas NO does not affect climate, ozone does. A typical sea-level mixing ratio of NO is 5 ppt (parts per trillion, $1{:}10^{12}$), but in urban regions, NO mixing ratios reach 0.1 ppm in the early morning, but it decreases to zero by midmorning due to reaction with ozone. A major source of NO_2 is oxidation of NO, with NO_2 being intermediary between NO emission and O_3 formation. Nitrogen dioxide is one of the six criteria air pollutants for which ambient standards are set by the US EPA under CAAA70 (Clean Air Act Amendments of 1970). In the urban regions, the mixing ratio of NO_2 ranges from 0.1 to 0.25 ppm. It is more prevalent during midmorning than during midday or afternoon because sunlight breaks down most NO_2 past midmorning. Exposure to high concentrations of NO_2 harms the lungs and increases respiratory infections. It may trigger asthma by damaging or irritating and sensitizing the lungs, making people more susceptible to allergens. At higher concentrations, it can result in acute bronchitis or death.

Sulfur dioxide (SO_2) is a strong-smelling, colorless gas that is formed by the combustion of fossil fuels, smelting, manufacture of sulfuric acid, conversion of wood pulp to paper, incineration of refuse, and production of elemental sulfur. Power plants, which may use coal or oil high in sulfur content, can be major sources of SO_2, accounting for about 50% of annual global emissions. SO_2 and other sulfur oxides contribute to the problem of acid deposition and can be major contributors to smog. Natural background levels of SO_2 are about 2 ppb. Hourly peak values can reach 750 ppb on infrequent occasions. Sulfur dioxide can lead to lung diseases. SO_2 is a criteria air pollutant.

Ozone (O_3) is not directly emitted from both anthropogenic and natural sources. Its only source into air is chemical reaction. In the urban air, ozone mixing ratios range from less than 0.01 ppm at night to 0.5 ppm (during afternoons in the most polluted cities worldwide), with typical values of 0.15 ppm during moderately polluted afternoons. Ozone causes headaches at concentrations greater than 0.15 ppmv, chest pains at mixing ratios greater than 0.25 ppm, and sore throat and cough at mixing ratios greater than 0.30 ppm. Exceeding the level 0.30 ppm, it decreases lung functions. Symptoms of a respiratory condition include coughing and breathing discomfort. Ozone can also accelerate the aging of lung tissue. It also interferes with the growth of plants and deteriorates organic materials, such as rubber, textiles, and some paints and coatings. Furthermore, ozone increases plant and tree stress and their susceptibility to disease, infestation, and death.

PM, frequently described simply as particle pollution, in the atmosphere arise from natural sources, such as wind-borne dust, sea spray, and volcanoes, and from anthropogenic activities, such as combustion of fuels. Particle pollution in the air includes a mixture of solids and liquid droplets and come in a wide range of sizes. Those less than 10 micrometers (μm) in diameter (PM10) are so small that they can get into the lungs, potentially causing serious health problems. Particles less than 2.5 pm in diameter are called fine particles. These particles are so small that they can be detected only with an electron microscope. Sources of fine particles include all types of combustion, including motor vehicles, power plants, residential wood burning, forest fires, agricultural burning, and some industrial processes. Particles between 2.5 and 10 μm in diameter are referred to as coarse. Sources of coarse particles include crushing or grinding operations and dust stirred up by vehicles traveling on roads. After releasing into air, particles can change their size and composition by condensation of vapor species or by evaporation, by coagulating with other particles, or by chemical reaction. Particles with a diameter smaller than 1 pm generally have atmospheric concentrations in the range from around 10 to several thousands per cubic meter; those exceeding 1 pm diameter are usually found at concentration less than 1 cm^{-3}.

Carbon monoxide (CO) is a colorless, odorless gas that is produced by the incomplete burning of carbon-based fuels including petrol, diesel, and wood. It is also produced from the combustion of natural and synthetic products such as cigarettes. Natural background levels of CO fall in the range of 10–200 ppb. Levels in urban areas are highly variable, depending upon weather conditions and traffic density. Eight-hour mean values are generally less than 10 ppm, but sometimes, they can be as high as 500 ppm. Carbon monoxide lowers the amount of oxygen that enters the blood. It can slow human reflexes and make people confused and sleepy.

Air Quality Monitoring

Design of Monitoring Networks for Air Pollution

For the purpose of monitoring and reporting air pollution, most industrialized countries have been divided into regions or zones and urban areas or agglomeration, e.g., in Europe, in accordance with EC Directive 96/62/EC.[19] This Directive sets a framework for ways how to monitor and report ambient levels of air pollutants. Other directives set ambient air limit values for particular pollutants:

- Directive 99/30/EC for nitrogen dioxide and oxides of nitrogen, sulfur dioxide, lead, and PM.
- Directive 2000/69/EC for benzene and carbon monoxide.
- Directive 2002/3/EC for ozone.

The monitoring sites are organized into automatic and non-automatic networks (regional and national) that gather a particular kind of information using a particular method. For example, across the U.K., there are more than 1500 monitoring sites that monitor air quality, and these are organized into networks (automatic and non-automatic) that gather a particular kind of information, using a particular method. The pollutants measured and method used by each network depend on the reason for setting

FIGURE 3 Stationary monitoring station of the ARMAAG network.

FIGURE 4 General view of an interior of a monitoring station.

up the network, and what the data are to be used for. In Poland, monitoring of air quality has been performed systematically since 1992, mainly by using automatic air monitoring stations. Air quality monitoring data are used at national, voivodship (provincial), and local scale. An exemplary air automated monitoring station, together with a general view of analyzers situated in them, is shown in Figures 3 and 4. This station belongs to the Agency of Regional Air Monitoring of the Gdansk Agglomeration (ARMAAG), which is Poland's first local government-owned air monitoring network. Foundation ARMAAG provides information of air condition in Gdansk agglomeration in real time from the automatic measurement network.

The obtained data from air monitoring are used in air quality inventories and bulletins.

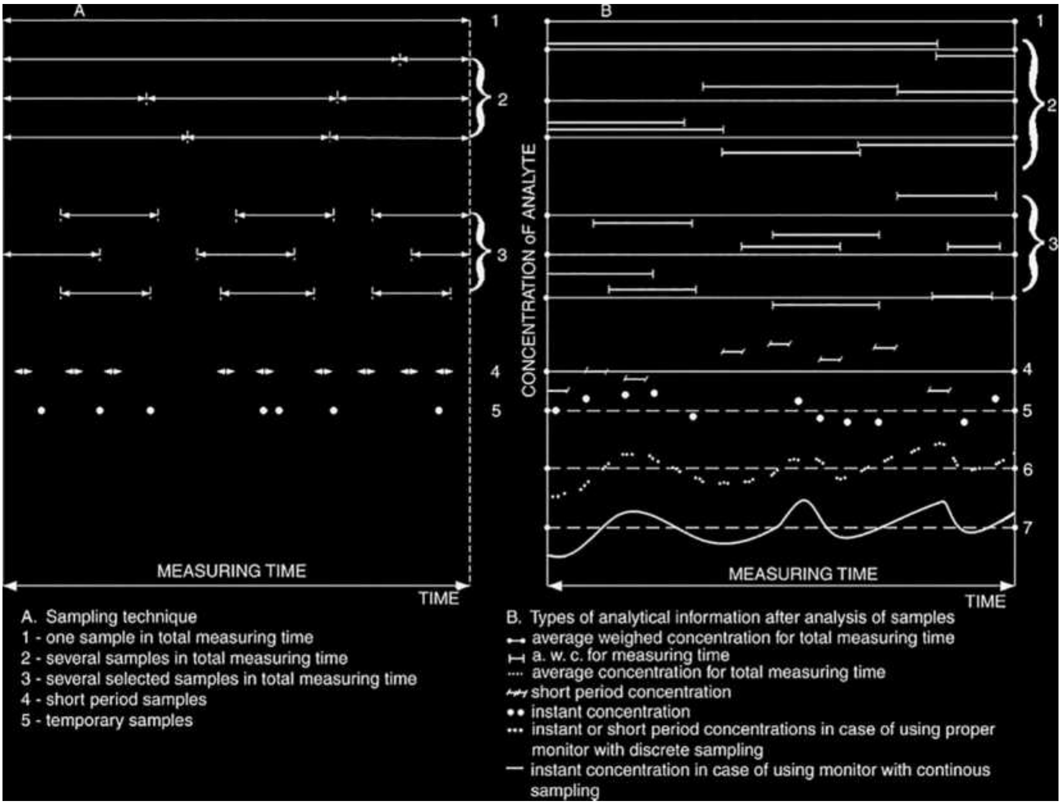


FIGURE 5 Schematic diagram of different sampling techniques used for getting information on concentration of pollutants in determined measuring time.

Types of Information Obtained from Air Monitoring

The obtained information concerns different types of concentration of investigated pollutants depending on applied sampling techniques and measuring period. The results of measurements may be referred to real time (instantaneous concentrations) or to a selected period of time (e.g., 1 hr, 8 hr, 24 hr, 1month, 1year). Final measurements represent averaged concentrations.

Figure 5 presents schematically different forms of concentration obtained as a function of sampling time.[20]

Considering the frequency of sampling, discrete, periodic, and instantaneous measurements are distinguished. Taking into account space, parameter measurements are divided to a point, averaged along a defined part of space, and averaged on the selected area. Point monitoring is inadequate to measure poorly mixed gases such as fugitive emissions over large areas. If the point instrument is wrongly placed, measurement results are not representative. Final measurements enable determination of weighted average concentrations over the sampling period.

General Requirements of the Instruments Used in Monitoring

Monitoring of air pollution is a prerequisite of air quality control and is carried out by a wide variety of analytical methods employing different measuring instruments (analyzers) that have different sensitivities and specificities.[21]

The basic requirement of the analyzers for air monitoring is high measurement sensitivity, i.e., the low limit of detection (LOD) and the low limit of quantitation (LOQ). It gives a chance to detect the pollutants at required levels. The instruments that should acquire analytical data in real time or only with a small time delay have to possess the following additional capabilities:

- Providing high data resolution (characterized by low response time).
- Providing automatic calibration and zeroing.
- Long functioning without service.

The last demand means that they should be equipped with an independent power supply and be able to automatically regenerate or exchange worn-out filters and, depending on the type of detector used (sensor), fulfill special demands, e.g., for electrochemical sensors, exchange or supplement the working solution and reagents, and in devices with Flame Ionisation Detector (FID) or Flame Photometric Detector (FPD) detection, protect against flame extinguishing.

Depending on the number of analytes that an instrument can determine in a single sample, they can be single-parameter (single gas) or multiparameter (multigas) instruments.

Based on sampling frequency, analyzers can be discrete (for single measurement), periodic (for measurements at preset intervals), or continuous (for permanent monitoring).

Classification of the Instruments Used for the Detection and Monitoring of Air Pollutants

The analytical instruments currently used for the detection and determination of atmospheric pollutants can be classified according to various criteria.

Recently, measuring techniques based on a physical (or physicochemical) principle are more frequently used in the assessment of air quality. Such methods involve direct determination of a physical property of a pollutant, sometimes after its interaction with another compound. In this approach, better stability, sensitivity, and reliability may be easily achieved. Furthermore, the practical application requires less maintenance. Instruments based on this principle can be easily automated, which enables their use in providing continuous measurements needed for up-to-date assessments of air quality. It is especially relevant to environmental monitoring because many existing standards refer to specified period of time, i.e., 1 hr, 24 hr, or 1 year.

According to the location where measurements are taken, instruments can be stationary or *on-site*. In the first case, analysis is performed in the laboratory and sophisticated instruments are applied, such as mass, electron mobility, or x-ray fluorescence spectrometers. *On-site* systems enable measuring of pollution levels in the field. Since access to a sophisticated laboratory is not required, the devices (usually uncomplicated, relatively cheap, and portable) hold great promise for use in remote locations. The main advantage of on-site analysis is the potential for rapid assessment and response to a particular problem.

All monitoring systems can be classified as mobile or stationary. Most existing systems monitoring gaseous pollutants of atmospheric air and ambient aerosols, both automatic and manual, usually perform stationary measurements; i.e., they are directly linked to a specific point or space in the vicinity of that point. Basing on the data obtained from single monitoring sites, it is not possible to assess spatial and temporal variations of air pollutants.

Mobile refers to a continuous-monitoring instrument that is portable or transportable. They are usually designed to perform analytical measurements without preliminary operations. Portable refers to self-contained, battery-operated, or worn or carried by the person using it, or may require the use of special vehicles for placement in a specific area to be monitored. Transportable gas monitors can be mounted on a vehicle such as a car, plane, balloon, ship, or space shuttle, but not to a mining machine or industrial truck.

For mobile systems, the registered values of pollutants have to be correlated with information about the geographical site and actual meteorological conditions (temperature and humidity).

Portable systems for field measurements should meet the following requirements:

- Compactness and robustness
- Ease of handling
- Adaptability to on-site measurements
- Automated operation with a long-lasting power supply
- Stability under aggressive environmental conditions

Several contributions published during the last decade have proven the advantages of mobile systems in getting information concerning the spatial and temporal distribution of atmospheric trace gases, without the need of a dense network of stationary stations. Most of the proposed systems are based on application of mobile laboratories,[21–23] equipped with appropriate monitors. There are also systems that allow to measure pollutants in a stream of vehicles but the measuring unit is installed on any vehicle[24–27] rather than attached to a dedicated van.[28,29]

The general trend in the field of creating instruments for air quality assessment is combining several instruments into one system and forming so-called hybrid multisensor systems, controlled by a microprocessor capable of transferring the obtained data to a central station, frequently using a wireless system. In the central station, the data are collected both from single objects (houses, plants) and from large areas. Many systems are equipped with devices for testing the sensors and for providing diagnosis of the whole instrument. Frequently, they have alarms that warn the user of any dangerous situation due to the breaching of some value limit. Such systems are battery powered and able to work continuously for several days or months.

The environments in which analyzers are used differ from the relative calm of the laboratory. Analyzers have to withstand wide ambient temperatures, fluctuations, and vibrations. Due to this, many systems are completely sealed so as to operate independent of outside conditions and be able to withstand the onslaught of monkey-wrench mechanics.

General Characteristics of the Methods and Analytical Instruments for Air Monitoring

Due to the complexity of environmental problems and the variety of pollutants and their different concentrations (typically parts per million or percentage levels in stack gases and parts per billion in air), there is a wide range of methods and instruments used in measuring ambient air quality.

Based on physicochemical principles, the monitoring instruments involve direct determination of the different physical properties of the pollutant or following its interaction with another compound. These methods allow determination of air pollutants in a continuous and automatic way. Such approach requires extremely sensitive instrumentation. Therefore, the most advanced techniques, comprising chemistry, physics, and microelectronics, should be used. As a result, the instruments are combinations of many different devices giving one measuring system. In developing such a system, it should be remembered that it will be exposed to environmental impact, such as changes in temperature, dustiness, humidity of air, aggressive components of air, vibrations, and transportation stress.

The typical instruments used for atmospheric ambient monitoring are based on optical, electrochemical, and semiconductor principles. Among spectroscopic techniques, chemiluminescence, infrared (IR), and fluorescence are the most frequently applied.

The range of typical measuring methods and techniques used in air monitoring is shown in Table 3.

Among the many different optical spectroscopic methods, differential optical absorption spectroscopy (DOAS) has found wide use in atmospheric research and air quality monitoring. The technique is based on the measurements of absorption features of gas molecules along a path of known length in the open atmosphere. The DOAS systems, due to the calibration-free absolute measurements and the unequivocal identification of many trace pollutants, such as CO, SO, NO, and VOCs, and highly

TABLE 3 Air Quality Monitoring Techniques Used in Air Monitoring

Pollutant	Emissions	Ambient Air
CO_2	FTIR, NDIR, TDLAS	NDIR, DOAS
CO	FTIR, NDIR, TDLAS, DOAS	NDIR (gas filter correlation variant)
SO_2		UV fluorescence
NO_x		Chemiluminescence
PM	Triboelectric, opacity, beta ray attenuation	Beta ray attenuation, oscillating microbalance, gravimetric
VOCs	FID, GC	FID, GC
CH_4	NDIR, FTIR,TDLAS, FID	FID, GC
O_3		UV absorption spectroscopy, DOAS, electrochemical sensors
NH_3	FTIR,TDLAS, chemiluminescence GC	Chemiluminescence, DOAS GC
Benzene, 1,3-butadiene	In situ GC-MS with continuous sampling	MS, GC-MS
Dioxins and furans	XRF, LIBS, cold vapor AFS, atomic	Sampling plus MS or GC-MS
Metals	Emission spectrometry	Sampling plus ICP-MS, DOAS (for Hg)

Source: Bogo.[29]

FTIR, Fourier transform infrared absorption spectroscopy; NDIR, non-dispersive infrared absorption; TDLAS, tunable laser diode absorption spectroscopy; DOAS, differential optical absorption spectroscopy; FID, flame ionization detector; GC, gas chromatography; MS, mass spectrometry; XRF, x-ray fluorescence; LIBS, laser-induced breakdown spectroscopy; ICP, inductively coupled plasma; AES, atomic emission spectrometry; AFS, atomic fluorescence spectrometry.

reactive radicals, e.g., OH, NO, and halogen oxide radicals, are exploited by air quality monitoring agencies around Europe and in the United States. The physical and chemical principles, the current state of this measurement method, and details for users are broadly presented in a recently published book.[30]

In Europe, standard/reference methods (EN) are provided by the European standardization body (CEN). The standard methods use the following principles:

Chemiluminescence for NO EN 14211

Nitrogen oxide reacts with ozone, generated within the instrument, produces an excited molecule of nitrogen dioxide, which emits light returning to its original state. A photomultiplier tube measures the emitted light that, if the volumes of sample gas and excess ozone are carefully controlled, is proportional to the concentration of NO in the gas sample.[31]

The chemiluminescent method used for nitrogen oxides is based on the following reactions:

$$NO + O_3 \rightarrow NO_2^* + O_3$$

$$NO_2^* \rightarrow NO_2 + h\nu$$

The chemiluminescence technique may be used to measure total oxides of nitrogen (NO_x) by passing the sample over a heated catalyst to reduce all oxides of nitrogen to NO. This is done within the instrument, just prior to the reaction chamber. Some instruments can perform the automatic switching of the catalyst in and out of the sample path so that the resulting signals may be compared to indirectly measure NO_2.

Ultraviolet Fluorescence for SO_2 EN 14212

Sulfur dioxide is measured without chemical pretreatment by gas-phase fluorescence spectrometry in the UV region. Molecules of SO_2 are excited by UV radiation (200–220 nm) into unstable forms, which return to a basic state, emitting radiation in the range of 240–420 nm according to following reactions:[32]

$$SO_2 + h\nu \rightarrow SO_2^*$$

$$SO_2^* \rightarrow SO_2 + (UV)$$

The intensity of fluorescence radiation is proportional to sulfur dioxide in the sample.

Non-Dispersive Infrared for CO EN 14626

The analytical principle is based on absorption of IR light by the CO molecule. NDIR-GFC (non-dispersive infrared-gas filter correlation) analyzers operate on the principle that CO has a sufficiently characteristic IR absorption spectrum such that the absorption of IR by the CO molecule can be used as a measure of CO concentration in the presence of other gases. CO absorbs IR maximally at 2.3 and 4.6 pm. Because many other molecules also absorb IR radiation in practice, different technical designs are proposed. The following approaches are typically applied:[33]

- Measurement of IR absorption at specific wave for CO (2.3 or 4.6 pm)
- Analyzers with two cells, one of which is filled with pure air (compensation of drift)
- Analyzers with turning circle (GFC)

Ultraviolet Photometry for O_3 EN 14625

Upon exposure to UV light, ozone absorbs some of the light and the intensity difference is directly proportional to the concentration of ozone. Frequently, the UV light source is a 254 nm emission line from a mercury discharge lamp.[34]

Online Gas Chromatography for Benzene EN 14662-Part 3

A measured volume of sample air is drawn or forced through a sorbent tube. Provided suitable sorbents are chosen, benzene is retained by the sorbent tube and thus is removed from the flowing air stream. The collected benzene (on each tube) is desorbed by heat and is transferred by inert carrier gas into a gas chromatograph equipped with a capillary column and a flame ionization detector or another suitable detector, where it is analyzed. Prior to entering the column, the sample is concentrated either on a cryo trap, which is heated to release the sample into the column, or on a pre-column, where higher boiling hydrocarbons are removed from the pre-column by back flush. Two general types of instruments are used. One is equipped with a single sampling trap and the other is equipped with two or more traps. The single-trap instrument samples for only part of the time in each cycle, whereas the multitrap instrument samples continuously. Typical cycling times are between 15 min and 1 hr.[35]

PM is usually determined using active filter method by gravimetry as the reference method.[36] In this method, the air is passed through a filter that stops particles above 10 pm (PM10) or 2.5 pm (PM2.5). Measurements are made over a period of 24 hr or longer. The filters are collected and the adsorbed particles are measured in the laboratory. Other methods use beta ray absorption or tapered element oscillating microbalance (TEOM) of PM. In beta gauge instruments, which are used for real-time measurements of particulate emissions from stationary sources, the mass of the sample deposited on the filter tape is automatically measured by beta ray attenuation. The measurement is made first on a blank, then on the particulate-laden filter. The range is 2–4000 mg/m^3 without interference or effect from color, size, or atomic mass of the dust.

An interesting review on *online* analyzers for monitoring of VOCs was recently published.[37]

Other Monitoring Approaches

Biomonitoring Using Plants

Modern air instruments cannot measure the effects air pollution has on living cells and are limited to measuring the present conditions. Biological materials can be an excellent basis for establishing a biomonitoring network on large areas for a long time. Biomonitoring, as a continuous observation of an area with the help of bioindicators, can allow a qualitative survey and quantitative estimation of the pollutants in the environment. Since the 19th century, biological monitoring as a rather simple observation has turned into a serious alternative if not a useful complement to the traditional methods of assessing contamination, from both natural and anthropogenic sources. In the case of airborne pollution, its heyday really began after World War II. The expensive growth of biomonitoring research works has gained momentum mainly from lower organisms such as lichens, bryophytes, and, to a lesser degree, fungi. The use of cosmopolite organisms for assessing pollution has developed notably during the last two decades.

Bioindicators are organisms or organs of such organisms that respond to a certain level of pollution by the change in their life cycle or accumulation of the particular pollutant. Bioindicators, in contrast to direct analysis, reflect complex effects of harmful substances, as such organisms show not only the synergistic effects of a sum of parameters but also a time-integrated picture of the history of their life span. Another advantage is the selective uptake of such substances, as an organism exposed to an environmental pollutant, either through air or via direct uptake, absorbs the bioavailability fraction only. They readily reflect the proportion of the pollutants, which may be dangerous to human beings as well.

In air monitoring, two organisms, i.e., lichens and mosses, have become the most popular bioindicators.

Lichens are unusual organisms because they consist of fungal threads and microscopic green alga living together and functioning as a single organism. Lichens grow on rocks, soil, trees, or artificial structures in unpolluted habitats. Lichens act like sponges, taking in everything that is dissolved in the rainwater and retaining it. Different species of lichens vary in sensitivity to air pollution. Lichens are commonly used as air quality indicators since some species are more pollution tolerant than others. The most sensitive lichens are shrubby and leafy, whereas the most tolerant lichens are crustose. In city centers, lichens may be entirely absent. If the air is clean, shrubby, hairy, and leafy lichens colonize every available surface.

Lichens may be used as bioindicators in two different ways:

- By mapping all species present in a specific area.
- Through the individual sampling of lichen species and measurement of the accumulated pollutants or by transplanting lichens from a clean environment to a contaminated one.

Lichens are used as biomonitors in the examination of the level of pollutants such as sulfur, nitrogen, and phosphorus compounds, as well as ozone, fluorides, chlorides, and heavy metals. Several biomonitoring methods using lichens have been described since the 1970s when Hawksworth and Rose[38,39] proposed a method based on the determination of zones (from 0—strong pollution to 10—clear air) with selected epiphytic lichens (on two different kinds of tree bark) that relate to levels of sulfur dioxide pollution. This method was widely adopted (both in the original scale and in relation to the real concentrations of sulfur dioxide) in many countries mainly due to its simplicity. Over the last decade, new techniques like the European method for mapping lichen diversity (LDV), as an indicator of environmental stress/quality, have been proposed. The procedure is based on the fact that epiphytic lichen diversity is impaired by air pollution and environmental stress. It provides a rapid, low-cost method to define zones of different environmental quality. In addition to information on the long-term effects of air pollutants, data on eutrophication, anthropization, and climatic change on sensitive organisms may likewise be obtained. Data quality depends on the uniformity of growth conditions, and usually standardization in sampling procedure is necessary.

The relative ease of sampling, the absence of any need for complicated and expensive equipment, and the accumulative and time-integrative behavior of the monitor organisms that make biomonitoring of atmospheric pollutants possible could be continued for the foreseeable future, especially in large-scale surveys.[40–45]

GIS in Air Quality Monitoring

Reliable information on air quality is needed not only on temporal trends in air pollution (as, for example, provided by data from fixed-site monitoring stations) but also on geographical variations. Maps are needed, for example, to identify so-called hot spots, to define at-risk groups, to show changes in spatial patterns of pollution, and to provide improved estimates of exposure for epidemiological studies. The development of GIS techniques offers considerable potential to mapping air pollution. GIS technology enables obtaining statistical and spatial data on air quality by estimation of environmental levels of regulated contaminants.[46]

Remote Monitoring Techniques

Remote monitoring techniques enable the measurement of atmospheric pollutants in remote, poorly accessible, and dangerous regions.[47–49] Remote sensing is especially recommended for the detection of diffuse emissions that are hard to quantify with typical ground-point measurements, but its use is restricted to specialized monitoring demands due to very considerable cost of the equipment.

The cheapest and most widely used methods are those of aerial photography, including IR sensing and optical spectroscopy. Different spectrometers, whether they are ground based in a mobile laboratory or airborne, are the most common instruments used in air pollution for determination of SO_2 and NO_2 concentrations in plumes from tall stacks. The results provide reliable data for studying the transport and dispersion of a plume.

Typical gaseous pollutants such as NO_x, SO_2, CO, and O_3 may be monitored using different types of lasers, which, due to long-path absorption measurements, enable the determination of pollutants at very low concentrations. In laser absorption methods, a detector is used to monitor absorption of specific wavelengths in light paths. Lidar (light detection and ranging) transmits light out to a target and part of this light is reflected/scattered back to the instrument where it is analyzed. The time for the light to travel out to the target and back to the lidar is used to determine the range to the target, allowing spatial resolution of pollutant concentration data within the light path; by monitoring back-scatter intensity at two close wavelengths, the species concentration as well as its spatial distribution may be inferred. This method has been successfully used for measurements of SO_2 up to a range of 2 km.

There are three basic generic types of lidars: range finder lidars, differential absorption lidars (dial), and Doppler lidars.

Range finder lidars are used to measure the distance from the lidar instrument to a solid or hard target. Dial is used to measure pollutant concentration in the atmosphere on the basis of the difference of the two return signals having two different wavelengths (one is absorbed by the molecule of interest while the other is not absorbed). The Doppler lidar is used to measure the velocity of a target. The target can be either a hard target or an atmospheric target.

Another type of device used for remote monitoring is sodar (sound detection and ranging). The Doppler sodar sends out sound pulses of several frequencies in slightly different directions. The acoustic signals are back-scattered by inhomogeneities in the atmosphere.

Further significant developments of laser techniques use the Raman back-scatter, which is highly characteristic of the scattering molecule.

Monitoring of Flue and Exhaust Gas Emissions

Different industrial branches, e.g., coal-fired power plants, chemical plants, petrochemical plants, oil refineries, PVC factories, heavy industries, and incinerators, are principal stationary pollution sources emitting, usually by chimneys, stack (flue) gases, containing different pollutants. Monitoring of such emission sources needs continuous, automatically acting systems that are usually a multielement, integrated, and cooperative set of measuring devices, auxiliary equipment, and calibration appliance.[50,51]

Systems for Continuous Monitoring of Stack Gases

The monitoring systems for continuous monitoring can be classified on the basis of different criteria. Depending on the way in which measurement is made, and especially on the applied sampling mode, extractive and in situ systems are distinguished.[52]

In extractive systems, as the name implies, the sample is extracted continuously from a duct or stack from a representative volume of stack gases and transported by transfer line to analyzers (one or more single-component analyzers or one multicomponent analyzer). However, in most cases, some conditioning of the sample is required to remove water vapor and PM. The two main types of extractive systems are fully extractive (sometimes called simply "extractive") and dilution extractive (also known as "dilution").

A typical extractive system (Figure 6) has a stainless steel probe, with a filter to remove coarse particulates. After filtration, a heated, unchanged sample is transferred to a sample conditioner located in the system enclosure. Calibration gas is delivered from the enclosure to the probe and back through the sample tubing to calibrate the system. The simplest sampling conditioning method is cooling the sample and allowing the moisture to condense and drain out of the system.

Instruments based on spectroscopic techniques (mainly UV and IR), paramagnetic properties, and solid electrolytes (zirconium dioxide) for oxygen determination are frequently used to monitor CO, CO_2, NO_x, and SO_2. For low concentrations of NO_x and SO_2, a chemiluminescent method can be applied. Fully extractive systems can be sometimes used without moisture removal, especially when the sample contains components that are easily soluble in water.

Fully extractive systems are recommended for monitoring of pollutants in stack gases with different physicochemical parameters of compounds. Another advantage of such systems is the possibility of monitoring several locations using one analyzer (time-share systems).

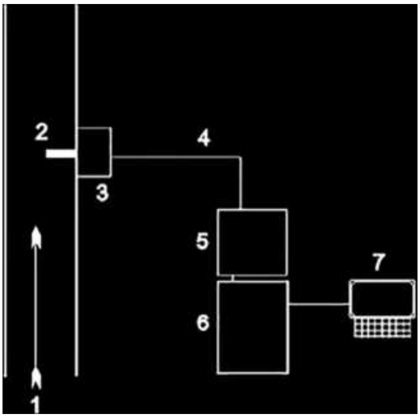


FIGURE 6 Schematic diagram of a fully extractive system for continuous emission monitoring of stack gases: 1, stack gases; 2, probe; 3, filter; 4, heated sampling line; 5, moisture removal; 6, analyzers; 7, storage unit.

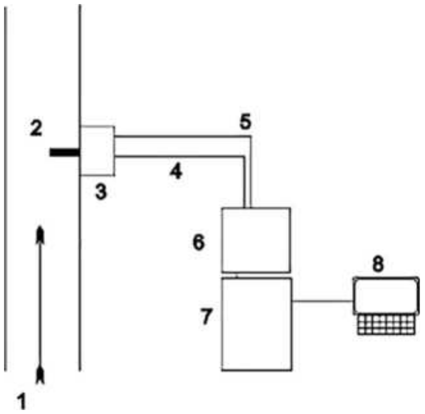


FIGURE 7 Schematic diagram of a dilution-extractive system for continuous emission monitoring of stack gases: 1, stack gases; 2, probe; 3, filter; 4, clean and dry dilution air; 5, diluted sample to analyzers; 6, dilution control unit; 7, analyzers; 8, storage unit.

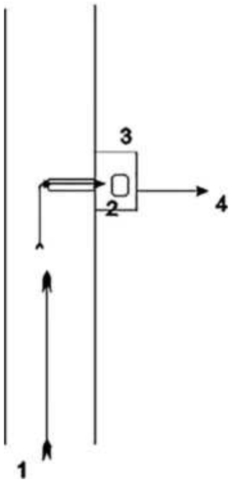


FIGURE 8 Schematic diagram for a point-type in situ system for continuous emission monitoring of stack gases (sensor mounted in the box with the sensor electronics): 1, stack gases; 2, sensor; 3, electronics in enclosure; 4, signal.

Dilution of the sample gas (Figure 7) with clean, dry air to the sample (usually from 50:1 to 250:1) considerably facilitates sample handling and reduces the dew point of the sample gas so that the sampling line can be unheated. Furthermore, the diluted sample is similar in respect of pollutant concentration to ambient air, enabling the use of ambient analyzers. Relatively small amounts of sampled gases increase the time between cleaning the filters. Because most dilution-extractive systems are affected by changes in temperature and barometric pressure, it is recommended that temperature and pressure sensors be installed at the sampling location to compensate these effects. The dilution systems are recommended for plants fueled with carbons when high levels of particulates are present in stack gases (0.1 g/m^3) and corrosive substances (e.g., HCl or SO_3).

In in situ systems (Figure 8), a gas probe is inserted into the wall. It allows monitoring the sample without removing it from the source and does not require sample conditioning or transport of the sample gas, thus minimizing the measurement errors during sampling, transferring, and conditioning

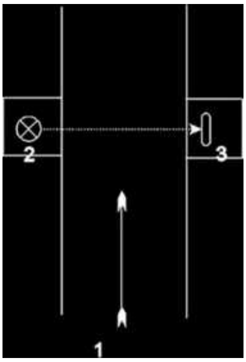


FIGURE 9 Schematic diagram of a single-path-type in situ system for continuous emission monitoring of stack gases: 1, stack gases; 2, light source; 3, receiver.

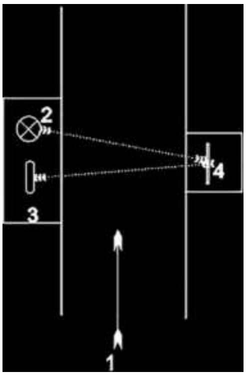


FIGURE 10 Schematic diagram of a dual-path-type in situ system for continuous emission monitoring of stack gases: 1, stack gases; 2, light source; 3, receiver; 4, reflector.

the sample. An optical beam is contained within the probe. This optical beam represents the absorption path that enables the analysis. The sample is drawn into the probe but remains under the conditions found in the stack. The sample is never removed or extracted from the stack.

In practice, two types of in situ systems are used: point and path monitoring systems (Figures 9 and 10). In point monitoring systems, a sample probe and analyzer are installed inside the stack. They are also called in-stack monitors and measure gas at a single point. Therefore, it is important to choose a location that is representative in terms of the components of interest. As analyzers in such systems, spectroscopic instruments are used (based on absorption of UV and IR radiation) as well as electrochemical devices. In situ monitoring systems are recommended for locations with easy stack access and for measuring SO_2 and O_2 in combustion sources because point monitors are very cost-effective for measuring only one or two components.

Path monitoring systems minimize errors that can arise when the location of a measuring point is not representative and due to the disturbances in the flow of stack gases. They measure gas concentration along a path, usually across the diameter of the stack or duct. A light source is mounted on one side of the stack and a beam is passed through to the other side. A single-pass system measures the light that reaches the other side of the stack, whereas a double-pass system uses a reflector and passes the light back across the stack before performing the measurements. Two parameters are limited in these systems: the length of the measuring path (no less than 0.5 m, no more than 8–10 m) and the temperature of stack gases (no more than 300°C).

In situ systems are usually mounted in the ducts after electrostatic precipitators or in chimney ducts.

The systems for stack sampling can be easily adapted to process monitoring or even ambient (closed or open path) monitoring.

Conclusions

Air quality has unquestionably adverse effects on human health. For example, air pollution is increasingly being cited as the main cause of lung conditions such as asthma—twice as many people suffer from asthma today compared to 20 years ago.

This is the main reason that the issue of air quality is now a major concern for many countries that have been working to improve air quality by controlling emissions of harmful substances into the atmosphere, improving fuel quality, and integrating environmental protection requirements into the transport and energy sectors. Despite these improvements in air quality over the last few years, the problem of air pollution still remains. Therefore, more needs to be done at the local, national, and international level. For example, a wide interest is observed to establish common criteria (e.g. AQIs-Air Quality Criteria (AQC)) to compare the state of the air for different countries that follow different directives.

During the last decade, different types of AQC were proposed in literature[53,54] and/or adopted by governments. An interesting review on AQI published recently online by the Italian Group of Environmental Statistics (GRASPA) (http://www.graspa.org) shows the lack of a common strategy to compare the state of the air for cities that follow different directives.[55] The major differences between the indices in the literature are found in the number of index classes (and their associated color) and relative descriptive terms (e.g., considered pollutants, averaging time, frequency). Also, the guidelines themselves are sometimes consistently different from state to state, not only in indicating the pollutants to be monitored but also in setting the threshold values and the number of exceedances per year. Furthermore, the way air quality is interpreted on the basis of a country-or city-specific AQI differs considerably.

Monitoring of air pollution is a prerequisite of air quality control and is carried out by a wide variety of analytical methods employing different measuring instruments that have different sensitivities and specificities. Monitoring plays a critical role in protecting the environment and is a key element of all actions related with management and protection of ambient air.

Every developed country has legislation to control or limit emissions of atmospheric pollution. The air quality standards that could be regulated by law and achieved were established. Concentrations of selected pollutants would have to be determined, and reliable analytical methods would be required to measure the levels of the pollutants. Monitoring actions are based on using stationary networks of measuring stations and/ or mobile laboratories equipped with proper instruments.

References

1. Atkisson, A.; Gaines R.S., Eds.; *Development of Air Quality Standards*; Charles E. Merril Publ. Co.: Columbus, Ohio, 1970.
2. California State Department of Public Health, Health and Safety Code, Section 426.1,1959.
3. Available at http://www.epa.gov/air/caa/ (accessed May 15, 2012).
4. World Health Organisation. *Air Quality Guidelines for Europe,* 2nd Ed.; WHO Regional Publications European Series No.91: Copenhagen, 2000.
5. Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe. Official Journal of the European Union 11.6.2008, L. 152/1–152/44.
6. WHO air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide—Global update 2005— Summary of risk assessment.
7. Available at http://armaag.gda.pl/normy.html (accessed May 15, 2012).

8. Available at http://citeair.rec.org/products.html (accessed May 15, 2012).
9. Available at http://www.airqualitynow.eu (accessed May 15, 2012).
10. Available at http://www.citeair.eu (accessed May 15, 2012).
11. Available at http://europa.eu/legislation_summaries/environment/air_pollution/l28026_en.htm (accessed May 15, 2012).
12. Griffin, R.D. *Principles of Air Quality Management;* Taylor & Francis Group: Boca Raton, 2007.
13. Jacobsen, M.Z. *Atmospheric Pollution. History, Science and Regulation*; Cambridge University Press: Cambridge, U.K., 2002.
14. Hobbs, P.V. *Introduction to Atmospheric Chemistry;* Cambridge University Press: Cambridge, U.K., 2000.
15. Friedrich, R.; Reis, S., Eds. *Emission of Air Pollutants;* Springer-Verlag: Berlin Heidelberg, 2004.
16. Seinfeld, J.H.; Pandis, S.N., Eds. *Atmospheric Chemistry and Physics. From Air Pollution to Climate Change*, 2nd Ed.; John Wiley and Sons, Inc.: New York, 2006.
17. Bell, J.N.B.; Treshow, M., Eds. *Air Pollution and Plant Life*; John Wiley & Sons, Ltd.: New York, 2002.
18. Bogue, R. Environmental sensing: Strategies, technologies and applications. Sens. Rev. **2008**, *28* (4), 275–282.
19. Council Directive 96/62/EC of 27 September 1996 on ambient air quality assessment and management. Official Journal L *296,* 21/11/1996 P, 0055–0063.
20. Michulec, M.; Wardencki, W.; Partyka, M.; Namiesnik, J. Analytical techniques used in monitoring of atmospheric air pollutants. Crit. Rev. Anal. Chem. **2005**, *35,* 1–17.
21. Wardencki, W.; Katulski, R.; Stefanski, J.; Namiesnik, J. The state of the art in the field of non-stationary instruments for the determination and monitoring of atmospheric pollutants. Crit. Rev. Anal. Chem. 2008, *38,* 1–10.
22. Bukowiecki, N.; Dommen, J.; Prevot, A.S.H.; Richter, R.; Weingartner, E.; Baltensperger, U. A mobile pollutant measurement laboratory—Measuring gas phase and aerosol ambient concentrations with high spatial and temporal resolution. Atmos. Environ. **2002**, *36,* 5569–5579.
23. Gouriou, F.; Morin, J.-P.; Weill, M.-E. On-road measurements of particle number concentrations and size distributions in urban and tunnel environments. Atmos. Environ. **2004**, *38* (18), 2831–2840.
24. Pirjola, L.; Parviainen, H.; Hussein, T.; Valli, A.; Hameri, K.; Aaalto, P.; Virtanen, A.; Keskinen, J.; Pakkanen, T.A.; Makela, T.; Hillamo, R.E. "Sniffer"—A novel tool for chasing vehicles and measuring traffic pollutants. Atmos. Environ. **2004**, *38*, 3625–3635.
25. Katulski, R.; Stefanski, J.; Wardencki, W.; Zurek, J. Concept of the mobile monitoring system for chemical agents control in the air, Proceedings of the IEEE Conference on Technologies for Homeland Security—Enhancing Transportation Security and Efficiency, Boston, USA, June 7–8, 2006; 181–184.
26. Katulski, R.; Stefanski, J.; Wardencki, W.; Zurek J.The Mobile Monitoring System (MMS)—A useful tool for assessing air pollution in cities, Proceedings of the Pittsburgh Conference on Analytical Chemistry and Applied Spectroscopy PITTCON'2007, February 25–March 2, 2007, Chicago, USA, 2007.
27. Katulski, R.; Namiesnik, J.; Sadowski, J.; Stefanski, J.; Szymanska, K.; Wardencki, W. Mobile system for on-road measurements of air pollutants. Rev. Sci. Instrum. **2010**, *81,* 045104.
28. Seakins, P.W.; Lansley, D.L.; Hodgson, A.; Huntley, A.; Pope, F. New directions: Mobile laboratory reveals new issues in urban air quality. Atmos. Environ. **2002**, *36,* 1247–1248.
29. Bogo, H.; Negri, R.M.; San Roman, E. Continuous measurement of gaseous pollutants in Buenos Aires city. Atmos. Environ. **1999**, *33*, 2587–2598.
30. Platt, U.; Stutz, J. *Differential Optical Absorption Spectroscopy. Principles and Applications*; Springer-Verlag: Berlin, Heidelberg, 2008.
31. EN 14211. *Ambient air quality.* Standard method for the measurement of the concentration of nitrogen oxide by chemilumiscence, 2005.

32. EN 14212. *Ambient air quality.* Standard method for the measurement of the concentration of sulphur dioxide by ultraviolet fluorescence, 2005.
33. EN 14626. *Ambient air quality.* Standard method for the measurement of the concentration of carbon monoxide by nondispersive infrared spectroscopy, 2005.
34. EN 14625. *Ambient air quality.* Standard method for the measurement of the concentration of ozone by ultraviolet photometry, 2005.
35. EN 14662–3. *Ambient air quality.* Standard method for measurement of benzene concentrations—Part 3: Automated pumped sampling with in situ gas chromatography, 2005.
36. Larssen, S.; de Leew, F. *PM10 and PM2.5 concentration in Europe as assessed from monitoring data reported to AirBase*, ETC/ACC; 2007.
37. Król, S.; Zabiegala, B.; Namiesnik, J. Monitoring of volatile organic compounds (VOCs) in atmospheric air. Part I. On-line gas analyzers. Trends Anal. Chem. **2010**, *29* (9), 1092–1100.
38. Hawksworth, D.L.; Rose F. Quantitative scale for estimating sulphur dioxide air pollution in England and Wales using epiphytic lichens. Nature, **1970**, *227,* 145–148.
39. Nash, T.H.; Gries C. Lichens as bioindicators of sulphur dioxide. Symbiosis **2002**, *33* (1), 1–21.
40. Svoboda, D. Evaluation of the European method for mapping lichen diversity (LDV) as an indicator of environmental stress in the Czech Republic. Biologia **2007**, *62* (4), 424–431.
41. Agraval, M.; Sing, B.; Rajput, M.; Marshall, F.; Bell, J.N.B. Effect of air pollution on peri-urban agriculture: A case study. Environ. Pollut. **2003**, *126,* 323–329.
42. Orendovici, T.; Skelly, J.M.; Ferdinad, J.A.; Savage, J.E.; Sanz, M.-J.; Smith, G.C. Response of native plants of northeastern United States and southern Spain to ozone exposures; determining exposure/response relationships. Environ. Pollut. **2003**, *125* (1), 31–40.
43. Wolseley, P.A. Using lichens on twigs to assess changes in ambient atmospheric conditions. In *Monitoring with lichens—Monitoring Lichens;* Nimis, P.L., Scheidegger, C., Wolseley, P.A., Eds.; NATO Science Series IV; Kluwer: Dordrecht, 1999; Vol. 7, 291–294.
44. Van Haluwyn, C.; van Hert, C.M. Bioindication: the community approach. In *Monitoring with lichens—Monitoring Lichens;* Nimis, P.L., Scheidegger, C., Wolseley, P.A., Eds.; NATO Science Series IV; Kluwer: Dordrecht, 2002; Vol. 7, 39–64.
45. Asta, J.; Erhard, W.; Ferretti, M.; Fornasier, F.; Kirschbaum, U.; Nimis, P.L.; Purvis, O.W.; Pirintsos, S.; Scheidegger, C.; van Haluwyn, C.; Wirth, V. Mapping lichens diversity as an indicator of environmental quality. In *Monitoring with Lichens—Monitoring Lichens;* Nimis, P.L., Scheidegger, C., Wolseley, P.A., Eds; NATO Science Series IV; Kluwer: Dordrecht, 2002; Vol. 7, 273–279.
46. Korte, G.B. *The GIS Book. How to Implement, Manage and Asses the Value of Geographic Information System*; Onward Press: Albany, NY, 2000.
47. Schroter, M.; Obermeier, A.; Bruggemann, D.; Plech-schmidt, M.; Klemm, O. Remote monitoring of air pollutant emissions from point sources by a mobile Lidar/Sodar System. J. Air Waste Manage. Assoc. **2003**, *53*, 716–723.
48. Available at http://www.ghcc.msfc.nasa.gov/sparcle_tutorial.html (accessed May 15, 2012.)
49. Weibring, P.; Andersson, M.; Edner, H.; Svanberg, S. Remote monitoring of industrial emissions by combination of lidar and plume velocity measurements. Appl. Phys. B, **1998**, *66*, 383–388.
50. Jahnke, J.A. *Continuous Emissions Monitoring*; Van Nostrand Reinhold: New York, 1993.
51. White, J.R. Technologies for enhanced monitoring. Pollut. Eng. **1995**, 27 (6), 46–50.
52. Walker, K. Select a continuous emissions monitoring system. Chem. Eng. Progress. **1996**, *92* (2), 28–34.
53. Mayer, H.; Holst, J.; Schindler, D.; Ahrens, D. Evolution of the air pollution in SW Germany evaluated by the longterm air quality index LAQx. Atmos. Environ. **2008**, *42,* 5071–5078.
54. Shooter, D.; Brimblecombe, P. Air quality indexing. Int. J. Environ. Pollut. **2009**, *36* (1–3), 19–29.
55. Plaia, A.; Rugierri, M. *Air quality indices: A review;* GRASPA Working paper no. 39; June 2010.

22

Air Pollution: Technology

Sven Erik Jørgensen

Sources of Particulate Pollution

When considering particulate pollution, the source should be categorized with regard to the contaminant type. Inert particulates are distinctly different from active solids in the nature and type of their potentially harmful human health effects. Inert particulates comprise solid airborne material, which does not react readily with the environment and does not exhibit any morphological changes as a result of combustion or any other process. Active solid matter is defined as particulate material that can be further oxidized or react chemically with the environment or the receptor. Any solid material in this category can, depending on its composition and size, be considerably more harmful than inert matter of similar size.

A closely related group of emissions come from aerosols of liquid droplets, generally below 5 μm. They can be oil or other liquid pollutants (e.g., freon) or may be formed by condensation in the atmosphere. Fumes are condensed metals, metal oxides, or metal halides, formed by industrial activities, predominantly as a result of pyro-metallurgical processes: melting, casting, or extruding operations. Products of incomplete combustion are often emitted in the form of particulate matter. The most harmful components in this group are particulate polycyclic organic matter (PPOM), which are mainly derivatives of benz[a]pyrene.

Natural sources of particulate pollution are sandstorms, forest fires, and volcanic activity. The major sources in towns are vehicles, combustion of fossil fuel for heating and electricity production, and industrial activity.

The total global emission of particulate matter is in the order of 10^7 t per year. Deposition of particles may occur by three processes:

1. Sedimentation (Stokes law may be applied, particles >20 μm)
2. Impaction (determined by differences in concentrations by use of Fick's law, particles between 5 and 20 μm)
3. Diffusion (particles <5μm)

Particles <20 μm are identified as suspended particulate matter. Particles >20 μm may be denoted dust, which will be deposited close to the source due to the high sedimentation rate. Dry deposition consists of gases or dry particles. Wet deposition is raindrops containing gases and particles. Particles may consist of minor concentrations of dissolved salts in water drops, crystals, or a combination of the two.

Particulate Pollution Problem

Particulate pollution is an important health factor, most crucially the toxicity and size distribution. Many particles are highly toxic, such as asbestos and those of metals such as beryllium, lead, chromium, mercury, nickel, and manganese. In addition, particulate matter is able to absorb gases, which enhances the effects of these components. In this context, the particle size distribution is of particular importance, as particles greater than 10 μm are trapped in the human upper respiratory passage and the specific surface (expressed as square meter per gram of particulate matter) increases with l/d, where d is the particle size. The adsorption capacity of particulate matter, expressed as grams adsorbed per gram of particulate matter, will generally be proportional to the surface area. Table 1 lists some typical particle size ranges. However, size as well as shape and density must be considered. Furthermore, particle size is determined by two parameters: the mass median diameter, which is the size that divides the particulate sample into two groups of equal mass, i.e., the 50% point on a cumulative frequency versus particle size plot, and the geometric standard deviation.

Control Methods Applied to Particulate Pollution

Particulate pollutants have the ability to adsorb gases including sulfur dioxide, nitrogen oxides, carbon monoxide, and so on. The inhalation of these toxic gases is frequently associated with this adsorption, as the gases otherwise would be dissolved in mouthwash and spittle before entering the lungs. Particulate pollution may be controlled by modifying the distribution pattern. This method is described in detail below. In principle, it represents an obsolete philosophy of pollution abatement, dilution, but it is still widely used to reduce the concentration of pollutants at ground level and thereby minimizes the effect of air pollution. Particulate control technology can offer a wide range of methods aimed at the removal of particulate matter from gas. These methods are settling chambers, cyclones, filters, electrostatic precipitators, wet scrubbers, and the modification of particulate characteristics.

Modifying the Distribution Patterns

Although emissions, gaseous or particulate, may be controlled by various sorption processes or mechanical collection, the effluent from the control device must still be dispersed into the atmosphere. Atmospheric dispersion depends primarily on horizontal and vertical transport. The horizontal transport depends on the turbulent structure of the wind field. As the wind velocity increases, so does the degree of dispersion and there is a corresponding decrease in the ground-level concentration of the contaminant at the receptor site.

TABLE 1 Typical Practical Size Ranges

	µm
Tobacco smoke	0.01–1
Oil smoke	0.05–1
Ash	1–500
Ammonium chloride smoke	0.1–4
Powdered activated carbon	3–700
Sulfuric acid aerosols	0.5–5

Source: Jørgensen[10]

The emissions are mixed into larger volume of air, and the diluted emission is carried out into essentially unoccupied terrain away from any receptors. Depending on the wind direction, the diluted effluent may be funneled down a river valley or between mountain ranges. Horizontal transport is sometimes prevented by surrounding hills that form a natural pocket for locally generated pollutants. This particular topographical situation occurs for instance in the Los Angeles area, which suffers heavily from air pollution.

The vertical transport depends on the rate of change of ambient temperature with altitude. The dry adiabatic lapse rate is defined as a decrease in air temperature of 1°C per 100 m. This is the rate at which, under natural conditions, a rising parcel of unpolluted air will decrease in temperature with elevation into the troposphere up to approximately10,000 m. Under so-called isothermal conditions, the temperature does not change with elevation. Vertical transport can be hindered under stable atmospheric conditions, which occur when the actual environmental lapse rate is less than the dry adiabatic lapse rate. A negative lapse rate is an increase in air temperature with latitude. This effectively prevents vertical mixing and is known as inversion.

The dispersion from a point source (a chimney for instance) may be calculated from the Gaussian plume model (see, for instance, Reible.[1])

These different atmospheric conditions are illustrated in Figure 1 where stack gas behavior under the various conditions is shown. Further explanations are given in Table 2. The distribution of particulate material is more effective the higher the stack. The maximum concentration, $C_{\max}$, at ground level can be shown to be approximately proportional to the emission and to follow approximately this expression:

$$C_{\max} = kQ/H^2 \tag{1}$$

where Q is the emission (expressed as grams per particulate matter per unit of time), H is the effective stack height, and k is a constant. The effective height is slightly higher than the physical height and can be calculated from information about the temperature, the stack exit velocity, and the stack inside diameter.

These equations explain why a lower ground-level concentration is obtained when many small stacks are replaced by one very high stack. In addition to this effect, it is always easier to reduce and control one large emission than many small emissions, and it is more feasible to install and apply the necessary environmental technology in one big installation.

Particulate Pollution Control Equipment

Environmental technology offers several solutions to the problem of particulate matter removal. The methods and their optimum particle size and efficiency are compared in Table 3. The cost of the various installations varies of course from country to country and is dependent on several factors (material applied, standard size or not standard size, automatized, and so on). Generally, electrostatic precipitators are the most expensive solution and are mainly applied for large quantities of air. Wet scrubbers also belong among the more expensive installations, while settling chamber and centrifuges are the most cost-effective solutions.

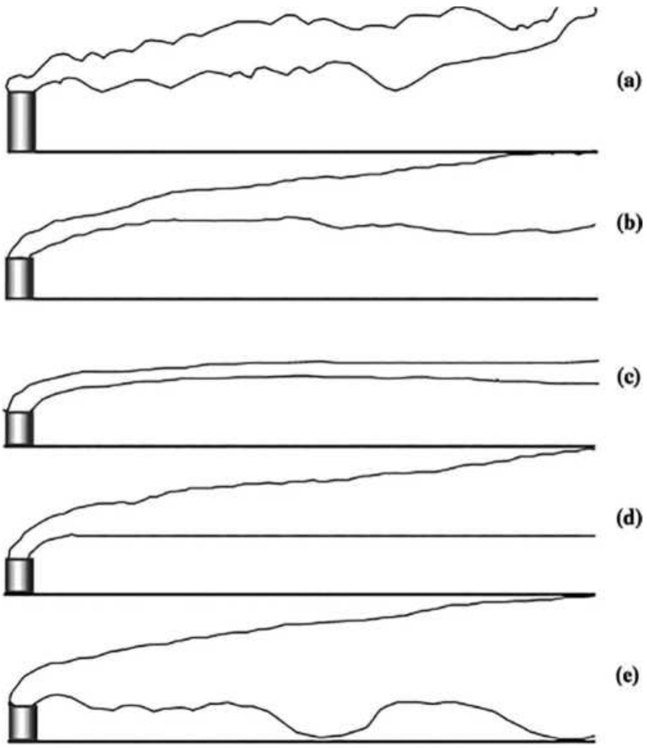


FIGURE 1 Stack gas behavior under various conditions. (a) Strong lapse (looping); (b) weak lapse (coning); (c) inversion (fanning); (d) inversion below, lapse aloft (lofting); (e) lapse below, inversion aloft (fumigation).

TABLE 2 Various Atmospheric Conditions

Strong lapse (looping)	Environmental lapse rate > adiabatic lapse rate
Weak lapse (coning)	Environmental lapse rate < adiabatic lapse rate
Inversion (fanning)	Increasing temperature with height
Inversion below, lapse aloft (lofting)	Increasing temperature below, App. adiabatic lapse rate aloft
Lapse below, inversion aloft (fumigation)	App. adiabatic lapse rate below, increasing temperature aloft

Source: Jørgensen.[10]

TABLE 3 Characteristics of Particulate Pollution Control Equipment

Device	Optimum Particle Size (μm)	Optimum Concentration(g–3)	Temperature Limitations (°C)	Air Resistance (mm H_2O)	Efficiency (% by Weight)
Settling chambers	>50	>100	–30 to 350	<25	<50
Centrifuges	>10	>30	–30 to 350	<50–100	<70
Multiple centrifuges	>5	>30	–30 to 350	<50–100	<90
Filters	>0.3	>3	–30 to 250	>15–100	>99
Electrostatic precipitators	>0.3	>3	–30 to 500	<20	<99
Wet scrubbers	>2–10	>3–30	0 to 350	>5–25	<95–99

Source: Jørgensen.[10]

Settling Chambers

Simple gravity settling chambers depend on gravity or inertia for the collection of particles. Both forces increase in direct proportion to the square of the particle diameter, and the performance limit of these devices is strictly governed by the particle settling velocity. The pressure drop in mechanical collectors is low to moderate, 1–25 cm water in most cases. Most of these systems operate dry, but if water is added, it performs the secondary function of keeping the surface of the collector clean and washed free of particles.

The settling or terminal velocity can be described by the following expression, which has general applicability:

$$V_t = \left(\partial_p - \partial\right) g \frac{d_p^2}{17\mu} \tag{2}$$

where V_t is terminal velocity, ∂_p is particle density, ∂ is gas density, d_p is particle diameter, and μ is gas viscosity.

This is the Stokes' law and is applicable to $N_{Re} < 1.9$,

Where

$$N_{Re} = d_p^* V_t^* \frac{\partial}{\mu} \tag{3}$$

The intermediate equation for settling can be expressed as:

$$V_t = \frac{0.153 * g^{0.71} * d_p^{1.14} \left(\partial_p - \partial\right)^{0.71}}{\partial^{0.29} * \mu^{0.43}} \tag{4}$$

This equation is valid for Reynolds numbers between 1.9 and 500, while the following equation can be applied for $N_{Re} > 500$ and up to 200,000:

$$V_t = 1.74 \left(d_p^* g \frac{\left(\partial_p - \partial\right)^{1/2}}{\partial} \right) \tag{5}$$

The settling velocity in these chambers is often in the range 0.3–3 m/sec. This implies that for large volumes of emission, the settling velocity chamber must be very large in order to provide an adequate residence time for the particles to settle. Therefore, the gravity settling chambers are not generally used to remove particles smaller than 100 µm (= 0.1 mm). For particles measuring 2–5 µm, the collection efficiency will most probably be as low as 1% –2%. A variation of the simple gravity chamber is the baffled separation chamber. The baffles produce a shorter settling distance, which means a shorter retention time. The shown equations can be used to design a settling chamber.

Cyclones

Cyclones separate particulate matter from a gas stream by transforming the inlet gas stream into a confined vortex. The mechanism involved in cyclones is the continuous use of inertia to produce a tangential motion of the particles towards the collector walls. The particles enter the boundary layer close to the cyclone wall and lose kinetic energy by mechanical friction. The forces involved are the centrifugal force imparted by the rotation of the gas stream and a drag force, which is dependent on the particle density, diameter, shape, etc. A hopper is built at the bottom. If the cyclone is too short, the maximum force will not be exerted on some of the particles, depending on their size and the corresponding drag forces.[2] If, however, the cyclone is too long, the gas stream might reverse its direction and spiral up the center.

It is therefore important to design the cyclone properly. The hopper must be deep enough to keep the dust level low. The efficiency of a cyclone is described by a graph similar to Figure 2, which shows the efficiency versus the relative particle diameter, i.e., the actual particle diameter divided by D_{50}, which is defined as the diameter corresponding to 50% efficiency. D_{50} can be found from the following equation:

$$D_{50} = K * \left(\frac{\mu Dc}{V_c * \partial_p} \right)^{1/2} \tag{6}$$

where D_c is the diameter of cyclone, V_c is inlet velocity, ∂p is the density of particles, μ is gas viscosity, and K is a constant dependent on cyclone performance.

If the distribution of the particle diameter is known, it is possible to calculate the total efficiency from a graph such as Figure 2:

$$\text{eff} = \sum \text{m}_i^* \text{eff}_i \tag{7}$$

where m_i is the weight fraction in the ith particle size range and eff_i is the corresponding efficiency.

The pressure drop for cyclones can be found from:

$$\Delta p = N^* \frac{V_c^2}{2_g} \tag{8}$$

From these equations, it can be concluded that higher efficiency is obtained without increased pressure drop if D_c can be decreased while velocity V_c is maintained. This implies that a battery of parallel coupled small cyclones will work more effectively than one big cyclone. Such cyclone batteries are available as blocks and are known as multiple cyclones. Compared with settling chambers, cyclones offer higher removal efficiency for particles below 50 µm and above 2–10 µm, but involve a greater pressure drop.

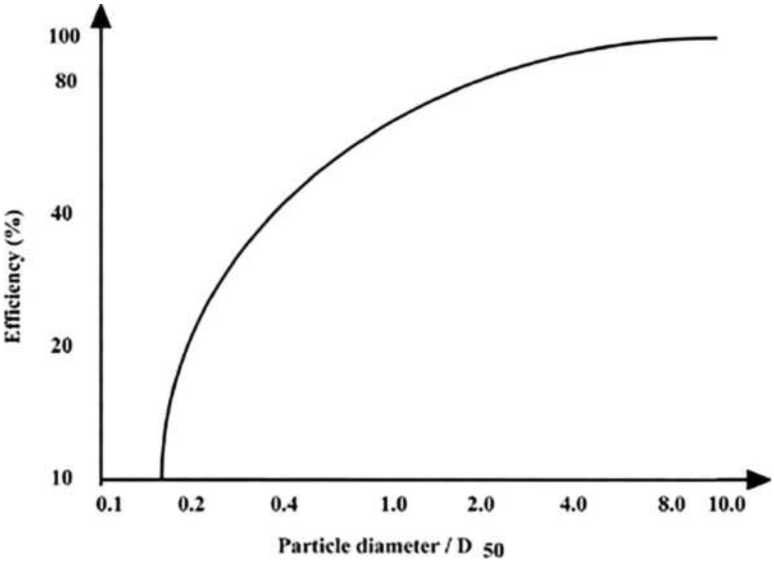


FIGURE 2 Efficiency plotted against relative particle diameter. Notice that it is a log–log plot.

Filters

Particulate materials are collected by filters by the following three mechanisms:

Impaction, where the particles have so much inertia that they cannot follow the streamline around the fiber and thus impact on its surface.

Direct interception, where the particles have less inertia and can barely follow the streamlines around the obstruction.

Diffusion, where the particles are so small (below 1 µm) that their individual motion is affected by collisions on a molecular or atomic level. This implies that the collection of these fine particles is a result of random motion.

Different flow patterns can be used, as demonstrated in Figure 3. The types of fibers used in fabric filters range from natural fibers, such as cotton and wool, to synthetics (mainly polyesters and nylon), glass, and stainless steel.

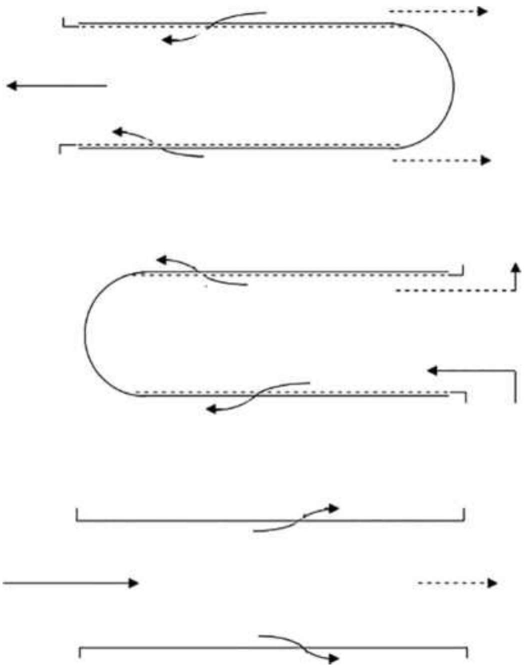


FIGURE 3 Flow pattern of filters.

TABLE 4 Properties of Fibers

Fabric	Acid Resistance	Alkali	Fluoride Strength	Tensile	Abrasion Resistance
Cotton	Poor	Good	Poor	Medium	Very good
Wool	Good	Poor	Poor	Poor	Fair
Nylon	Poor	Good	Poor	Good	Excellent
Acrylic	Good	Fair	Poor	Medium	Good
Polypropylene	Good	Fair	Poor	Very good	Good
Orlon	Good	Good	Fair	Medium	Good
Dacron	Good	Good	Fair	Good	Very good
Teflon	Excellent	Excellent	Good	Good	Fair

Some properties of common fibers are summarized in Table 4. As can be seen, cotton and wool have a low temperature limit and poor alkali and acid resistance, but they are relatively inexpensive. The selection of filter medium must be based on the answer to several questions:[3,4]

What is the expected operating temperature?

Is there a humidity problem that necessitates the use of a hydrophobic material, such as, e.g., nylon?

How much tensile strength and fabric permeability are required?

How much abrasion resistance is required?

Permeability is defined as the volume of air that can pass through 1 m^2 of the filter medium with a pressure drop of no more than 1 cm of water.

The filter capacity is usually expressed as cubic meter of air per square meter of filter per minute. A typical capacity ranges between 1 and 5 $m^3/m^2/min$.

The pressure drop is generally larger than for cyclones and will in most cases be 10–30 cm of water, depending on the nature of the dust, the cleaning frequency, and the type of cloth.

There are several specific methods of filter cleaning. The simplest is backwash, where dust is removed from the bags merely by allowing them to collapse. This is done by reverting the airflow through the entire compartment. The method is remarkable for its low consumption of energy. Shaking is another low-energy filter-cleaning process, but it cannot be used for sticky dust. The top of the bag is held still and the entire tube sheath at the bottom is shaken. The application of blow rings involves reversing the airflow without bag collapse. A ring surrounds the bag; it is hollow and supplied with compressed air to direct a constant stream of air into the bag from the outside.

The pulse and improved jet cleaning mechanism involves the use of a high-velocity, high-pressure air jet to create a low pressure inside the bag and induce an outward airflow, cleaning the bag by sudden expansion and reversal of flow.

In some cases, as a result of electrostatic forces, moisture on the surface of the bag, and a slight degree of hygroscopicity of the dust itself, the material forms cakes that adhere tightly to the bag. In this case, the material must be kept drier, and a higher temperature on the incoming dirty airstream is required. Filters are highly efficient even for smaller particles (0.1–2 m), which explains their wide use as particle collection devices.

Electrostatic Precipitators

The electrostatic precipitator consists of four major components:

1. *A gas-tight shell with hoppers* to receive the collected dust, inlet and outlet, and an inlet gas distributor
2. *Discharge electrodes*
3. *Collecting electrodes*
4. *Insulators*

The principles of electrostatic precipitators are illustrated in Figure 4. The dirty airstream enters a filter, where a high, 20–70 kV, usually negative voltage exists between discharge electrodes. The particles accept a negative charge and migrate towards the collecting electrode. The efficiency is usually expressed by use of Deutsch's equation (see the discussion including correction of this equation in the work of Gooch and Francis[5]). This equation entails that the relationship between migration velocity and particle diameter has a minimum between 0.1 and 1.0 μm.

The operation of an electrostatic precipitator can be divided into three steps:

1. The particles *accept a negative charge.*
2. The charged particles *migrate towards the collecting electrode* due to the electrostatic field.
3. The collected dust *is removed from the collecting electrode* by shaking or vibration, and is collected in the hopper.

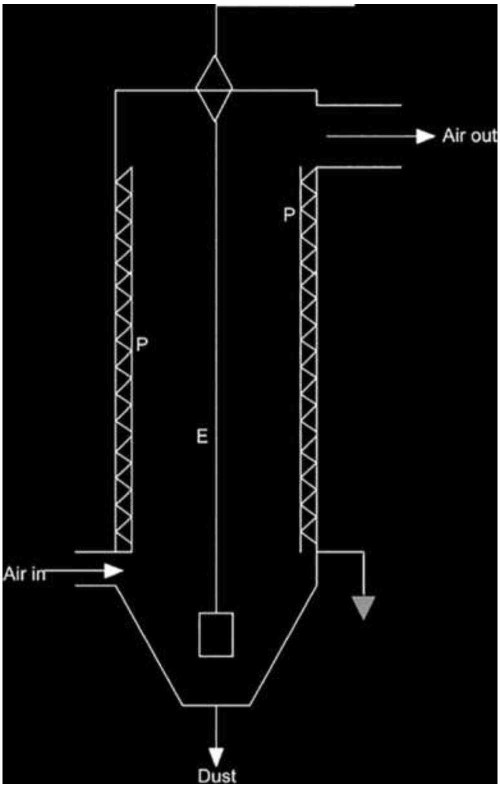


FIGURE 4 The dust is precipitated on the electrode P. E has a high, usually negative voltage and emits a great number of electrons that give the dust particles a negative charge. The dust particles will therefore be attracted to P.

r, the specific electrical resistance, measured in ohm meter, determines the ability of a particle to accept a charge. The practical specific resistance can cover a wide range of about four orders of magnitude, in which varying degrees of collection efficiencies exist for different types of particles. The specific resistance depends on the chemical nature of the dust, the temperature, and the humidity.

Electrostatic precipitators have found wide application in industry. As the cost is relatively high, the airflow should be at least 20,000 m^3/hr; volumes as large as 1,500,000 m^3/hr have been treated in one electrostatic precipitator.

Very high efficiencies are generally achieved in electrostatic precipitators and emissions as low as 25 mg/m^3 are quite common. The pressure drop is usually low compared with other devices—25 mm water at the most. The energy consumption is generally 0.15–0.45 Wh/m^3/hr.

Wet Scrubbers

A scrubbing liquid, usually water, is used to assist separation of particles, or a liquid aerosol from the gas phase. The operational range for particle removal includes material less than 0.2 μm in diameter up to the largest particles that can be suspended in air. Gases soluble in water are also removed by this process. Four major steps are involved in the collection of particles by wet scrubbing. First, the particles are moved to the vicinity of the water droplets, which are 10–1000 times larger. Then, the particles must collide with the droplets. In this step, the relative velocity of the gas and the liquid phases is very important: If the particles have an excessively high velocity in relation to liquid, they cannot be retained by the droplets unless they can be wetted and thus incorporated into the droplets. The last step is the removal of the droplets containing dust particles from the bulk gas phase. Scrubbers are generally very flexible. They are able to operate under peak loads or reduced volumes and within a wide temperature range.[6] They are smaller and less expensive than dry particulate removal devices, but the operating costs are higher. Another disadvantage is that the pollutants are not collected but transferred into water, which means that the related water pollution problem must also be solved.[7]

Several types of wet scrubbers are available and their principles are outlined below:

1. *Chamber scrubbers* are spray towers and spray chambers that can be either round or rectangular. Water is injected under pressure through nozzles into the gas phase.
2. *Baffle scrubbers* are similar to a spray chamber but have internal baffles that provide additional impingement surfaces. The dirty gas is forced to make many turns to prevent the particles from following the airstream.
3. *Cyclonic scrubbers* are a cross between a spray chamber and a cyclone. The dirty gas enters tangentially to wet the particles by forcing its way through a swirling water film onto the walls. There, the particles are captured by impaction and are washed down the walls to the sump. The saturated gas rises through directional vanes, which are used solely to impart rotational motion to the gas phase. As a result of this motion, the gas goes out through a demister for the removal of any included droplets.
4. *Submerged orifice scrubbers* are also called gas induced scrubbers. The dirty gas is accelerated over anaerodynamic foil to a high velocity and directed into a pool of liquid. The high velocity impact causes the large particles to be removed into the pool and creates a tremendous number of spray droplets with a high amount of turbulence. These effects provide intensive mixing of gas and liquid and thereby a very high interfacial area. As a result, reactive gas absorption can be combined with particle removal.
5. The *ejector scrubber* is a water jet pump (see Figure 5). The water is pumped through a uniform nozzle and the dirty gas is accelerated by the action of the jet gas. The result is aspiration of the gas into the water by the Bernoulli principle and, accordingly, a lowered pressure. The ejector scrubber can be used to collect soluble gases as well as particulates.
6. The *venturi scrubber* involves the acceleration of the dirty gas to 75–300 m/min through a mechanical constriction. This high velocity causes any water injected just upstream of or in the venturi throat to be sheared off the walls or nozzles and atomized. The droplets are usually 5–20 µm in size and form

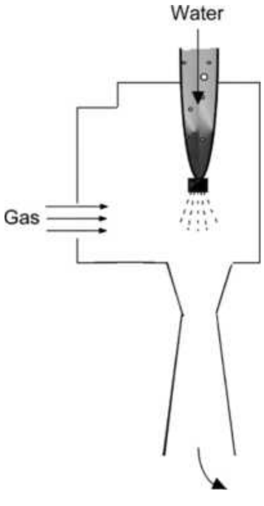


FIGURE 5 Principle of ejector scrubber.

into clouds from 150 to 300 μm in diameter, depending on the gas velocity. The scrubber construction is similar to that of the ejector scrubber, but the jet pump is replaced by a venturi constriction.

7. *Mechanical scrubbers* have internal rotating that which break up the scrubbing liquid into small droplets and simultaneously create turbulence.
8. *Charged-droplet scrubbers* have a high voltage ionization section where the corona discharge produces air ions (as in electrostatic precipitators). Water droplets are introduced into the chamber by the use of spray nozzles or similar devices. The additional collection mechanism provided by the induction of water droplets increases the collection efficiency.
9. *Packed-bed scrubbers* have a bottom support grid, and a top retaining grid (see Figure 6). The fluid (often water or a solution of alkali or acid) is distributed as shown in the figure over the top of the packed section, while the gas enters below the packing.

The flow is normally counter current. Packed-bed scrubbers offer the possibility of combining gas absorption with the removal of particulate material. The pressure drop is often in the order of 3 cm water per meter of packing. If the packing consists of expanded fiber, the bed scrubber is known as a fiber-bed scrubber. The packed-bed scrubber has a tendency to clog under high particulate loading, which is its major disadvantage.

Common packing includes saddles, rings, etc., like those used in absorption towers. Some important parameters for various scrubbers are plotted in Figure 7, which demonstrates the relationship between pressure drop, energy consumption, and D_{50} (the diameter of the particles removed at 50% efficiency).

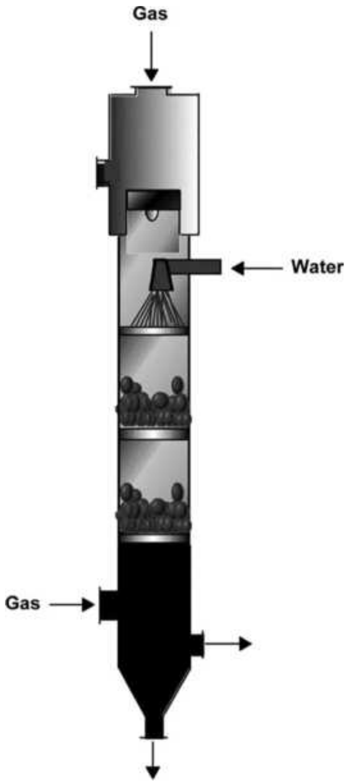


FIGURE 6 Packed-bed scrubber.

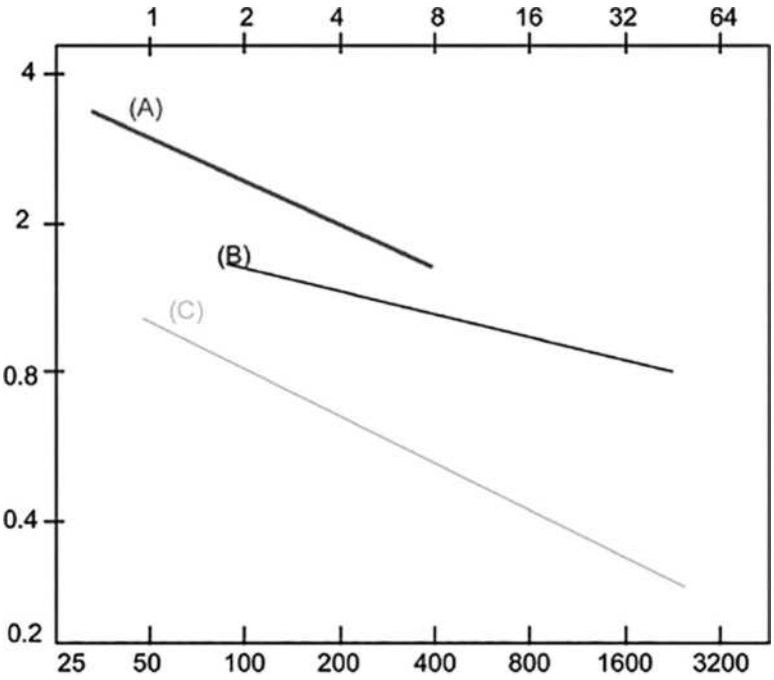


FIGURE 7 Relationship between D_{50} (μm), pressure drop (mm H_2O, lower axis), and energy consumption (kW/m^3/sec; upper axis). A: Packed-bed scrubber. B: Baffled scrubber. C: Venturi scrubber.

Air Pollution Problems of Carbon Hydrides and Carbon Monoxide: Sources of Pollutants

All types of fossil fuel will produce carbon dioxide on combustion, which is used in the photosynthetic production of carbohydrates. As such, carbon dioxide is harmless and has no toxic effect, whatever the concentration levels, but see the entries about the greenhouse effect of carbon dioxide. An increased carbon dioxide concentration in the atmosphere will increase absorption of infrared radiation and the heat balance of the earth will be changed.

Carbon hydrides are the major components of oil and gas, and incomplete combustion will always involve their emission. Partly oxidized carbon hydrides, such as aldehydes and organic acids, might also be present.

The major source of carbon hydrides pollution is motor vehicles.

In reaction with nitrogen oxides and ozone, they form so-called photochemical smog, which consists of several rather oxidative compounds, such as peroxyacyl nitrates and aldehydes. In areas where solar radiation is strong and the atmospheric circulation is weak, the possibility of smog formation increases as the processes are initiated by ultraviolet radiation.

Incomplete combustion produces carbon monoxide. By regulation of the ratio of oxygen to fuel, more complete combustion can be obtained, but the emission of carbon monoxide cannot be totally avoided.

Motor vehicles are also a major source of carbon monoxide pollution. On average, 1 L of gasoline (petrol) will produce 200 L of carbon monoxide, while it is possible to minimize the production of this pollutant by using diesel instead of gasoline.

The annual production of carbon monoxide is more than 200 million tons, of which 50% is produced by the United States alone.

In most industrial countries, more than 75% of this pollutant originates from motor vehicles.

Pollution Problem of Carbon Hydrides and Carbon Monoxide

Carbon hydrides, partly oxidized carbon hydrides, and the compounds of photochemical smog are all more or less toxic to man, animals, and plants. Photochemical smog reduces visibility, irritates the eyes, and causes damage to plants, with immense economic consequences, for example, for fruit and tobacco plantations. It is also able to decompose rubber and textiles.

Carbon monoxide is strongly toxic as it reacts with hemoglobin and thereby reduces the blood's capacity to take up and transport oxygen. Ten percent of the hemoglobin occupied by carbon monoxide will produce symptoms such as headache and vomiting. It should be mentioned here that smoking also causes a higher carboxyhemoglobin concentration. An examination of policemen in Stockholm has shown that non-smokers had 1.2% carboxyhemoglobin, while smokers had 3.5%.

Control Methods Applied to Carbon Dioxide, Carbon Hydrides, and Carbon Monoxide Pollution

Carbon dioxide pollution is inevitable with the use of fossil fuels. Therefore, it can only be solved by the use of other sources of energy.

Legislation plays a major role in controlling the emission of carbon hydrides and carbon monoxide. As motor vehicles are the major source of these pollutants, control methods should obviously focus on the possibilities of reducing vehicle emission. The methods available today are as follows:

1. Motor technical methods
2. Afterburners
3. Alternative energy sources

The first method is based upon a motor adjustment according to the relationship between the composition of the exhaust gas and the air/fuel ratio. A higher air/fuel ratio results in a decrease in the carbon hydrides and carbon monoxide concentrations, but to achieve this, a better distribution of the fuel in the cylinder is required, which is only possible through the construction of another gasification system. This method may be considered cleaner technology.

At present, two types of afterburners are in use—*thermal* and *catalytic afterburners*. In the former type, the combustible material is raised above its autoignition temperature and held there long enough for complete oxidation of carbon hydrides and carbon monoxide to occur. This method is used on an industrial scale[8,9] when low-cost purchased or diverted fuel is available; in vehicles, a manifold air injection system is used.

Catalytic oxidation occurs when the contaminant-laden gas stream is passed through a catalyst bed, which initiates and promotes oxidation of the combustible matter at a lower temperature than would be possible in thermal oxidation. The method is used on an industrial scale for the destruction of trace solvents in the chemical coating industry. Vegetable and animal oils can be oxidized at 250–370°C by catalytic oxidation. The exhaust fumes from chemical processes, such as ethylene oxide, methyl methacrylate, propylene, formaldehyde, and carbon monoxide can easily be catalytically incinerated at even lower temperatures. The application of catalytic afterburners in motor vehicles presents some difficulties due to poisoning of the catalyst with lead. With the decreasing lead concentration in gasoline, it is becoming easier to solve that problem, and the so-called double catalyst system is now finding wide application. This system is able to reduce nitrogen oxides and oxidize carbon monoxide and carbon hydrides simultaneously. New catalysts are currently coming on the market and offer a higher efficiency.

Lead in gasoline has been replaced by various organic compounds to increase the octane number. Benzene has been applied, but it is toxic and causes air pollution problems because of its high vapor pressure. MTBE (methyl tertiary butyl ether) is another possible compound for increasing the octane number. It is, however, very soluble and has been found as a groundwater contaminant close to gasoline stations.

Application of alternative energy sources is still at a preliminary stage. The so-called Sterling motor is one alternative, as it gives more complete combustion of the fuel, but there have been several improvements of the efficiency of the motor during the last years due to the increasing cost of fossil fuel. Much interest has, however, been devoted to electric and hybrid vehicles.

Air Pollution of Sulfur Dioxide: Sources, Problems, and Solutions

Fossil fuel contains approximately 2%–5% sulfur, which is oxidized by combustion to sulfur dioxide. Although fossil fuel is the major source, several industrial processes produce emissions containing sulfur dioxide, for example, mining, the treatment of sulfur containing ores, and the production of paper from pulp. The total global emission of sulfur dioxide has been decreasing during the last 25 years due to the installation of pollution abatement equipment, particularly in North America, the European Union, and Japan. The concentration of sulfur dioxide in the air is relatively easy to measure, and sulfur dioxide has been used as an indicator component. High values recorded by inversion are typical.

Sulfur dioxide is oxidized in the atmosphere to sulfur trioxide, which forms sulfuric acid in water. Since sulfuric acid is a strong acid, it is easy to understand that sulfur dioxide pollution indirectly causes the corrosion of iron and other metals and is able to acidify aquatic ecosystems.[10]

The health aspects of sulfur dioxide pollution are closely related to those of particulate pollution. The gas is strongly adsorbed onto particulate matter, which transports the pollutant to the bronchi and lungs. There is a clear relationship between concentration, effect, and exposure time, which is reflected in the emission standards for sulfur dioxide (see Table 5).

Clean Air Acts were introduced in all industrialized countries during the 1970s and 1980s. Table 5 illustrates some typical sulfur dioxide emission standards, although these may vary slightly from country to country.

The approaches used to meet the requirements of the acts as embodied in the standards can be summarized as follows:

1. Fuel switching from high to low sulfur fuels.
2. Modification of the distribution pattern—use of tall stacks.
3. Abandonment of very old power plants that have a particular high emission.
4. Flue gas cleaning.

Desulfurization of liquid and gaseous fuel is a well-known chemical engineering operation. In gaseous and liquid fuels, sulfur either occurs as hydrogen sulfide or reacts with hydrogen to form hydrogen sulfide. The hydrogen sulfide is usually removed by absorption in a solution of alkanol- amine and then converted to elemental sulfur. The process in general use for this conversion is the so-called Claus process. The hydrogen sulfide gas is fired in a combustion chamber in such a manner that one-third of the volume of hydrogen sulfide is converted to sulfur dioxide. The products of combustion are cooled and then passed through a catalyst-packed converter, in which the following reaction occurs:

$$2H_2S + SO_2 = 3S + 2H_2O \tag{9}$$

TABLE 5 SO_2 Emission Standards

Duration	Concentration (ppm)	Comments
Month	0.05	
24 hr	0.10	Might be exceeded once a month
30 min	0.25	Might be exceeded 15 times/month

Source: Jorgensen.[10]

The elemental sulfur has commercial value and is mainly used for the production of sulfuric acid.

Sulfur occurs in coal both as pyritic sulfur and organic sulfur. Pyritic sulfur is found in small discrete particles within the coal and can be removed by mechanical means, e.g., gravity separation methods. However, 20%–70% of the sulfur content of coal is present as organic sulfur, which can hardly be removed today on an economical basis. Since sulfur recovery from gaseous and liquid fuels is much easier than that from solid fuel (which has other disadvantages as well), much research has been and is being devoted to *the gasification or liquefaction of coal.* It is expected that this research will lead to an alternative technology that will solve most of the problems related to the application of coal, including sulfur dioxide emission. Approach (2) listed above has been mentioned earlier in this entry, while approach (3) needs no further discussion. The next subsection is devoted to (4) flue gas cleaning.

Flue Gas Cleaning of Sulfur Dioxide

When sulfur is not or cannot be economically removed from fuel oil or coal prior to combustion, removal of sulfur oxides from combustion gases will become necessary for compliance with the stricter air pollution control laws.

The chemistry of sulfur dioxide recovery presents a variety of choices and five methods should be considered:

1. Adsorption of sulfur dioxide on active metal oxides with regeneration to produce sulfur.
2. Catalytic oxidation of sulfur dioxide to produce sulfuric acid.
3. Adsorption of sulfur dioxide on charcoal with regeneration to produce concentrated sulfur dioxide.
4. Reaction of dolomite or limestone with sulfur dioxide by direct injection into the combustion chamber. A lime slurry is injected into the flue gas beyond the boilers.
5. Fluidized bed combustion of granular coal in a bed of finely divided limestone or dolomite maintained in a fluid-like condition by air injection. Calcium sulfite is formed as a result of these processes.

In particular, the two latter methods have found wide application, particularly to large industrial installations. It is possible to recover the sulfur dioxide or elemental sulfur from these processes, making it possible to recycle the spent sorbing material.

Air Pollution Problems of Nitrogenous Gases: Sources, Problems, and Control

Seven different compounds of oxygen and nitrogen are known: N_2O, NO, NO_2, NO_3, N_2O_3, N_2O_4, and N_2O_5— often summarized as NO_x. From the point of view of air pollution, it is mainly NO (nitrogen oxide) and NO_2 (nitrogen dioxide) that are of interest. Nitrogen oxide is colorless and is formed from the elements at high temperatures. It can react further with oxygen to form nitrogen dioxide, which is a brown gas. The major sources of the two gases are combustion of gasoline and oil (nitrogen oxide) and combustion of oil, including diesel oil (nitrogen dioxide). The production of NO is favored by high temperature. In addition, a relatively small emission of nitrogenous gases originates from the chemical industry. The total global emission is approximately 10 million tons per year. This pollution has only local or regional interest, as the natural global formation of nitrogenous gases in the upper atmosphere by the influence of solar radiation is far more significant than the man-controlled emission.

As mentioned above, nitrogen oxide is oxidized to nitrogen dioxide, although the reaction rate is slow—in the order of 0.007/hr. However, it can be accelerated by solar radiation. Nitrogenous gases take part in the formation of smog, as the nitrogen in peroxyacyl nitrate originates from nitrogen oxides. They are highly toxic but as their contribution to global pollution is insignificant, local and regional problems can partially be solved by changing the distribution pattern (see the section on *Control Methods Applied to Particulate Pollution).*

The emission from motor vehicles can be reduced by the same methods as mentioned for carbon hydrides and carbon monoxide. The air/fuel ratio determines the concentration of pollutants in exhaust gas. An increase in the ratio will reduce the emission of carbon hydrides and carbon monoxide, but unfortunately will increase the concentration of nitrogenous gases. Consequently, the selected air/ fuel ratio will be a compromise. A double catalytic afterburner is applied today, and it is able to reduce nitrogenous gases and simultaneously oxidize carbon hydrides and carbon monoxide. The application of alternative energy sources will, as for carbon hydrides and carbon monoxide, be a very useful control method for nitrogenous gases at a later stage.

Between 0.1 and 1.5 ppm of nitrogenous gases, of which 10%–15% consists of nitrogen dioxide, are measured in urban areas with heavy traffic. On average, the emission of nitrogenous gases is approximately 15 g per liter of gasoline and 25 g per liter of diesel oil.

Nitrogenous gases in reaction with water form nitrates that are washed away by rainwater. In some cases, this can be a significant source of eutrophication. For a shallow lake, for example, the increase in nitrogen concentration due to the nitrogen input from rainwater will be rather significant. In a lake with a depth of 1.7 m and an annual precipitation of 600 mm, which is normal in many temperate regions, the annual input will be as much as 0.3 mg/L.

The methods used for control of industrial emission of nitrogenous gases, including ammonia, will be discussed in the next section that discusses industrial air pollution, but as pointed out above, industrial emission is of less importance, even though it might play a significant role locally. The emission of nitrogenous gases by combustion of oil for heating and the production of electricity can hardly be reduced.

Industrial Air Pollution, Overview, and Control Methods

The rapid growth in industrial production during recent decades has exacerbated the industrial air pollution problem, but due to increased application of continuous processes, recovery methods, air pollution control, use of closed systems, and other technological developments, industrial air pollution has, in general, not increased in proportion to production.

Industry displays a wide range of air pollution problems related to a large number of chemical compounds in a wide range of concentrations.

It is not possible in this context to discuss all industrial air pollution problems; instead, we shall touch on the most important problems and give an overview of the control methods applied today. Only the problems related to the environment will be dealt with in this context.

A distinction should be made between air quality standards, which indicate that the concentration of a pollutant in the atmosphere at the point of measurement should not be greater than a given amount, and emission standards, which require that the amount of pollutant emitted from a specific source should not be greater than a specifically indicated amount.

The standards reflect, to a certain extent, not only the toxicity of the particular component but also the possibility for its uptake.

Here, the distribution coefficient for air/water (blood) plays a role. The more soluble the component is in water, the greater the possibility for uptake. For example, the air quality standard for acetic acid, which is very soluble in water, is relatively lower than the toxicity of aniline, which is almost insoluble in water.

Since industrial air pollution covers a wide range of problems, it is not surprising that all three classes of pollution control methods mentioned previously have found application: modification of the distribution pattern, alternative (cleaner) production methods, and particulate and gas/vapor control technology.

All the methods mentioned in the section on *Control Methods Applied to Particulate Pollution* also apply for industrial air pollution control.

In gas and vapor technology, a distinction has to be made between condensable and non-condensable gaseous pollutants. The latter must usually be destroyed by incineration, while the condensable gases can be removed from industrial effluents by absorption, adsorption, condensation, or combustion.

Recovery is feasible by the first three methods.

Gas Absorption

Absorption is a diffusion process that involves the mass transfer of molecules from the gas state to the liquid state along a concentration gradient between the two phases. Absorption is a unit operation that is enhanced by all the factors generally affecting mass transfer, i.e., high interfacial area, high solubility, high diffusion coefficient, low liquid viscosity, increased residence time, turbulent contact between the two phases, and possibilities for reaction of the gas in the liquid phase. This last factor is often very significant and almost 100% removal of the contaminant is the result of such a reaction. Acidic components can easily be removed from gaseous effluents by absorption in alkaline solutions, and, correspondingly, alkaline gases can easily be removed from effluent by absorption in acidic solutions (Table 6).

Carbon dioxide, phenol, and hydrogen sulfide are readily absorbed in alkaline solutions in accordance with the following processes:

$$+CO_2 + 2N_aOH \rightarrow 2N_a^+ + CO_3^{2-} \tag{10}$$

$$H_2S + 2NaOH \rightarrow 2N_a^+ + S_{2-} + 2H_2O \tag{11}$$

$$C_6H_5OH + NaOH \rightarrow C_6H_5O^- + N_a^+ + H_2O \tag{12}$$

Ammonia is readily absorbed in acidic solutions:

$$NH_3 + H_2SO_4 \rightarrow 2NH_4 + SO_4^{2-} \tag{13}$$

Gas Adsorption

Adsorption is the capture and retention of a component (adsorbate) from the gas phase by the total surface of the adsorbing solid (adsorbent). In principle, the process is the same as when dealing with wastewater treatment; the theory is equally valid for gas adsorption.

Adsorption is used to concentrate (often 20 to 100 times) or store contaminants until they can be recovered or destroyed in the most economical way. Figure 8 illustrates some adsorption isotherms applicable to practical gas adsorption problems. These are often described as either Langmuir's or Freundlich's adsorption isotherms. Adsorption is dependent on temperature: increased temperature

TABLE 6 Absorber Reagents

Reagents	Applications
$KMnO_4$	Rendering, polycyclic organic matter
NaOCl	Protein adhesives
Cl_2	Phenolics, rendering
Na_2SO_3	Aldehydes
NaOH	CO_2, H_2S, phenol, Cl_2, pesticides
Ca(OH)2	Paper sizing and finishing
H_2SO_4	NH_3, nitrogen bases

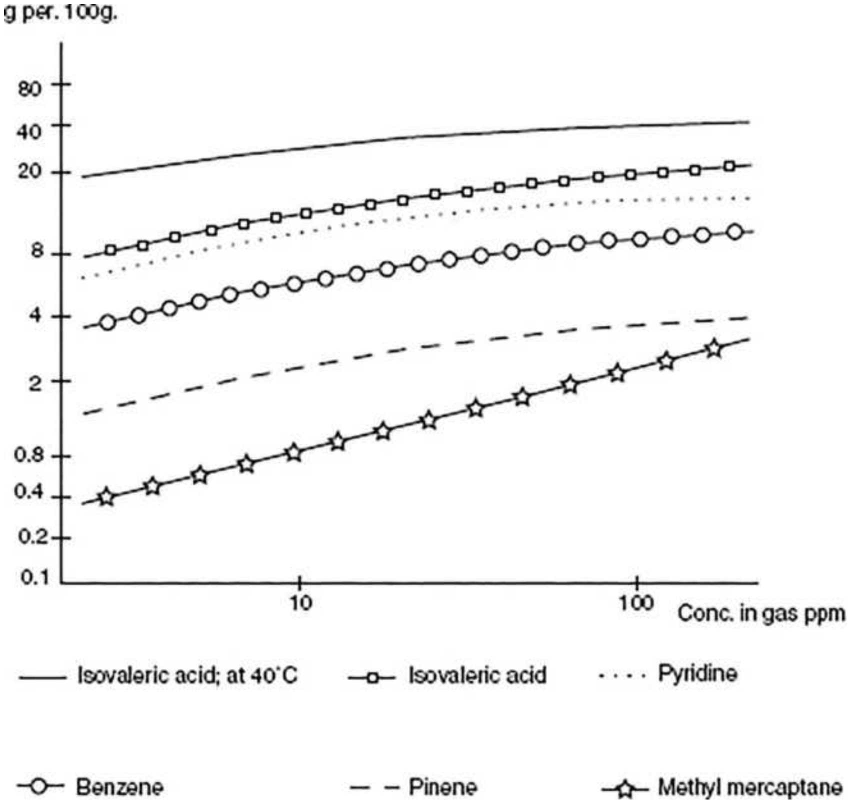


FIGURE 8 Adsorption isotherms, at 20°C.

means that the molecules move faster and therefore it is more difficult to adsorb them. There are four major types of gas adsorbents, the most important of which is activated carbon, but aluminum oxide (activated aluminum), silica gel, and zeolites are used as well.

The selection of adsorbent is made according to the following criteria:

1. High selectivity for the component of interest.
2. Easy and economical to regenerate.
3. Availability of the necessary quantity at a reasonable price.
4. High capacity for the particular application so that the unit size will be economical. Factors affecting capacity include total surface area involved, molecular weight, polarity activity, size, shape, and concentration.
5. Pressure drop, which is dependent on the superficial velocity.
6. Mechanical stability in the resistance of the adsorbent particles to attrition. Any wear and abrasion during use or regeneration will lead to an increase in bed pressure drop.
7. Microstructure of the adsorbent should, if at all possible, be matched to the pollutant that has to be collected.
8. The temperature, which has a profound influence on the adsorption process, as already mentioned.

Regeneration of the adsorbents is an important part of the total process. A few procedures are available for regeneration:

1. *Stripping* by use of steam or hot air.
2. *Thermal desorption* by raising the temperature high enough to boil off all the adsorbed material.

3. *Vacuum desorption* by reducing the pressure enough to boil off all the adsorbed material.
4. *Purge gas stripping* by using a non-adsorbed gas to reverse the concentration gradient. The purge gas may be condensable or non-condensable. In the latter case, it might be recycled, while the use of a condensable gas has the advantage that it can be removed in a liquid state.
5. *In situ oxidation* based on the oxidation of the adsorbate on the surface of the adsorbent.
6. *Displacement* by use of a preferentially adsorbed gas for the desorption of the adsorbate. The component now adsorbed must, of course, also be removed from the adsorbent, but its removal might be easier than that of the originally adsorbed gas, for instance, because it has a lower boiling point.

Although the regeneration is 100%, the capacity of the adsorbent may be reduced 10%–25% after several regeneration cycles, due to the presence of fine particulates and/or high molecular weight substances that cannot be removed in the regeneration step. A flowchart of solvent recovery using activated carbon as an adsorbent is shown in Figure 9 as an illustration of a plant design.

Combustion

Combustion is defined as rapid, high-temperature gas-phase oxidation. The goal is the complete oxidation of the contaminants to carbon dioxide and water, sulfur dioxide, and nitrogen dioxide.

The process is often applied to control odors in rendering plants, paint and varnish factories, rubber tire curing, and petrochemical factories. It is also used to reduce or prevent an explosion hazard by burning any highly flammable gases for which no ultimate use is feasible. The efficiency of the process is highly dependent not only on temperature and reaction time but also on turbulence or the mechanically induced mixing of oxygen and combustible material. The relationship between the reaction rate, *r,* and the temperature can be expressed by Arrhenius' equation:

$$r = A^*e - \frac{E}{RT} \tag{14}$$

where A is a constant, E is the activation energy, R is the gas constant, and T is the absolute temperature. A distinction is made between combustion, thermal oxidation, and catalytic oxidation, the latter two being the same in principle as the vehicle afterburners.

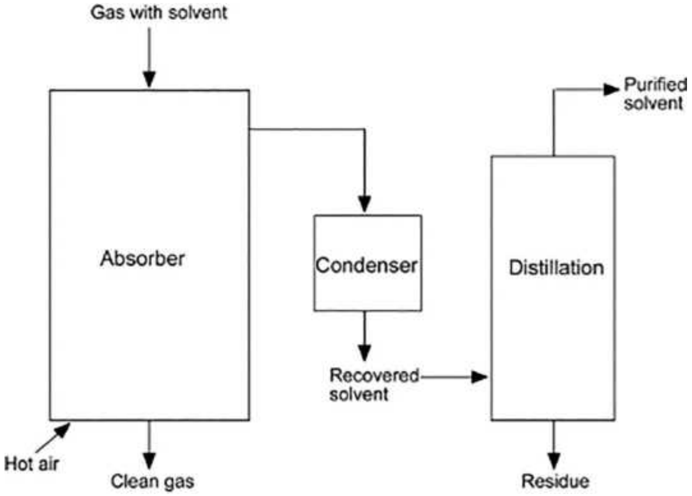


FIGURE 9 Flow chart of solvent recovery by the use of activated carbon.

Heavy Metals as Air Pollutants

Heavy metals, which may be defined as the metals with a specific gravity >5.00 kg/L, comprise 70 elements. Most of them are, however, only rarely found as pollutants. The heavy metals of environmental interest form very heavy soluble compounds with sulfide and phosphate and form very stable complexes with many ligands present in the environment. It means, fortunately, that most of the heavy metals are not very bioavailable in most environments (see also *Bioremediation,* p. 408).

A number of enzymes activated by metal ions and metalloenzymes are known. Members of the first mentioned group, comprising iron, cobalt, chromium, vanadium, selenium, copper, zinc, iron, cobalt, and molybdenum, are able with a stronger bond to form metalloenzymes: metalloproteins, metalloporphyrins, and metalloflavins.

Heavy metals are emitted to the atmosphere by energy production and a number of technological processes (see Table 7). It makes the atmospheric deposition of heavy metals originating from human activities the dominant pollution source for the vegetation of natural ecosystems— forests, wetlands, peat lands, and so on. The heavy metal content in sludge and fertilizers plays a more important role for agricultural land where also the inputs of heavy metals by irrigation, natural fertilizers, and application of chemicals including pesticides may add to the overall pollution level. The atmosphere and hydrosphere both have a well- developed ability for "self-purification"—for heavy metals by removal processes, for instance, sedimentation. The lithosphere has a high buffer capacity toward the effects of most pollutants and also has an ability to self-purify, for instance, by runoff and uptake by plants, although the rates usually are much lower than in the two other spheres. Table 8 illustrates the removal rates.

Heavy metals are bound to clay particles due to their ion-exchange capacity and to hydratized metal oxides, such as iron sesquioxide (As, Cr, Mo, P, Se, and V) and manganese sesquioxides (Co, Ba, Ni, and lanthanides). Calcium phosphate is further better able to bind As, Ba, Cd, and Pb in alkaline soil. Fulvic acids (molecular weight about 1000) and humic acid (molecular weight about 150,000) are able to form complexes with a number of heavy metals, Hg(II), Cu(II), Pb(II), and Sn(II). The mobility of heavy metals is dependent on a number of factors. The soil pore water contains soluble organic compounds (acetic acid, citric acid, oxalic acid, and other organic acids), partly excreted by the roots. These small

TABLE 7 Important Atmospheric Pollution Sources of Heavy Metals

Source	Heavy Metals
Incineration of oil	V, Ni
Incineration of coal	Hg, V, Cr, Zn, As
Gasoline	Pb (leaded gasoline)
Metal industry	Fe, Cu, Mn, Zn, Cr, Pb, Ni, Cd, and others
Application of pesticides	Hg, Cr, Cu, As
Incineration of solid waste	Hg, Zn, Cd, and others

Source: Jørgensen.[10]

TABLE 8 Removal of Heavy Metals by Runoff and Drainage from a Typical Cultivate Clay Soil

Metal	Removal (mg/m^2/yr)	Removal % of Pool 1–3
Pb	0.5	1–3
Cu	1.2	2–3
Zn	15.9	30–50
Cd	0.07	15–30

Source: Waid.[8]

organic molecules form chelated, soluble compounds with metal ions such as Al, Fe, and Cu. Activity of living organisms in soil may also enhance the mobility of heavy metal ions. Fungi and bacteria may utilize phosphate and thereby release cations. Formation of insoluble metal sulfide under anaerobic conditions from sulfate implies a reduced mobility. The lower oxidation stages of heavy metals are generally more soluble than the higher oxidation stages, implying increased mobility.

The many possibilities of binding heavy metals in soil explain the long residence time. Cadmium, calcium, magnesium, and sodium have the most mobile metal ions with a residence time of about 100 years. Mercury has a residence time of about 750 years, while copper, lead, nickel, arsenic, selenium, and zinc have residence times of more than 2000 years under temperate conditions. Tropic residence times are typically lower (for all heavy metals, about 40 years).

The biological effect of heavy metal pollution occurs in accordance with Sections 4.4 and 4.5 on two levels: on an organism level and on the higher level—the ecosystem level.

Plant toxicity is very dependent on the presence of other metal ions. For instance, Rb and Sr are very toxic to many plants, but the presence of the biochemically more useful K and Ca is able to reduce or eliminate toxicity. The toxicity of arsenate and selenate can be reduced in the same manner by sulfate and phosphate.

Formation of complexes by reaction with organic ligands also reduces toxicity due to reduced bioavailability. The plant toxicity of heavy metals in soil is consequently also correlated with the concentration of heavy metal ions in the soil solution.

The heavy metals that are most toxic to plants are silver, beryllium, copper, mercury, tin, cobalt, nickel, lead, and chromium. With the exception of silver and chromium, the divalent form is most toxic. For silver, it is Ag^+, and for chromium, it is chromate and dichromate that are most toxic. Silver and mercury ions are very toxic to fungus spores, and copper and tin ions are very toxic to green algae; lethal concentrations may be as low as 0.002–0.01 mg/L.

One of the key processes on ecosystem level is the mineralization process, because it determines the cycling of nutrients. Heavy metals can inhibit the mineralization due to the blocking of enzymes. The effect is known not only for the enzymes produced in the organisms but also for extracellular enzymes—exoenzymes—originated from dead cells or excreted from roots and living microorganisms. As the various processes forming the cycling of nutrients are coupled, the entire mineralization cycle is disturbed if only one process is reduced. It is therefore possible to determine the change of the mineralization cycle by measuring the respiration, the transformation of nitrogen, and the release of phosphorus. As low a concentration of copper as 3–4 times the background concentration may imply a reduced soil respiration. A few hundred milligrams of copper per kilogram of soil is furthermore able to diminish the nitrogen release rate by one half.

The most sensitive mineralization process is phosphorus cycling. Biological material binds phosphorus as esters of phosphoric acid. The phosphate is released by the hydrolysis of the ester bond, a process catalyzed by phosphatase. This process is inhibited by the presence of heavy metals. The inhibition is decreasing in the following sequence: molybdenum (VI) > wolframate (VI) > vanadate (V) > nickel (II) > cadmium > mercury (II) > copper (II) > chromate (VI) > arsenate (V) > lead (II) > chromium (III).

The inhibition of exo-enzymes by heavy metals does not form a clear pattern. It is therefore difficult to generalize. Most experiments, however, give a clear picture of the influence of heavy metals on mineralization: the rate of mineralization may be reduced significantly with a consequent reduction of the productivity of the entire ecosystem.

In Denmark (a country with relatively little heavy industry and good pollution control), atmospheric deposition causes an average annual increase of the total content of heavy metal in soil between 0.4% and 0.6%, but it varies very much from location to location. In accordance with the many possibilities for side reactions of heavy metals in soil, including adsorption to the soil particles, the amount of heavy metal ions that are available to plants is only a fraction of the total content. If only the bioavailable heavy metals are used as the basis, the annual percentage increase in the soil concentration due to atmospheric deposition is probably higher.

Most lead in soil is not mobile and cannot be transported via the root system to the leaves and stems. This is in contrast to cadmium, which is very mobile. About 50% of the cadmium in soil will be found in the plants after the growth season, although the concentration may be very different in different parts of the plants. The cadmium in grains for instance has not increased parallel to the increased atmospheric deposition of cadmium.

The heavy metal pollution of soil is one of the major challenges in environmental management in industrialized countries. Due to the many diffuse sources of heavy metal pollution, the solution of the problem requires a wide spectrum of methods, the first of which is application of cleaner technology (see the section on *Industrial Air Pollution, Overview, and Control Methods*). It is in other words necessary to reduce the total emission of heavy metals. Dilution (for instance, higher chimneys) is not an applicable solution. Moreover, as pollution, particularly air pollution, has no borders, it is necessary to take international initiatives and agree on international standards, particularly for the most problematic heavy metals, i.e., cadmium, mercury, nickel, chromium, and vanadium. A three-point program must be adopted:

- A nationally and internationally accepted environmental strategy.
- Agreed international standards and long-term goals.
- A monitoring program to assess the pollution level and compare the measured concentrations with standards.

References

1. Reible, D.D. *Fundamentals of Environmental Engineering*;; Lewis Publ.: Boca Raton, New York, Washington, DC, 1998; 526 pp.
2. Leith, D.; Licht, W. The collection of cyclone type particles—A new theoretical approach. AICHE Symp. Ser. **1975**, *68,* 196–206.
3. Pring, R.T. Speciation considerations for fabric collectors. Pollut. Eng. **1972**, *4,* 22–24.
4. Rullman, D.H. Backhouse technology: A perspective. J. Air Pollut. Control Assoc. **1976**, *26,* 16–18.
5. Gooch, N.P.; Francis, N.L. A theoretically based mathematical model for calculations of electrostatic precipitators performance. J. Air Pollut. Control Assoc. **1975**, *25,* 106–113.
6. Onnen, J.H. Wet scrubbers tackle pollution. Environ. Sci. Technol. **1972**, 6, 994–998.
7. Hanf, E.B. A guide to a scrubber selection. Environ. Sci. Technol. **1970**, *4*, 110–115.
8. Waid, D.E. Controlling pollutants via thermal incineration. Chem. Eng. Prog. **1972**, *68*, 57–58.
9. Waid, D.E. *Thermal Oxidation or Incineration*; Pollut. Control Assoc.: Pittsburgh, PA, 1974; 62–79.
10. Jørgensen, S.E. *Principles of Pollution Abatement*; Elsevier: Amsterdam, 2000; 520 pp.

23

Alternative Energy: Hydropower

Andrea Micangeli,
Sara Evangelisti, and
Danilo Sbordone

Introduction

Approximately 70% of the earth's surface is covered with water, a resource that has been exploited for many centuries. Hydropower is currently the most common and the most important renewable energy source: throughout the world, it produces 3288 TWh, just over 17% of global production and the 84% of energy produced by renewable energy sources[1] from an installed capacity of about 850 GW.[2]

The International Energy Agency (IEA) has developed a number of scenarios that describe the efforts needed to reduce carbon dioxide emissions. The "business-as-usual" baseline scenario foreshadows the situation in the absence of policy change and major supply constraints leading to increases in oil demand and CO_2 emissions. The "**BLUE**" scenario is the most ambitious, bringing emissions at 50% of the 2005 level by 2050. This implies of course higher investment costs, as well as greater needs in technological and policy developments. In *Energy Technology Perspectives 2010*, it states that hydro could produce up to 6000 TWh in 2050.

The main characteristics that make hydropower a successful energy source are its plant storage capacity and fast responses to meet sudden fluctuations in electricity demand. Global hydropower generation has increased by 50% since 1990, with the highest absolute growth in China, as shown in Figure 1.

IEA estimates the *global technically exploitable hydropower potential* (the *technically exploitable potential* is the annual energy potential of all natural water flows that can be exploited within the limits of current technologies[3]) at more than 16,400 TWh per year.[1] However, hydroelectric plants of big dimensions, with million cubic meter water basins, have negatively affected the natural and social environment of the territories. Small plants are characterized by a different management, distributed on the territory, managed in small communities, integrated in the multiple and balanced use of the water resources.

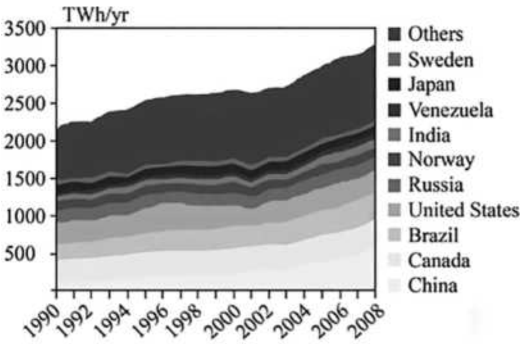


FIGURE 1 Evolution of global hydropower generation, 1990–2008.
Source: International Energy Agency.[1]

The contribution of *small hydropower* (*SHP*) *plants*, defined as those with installed capacity of up to 10 MW, to the worldwide electrical supply is about 1%–2% of the total one, amounting to about 61 GW.[2] Europe with about 13 GW installed capacity has the second biggest contribution to the world's installed capacity, just behind Asia. Moreover, the SHP potential is estimated in 180,000 MW.

SHP has a key role to play in the development of renewable energy resources and an even greater role in developing countries. In the face of increasing electricity demand, international agreements to reduce greenhouse gases, environmental degradation from fossil fuel extraction and use, and the fact that, in many countries, large hydropower sites have been mostly exploited, there is an increasing interest in developing SHP. Indeed, SHP has a huge, as yet largely untapped potential, which will enable it to make a significant contribution to future energy needs, offering a very good alternative to conventional sources of electricity, not only in the developed world but also in developing countries.

A hydropower sector technological maturity has already been reached during the last century, but only big plants have received all the benefits from technological development, while those of smaller dimension have been neglected. Nowadays the economy of scale, social and environmental implications, suggest this solution due to their economical feasibility and environmental respectful, allowing sustainable distributed production with an easy installation and great applicability in developing countries.

This entry is organized as follows: first, a classification of the hydropower plants is given. Then, the basics of the technology of mini and micro-hydro plants are illustrated, together with a description of the main civil works that occurred in a hydropower scheme. Finally, a conclusion on the potential and shortcomings of the hydropower technology is drawn.

Classification of the Hydroelectrical Plants

Hydropower plants can be generally classified in terms of power outputs:

Micro-hydro plant, with a nominal power lower than 100 kW, subdivided into *low-head plants,* when the vertical drop is lower than 50 m, and *low-flow rate plants,* when the water flow is lower than 10 m^3/sec.

Mini-hydro plant, with a nominal power between 100 kW and 1000 kW, subdivided into *mini-head plants*, when the vertical drop is between 50 and 250 m, and *mini-flow rate plants*, if the water flow is between 10 and 100 m^3/sec.

Small hydro plants, with a nominal power between 100 kW and 10 MW, subdivided into *medium-head plants,* when the drop is between 250 and 1000 m, and *medium flow-rate plants*, when the water flow is between 100 and 1000 m^3/sec.

Big hydro plants, as shown in Figure 2, with a nominal power of more than 10 MW, defined as *high-head plants* if the drop is higher than 1000 m and as *high-flow rate plants* with a water flow of more than 1000 m^3/sec.

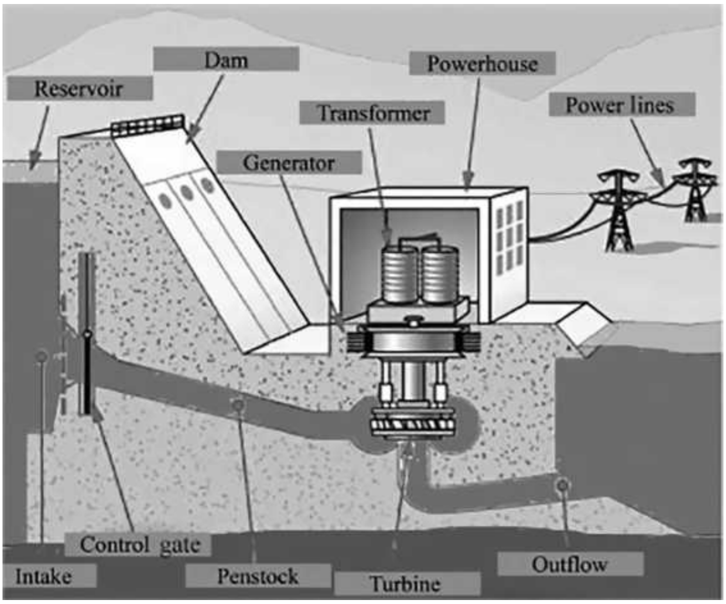


FIGURE 2 Big hydro storage scheme.

Another important way to classify hydroelectric plants is on the basis of their typology. In particular, they can be classified as follows:

Run-of-river scheme: or fluent water plants, they take a portion of a river through a canal or penstock. They do not require the use of a dam or catch basin. Because of that, they aim to affect upstream water levels and downstream stream flow less than any other power plants. Electricity generation from these plants could change in the amount of water flowing in the river.

Storage scheme: an impound water behind a dam, as a reservoir. Water is released through turbine generators to produce electricity. The water storage and release cycles can be relatively short, for instance, storing water at night for daytime power generation, or the cycles can be long, storing spring runoff for generation in the summer, when air conditioner use increases power demand. Some projects operate on multiyear cycles carrying over water in a wet year to offset the effects of dry years.

Pumped-storage scheme: these plants use off-peak electricity to pump water from a lower reservoir to an upper reservoir. During periods of high electrical demand, the water is released back to the lower reservoir to generate electricity.

The following sections focus on mini and micro-hydropower plants and run-of river scheme.

Mini-Hydro Plants

Mini-hydropower (MHP) plants have an installed power lower than 1 MW actually in Europe is 3 MW but in many country (in particular USA 5 MW) it may be more. Generally, these plants need less civil works, consistently reducing the costs connected to the realization of the plant and justifying their realization also under an economic point of view. If the plants are well planned and placed, their environmental impact is reduced for their limited dimensions. The simplicity of construction allows them to be introduced in contexts where the technology of the sector is not yet developed and there is a strong need for mechanical or electrical power. Also in those cases, MHP plants can be operated and maintained locally, even with less-specialized technicians. Changing hydropower plant size or typology, many things are the same, such as a turbine installation. This entry focuses on mini-hydro due to its low environmental impact and its opportunities of developing in the future.

Water Resource

Hydraulic energy, as almost every forms of energy on the earth, comes from the sun, which is the "engine" of the hydrological cycle. The sun, irradiating and warming up the atmosphere, makes seas and lakes evaporate; the water vapor rises up and thickens the clouds that move because of the wind, also generated by the sun; the clouds then produce precipitations in the form of snow, hail, and rain. When the rainfall ends up in the natural basin situated at a higher level, energy is transformed to *potential energy*. This energy is naturally stocked in rivers and in creeks that flow into the sea, closing in this way the hydrological cycle (see Figure 3).

The amount of available energy, which the water basin can produce at a given height, comes from the water level reached at the end of the cycle. In other words, to know the potential energy of a basin, it is necessary to evaluate the available rise, depending on the orography of the territory and on specific water works such as dams or small barriers.

The amount of water available is defined as the mass of water flowing per time unit (*flow rate*). In general, the potential power from a reservoir can be calculated by the following equation:

$$P_0 = \rho \cdot g \cdot Q \cdot H_0 \tag{1}$$

where

P_0 = theoretical power (W)
ρ = water density (1000 kg/m^3)
g = gravitational acceleration (9.8 m/sec^2)
Q = flow rate (m^3/sec)
H_0 = net head (m)

The output power of the plant is a percentage of it, due to mechanical, electrical, and friction losses.

Hydrology and Rain Measurements

In order to exploit the energy of water for power purposes, a *hydro geological analysis* of the territory is needed, particularly for catch basins. The analysis is based on the evaluation of the supply of the basin and its outflow. To obtain a balanced catch basin, it has to take into consideration the *meteoric flow rate,*

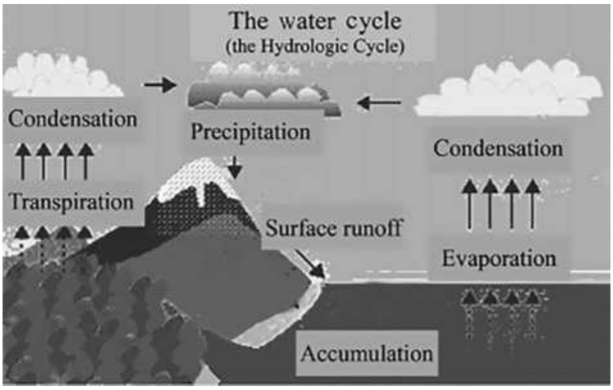


FIGURE 3 Hydrological cycle. (Source: Harvey et al.[4])

evaporation, and both *superficial* and *underground circulation*. The quantity of water from the rainfall to the basin, the *meteoric supply*, must be evaluated not only on the water surface of the basin itself but also on the whole area in which the rainfall is collected as well as the streams towards the basin (see Equation 2). The flows, which depend on soil permeability, are essentially of two types: *superficial* and *underground*, i.e., when the water filters through the soil and supplies underground basins and water-bearing stratum.

It is possible to evaluate the *meteoric supply*—superficial flow, as:

$$P = E + D + (I - C) \tag{2}$$

where P is the *meteoric supply* to the basin, E is the contribution given to the evaporation, D is the outflows, and I and C are the increase and decrease of the basins, respectively.

The term E is the amount of the following different contributions:

- Evaporation of water from the soil
- Transpiration of plants
- Evaporation of water intercepted by vegetation
- Evaporation of internal basins

Similarly, the term in relation to draining—underground flow—can be subdivided into the following:

- Natural water draining underground toward the external (*groundwater*)
- Artificial draining water toward the external (*inversion*)
- Natural superficial water inflow from the external (*water flows*)
- Underground natural water inflow (*water-bearing stratum*)
- Artificial inflow from the external (*adduction*)

Joining of different rivers has to be considered as well. The evaluation of the meteoric intake is usually performed through specialized devices, such as *rain gauges,* very common all over the world (see Figure 4).

The intensity of the rainfall flow rate is not constant through time and it has to be referred to different periods of the year (usually a multiyear). Not all the water from the rainfall ends up in the catch basin as

FIGURE 4 Example of a rain gauge.

shown before. Generally, the phenomenon is estimated by introducing a *coefficient of draining,* depending on the waterfall and the water collected into the catch basin, as shown in Equation 3:

$$C = \frac{V}{V_0} \tag{3}$$

where V is the real caught volume and V_0 is the waterfall.

Once V is determined, it is possible to calculate directly the energy exploitable from the plant in a given site:

$$E = 0.00273 \cdot \eta \cdot H_0 \cdot V \tag{4}$$

where H_0 indicates the *net drop*, in other words, the available drop minus the losses in the work of adduction with η output efficiency of the turbine.

Within a natural *hydrological basin,* it is necessary to analyze the head and the flow rate available along the whole river bed, through the *hydrodynamic curve* (see Figure 5). It shows the surfaces of the catch basin on the horizontal axes and the height of the water flow on the vertical one. Through the hydrodynamic curve, it is possible to optimize the entire use of a catch basin, while for the realization of a single plant, without having the intention to optimize the use of the resource along the whole river bed, it is enough to measure the flow rate of the river and the consequent evaluation of the quantity of water that can be taken from the basin or from the available drop.

Measure of the Water Flow

The determination of the water flow of a catch basin can be done by using specific devices, but it has to be undertaken only in absence of historical data of the course along the years. Different methods exist.[4] However, for each of them, it would be necessary to repeat the measures along a period of time to obtain the variation that occur throughout the year.

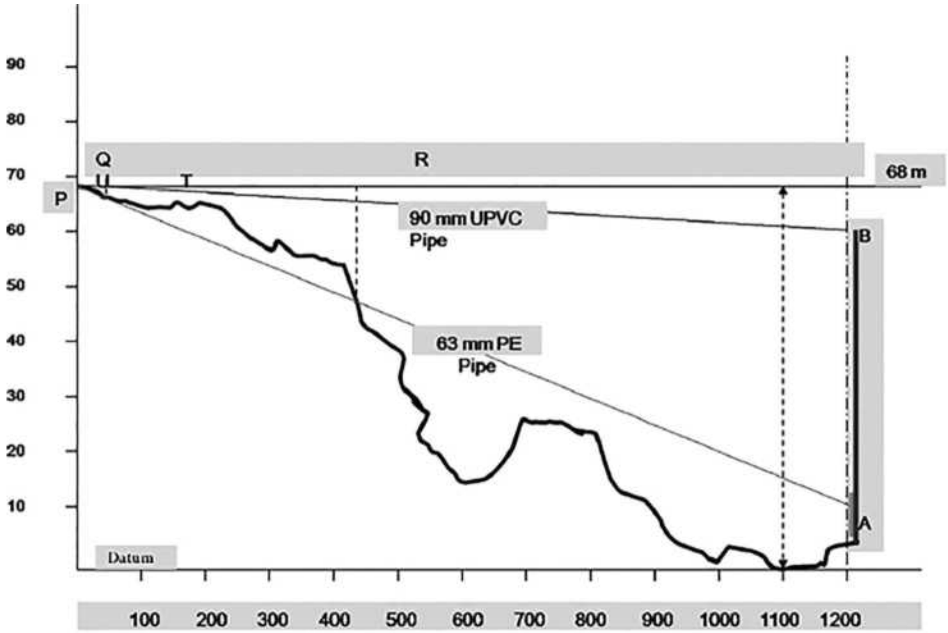


FIGURE 5 Example of a hydrodynamic curve.

One of the simplest methods that can be used is to force the flow to get into a container of known dimensions, measuring the necessary time to fill it up. This method is known for its simplicity; however, it is limited in that it can be applied only in rivers with small water flows.

A second method (Figure 6) consists of the realization of a weir of known dimensions, in which the river is forced to get into—the *weir method*. This method can be used to bring up to around 1 m^3/sec. To measure how much is carried, the second level reached by the water is taken into consideration as illustrated in Figure 7.

The water flow Q can be calculated using:

$$Q = 0.41 \cdot B \cdot H \cdot \sqrt{2 \cdot g \cdot H} \tag{5}$$

In this formula, H can be calculated as the difference between H_2 and H_1 (see Figure 7), B is the width of the weir, and g is the gravitational acceleration.

Another method of measuring water flow involves the evaluation of water velocity on the cross-sectional area. Velocity measurements can be made through a very simple method: with a floating, not too light in order to avoid the friction with the air, in a place where the river is pretty regular and flat, measuring the time that the floating takes to cover a specific distance. The measurement must be repeated more than once and the sought value taken into consideration must be the average between the distance covered and the time spent to cover it. The final velocity must be corrected with a factor between 0.75 and 0.85, given by the losses due to the friction with the sides of the canal.

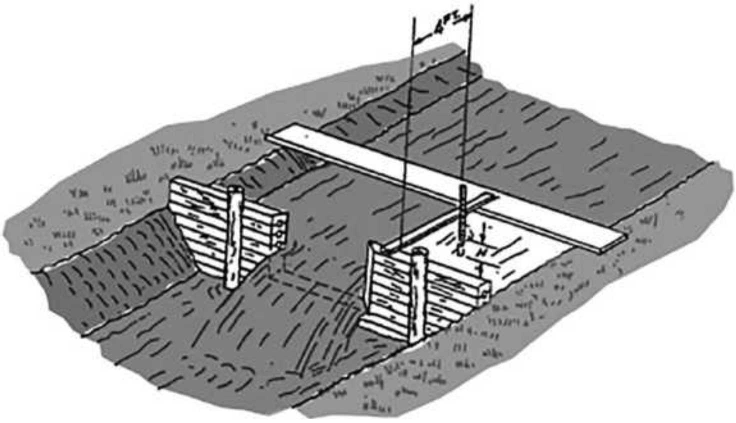


FIGURE 6 Weir method.

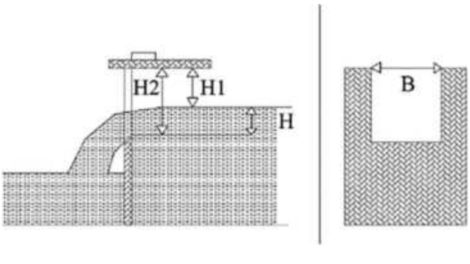


FIGURE 7 Weir realization.

TABLE 1 Value of the *Factor of Manning* (n) for Different River Bed Typologies

River Bed Typology	n
Regular river bed with a minimum annual flow	0.030
Stable flow condition	0.035
River with stagnant water, aquatic vegetation, and meanders	0.045
River with stones and shrubs with shallow pools and lush vegetation	0.060

The velocity can be obtained also through different methods, such as the use of "titled solutions," taking into account the variations on the electricity conductivity of the river when it flows with a known quantity of salt inside.

Once the cross section is evaluated, the water flow can be then calculated through the *formula of Manning*:[5]

$$Q = \frac{A \cdot R^{\frac{2}{3}} \cdot S^{\frac{1}{2}}}{n} \tag{6}$$

where A represents the cross section, R is the hydraulic radius, and S is the slope of the water surface. The value of n can be obtained from Table 1.

Flow Duration Curve

Water flow measurements are always referred to a specific period, as the water of the river changes during the year, passing several times from huge quantities of water to smaller ones. In general, the curve is uneven as it reports the water flow rate throughout the year as shown in Figure 8.

To organize the collected measurement data, it is possible to use another graphic that puts them all together. Indeed, in Figure 8, it is possible to note that there are two evident points of absolute maximum and absolute minimum, corresponding to the maximum quantity of water that occurs for a very short period and the minimum quantity of water that is the quantity available all year. Moreover, Figure 9 shows a *flow duration curve*, which presents the duration of each amount of the flow rate.

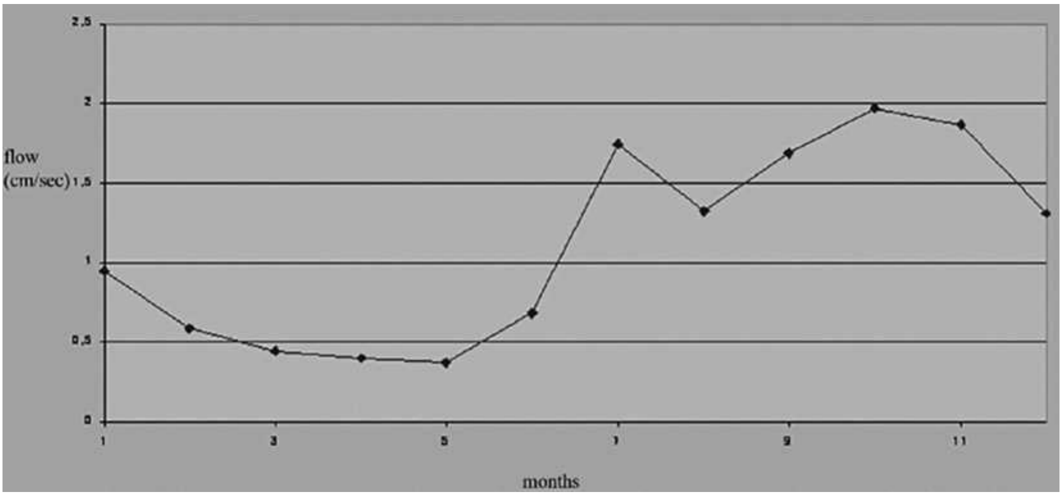


FIGURE 8 Example of a daily flow curve.

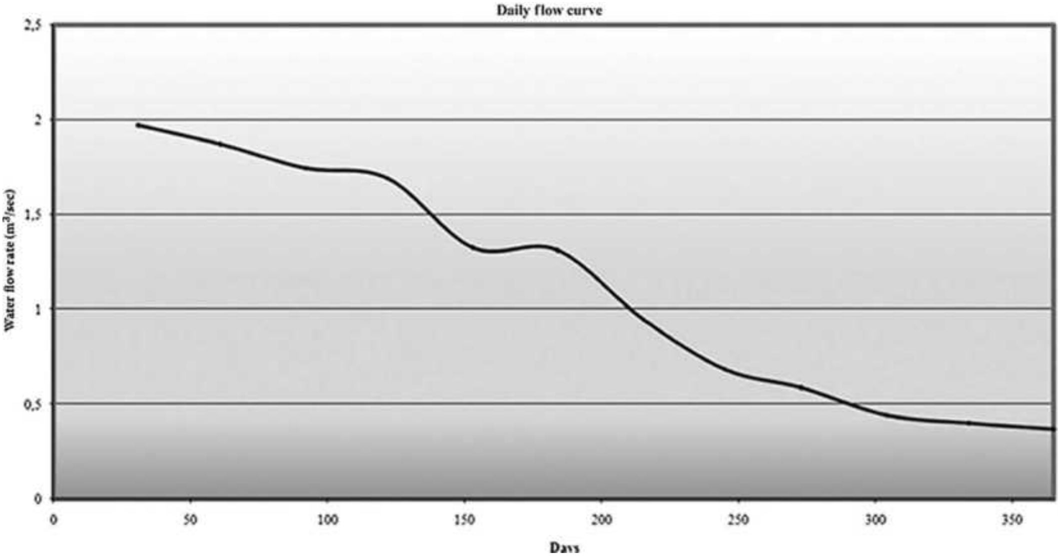


FIGURE 9 Example of a flow duration curve.

Measurements of the Gross Head

To measure the height difference between two points, it is necessary to utilize a level and to follow the scheme in Figure 10. The operator must simply read the values of each ruler to come out with the height by computing the difference between the two values. This procedure can be repeated until the final point is reached. If a level is not available, it is possible to use a table with a carpenter level—even if it requires a lot of patience—or to proceed with a plastic transparent pipe filled up with water, which fulfills the same characteristics of a level. While measuring from the available head, it is wise to also calculate the length of the *forced penstock* as distance from the hold point to the arrival of the penstock itself. This is essential both for the choice of the material of the penstock and for the evaluation of the pressure drop.

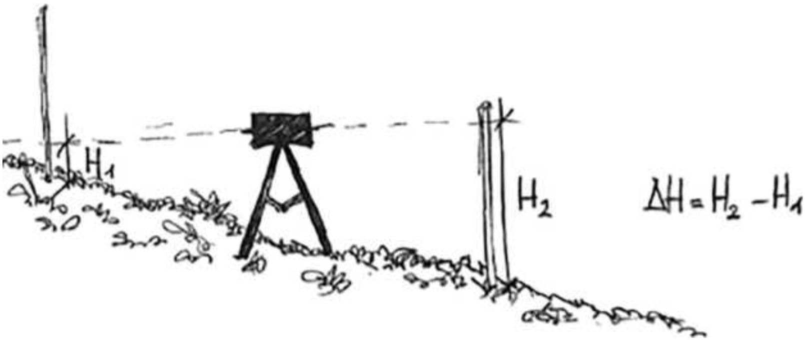


FIGURE 10 Example of measurement of the gross head.

Instream Flow and Environmental Impact

The balance of any catch basin is connected not only to the water balance but also to the real possibility of exploitation, characterized by other aspects:

Rivers can be used not only for power purposes. Before proceeding to the derivation of the outflow, it is necessary to be sure that it will not have a negative impact on further communities. It is necessary to verify all the aspects connected with the multiple uses of the water resource.
The subsistence of the natural balances involved in the river and in the catch basin itself. An example, to adduce water to a riverbed of a catch basin, can be useful for the fish fauna, although the flow rate is high and it comes from nearby.

The use of the water resource of a catch basin refers to a wider issue that generally is approached letting a minimum natural course of the river, generally indicated as *minimum instream flow* (*IF*), defined as "the minimum height of water needed to maintain the values of the basin at an acceptable level." The calculation of the IF is essential: in fact, if the minimum flow rate is lower than the IF during the planning stage, the no-working periods of the plant can be estimated.

To guarantee a minimum flow means to preserve the biological balance and the need of the use of a civil work as a caption downriver. The derivation established on the basis of this context must be lower than the limit beyond which it may influence the river ecosystem and may cause the entering in crisis however the natural water regime must be guaranteed.

The river regime model must therefore take into account the following aspects:

Biological species that would suffer from the uncontrolled derivation of water.
Hydrologic characterization for the protection of water balance (equilibrium) and the defense of soil.
The use of the water resource represented by the social and economic wardship of needs.

Two different relations are commonly used for the calculation of the minimum IF:

$$Q_{\mathrm{IF}} = \frac{15 \cdot \alpha \cdot Q_{\mathrm{media}}}{\left[\ln\left(\alpha \cdot Q_{\mathrm{media}}\right)\right]^2} \tag{7}$$

where α is the *coefficient of perpetuity* given by the relation between low intake and the medium IF values, expressed in liters per second:

$$\mathrm{IF}_{\mathrm{hydrol.}} = 6.\mathrm{IF}_{\mathrm{microhabitat}} \tag{8}$$

The equation shows that conserving the IF from a hydrological point of view also means, with a big margin, that the conservation of IF from a biological point of view is connected with the microhabitat. If the value of the low intake is unknown and only the average is known, it is possible to assume that the coefficient of perpetuity α is equal to 0.24, showing the outflow of hydropower plant that guarantees the preservation of the fish fauna and hydrology. Figure 11 shows a typical example of the realization of a canal to guarantee the conservation of hydrology and of the fish fauna in a hydropower plant.

Civil Works in MHP Plants

The *civil works* for *hydropower facilities* have the functions to capture, to exploit, and to return the water downriver. In small rivers, the realization of those works is done by deflecting the water flow for a short period of time, to operate in dry conditions. To deflect the river, a provisional river bed has to be built or the plant must be realized during the dry season, if feasible.

FIGURE 11 Example of a system to conserve fish fauna.

Dams and Weirs

The choice of the typology of the dam must be made taking into consideration the orography of the territory and the water resource to exploit.[6] Generally, hydraulic works must create a good drop between mountain and valley, to have the possibility to exploit the potential energy from the mass of water. The structure can be *dams* (Figure 12) or *catch basins* (Figure 13).

A *dam* is a "lung" of water available that is able to compensate, partially and totally depending on the volume created, the variation of the flow rate of the river during the year. Although the creation of a dam needs a big investment, this makes it practically not usable in MHP.[7]

A type of barrier—often installed in small plants—is the *derivation weir*. The derivation weir is smaller and more cost-effective than a traditional dam and generally the water overflows the traverse, reversing in the natural path of the river. If the plant is big enough, the derivation weir turns into a small dam that basically does not function like a water accumulator, but it aims at raising the level of the water flow.

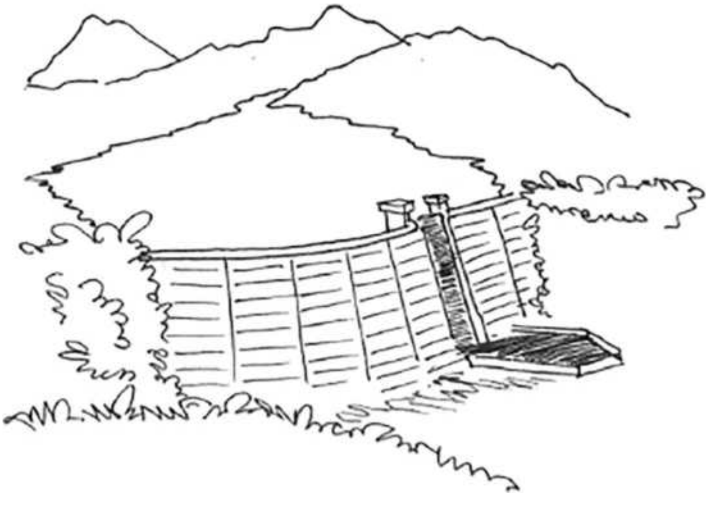


FIGURE 12 Example of a dam.

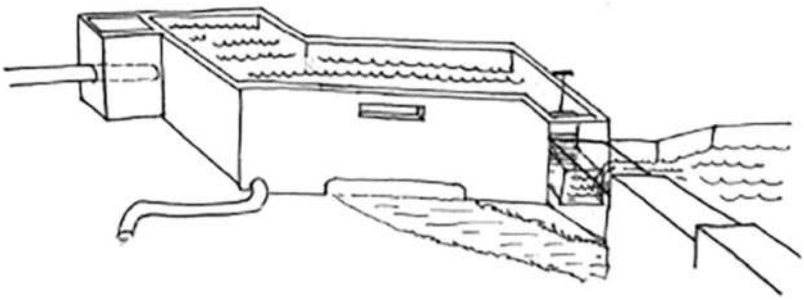


FIGURE 13 Example of a catch basin.

The derivation weir can be realized with several different materials and, when possible, it is suggested that locally available material, such as rocky materials, be used. Otherwise, it is possible to use blocks of flat rocks and soil kept altogether by metal nets to constitute the *barrage.*

It is important that the barrage is realized with a central waterproof nucleus, made with clay, and supported with soil or minerals. If there sand or gravel is available, a barrier made of concrete can be also taken into consideration.

If the regimes of full flow are huge, it is necessary that the intake must be drained through ad hoc dischargers not easy to realize when the barrage has been made in the river. Instead, in places with much seismic activities or in very cold climates, it is better to avoid rigid structures, so floor barriers are preferred.

Stability of the Dam

In small plants, the stability of the dam generally depends on the weight of the dam itself. The strains that the dam applies to the ground through its own weight can be higher than those that the ground can take. It is necessary to minimize the infiltration of water underneath the dam with the use of spillway drain and erosion or diaphragms.

Also, the stability to the overturning of the dam must be guaranteed in all the conditions of charge concerning the hydrostatic push related to the dragging of solid material or to seismic waves. In order to guarantee that, it is enough to ensure that all the strains against the ground are not negative, in order to avoid dangerous situations that can bring the structure to overturn. This entails designing a dam where all the horizontal and vertical strains fall internally in the central part of its own basement (see Figure 14).

As the hydrostatic strains generally go from upstream to downstream, the side of the dam facing downstream is more sloping as compared to the perpendicular of the basement; thus, the profile of the dam upstream has to be more vertical than the downstream.

System of Elevating the Free Water Surface

To control the water flow, *floodgates* are generally used. Given that most of these devices are used to control the free water surface elevation being stored or routed, they are also known as *crest gates.*

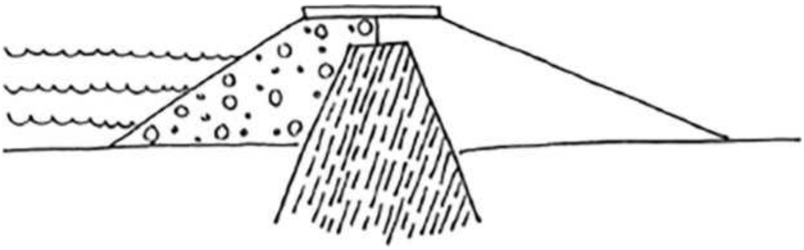


FIGURE 14 Stability of the dam.

A removable type of *floodgates* is *flashbooks gates*— wooden panels usually installed on the cap of the barrier that allow the increase of the water surface and that can be removed during floods, avoiding the inundation of the upstream fields.[8]

Generally, to avoid the manual intervention and to check upstream flooding, it is possible to install a gate that can be opened progressively during full intakes. Another solution is to install *fusegates*, concrete crates that flip over when the water level is reached (Figure 15). Finally, another type is the *dinghy*, anchored to the cap of the barrier that blows up during full intakes of water, only to deflate again during the rest of the year.

Spillway

If the flow rate varies during the year, it could be that the floodgates systems are not enough to guarantee the integrity of the plant. In this case, it is necessary to foresee a system that takes the surplus flow rate to downstream— the *spillway*. Except during flood periods, water does not normally flow over a spillway. The surplus water usually flows at a high velocity and often it is necessary to insert a system to reduce its kinetic energy. In small plants, the introduction of a drainage, which allows emptying the loading tank to help in the maintenance of the plant, is always considered.

The spillway is also used in times of emergency, i.e., shutting down the plant.

Intake

The *water intake* (see Figure 16) is a structure in which water is adduced in order to bring it to the *forebay tank*. It must be able to address in the penstock or in the drainage canal the amount of water estimated in the project. It is pleonastic to underline how the civil work has to be studied and realized in a way to minimize friction losses and the impact on the environment. It must also be designed to minimize maintenance and to reduce the costs. Practical aspects of the project related to civil work therefore must follow the following criterion: Hydraulics and structural works have to guarantee the resistance of the pipes to minimize the waste of energy and to be cost-effective, to avoid transportation of solid material inside the pipes to the powerhouse and for easier maintenance, and to reduce the passage of fishes and to not compromise the ecosystem of the area in which it has been realized.

FIGURE 15 Example of fusegates: “Sant’Antonio 1” IdroPower Station (Italy).

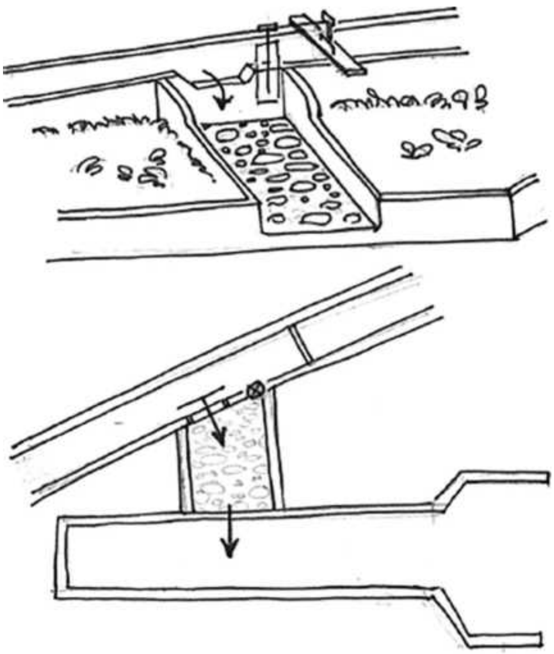


FIGURE 16 Example of intake works.

The water intake works can be realized in different ways, depending on the peculiarity of the orography of the territory. Often it is a channel that brings the water to the point where the forced pipe is. In this case, the channel is realized in a way that the water flows slowly with a contained slope of the pipe in order to contain the losses and the erosion of the walls and to preserve the jump.

The choice of the dimension of the intake channel and the water velocity are results of a compromise to avoid frequent maintenance due to the deposits of sand and slime, to avoid losses and erosion against the wall. Generally, the water flows along the channel with a velocity between 0.1 and 0.4 m/sec. If the quantity of solid materials to be transported is huge, then a bigger tank, where the deposits end up due to the reduction of the speed of the water as a result of a bigger section, is needed, based on the fundamental hypothesis of the continuity of the intake.

The *sedimentation tank* is necessary if the channel has not an open surface, due to the costs connected to the maintenance of the closed channel. In small plants, the drainage tank usually works as a sedimentation tank as well; however it must be cleaned more frequently and it will have a bigger dimension due to the sedimentation process. The transportation of solid material is very deleterious not only for turbine performance but also for the life of the device itself.[5]

The *orientation* of the *intake* is crucial when choosing a project, as it can reduce the accumulation of material over the grid itself and the frequency of the intervention of maintenance. The best position is parallel to the flow letting the full flows the task of removing the material stuck in front of the grid. Anyway, it does not have to be located in areas of stagnant water as the whirlpool and the parasite flows tend to accumulate solid material in front of the grill. If there is a discharger, it is good to place the grill next to it to simplify maintenance as the deposits can be pushed to the discharger as well.

Forebay Tank

At the end of the channel or coincident with the intake, replacing the channel of charge, a little tank, known as *forebay tank* or *basin of charge,* has to be realized. Its function is to guarantee the presence of upstream water in the penstock, in order to avoid the entrance of air along the pipe and the formation of whirlpools.

Channel

If the plant needs it, the sampling of the flow to adduce in the central, is done by using a channel. The channel can be realized both as open channel and as under pressure pipe. In the small plant, the technical solution is oriented towards an open surface channel, the sizing of which is done considering first the intake to derive. The intake is the function of the section of the channel as well as of the slope and the roughness that depends on the material used and on the degree of finishing of the wall. The channel can be made from different materials such as soil, wood, and concrete.

Generally, for small plants where the banks have an inclination of 45° with the base, if the width is L, the width of water surface is $2L$ and the height is $L/2$. Concerning concrete structures, using one that is rectangular shaped, which helps in the cleaning of the channel, is usually preferred.

Penstock and Pressure Drop

The *penstock* (see Figure 17) takes the water from the load tank and pushes it to the turbine. It can be realized with the use of different materials. One should take into account the cost, weight, type of joints, and the conditions of the ground when choosing the type of material to be used. The *penstock* is also characterized by the diameter of the conduit itself that must contain the loss of load.

The choice of the pipe diameter has to be made as a compromise between three needs:

1. Keeping the costs down and therefore realizing a small-diameter pipe.
2. Containing load losses.
3. Realizing a bigger pipe to increase the energy.

The first head measure can be seen as a gross head, keeping into consideration the losses inside the conduits and all the other works of adduction.

The real *head* exploited by the turbine is lower than the first value above. Indeed, the definition of drop goes together with the *net head* that identifies the usable jump by the turbine; thus, the gross head minus the losses of the adduction works. Such a definition allows dividing the losses of the hydraulic parts from those related to the turbine. The value of the net head obviously depends on the pressure

FIGURE 17 Example of a penstock.

drops occurring inside the penstock. Through the definition of *net head*, it is also possible to define the output of the section of the hydraulics work simply as:

$$\eta_{\mathrm{idr}} = \frac{H_{\mathrm{n}}}{H_0} \tag{9}$$

where n and 0 refer to the net head and the gross head, respectively. The bigger is the hydraulic output of the plant, the better will be the exploitation of the water resource as higher power can be obtained with the same load or the same power can be obtained with less load. Moreover, if the entity of the load losses compared to the available drop is modest in high- and very-high-fall plants, in the low-fall plant with 6 m of available drop, a load loss of 1 m is almost 20% of the produced power. Thus, the hydraulic works of an adduction channel have to be realized with focus on MHP plants.

The penstock is the part of the plant in which the water flows faster. Considering that the losses are proportional to the square of the speed, the realization of the penstock is very important in terms of hydraulic performances of work of adduction.

The amount of water that flows inside the penstock is functional to the section of the pipe, its diameter and the water velocity. Once the flow is designed, a relation between speed flooding and penstock diameter is needed.

The problem can be solved with a dimensional analysis that puts into evidence how the *Fanning factor* is a function of the *Reynolds number* and *relative roughness,* known from fluid dynamic theory. The Fanning factor is connected to load losses and it represents their adimensionalization; the Reynolds number comes from the relation between inertia forces, viscosity, and velocity. Relative roughness is connected to the choice of material and the level of superficial finishing. Generally speaking, it is verified that, depending on the fluid regime, the Fanning factor tends to depend only on one of the variables.[8]

Walls

The thickness of the walls and veins are subject to the pressure of the impulse load, which also includes a water hammer. Nevertheless, in the case of a water hammer, plastic pipelines react better than iron ones, because the elasticity of the plastic tends to absorb overpressure better than other materials.

Once the ideal material is selected, the formula of the thickness can be found using the *Mariotte's formula*:[9]

$$t = \frac{P \cdot D}{2\sigma_{\mathrm{f}}} \tag{10}$$

where t is the thickness of the pipe, P stands for the hydrostatic pressure, D is the diameter, and σ is the allowable stress.

Equation 10 is only valid for stationary systems, where both capacity reductions and closure operations are not verified. Moreover, it does not take into account the problems that occur in iron pipes. Therefore, Equation 11 should be amended, and, taking into account the types of joints, it becomes:

$$t = \frac{P \cdot D}{2\sigma_{\mathrm{f}} \cdot k_{\mathrm{f}}} + t \cdot s \tag{11}$$

where k_{f} is the efficiency of the welding and $t \times s$ represents the overpressure due to corrosion. k_{f} can be derived using Table 2.

In general, the value obtained as the thickness is always corrected when it is too low to take into account other factors: the tube must have achieved a sufficient rigidity to be moved without deformation. If the plant has a high fall, then a conduct with variable thicknesses (based on the pressure) can

TABLE 2 Value of k_f for Different Types of Joints

Type of Joint	k_f
Without welding	1
Welding checked with x-ray	0.9
Welding checked with x-ray and subjected to a relaxation	1

be used in order to reduce the cost of the materials. In addition to resistance to pressure increases, a conduct has also to withstand internal depressions to avoid collapsing:

$$P_c = 882.500 \cdot \left(\frac{t}{D}\right)^3 \tag{12}$$

where P_c is the pressure of collapse.

The depressions can be avoided through an *aerophore* with a minimum diameter:

$$d = 7.47\sqrt{\frac{Q}{\sqrt{P_c}}} \tag{13}$$

where d is the diameter of the aerophore.

Finally, to conclude the calculation of the wall thickness, a water hammer has to be considered. The *Allievi-Michaud formula* can be modified if the pressure is expressed in water column, as:

$$\Delta P = c\frac{\Delta V}{g} \tag{14}$$

where c is the propagation speed in the middle of the pressure wave that depends on the water density and the elasticity of the material:

$$c = \sqrt{\frac{k}{\left(1 + \frac{kD}{Et}\right)}} \tag{15}$$

where k is the water cubic compression module (2.1×10^9 MPa); E is Young's modulus of the conducting material; t and D are the thickness and diameter of the tube, respectively; and ρ is the water density. By applying the relationships (Equations 14 and 15) to PVC and iron pipes, it is possible to calculate for an instant closure (d of 400 mm, PVC thickness of 14 mm, and iron thickness of 4 mm):

$$c_{\text{pvc}} = 305 \text{ m/sec}$$

$$c_{\text{acciaio}} = 1023 \text{ m/sec}$$

$$\Delta P_{\text{pvc}} = 123 \text{ m}$$

$$\Delta P_{\text{acciaio}} = 417 \text{ m}$$

This provides a quantitative demonstration of the previously described nature of the two materials. If the operating time increases, the pressure is drastically reduced. Indeed, the maneuvering speed plays a crucial role in the generation of the overpressure. In large systems, it is common not to install pipes that can withstand overpressure, or water hammer, but rather to install mechanisms for the exclusion of the load to prevent the turbine from going into "overspeed."

Another device that serves to absorb the pressure waves that can occur within the pipes is the *piezometric borehole.* To evaluate if its installation is needed, the following formula can be considered:

$$I = \frac{V \cdot L}{gH} \tag{17}$$

where I is the constant acceleration. If I is less than 3, it can be assumed that the piezometric borehole is not necessary.[5]

The penstocks are also anchored to the ground, or supported on special works such as anchor blocks or saddles. The distance between two saddles or between two anchor blocks has to be as much as to make the pipe's arrow acceptable when it is full.

Water Turbines

A turbine converts energy in the form of falling water into rotary shaft power. The selection of the best turbine depends on several factors: the net head of the plant, the nominal flow, the power rating, and the shape of the turbine. The most installed turbine models, mainly used in big hydro plants, are three different typologies: Pelton, Francis, and Kaplan. In MHP, the turbine choice is made on different considerations, not only economical. Indeed, its construction and operational simplicity become essential, especially in developing contexts. Often, it is possible to install simpler versions of big hydro plant turbines—not for the Pelton model, which is the simplest one yet, for example, changing the blade edges.

As stated before, the choice of the turbine depends on the net head and the flow rate, as well as on the available water resources and the plant typology. A good criterion to select the turbine is resumed in the following well-known diagram, where they have a range in head and flow (Figure 18).

It should be stressed that the fields of employment are not very narrow and are only suggestions for the best choice. In fact, a Pelton turbine could be installed in a low head-high flow rate plant, even if the turbine efficiency will be strongly penalized. This results in areas of the diagram where different typologies of turbines can be used at the same time. The final choice has to be taken considering also other factors, such as the operation of the plant.

Conclusion

Hydropower has been used as far back as the Roman empire and through history has been used to power water mills, textile machines, sawmills, and irrigation systems. In the early 1800s, however, people started to see that the use of water to power small factories and machines is but a minor application of its potential. As early as the 19th century, waterpower was being used as a source of electricity. Though primitive hydropower technology only consisted of wheels, buckets, and river flow, it was from this point on that waterpower's potential as one of the most efficient and abundant sources of renewable energy became apparent.

Mini-hydropower is probably the least common of the three readily used renewable energy sources (i.e., water, sun, and wind), but it has the potential to produce the most power, more reliably than solar or wind power if you have the right site. Small-scale hydro is in most cases run-of-river, without dam or water storage, and is one of the most cost-effective and environmentally benign energy technologies for developing countries and further hydro development in Europe.[17]

In this entry, a summary of the main advantages and shortcomings of small-scale hydropower has been presented. The hydro resource is a much more concentrated energy resource than either wind or solar power and the energy available is readily predictable. Moreover, no fuel and only limited maintenance

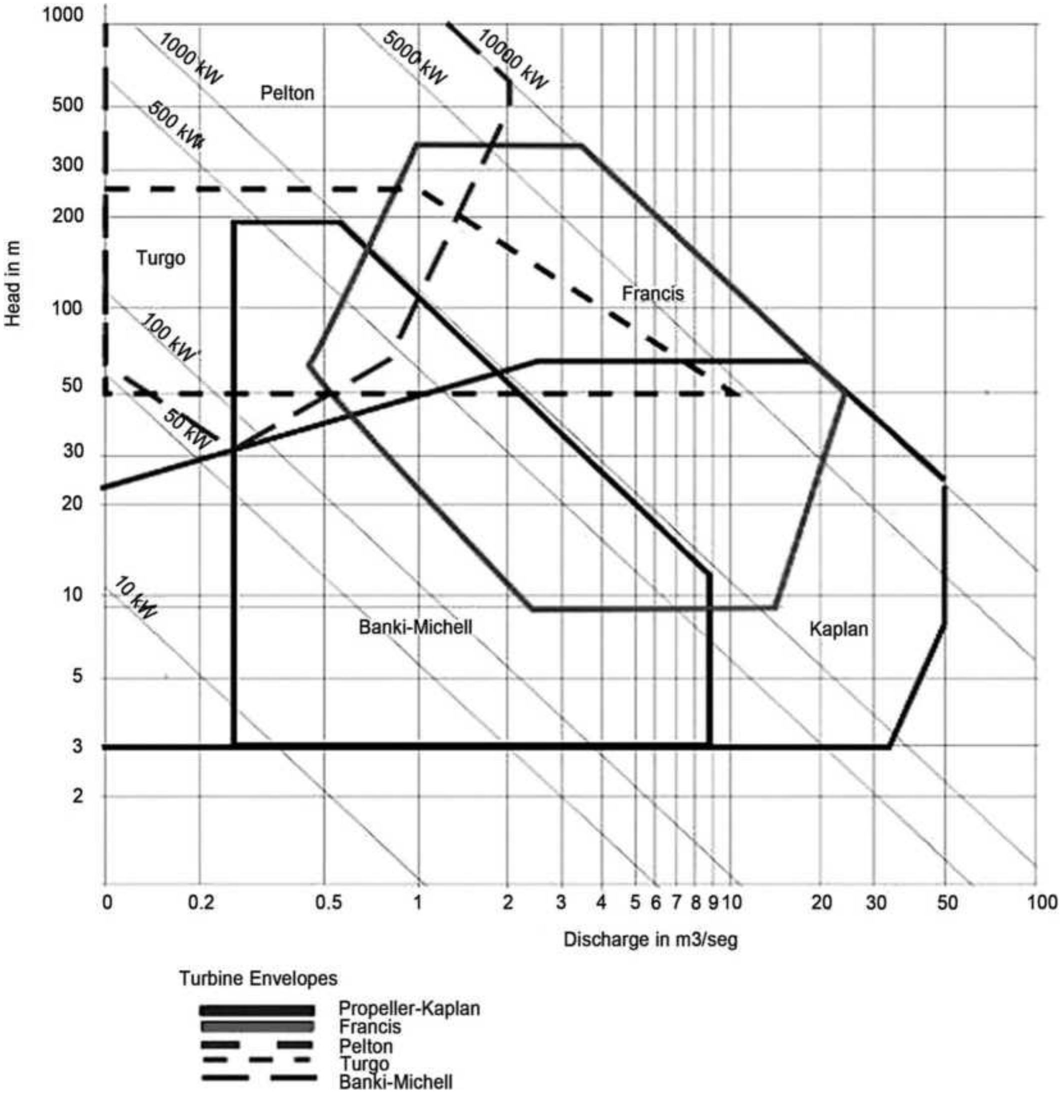


FIGURE 18 Values of flow and height for each field of application of different turbines.

are required and, if well designed, it has almost no environmental impact. On the other hand, it has to be considered that it is a site-specific technology and no general consideration has to be taken in the design of the plant; otherwise, environmental, social, or economical problems could occur, such as conflicts with fisheries interests on low-head plants and with irrigation needs on high-head plants. Furthermore, river flows often vary considerably with the seasons, especially where there are monsoon-type climates, and this can limit the firm power output to quite a small fraction of the possible peak output.

However, where a hydropower resource exists, experience has shown that there is no more cost-effective, reliable, and environmentally sound means of providing power than a hydropower system. Even with the various advantages of hydropower, it is still an underused alternative energy source. As of 2008, only 6% of the United States' electricity production came from hydropower, while nearly 50% came from the non-renewable source that is coal.[12] Due to a lack of economic speculation, a vast amount of potential for renewable hydropower remains untapped. Third world countries and underdeveloped areas have many areas that would be highly conducive to hydropower. The construction and use of hydropower facilities in these countries/areas, along with an increase in hydropower in the United States, could result in a great increase in renewable, affordable, and non-polluting energy.

However, if the prospected potential has to be realized, significant challenges have to be faced, in terms of *decision-making process,* establishing an equitable, credible, and effective environmental assessment procedure that takes into account both environmental and social concern and that takes into consideration the share of the benefits with local communities, both in the short term and in the long term. Finally, increasing efficiency, developing high-tech turbines, and reducing the costs of very low-head schemes, along with proper technology transfer of appropriate turbines to local manufacturers and technical support to the developers, will help realize our long-term objectives.

References

1. International Energy Agency. *Renewable Energy Essentials: Hydropower*; OECD/IEA, 2010.
2. *State of the art of small hydro power in the EU 25*; Thematic Network of Small Hydropower Project, European Small Hydro Power Association, 2005; 1–20.
3. *2007 Survey of Energy Resources*; World Energy Council, ISBN: 0 946121 26 5, 2007; 1–600.
4. Harvey, A., et al. *Micro-Hydro Design Manual*; IT Publications Ltd.: London, 1993.
5. Caputo, *C. Gli impianti convertitori di energia;* Casa edi-trice Ambrosiana, ristampa, 2008.
6. *International Journal of Hydropower and Dams, World Atlas;* Aquamedia Publications: Sutton, 2000.
7. Khennas, S.; Barnett, A. Best practices for sustainable development of micro-hydro in developing countries.
8. Available at http://www.friulanacostruzioni.it/pages/ita/prodotti/paratoie/paratoia-a-tenuta-su-tre-lati-manuale.php.
9. Churchill, S.W. Friction factor equations spans all fluid-flow ranges. Chem. Eng. **1977**, 91.
10. Arrighetti, C. *Dispense del corso di Macchine II; 2007.*
11. Paish, O. Micro-Hydro Power: Status and Prospects. Journal of Power and Energy, Professional Engineering Publishing, 2002.
12. Energy Technology Perspective 2010, IEA, Paris.

24

Alternative Energy: Photovoltaic Solar Cells

Ewa Klugmann-Radziemska

Introduction

Nowadays, fossil fuels are the main sources of energy from which electricity is obtained, but these sources will not last forever, so in due course renewable energies will have to replace them in this role. One of these new sources is solar energy. Each year, the Earth receives around 1×10^{18} kWh of solar energy, which is more than 1000 times the current global energy demand. This is therefore a vast source of energy that can be tapped to satisfy human energy requirements. To generate electricity from sunlight, solar cells (photovoltaic cells) are used. These devices are based on the photovoltaic effect, in which a p–n semiconductor is exposed to light, and photons are absorbed by electrons, providing an electric current. The electrons that are set free are pulled through the electric field and into the n-area. The holes produced move in the other direction, into the p-area.

The use of solar energy releases no CO_2, SO_2, or NO_2 gases and does not contribute to global warming. Photo-voltaics is now a proven technology that is inherently safe, as opposed to some dangerous electricity-generating technologies. Over its estimated life, a photovoltaic module will produce much more electricity than was used in its production. A 100 W module will prevent the emission of more than 2 tons of CO2. Photovoltaic systems make no noise and cause no pollution while in operation.

Photovoltaic Effect

The solar cells now in use are the practical application of fundamental physical phenomena observed already in the 19th century (see Table 1).

The absorption of light in semiconductors takes place when electrons are released from interatomic chemical bonds. In order to produce a free electron in a given semiconductor material, a certain quantity of energy must be supplied, equal at least to that of the energy band gap, which in the case of silicon at a temperature of 300 K is $E_g = 1.12\,\text{eV}$. The liberated electron leaves behind it a hole that can move about by diffusion or drift under the influence of an electrical field owing to its being positively charged (Figure 1).

The introduction of other atoms in place of the parent atoms (or at interstitial positions) of a pure intrinsic semiconductor considerably improves its electrical conductivity. Energy levels of elements with one valence electron more than the semiconductor atoms form donor levels in the neighborhood

TABLE 1 The Beginnings of Photovoltaics Worldwide

Year	Achievement
1839	Alexander Edmund Becquerel observed the photovoltaic effect in a circuit of two illuminated electrodes immersed in an electrolyte.
1843	Fritts produced the first tin–selenium solar cell.
1879	Adams and Day observed the photovoltaic effect at the interface between two solid bodies (selenium–platinum).
1930	The first copper/copper oxide solar cell.
1941	Ohl patented the silicon cell (monocrystalline silicon doped during its growth).
1954	At the Bell laboratory (United States), Chapin, Fuller, and Pearson developed a cell on monocrystalline silicon with a diffusion p–n junction of 6% efficiency, which was subsequently manufactured by two companies.
1954	Lindmayer and Allison obtained a cell with an efficiency of 16% (radiation intensity, 1000 w/m^2).
1954	Reynolds produced the first multiple-junction Cu_2S/ CdS cell.
1958	Monocrystalline solar cells were used for the first time in the Vanguard I satellite, where modules of six cells supplying 5 W of electrical power were installed; the traditional batteries ran out after a few months, but the photovoltaic panels powered the satellite's transmitter for another 6 years.
1962	The first thin-layer Cu_2S/CdS photocell was obtained.

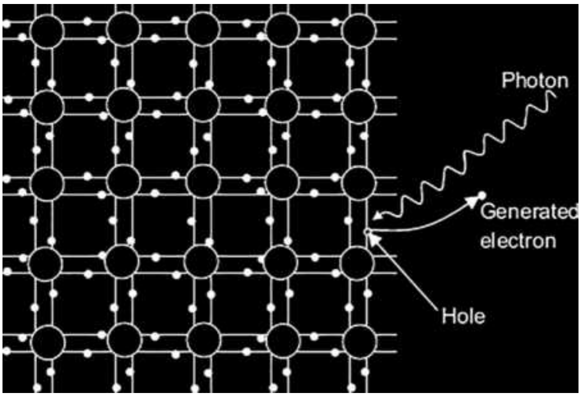


FIGURE 1 The generation of an electron–hole pair by a photon of energy $h\nu > E_g$.

of the conduction band (n-type). Energy levels of elements with one valence electron less than the semiconductor atoms form acceptor levels in the vicinity of the valence band (p-type).

If a p –n junction is formed from the p- and n-type areas of the semiconductor, then the charge carriers move around in such a way that the Fermi level will be identical throughout the crystal (Figure 2b). At room temperature (300 K), practically all donor and acceptor dopants are ionized; hence, the concentrations of majority carriers (electrons in the n-type area and holes in the p-type area) are approximately equal to the concentrations of the relevant dopants (Figure 2a).

At the instant these two areas are brought into intimate contact, a very large concentration gradient of electrons and holes across the boundary between them comes into existence. This gradient causes electrons to diffuse from the n-type area to the p-type area and the holes to move in the opposite direction. As a result of this diffusion, a space charge region comes into being near the junction: on the n-type side, this is positive, since electrons have left this area, while the uncompensated positive charges of immobile donor ions remain along with the holes newly arrived from the p-type area; on the p-type side, it is negative, because in the same way carrier diffusion has given rise to an area of negative charge consisting of immobile acceptor ions and electrons newly arrived from the n-type area. In this way, a dipole space charge layer is formed in the area around the p–n junction (Figure 3). A potential barrier and electric field are formed within this layer that counteract further diffusion and restrict the diffusion

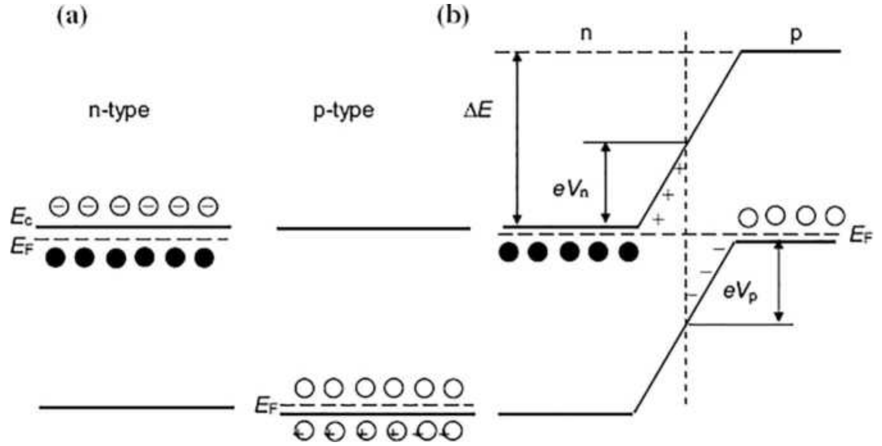


FIGURE 2 Formation of an abrupt p–n junction (b) as a result of the juxtaposition of n- and p-type areas (a), $\Delta E = e(V_n + V_p)$.

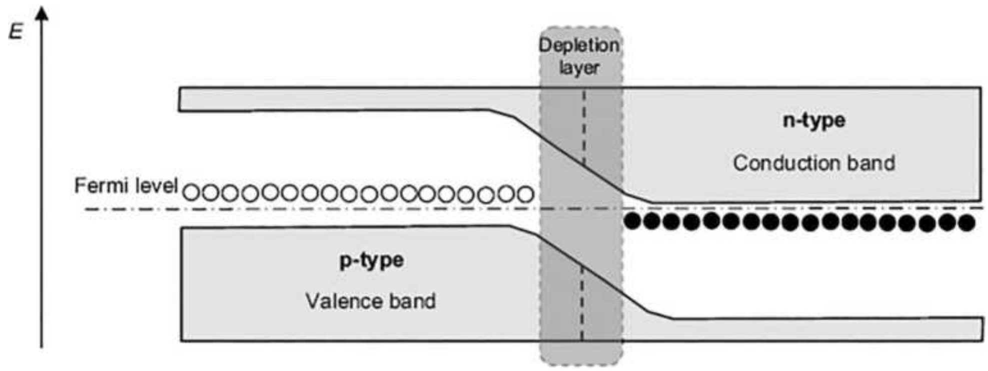


FIGURE 3 Equilibrium in the p–n junction region.

current. Apart from majority carriers, there are minority carriers in the two areas on either side of the junction, which come about as a result of the thermal generation of electron–hole pairs. The potential barrier formed as a result of majority carrier diffusion favors the outflow of minority carriers from both areas. The movement of these carriers creates a dark current, which flows in the opposite direction to that of the diffusion current.

If the p–n junction is illuminated by photons with an energy equal to or greater than the band gap width E_g ($h\nu \geq E_g$), then electron–hole pairs form on either side of the junction, as in the case of thermal generation (Figure 4).

Carriers forming no farther from the potential barrier than the diffusion length of minority carriers will diffuse towards the potential barrier and will be distributed there by the electric field due to the presence of the junction (the diffusion length is the mean distance that minority carriers have to move before they recombine with majority carriers). This field causes the carriers to move in opposite directions—electrons to the n-type area and holes to the p-type area. If an electron–hole pair forms on the p-type side of the junction, the electron reaches the junction before it has any chance of recombining with the hole (if recombination does occur, the resultant energy is emitted in the form of heat, and the effect is entirely useless as far as the photovoltaic effect is concerned), and the hole in this pair stays on the p-type side since it is repelled by the barrier in the junction. There is no danger of recombination here as there is an excess of holes in this region. The same thing happens when an electron–hole pair is generated by light on the n-type side of the junction. Then, the liberated electron remains on the n-type side, as it is repelled by the barrier. On the p-type side, however, we now have very few free electrons capable of recombining. This causes an increase in negative charge on the n-type side and of positive charge on the p-type side, which leads to a charge imbalance in the cell. This charge separation gives rise to a potential difference across the junction. As a result, a photoelectric current I_f comes into being in a closed circuit, regardless of the height of the potential barrier.

The generation of a photoelectric current I_f by a stream of photons in a solar cell can be demonstrated using the model of a current generator connected in parallel with a diode representing the p–n junction of the cell. As Figure 5 shows, the output current I flowing through the series resistance r_s of the cell and the load resistance is equal to the difference between the generated photoelectric current I_f and the diode current I_d.

$$I = I_f - I_d = I_f - I_s\left(\exp\frac{eU}{mkT} - 1\right),$$

where I_s is the saturation current and m is the diode ideality factor.

It emerges from the above equation that for cell operated at open circuit ($I = 0$):

$$I_d = I_f \text{ and } U = U_{oc} = \frac{mkT}{e}\text{In}\left(\frac{I_f}{I_s} + 1\right).$$

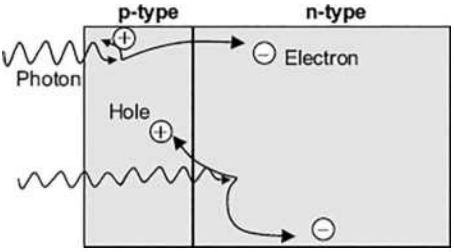


FIGURE 4 The potential barrier in a solar cell distributes the charge carriers generated by light.

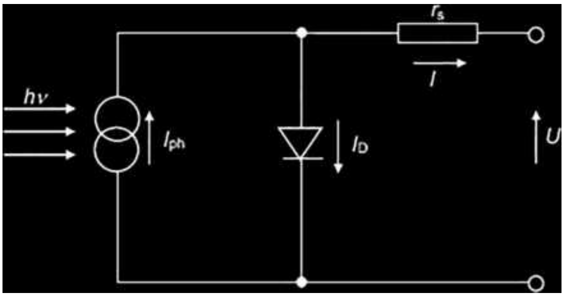


FIGURE 5 Electrical model of a solar cell.

From this last relationship, we obtain:

$$I_f = I_s\left(\exp\frac{eU_{oc}}{mkT}-1\right)\text{ and}$$

$$\text{finally}: I = I_s\left(\exp\frac{eU_{oc}}{mkT}-\exp\frac{eU}{mkT}\right).$$

For an exact description, however, we replace the single-diode electrical model with a two-diode equivalent circuit, which has two resistors: r_s is the series resistance of the cell, which consists of a number of components, and r_p is the effect of all defects in the crystal in the p–n junction area and is a shunt resistor (Figure 6).

In this model, the current generated is described as a function of the cell voltage as follows:[2]

$$I = I_{s1}\left(\exp\frac{e(U-Ir_s)}{m_1kT}-1\right)+I_{s2}\left(\exp\frac{e(U-Ir_s)}{m_2kT}-1\right)+\frac{U-Ir_s}{r_p}-I_{ph}$$

where I_{ph} is the photoelectric current, I_{s1} and I_{s2} are saturation currents, and m_1 and m_2 are non-ideality factors of the characteristics of the two diodes.

The parameters of the model are defined in such a way as to ensure that the above equation gives a good description of the real characteristic of a photovoltaic cell. The first exponential term in the characteristic equation represents the diffusion current, whereas the second one represents the recombination currents in the entire cell, particularly in the space charge region. The characteristics, which also enable the parameters of the two-diode model to be determined, are measured when the cell is polarized in the

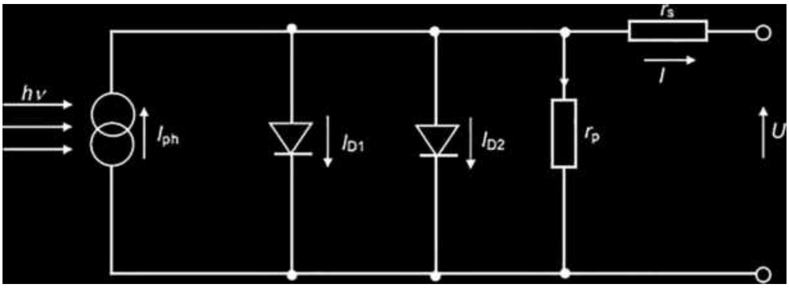


FIGURE 6 Equivalent circuit of a two-diode model of a solar cell.
Source: Stutenbaeumer and Masfin.[1]

forward direction and in the complete absence of any illumination, the dark current being measured as a function of the external voltage.

Solar Cell Characteristics

The usable voltage from solar cells depends on the semiconductor material. In silicon cells, it amounts to approximately 0.6 V.

Under illumination, the fourth quadrant of the light I–U is the region of interest (Figure 7), and the figures of merit for the device are the following:

1. The open-circuit voltage (U_{oc}) is the maximum voltage obtainable under open-circuit conditions (Figure 8).
2. The short-circuit current (I_{sc}) is the maximum current through the load under short-circuit conditions (Figure 8).
3. Fill factor (*FF*).

The output voltage of the photovoltaic cell is only slightly dependent on irradiance, while the current intensity increases with intensity of insolation. The working point of the solar cell therefore depends on load and insolation. In addition, the output voltage of a solar cell is temperature dependent. A higher cell working temperature leads to lower output and, hence, to lower efficiency (Figure 9)

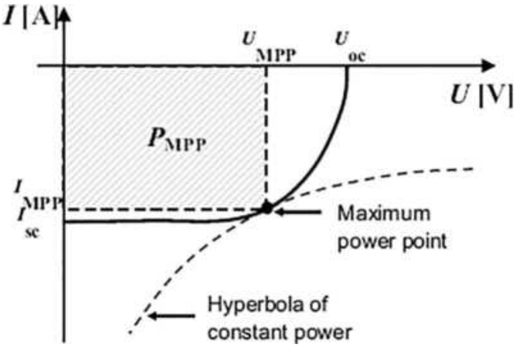


FIGURE 7 Current–voltage characteristic of illuminated photovoltaic solar cell.

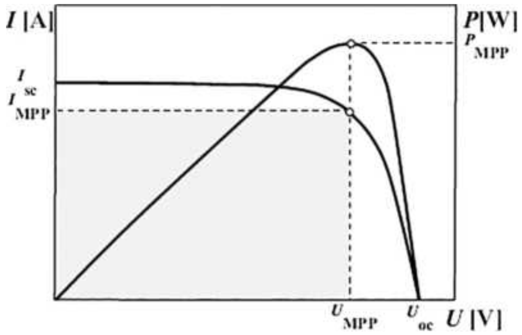


FIGURE 8 Current–voltage and power–voltage characteristics of the solar cell.

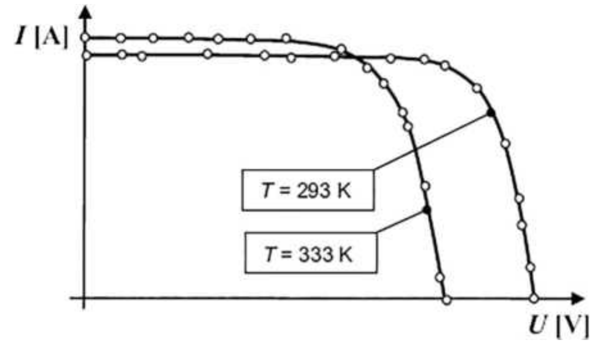


FIGURE 9 Temperature dependence of the solar cell characteristic.
Source: Radziemska and Klugmann.[3]

An important parameter as regards the application of a PV module in photovoltaics is the peak power obtainable from the module at the load resistance R_{opt}, where the rectangle under the characteristic $I(U)$ has the maximum area equal to the maximum power $P_{\text{mpp}} = I_{\text{MPP}}\, U_{\text{MPP}}$, and the point of intersection with the curve of $I(U)$ is in this case the maximum power point (MPP). The load resistance R in the cell circuit or PV module should be chosen such that the power it generates takes the maximum value, i.e., $P = \text{P}_{\text{Mpp}}$.

The MPP is the point at which the coordinates I_{MPP} and U_{MPP} form a rectangle with the largest possible area under the $I(U)$ curve.

The level of efficiency indicates how much of the radiated quantity of light is converted into usable electrical energy. The photovoltaic conversion efficiency of the cell η_{PV} is calculated from the maximum output power point (MPP) in the $I(U)$ curve:

$$\eta_{\text{pv}} = \frac{I_{\text{MPP}} I_{\text{MPP}}}{E.S_{\text{C}}}.100\%,$$

where S_{C} is the total surface of solar cell and E is the irradiance (W/m^2).

To describe solar cell quality, a special parameter—the fill factor (*FF*)—is used. It can be calculated from the following equation:

$$FF = \frac{I_{\text{MPP}} U_{\text{MPP}}}{I_{\text{sc}}.U_{\text{oc}}},$$

where I_{MPP} is the MPP current, U_{MPP} is the MPP voltage, I_{sc} is the short-circuit current, and U_{oc} is the open-circuit voltage.

For the ideal solar cell, the fill factor is a function of open-circuit parameters and can be calculated as follows:

$$FF \approx \frac{v_{\text{oc}-\ln(v_{\text{oc}}+0.72)}}{v_{\text{oc}}+1},$$

where v_{oc} is the voltage, calculated from the equation:

$$v_{\text{oc}} = U_{\text{oc}} \frac{e}{mkT},$$

where k is the Boltzmann constant = 1.38×10^{-23} J K^{-1}, T is the temperature (K), e is the charge on an electron = 1.6×10^{-19} C, and m is the diode ideality factor (–).

Solar Cell Materials: Production and Features

Silicon

The most important material for solar cell production is silicon. At the present time, it is almost the only material used for the mass production of solar cells. Being the most often used semiconductor material, it has some important advantages.

In nature, it is readily found in large quantities. Silicon dioxide forms one-third of the Earth's crust. It is environmentally friendly and not poisonous, and its waste does not cause any problems. It is easily melted and handled and it is fairly easy cast into its monocrystalline form. Its electrical properties, which remain unchanged up to temperatures of 125°C, allow the use of silicon semiconductor devices even in the harshest environments and applications.

In technology, pure silicon is the only widely used chemical element produced at such a high level of purity. The percentage of pure silicon in "pure silicon" is at least 99.9999999%. The concentration of silicon is 5×10^{22} atoms/cm^3, which means 5×10^{13} impure atoms/cm^3. Quantities of impure atoms are measured using sophisticated physical methods like mass spectrometry.

Pure silicon is produced from sand (silicon dioxide—SiO_2) by reduction at carbon electrodes at 1800°C in specially designed furnaces. The final material contains 98%–99% pure silicon. The complete reaction is:

$$\mathrm{SiO_2 + C \rightarrow Si + CO_2}.$$

Such silicon is the raw material for the production of pure silicon. It is also used in steel and aluminium production as a supplementary material. The most important producers of raw silicon are Canada, Norway, and Brazil. Fifteen to twenty-five kilowatt-hours of electrical energy is needed to produce 1kg of silicon. Silicon tetrachloride (tetrachlorosilane) gas is obtained by the chlorination of finely ground metallurgical-grade silicon in a special reactor. Additions or impurities are eliminated in the form of chlorine salt.

$$\mathrm{Si + 2CI \rightarrow SiCI_4}.$$

The following reaction produces trichlorosilane gas:

$$\mathrm{SiCI_2 + HCI \rightarrow SiHCI_3} \qquad .$$

This gas is then further purified with the removal of any remaining tetrachlorosilane and other silanes. The purification is followed by reduction in a hydrogen atmosphere at 950°C:

$$\mathrm{4SiHCl_3 + H_2 \rightarrow 2Si + SiCl_4 + SiCl_2 + 6HCl}.$$

Besides pure silicon, the procedure yields a number of gaseous by-products, which condense outside the reactor. Tetrachlorosilane is one of these by-products. At 1200°C, it can be converted into trichlorosilane using the following reaction:

$$\mathrm{SiCl_4 + H_2 \rightarrow SiHCl_3 + HCl}$$

This example illustrates one possible way of producing pure silicon. There are other procedures using different chemical reactions, but the end product is the same—pure silicon.

Crystalline Silicon Solar Cells

Polycrystalline as well as monocrystalline solar cells belong to this group. The basic form for crystalline solar cell production is the silicon ingot (see the description of the production procedure above). The ingot (block of silicon), cut with a diamond saw into thin wafers, is the basis of solar cell production. One-millimeter-thick wafers sawn accurate to 1/10 mm are placed between two plane-parallel metal plates rotating in opposite directions. This procedure enables the wafer thickness to be adjusted to within 1/1000 mm. The subsequent procedure for solar cell production consists of the following steps:

- Doped wafers are first etched some micrometers deep. The procedure removes crystal structure irregularities caused by sawing and cleans the wafer. During the extraction of pure silicon, the material is doped either as powdered polycrystalline silicon or by the addition of a suitable gas. This is then followed by diffusion. Phosphorus, supplied inside the material in gaseous form, diffuses at 800°C. The n-doped layer and the p-rich oxide layer form on top of the wafers as a result of reaction with oxygen.
- The wafers are then folded to form a cube and etched in oxygen plasma, which removes the n-doped layer from the edges. Wet chemical etching then removes the oxide layers from the top of the wafer.
- At the rear, the contact surface is produced from silver containing 1% aluminium. Special procedures enable silver to be printed over mask on cell surface. The pressed cells are then sintered at high temperatures. A similar procedure is used to print the contacts on the front cell surface, and the anti-reflex layer is applied likewise. In this case, titanium paste is used, which, on sintering, forms titanium dioxide (TiO_2) or silicon nitride (Si_3N_4).

Polycrystalline Silicon Production

The extraction of pure polycrystalline silicon from trichlorosilane can be carried out in special furnaces, such as those developed by Siemens. The furnaces are heated by electric current, which, in most cases, flows through silicon electrodes. These 2 m long electrodes are 8 mm in diameter. The current flowing through the electrodes can be as much as 6000 A. The furnace walls are cooled to prevent the formation of unwanted reactions producing gaseous by-products. The procedure yields pure polycrystalline silicon, which is used as a raw material for solar cell production. Polycrystalline silicon can be extracted from silicon by heating it up to 1500°C and then cooling it down to 1412°C, which is just above the melting point of the material. As material cools, a 40 × 40 × 30 cm ingot of fibrous polycrystalline silicon forms.

Monocrystalline Silicon Production

Two different technological procedures are used to produce monocrystalline silicon from pure silicon.

Czochralski's Method

In 1918, the Polish scientist Jan Czochralski discovered a method for producing monocrystalline silicon, from which monocrystalline solar cells could be manufactured. The first monocrystalline silicon solar cell was constructed in 1941. In Czochralski's method, silicon is extracted from the melt in a graphite-lined induction oven at a temperature of 1415°C. A silicon crystal with a set orientation is placed on a rod. Spinning the rod in the melt makes the crystal grow. The rod spins at 10 to 40 revolutions per minute and grows in length at a rate of between 1 μm and 1 mm per second. This allows the production of rods measuring 30 cm in diameter and several meters in length. The whole process takes place in an inert atmosphere. Possible impurities are burnt or eliminated in the melt.

Float Zone Method

With this method, monocrystalline silicon is produced from polycrystalline silicon. The main advantage of this procedure over the previous one is the better yield of pure silicon.

The silicon rods produced measure 1 m in length and 10 cm in diameter. This procedure, in which an induction heater travels along the rod melting the silicon, also takes place in an inert atmosphere. Monocrystalline silicon is produced during the cooling stage. Monocrystalline or polycrystalline silicon ingots are then sawn and the wafers are worked upon until they can serve as a foundation for solar cell production. Sawing causes approximately 50% of the material to be wasted.

Amorphous Solar Cells

Amorphous silicon is produced in high-frequency furnaces under partial vacuum. In the presence of a high-frequency electrical field, gases like silane, B_2H_6, or PH_3 are blown through the furnaces, supplying silicon with boron and phosphorus.

Amorphous solar cells are produced with technologies similar to those used in the manufacture of integrated circuits. Due to this procedure, these modules are also known as thin-film solar cells (thin-film modules). Here is a brief summary of amorphous solar cell production:

- The glass substrate is cleaned thoroughly.
- The lower contact layer is applied.
- The surface is then structured—it is divided into bands.
- The amorphous silicon layer is applied under vacuum and in the presence of a high-frequency electric field.
- The surface is rebanded.
- The upper metal electrodes are fixed.

Other Solar Cells

The less frequently used solar cell types include solar cells produced by the EFG (edge-defined film-fed growth) method, as well as Apex solar cells made from silicon, cadmium telluride (CdTe) solar cells, and copper-indium selenide (CIS) solar cells. EFG monocrystalline solar cells are produced directly from the silicon melt, which eliminates wafer sawing; production costs are thus lower and material is saved since there is no “sawdust.” In the EFG procedure, an octagonal tube of silicon, several meters long, is extruded from the silicon melt. The flat sides of this tube are then laser sawn into separate solar cells. Most solar cells are square in shape and 100 × 100 mm in size. Consequently, the module power is greater with a smaller surface compared to crystal modules of square cells with truncated sides. Contacts take the form of copper bands. The separate cells are then combined in the same way as other cell types.

EFG cells are produced by Schott Solar. In contrast to EFG cells, Apex cells are polycrystalline. Their production procedure is protected by patent. The production procedure was developed by Astropower Inc.

Cadmium Telluride

Thin-film material produced by deposition or by sputtering is a promising low-cost foundation for photovoltaic applications in the future. The disadvantage of this procedure, however, is that the materials used in production are toxic. The efficiency of solar cells in the laboratory is as high as 16%, but that of commercial types is only 8%.

Copper-Indium-Diselenide ($CuInSe_2$, or CIS)

$CuInSe_2$ is a thin-film material with an efficiency of up to 17%. The material is promising, but not yet widely used owing to production-specific problems. CdTe and CIS cells have so far been used mostly in laboratory research. Commercial modules made from these materials are still hard to find.

Gallium Arsenide (GaAs)

GaAs is used in the production of high-efficiency solar cells. It is often utilized in concentrated PV systems and space applications. Their efficiency is as good as 25%, and even 28% at concentrated solar radiation. Special types have an efficiency of more than 30%.

Structure and Manufacture of Photovoltaic Cells from Crystalline Silicon

In principle, a photovoltaic cell consists of the following elements (see Figure 10):

- A mono-or polycrystalline silicon wafer in which a p–n junction has been formed.
- Contacts, i.e., the front and rear electrodes; the front one should be shaped in such a way that the maximum amount of incident radiation can reach the junction region, the depth of which is limited by the permeability of silicon to radiation.
- An antireflective coating (ARC) on the front side of the cell.

The manufacture of a crystalline cell takes place in the following stages.

Surface Preparation

The surfaces of silicon wafers, cut from monocrystalline ingots, are degreased, cleaned, polished (mechanically or chemically), and etched in an aqueous solution of sodium hydroxide (40% NaOH) at a temperature of 383 K. A pyramidal surface structure is thereby obtained, which is then rinsed in hydrochloric or nitric acid. Dry (plasma) etching is also possible.

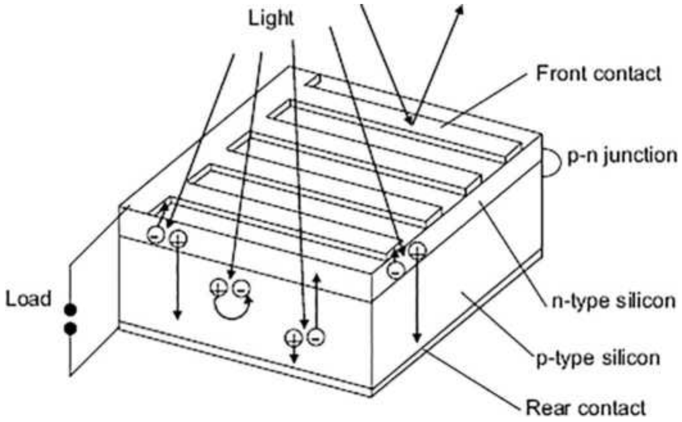


FIGURE 10 The structure and functioning of a photovoltaic cell (not to scale).

Diffusion Formation of a p–n Junction

A dopant, usually an acceptor, is added to the silicon base during crystal growth, whereas the n–p junction is produced by the diffusion of a dopant (usually a donor) to the p-type base wafer across one of its surfaces. If the donor concentration in the subsurface layer of the silicon (initially p-type) is greater than the acceptor concentration, this layer then becomes an n-type semiconductor.

The source of the dopant may be a solid or a gas. There are a number of techniques involving diffusion from the solid phase:

- Vacuum deposition of a thin layer of dopant
- Doping a vacuum-deposited layer of SiO_2
- Doping a mechanically deposited or screen-printed layer of SiO_2
- Coating the silicon wafer with a material containing phosphorus and silicon dioxide

If the diffusion of phosphorus to silicon is carried out at a temperature of 1220 K, 10 min are sufficient to obtain an n-type layer 0.25 μm in thickness.

Formation of an p–n Junction by Ion Implantation

Ion implantation is a technique for obtaining shallow, abrupt n–p junctions; it allows precise control of junction depth and dopant concentration. It is based on the implantation of dopant ions into silicon when its surface is bombarded with a beam of ions (of the order of 10^{16} ions/cm^2) of energy 5–300 keV.

The depth of the junction depends on the energy of the ions. A drawback of this method it that its gives rise to a large number of structural defects in the silicon; these can be removed after implantation by either heating the silicon wafer or irradiating it with an electron or laser beam. Ion implantation is quite a costly method—the necessary equipment (ion source, ion separator and accelerator, vacuum) is expensive.

Passivation of the Silicon Surface

Both the quantum efficiency for blue light and the open circuit voltage can be considerably enhanced if the front surface of the silicon is passivated. This is easily done with a layer of silicon dioxide, which forms on the surface when this oxidizes. Such a thin layer of SiO_2 of controlled thickness can be obtained by heating the silicon wafer in a stream of neutral gas like nitrogen containing a small quantity of dry oxygen. In the cells designed by Green, there is also a very thin layer of SiO_2 (ca. 2 nm thick) underneath the metal front electrode.

Metal Contacts

Since cells are an integral part of an electrical installation, metal contacts are made on either side of the cell. The front electrode is a fine grid, so as to reduce shading of the light-sensitive surface to a minimum. The “fingers” are ca. 0.1 mm wide, and the bus bar is from 1.5 to 2 mm wide. During the module’s construction, the bus bar is connected to the rear electrode of the adjacent cell by means of copper strips soldered to it.

The metallic layers in a cell should be in ohmic contact with the silicon and have a low contact resistance, good adhesion to the silicon surface, and good soldering properties. Various techniques are used worldwide for producing metal contacts fulfilling these requirements.

Unlike the front electrode, the rear one can cover the entire area of the wafer. To improve efficiency, the rear electrode is made from a layer of aluminium vacuum deposited on the silicon surface between silver contacts in the form of strips or squares 2.5–6 mm wide. When this electrode is heated at temperatures from 770 to 1070 K, aluminium diffuses to the silicon forming a thin p^+ layer. To obtain a $p–p^+$

junction, the p^+ layer must be far thicker (0.2 μm); this is achieved by heating the cell to a temperature of 970–1070 K.

A very much cheaper method of making the two electrodes is the chemical deposition of nickel, or screen printing using a paste containing silver, aluminium, copper, or some other metal. Screen printing is a method that was used to produce lettering with a stencil over a thousand years ago. In 1975, it was first used in silicon cell technology to deposit the front and rear electrodes, thus replacing the costly vacuum deposition technique.

Silver paste is used for producing the front electrode, while aluminium paste combined with a small quantity of silver is usually used for the rear one. The silver pastes used for screen printing metal contacts on a silicon surface consist of a conducting phase (powdered silver of grain size 1–3 pm), an auxiliary phase (assisting the sintering of solid-phase grains—an enamel formed from the melting of a mixture of inorganic oxides), and an organic carrier facilitating the screen-printing process, which is burnt off when the layer is fired. The silver layers are usually sintered at 850°C.

About 86% of the silicon cells produced in 2006 had screen-printed metal contacts. At present, standard silicon cells with screen-printed electrodes achieve an efficiency of around 15% if they are polycrystalline and ca. 18% if they are monocrystalline (produced by Czochralski's method). During the screen printing of metal contacts, the mesh must be placed at a constant distance from the front side of the wafer. Silver paste is applied to the mesh and then imprinted using two squeegees. To ensure that the paste properly fills all the openings in the mesh, one squeegee moves along it, spreading the paste down its whole length. The other squeegee then applies just enough pressure to force the paste out of the mesh openings onto the wafer surface. After drying at 120°C for ca. 60 min, the printed layer consists of an aggregation of loose grains 1–2 μm in size; this must now be fired in order to impart stability to the layer.[4]

Cells with Rear Contacts

In cells of this type, both the positive and negative contacts are made on the rear surface of the cell. In this way, the whole of the front light-sensitive surface can be used to harvest light and the space between the cells can be minimized. The SunPower company produces commercial cells and modules of this type from n-type monocrystalline silicon—the efficiency of the cells is 21.5%, and that of the modules is 18.6%. The p–n junction is produced in the lower layer in the form of bands. This means that the photogenerated charges have to cover quite a long distance to reach the junction region, so only high-quality silicon is suitable for this type of cell.

Deposition of *the ARC*

One of the most significant parameters affecting the efficiency of a cell is the coefficient of light reflection from its surface—in the case of silicon, this is 33%–54%. This can be minimized by applying a transparent ARC to the cell's active surface.

ARCs can be applied in various ways: chemical vapor deposition (CVD), spraying, spin-on, and screen printing. The spin-on technique is the simplest one, as it does not require expensive equipment and is very efficient, but it can only be usefully applied to silicon wafers with a smooth, polished surface. If the wafer surface is textured, the ARC obtained in this way will not be of uniform thickness. Plasma-enhanced chemical vapor deposition (PECVD) produces ARCs with very good refractive index, photonic band gap, homogeneity, chemical composition, and controlled thickness. The surfaces of silicon wafers are usually coated with one or two antireflective layers.

The presence of an ARC is responsible for the color of the cells: polycrystalline cells are blue and monocrystalline ones are dark blue to black. By optimizing the thickness of the ARC, it is now possible to produce cells that are green, gold, brown, and violet in color, but this is only at the expense of their efficiency. One can, of course, do without the ARC and apply the cells in their original silver (polycrystalline) or dark gray (monocrystalline) colors; depending on architectural requirements, solar panels without an ARC can be integrated, for example, into the façades of buildings.

Dye-Sensitized Solar Cells

Dye-sensitized cells (DSCs) imitate the way that plants and certain algae convert sunlight into energy. The cells are inexpensive, easy to produce, and can withstand long exposure to light and heat compared with traditional silicon-based solar cells. This is a relatively new class of low-cost solar cell. It is based on a semiconductor formed between a photosensitized anode and an electrolyte, a photoelectrochemical system.

The fruits, flowers, and leaves of plants are tiny factories in which sunlight converts carbon dioxide gas and water into carbohydrates and oxygen. Although not very efficient, they are very effective over a wide range of sunlight conditions. Despite the low efficiency and the fact that the leaves must be replaced, the process has worked for hundreds of millions of years and forms the primary energy source for all life on earth. On the basis of this principle, there were early attempts to cover crystals of semiconductor titanium dioxide with a layer of chlorophyll. Unfortunately, the efficiency of the first solar cells sensitized in this way was about 0.01%. In 1991, Michael Grätzel and Brian O'Regan at the École Polytechnique Fédérale de Lausanne (Switzerland) used a sponge of small particles, each about 20 nm in diameter, coated with an extremely thin layer of pigment to obtain the effective surface area available for absorbing light.[5] These dye-sensitized solar cells (DSSCs or DSCs) are also known as Grätzel cells. Following much academic research, the energy conversion efficiency of laboratory cells made on glass substrates with liquid electrolytes has steadily increased to around 10% at air mass 1.5, 1 Sun conditions (for testing, the cells, regardless of design and active material, are typically insolated at a constant density of roughly 1000 W/m^2, which is defined as the standard 1 Sun value).

Dye-sensitized photoelectrochemical solar cells differ from conventional photovoltaic solar cells in that they separate the function of light absorption from charge carrier transport. The cells are made up of a porous film of tiny (nanometer sized) white pigment particles of titanium dioxide. These are covered with a layer of dye that is in contact with an electrolyte solution. Photoexcitation of the dye results in the injection of an electron into the conduction band of the oxide. The original state of the dye is subsequently restored by electron donation from a redox system, such as the iodide/tri-iodide couple. This process results in the conversion of sunlight into electrical energy.

In the case of the original Grätzel design, the cell has three primary parts. On the top is a transparent anode made of fluorine-doped tin dioxide (SnO_2:F) deposited on the back of a (typically glass) plate. On the back of the conductive plate, a thin layer of titanium dioxide (TiO_2) is deposited, which forms into a highly porous structure with an extremely high surface area. The plate is then immersed in a mixture of a photosensitive ruthenium-polypyridine dye and a solvent. After the film has been soaked in the dye solution, a thin layer of the dye is left covalently bonded to the surface of the TiO_2. A separate backing is made with a thin layer of the iodide electrolyte spread over a conductive sheet, typically platinum metal. The front and back parts are then joined and sealed together to prevent the electrolyte from leaking. The construction is simple enough that there are hobby kits available for hand-constructing them. Although they use a number of advanced materials, these are inexpensive compared to the silicon needed for normal cells because they require no expensive manufacturing steps. TiO_2, for instance, is already widely used as a paint base.

Principles

DSSCs separate the two functions typical of a traditional silicon cell design. In the crystalline silicon solar cells, the silicon acts as the source of photoelectrons and provides the electric field to separate the charges and produce the current.

In the DSSC, the bulk of the semiconductor is used solely for charge transport, while the photoelectrons are provided from a separate photosensitive dye. Charge separation occurs at the semiconductor/dye/electrolyte interface.

Although photoelectrochemical cells can operate without an organic dye, the efficiency of such cells is very low due to the low light-harvesting ability of inorganic n-conductors, which normally absorb light only from the high-energy ultraviolet region of the solar spectrum. The introduction of an organic dye makes for a significant increase in the absorption ability of the cells that extends across almost the entire solar spectrum.

DSC Structure

The structure of currently produced DSSCs is similar to those produced by Grätzel and O'Regan: on the top, there is a transparent anode made of indium tin oxide (ITO), deposited on the back of a glass plate. ITO (or tin-doped indium oxide) is a solid solution of indium (III) oxide (In_2O_3) and tin (IV) oxide (SnO_2), typically 90% In_2O_3 + 10% SnO_2 by weight. In thin layers, it is transparent and colorless, but in bulk form, it is yellowish to gray. In the infrared region of the spectrum, it is a metal-like mirror. ITO's main feature is the combination of electrical conductivity and optical transparency. A compromise has to be reached during film deposition, as a high concentration of charge carriers will increase the material's conductivity but decrease its transparency. Thin films of ITO are most commonly deposited on surfaces by electron beam evaporation or a range of sputter deposition techniques.

The thin oxide coating on one side of the glass makes the glass surface electrically conducting. On the back of the conductive plate, a thin layer of titanium dioxide (TiO_2) is deposited, which forms into a highly porous structure with an extremely high surface area—a 10 μm layer of randomly stacked nanoparticles (ca. 20 nm in diameter) (Figure 11).

The plate is then immersed in a mixture of a photosensitive ruthenium–polypyridine dye and a solvent. After the film has been soaked in the dye solution, a thin layer of the dye is left covalently bonded to the surface of the TiO_2. A separate backing is made with a thin layer of the iodide electrolyte spread over a conductive sheet, typically platinum metal. The front and back parts are then joined and sealed together to prevent the electrolyte from leaking. Sunlight enters the cell through the transparent top contact, striking the dye on the surface of the TiO_2.

Titanium Dioxide

Titanium dioxide, also known as titanium(IV) oxide or titania, is the naturally occurring oxide of titanium TiO_2. When used as a pigment, it is called titanium white.

Titanium dioxide occurs in nature as the well-known, naturally occurring minerals rutile (the most common and stable form), anatase, and brookite. Crude titanium dioxide is purified via titanium tetrachloride in the chloride process. In this process, the crude ore (containing at least 90% TiO_2) is reduced with carbon, oxidized with chlorine to give titanium tetrachloride.

This titanium tetrachloride is distilled and then reoxidized with oxygen to give pure titanium dioxide. Another widely used process utilizes ilmenite as the titanium dioxide source, which is digested in sulfuric acid (as in Millennium Inorganic Chemicals). The by-product, iron(II) sulfate, is crystallized

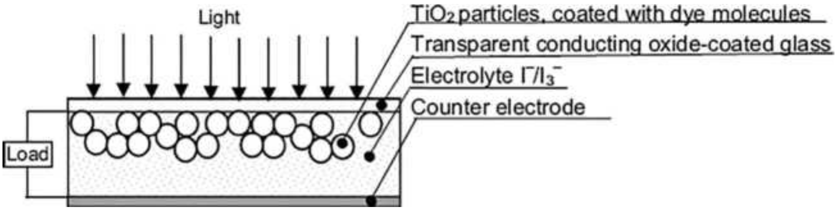


FIGURE 11 DSC structure.

and filtered off to yield only the titanium salt in the digestion solution, which is processed further to yield pure titanium dioxide.

The TiO_2 semiconductor has three functions in the DSSC:

1. It provides the surface for the dye adsorption.
2. It functions as electron acceptor for the excited dye.
3. It serves as electron conductor.

Colloid preparation and layer deposition have been developed to optimize the TiO_2 for these functions. Most important for the performance of desensitized solar cells was the development of a mesoporous semiconductor structure. This becomes evident considering the limited light capture of a dye monolayer on a flat surface.

The conductivity of nanophase TiO_2 films in a vacuum has been found to be very low, $\sim 10^{-9}$ $(\Omega \text{ cm})^{-1}$ at room temperature.[6] However, on exposure to UV light, the conductivity is much increased, indicating that the low conductivity in the dark is due to the low electron concentration in the conduction band rather than to poor electrical contacts between the particles.

Untreated TiO_2 is an insulator that becomes "photodoped" and therefore conductive following electron injection of the adsorbed dye. Electronic contact between the nanoparticles is established by sintering the nanoparticles together, which enables the entire surface-adsorbed molecular layer to be accessed electronically. The interconnection of the nanoparticles by the sintering process allows the deposition of a mechanically stable, transparent film, typically a few microns thick. It is not necessary to increase the free electron concentration in the dark; indeed, this may even be detrimental to the photoelectrochemical behavior of the TiO_2. Among several semiconductors studied for photoelectrochemical applications, TiO_2 is by far the most commonly used, because of its energetic properties, its stability, and the ability to attach dyes. It is, furthermore, a low-cost material that is widely available. TiO_2 is used in its low-temperature stable form anatase (pyramid-like crystals), as rutile shows non-negligible absorption in the near-UV region (350–400 nm). This excitation within the band gap leads to the generation of holes, which are strong oxidants and cause long-term instability issues in the solar cell.

Mechanism of Operation

Photons striking the dye with enough energy to be absorbed will create an excited state in the dye, from which an electron can be injected directly into the conduction band of the TiO_2, and from where it moves by diffusion (as a result of an electron concentration gradient) to the clear anode on top (Figure 12).

Meanwhile, the dye molecule has lost an electron and the molecule will decompose if another electron is not provided. The dye strips one from the iodide in the electrolyte below the TiO_2, oxidizing it to tri-iodide. This reaction occurs quite quickly compared to the time that it takes for the injected electron to recombine with the oxidized dye molecule, thus preventing the recombination reaction that would

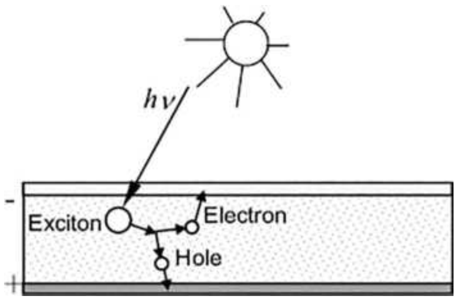


FIGURE 12 Photogeneration of charge carriers in DSC.

effectively short-circuit the solar cell. The tri-iodide then recovers its missing electron by mechanically diffusing to the bottom of the cell, where the counter electrode reintroduces the electrons flowing through the external circuit.

The dye-sensitized oxide is usually deposited on a highly doped transparent conducting oxide (TCO), which allows light transmission while providing sufficient conductivity for current collection. Recently, high-conductivity organic polymers deposited onto plastic foil have found increasing application as substrates for flexible devices. The conductivities of metal foils are superior to those of TCOs and polymers. Because of their opacity, the illumination of the cell has to be established through the counter electrode. The surface of TCO should make good mechanical and electrical contact with the porous TiO_2 film.

To reduce dark current losses due to the short-circuiting of electrons in the substrate with holes in the hole conductor, a thin underlayer of TiO_2 is introduced between the SnO_2 layer and the nanocrystalline TiO_2 layer. This thin compact layer improves the mechanical adhesion of the porous TiO_2 film to the substrates, especially to SnO_2 layers of low haze, i.e., layers with less surface roughness and thus less contact area. Figure 13 illustrates a model of charge carrier separation and charge transport in a nanocrystalline film.[7] The electrolyte is in contact with the individual nanocrystallites. Illumination produces an electron–hole pair in one crystallite. The hole transfers to the electrolyte and the electron traverses several crystallites before reaching the substrate.

The photogenerated hole always has a short distance (roughly equal to the particle radius) to cover before reaching the semiconductor–electrolyte interface whenever an electron–hole pair is created in the nanoporous film. However, the probability that the electron will recombine depends on the distance between the photoexcited particle and the back contact.

The mechanism for converting solar energy into electrical energy in a DSSC is a five-step process:

1. Solar energy (photons of light *hv*) causes electrons in the molecular orbitals within the adsorbed dye sensitizer (S) molecules to become photoexcited (S*):

   S (adsorbed on TiO_2) + $hv \rightarrow$ S* (adsorbed on TiO_2)

   The trapping of solar energy by a sensitizer molecule is analogous to the light-absorbing chlorophyll molecule found in nature, which converts carbon dioxide and water to glucose and oxygen.

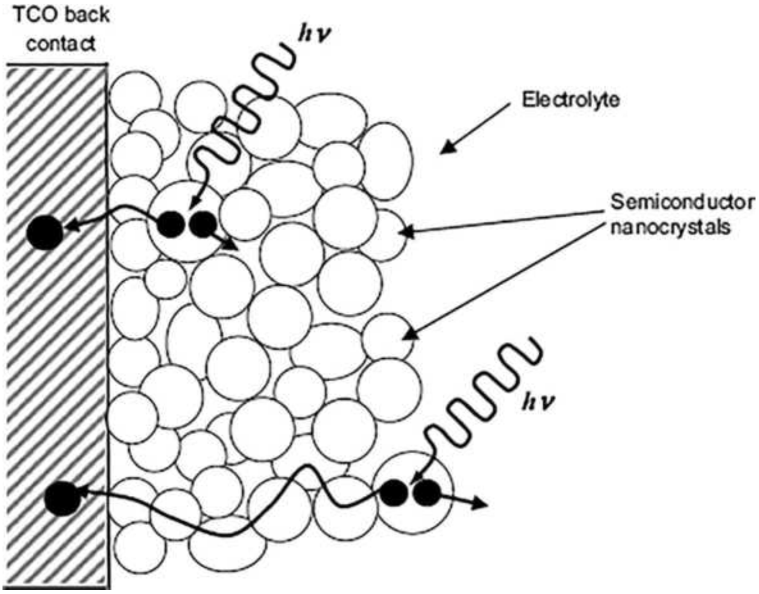


FIGURE 13 Qualitative model of photocurrent generation in nanocrystalline films.

2. The excited electrons escape from the dye molecules: S* (adsorbed on TiO_2) → S^+ (adsorbed on TiO_2) + e^-
3. The free electrons then move through the conduction band of TiO_2, gather at the anode (the dyed TiO_2 plate), and then start to flow as an electric current through the external load to the counter electrode.
4. The oxidized dye (S^+) is reduced to the original form (S) by regaining electrons from the organic electrolyte solution that contains the iodide/tri-iodide redox system, with the iodide ions being oxidized (loss of electrons) to tri-iodide ions:
   S^+ (adsorbed on TiO_2) + 3/2 I^- → S (adsorbed on TiO_2) + 1/2 I_3^-
5. To restore the iodide ions, free electrons at the counter (graphite) electrode (which have travelled around the circuit) reduce the tri-iodide molecules back to their iodide state. The dye molecules are then ready for the next excitation/oxidation/reduction cycle.

Initially, the solutions of iodine–iodide mixtures in volatile solvents, usually acetonitrile, were used as redox electrolytes:

Electrolyte:	I2 + I– ↔ I3–
Anode (Dye):	2 Dye+ + 3 I– → 2 Dye0 + I3–
Cathode:	I3– + 2e– → 3 I–

On the basis of different measurements, it is possible to indicate the orders of magnitude for the rate of the reaction steps. Upon illumination, the sensitizer is photoexcited in a few femtoseconds and electron injection is ultrafast from S* to TiO_2 on the subpicosecond time scale, where they are rapidly (less than 10 fs) thermalized by lattice collisions and phonon emissions. The nanosecond-ranged relaxation of S* is rather slow compared to injection, which ensures that the injection efficiency is unity. The ground state of the sensitizer is then recuperated by I^- in the microsecond domain, effectively annihilating S^+ and intercepting the recombination of electrons in TiO_2 with S^+, which happens in the millisecond time range. This is followed by the two most important processes—electron percolation across the nanocrystalline film and redox capture of the electron by the oxidized relay, I_3^-, within milliseconds or even seconds.[8]

Recently, solvent-free redox electrolytes, prepared from ionic liquids (liquid ionic organic compounds) or from ionically conducting polymer–nanocrystal blends were found to be very efficient. Figure 14 shows the processes taking place during the conversion of light into electrons in a DSSC.

Upon excitation, the dye injects electrons into the conduction band of the titanium oxide. The photo-induced electrons diffuse through the porous TiO_2 network and are extracted at the SnO_2 substrate. The dye itself is regenerated by the electrolyte containing the I^-/I_3^- redox pair. In the most efficient DSCs, the dye is regenerated by I^-, present in an external electrolyte at high (~0.45 M) concentrations.[10] The electronic circuit is closed by the reduction of the iodide couple at the platinized SnO_2 counter electrode.

Titania Solar Cells: Manufacturing Process

Manufacturing processes utilize the production methodologies and equipment already in use in the manufacture of electronic components. Production of cells needs a relatively low capital investment, the equipment does not rely on highly skilled operators, and—most importantly—the technology is environmentally friendly. A number of these manufacturing processes are protected by patents.

During production, a piece of glass is coated with fluorine-doped tin oxide and then sputtered onto a 500 nm thick layer of titanium. The layer is then anodized by placing it in an acidic bath with a mild electric current, and the titanium dioxide nanotube arrays grow to about 360 nm. The tubes are then

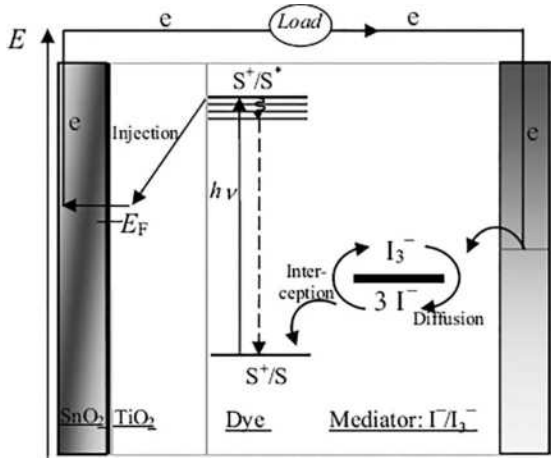


FIGURE 14 The electron processes taking place in an illuminated DSSC under open-circuit conditions.
Source: Anandan.[9]

heated in oxygen so that they crystallize. The process turns the opaque coating of titanium into a transparent coating of nanotubes.

In the next step, the nanotube array is coated with a commercially available dye—this becomes the negative electrode. The cell is sealed with a positive electrode, which contains an iodized electrolyte.

The most important advantages of the DSC manufacturing processes are as follows:

- The materials can be produced cheaply in large quantities.
- They use standard processing and assembly equipment.
- They are not very energy intensive (about 32 kWh/m^2).
- They can be automated.
- The equipment consists of a number of processing stations, each easily reprogrammable to adopt a wide range of DSC designs.

DSSCs can be connected in series or parallel and integrated into solar panels (modules). The tiles are normally ochre, but other colors such as gray, green, and blue are to be introduced shortly.

Advantages of DSC in Comparison with Other Solar Cells

The important features that distinguish a DSC from a conventional silicon crystalline solar cell are as follows:

- It is a photoelectrochemical cell: charge separation occurs at the interface between a wide band gap semiconductor and an electrolyte.
- It is a nanoparticulate titania cell: it is not a dense film like amorphous silicon, but a “light sponge.”
- In a DSC, the dye monolayer chemically absorbed on the semiconductor is a primary absorber of sunlight; free charge carriers are generated by electron injections from a dye molecule, excited by visible radiation.
- In a DSC, light absorption and charge carrier transport are separate, whereas in a conventional cell, both processes are performed by the semiconductor.
- An electric field is necessary for charge separation in the p–n junction cell. Nanoparticles in the DSC are too small to sustain a built-in field; accordingly, charge transport occurs mainly via diffusion.

- Inside a p–n junction, minority and majority charge carriers coexist in the same bulk volume. This makes conventional solar cells sensitive to bulk recombination and demands the absence of any recombination centres such as trace impurities. DSCs are majority charge carrier devices in which the electron transport occurs in the TiO_2 and the hole transport in the electrolyte. Recombination processes can therefore only occur in the form of surface recombination at the interface.
- The maximum voltage generated by DSC in theory is the difference between the Fermi level of the TiO_2 and the redox potential of the electrolyte, equal to about 0.7 V (V_{oc})—it is slightly higher than V_{oc} for silicon, which has an open cell voltage equal to 0.6 V. However, the most important differences are dominated by current production.
- Compared to other solar cells, the titania solar cell has the following advantages:
- It is much less sensitive to the angle of incidence of radiation—it can therefore utilize refracted and reflected light.
- It performs over a much wider range of light conditions owing to the high internal surface of the titania (light sponge)—it can thus be designed for operation under very poor light conditions.
- It can be designed to operate optimally over a wide range of temperatures; unlike silicon solar cells, whose performance declines with increasing temperature, DSSC devices are only negligibly influenced when the operating temperature increases from ambient to 60°C.
- There is an option for transparent modules—they can be used for daylighting, roof lighting, and displays.
- DSC production needs only commonly available (nonvacuum) processing equipment, making it appreciably cheaper to set up facilities.
- DSC has a significantly lower embodied energy than all other forms of solar cell.
- Crystalline silicon PV modules are suited to full sun applications, particularly sun tracking systems and roofs, and are best suited to cold climates and clear sky conditions.
- In contrast, the DSC is particularly suited to target markets in temperate and tropical climates, because of its good temperature stability and excellent performance under indirect radiation, during cloudy conditions, and when temporarily or permanently partially shaded.
- The DSC is 7also particularly suited to indoor applications that require stability of voltage and power output over a wide range of low-light conditions.

Comparison of Different Types of Solar Cell

Table 2 compares the different solar cell types.

Conclusion

In many countries, the decentralization of power production, the increasing proportion of energy generated from renewable sources, and the development of cogenerative systems for producing heat and electricity are viewed as the paths to be taken by energy production in the future. A secure, long-term plan for energy development that takes environmental conservation requirements into account should therefore fulfill two conditions:

- The supply of power from conventional sources should decrease.
- The proportion of energy generated from renewable sources should increase.

Electricity consumption is increasing by 1% per annum in developed countries and by ca. 5% per annum in developing countries.[11] This implies the necessity to look for sources of electrical energy other than the traditional ones. Photovoltaics is one such non-traditional source, which satisfies all the criteria now required of energy sources: solar energy is universally available, and photovoltaic cells and modules are some of the environmentally safest devices for energy conversion.

TABLE 2 Comparison of Different Types of Solar Cell (copyright: ©pvresources.com)

Material	Thickness	Efficiency (%)	Color	Features
Monocrystalline Si solar cells	0.3 mm	15–18	Dark blue, black with AR coating; gray, without AR coating	Lengthy production procedure, wafer sawing necessary. Best researched solar cell material—highest power/area ratio
Polycrystalline Si solar cells	0.3 mm	13–15	Blue, with AR coating; silver-gray, without AR coating	Wafer sawing necessary. Most important production procedure, at least for the next 10 years
Polycrystalline transparent Si solar cells	0.3 mm	10	Blue, with AR coating; silver-gray, without AR coating	Lower efficiency than monocrystalline solar cells. Attractive solar cells for different BIPV (Building Integrated Photovoltaics) applications
EFG	0.28 mm	14	Blue, with AR coating	Limited use of this production procedure. Very fast crystal growth; no wafer sawing necessary
Polycrystalline ribbon Si solar cells	0.3 mm	12	Blue, with AR coating; silver-gray, without AR coating	Limited use of this production procedure; no wafer sawing necessary. Decrease in production costs expected in the future
Apex (polycrystalline Si) solar cells	0.03 to 0.1 mm + ceramic substrate	9.5	Blue, with AR coating; silver-gray, without AR coating	Production procedure used only by one producer; no wafer sawing necessary; production in band form possible. Significant decrease in production costs expected in the future
Monocrystalline dendritic web Si solar cells	0.13 mm including contacts	13	Blue, with AR coating	Limited use of this production procedure; no wafer sawing necessary; production in band form possible
Amorphous silicon	0.0001 mm + 1 to 3 mm substrate	5–8	Red-blue, black	Lower efficiency; shorter life span. No sawing necessary; possible production in band form
CdTe	0.008 mm + 3 mm glass substrate	6 –9 (module)	Dark green, black	Toxic raw materials; significant decrease in production costs expected in the future
CIS	0.003 mm + 3 mm glass substrate	7.5 –9.5 (module)	Black	Limited supply of indium in nature. Significant decrease in production costs possible in the future
Hybrid silicon (HIT) solar cell	0.02 mm	18	Dark blue, black	Limited use of this production procedure; higher efficiency; better temperature coefficient and lower thickness

References

1. Stutenbaeumer, U.; Masfin, B. Equivalent model of monocrystalline, polycrystalline and amorphous silicon solar cells. Renewable Energy **1999**, *18,* 501.
2. Abdel Rassoul, R.A. Analysis of anomalous current–voltage characteristics of silicon solar cells. Renewable Energy **2001**, *23,* 409.

3. Radziemska, E.; Klugmann, E. Photovoltaic maximum power point varying with illumination and temperature. J. Sol. Energy Eng. ASME **2006**, *128/1,* 34–39.
4. Panek, P.; Lipinski, M.; Beltowska-Lehman, E.; Drabczyk, K.; Ciach, R. Industrial technology of multicrystalline silicon solar cells. Opto-Electron Rev. **2003**, *11* (4), 269–275.
5. Lund, J.W.; Freeston, D. World-wide direct uses of geothermal energy 2000. Geothermics **2001**, *30,* 29–68.
6. Hagfeldt, A.; Grätzel, M. Light redox reactions in nanocrystalline systems. Chem. Rev. **1995**, *95,* 49–68.
7. O'Regan, B.; Grätzel, M. A low-cost, high-efficiency solar cell based on dye-sensitized colloidal TiOz films. Nature **1991**, *353,* 737–740.
8. Zhang, Z. PhD Thesis N°4066, École Polytechnique Fédérale de Lausanne: Suisse, 2008.
9. Anandan, S. Recent improvements and arising challenges in dye-sensitized solar cells. Sol. Energy Mater. Sol. Cells **2007**, *91,* 843–846.
10. Marton, C.; Clark, C.C.; Srinivasan, R.; Freundlich, R.E.; Narducci Sarjeant, A.A.; Meyer, G.J. Static and dynamic quenching of Ru(II) polypyridyl excited states by iodide. Inorg. Chem. **2006**, *45,* 362–369.
11. Muneer, T.; Asif, M.; Munawwar, S. Sustainable production of solar electricity with particular reference to the Indian economy. Renewable Sustainable Energy Rev. **2005**, *9*, 444.

25

Alternative Energy: Solar Thermal Energy

Andrea Micangeli,
Sara Evangelisti, and
Danilo Sbordone

Introduction

The sun produces energy through nuclear fusion at its core, where tremendous amounts of energy are released by the fusion of nuclei into more massive nuclei under extreme conditions such as extreme pressures and temperatures. In this way, the conversion of hydrogen nuclei into the much heavier helium causes neutrinos and photons to be discharged. This energy travels through space by radiation that is a form of energy transmitted in electromagnetic waves. The energy is transferred from the core to the photosphere through the convection zone. The photosphere, or the radiating surface of the sun, is where the energy will be radiated into space. Electromagnetic radiation is mostly emitted here, though sometimes small amounts of microwave, radio, and x-ray emissions can also be emitted. A typical photon journey lasts about 100,000 years from the core of the sun to its surface, while a photon will take only 8 min from the sun's surface to reach the Earth.

Solar intensity can be measured by the Inverse Square Law, where the intensity of radiation hitting or striking objects in space, such as planets, asteroids, or dust, can be quantitatively assessed. The total energy given off by the sun is very high, owing through the tremendous energy released by nuclear fusion. Scientists estimate the energy output at 63,000,000 W/m^2 (watts per square meter).

Given the distance of the Earth from the sun, and due to the intercepting atmosphere, this figure is definitely much lower. Radiation in the outer atmosphere amounts to approximately 1367 W/m^2. Of these, only about 40% will reach the surface of the Earth as shown in Figure 1.

The solar energy that reaches the surface of the Earth can be used by a photovoltaic system to produce electricity or by a thermal system to produce heat. When the solar energy is used to produce heat, it is called *solar thermal energy* technology. In this case, the solar radiation is used to warm up a fluid (water, air, solutions appropriated to each system) that can circulate, mainly:

- In heat exchangers at the beginning of the circuit that will use the returned available heat.
- In pipes and radiant objects put in place to warm up.
- In the refrigeration cycles to evaporate the volatile substances that are used in the condensation phase.

Through the surface, rarely adjustable, the solar energy must be accumulated and transformed into thermal energy, so it is possible:

- To concentrate it through mirrors or lenses that reflect the radiation towards panels and a boiler to be able to use directly to warm water or to the production of vapor in the pipe and into a turbine.
- To collect it from applied panels or integrated panels in closure of buildings (walls, roofs, parapet, etc.)

The most important application that can be used by everyone is a *low-temperature solar thermal technology*. It consists of a system that uses a solar collector to warm up a liquid. The purpose is to capture and to transfer solar energy to produce *domestic hot water* (DHW) or to control the temperature inside a

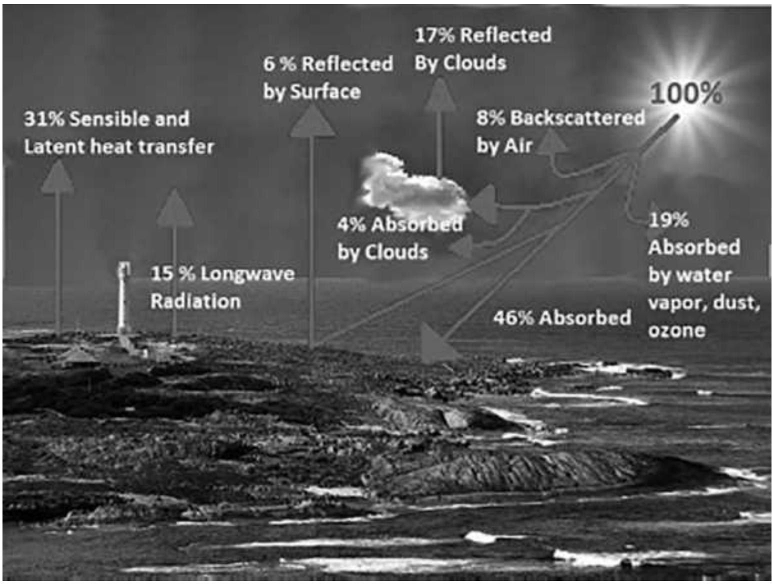


FIGURE 1 Earth's energy budget.

building. The term “low temperature” refers to the temperature that a working fluid can reach, generally up to 100–120°C.

This technology is ideal for application on a small scale, because it is cost-effective and it is simple to install and to manage. Low-temperature solar collectors can be classified as follows:

- *Flat plate collectors,* the most common type that consists of a dark flat-plate absorber of solar energy, a transparent cover that allows solar energy to pass through but reduces heat losses, a heat-transport fluid (air, antifreeze, or water) to remove heat from the absorber, and a heat-insulating backing.
- *Unglazed collectors* that are realized with tubes in plastic materials such as propylene, neoprene, synthetic rubber, and PVC. These are cheap because there is no insulation or transparent coverage. They have good performance only during summertime. They are recommended only if thermal energy for open swimming pools (and the like) is requested for.
- *Integrated collector storage,* recommended only for temperate climates, where the collector itself is the storage.
- *Evacuated tube panels*, where it is possible to eliminate the air between the capturing plate (reduced to a strip) and the evolved transparent sheet in a glass cylinder to resist the pressure difference.

The principal elements that are used by solar collectors are as follows:

- *The capturing plate*
- *The insulated material*
- *The transparent coverage*
- *The external casing*
- *The working fluid*

Solar Collector Technology

Flat-Plate Solar Collector

The *capturing plate* is realized with copper or steel and it is treated with satin and dark paint to reduce the reflection losses to bost the absorption capability to the wavelength of solar radiations with a low emissivity in the infrared radiations.

Normally, the canalizations on the plate are built to resist a pressure of about 6–7 bar; some collectors guarantee the resistance to a pressure of up to 10 bar.

The *insulated material* is a barrier against the conduction losses of the plate toward the external part of the collector. The materials used are always characterized by a porous or alveolar conformation in order to create microscopic motionless air spaces (that constitute a perfect barrier to the heat transmission). Polyurethane, polyester wool, glass wool, and rock wool are the most used materials. An enemy of the insulating material is humidity, which can appear for many reasons inside the collector (moisture, rain due to the gasket caused by little leakage in the pipes); often, the above-mentioned materials are covered with a thin aluminum layer that acts as a barrier to humidity and, at the same time, reflects towards the absorption plate.

The *transparent coverage* has the dual function of limiting the loss of energy towards the outside of the collector and facilitating the penetration of the radiation inside the collector. To satisfy this request, the coverage should be the most transparent as possible to the wavelength typical of the solar radiation (approximately between 0.2 and 2.5 mm) and at the same time it should be matt to the infrared radiation coming from the pipe-table while their temperature rises. The material that meets these qualifications is glass, above all if treated to acquire more transparency. Sometimes, due to some adverse elements, such as fragility and weight, plates made of plastic materials (polycarbonate) are preferred to glass plates.

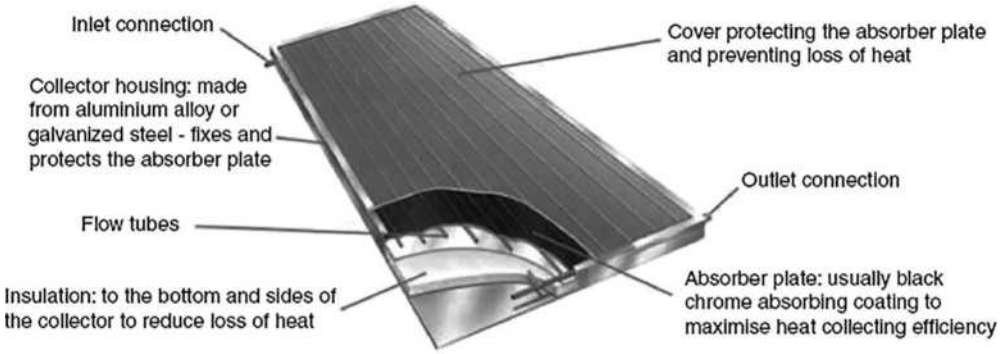


FIGURE 2 Flat plate solar collector scheme.

The *external casing* has the dual function of contributing consistency and mechanical solidity to the collector and protecting the internal elements from dirt and atmospheric agents. It can be made of stainless steel (zinced) or aluminium.

The *working fluid* that flows along the pipe system must take the largest quantity as much as possible. The fluid should also have a high density even at high temperature. It is important that it does not have a corrosive effect along the wall of the circuit; it must be inert and stable at temperatures below 100°C, and it should also have limited hardness to avoid limestone deposits. The hardness of the water refers to the quantity of magnesium and calcium salts in the water. The fluid should have a low freezing point. The option used among the producers of solar panels is a water solution of propylene glycol (not toxic and has a good anti-freezing action). In Figure 2, a section of a flat plate collector with all components is shown.

Unglazed or Open Collectors

An *unglazed collector* is a simple form of flat-plate collector without a transparent cover. Typically, polypropylene, Ethylene-Propylene Diene Monomer (EPDM) rubber, or silicone rubber is used as an absorber as shown in the left panel of Figure 3. Used for pool heating, they can work quite well when the desired output temperature is near the ambient temperature (i.e., when it is warm outside); moreover, they are cost-effective. As the ambient temperature gets cooler, these collectors become less effective. They can be used as preheat make-up ventilation air in commercial, industrial, and institutional buildings with a high ventilation load. They are called “transpired solar panels,” and they employ a painted perforated metal solar heat absorber that also serves as the exterior wall surface of the building. Heat conducts from the absorber surface to the 1mm thick thermal boundary layer of air on the outside of the absorber and to air that passes behind the absorber. The boundary layer of air is drawn into a nearby perforation before the heat can escape by convection to the outside air. The heated air is then drawn from behind the absorber plate into the building’s ventilation system (Figure 3, right panel).

Integrated Collector Storage

An *integrated collector* storage system is constituted by a unique element that assumes the role of capturing plate, absorber, and external accumulation. In this type of solar collector, the storage is located in the collector itself, and it is exposed to a slow heating process.

The water is placed inside the insulated tank and in this case the heat losses due to the exposure surface cannot be ignored. The integrated collector storage works without pumps or electrical devices: the panel absorbs the solar energy and the water inside the collector rises up for conduction, moving towards the outlet pipe and reaching the domestic network when hot water is required (see Figure 4).

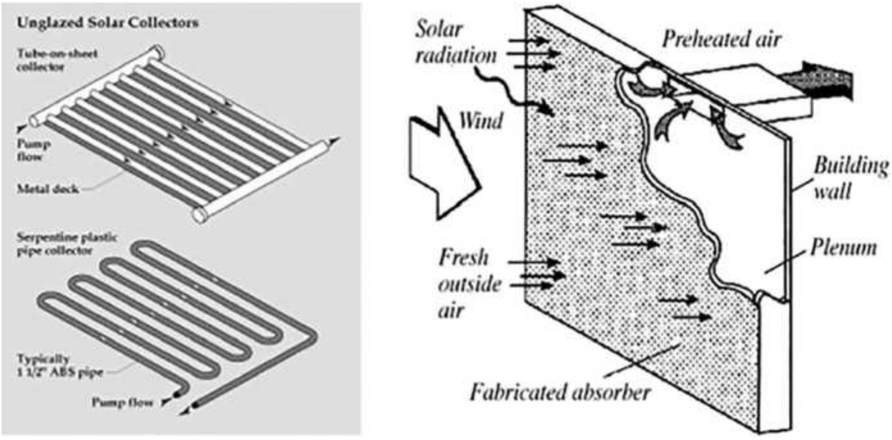


FIGURE 3 Unglazed solar collector scheme.

FIGURE 4 Example of an integrated collector storage system.

The performance of such equipment is not fully satisfactory, because during the discharge phase, the water temperature decreases rapidly, reducing the overall usability of the collector.

Evacuated Tube Collectors

The *evacuated tube collectors* are obtained, reducing the presence of air in the space between the plate and the transparent cover, thus avoiding losses caused by convective movements. In spite of their higher cost, these collectors are able to perform well even when the environment temperature is low. Among the possible technological solutions that can be used to build these collectors, the *heat pipe* technology does not limit the exchange of heat between liquids but in the case the fluid flows in thin tube pipes, also between vapor and liquid, taking advantage of the heat for the condensation along its way, resulting more efficient although more complex. Taking advantage of this type of heat exchange, the pipe system where the thermal vector liquid flows should be depressurized to decrease the evaporation temperature. A collector is built by the plate in long metal cylinders (copper) superficially covered with black and selected paint; this tube is in a second glass tube, in a way that surface of the first one is perfectly tallying with the internal of the second one; it is concentrically inserted in a second glass pipe. Figure 5 shows a section of the collector.

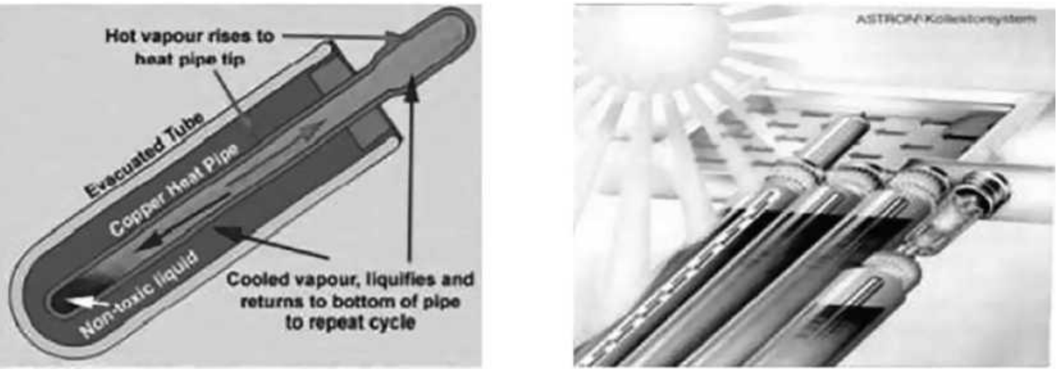


FIGURE 5 Section of an evacuated tube collector.

The air between the two tubes is vacuumed out, until a pressure of $P = 5 \times 10^{-3}$ *Pa* is reached. A small tube goes lengthwise through the copper cylinder following a U path, inside the thermal vector fluid flows reaching a temperature close to 100°C. In order to maximize the use of the heat pipe, a bigger pipe can be used in a concentric position; the heating of the fluid in the pipe will increase because the captured radiation increases. Heater exchangers are placed at the ends of the pipes, in order to transfer heat to the users. The evacuated tube collectors are divided into two main types:

1. *Evacuated pack collectors* with direct circulation of thermal vector fluid.
2. *Heat pipe evacuated pack collectors:* the fluid inside the pipe system evaporates along the way and give its heat, due to the condensation process.

In the first type, the plate is divided in long metal cylinders superficially treated with selected black paint, with each of these tubes inserted in a glass tube, which is also inserted in a larger glass tube, and then vacuum packed. A small tube goes through the copper cylinder following a U path.

In the heat pipe evacuated pack collectors, the little tube under pressure (heat pipe) receives heat from the capturing plate. The tube contains water or alcohol that evaporates at around 25°C under pressure. The vapor goes up until the head, where it exchanges heat through the condensation phase giving heat to another external fluid.

Energy Balance of a Collector

The phenomena that interact within a manifold are multiple and interconnected: the main energy exchanges between solar radiation and the various elements of the system are here described.

As shown in Figure 6, the solar radiation (E_0) hits the glass cover. A small amount of radiation (E_1) is reflected and absorbed by the transparent cover. The copper absorber does not absorb all the remaining radiation into useful heat and partly reflects and dissipates heat (E_2) by convection, conduction, and radiation to the outside. Treatment with selective coatings, as mentioned above, reduces leakages. On one side, the transparent cover prevents the reflection of the solar radiation from the plate to disperse outwards favoring the greenhouse effect inside the collector, and on the other side, it limits the heat convection dispersion (Q_1). If a good thermal insulation in the back and sides of the collector is designed, with standard insulating materials such as rock wool or polyurethane foam, the energy losses by thermal conduction are reduced to a minimum (Q_2).

Only a portion of the incident solar energy (E_0) is transferred to the fluid as useful heat (Q_3) due to the different energy losses (E_1, E_2, Q_1, and Q_2).

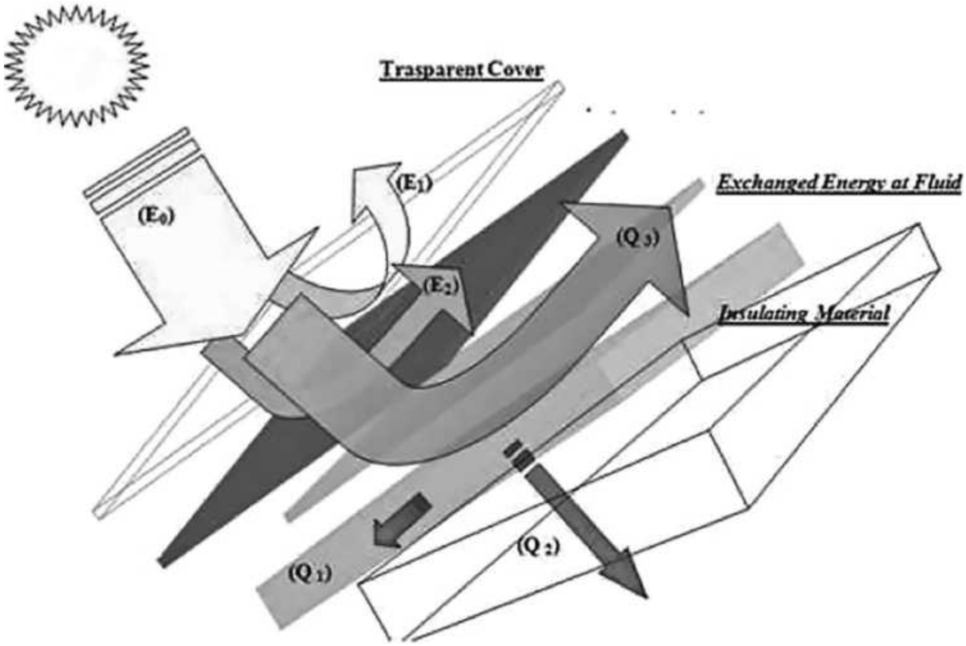


FIGURE 6 Energy balance of a solar thermal collector.

Collector Efficiency Curves

The thermal energy transferred to the fluid per time unit is calculated as the difference between solar radiation captured by the plate and converted into heat, and the heat losses by convection, conduction, and radiation.

$$\mathrm{Q_n} = E_c - \mathrm{Q_p} \tag{1}$$

E_c takes into account the absorbance and transmittance of the glass plate, and it is calculated as the product of the irradiance E, the transmittance τ, and absorbance of the plate α, as follows:

$$E_c = E^{*}\tau^{*}\alpha \tag{2}$$

The heat loss depends on the temperature difference between the plate and the environment. As a first approximation (for low temperatures of the plate), this relationship is linear and it can be described through the coefficient of total losses:

$$Q_p = k^{*}\Delta T \tag{3}$$

where $\Delta T = T_p - T_a$, T_p is the average temperature of the plate, and T_a is the ambient temperature.

By substituting these relations in the efficiency of the collector, the following are obtained:

$$\eta = \left(E^{*}\tau^{*}\alpha - k^{*}\Delta T\right)/E, \tag{4}$$

$$\eta = \left(E^{*}\tau^{*}\alpha / E\right) - \left(k^{*}\Delta T/E\right), \tag{5}$$

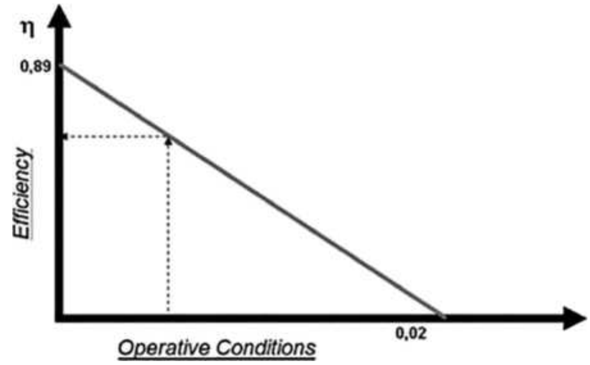


FIGURE 7 Efficiency of a solar thermal collector at different operative conditions.

and

$$\eta = \tau^{*}\alpha - \left(k^{*}\Delta T/E\right), \tag{6}$$

Whereas

$$\eta_0 = \tau^{*}\alpha. \tag{7}$$

It follows that the efficiency of the collector is equal to:

$$\eta = \eta_0 - \left(k^{*}\Delta T/E\right). \tag{8}$$

Conventionally, it is defined:

$$\Delta T^{*} = \left(T_{\mathrm{mf}} - T_a\right)/E. \tag{9}$$

Finally, the efficiency of a collector can be expressed as:

$$\eta = F'^{*}\eta_0 - K\Delta T^{*}. \tag{10}$$

In Figure 7, the efficiency curve of a collector is shown.

Instantaneous Efficiency of a Collector

The instantaneous efficiency depends on the optical losses (E_1 and E_2) and on the temperature (Q_1 and Q_2).

The overall losses of the total heat occurring in the collector by conduction, convection, and radiation can be expressed as the coefficient of total loss in K (W/m^2 * °C).

The graph shows that, with constant irradiance, the higher is the difference between the average temperature of the fluid and the ambient temperature, the heat losses also increase and, consequently, the efficiency of the collector decreases.

Comparison between Different Types of Collectors

By comparing the two efficiency curves, the principal characteristics of the collectors can be classified:

- *No glass collector* is the one with the best possibility to absorb the incidental radiations; its efficiency though decreases fast until zero in situations where the other collectors still have valid performances.
- *Flat-plate collector* with a selective plate has a better performance than the open one, practically in every working condition.
- *The evacuated collector* has a more stable efficiency curve, and it guarantees good performance even during bad weather conditions.

From Figure 8, which shows the efficiency curve of different collector types, it is possible to observe that the unglazed collectors have better optics than others. In fact, the absence of covering helps eliminate the untransparency losses and reflection that often occur with the glass covering.

Natural Circulation Systems

Once the differences between an open and a closed circuit have been examined, it is possible to analyze the circulation of the thermal vector fluid of the plant.

Natural circulation and closed circulation systems are based on the convective movements flown from the thermal vector fluid caused by a difference of temperature in the fluid itself. Indeed, the fluid warms up inside the serpentine of the capturing plate and it naturally goes up to the top of the collector. It will need an inclination compared to the floor to maximize the quantity of transferred energy.

Over the collector, a “storage tank” is placed with the heat exchanger inside. In the tank for the accumulation, there are two separate flows: the closed circuit of the collector with the thermal vector fluid and the water net of the running system designated to the final users.

The water, in contact with the heat flow, has a lower temperature, with higher heat absorption capacity, thanks to its stratification. While liberating heat to the running water, the fluid cools off moving to the lower part of the capturing plate; meanwhile, the part of the fluid that was at a colder temperature, being at this point at a higher temperature, tends to go up to the accumulation tank, cooling off while transferring heat to the running water. The quantity of fluid inside the solar collector remains constant and does not need any regulation pump for the circulation of the water, because a self-regulation natural mechanism works thanks to the trigger of the convective movement (see Figure 9).

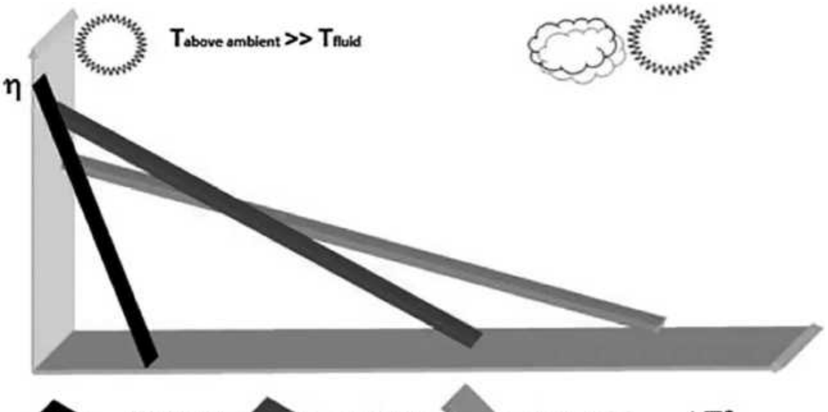


FIGURE 8 Efficiency curves of different collector systems.

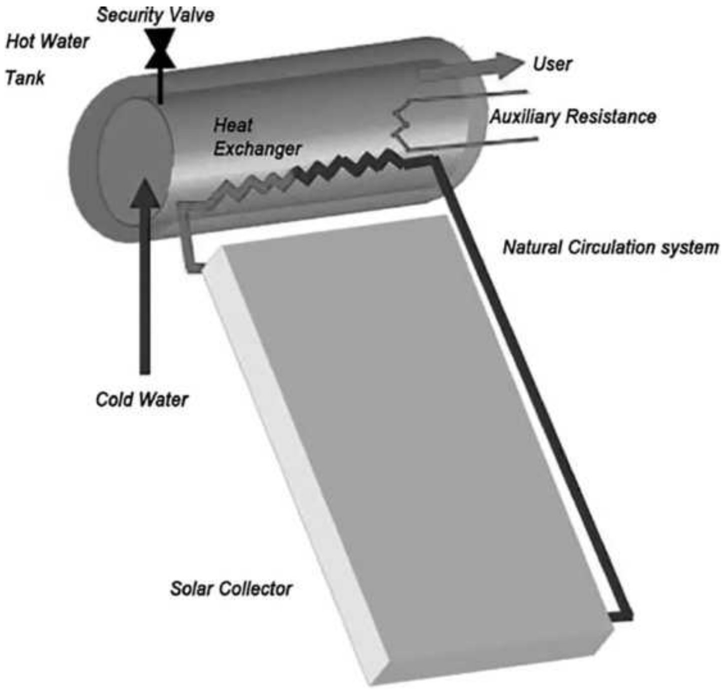


FIGURE 9 Natural circulation system.

Forced Circulation Systems

In the system where the circulation is forced, the presence of the circulation pump, driven by a differential thermostat, allows the fluid to circulate inside the pipes. After a selected difference between the water temperature and the fluid, it activates automatically. The tube in which the thermal vector fluid circulates represents a primary circuit. In the highest part of the accumulation tank, there is an exchange of heat coming from an auxiliary traditional circuit that starts functioning when the water temperature designated to the users does not arrive to the requested one. The fluid that liberates its heat inside the tank of accumulation cools off through a pump of circulation that is sent to the lower part of the collector usually sited on the roof. See Figure 10 for the scheme.

Open Circuit Systems

Open circuit systems have some advantages, such as the simplicity of the hydraulic circuit realization and the lowest thermal losses that always occur when heat goes from one fluid to another. There are two problems that put limitations to this type of plant: the possible freezing of the water in case the temperature reaches values below 0°C and the calcium deposits along the tube system of the collector. In both cases, the collectors could go out of service.

Closed Circuit Systems

The closed circuit is the most common solution. In this case, there are two different hydraulic systems: the main one, in which the thermal vector fluid circulates, and the secondary one, where the water coming from the hydro net is used (a third circuit is foreseen in the large-scale plans for the heat storage).

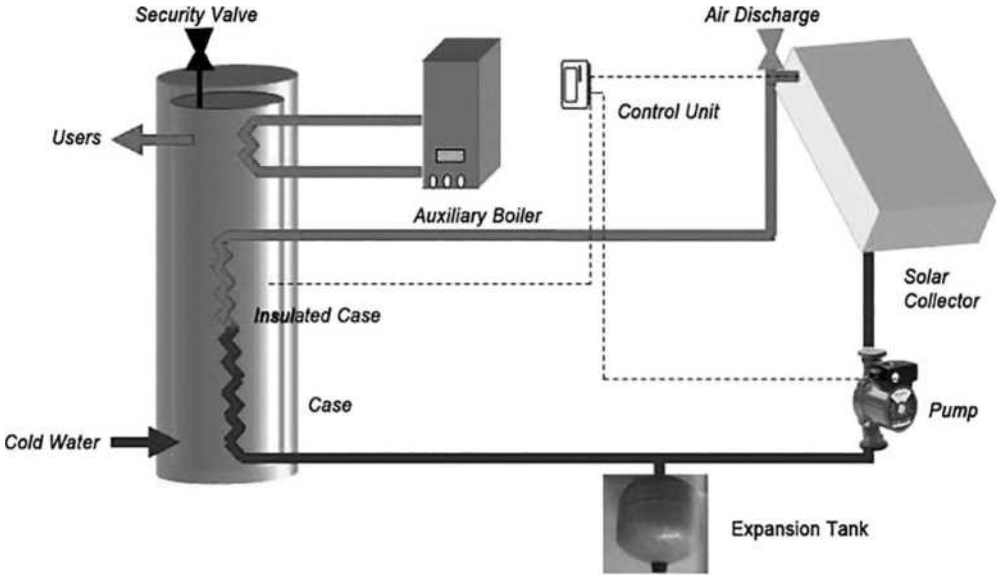


FIGURE 10 Forced circulation system.

The thermal vector fluid must respond to specific functions:

1. To increase density and specific heat capacity to be able to use the smaller pipe systems.
2. To avoid limestone deposits due to hardness of the water.
3. To reduce freezing points and viscosity.
4. To be not toxic (in the case of plant for sanitary warm water).
5. To be chemically inert, stable, and not corrosive.

The option adopted by most of the producers is a solution of water and polyethylene glycol (usually 25%–45% of glycol).

The purpose of the thermal vector fluid is to take thermal energy captured by plane and transfer it to the water to heat up. In this energy transfer, the fluid can give some heat to the cold water through the exchanger in a proportional measure depending on the difference of temperature between the two fluids. In Figure 11, the plate exchanger of heat is shown together with the tank with a double exchanger and a serpentine plunged. The larger the interface, the greater the amount of energy exchanged. The differences in temperature are of the utmost importance. To unify the need of big exchange surfaces with compact exchangers, the immersion serpentines are used.

Solar Cooling

The physical principle of generating solar cooling power is almost similar to the operating principle of conventional air-conditioning systems (condenser-compressor type) air-conditioning. Both systems rely on systems full vacuum single glass heat pipe collector picture liquid-to-vapor phase change energy of the refrigerant to attract heat (i.e., to produce cooling effect). The way the two systems achieve this is quite different because the condenser machine achieves a cooling effect by expanding compressed refrigerant into a low-pressure chamber, while the solar cooling machine relies on the absorbing action of the absorbent to create near-vacuum inside its chamber. In near-vacuum, the refrigerant will evaporate at a very low temperature, removing latent heat from the refrigerant (i.e., producing a cooling effect). This happens at a temperature significantly lower than the refrigerant's evaporation temperature

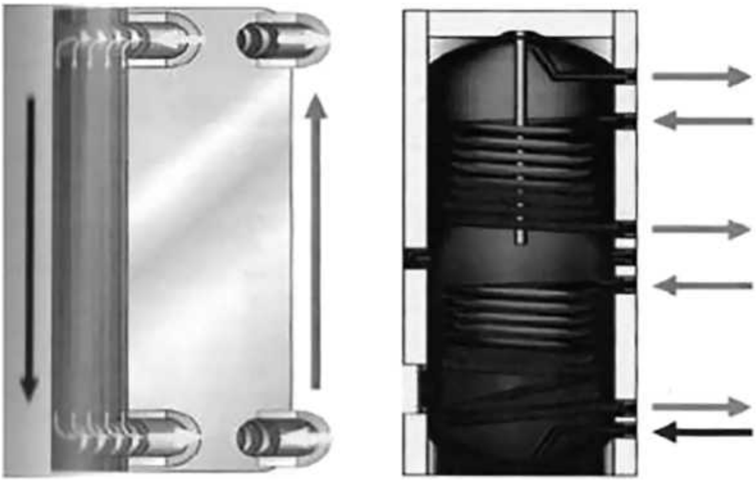


FIGURE 11 The plate exchanger of heat is shown together with a tank with a double exchanger and a serpentine plunged.

at atmospheric pressure. A cooling effect is thus achieved at usefully low refrigerant temperatures, making this principle practical for commercial use.

Heat supply from a field solar collector to the solar air-conditioning system is required not to directly provide the cooling action directly but to maintain the absorbent concentration. This ensures that low chamber pressure and low evaporation temperature of the refrigerant are maintained.

Preliminary Analysis and Solar Thermal Plant Design

Due to the abundance and benefits of solar energy, solar thermal plants can be very useful especially if there is a willingness to invest some capital into them, but the choice should be made taking into account the quality of the project and the devices.

The covered need can be high (60%–70% or more), for various reasons such as the randomness of the sunshine and the urban or market situations; it will not be able to meet the demand of the users. Most of the time, the plant will limit the use of fuel for heat production.

In the supply of DHW (domestic home water) through this type of plant, it must take into consideration that the weather conditions will not always be able to provide a sufficient quantity of energy to satisfy the demand and the complete comfort of the users. Demand and supply of heat will not be equal at the same time, and so to obtain independent plants in an annual scale, the capturing surface should be overestimated for summer (to supply the energy needed during winter).

Nowadays, investing in large seasonal storage is still in an experimental phase and involves very interesting cases. Figure 12 shows the offset of the thermal charge compared to the availability of the energetic source as the solar radiation incident. The latter is present in a solar thermal plant.

Matching Energy Availability and Thermal Energy Need

The way to estimate the energy needed to consume hot water is simple, but it is important to pay attention to the details involved.

Previously collected data can be used, but it is important to verify who collected them and what the original purpose was. If purposes are not reliable, it is necessary to go through monitoring to carry out data collection using a manual or digital device (calorie counter or a simple flow meter), gathering the data daily for at least 1 mo. In the sizing of the solar plants, the most important element is the energy

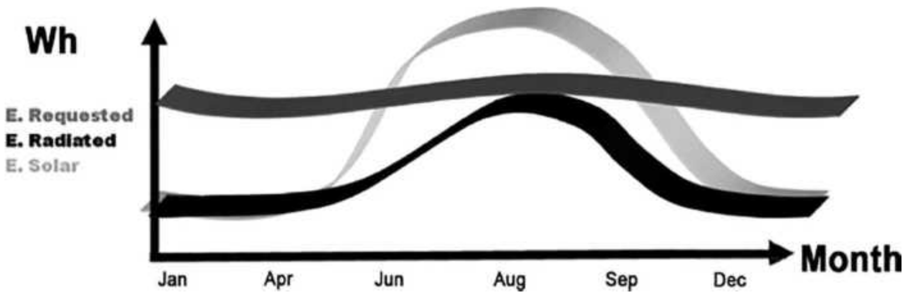


FIGURE 12 Offset of the thermal charge compared to the availability of the energetic source as the solar radiation incident.

need that will be accumulated and not the power need, which will be obtained at the right moment, from the mix of the solar heat storage and the auxiliary heater. Moreover, it is worth to invest on a thermal solar plant, as in all the renewable energies, only after a strong initiative to reduce the energy consumption, for example the aerator to be put at the final tap, can reduce up to 50% of the consume of hot water.

The Design Phase

The main design activities for a thermal solar system will be described here. The parameters and the different components that must be chosen for the plant and its dimensioning will be analyzed by the on-site procedure. It is important to note that in order to obtain the estimation of the costs of a solar plant, there is no need to proceed with detailed dimensioning, but it is sufficient to make an estimation of the following data:

- The surface of the collectors
- The volume of the tanks
- The thermal energy needs of the user

The phases of the project can be organized in the following way:

1. On-site investigation
2. Selection of the typology of plant and collectors
3. Dimensioning of the capturing surface
4. Dimensioning of the different components

On-Site Investigation

This is the first step in the design of thermal solar plants that includes three main targets:

- A detailed analysis of user energy consumption to estimate energy need (in joules, kilowatt-hours, or calories).
- A set of possible solutions to reduce the consumptions through energy saving.
- A check on possible solutions considering the logistic realization of the plant, with particular reference to the collectors' colocation.

Analysis of the Users' Consumptions

The first step toward a correct project is to estimate the exact heat consumptions; this allows the dimensioning of the solar system to satisfy the user energy needs through solar energy. There are two rapid assessment methods through which the evaluation of the users on thermal energy consumption is made:

1. The study of energy bills of the previous years
2. The study of the consumption habit of the users

Saving Energy Interventions

The dimensioning of a solar system cannot be based exclusively on the previous analysis; it needs to examine the opportunity to minimize each and every cost, to utilize clean energy after dispersion and waste reduction. Saving energy represents an objective of main importance, which can be realized in two ways:

1. Working on the demand level, boosting the users to modify their habits (this is called energy sacrifice), not always accepted but free of cost.
2. Working on the level of the offer, promoting the substitution of old and common devices with high efficiency products having the same performance but lower consumption.

The realization of a recovery intervention implies almost all the time an economic investment, the convenience of which must be analyzed before comparing the saved energy that would be obtained. It has to be found out *how much is saved*.

In the first case, the person who is taking care of the project must underline the waste, applying devices of energy saving. For example, in the case of DHW, both flow reducers and an air–water mixer limit the flow of hot water. In this way, users have the impression that the jet of water is the same although the quantity is much lower.

Logistic Aspects

The purpose of the on-site visit is to analyze some logistic aspects because it is very important to evaluate the feasibility of the solar plant. This phase is often ignored but very important. The first examination should verify if any historical constraint, related to the buildings or to the landscape in the area, exists. There could be situations where it is not possible at all to install a solar plant because it should be a waste of time to proceed. Under a technical point of view, it is necessary to find out a free area in which the solar panel could be installed, on the surface of which it is necessary to measure:

- Typology and material of the surface (characteristic of the roof, for instance)
- Gross area available
- Obstacles (antenna, chimney, others)
- Shade elements nearby or far away (buildings, trees, etc.)
- Accessibility for the installation and for the following maintenance operations
- Azimuth (orientation compared to the south) and inclination compared to the horizon

With reference to the specific problem of the shade, it can be useful in areas that are growing, to evaluate the local urbanization policies on the short and medium terms and to avoid that some years after the installation, the panels might fall in the shade of a new building.

Choice of Solar Plant Type

From an initial analysis, it is possible to distinguish the solar thermal plants on the basis of the hydraulic circuit. In an open circuit, the thermal vector fluid coincides with the thermal vector fluid used by the final users. A closed circuit consists of three hydraulic circuits divided into

1. Thermal vector fluid
2. Accumulation fluid
3. Water from the user

Analyzing the system circulation two categories of plants can be underlined: systems of natural circulation and systems of forced circulation. In most cases, the plant is designed with a closed circuit because there is the possibility of using a thermal vector fluid different from simple water.

The use of water as a fluid has some disadvantages like the presence of calcium and a low freezing point on one hand, and the open circuit is more economic and rapid to be installed on the other hand.

Estimation of the DHW Need

The production of DHW represents one optimum solution for the use of solar energy because it allows using solar radiation also when it is at its maximum power. A reason that has increased their new diffusion is the economic return in terms of capital investment due to the costs relatively contained for the installation and the good reliability of the plant. The evaluation of the theoretical thermal need in the systems of DHW can be quantified through the following expression:

$$Q_{\mathrm{ac}} = V^{*}\left(t_{\mathrm{ut}} - t_{\mathrm{al}}\right)^{*} C_{s} \tag{11}$$

where Q_{ac} is the needed daily thermal energy (kcal/g), V is the requested water volume (L/g), t_{al} is the temperature of water source (°C), t_{ut} is the temperature of water output (°C), and C_{s} is the specific heat (kcal/kg * °C).

The supplied temperature depends on the place of the plant installation and on the water net from which it is taken, as it is possible to have differences of temperature depending on the period of the year. The temperature of warm water supply depends on the user considering that the maximum temperature of the water to the thermal generation should be $t_{\mathrm{ut}} = 48 + 5$°C, where the second term indicates the maximum tolerance. To be able to guarantee this temperature until the water reaches the user, it is necessary that the distribution system must be suitably insulated. For the buildings that have no isolated distribution system, the temperature of supply would be higher than 50–55°C with peaks of 60°C.

The effective calculation takes into consideration the thermal losses that occur in the distribution network that brings the water from the production point to where it should be used. These losses make one think of the necessity of an effective thermal load higher than the theoretical one, to be able to satisfy the users' needs.

The total output of the distribution network depends on the grade of the insulation of the network itself. The average of the final output can have values in the following ranges:

- 0.85 to 0.90 in the case of recent constructions with insulated pipe system and with the recirculation.
- 0.75 to 0.85 for the constructions with plants without recirculation or with a cycle that works only during the day.
- 0.65 to 0.705 for the construction of plants with a working recycle during full time.

Sizing the Collector Field Surface

The first step to designing a solar plant system is to evaluate the surface needed for the plant as a compromise between the technical need and the economic need: unless a solar roof is used, it is necessary to think about the tilt angle and its right positioning. Basically, the exposure should not face north and the surface should be enough to have all the collectors needed. That surface of collectors must avoid getting over economic convenience, even only on log term. The dimension of the capturing also depends on the type of chosen collectors, particularly on their performance and on the orientation of the tilt in the roof.

The quantity of energy on the surface of the collectors varies in function of two angles: the azimuth angle θ and the tilt angle φ. The azimuth angle θ is the projection on the horizon plane surface and the tilt angle φ defines the inclination of the collector to the surface as shown in Figure 13.

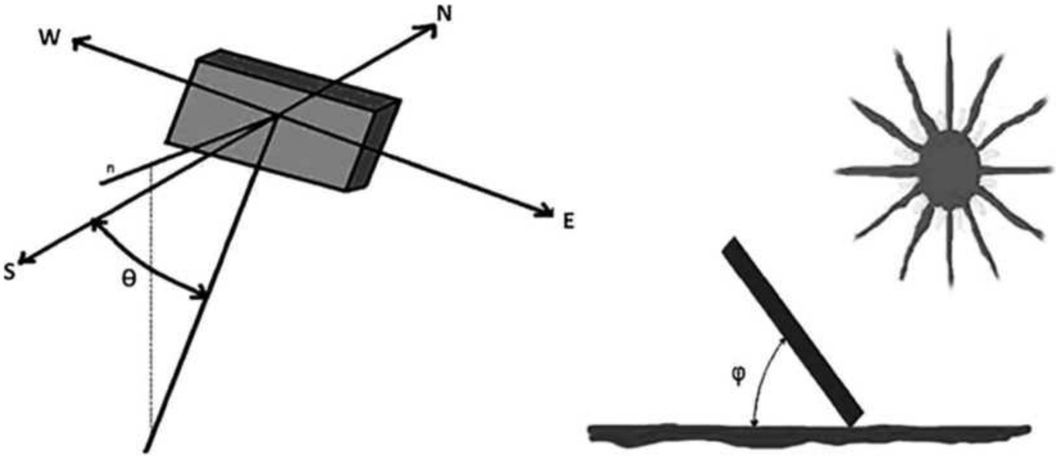


FIGURE 13 Azimuth and tilt angles.

TABLE 1 Increment of the Surface of the Collectors under Suboptimal Conditions

	Tilt Angle						
	0°	**15°**	**30°**	**45°**	**60°**	**75°**	**90°**
South	12%	3%	0%	1%	8%	20%	45%
South/East or South/West	12%	6%	3%	5%	11%	23%	43%
East and west	12%	14%	15%	20%	28%	41%	61%

The tilt angle depends on the building on which the plant is installed and on the orientation of the building itself as well. Since it is often not economically and aesthetically convenient, they are fixed directly on the top of the roof to create an ad hoc structure to hold the collectors.

About the typology of collectors, the choice depends on the conditions of the performance that will influence the output, and so from

- The internal temperature: the temperature to which the water is heated up.
- The external temperature: it depends on the period of the year in which it will be used.

The most important data (for the output) are the difference between the temperatures of the collector and the external environment because it characterizes the thermal losses. If data on consumptions are not yet available (for example, the building has just been built), an estimation on the energy needed (referring to average values) must be done. The corrections match the increases of the surface in bad orientation conditions, as seen in Table 1.

For combined plants, used for house heating and for the production of DHW would be appropriate to increase the value of the inclination to reduce the difference between the summer and winter production as underlined in Figure 14.

Heat Storage System

The energy needed for a large number of applications depends on the time, but often in a different way from solar energy. As a consequence, it is necessary that the accumulation of thermal energy produced with solar panels firstly in those cases in which the solar energy must cover an important fraction of consumes. The optimal skill of the heat storage system depends on

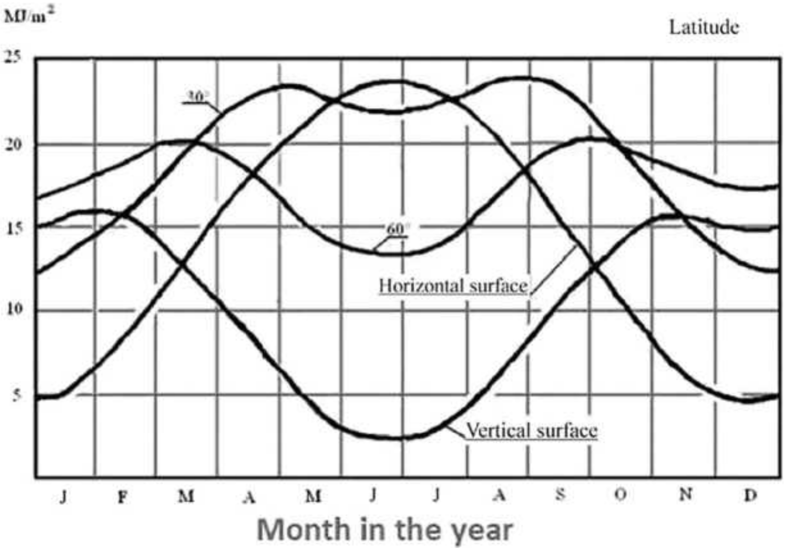


FIGURE 14 Monthly average solar radiation during a year at 40° latitude.

1. The availability of solar energy.
2. The nature of the loads.
3. The level of reliability needed in supply processes.
4. The economic analysis that sets the optimal percentage of traditional means.

Normally, a thermally isolated tank and a boiler are used. The boiler should store the solar energy when available and give it back when requested.

The most common energy discrepancy include alternation of day and night, sunny days as against cloudy days, and the summer season compared to the winter season. The latter is very important under a technical point of view because it reduces the technical problems connected to the summer overproduction as well.

A typical situation is a plant with accumulation. Paste Figs. 9 and 10:

Heat Storage Systems Dimensioning

Considering the time course of thermal loads and solar energy available:

- G_c represents the incident solar power.
- q_u represents the power delivered to the fluid.
- L is the power required by the loads that vary over time.

The areas above the red line and below the green one represent the time intervals that exceed the energy needs and must be accumulated. In contrast, the areas above the line in red and green show the time intervals in which the heat must be supplied to the loads from the storage system as in Figure 15.

The storage systems can be divided into two broad categories:

- Sensible heat storage systems
- Latent heat storage systems

In the sensitive system, energy is stored by raising the temperature of a suitable material. Latent heat storage systems take advantage of the latent energy at the phase change (usually the liquefaction) of a substance, and in this case, the process takes place at constant temperature.

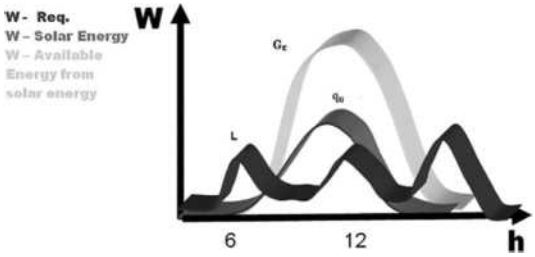


FIGURE 15 Energies during the day for a typical plant with storage.

In the first type, energy is taken by decreasing the temperature of the substance in storage while in the second type, stored energy is made available, causing a phase change.

In the build-sensitive systems, adopted in most cases, the energy is proportional to the temperature change of the substance contained in the batteries themselves:

$$Q_s = Mc_p\Delta T_s \tag{12}$$

where Q_s represents the accumulated energy, M is the mass storage, c_p is the specific heat of substance accumulation, and ΔT_s is the temperature variation.

Often a single temperature does not characterize the accumulation, because it cannot be considered perfectly mixed; it is actually layered. The top tank is the hottest part and the bottom is colder. The stratification is beneficial because of the collectors; reducing the average temperature of the absorber plate improves the efficiency to capture solar energy.

The best storage material in liquid systems is certainly water, since it has a low cost, has high specific heat, and is not toxic, and its boiling temperature at atmospheric pressure is high enough. The size of the storage system depends on the absorbing surface. Recommended values for solar systems for hot water are 100 L/m^2. The most important aspect to consider is, as already mentioned, the stratification of the water tank. As an index to assess the extent of the stratification of the water in a reservoir, the extraction efficiency is defined as:

$$\eta = \left(Qt^{*}\right)/V \tag{13}$$

where Q is the volumetric tank drain, V is the volume of accumulation, and t^* is the time required for starting from a completely mixed storage; the temperature difference of input-output has fallen to 90% of initial value.

Since the tank has a considerable cost, mainly due to heat insulation, sizing is necessary to make technical evaluations of economics. The storage tanks can be recharged in a sunny day or in an entire season. The first are those most commonly used and consist of an insulated tank to maintain hot water temperature. The degree of isolation of the accumulation should be as high as possible if the tank is installed outside, as often happens in systems with forced circulation. The storage tanks must also possess other important characteristics:

- They must be suitable for containment of potable water and must have had internal anticorrosion treatment.
- They must be resistant to high pressure (6 bar).
- They must be equipped with safety devices such as air vent, expansion vessel, and safety valve.
- They must be equipped with the following measuring devices: temperature gauge and pressure gauge for measuring pressure.

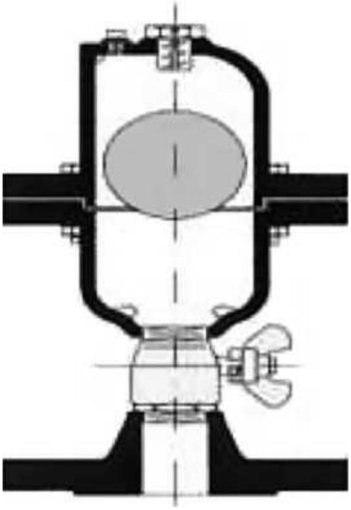


FIGURE 16 Example of an expansion tank.

When the heat demand is roughly constant during the day, usually about 20 L/m^2 of collector is sufficient, while this figure varies between 50 and 100 L/m^2 of collector daily batch loads such as residential use, where consumption of water for showers are concentrated at certain times of day (morning or evening).

An important phenomenon is the *stagnation* of the system. In summer, the energy produced is often greater than the amount of thermal energy users and, if the accumulation is too small or absent, the excess energy increases the temperature inside the collector to allow evaporation of the fluid heat transfer. This leads to the thermal gradients and overpressure under serious problem the hydraulic components of the system, and it can compromise the integrity of the plant. Also, bear in mind that the plant also has air release valves to eliminate the air that enters the piping system itself, inevitably compromising the thermal exchanges and thus the operation. If it is used, an automatic valve (Jolly) is strongly discouraged, due to the stagnation of the primary because the emptying of the liquid transfer medium such as air comes out unwanted in automatic air vents. The sizing of the accumulation is performed for the reasons outlined above, depending on the surface of the manifold: in practice, it takes an average of 50 to 90 L of storage per square meter of collector area (see Figure 16).

Size of the Auxiliary

The sizes of the auxiliary are dependent on the diameter pipes that are obviously related to flow values (Table 2).

Another very important element, which has a major impact on system performance, is the pressure drop. The volume flow can be calculated by the following equation:

TABLE 2 The Size of the Pipes according to the Scale

Flow (L/hr)	External Diameter per Thickness (mm)
<240	15 × 1
240–410	18 × 1
410–570	22 × 1
570–880	28 × 1.5
880–1450	35 × 1.5

$$P_v = Q/\left(c_g \Delta T\ m_v\right) \tag{14}$$

where Q is the thermal power made available by the solar collector (W/m^2); $c_g = 1.3$ Wh/(kg °C), the specific heat of the fluid; m_v is the density of the fluid (1 kg/L,); and ΔT is the temperature difference between inlet and outlet collectors (10°C).

The section of pipe can be calculated from the volumetric flow rate and velocity of the fluid P_v.

Prevention and Control of Legionella Exposure Risk

A problem that needs to be focused on in water systems for large facilities (hotels, prisons, hospitals, etc.,) is the presence of *Legionella*.

During the design phase of some facilities, a plant configuration to eventually prevent *Legionella bacteria* has already been furnished. Legionnaires' disease is contracted by breathing, by inhalation, or by aerosol microaspiration that contains the bacterium. The aerosol is formed by the droplets generated by water spray or the impact of water on solid surfaces. Most droplets are small; the more dangerous water droplets are those with a diameter less than 5 microns as they can more easily reach the lower respiratory tract. To ensure a reduction in the risk of legionellosis, the following preventive measures should be followed:

- Always keep hot water at a temperature above 55°C.
- Slide the water (either hot or cold) taps and showers that are not used for a few minutes at least once a week.
- Keep showers, jet showers, and speaker of the taps clean and free of fouling; replace as needed.
- Clean and disinfect all water filters regularly every 1–3 mo.
- Ensure that any changes made to the system, including new installations, do not create dead arms or pipes with no water or stream flows intermittently.
- Adhering to a point, the aim is to ensure a continuous thermal disinfection of hot water.
- Heating the water to 55°C results to, in fact, short-term elimination of *Legionella bacteria,* as evidenced by the Hodgson-Casey diagram taken as a point of reference for setting temperature and the time needed to implement thermal disinfection against *Legionella*.

However, in order to properly set up a continuous thermal disinfection, it is necessary that the water plant is at least maintained at 55°C.

It should be considered, in fact, that if the thermal head high facilities are not well insulated and well balanced between supply points and certain areas of the circulation thermal head high can be determined, as shown in the diagram below (Figure 17). So, even if hot water is supplied at a temperature below 50°C, there may be circulating in the networks of temperatures and can encourage the development of Legionella.

Large Systems

Solar systems are designed to be large to produce hot water for space heating. Possible applications include the following (see Figure 18): hospitals, home care for seniors, military barracks, hotels, gyms, prisons, and residential complexes—situations where the demand for heat is constant both during the day and at different seasons of the year.

They are called large systems when the surface of the solar field is greater than 100 m^2. The investments made to realize these plants are vital because of the current costs of the collectors; thus, before spending a lot of money, an in-depth energy audit to evaluate the requests of the users and assess the economic breakeven point should be undertaken. These systems, of course, are not able to meet the total energy demand, so they must be integrated into a conventional source.

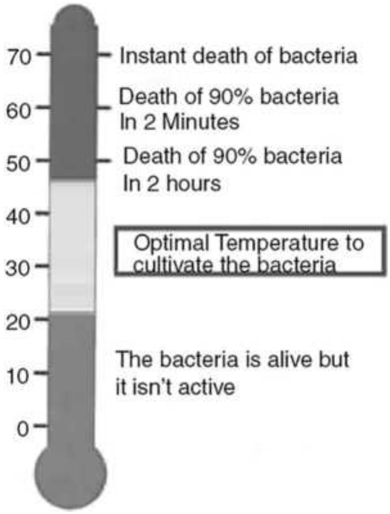


FIGURE 17 Behavior of the *Legionella* bacteria, depending on different temperatures.

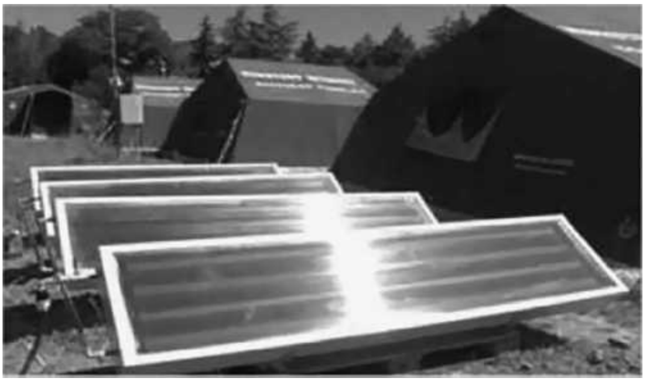


FIGURE 18 Abruzzo region post-earthquake emergency camps (2009).

In support of these investments often involved in campaign financing from the state, such as loans to grants and tax deductions on the cost of the material. In recent years, the problem of pollution and continuous increases in the cost of traditional energy sources have caused people to engage in new ways, i.e., try to change their old habits, often met by indiscriminate consumption of energy.

Solar Integrated Collector Storage System Innovations

The use of a solar integrated collector storage (ICS) system (see Figure 19) represents a well-established technology for heat storage in civil and industrial applications.

An innovative solar thermal device has been used as an integrated collector storage providing DHW (up to 50–80°C). Here, the collector also acts as a storage unit, without requiring an external vessel. This device was successfully used in several circumstances, especially in extreme situations such as in the post-earthquake tent cities or to feed remote users in developing countries.

The efficiency of this device is strongly related to the draw-off curve. In fact, it is strongly desirable that during the draw-off, the water temperature remains as high as possible. Presently, this aspect is not optimal, leading to a strong reduction of the water temperature (more than 50% when 50% of the hot water filling the collector was discharged).

The collector was designed for use in emergency situations or for feeding remote users. Then, very simple configuration was adopted.

A series of eight J-type thermocouples have been placed to measure the temperature during the thermal energy storage phase and during discharge. The accuracy of the thermocouples was ±1.5°C and the acquisition frequency was 0.88 Hz. Solar radiation fluctuations have been followed by a pyranometer system with an accuracy of ±10 μV/(W m^2). The eight thermocouples (seven have been used for monitoring the pipe and one for the ambient temperature) and the pyranometer have been connected through a thermocouple and RTD modules, with a National Instrument field point with Ethernet connection, and connected to a data processing system through LabVIEW, as shown in Figure 19.

During the charge phase, the pipe has been placed in vertical position in order to be completely filled with water and to eliminate the air. Then, it has been placed on its supports, with a tilt of 42°.

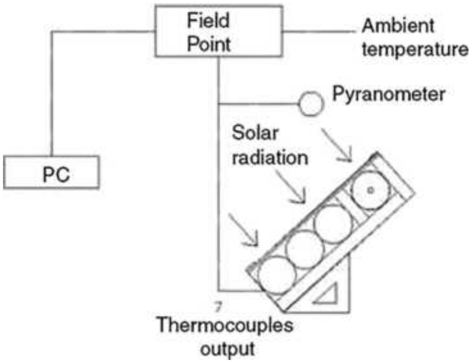


FIGURE 19 Experimental device used for the ICS prototype.

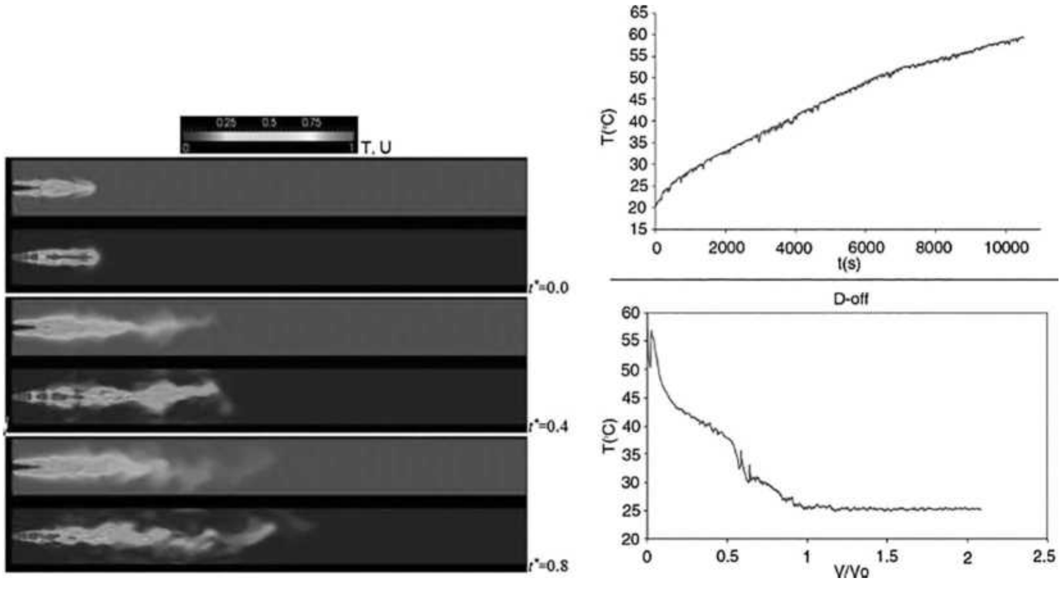


FIGURE 20 Longitudinal view of the temperature (top) and velocity magnitude (bottom) fields in three successive time steps.

The experimental analysis demonstrates that in the present configuration, the ISC has a reasonable performance in supplying hot water during the discharge phase.

Conclusions

The investments made to realize the plants are important because of the current costs of the collectors; therefore, before spending such quantity of money, it is better to undertake an in-depth energy audit to evaluate not only the requests of the users but also the breakeven point.

These plants are not able to supply the total energy demand; therefore, they should be integrated with a traditional source.

To support such investments, often there are financial aids from the state, such a free grant loan or allowances on the V.A.T. on the costs of the material.

In recent years, the pollution problem and the constant rise in the cost of traditional energy sources have resulted to users trying to find new solutions, changing old habits, often entailing excessive energy consumption.

References

1. Weiss, W.; Biermayr, P. Potential of solar thermal in Europe. European Solar Thermal Industry Federation (ESTIF), RESTMAC 6th Framework EU founded project, 2009, available at http://www.solarthermalworld.org/node/878.
2. Kalogirou, S.A. Solar thermal collectors and applications. Progress Energy Combustion Sci. **2004**, *30,* 231–295.
3. Hazami, M.; Kooli, S.; Lazaar, M.; Farhat, A.; Belghith, A. Performance of a solar storage collector. Desalination **2005**, *183,* 167–172.
4. Battisti, R.; Corrado, A. Environmental assessment of a solar thermal collectors with integrated water storage. J. Cleaner Prod. **2005**, *13,* 1295–1300.
5. Micangeli, A. *Design and implementation of innovative systems for the exploitation of renewable sources in social conflict areas: Desalination with heat recovery and hydroelectric energy production.* Sapienza University Doctoral Thesis, 2000.
6. Lesieur, M.; Metais, O.; Comte, P. *Large Eddy Simulations of Turbulence;* Cambridge University Press: New York, 2005; ISBN 0–521–78124-8.
7. Borello; Corsini; Delibra; Evangelisti; Micangeli. *Experimental and computational investigation of a new solar integrated collector storage system*, Third International Conference on Applied Energy, May 16–18, 2011, Perugia, Italy.

26

Alternative Energy: Wind Power Technology and Economy

K.E. Ohrn

Introduction (Web Preview)

This entry is a broad overview of wind power. It covers a range of topics, each of which could be expanded considerably. It is intended as an introductory reference for engineers, students, policy-makers, and the lay public.

Wind power is a small but growing source of electrical energy. Its economics are well known; there are several large and competent manufacturers, and technical problems are steadily being addressed. Wind power can now be considered as a financially and operationally viable alternative when planning additional electrical capacity. However, as shown below, wind power is only a minor component of present energy sources.

World final energy consumption (2002)[1]: 100%
World final electrical consumption (2002)[2]: 16.1%
World final wind power consumption (2004)[3]: 0.15%

Wind power's main deficiency as a power source is variability. Since wind velocity cannot be controlled or predicted with pinpoint accuracy, alternatives must be available to meet demand fluctuations.

Wind power carries few environmental penalties and makes use of a renewable resource. It has the potential to become a major but not dominant part of the future energy equation.

History

People have used wind to move boats, grind grains, and pump water for thousands of years. Wind-powered flour mills were common in Europe in the 12th century. In the 1700s, the Dutch added technical sophistication to their windmills with improved blades and a method to follow the prevailing wind. Isolated farms in the last century used windmills to generate electricity until the availability of the electrical power grid became widespread.

Past interest in wind power has tended to rise and fall with fuel prices for the predominant method of electrical production—thermal plants burning oil, natural gas, and coal.

Current

Electrical Production

Although small, wind power is a fast-growing part of the energy picture. Since 1990, worldwide installed capacity has grown about 27% (Table 1).[3]

Business is good for the leading manufacturers of wind power devices. Sales have increased; the technology is stable and predictable, with low maintenance and high availability.

Geographical Distribution (Countries)

The European Union had around 72% of installed capacity in 2004, and Germany, Spain, and the United States accounted for 66.1%. Denmark, Spain, and Germany had by far the largest 2004 capacity in terms of MW per million populations, and 10.6% of the world's population had 81.9% of its wind power capacity. In 2004, Denmark produced about 20% of its electrical power from wind power in 2004 (Table 2).[3]

TABLE 1 Capacity Growth

Year	Capacity (MW)	Growth Rate Year-over-Year (%)
1990	1,743	13.8
1991	1,983	17.0
1992	2,321	20.7
1993	2,801	26.1
1994	3,531	36.5
1995	4,821	26.6
1996	6,104	25.1
1997	7,636	33.0
1998	10,153	33.9
1999	13,594	27.7
2000	17,357	66.3
2001	28,857	7.9
2002	31,128	26.9
2003	39,500	20.3
2004	47,500	13.8
	Average Growth Rate	26.7

Source: Reprinted with permission from European Wind Energy Association.[3]

TABLE 2 Capacity Distribution

	Wind Power Installed Capacity (MW)								
Country	2001	2002	2003	2004	One-Year (%)	Three-Year (%)	Population (millions)	Capacity (MW/Million)	Percent of World Capacity
Denmark	2,456	2,880	3,076	3,083	0.2	7.9	5.4	570.9	6.4
Spain	3,550	5,043	6,420	8,263	28.7	32.5	40.3	205.0	17.2
Germany	8,734	11,968	14,612	16,649	13.9	24.0	82.4	202.1	34.7
Netherlands	523	727	938	1,081	15.2	27.4	16.4	65.9	2.3
USA	4,245	4,674	6,361	6,750	6.1	16.7	295.8	22.8	14.1
Italy	700	806	922	1,261	36.8	21.7	58.1	21.7	2.6
UK	525	570	759	889	17.1	19.2	60.4	14.7	1.9
Japan	357	486	761	991	30.2	40.5	127.4	7.8	2.1
India	1,456	1,702	2,125	3,000	41.2	27.2	1080.4	2.8	6.3
China	406	473	571	769	34.7	23.7	1306.4	0.6	1.6
Total	22,952	29,329	36,545	42,736	16.9	23.0	3073		89.2

Manufacturing capacity in 2002 was largely confined to this group of countries, with five big vendors accounting for 76% of sales. European Union vendors accounted for 85% of manufacturing market share.[3]

Economics

Cost per kWh

Wind power is a viable method of producing electricity that is capital intensive with low operating costs. The cost of production compares favorably with traditional fossil fuel or nuclear plant costs.

The major cost elements of a modern wind power installation are as follows[3,5,9]:

Capital

Onshore: 1200–1500 USD/kW
Offshore: 1700–2200 USD/kW

Operating: usually about 1.5%–2.0% of capital cost per year.[12]

Capital costs include wind capacity survey and analysis, land surveying, permits, roads, foundations, towers and turbines, sensors and communications systems, cabling to transformers and substations, maintenance facilities, testing, and commissioning. By far the largest individual capital cost is the turbine (up to 75%).

Operating and maintenance costs include management fees, insurance, property taxes, rent, and both scheduled and unscheduled maintenance.

Financing costs are a major portion of energy production costs, making them very sensitive to interest rates, incentives and subsidization.

Energy production cost estimates vary considerably. Optimists in the industry, such as the British Wind Energy Association, quote a low of 4.8 USD cents per kWh for an onshore plant in an optimal location. Pessimists, like the Royal Academy of Engineering in the U.K., quote up to 13.2 USD cents per kWh for an offshore plant, partly by including a controversial 3.1 USD cents per kWh cost for “standby capacity” required to supply demand when wind power is not available (Table 3).

Capacity Factor

The power generated by a wind turbine depends on the speed of the wind, and on how often it is available. At any given site, this is measured by the capacity factor, or the ratio of actual generated energy

TABLE 3 Windower Costs

	Cents (US) per kWh			
Wind Power Costs	Wind Onshore	Wind Offshore	Coal CFB	Gas CCGT
RAE [15]	6.8	9.9	4.8	4.0
RAE 2[a]	10.1	13.2	9.2	6.0
EWEA [3]	5.5	8.5		
AWEA (5)	6.6			
AWEA 2[b]	5.4			
BWEA (minimum)[4]	4.8	6.9	4.8	4.8
BWEA (maximum)[c]	6.8	9.1	6.8	5.9
Euro to USD conversion (2004)				1.22
GBP to USD conversion (2004)				1.83

[a] Adds 3.1 cents per kWh for wind power backup capacity and 1.9–4.6 cents/kWh for coal and gas carbon capture.

[b] AWEA figures adjusted to delete 1.8 cents US/kWh production tax credit and are for onshore sites with different average wind speeds.

[c] BWEA figures for a range of site types in November 2004 and 1–2 cents/kWh for carbon capture.

Source: RAE, Royal Academy of Engineering[15]; BWEA, British Wind Energy Association[4]; EWEA, European Wind Energy Association[3]; AWEA, American Wind Energy Association.[5]

to the theoretical maximum. Wind power turbine electrical output rises as the cube of the wind speed. When wind speed doubles, energy output increases eightfold. A typical turbine begins to turn when wind speed is at 9 MPH and will cut out at 56 MPH for safety reasons.

Capacity factors vary by site but are typically in the range of 20%–30% with occasional very good offshore sites reaching 40%. The yearly energy output from a wind farm is given by the following formula:

Site

Power production costs, site size, site design, and energy output and variability will depend mainly on details about the wind. These details include wind speed, wind direction, and the geographical distribution of favorable wind profiles. During analysis of potential sites, most planners use high (60 m plus) anemometer towers—often several of them—to gather at least one year's data per site. These data are usually correlated with national meteorological observations, if these are available and suitable. If not available, it would be prudent to gather site data for a longer period of up to three years.

Investors and regulators are increasingly aware of the crucial nature of wind data in estimating the quantity and timing of potential power production at a specific site. This research is crucial to the financial analysis of a potential wind power venture.

Other site analysis factors are accessibility via road for heavy equipment, electrical grid proximity and capacity, land ownership, and environmental impact.

Location

Favored Geography

Many countries have developed wind charts of broad areas based on meteorological data gathered for weather and aviation purposes. These charts show potential areas for investigation, where wind strength is high and constant over long periods of the year. Once potential sites, and their extent, have been identified, on-site data measurements provide the basis for analysis and modeling of potential energy production for a specific site.

After wind modeling, the site's geographical, environmental, financing, and ownership issues can be explored in detail.

Generally, sites are either onshore or offshore. Onshore sites are cheaper to construct, but have lower capacity factors due to wind turbulence from nearby hills, trees, and buildings. Offshore sites can have more potential energy available due to higher wind speeds and lower turbulence, which also reduces turbine component wear. Good offshore sites can be near high-demand load areas such as coastal cities, which also increases transmission options. Aesthetic and noise concerns are often fewer offshore, and sea-bed environmental concerns can be lower than landuse concerns for an onshore site.

Sizing a Location

Wind farm towers are usually spread over a large area in order to minimize wake losses. A spacing of five rotor diameters is often recommended. In a typical wind farm, the land physically occupied by tower foundations, buildings, and roads is often less than 2% of overall land area.[6] The remaining land is quite suitable for agriculture and other uses.

Limits to Maximum Production

How much capacity exists to generate electricity from wind? Is it possible that we will require more energy than wind can provide? After surveying wind patterns in the United States and applying energy density and extraction calculations, Elliott and Schwartz[14] concluded in 1993 that 6% of the available U.S. land mass could provide 150% of then-current U.S. electrical consumption. Furthermore, the needed land would be sparsely affected by the wind farm installation, with the vast majority of it (95%–98%) unoccupied by tower foundations, roads, or ancillary equipment and suitable for farming, ranching, and other uses. This study excluded land that is environmentally or otherwise unsuitable, such as cities, forests, parks, wildlife refuges, and environmental exclusion areas.

In the European Union, potential wind power capacity is also larger than current electrical consumption.

Future

Projected Growth

Thanks to increasing concern over the environmental effects of greenhouse gas emissions, the rising cost of fossil fuels, and the impending decrease in availability of oil and natural gas, wind power has a bright future.

Current 25%–30% growth rates are likely not sustainable, due to equipment production volume constraints and limits to perceived need for further capacity. Given the Eurocentric, highly clustered nature of current installed capacity, there is significant potential for high-rate growth elsewhere. However, even in European countries like Germany and Denmark, steady growth will be driven by predicted cost reductions in the 10%–20% range and by regulatory and governmental initiatives aimed at reducing emissions from electrical energy production and transitioning to renewable resources.

Projected Cost

Wind power technology is well down the cost improvement curve, with costs having fallen to present levels, below ten cents USD per kWh, from over $1.00 U.S. in 1978. Costs for a medium-sized turbine have dropped 50% since the mid-1980s, reflecting increasing maturity in the market. Cost projections range from a further 9%–17% drop as installed capacity doubles in the near-term future.

Projected Production

With increasing governmental policy support and commitment, growth rates of 15%–20% appear achievable in five to ten years. But there is likely an upper limit to the amount of electrical energy that can be produced from the wind.

Reaching Maximum Production

Production limits for wind power are based on its variable nature. Other types of electrical production capacity will be needed to provide base-load electrical capacity in the event that there is little wind available. Wind power will then become one player in a mix of generating technologies.

Hydrogen Economy

As wind power becomes a larger portion of electrical supply, occasionally its supply will exceed demand. Rather than simply curtailing wind plant production, it is attractive to think of using this excess electrical power capacity to generate hydrogen via electrolysis. This has the effect of storing wind energy that would otherwise not be harvested. This energy, in the form of hydrogen, can be used directly as a nonpolluting fuel or as an input source to fuel cells to produce electricity at a later time.

When there is a significant hydrogen economy, with transmission lines, storage, and fuel cell capacity, this use for wind power will become a very attractive scenario.

Other Issues for the Future

Learning more about wind and forecasting—predicting the best locations, wind farm output, gusts, and directional shear.[10] This will help reduce financing costs when wind power plant output and impact on the grid are better understood and more predictable.

- Improving the control of demand through incentives around end-user load shedding, rescheduling and simple conservation methods. This could be used to offset wind power production shortfalls as an alternative to other forms of generation.
- Advancing aerodynamics specific to wind turbine blades and control systems.
- Designing extremely large wind tunnels to study wake effects minimization, structural load prediction, and energy output maximization at lower wind speeds.
- Enhancing power system capacity planning models to include wind farm components.
- Re-planting, or upgrading older mechanical and electrical components at existing wind farms.
- Wind farm siting further offshore and on floating platforms.
- Combining wind power and hydroelectric capacity by using surplus wind power to pump water behind dams and so store power that might otherwise be wasted.
- Determining how and whether to allocate full costs of environmental impact to fossil and nuclear plants.

Strengths and Weaknesses

Strengths

Environment

Wind power installations do not emit air pollution in the form of carbon dioxide, sulfur dioxide, nitrogen oxides, or other particulate matter such as heavy metal air toxins. Wind power installations do not use water or discharge any hazardous waste or heat into water. Conventional coal, oil, and gas electric power plants produce significant emissions of all kinds. Nuclear power plants produce dangerous and long-lasting radioactive waste. Greater use of wind power means less impact on health and the environment, particularly regarding climate change due to greenhouse gas emissions.

Renewable

Wind power produces energy from a resource that is constantly renewed. The energy in wind is derived from the sun, which heats different parts of the earth at different rates during the day and over the seasons. Unlike fossil and nuclear plants, the source of energy is essentially inexhaustible.

Costs

Wind power's costs are well known and are dropping to the point at which this technology is very competitive with other means of production. Fuel costs are nil, meaning that fuel costs have no uncertainty. Wind power costs should be more stable and predictable over the lifetime of the plant than power costs for fossil fuel plants.

Local and Diverse

Wind power plants provide energy source diversity and reduce the need to find, develop, and secure sources of fossil or nuclear fuel. This reduces foreign dependencies in energy supply, and reduces the chances of a political problem or natural disaster interfering with and diminishing the supply of electricity.

Quick to Build, Easy to Expand

Wind power plants of significant capacity can be constructed and installed within a year, a much shorter time than conventional plants. The planning time horizon is similar to conventional plants, given the need to accurately survey site wind characteristics and deal with normal environmental and related site issues. This means that capacity can be increased in closer step with demand than with conventional plants. With the right site and design, a wind power plant can be incrementally expanded very quickly.

Weaknesses

Natural Variability

A single wind farm produces variable amounts of energy, and its output is not yet as predictable as a traditional plant. As the geographical distribution and number of wind plants increases, and as research into predicting wind continues, these problems should be minimized, allowing cost-effective and orderly scheduling and dispatch of total grid capacity sources—but it is difficult to see traditional power sources disappearing entirely.

Connection to Grid

As the amount of electrical power supplied by wind power plants increases, concern increases over its effects on the electrical grid.

In order to maintain a reliable supply of electricity that matches demand, utility operators maintain emergency reserve capacity in order to deal with plant outages (failures) and unexpected demand across their entire system. This reserve is in the form of purchased power, unused capacity at conventional plants running below their maximum, or quick-start plants such as gas-fired turbines. Often, conventional plants on the grid are allocated a cost to cover this reserve based on their capacity (large plant, large reserve) and reliability (more outages, more reserve).

The industry is working on ways to determine and allocate this reserve cost for wind power plants. Yet to be agreed upon is the statistical basis for calculating such wind power plant reserve costs. Improvements in day-ahead wind forecasting will greatly reduce the uncertainty around wind plant output, and so decrease the cost burden to provide this reserve.

Several current estimates prepared for U.S. utilities show this reserve cost burden (or ancillary services cost) to increase with the amount of capacity provided by wind power, and to be in the range of 0.1–0.5 cents USD per kWh for penetrations between 3.5 and 29%. In no case was it thought necessary to allocate a reserve equal to 100% of the wind power capacity.[11] German experience is similar,[8] with no additional reserve capacity required for the 14% wind energy share of the national electrical consumption forecast for 2015.

When wind power supplies less than 20% of electrical consumption, these problems are not severe. At larger penetrations, reserves become a major issue. Interestingly, wind power plants may be subject to shutdown or voluntary power reductions in the event of coincident high wind, low demand situations. This is occasionally the case today in Denmark and Spain.

In some cases, wind power sites are situated far from the location of high electrical power demand, placing strain or potential overload on existing transmission facilities. In these cases, there are often cost, ownership and responsibility issues yet to be resolved.

Power quality problems around power factor, harmonic distortions, and frequency and voltage fluctuations are being successfully addressed in modern large production wind farms.[8]

This is one of the most difficult sets of issues facing the future of wind power as it matures from small-scale and local to large-scale penetration.

Local Resource Shortage

In a few places, high-quality wind power sites are not available or are already in production, leaving these places to import electrical power or use traditional sources.

Noise

Noise levels have decreased and are now confined to blade noise in modern units. Generator and related mechanical noise has been effectively eliminated. Noise, however, will always be a significant factor. Blade noise is described as a “whoosh, whoosh” sound, and is in the 45–50 decibel range at a distance of 200–300 m. This noise level is consistent with many national noise level regulations. However, this noise buffer zone adds to the overall land requirement for a wind power plant and so increases costs.

Visual Impact

Onshore wind farms are highly visible due to the height of towers and the size of the blades and generator. The impact of this varies with each person. Each wind plant operator needs to determine the levels of support and opposition from those who live and work within sight of the plant. Offshore plants attract fewer detractors than onshore plants—one of the reasons for their increasing popularity.

Offshore wind plants are less likely to cause unwanted noise since they are far from human habitation. This reduces turbine and blade design constraints and can lead to higher capacity factors.

Bird Impact

Bird deaths are a regrettable reality. The bird death rate at a specific wind farm project is quite variable. Several early wind farms (Altamont Pass, California, and La Tarifa, Spain) caused concern over death rates. The California Energy Commission estimates the death rate at Altamont (5400 turbines) to be 0.33–0.87 bird deaths per turbine per year.[16] The overall recent U.S. national average[13] is 2.3 bird deaths per turbine per year. Prudently located sites are off migration routes and not in nesting, overwintering, or feeding areas. Their tower designs do not offer nesting or even roosting places. In such locations, death rates are lower, and overall impact is much lower than that caused by other types of human activity.

Since climate change is a very serious environmental problem faced by bird populations, wind power and other renewables are an important part of the solution.

Technology

Overview

Wind turbine design has three major components, and there are large economies of scale in design.

- *Tower height:* Wind turbine energy output is proportional to the cube of wind speed. Since moving air (wind) is subject to drag and turbulence from its contact with the earth and the objects on the earth, wind speed increases with height (vertical shear). The higher the tower, the more advantage there is for power generation. The tradeoff is between tower costs and increase in power generation. Typically, tower heights are rising, and are currently in the 100 m range. Off shore, vertical wind shear is generally less than onshore, so towers can be shorter, with wave height clearance being the factor that determines tower height (Figure 1).
- *Blade diameter:* The power capacity (watts) of a wind turbine varies with the square of its blade diameter, because a blade with a larger diameter has a larger area available for harvesting the wind energy passing through it. The coefficient of performance defines the actual power capacity compared to the maximum—how much energy can be extracted from the wind compared to the available energy. Modern wind turbines can achieve a coefficient of performance approaching 0.5, very close to the theoretical maximum of 0.59 derived by Betz.[3,18] This maximum is derived from the concept that if 100% of wind energy were extracted, the wind exiting the turbine would be at zero speed, so no new air could enter the turbine. Larger capacity turbines benefit significantly from economies of scale in foundation and support costs as well as swept area (Figure 2).
- *Controls and generating equipment:* The turbine's hub (or nacelle) is the most costly component and contains the generator, gear boxes, yaw controls, brakes, cooling mechanism, computer controls, anemometer, and wind directional vane.

Generators

As the blades turn, they drive a generator to produce electricity. Generating capacity ranges from a few hundred kW to over 3 MW. In older designs, there is usually a 40:1 gearbox to match low, fixed rotor speeds (~30 RPM) to required generator speed (1200 RPM for a 60 Hz output, 6-pole generator).

FIGURE 1 Typical large wind turbine. Note entrance steps and utility pole at base for scale.
Source: Photo courtesy of Suncor Energy, Inc.

FIGURE 2 Site assembly of large wind turbine nacelles and blades.
Source: Photo courtesy of Suncor Energy, Inc.

The gearbox often incorporates brakes as a part of the overall wind turbine control system. Generators that operate at low RPM are available and are called direct drive generators. These would eliminate the gearbox.

In more modern designs, rotor speed is variable and controlled to optimize power extraction from the available wind. Generator output is converted to d.c. and then back to a.c. at the required grid frequency and voltage. The conversion equipment is sometimes located at a central part of the wind power plant.

This is an active area for ongoing technical innovation.

Blades

In order to maximize power capacity through size, blades are very long, up to 50 m. To minimize noise, they must turn slowly so as to reduce tip speed, the primary blade noise source. Typical rotation speeds are in the 10–30 RPM range. Blades are increasingly made from composites (carbon fibre reinforced epoxy resins).

Rotor blade aerodynamics[19] have much in common with the aerodynamics of a propeller or a helicopter blade, but they are sufficiently different that the aerodynamics of wind turbine blades is an evolving field. The difference is that wind turbine airflows are unsteady due to gusting, turbulence, vertical shear, turbine tower upstream shadow, yaw correction lag, and the effects of rotation on flow development. For example, at present it can be difficult to predict rotor torque (and therefore power output) accurately for normal turbine operating conditions. Further development of theory and modeling tools should allow the industry to improve rotor strength, weight, power predictability, power output, and plant longevity while controlling cost and structural life.

For a given site wind velocity, the rotor blade’s tip has very different air flow than its root, requiring the blade to be designed in a careful twist. The outer third of the blade generally produces two-thirds of the rotor’s power. The third nearest the hub provides mechanical strength to support the tip, and also provides starting torque in startup situations.

Each blade generally has lightning protection in the form of a metallic piece on the tip and a conductor running to the hub.

Some manufacturers place Whitcomb winglets at the blade tips to reduce induced drag and rotor noise, in common with aircraft wing design.

In order to control blades during high wind speeds, some are designed with a fixed pitch that will progressively stall in high wind speeds. Others incorporate active pitch control mechanisms at the hub. Such control systems use hydraulic actuators or electric stepper motors and must act very quickly to be effective.

Wind Sensors

Wind turbines incorporate an anemometer to measure wind speed and one or more vanes to measure direction. These are primary inputs to the control mechanism and data gathering systems usually incorporated into a wind turbine.

Control Mechanisms (Computer Systems)

Control systems are used to yaw the wind turbine to face into the wind, and in some designs to control blade pitch angle or activate brakes when wind gets too strong. In sophisticated cases, the controllers are redundant closed-loop systems that operate pumps, valves and motors to achieve optimum wind turbine performance. They also monitor and collect data about wind strength and direction; electrical voltage, frequency and current; nacelle and bearing temperatures; hydraulic pressure levels; and rotor speeds, vibration, yaw, fluid levels, and blade pitch angle. Some designs provide warnings and alarms to central site operators via landline or radio. Manufacturers do not release much detail about these systems, since they are a critical contributor to a wind turbine's overall effectiveness, safety, and mechanical longevity.

Role of Governments and Regulators

Governments play a large part in determining the role and scale of wind energy in our future mix of energy production capabilities.

Subsidies, Tax Incentives

As part of programs to encourage wind power production, the following are used in varying ways[7]:

- Outright subsidies, grants and no-interest loans.
- Tax incentives such as accelerated depreciation.
- Fixed prices paid for produced electrical power.
- Renewable energy quantity targets imposed on power utility operators.

Grid Interconnection and Regulatory Issues

Since many power utility operators are owned by governments, and most are regulated heavily, governments have a role to play in encouraging solutions to grid interconnection issues. There must be a political will to address issues, find solutions, and develop practices and different management strategies that will allow greater penetration of wind power into the electrical supply.

Improving Wind Information

Climate and environmental information is most often collected and supplied by national governments in support of weather and aviation services. Wind atlases are an invaluable resource to the wind plant planning process. National efforts to improve long-and short-term wind forecasting, atmospheric modeling tools, and techniques will benefit wind power projects' ability to forecast power output for long-term and short-term planning purposes.[10]

Environmental Regulation

In this controversial area, government can tighten its regulation of air quality, carbon emissions and other environmental areas. This would have the effect of increasing the apparent cost of conventional thermal electrical power, which is responsible for significant emissions. It is often argued that wind power would already be cost competitive if environmental and health costs were to be fully allocated to conventional oil, gas, and particularly coal-powered plants, or if such plants were required to make investments to significantly reduce emissions.

Conclusion

Western societies depend on a steady supply of energy, much of it in the form of electricity. Most of that supply comes from thermal plants that burn oil, natural gas, and coal, or from nuclear plants. Where will our electrical energy come from in the future? How will we keep our environment livable and healthy?

One part of this answer lies in wind power. Its costs are within reason; the technology has matured with some gains yet to be realized; it carries little environmental penalty; and the source of its energy is renewable. As long as the sun heats the earth, wind power will be available.

Wind power will not likely be the complete answer; it is an intermittent source because the wind doesn't always blow. But there is a very large amount of it available for us to harvest. As wind power moves quickly from small-scale to large-scale, its future path depends on governments and regulators as much as it does on technical innovators and manufacturers.

References

1. Annual Energy Review 2003, DOE/EIA-0384(2003), Energy Information Administration, U.S. Department of Energy, Washington, DC, September 2004.
2. Key World Energy Statistics -2004 Edition, International Energy Agency, OECD/IEA -2, rue André-Pascal, 75775 Paris Cedex 16, France or 9, rue de la Fédération, 75739 Paris Cedex 15, France, 2004.
3. Wind Energy, The Facts, Analysis of Wind Energy In the EU-25, European Wind Energy Association, Brussels, Belgium, 2004.
4. BWEA Briefing Sheet, Wind and the U.K.'s 10% Target, The British Wind Energy Association, http://www.bwea.com/energy/10percent.pdf (accessed July 2005).
5. The Economics of Wind Energy, American Wind Energy Association, http://www.awea.org/pubs/factsheets/Economics OfWind-Feb2005.pdf (accessed July 2005).
6. Saddler, H., Dr.; Diesendorf, M., Dr.; Denniss, R. A Clean Energy Future For Australia, A Study By Energy Strategies for the Clean Energy Future Group, Australia, March 2004.
7. Policies to Promote Non-hydro Renewable Energy in the United States and Selected Countries, Energy Information Administration, Office of Coal, Nuclear, Electric and Alternate Fuels, United States Department of Energy, Washington, DC, February 2005.
8. BRIEFING May 10th, 2005. German Energy Agency DENA study demonstrates that large scale integration of wind energy in the electricity system is technically and economically feasible. http://www.ewea.org/documents/0510_EWEA_BWE_VDMA_dena_briefing.pdf (accessed July 2005).
9. Offshore Wind Experiences, International Energy Agency, 9, rue de la Fédération, Paris Cedex 15, 2004.
10. Milborrow, D. Forecasting for Scheduled Delivery, Windpower Monthly, December 2003.
11. Utility Wind Interest Group—Wind Power Impacts On Electric-Power-System Operating Costs, November 2003.
12. Danish Wind Industry Association, http://www.windpower. org/en/tour.htm (accessed July 2005).
13. Wind Turbine Interactions With Birds and Bats: A Summary of Research Results and Remaining Questions, Fact Sheet, Second Edition, National Wind Coordinating Committee, Washington, DC, November 2004. http://www.nationalwind.org/publications/avian/wildlife_factsheet.pdf (accessed July 2005).
14. Elliott, D.L.; Schwartz, M.N., Wind Energy Potential In the United States, Pacific Northwest Laboratory, PNL-SA-23109, Richland, Washington, USA, September 1993.

15. The Cost of Generating Electricity, Royal Academy of Engineering, http://www.raeng.org.uk/news/publications/list/reports/Cost_of_Generating_Electricity.pdf (accessed July 2005).
16. Developing Methods to Reduce Bird Mortality in the Al- tamont Pass Wind Resource Area, August 2004, http://www.energy.ca.gov/reports/500-04-052/500-04-052_00_ EXEC_SUM.PDF (accessed July 2005).
17. Wind Force 12, Global Wind Energy Council, June 2005, http://www.ewea.org/03publications/WindForce12.htm (accessed July 2005).
18. Proof of Betz Theorem, Danish Wind Industry Association, http://www.windpower.org/en/stat/betzpro.htm (accessed July 2005).
19. National Wind Technology Center, http://www.nrel.gov/wind/about_aerodynamics.html (accessed July 2005).

27

Electric Power: Microgrids

Ryan Hanna

Introduction

Microgrids are insular power systems embedded within the bulk power grid that use distributed energy resources (DERs), such as fossil fuel generators, renewable generators, and energy storage, to provide energy locally near the point of consumption (Figure 1). They may be small or large, for example, a single building or a campus with dozens of buildings, and they may be simple or complex, for example, consisting of a single backup generator or numerous DERs working in concert to provide primary power. In any case, their defining attribute is their ability to disconnect, or "island", from the larger grid during power outages, thereby shielding customers within from disruptions and improving the reliability of electric service. (Some taxonomies include remote or "off-grid" systems—those not embedded within a bulk grid.) Although microgrids vary widely in size, from tens of kilowatts to tens of megawatts, those powering a small commercial building (approximately hundreds of kilowatts) are indeed "micro" or on the order of 10^{-6} the capacity of the bulk U.S. power grid.

Since their inception, power systems have been built to provide electricity that is reliable and economic—two competing design criteria that require balancing (Billinton and Allan 1996). Today, a third criterion—that power systems provide carbon-free electricity in accordance with long-term climate stabilization goals—is now widely accepted. The climate criterion presents new and interesting design trade-offs for microgrids. Historically, the fossil fuel pathway—i.e., the use of fossil fuel generators, first diesel but more recently natural gas—has provided high levels of reliability at low cost. Given future projections for very low gas prices, it is likely to remain the status quo absent new carbon-minded policy. Some policies are emerging—typically in the form of restrictions on diesel use—while governments have variously supported low-carbon pathways such as combined heat and power (CHP), renewables with energy storage, biofuels, and hydrogen. Although some options remain far-off, consideration is important today because each pathway offers a fundamentally different vision for how microgrids interact with a decarbonized energy system. Pathways can differ in use of fuels, generators, load electrification, outside requirements for energy infrastructure, and, importantly, outcomes for carbon emissions and pollution.

Recent support for microgrids has stemmed from concern about potential impacts to power systems from an increasing frequency of severe weather, including wildfires, to which climate change is contributing. The conventional wisdom is that microgrids can increase grid reliability and resiliency while also

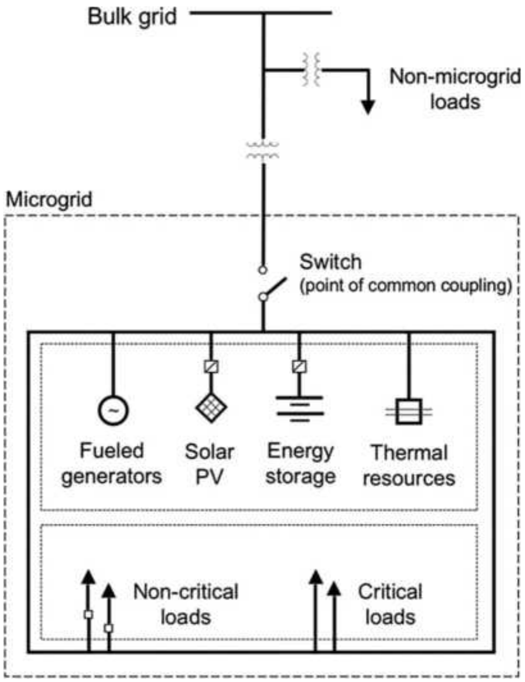


FIGURE 1 A simple representation of a grid-connected microgrid. Distributed energy resources (DERs) and loads comprise the microgrid and lie behind a single point of common coupling to the bulk grid. The microgrid can disconnect from and reconnect to the bulk grid and operate autonomously. (Fossil fuel and thermal energy networks are present but, for clarity, not pictured.)

helping to address climate change. An appraisal of decarbonization pathways, however, shows that this is not always so (Section "Decarbonization Pathways for Microgrids"), and comparison of cost, emissions, and performance helps explain where co-benefits and trade-offs exist. A picture of historical deployments and present-day activities around microgrids provides further context (Section "Microgrids, Past and Present"). Although the focus here is the U.S.A, deployments are on the rise globally, particularly in Asia Pacific (e.g., Australia, India, China, and Japan) where forecasted growth is in fact greatest and driven by deployment of remote systems in developing areas that lack grid infrastructure (Willette and Asmus 2019).

Microgrids, Past and Present

Microgrids are not a new phenomenon. Today, we recognize the earliest power systems—pioneered by Thomas Edison during the 1880s that provided lighting for hundreds of customers across a few New York City blocks—as microgrids. These grids, and hundreds in other cities, would first experience decades of expansion, as industry pursued larger customer bases and scale economies and governments mandated rural electrification in the 1930s–1940s to combat poverty (Tuttle et al. 2016), and later interconnection to increase redundancy and power plant utilization rates.

Centralized grids now underpin nearly all of the industrialized world, limiting the role for microgrids to niche applications. The high cost of microgrids has been justified where interconnection to the bulk grid is impractical or uneconomic, such as with island villages, military bases, and remote industrial facilities. In Alaska, diesel and hydropower microgrids power over 200 remote communities. It has also been justified where services within the bulk grid are deemed critical. Microgrids provide backup power for the health care industry, where hospitals must comply with emergency generator standards and maintain onsite a 96-hour supply of diesel fuel, as well as for the telecommunications industry's enormous network of cell towers, where the use of small (<10 kW) diesel generators and lead-acid batteries has dominated. More recently, incentives have supported the use of fuel cells—over 3,000 have been deployed since 2007 (Ma et al. 2017). Microgrids have also been economic in campus settings that have space and the institutional capacity for ownership and operation. Prominent examples include those at New York University, Princeton University, the Illinois Institute of Technology, the University of California San Diego, and the University of California Irvine, among others. Natural gas cogeneration forms the physical and economic backbone of many, providing significant cost savings on electricity, heating, and cooling.

Today, the role for microgrids is expanding as governments begin to take a proactive approach in guiding their development. State governments in the northeast U.S.A are seeking to create a role for "community" microgrids that embed pockets of resiliency into the grid and serve the public interest. Following severe weather in 2012 that caused widespread blackouts, Connecticut in 2013, Massachusetts and New Jersey in 2014, New York in 2015, and Maryland in 2019 each enacted a community microgrid grant program to increase resiliency at critical infrastructure such as hospitals, police and fire stations, wastewater treatment plants, public shelters, and municipal commercial areas. Recent power outages at the Atlanta, Las Vegas, Washington D.C., Los Angeles, Hamburg, and Amsterdam international airports have spurred funding for microgrids that support critical airport services.

Other jurisdictions are pursuing a market-based approach with the development of commercialization pathways that tie in to broader efforts to decentralize electricity generation. California is leading, having recently awarded $25.7 million for seven microgrids in 2014 that demonstrate "low-carbon" and "renewables-based" configurations and $45 million for nine microgrids in 2018 that demonstrate "repeatable, commercial-scale" configurations that support the state's emissions reduction goals. Hawaii passed legislation in 2018 to establish a tariff that defines the value of services exchanged between a microgrid and the bulk grid. In the Caribbean, Puerto Rico's post-Hurricane Maria reconstruction plans aim to decentralize generation on the island, and part of that vision includes a new regulatory framework for microgrids that encourages deployment by third-party (non-utility) providers. The island's goals are to increase reliability, and reduce fossil fuel consumption and carbon dioxide emissions, among others. The U.S. Virgin Islands is also rebuilding through microgrids, deploying two on the island of St. John comprised of diesel generators, solar photovoltaics (PV), and battery energy storage.

To what extent are these microgrids low-carbon? In all historical use cases, fossil fuel generators have been the preferred technology. This holds true today in the emerging community microgrid paradigm, despite falling prices for renewable energy and energy storage. For the programs in Connecticut (Microgrid Pilot Program Round 1 and 2) and New York (NY Prize Stage 2), for example, for which good data on installed or planned DERs exists, natural gas generators comprise 78% of total capacity, of which 61% is CHP, while PV comprises 4%, which is on par with diesel generators (also 4%). Battery storage, on a power capacity basis, comprises 3%. Five of the six microgrids supported by the New Jersey Energy Resilience Bank are based entirely on natural gas CHP. By contrast, California has set out to support renewable microgrids and its 2014 program produced 73% solar PV, 11% battery storage, and 15% fossil fuel generation by capacity. Programs in other jurisdictions remain too nascent to draw conclusions.

Each of these jurisdictions pushing microgrids has enacted climate policies that aim to decarbonize the power sector through mandated cuts in greenhouse gas emissions, procurement targets for renewable energy, or both. Puerto Rico and Hawaii have passed into law 100% renewable energy portfolio standards (RPS), for instance, while California has passed a 100% clean energy standard. The five northeast states with microgrid grant programs have RPS targets ranging from 25% to 50% by 2030. Accordingly, some have acted to limit fossil fuel use in microgrids. Puerto Rico's 2018 rules, for example, mandate renewable energy and restrict fossil fuel use in non-CHP generators in three ways: to 25% of the total microgrid energy output on an annual basis, by generator heat rate (not greater than 13,000 btu/kWh), and by the total microgrid energy input–output ratio (not greater than 2,500 btu/kWh). CHP generators qualify but must adhere to efficiency standards (useful thermal energy must be greater than 50%

of the total energy output) and generator heat rate (not greater than 7,000 btu/kWh). California's 2018 microgrid legislation (SB 1339) prohibits compensation for all diesel use and natural gas generators that fail to meet strict emission standards, effectively precluding non-CHP fossil fuel generators. By contrast, Hawaii, which is in the process of establishing a microgrid services tariff, has not set any prohibitions on fossil fuel use.

Decarbonization Pathways for Microgrids

Although the merits of particular choices for technology and fuel in microgrids are case-specific (Hanna et al. 2017; Flores and Brouwer 2018), an assessment of pathways generally should address the three power system design criteria—reliability, cost, and emissions. Table 1 summarizes that assessment qualitatively for the present day, with characteristics scored from very poor (––) to very good (++). It also includes the practicability of deployment, or dependency on outside energy infrastructure, as well as scalability, or the capability of the pathway to scale up to meet a significant portion of society's electricity needs. Power density, a watt per square meter (W/m^2) measure of the concentration of energy generation or consumption, is a useful metric with which to assess the capabilities of distributed generators that, unlike rural power plants, are confined by limited space in built environments. Here, it is a proxy for reliability.

The diesel pathway, or use of diesel reciprocating engines, is the historical pathway and taken in this assessment as the base case against which other pathways are scored. Diesel generators are technologically mature, highly versatile, and have high power densities and low cost. Overuse can significantly degrade local air quality, however, and many jurisdictions limit their application to emergency operation only, typically a few hundred hours per year. Thus, they are neither scalable nor desirable options for widespread decentralization; other solutions are needed. Figure 2 shows the physical features for technology choice and fuel that distinguish pathways.

Combined Heat and Power

Over 2,000 CHP systems across the U.S.A provide nearly 2.3 GW of capacity (Darrow et al. 2017). These systems are fossil fuel generators like reciprocating engines and turbogenerators that capture and use engine waste heat to supply building heating and cooling needs. In making use of heat that is otherwise dissipated to the environment, CHP provides two benefits: it increases the efficiency of fuel use, from 25–45% in standalone (i.e., non-CHP) generators to 65–80% in generators with heat recovery, and it reduces (and can eliminate) direct gas combustion for heating as well as electricity for cooling, thereby reducing greenhouse gas emissions in the commercial and residential sectors. CHP is successful in reducing emissions today because the combustion of gas to meet electrical and thermal energy demand in concert results in a lower carbon intensity (gCO$_2$/MJ) than the separate production of electricity from the bulk grid and heat from a gas boiler or furnace.

TABLE 1 Microgrid Pathways—Characteristic Technology Choice, Fuel Use, and Assessment[a]

Pathway	Technology	Fuel	Power Density	CO$_2$ Emissions	Pollutant Emissions	Cost	Practicability	Scalability
Diesel	ICE	Diesel	++	––	––	++	++	––
CHP	ICE-CHP	NG	++	–	–	+	++	+
All-electric (solar-plus-storage)	PV, BS, HP	None	–	++	++	–	++	+
Biomethane	ICE-CHP	Biomethane	++	+	–	–	+	––
Hydrogen	FC-CHP	Hydrogen	++	++	++	––	––	++

[a] *Abbreviations:* ICE (internal combustion engine), CHP (combined heat and power), PV (photovoltaics), BS (battery storage), HP (heat pumps), FC (fuel cell), and NG (natural gas).

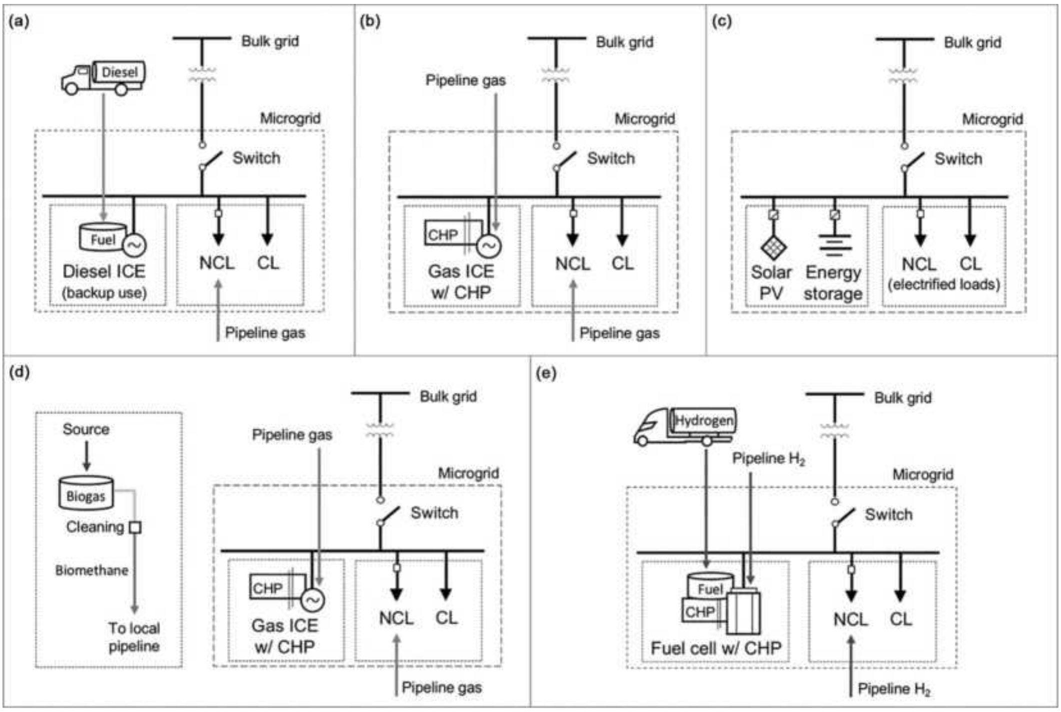


FIGURE 2 Physical features for technology choice and fuel that distinguish pathways from one another. The five pathways are (a) a diesel genset, (b) gas-fired genset, (c) solar-plus-storage (all-electric), (d) biomethane-fired genset, and (e) hydrogen fuel cell. All fueled gensets (less the diesel genset; b–d) make use of CHP. NCL denotes non-critical load, CL critical load, ICE internal combustion engine, CHP combined heat and power (i.e., infrastructure for capturing and using waste heat), and H_2 hydrogen.

Governments have supported CHP for decades as non-utility generators. For example, the Obama Administration in 2012 set a nationwide target of 40 GW of new capacity by 2020, while in 2010 the state of California set a target of 6.5 GW by 2030. Governments are now coupling support of CHP to microgrids. California and Puerto Rico, through new microgrid rules, are discouraging (or banning outright) the use of standalone generators while encouraging CHP use, while the community microgrid programs emerging in the northeast U.S.A are largely supporting natural gas CHP.

The central issue with CHP is its reliance on fossil fuels. The CHP pathway is therefore at most a low-carbon pathway. For this reason, it also lacks scalability, as local air quality regulations limit criteria pollutant emissions. Its ability to reduce carbon emissions depends on the carbon intensity of grid electricity, which is expected to decarbonize over time, as well as the presence of gas combustion in buildings, which are expected to electrify over time. The fossil CHP pathway is therefore limited in efficacy: it locks in long-lived fossil generation that is effective in reducing carbon emissions in the short term but whose effectiveness declines as the bulk grid decarbonizes. (Fuel switching from fossil gas to renewable gas could make CHP carbon-neutral—but that is a distinct pathway considered separately.)

All-Electric: Solar-plus-Storage

One alternative to the use of fossil fuels is to discard them entirely via a fuel-free "all-electric" pathway that pairs solar PV with electric storage such as lithium-ion batteries and that might further include electrified loads and heat pumps, deep energy efficiency, and demand response. PV and storage provide power and substitute for gas generators, electrified loads obviate gas combusted for heating, and responsive demand provides operational flexibility. Combinations of these technologies offer a legitimate path to zero carbon and zero pollution in microgrids.

Of all pathways, the solar-plus-storage option is perhaps the most touted by advocates of decentralization. It maps technologically to other transitions taking shape in the power sector, where utility-scale PV and batteries are emergent and electrification of load and responsive demand are expected to assume important roles in decarbonization. Like CHP, solar-storage microgrids are not dependent on technological breakthrough nor new energy infrastructure; they are proven solutions with numerous operational today, including three funded in 2014 by the State of California as well as many financed privately (Asmus et al. 2017).

Yet the solar-storage pathway faces fundamental challenges related to performance and applicability, while high costs remain an issue. Fundamentally, the use case for solar-storage microgrids is limited by the high power demand of modern cities and low power density of PV (Smil 2019). For example, an analysis of energy use for a selection of the most common types of buildings in the U.S.A shows power densities of 10–250 Wm^{-2}, with warehouses (5–10 Wm^{-2}), retail spaces (15–20 Wm^{-2}), schools (15–60 Wm^{-2}), and smaller office buildings (15–55 Wm^{-2}) constituting the lower end and hotels (150–165 Wm^{-2}), large offices (190–230 Wm^{-2}), and hospitals (235–260 Wm^{-2}) constituting the higher. Variation derives from thermal loads that differ by climate zone—in this example bounded by temperate coastal and hot inland conditions in southern California. In comparison, average annual solar irradiance in coastal and inland California is approximately 210 Wm^{-2} and 240 Wm^{-2}, respectively. With 19% efficient PV cells, PV power densities are 40–46 Wm^{-2}. Thus, only for a subset of these buildings—warehouses, retail spaces, schools, and small office buildings—is PV able to provide a surplus of energy on an annual basis. For the remaining buildings, 100–600% more space is needed for PV installations to achieve parity between annual energy demand and generation. In all cases, substantial energy storage is required to shift excess PV generation to the nighttime when operating in island mode. Demand response can reduce building power density during times of stress by turning off or delaying non-critical load, whereas building electrification increases building power density because it electrifies loads otherwise met with gas.

The solar-storage pathway also faces challenges of cost and performance. Compared to solar-storage microgrids, microgrids that further include fossil fuel generators can provide higher levels of reliability at lower cost because fossil fuel generators are inexpensive by comparison and with fuel reserves are able to supply power day and night. Because PV power is weather-dependent, microgrids depend heavily on storage during times of stress, possibly requiring significant oversizing of storage to ensure islanded operation. As such, microgrids with fossil generators outperform solar-storage microgrids on a reliability-per-cost basis.

Ultimately, low power density and lack of space for PV installations in urban environments limit scalability. Barring radical breakthroughs in cell efficiency or cost decreases for high-efficiency cells, distributed PV will remain unable to meet the power demand of much of modern civilization. A restriction on fossil generators achieves climate benefits but limits applicability because fossil generators are high–power-density DERs that are best suited to serving high-density urban areas. Absent new carbon-minded policy, such as carbon pricing, addressing resiliency and climate change simultaneously with solar-storage microgrids, requires sacrificing cost, performance, or both.

Biomethane

A second alternative to the use of fossil fuels is to substitute them with biomethane while keeping the generator. Biomethane, or renewable natural gas, is biogas that has been cleaned of impurities, upgraded to pipeline quality, compressed, and injected into the natural gas network. Biogas can be sourced from landfills, wastewater treatment plants, municipal solid waste, and manure, among others. Although its emissions when combusted are similar to those of fossil natural gas, biomethane is considered carbon-neutral because it displaces fossil gas in the power sector and when sourced from waste streams reduces emissions from the source sector (either industry or agriculture).

Although in some cases microgrids are collocated with a source of biogas—the MCAS Miramar microgrid in San Diego, California, which uses landfill gas, is one such example—production and consumption need not be physically linked. With such "directed" biomethane, a source facility captures and treats previously vented or flared biogas and injects it into the local gas network, displacing an equal volume of fossil gas as a result. The consuming generator, located perhaps tens or hundreds of kilometers distant, combusts pipeline gas locally but through reporting methodologies is credited with the combustion of the injected, carbon-neutral biomethane. Allowing for directed biomethane is practical because the load centers that microgrids serve may be located far from contracted biogas sources. It also enables market forces to identify the cheapest sources of supply. The University of California, as part of its 2025 carbon neutrality initiative, has contracted for directed biomethane to permit continued use of its CHP gas turbines (Meier et al. 2018).

The biomethane pathway is a carbon-neutral pathway and allows microgrids to retain use of fossil fuel generators. Microgrids can therefore achieve high levels of reliability and are not limited in applicability due to issues of power density. Nevertheless, issues of cost, scalability, and pollutant emissions constrain and limit its potential. In the short term, biomethane commands a price premium relative to fossil gas: prices for pipeline-ready biomethane are $10–24/mmbtu depending on the source (Sheehy and Rosenfeld 2017), while U.S. fossil gas prices stand near $3/mmbtu. Whether biomethane microgrids are cost competitive with fossil CHP microgrids depends on the contracted price for biomethane and prevailing carbon pricing; whether they are cost competitive with solar-storage microgrids requires granular analysis of performance and cost that considers local conditions such as electric tariff structures and rules limiting air pollution or pricing carbon. Although biomethane prices exceed those of fossil gas, governments have created policies and incentives to encourage the development of biogas resources, such as methane emission reduction targets and direct financial support for projects. However, the bulk of these incentives, notably the federal Renewable Fuel Standard (RFS) and Low-Carbon Fuel Standard (LCFS) in California, place emphasis on the use of biogas as an alternative transportation fuel rather than as a fuel for electricity generation. Competition with the transportation sector may constrict supply or increase price.

A second major issue is that long-term biogas resources are limited. The most recent estimate of U.S. biogas availability is 331 billion cubic feet per year or about 1–1.5% of U.S. fossil gas consumption (National Renewable Energy Laboratory 2013). Biogas may therefore provide for early adopters, but it lacks scalability to support a large customer base. As with fossil CHP microgrids, scalability might also become constrained jurisdictionally by air quality regulations. Developers could instead use fuel cells, which would emit only water vapor and carbon dioxide, but, at present, these command a significant cost premium compared to internal combustion engines and turbines.

Hydrogen

A fourth pathway, and the final discussed here, is the hydrogen pathway—the use of hydrogen fuel with engines and fuel cells. Although long-term visions for transitioning to hydrogen (Staffell et al. 2019) remain distant, the hydrogen pathway merits inclusion because it is the most promising in theory: it offers a path to reliable, scalable, pollution-free, carbon-free electricity. It is similar technologically to the CHP and biomethane pathways in its use of fueled generators and can further make use of fuel cells to provide pollution-free electricity. In either case, it is a zero-carbon pathway. Although today the majority of hydrogen is produced via steam methane reforming of natural gas, an underlying assumption of a future hydrogen system is production from renewable energy and electrolysis.

Hydrogen can be combusted in conventional spark-ignition engines and gas turbines with only small modification and efficiency losses and with emissions limited to water vapor and nitrous oxides (no carbon dioxide or carbon monoxide). A more efficient, cleaner, and ideal use of hydrogen is via electrochemical oxidation in fuel cells, where water vapor is the sole product. Hydrogen microgrids can provide high levels of reliability (i.e., they are not limited by low power densities) and are carbon-free.

They therefore match CHP and biomethane microgrids for provision of reliability and exceed them for emissions reductions, while the reverse is true in comparison with solar-storage microgrids.

Though the most promising, the hydrogen pathway faces the largest hurdles to implementation. Cost is a major issue in the short term, as hydrogen fuel commands a cost premium relative to fossil gas and the cheapest form of production remains steam methane reforming, which is carbon-intensive. Fuels cells also remain expensive relative to engines for prime power generation applications, although smaller deployments do exist today—notably over 3,000 hydrogen fuel cells that provide backup power for the telecommunications sector (Ma et al. 2019).

In the long term, the use of hydrogen places the largest demand of any pathway on new infrastructure outside the microgrid. Although hydrogen can be blended into the natural gas network in small amounts (concentrations larger than a few percent on a volume basis may embrittle pipeline materials), blending limits are a constraint and require that a future system of centralized hydrogen production use a dedicated transmission network. The development of hydrogen microgrids is therefore reliant on substantial build-out of infrastructure, including pipelines, production facilities, and storage sites—little of which exists today—and would require coordinated buy-in from the public, private industry, and government. Centralized hydrogen production also requires high penetrations of utility-scale renewable energy.

A ubiquitous hydrogen network for production and distribution, as envisioned in a future hydrogen economy, would make hydrogen microgrids fully scalable. In its absence, however, incremental scale-up is a challenge and the need today for distributed refueling networks limits applicability geographically. Experimentation with distributed refueling networks is ongoing, for example, with the U.S. Department of Energy's support for trigeneration fuel cells (U.S. Department of Energy 2016).

Conclusion

In the context of long-term climate stabilization goals, sustainable management and use of microgrids requires zero-carbon pathways for technology and fuel use. Competing objectives for reducing cost and emissions while increasing reliability preclude a single "best" pathway, while applicability and scalability should also be considered when planning long-term support for pathways. With respect to climate goals, the all-electric and hydrogen pathways are ideal because they make use of carbon- and pollution-free generation. The biomethane pathway is also ideal because it incents reductions in industrial and agriculture sector emissions, which must be brought to near-zero. Today, these pathways incur high costs, however, while low-cost options (diesel and CHP) are the least ideal. Aligning microgrid deployments with climate goals therefore requires action by policymakers. Carbon and pollution taxes, bans, or mandates for technology choice, and technology runtime limits are all policy options that can help encourage low-carbon investment choices. It is not the intent here to provide a ranking, however, and further work is needed to understand which options are best and why, or whether such policies are in fact sensible at this time given the nascency of the microgrid industry.

That microgrids be decarbonized to support climate stabilization goals assumes microgrids will play a non-negligible role in the future power system. Their future role remains uncertain, however, and will be guided by policy and market forces. There is another distinct possibility that microgrid deployments will remain limited in application, for example, to only the most critical of infrastructure. In such a scenario, it is possible that negative emissions in other sectors of the economy permit the continued use of fossil fuels in a relatively small number of strategic microgrids.

References

Asmus, P., Forni, A., and L. Vogel. 2017. Microgrid Analysis and Case Study Report. Navigant Consulting, Inc., Report No. CEC-500-2018-022.

Billinton, R., and R.N. Allan. 1996. *Reliability Evaluation of Power Systems*. Springer US.

Darrow, K., Tidball, R., Wang, J., and A. Hampson. 2017. Catalog of CHP Technologies. Environmental Protection Agency.

Flores, R. J., and J. Brouwer. 2018. Optimal design of a distributed energy resource system that economically reduces carbon emissions. *Applied Energy* 232:119–138.

Hanna, R., Ghonima, M., Kleissl, J., Tynan, G., and D.G. Victor. 2017. Evaluating business models for microgrids: Interactions of technology and policy. *Energy Policy* 103:47–61.

Ma, Z., Eichman, J., and J. Kurtz. 2017. Fuel Cell Backup Power Unit Configuration and Electricity Market Participation: A Feasibility Study. National Renewable Energy Laboratory, Report No. NREL/TP–5400–67408.

Ma, Z., Eichman, J., and J. Kurtz. 2019. Fuel cell backup power system for grid service and microgrid in telecommunication applications. *Journal of Energy Resources Technology* 141(6):062002.

Meier, A., S.J. Davis, D.G. Victor, K. Brown, L. McNeilly, M. Modera, R.Z. Pass, J. Sager, D. Weil, D. Auston, A. Abdulla, F. Bockmiller, W. Brase, J. Brouwer, C. Diamond, E. Dowey, J. Elliott, R. Eng, S. Kaffka, C.V. Kappel, M. Kloss, I Mezić, J. Morejohn, D. Phillips, E. Ritzinger, S. Weissman, J. Williams. 2018. University of California Strategies for Decarbonization: Replacing Natural Gas. UC TomKat Carbon Neutrality Project.

National Renewable Energy Laboratory. 2013. Biogas Potential in the United States. Report No. NREL/FS-6A20–60178.

Sheehy, P., and J. Rosenfeld. 2017. Potential to Develop Biomethane, Biogas, and Renewable Gas to Produce Electricity and Transportation Fuels in California. 2017 Joint Agency IEPR Workshop on Renewable Gas.

Smil, V., 2019. Distributed Generation and Megacities: Are Renewables the Answer? *IEEE Power and Energy Magazine* 17(2):37–41.

Staffell, I., Scamman, D., Abad, A.V., Balcombe, P., Dodds, P.E., Ekins, P., Shah, N., and K.R. Ward, 2019. The role of hydrogen and fuel cells in the global energy system. *Energy & Environmental Science* 12(2):463–491.

Tuttle, D.P., Gülen, G., Hebner, R., King, C.W., Spence, D.B., Andrade, J., Wible, J.A., Baldwick, R., and R. Duncan. 2016. The History and Evolution of the U.S. Electricity Industry. White Paper UTEI/2016-05-2. http://energy.utexas.edu/the-full-cost-of-electricity-fce.

U.S. Department of Energy Fuel Cell Technologies Office. 2016. Tri-Generation Success Story: World's First Tri-Gen Energy Station–Fountain Valley.

Willette, S., and P. Asmus. 2019. Microgrid Deployment Tracker 2Q19: Project and Trends in the Global Microgrid Market by Region, Segment, Business Model, and Top Vendors. Navigant Consulting, Inc.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

28

Energy Conservation: Benefits

Eric A. Woodroof,
Wayne C. Turner,
and Steven D. Heinz

Executive Summary

"It's not the age … it's the mileage" …

It is logical that a car driven 25% less each year will last longer. The same is true for most energy-consuming equipment, such as lights, motors, and even digital equipment. By turning "off "energy-consuming equipment when it is not needed, an organization can find a financial jackpot, which extends beyond the utility budget. *It doesn't matter how energy-efficient an organization is, there are savings from turning equipment "off" when it is not needed.* Listed here are some "secret" benefits of energy conservation and these are benefits that can be attained without a negative impact on productivity.

Budgetary Improvements

1. Efficient Net Income: When energy is conserved, utility budgets are reduced. This is no secret, but what is noteworthy is that conservation savings impact a bottom line far more efficiently than many other investment initiatives. *For example: an energy conservation program that saves $100,000 in operating costs is equivalent to generating $1,000,000 in new revenue (assuming the organization has a 10% profit margin). It is more difficult to generate $1,000,000 in new revenue, and would require more marketing, infrastructure, etc. Thus, the energy conservation/efficiency program is an investment with less risk and quickly improves cash flow.*

2. Extended Equipment Lives: If assets are lasting longer (owing to reduced operation per year), replacements are less frequent, thereby reducing capital budget requirements. *For example, if a lighting system is operating 30% fewer hours per year, it could last up to 30% longer. A 15-year replacement policy could be changed to 20 years.* Further savings could result from considering that if equipment lasts longer, then staff/engineering/project management time is reduced for reviewing new equipment proposals, evaluating competing bids, overseeing installation efforts, coordinating invoices and payables with accounting.
3. Reduced Maintenance Costs: When equipment runs fewer hours per year, maintenance material/labor requirements are reduced. *For example, if maintenance on a motor is done on a "run hour" basis and there are less "run hours" per year, there should be fewer maintenance visits. Further, if the motor is part of a ventilation system, air filter replacements would occur less often, reducing material and labor costs. Predictive maintenance technologies can also assist in this strategy and reduce the cost per horsepower by 50%. (Ameritech)*
4. Reduced Risk to Energy Supply Price Spikes*: For example, if less energy is consumed; the operational budget is less vulnerable when electric/gas/heating oil prices hit their seasonal spikes. The avoided costs can be worth millions to a large organization.*

Beyond the large financial benefits mentioned here, there are many strategic benefits of energy conservation, which can significantly add to your organization's "jackpot."

Long-Term Strategic Benefits

1. Ability to Sell "Carbon Credits": Organizations can claim emissions reductions from energy conservation. When energy is saved, power plants do not have to produce as much electricity, thereby reducing "smoke stack" emissions. *Emission benefits from energy conservation can be expressed in terms of "equivalent trees planted," or "equivalent barrels of oil not consumed."* There are environmental markets where "emissions credits" (from energy conservation) can be sold, generating revenue for an organization. These markets are already liquid in Europe (and are motivated by carbon-related legislation). California and other states already require emissions reporting and reductions, and federal regulations are in process that will open the door to a similar trading environment in the United States.[1]
2. Enhanced Public Image: Organizations that conserve/manage energy (thereby reducing emissions) can differentiate themselves as "environmentally friendly" and "good" members of a community. This can have tremendous political, strategic, competitive, and morale-building value for organizational leaders. Many benefits (such as attracting and retaining better employees, faculty, students, clients, suppliers, etc.,) result from being the "leader" in your field. A recent study showed that 92% of young professionals want to work for an organization that is environmentally friendly.[2] Even stock prices of corporations have been proven to improve dramatically (21.33% on average) when energy management programs are announced.[3]
3. Reduced Risk to Environmental/Legal Costs: If assets are replaced less frequently, an organization will generate less waste and be less vulnerable to environmental regulations governing disposal. (Disposal of batteries and fluorescent lamps is already regulated in most states) Greater environmental regulations are inevitable and unforeseen legal costs can pose a significant expense and political risk.[4,5]

As will be shown in the example on the next page, benefits 2–7 represent a significant improvement (18%–50%) to the original savings estimates.

Climate Change and Its Effect on Energy Conservation Approaches

"The Writing Is on the Wall..."

The glaciers are melting and climate change is here. Climate change is a result of changes in the Earth's atmosphere. The growth of "greenhouse gases" between 10,000 years ago and the 1800s was approximately 1% for that period. Since the 1800s, greenhouse gases have increased 33%. Thus, it is logical that this growth is due to human-caused activities with the dawn of the industrial age.

— Time Magazine-Special Report, December 2007. Also quoted by the UN Intergovernmental Panel on Climate Change-February, 2007 as well as the US EPA website: http://www.epa.gov.

The data is compelling and creating change in consumer choices.[6] Consumers are becoming more "green-minded" in their purchases, especially young people and college students. Studies show that more consumers are choosing to reduce their "carbon footprint," and thereby are choosing products, companies, and colleges that are more environmentally friendly. Carbon offset trading growth is greater than 200%: "in June 2007, the Chicago Climate Exchange reported that in the past 6 months, it had already traded 11.8 $MtCO_2e$—more than had been traded in the entire year of 2006"[7] and 69% of consumers shop for brands aligned with a social cause.[8] Federal and state governments are introducing legislation that will mandate carbon emissions reporting and management.[9] In September 2006, Governor Schwarzenegger signed the California Global Warming Solutions Act, which mandates a 25% cut in emissions by 2020 and an 80% cut by 2050. In summary, the need for a "carbon diet" is driving activity in the energy-conservation industry.

The "Good" News ...

Companies, colleges, and governments are responding to this growing "green" consumer market and competitors are innovating to be the "environmental leaders" of their fields. Hewlett-Packard says that in 2004, $6 billion of new business depended on answers to customer questions about the company's environmental record—a 660% growth from 2002.[10] "Sustainability efforts protect our license to grow" said Wal-Mart CEO Lee Scott in 2005.[11] Energy efficiency/conservation is ranked by corporate executives as the no. 1 way to reduce emissions in a cost-effective manner.[12] Because buildings contribute approximately 43% of the carbon emissions in the USA, an opportunity exists to reduce a large part of these emissions and become "environmental heros."[13] In addition, organizations perceived as more "environmentally friendly" can recruit better faculty, students, suppliers, and employees.[14] Finally, the "secret benefits" (discussed in the Executive Summary) are increasing in value and importance. *An energy conservation program is more valuable today because the material, waste, labor, emissions, and risk savings are more valuable in today's economy.*

A Simple Example to Demonstrate the "Secret Benefits"

A lighting conservation measure will serve as the example, although similar calculations could be applied towards motor systems. Motors and lights consume the majority of electricity in a typical building.[15] Computers and other digital equipment are also worth mentioning, because they can consume considerable amounts of energy, and it is noted that "plug loads" (computers, printers, and other digital equipment) have increased significantly during the past 20 years.

For this example, consider a large school with 10,000 light fixtures. Through a variety of energy conservation measures, it is common to reduce consumption by 25%.[16] First, we will calculate the dollar savings from electricity conservation. Then, we will show the "secret benefits," which have impacts beyond the utility budgets. A spreadsheet will illustrate the total savings/benefits.

Benefit 1: Reduced Utility Budget from Lighting Conservation

Assume the fluorescent lights are relatively new and consume 60 watts per 2-lamp fixture and operate 5000 hr/yr. (This example uses a standard T-8 lighting system, although the energy conservation savings would be even greater with a less efficient lighting system, such as a T-12.) Our baseline energy consumption is:

$$= \left(5000 \text{ hr/yr}\right)\left(0.060 \text{ kW/fix}\right)\left(10{,}000 \text{ fix}\right)$$

$$= 3{,}000{,}000 \text{ kWh/yr}$$

If the school pays approximately $.08/kwh, then the dollars spent on electricity for this lighting system:

$$= \$240{,}000/\text{yr}$$

Thus, a 25% reduction from the baseline usage would equal: 750,000 kWh/yr, or $60,000/yr in savings, *which goes immediately to the bottom line* and improves cash flow. Note that we will not count demand (kW) savings as the electrical load reduction would most likely occur during non-occupied hours (off-peak electrical rates). *However, it is not unusual for conservation programs to reduce both kWh and kW.* In addition to "direct dollar savings," there are tax rebates and credits available that can further improve the financial results from energy conservation/ efficiency programs/projects.

Benefit 2: The Value of Extended Equipment Lives (Reducing Capital Budgets)

If lights are used 25% less, the lighting system (ballasts) should last about 25% longer. Fluorescent fixture and wiring replacement costs are not included, as these components typically last longer than the ballast. We will address lamp life as a part of "maintenance costs" in Benefit 3. A lighting ballast is rated for 60,000 hr of operation. If the school operates the lights 5000 hr/yr, they would need to replace the ballasts at the 12th year and dispose of the old ballasts. If there are 5000 ballasts, each costing $25-$55 (material, installation, and disposal costs vary by geographic location), the replacement cost (minimum) at the 12th year would be:

$$= \left(\$25/\text{ballast}\right)\left(5000 \text{ ballasts}\right)$$

$$= \$125{,}000$$

Annualized replacement cost would be:

$$= \left(\$125{,}000\right)\left(1/12 \text{ yr}\right)$$

$$= \$10{,}417/\text{yr}$$

With a use rate of only 3750 hr/yr (a 25% reduction), the ballasts should last 16 years. If replacement occurs at failure or based on run time, these savings automatically occur. If replacements are planned in advance, planners should adjust their schedules to insure that savings are captured from extended equipment lives (not replacing assets pre-maturely). This would reduce the annualized replacement cost to:

$$= \left(\$125{,}000\right)\left(1/16 \text{ yr}\right)$$

$$= \$7813/\text{yr}$$

Thus, the Annualized Savings (calculated as the difference between the original replacement cost minus the reduced replacement cost) are:

$$= \$10{,}417/\text{yr} - \$7813/\text{yr}$$

$$= \$2604/\text{yr}\ (\text{at } \$25 \text{ per ballast})$$

Using the same equations, at $55/ballast, the annualized savings, (from replacing at 16 years instead of 12 years) would be:

$$= \$5729 \text{ per year}$$

Thus, due to extended equipment life, we have reduced the annualized replacement cost by a minimum of $2604/ yr to a maximum of $5729/yr.

Benefit 3: The Value of Reduced Maintenance Costs (Operating Expenses, Not Capital Replacements)

If the lights are used 25% less, the lamps should last about 25% longer. (Note that if lamps are turned "on" and "off " frequently at less than 3 hr intervals, the lamp's expected life will be reduced by approximately 25%, which would erode the savings in this category. Lamp life is rated at the factory by turning lamps on and off every 3 hr until they burn out. If the frequency of on/off cycling is less than 3 hr, lamp lives will decline by 25% on average. Therefore, turning a lamp off for longer periods is better than shorter periods. For example, it is better to find locations where you can turn off lamps for 5 hr out of 15 hr, instead of 1 min out of every 3 min, although the % time off is the same.) A typical fluorescent lamp life is 20,000 hr. With a use rate of 5000hr/yr, the school would need to replace lamps at the 4th year. If there are 10,000 lamps, each costing $3–$5 (material, installation, and disposal costs vary by location), the replacement cost (minimum) at the 4th year would be:

$$= (\$3/\text{lamp})(10{,}000 \text{ lamp}) = \$30{,}000$$

Annualized replacement cost would be $30,000/4 = $7500.

With a use rate of only 3750 hr, the lamps should last 5.3 years, thereby reducing the annualized replacement cost to:

$$= \$30{,}000/5.3 \text{ yr}$$

$$= \$5660/\text{yr}$$

Thus, Annualized Savings are:

$$= \$7500 - \$5660/\text{yr}$$

$$= \$1840/\text{yr}\ (\text{at } \$3/\text{lamp})$$

Using the same equations, at $5/lamp, the re-lamping cost would be $50,000 and the annualized savings from replacing at 5.3 years instead of at 4 years would be = $3066/yr.

Thus, due to extended lamp life, we have reduced the annualized maintenance cost by a minimum of $1840/yr to a maximum of $3066/yr.

Benefit 4: The Value of Reduced Risk to Energy Supply Price Spikes

Assume that on average, for one-quarter of the year, energy prices are 25%–50% higher ($.02–$.04 more per kWh) due to seasonal/supply spikes. Similar calculations could be used for systems that use natural gas, owing to its seasonal volatility.

If we are using less energy, we will pay less of a premium for the price spike. The avoided price spike premium is equal to:

$$= (\text{price premium})(\text{kWh saved})(\text{premium period})$$

$$= (\$.02/\text{kWh})(750{,}000\ \text{kWh/yr})(1/4)$$

$$= \$3750/\text{yr}$$

Using the same equations, a 50% price spike would represent an avoided premium worth:

$$= (\text{price premium})(\text{kWh saved})(\text{premium period})$$

$$= (\$.04/\text{kWh})(750{,}000\ \text{kWh/yr})(1/4)$$

$$= \$7500/\text{yr}$$

Thus, owing to reduced risk from price spikes, the avoided premiums are $3750–$7500/yr.

Benefit 5: The Value of Carbon Credits

According to the Environmental Protection Agency,[17] 1.37 lbs of CO_2 are created for every kWh burned. So, if we are saving 750,000 kWh/yr, the avoided power plant emissions would be equivalent to:

$$= (750{,}000\ \text{kWh saved})(1.37\ \text{Ibs of}\ CO_2/\text{kWh})$$

$$= 1{,}027{,}500\ \text{Ibs of}\ CO_2\ \text{saved per year F}$$

Translating lbs to Metric Tons:

$$= (1{,}027{,}500\ CO_2)(.000454\ \text{Metric Tons/Ib})$$

$$= 466.5\ \text{Metric Tons of}\ CO_2\ \text{saved per year}$$

These avoided power plant emissions could be claimed as "carbon credits" and sold to another party who wants to buy "carbon credits." An "aggregator" may be required to trade carbon credits in small quantities. Note that as of this printing, European prices for carbon credits are well over five times the price of carbon credits in the USA. The US carbon market is expected to follow Europe's lead as US regulations begin to take effect. Therefore, it is logical that the US prices will approach the European prices, which are currently at $34/metric ton.

Assuming for now a market price of $6 per metric ton, the additional revenue generated by selling the carbon credits would be:

$$= (466\ \text{Metric Tons of}\ CO_2/\text{yr})(\$6/\text{M} - \text{Ton})$$

$$= \$2799/\text{yr}$$

Using the same equations, at \$30 per metric ton, the additional revenue generated by selling the carbon credits would be:

$$= (466 \text{ Metric Tons of CO}_2/\text{yr})(\$30/\text{M} - \text{Ton})$$

$$= \$13{,}980 \text{ per year}$$

Thus, due to the new carbon market, there is a possible additional revenue stream worth a minimum of \$2799 to a maximum of \$13,980 per year from selling carbon credits. In addition, as carbon prices go higher … so does the value of this new revenue stream.

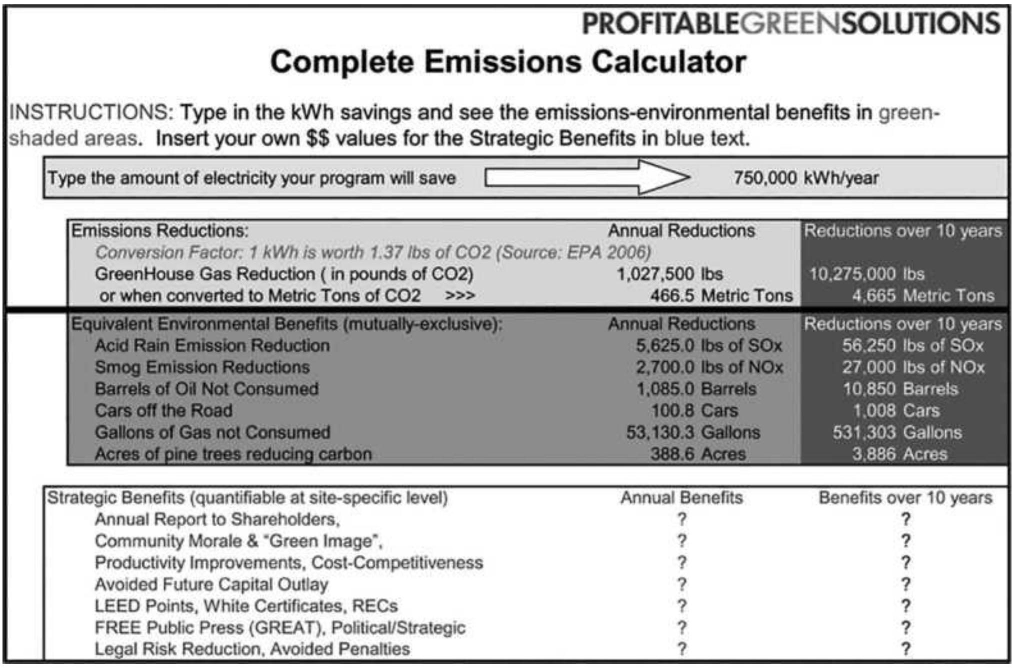
Benefit 6: The Value of Enhanced Public Image

Although calculation of this value is difficult and is not generalized here, it can be far greater than any of the benefits mentioned above. In today's "green-minded" economy, many organizations such as Patagonia, Google, GE, and Home Depot, have used "green" programs as a very effective marketing tool to differentiate themselves from the competition, achieve business objectives, secure and retain talent, improve productivity, and capture a greater market share.

The green-shaded area of Figure 1 shows the "equivalent environmental benefits" from avoided power plant emissions. These reductions/benefits can be published in various places to improve the organization's green image with employees, clients, students, suppliers, distributors, shareholders, and other groups relevant to the success of an organization.

Thus, due to energy conservation program, the school can claim environmental benefits equivalent to removing 1008 cars off the road, thereby improving the school's public image. Although not calculated here, the benefits of attracting better faculty, students, employees, etc., could far out-weigh all the benefit estimates in this entry.

See Figure 1 for additional expressions of environmental benefits.



PROFITABLEGREENSOLUTIONS

Complete Emissions Calculator

INSTRUCTIONS: Type in the kWh savings and see the emissions-environmental benefits in green-shaded areas. Insert your own \$\$ values for the Strategic Benefits in blue text.

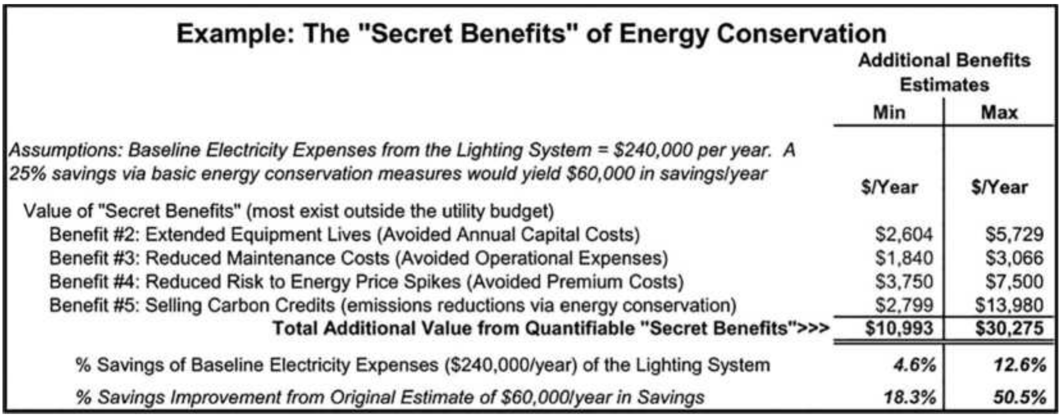
Type the amount of electricity your program will save ⇨ 750,000 kWh/year

Emissions Reductions:	Annual Reductions	Reductions over 10 years
Conversion Factor: 1 kWh is worth 1.37 lbs of CO2 (Source: EPA 2006)		
GreenHouse Gas Reduction (in pounds of CO2)	1,027,500 lbs	10,275,000 lbs
or when converted to Metric Tons of CO2 >>>	466.5 Metric Tons	4,665 Metric Tons

Equivalent Environmental Benefits (mutually-exclusive):	Annual Reductions	Reductions over 10 years
Acid Rain Emission Reduction	5,625.0 lbs of SOx	56,250 lbs of SOx
Smog Emission Reductions	2,700.0 lbs of NOx	27,000 lbs of NOx
Barrels of Oil Not Consumed	1,085.0 Barrels	10,850 Barrels
Cars off the Road	100.8 Cars	1,008 Cars
Gallons of Gas not Consumed	53,130.3 Gallons	531,303 Gallons
Acres of pine trees reducing carbon	388.6 Acres	3,886 Acres

Strategic Benefits (quantifiable at site-specific level)	Annual Benefits	Benefits over 10 years
Annual Report to Shareholders,	?	?
Community Morale & "Green Image",	?	?
Productivity Improvements, Cost-Competitiveness	?	?
Avoided Future Capital Outlay	?	?
LEED Points, White Certificates, RECs	?	?
FREE Public Press (GREAT), Political/Strategic	?	?
Legal Risk Reduction, Avoided Penalties	?	?

FIGURE 1 Complete emissions calculator.



Example: The "Secret Benefits" of Energy Conservation

	Additional Benefits Estimates	
	Min	Max
Assumptions: Baseline Electricity Expenses from the Lighting System = $240,000 per year. A 25% savings via basic energy conservation measures would yield $60,000 in savings/year	$/Year	$/Year
Value of "Secret Benefits" (most exist outside the utility budget)		
Benefit #2: Extended Equipment Lives (Avoided Annual Capital Costs)	$2,604	$5,729
Benefit #3: Reduced Maintenance Costs (Avoided Operational Expenses)	$1,840	$3,066
Benefit #4: Reduced Risk to Energy Price Spikes (Avoided Premium Costs)	$3,750	$7,500
Benefit #5: Selling Carbon Credits (emissions reductions via energy conservation)	$2,799	$13,980
Total Additional Value from Quantifiable "Secret Benefits">>>	**$10,993**	**$30,275**
% Savings of Baseline Electricity Expenses ($240,000/year) of the Lighting System	***4.6%***	***12.6%***
% Savings Improvement from Original Estimate of $60,000/year in Savings	***18.3%***	***50.5%***

FIGURE 2 Secret benefits of energy conservation. (Note: Estimates are Conservative because Dollar Values for Benefits #6 and #7 were not included here).

Benefit 7: The Value of Reduced Risk of Environmental/Legal Costs

Although calculation of this value is also difficult and is not generalized here, it can be very significant. The risk is real, but unknown. This is demonstrated by the following environmental disasters that significantly crippled or destroyed the organizations deemed responsible:

- The Union Carbide accident in Bhopal
- Love Canal's hazardous waste
- Mercury poisoning at Alamogordo, NM

It is also interesting to note that Exxon's penalties and fees were four times the actual clean-up costs for the Valdez oil spill.

More relevant to this entry is that emissions regulations are quite likely to become a standard in the United States. Organizations that are implementing energy conservation programs will have a regulatory advantage over those that do not. Inaction could pose legal risks.

Thus, due to its energy conservation program, the school in this example can reduce its risk from unknown environmental and legal risks that may arise in the future.

Fi@@g. 2 summarizes the dollar value from the benefits mentioned in this entry. The approach and calculations for these benefits could be used as a guide to identify the "secret benefits" of other energy consuming systems, such as HVAC and motors, etc.

Conclusion

This entry has presented additional benefits from energy conservation. The example described an energy conservation project that was achieving a 25% reduction in electrical consumption from the lighting system. Beyond obvious energy savings, the "secret benefits" 2–5 yield additional value worth $10,993 to $30,275 per year. *In other words, if energy conservation saves 25% of a utility budget, the "secret benefits" are worth an additional 4.6%–12.6%.*

Looking at this in a different way, the "secret benefits" contribute additional value worth a minimum 18% improvement from the original estimated savings of $60,000 per year. *In other words, if we value the secret benefits as worth only an additional $10,993, this represents a minimum improvement of 18% to our energy savings of $60,000. In addition, there is a $4660 value improvement for each $10 rise in US carbon prices.*

Finally, all estimates in this entry only included the quantifiable "secret benefits" (benefits 2–5). Actual values could be much higher when accounting for enhanced public image and a reduction in legal and environmental risks (benefits 6 and 7).

We hope that this entry motivates additional action for energy conservation, dollar savings, and environmental benefits.

Acknowledgments

This entry originally appeared as "Overcoming Barriers to Approval for Energy and Green Projects" in *Strategic Planning for Energy and Environment,* Vol. 27, No. 3, 2008. Reprinted with permission from AEE/Fairmont Press.

References

1. Senate pases cap and trade legislation. New York Times, December 6, 2007.
2. How going green draws talent, cuts costs. Wall Street Journal, November 13, 2007.
3. Wingender, J.; Woodroof, E. When firms publicize energy management projects: Their stocks prices go up. Strategic Planning for Energy and the Environment 1997 (Summer Issue).
4. McCain-Lieberman Senate Proposal, 110th Congress, 8/2/2007.
5. Mayor Bloomberg calls for carbon tax. New York Times Articles, November 7, 2007.
6. If the entire world lived like North Americans, it would take three planet Earths to support the present world population (in 2006)—Source: Harvard Business Review on Business and the Environment.
7. State of the voluntary carbon market 2007, The Chicago Climate Exchange, July 18, 2007.
8. Survey data from United States Green Building Council.
9. Bingaman-Specter, Kerry-Snowe, Standers-Boxer Senate Proposals, 110th Congress, 8/2/2007.
10. Daniel, E. Green to Gold; Yale University Press, 2006.
11. Harvard Green Campus Initiative, UCSB Sustainability Program.
12. Getting Ahead of the Curve: Corporate Strategies That Address Climate Change. Pew Center on Global Climate Change, 2006.
13. Pew Center on Global Climate Change. The U.S. Electric Power Sector and Climate Change Mitigation and Towards a Climate Friendly Built Environment.
14. How going green draws talent, cuts costs. Wall Street Journal, November 13, 2007.
15. Association of Energy Engineers. Certified Energy Manager Program Workbook, 2008.
16. Gregerson, J. Cost effectiveness of commissioning 44 existing buildings, 5th National Conference on Building Commissioning.
17. EPA. Avoided power plant emissions calculations (updates October 2006. Available at http://www.ProfitableGreenSolutions.com ("Resources" tab).

29

Energy Conservation: Industrial Processes

Harvey E. Diamond

Introduction and Scope

Energy conservation is a broad subject with many applications in governmental, institutional, commercial, and industrial facilities, especially because energy costs have risen so high in the last few years and continue to rise even higher. Energy conservation in industrial processes may well be the most important application—not only due to the magnitude of the amount of potential energy and associated costs that can be saved, but also due to the potential positive environmental effects such as the reduction of greenhouse gases associated with many industrial processes and also due to the potential of the continued economic success of all of the industries that provide jobs for many people.

This entry will focus on energy conservation in industrial processes—where energy is used to manufacture products by performing work to alter feedstocks into finished products. The feedstocks may be agricultural, forest products, minerals, chemicals, petroleum, metals, plastics, glass, or parts from other industries. The finished products may be food, beverages, paper, wood building products, refined minerals, refined metals, sophisticated chemicals, gasoline, oil, refined petroleum products, metal products, plastic products, glass products, and assembled products of any kind.

This entry will distinguish industrial processes and the characteristics that differentiate them in order to provide insight into how to most effectively apply energy conservation within industries. The level of applied technology, the large amount of energy required in many cases to accomplish production, the extreme conditions (e.g., temperature, pressure, etc.,) that are frequently required, and the level of controls that are utilized in most cases to maintain process control will be addressed in this entry.

This entry will outline the analytical procedures needed to address energy conservation within industrial processes and will comment on general analytical techniques that will be helpful in analyzing energy consumption in industrial facilities.

Many of the main energy intensive processes, systems, and equipment used in industries to manufacture products will be identified and discussed in this entry and some common ways to save energy will be provided.

This entry will cover main energy intensive processes, systems, and equipment in a general format. If more in-depth instruction is needed for explanation of a particular industrial process, system, or type of equipment or regarding the analytical procedures required for a specific process, then the reader should refer to the many other articles included in this Encyclopedia of Energy Engineering and Technology, to references at the Association of Energy Engineers, to the references contained in this entry, and if further detail is still needed, then the reader should contact an applicable source of engineering or an equipment vendor who can provide in-depth technical assistance with a specific process, system, or type of equipment.

In addition to the analytical methods of energy conservation, managerial methods of energy conservation will be briefly discussed. The aspects of capital projects versus managerial and procedural projects will be discussed. The justification of managerial efforts in industrial processes will be presented.

Industrial Processes—Differentiation

Industrial processes require large amounts of energy, sometimes the highest level of technology, and often require very accurate process controls for process specifications, safety, and environmental considerations.

Industrial processes utilize an enormous amount of energy in order to produce the tons of production that are being produced within industrial facilities. Industrial processes utilize over one-third of the total energy consumed in America.[1] Consider the amount of energy that is required to melt all of the metals being manufactured, to vaporize all of the oil and gasoline being refined, to dry all of the finished products that are made wet, to heat all of the chemicals that react at a certain temperature, to vaporize all of the chemicals that must be distilled for purity, to vaporize all of the steam that is used to heat industrial processes, to mechanically form all of the metal objects that we use, etc.,—this list is too long to be fully included in this entry. This is an enormous amount of energy that produces all of the things that humans need and use—food, clothes, homes, appliances, cars, municipal facilities, buildings, roads, etc.

The level of technology required by current industrial processes is the highest in many cases and it is always at a high level in most industrial processes. Most industrial processes are utilizing technology that has been developed in the last 100 years or so, and consequently it has been further improved in the most recent years. Industrial processes most often utilize aspects of chemistry and physics in a precise manner in order to produce the sophisticated products that benefit people in our culture today. Very often, industrial processes require a very high or low temperature or pressure. Often they require a very precise and sophisticated chemistry and commonly they require highly technical designed mechanical processes. The application of electrical equipment and facilities in industrial processes is the highest level of technology for electrical power systems.

Industrial processes often require the highest level and accuracy of controls in order to produce products that meet product specifications, keep processes operating in a safe manner, and maintain environmental constraints. Due to each of these requirements or due to a combination of these requirements, the process controls for the processes within industrial facilities are often real-time Distributed Control Systems (DCSs), that are of the most sophisticated nature. A typical DCS for industrial processes functions to control process variables instantaneously on a real-time basis, whereby each process variable is being measured constantly during every increment of time and a control signal is being sent to the control element constantly on a real-time basis. The accuracy of a DCS in an industrial facility today is comparable to that of the guidance systems that took the first men to the moon

Most industrial facilities with DCS controls also utilize a process historian to store the value of most process variables within the facility for a certain increment or period of time. The stored values of these process variables are used for accounting purposes and technical studies to determine optimum operating conditions and maintenance activities.

Energy Conservation Analyses for Industrial Processes

In any industrial facility, the first analysis that should be performed for the purpose of energy conservation should be that of determining a balance of the energy consumed for each form of energy. This balance is used to determine how much energy is consumed by each unit, area, or division of the plant or possibly by major items of equipment that are consuming major portions of the energy consumed by the plant. This balance should be determined for each form of energy, whether it is for natural gas, electricity, coal, fuel oil, steam, etc., (see Table 1 for an example of an energy balance). It might be best that this determination not be called a balance (in that the numbers might not exactly come to a precise balance) but that it sufficiently quantifies the amount of energy consumed by each unit, area, or division of the plant. A better term for this determination might be an "Energy Consumption Allocation." The term balance is more usually applied to chemical and thermodynamic processes where heat and material balances are worked together mathematically to determine a calculated variable and the numbers have to exactly balance in order to arrive at the correct mathematical solution.

Once the amount of energy that is being actually consumed by each part of a plant has been determined, an energy consumption analysis should be performed for each item of energy consuming equipment and each major energy consuming facility in order to determine how much energy should realistically be consumed by each part of the plant. Notice that these calculations are called "realistic" as opposed to just theoretical because the object of these calculations is to determine as closely as possible how much actual energy each item of equipment or part of the plant should be consuming. By comparing these calculations with the actual energy consumption allocations mentioned above, it should be possible to obtain at least an initial indication of where energy is being wasted in a plant (see Table 2 for an example of an energy consumption analysis for a process feed heater). During the course of obtaining the values of process variables that are required to make these energy consumption analyses, it is

TABLE 1 Energy Balance

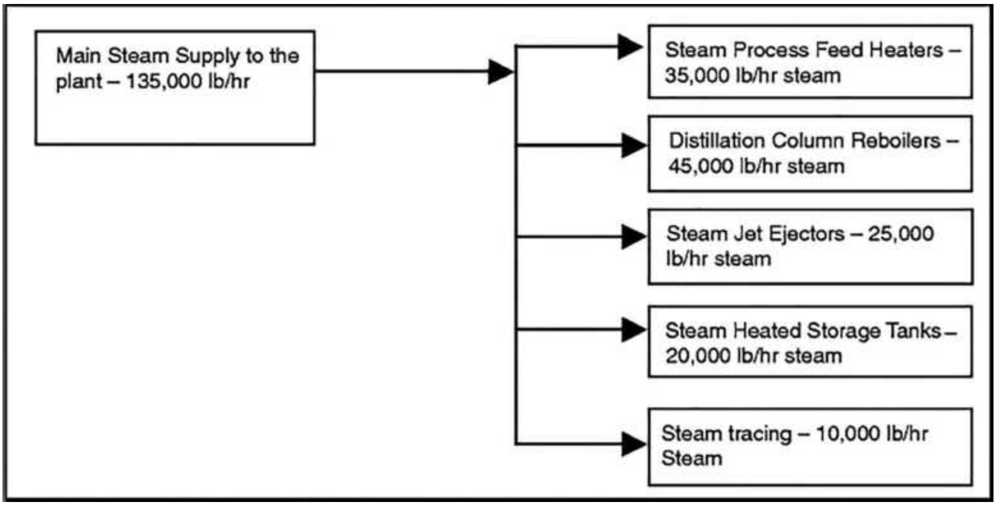


TABLE 2 Example of Energy Consumption Analysis: Process Feed HTR

- A process feed heater heats 1,199,520 lb/day of liquid feed material with a specific heat of 0.92 Btu/lb-°F, from 67 to 190°F. A realistic heater efficiency for this type of heater is determined to be 88%. The amount of realistic heat required for this heater is calculated to be: Q = 1,199,520 lb/day × 0.92 Btu/lb-°F × (190°F – 67°F) ÷ 0.88 = 154,247,367.3 Btu/day of realistic heat consumption.
- It is observed that this feed heater is consuming 186,143,720 Btu/day.
- This feed heater is being operated in a wasteful way and is wasting over 20% of its heat.

possible that indications will be observed of energy wastage due to the presence of an inordinate value of some process variable, such as too high or low of a temperature or pressure. When this type of indication is discovered, it usually also provides insight into what is operating in the wrong way to waste energy. There are numerous instances of energy wastage that can be discovered during these analyses, such as the inordinate manual control of a process, loose operational control of a process variable, or simply not shutting down a piece of equipment when it is not needed.

The analyses discussed in the above paragraph encompass all technical engineering science subjects, such as chemistry, thermal heat transfer, thermodynamics, fluid mechanics, mechanical mechanisms, and electrical engineering.

The next set of energy conservation analyses that should be performed are used to calculate the efficiencies of each item of equipment or facility to which efficiency calculations would be applicable, such as boilers, fired heaters, furnaces, dryers, calciners, and all other thermodynamic processes (efficiency calculations for boilers and other combustion equipment is available in the *Energy Management Handbook* by Wayne C. Turner and Steve Doty[2] and in the *Guide to Energy Management* by Barney L. Capehart, Wayne C. Turner, and William J. Kennedy)[3] and for electrical and mechanical equipment such as motors, pumps, compressors, vacuum pumps, and mechanical machinery. Once the actual efficiencies of any of the above have been determined, these numbers can be compared to the realistic efficiency for the type of equipment or facility that is prevalent throughout industry. These calculations and comparisons will also reveal wastage of energy and will frequently identify the causes of energy wastage and the possible issues to be corrected.

The next level of energy conservation analysis that may be performed is process analysis that can be conducted on a particular chemical, thermodynamic, thermal, fluid flow, mechanical, or electrical process. These analyses are usually performed by experienced engineers to examine the process itself and the process variables to determine if the process is being operated in the most effective and efficient manner. Here again, an indication will be provided as to whether or not energy is being wasted in the actual operation of the process. Chemical, thermo-dynamic, thermal, fluid flow, and other processes, as well as combinations of any of these processes can often require process simulation software such as PROMAX by Bryan Research and Engineering, Inc.,[4] in order to properly analyze these processes. The analysis of distillation columns, evaporators, and dryers can fall into this category. A good example would be the process analysis of a distillation column to determine if an effective and efficient level of reflux to the column and reboiler duty is being used.

Another analysis that has been very useful in the past few years in identifying energy conservation projects is Pinch analysis. This analysis is performed on thermodynamic and thermal processes in order to identify sources of energy within existing processes that can be used to supply heat for these processes instead of having to add additional heat to the entire process. The net effect is to reduce the amount of energy required for the overall process. The performance of a Pinch analysis on a particular process or facility will usually identify capital projects where revisions to the facility can be made to decrease the total amount of energy required. These are very often waste heat recovery projects. See "Use Pinch Analysis to Knock Down Capital Costs and Emissions" by Bodo Linnhoff, Chemical Engineering, August 1994[5] and "Pinch Technology: Basics for the Beginners."[6]

Main Industrial Energy Processes, Systems, and Equipment

This section provides an overview and a list of the more common energy intensive industrial processes that are used to manufacture products in industrial facilities. Most energy intensive industrial processes can be classified into about eight general process categories—process heating, melting, chemical reactions, distillation-fractionation, drying, cooling, mechanical processes, and electrical processes. These processes are intended to be the main general energy intensive processes that are most commonly used and to which variations are made by different industries in order to make a specific product. In this regard, this is an overview—these processes are often not the specific process but a general category to which variations can be made to achieve the specific process.

In the following paragraphs, each process will be discussed by addressing its description, what systems it utilizes, what products are generally made, how it uses energy, and frequent ways that energy can be saved.

Common energy consuming systems and equipment that work to manufacture products in industrial facilities are also listed below and discussed in the same manner as the main industrial processes, as they are also common to industrial facilities and are related to these processes.

Process Heating

- *Description.* The addition of heat to a target in order to raise its temperature. Temperatures can range from the hundreds to the thousands in industrial process heating.
- *Energy form.* Heat must be generated and transferred to the intended object or medium.
- *Energy unit.* Btu, calorie, joule, therm or watt-hour.
- *Examples.* The application of heat in order to heat feed materials, to heat chemical processes, to heat metals for forming, to heat materials for drying, to heat materials in a kiln or calciner, to heat minerals and metals for melting.
- *Applied systems.* Combustion systems, steam systems, thermal systems, hot oil systems, heating medium systems such as Dowtherm[7] or Paracymene,[8] and electrical resistance or induction heating systems.
- *Common equipment.* Boilers, furnaces, fired heaters, kilns, calciners, heat exchangers, waste heat recovery exchangers, preheaters, electrical resistance heaters, and electrical induction heaters.
- *Common energy conservation issues.* Keeping the heat targeted at the objective—proper insulation, seals on enclosures, eliminating leakage, and eliminating unwanted air infiltration. Control issues—maintaining sufficient control of the heating process, temperatures, and other process variables to avoid waste of heat. Management issues—shutting down and starting up heating processes at the proper times in order to avoid waste of heat and management of important process variables to reduce the amount of heat required to accomplish the proper process. Application of Pinch Technology—identify process areas where heat can be recovered, transferred, and utilized to reduce the overall process heat requirement. Waste heat recovery.

Melting and Fusion

- *Description.* The addition of heat or electrical arc energy at a high temperature in order to melt metals, minerals, or glass. The melting process involves more than just process heating, it involves fluid motion, fluid density equilibrium, chemical equilibrium, cohesion, and sometimes electromagnetic inductance. Reference: “The study showed that the fluid equations and the electromagnetic equations cannot be decoupled. This suggests that arc fluctuations are due to a combination of the interactions of the fluid and the electromagnetics, as well as the rapid change of the boundary conditions.”[9]

- *Energy forms.* Heat at high temperatures or electrical arc energy in the form of high voltage and high current flowing in an arc.
- *Energy units.* Heat, Btu, calorie, joule, or therm.
- *Electrical arc. kWhrs.*
- *Examples.* Melting of ores in order to refine metals such as iron, aluminum, zinc, lead, copper, silver, etc. Melting of minerals in order to refine minerals such as silica compounds, glass, calcium compounds, potassium alum, etc.
- *Applied systems.* Combustion systems, chemical reactions, and electrical systems.
- *Common equipment.* Blast furnaces, arc furnaces, electrical resistance heaters, and electrical induction heaters.
- *Common energy conservation issues.* Pre-condition of feed material—moisture content, temperature, etc. Feed method—efficiency of melting process effected by the feed method, feed combinations, and feed timing. Control of electromagnetics during melting and use of magnetic fields during separation. Over-heating can waste energy without yielding positive process results. Heat losses are due to poor insulation, the failure of seals, or lack of shielding or enclosure.

Chemical Reactions

- *Description.* Chemicals react to form a desired chemical, to remove an undesired chemical, or to break out a desired chemical. The chemical reaction can involve heat, electrolysis, catalysts, and fluid flow energy.
- *Energy forms.* Heat, electrolysis, and fluid flow.
- *Energy units.* Heat, Btu, calorie, joule, or therm.
- *Electrolysis. kWhrs.*
- *Fluid flow.* ft-lbs or kg-m.
- *Examples.* Reaction of chemical feed stocks into complex chemicals, petrochemical monomers into polymers, the oxidation of chemicals for removal, dissolving of chemicals to remove them, reaction of chemicals with other chemicals to remove them, the reaction of lignin with reactants in order to remove it from cellulose, the electroplating of metals out of solution to refine them.
- *Applied systems.* Feed systems, catalysts systems, heating systems, cooling systems, vacuum systems, run-down systems, separation systems, filtering systems, and electroplating systems.
- *Common equipment.* Reactors, digesters, kilns, calciners, smelters, roasters, feed heaters, chillers, pressure vessels, tanks, agitators, mixers, filters, electrolytic cells.
- *Common energy conservation issues.* Close control of heating and cooling for chemical reactions. Close control of all reaction process variables—balance of all constituents, amount of catalyst, proper timing on addition of all components. Management of feedstocks, catalysts, and run-down systems for proper timing and correct balance for highest efficiency. Pinch analysis of feed heating, run-down products, cooling system, etc. Conservation of heating and cooling—proper insulation, sealing, and air infiltration. Waste heat recovery.

Distillation-Fractionation

- *Description.* A thermo-dynamic and fluid flow equilibrium process where components of a mixture can be separated from the mix due to the fact that each component possesses a different flash point.
- *Energy form.* Heat and fluid flow.
- *Energy units.* Heat, Btu, calorie, joule, or therm.
- *Fluid flow.* Ft-Lbs or KG-M.
- *Examples.* Distillation-fractionation of hydrocarbons in oil and gas refineries and chemical plants. Distillation of heavy hydrocarbons in gas processing plants where natural gas is processed to remove water and heavy hydrocarbons.

- *Applied systems.* Feed heating systems, over-head condensing systems, reflux systems, reboil systems, vacuum systems.
- *Common equipment.* Distillation columns or towers, over-head condenser heat exchangers and accumulators—vessels, reflux pumps, reboiler heat exchangers, feed pumps, feed—effluent heat exchangers, vacuum steam jet ejectors.
- *Common energy conservation issues.* Feed temperatures, reflux ratios, reboiler duty. Close control on pressures, temperatures, feed rates, reflux rates, and reboil duty. Management of overall operation timing—running only when producing properly. Concurrent use of vacuum systems—only when needed. Pinch analysis for feed and effluent streams and any process cooling systems. Proper insulation and elimination of lost heat for fired heater reboilers.

Drying

- *Description.* The use of heat and fluid flow to remove water or other chemical components in order to form a more solid product.
- *Energy forms.* Heat and fluid flow.
- *Energy units.* Heat, Btu, calorie, joule, or therm.
- *Fluid flow.* Ft-Lbs or KG-M.
- *Examples.* Spray dryers that dry foods, sugar, fertilizers, minerals, solid components, and chemical products. Rotary dryers that dry various loose materials. Line dryers that dry boards, tiles, paper products, fiberglass products, etc. Other dryers that dry all kinds of products by flowing heated air over finished products in an enclosure.
- *Applied systems.* Combustion systems, steam systems, thermal heating systems, cyclone systems, air filter systems, incinerator systems, Regenerative Thermal Oxidizer (RTO) systems.
- *Common equipment.* Spray dyers, spray nozzles, natural gas heaters, steam heaters, electrical heaters, blowers, fans, conveyors, belts, ducts, dampers.
- *Common energy conservation issues.* Efficient drying process for the components being eliminated. Proper amount of air flowing through dryer for drying. Proper insulation, seals, and elimination of lost heat due to infiltration. Waste heat recovery.

Process Cooling

- *Description.* The removal of heat by a cooling medium such as cooling water, chilled water, ambient air, or direct refrigerant expansion.
- *Energy form. Heat.*
- *Energy unit.* Btu, calorie, joule, or therm.
- *Examples.* Cooling water or chilled water circulated through cooling heat exchanger, an air cooled heat exchanger, or a direct expansion evaporator that cools air for process use.
- *Applied systems.* Cooling water systems, chilled water systems, refrigerant systems, thermal systems.
- *Common equipment.* Cooling towers, pumps, chillers, refrigeration compressors, condensers, evaporators, heat exchangers.
- *Common energy conservation issues.* Use evaporative cooling as much as possible. Keep chillers properly loaded. Restrict chilled water flow rates to where 10°F temperature difference is maintained for chilled water. Limit cooling water pumps to the proper level of flow and operation. Apply Pinch analysis to achieve most efficient overall cooling configuration. Proper insulation, seals, and elimination of air infiltration.

Mechanical Processes

- *Description.* Physical activities that involve force and motion that produce finished products. Physical activities can be discrete or can be by virtue of fluid motion.
- *Energy form.* Physical work.
- *Energy unit.* Ft-Lbs or KG-M.
- *Examples.* Machining of metals, plastics, wood, etc.; forming or rolling or pressing of metals, minerals, plastics, etc.; assembly of parts into products; pumping of slurries thru screens or filters for separation; cyclone separation of solids from fluids; pneumatic conveyance systems that remove and convey materials or products and separate out solids with screens or filters.
- *Applied systems.* Machinery, electrical motors, hydraulic systems, compressed air systems, forced draft or induced draft conveyance systems, steam systems, fluid flow systems.
- *Common equipment.* Motors, engines, turbines, belts, chains, mechanical shafts, bearings, conveyors, pumps, compressors, blowers, fans, dampers, agitators, mixers, presses, moulds, rolls, pistons, centrifuges, cyclones, screens, filters, filter presses, etc.
- *Common energy conservation issues.* Equipment efficiencies, lubrication, belt slippage, hydraulic system efficiency, compressed air system efficiency. Control of process variables. Application of variable speed drives and variable frequency drives. Management of system and equipment run times.

Electrical Processes

- *Description.* The application of voltage, current, and electromagnetic fields in order to produce products.
- *Energy form.* Voltage-current over time; electromagnetic fields under motion over time.
- *Energy units. KWh.*
- *Examples.* Arc welding, arc melting, electrolytic deposition, electrolytic fission, induction heating.
- *Applied systems.* Power generator systems, power transmission systems, amplifier systems, rectifier systems, inverter systems, battery systems, magnetic systems, electrolytic systems, electronic systems.
- *Common equipment.* Generators, transformers, relays, switches, breakers, fuses, plates, electrolytic cells, motors, capacitors, coils, rectifiers, inverters, batteries.
- *Common energy conservation issues.* Proper voltage and current levels, time intervals for processes, electromagnetic interference, hysteresis, power factor, phase balance, proper insulation, grounding. Infrared scanning of all switchgear and inter-connections.

Combustion Systems

Combustion systems are found in almost all industries in boilers, furnaces, fired heaters, kilns, calciners, roasters, etc. Combustion efficiency is most usually a prime source of energy savings.

Boilers and Steam Systems

Boilers and steam systems may well be the most widely applied system for supplying process heat to industrial processes. "Over 45% of all the fuel burned by U.S. manufacturers is consumed to raise steam."[10] Boiler efficiencies, boiler balances (when more than one boiler is used), and steam system issues are usually a prime source of energy savings in industrial facilities.

Flare Systems and Incinerator Systems

Flare and incinerator systems are used in many industrial facilities to dispose of organic chemicals and to oxidize volatile organic compounds. Proper control of flares and incinerators is an issue that should always be reviewed for energy savings.

Vacuum Systems

Vacuum systems are used to evaporate water or other solvents from products and for pulling water from products in a mechanical fashion. Vacuum systems are also used to evacuate hydrocarbon components in the petroleum refining process. Vacuum systems are frequently used in the chemical industry to evacuate chemical solvents or other components from a chemical process. Steam jet ejectors and liquid ring vacuum pumps are commonly used to pull vacuums within these systems. The efficiencies of the ejectors and the liquid ring vacuum pumps can be a source of energy savings as well as the management of vacuum system application to production processes. Pneumatic conveyance systems that utilize a fan or blower to create a low-level vacuum are sometimes used to withdraw materials or products from a process and separate the matter within a screen or filter. For large conveyance systems, the efficiencies of the equipment and the management of their operation can be a source of energy savings.

Furnaces, Fired Heaters, Kilns, Calciners

The above comments on combustion systems are applicable to these equipment items and additional energy savings issues can be found relative to them.

Centrifugal Pumps

Centrifugal pumps are used widely in industries. The flow rate being pumped is a primary determining factor for the amount of power being consumed and it is sometimes higher than required. Good control of the pumping rate is an important factor in saving energy in centrifugal pumps. The application of variable frequency drives to the motor drivers can be a good energy saving solution for this issue.

Fans and Blowers

The flow rate for fans and blowers is analogous to the pumping rate above for centrifugal pumps. Good control of the flow rate and the possible application of Variable Frequency Drive (VFD) apply here as well for fans and blowers.

Centrifugal Compressors

Compressors are used widely in industry. The above discussions of flow rates, control of flow rates, and application of VFDs apply here as well. Centrifugal compressors frequently will have a recycle flow that is controlled in order to prevent the compressor from surging. Close control of this recycle flow at its minimum level is very important for compressor efficiency.

Liquid Ring Vacuum Pumps

As mentioned above in several places, liquid ring vacuum pumps are used widely in industry. The amount of sealing liquid that is recycled to the pump and the temperature of the sealing liquid are important determinates of the efficiency of the Liquid Ring Vacuum Pump (LRVP).

The above overview and list of industrial processes, systems, and equipment has been general in nature due to the limitations of this entry. Greater and more specific familiarity with each of these industrial energy intensive processes, systems, and equipment will yield greater applicable and effective insight into ways to save energy related to each of these items.

Capital Projects versus Improved Procedures

Energy conservation effort applied in industrial facilities can identify capital projects whereby the facilities can be changed in order to achieve greater overall energy efficiency or the efforts can identify changes to in day-to-day operating and maintenance procedures that can reduce waste of energy and also improve the overall efficiency of the facility. Frequently, energy-saving procedural changes to day-to-day operations and maintenance activities within an industrial plant can be identified by taking and recording operating data once the processes, systems, and equipment have been studied and analyzed for energy consumption. Procedural changes to operations and maintenance within an industrial plant can often amount to low costs or possibly no costs to the facility. This aspect of energy conservation is often overlooked by highly technical personnel that have worked hard to design industrial facilities because they have technically designed the facility very well for energy consumption considerations and the more mundane activities related to day-to-day operation and maintenance tend to not register in their highly technical perspective. None-the-less, a considerable amount of energy can usually be saved within most industrial processes, systems, and equipment due to changes in the way they are operated and maintained. A general tendency within industrial plants is that operations will often operate the processes and systems at a point that provides a comfortable separation between an operating variable and its limitation in order to understandably ensure no upsets occur within the process or system. However, with the cost of energy being what it is today, it is frequently found that a significant amount of energy can be saved by operating processes and systems more tightly and efficiently, even though it may require more attention, increased control, and the monitoring of process variables.

Effective Energy Management Systems

Another aspect of energy conservation that can be very productive in saving energy within industrial processes is that of an effective energy management system. An effective energy management system is comprised of operational and maintenance managers functioning in conjunction with an accurate and concurrent data collection system in order to eliminate waste and improve overall efficiency of industrial processes. It is not possible to manage any activity unless the activity is being properly monitored and measured with key performance metrics (KPMs). The data collection system part of an effective energy management system within any industrial facility provides the accurate and concurrent measurement data (KPMs) that is required in order to identify actions that are needed to eliminate waste of energy and improve overall efficiency of the facility. An effective energy management system is first built upon acquiring total knowledge of the facility down into every level of operation and maintenance of the facility. Such a level of thoroughness and complete analysis of energy consumption within a facility is sometimes referred to as *Total Quality Energy* management.[11] Once an effective energy management system has been established and is effectively controlling energy consumption of an industrial facility, it should be maintained, in effect, so that it will continue to monitor KPMs to maintain energy conservation for the facility. An effective energy management system within an industrial manufacturing facility can eliminate as much as two to three percent of the energy costs by eliminating waste of energy on a day-to-day operational and maintenance basis. In most industrial facilities, this level of cost reduction is significant and will justify an effective energy management system.

Current Need for Greater Energy Conservation in Industry

With the present cost of all forms of energy today, it would certainly seem logical that all of industry would be seeking greater energy conservation efforts within their facilities. Unfortunately, many corporate industrial managers are not aware of the true potential of conserving energy within their processes and facilities. Greater awareness of the ability to conserve energy on the basis of increased efficiencies of processes, systems, and equipment is needed; and also due to the application of an effective energy management system. For the good of society and environment, corporate industrial managers should be more open to the possibility of the improvement of industry that will work to sustain their business and improve the world that we live in. This is in opposition to corporate political thinking, which does not want to consider making changes and wants no one to interfere with their present activities. Human beings should be willing to examine themselves and make changes that will make things better. The same outlook should be applied to businesses and industry in order to make things better. Greater management support is needed in industry today to accomplish greater and very much needed increased energy conservation.

Current Application of Increased Energy Management

With the recent technological advancements that have been made in digital computer and communications systems, data collection systems can be implemented in industrial facilities in a much more cost effective manner. Wireless communication systems for metering and data collection systems have advanced dramatically in the last few years and network-based computer communication has enabled whole new systems for measurement and control. With all of these new fields of configuration for data collection systems, with the increased technology, and with the lower costs to accomplish data collection systems, it is now possible to apply energy management systems to industry today with much greater applicability. Hopefully this will be recognized and result in greater applications of effective energy management systems.

From recent observations, it appears that most of industry today is a candidate for improved and more effective energy management systems. In conjunction with the increased technology and lower cost potentials, it seems that there is a definite match between supply and need for the application of increased energy management systems.

Summary

Industrial processes have commonality in processes, systems, and equipment. There are logical and systematic analyses that can be performed in industrial processes that can identify ways to save energy. Effective energy management systems are needed in industry today and there are great possibilities to save energy in industrial processes. Energy can be conserved in industrial processes by analyses that will improve efficiencies, by implementation of procedures that eliminate waste, and by application of an effective energy management system.

References

1. U.S. Department of Energy–Energy Efficiency and Renewable Energy, available at http://www.eere.energy.gov/EE/industry.html (accessed on 2006).
2. Turner, W.C.; Doty, S. *Energy Management Handbook,* 6th Ed.; Fairmont Press: Lilburn, GA, 2005.
3. Capehart, B.L.; Turner, W.C.; Kennedy, W.J. *Guide to Energy Management,* 5th Ed.; Fairmont Press: Lilburn, GA, 2004.
4. Bryan Research and Engineering, Inc. PROMAX; BRE, Bryan, TX; available at http://www.bre.com (accessed on 2006).

5. Linnhoff, B. *Use pinch analysis to knock down capital costs and emissions.* Chemical Engineering August 1994 http://www.che.com.
6. Solar Places Technology. *Pinch Technology: Basics for the Beginners,* available at http://www.solarplaces.org/pinchtech.pdf (accessed on 2006).
7. Dowtherm, Dow Chemical http://www.dow.com/heattrans/index.html (accessed on 2006).
8. Paracymene, Orcas International, Flanders, NJ 07836 available at http://www.orcas-intl.com (accessed on 2006).
9. King, P.E. *Magnetohydrodynamics in Electric Arc Furnace Steelmaking.* Report of Investigations 9320; United States Department of the Interior, Bureau of Mines, available at http://www.doi.gov/pfm/ar4bom.html (accessed on 2006).
10. U.S. Department of Energy-Energy Efficiency and Renewable Energy, available at http://www.eere.energy.gov/EE/industry.html the common ways (accessed on 2006).
11. Energy Management International, Inc. Total Quality Energy, available at http://www.wesaveenergy.com (accessed on 2006).

30

Energy Master Planning

Fredric S. Goldner

Introduction

Developing an energy management program or more broadly energy master planning (EMP) is the process of transitioning an organization's culture from the traditional "fixed cost, line item" view of energy to one in which energy is recognized as the opportunity and risk that it has become.

An EMP can guide an organization in longer-range planning of energy cost reduction and control as part of their facility maintenance, management, and design. An EMP can even lead the energy budget to be recognized as a potential profit center and source of opportunity rather than just another business expense. An EMP moves beyond the confines of traditional engineering to include energy procurement, energy-related equipment purchasing, measurement and verification (M&V), staffing and training, communications, and setting energy consumption targets and tracking/feedback loop systems. The long-term perspective goes beyond simply cutting last year's energy use. It makes energy awareness part of the everyday operation and "mindset" of the organization.

If you're thinking this doesn't apply to your firm or clients because you're too small, think again. This approach works for an organization as small as a single site to an owner with half a dozen small buildings to Fortune 100 companies. The effort and level of detail vary, respectively, but the approach is basically the same. The good news is that there are resources available to assist professionals and the organizations they serve to understand the EMP process and get started on this path.

To be successful an organization must treat energy in the same business-like manner that they do all other major expenses, such as labor and materials. "If you can measure it, you can manage it"[1] is the catch phrase of Paul Allen, the energy manager at Disney World, who has been instrumental in implementing one of the most successful EMPs in the country. "Energy is a competitive opportunity

… Winners manage it effectively!" is the driving force of another highly successful program at Owens Corning. Many other organizations throughout this country and across the globe have recognized that to achieve significant and sustained energy cost control, organization needs to make energy management an integrated part of their business/operations. There are various ways to pursue the process, but a key requirement is an interdisciplinary mix of engineering/technical, behavioral, and organizational or management components. The EMP must be integrated into the basic business operations.

Unexpected Benefits

One of the most potent driving factors in many organizations' efforts to address energy issues is increased profitability that can be realized through reduced/optimized energy expenditures. Beyond the "bottom line" impacts, an EMP can also provide an organization with a more secure energy supply, reduced downtime of systems, improved equipment availability, reduction in maintenance costs/ premature system replacement expenditures, and overall productivity gains. Additional benefits that have been documented include quality-of-life improvements, enhanced product quality, better operational safety, reduced raw material waste in industrial plants, and increased rentability in commercial facilities. An often overlooked outcome of an EMP is the reduction of emissions, among other environmental impacts that help organizations become perceived as better corporate citizens.

As more companies move toward an integrated corporate strategy that links environmental, economic, and social considerations, the results of an EMP can be used to considerable public relations advantage. Ratings in one of the sustainability indices and publication of an annual sustainability report (using, for example, the Global Reporting Initiative guidelines) can give an organization a higher standing in the business community and can result in a higher level of trust by stockholders.

A Process for Optimization

Though energy prices are volatile, and energy security is often far from reliable, facilities now face leaner operating funds and increased directives to do more with less. Optimizing a facility's operations budget frequently means cutting energy costs. But how do you do so without cutting occupant comfort or productivity? How do you know where to start and what steps to take? And how do you persuade upper management that energy costs can be controlled?

While the general goal of an EMP is the same as that of conventional energy management, the two disciplines are far from identical. Traditional energy management, which is technically oriented, is essentially centered around the boiler or mechanical room. Energy master planning, on the other hand, is a business management procedure for commercial, institutional, and industrial operations. With this approach, it is not enough simply to manage installations. The process involves:

- Developing strategies
- Creating processes to fulfill those strategies
- Identifying barriers and finding procedures to overcome them
- Creating accountability
- Providing feedback loops to monitor and report progress

Clarification of the terminology for these disciplines is important, as the terms mean different things to different people. In other English-speaking nations, for example, the defining feature of "energy management" is the emphasis on integration with business practices to analyze, manage, and control energy. In the U.S., however, "energy management" has traditionally referred to developing technical and operational measures involving equipment handled by facility managers—not processes such as energy procurement and business planning usually handled by purchasing agents, production personnel, and corporate economists. A typical U.S. energy manager's responsibility rarely extends much beyond utility bill analysis, an occasional energy audit, or managing installation of system upgrades.

Though an EMP encompasses traditional efforts to cut energy costs, it also includes many steps not usually taken under standard energy management. Rather than being just equipment-oriented, an EMP starts long before a comprehensive energy audit and extends beyond commissioning of new systems. An EMP may be thought of as a road map to savings that starts before and continues after energy- efficiency measures are involved. Why do you need such a map? Because you can't get there if you don't know where "there" is. How many of us are willing to undertake a trip that will have costs and risks (like any business or personal decision) if we don't know where we are going? A map makes clear not just your final destination, but also how to get there—and it leaves no question about the starting point.

An EMP is a process to organize and improve your existing energy-related resources and capabilities. Resources, in this case, include standard operating procedures, institutional memory, and actual records (such as energy bills, plans and blueprints, and energy contracts). Capabilities include facilities staff familiar with mechanical room equipment, consultants for energy costs or usage, energy cost accounting and management systems, and meters and software that monitor them. Once organized and integrated, these resources and capabilities become powerful tools for managing energy and producing savings.

Today's Practices Found Wanting

Current thinking about managing energy often falls short of the EMP perspective. Today's energy management frequently reflects a short-term crisis mentality: A facility manager or energy manager concentrates on whatever immediate 'fires' have to be put out at his facility. By contrast, with an EMP mindset, the energy manager might first look to increase the efficiency of systems already in place and then move ahead to lay a solid foundation for improving performance via tight energy specs and training.

Current thinking is also often characterized by piecemeal 'solutions' that lead to short-sighted component replacement. When equipment breaks down, it gets replaced without anyone's asking whether this is the best option, long-term. Typically, business thinks in a quarterly mindset because of the short budget cycles of our economic system. That kind of "right now and right here" viewpoint creates situations in which life-cycle thinking is not possible because potentially higher initial costs are visible, but potential benefits tend to be invisible. As a result, when first cost becomes the main criterion for purchasing, such a focus distorts planning and decision-making. Too often, facilities choose the cheapest solution, based on the current quarter's budget, without realizing it could cost them more later.

A critical difference between energy master planning and conventional energy management is an orientation toward the future. With an EMP, you don't just look to increase the efficiency of systems already in place. Instead, you plan for new or changed loads based on your detailed information about the facility's long-term business strategy and projected growth.

Energy master planning deals with a longer timeframe than just simple payback periods. For example, typical financial constraints for a commercial building upgrade often dictate a 2- to 3-year timeframe. Energy master planning, however, looks deeper than simple payback and goes beyond merely reducing energy use. Therefore, an energy professional with an EMP considers life-cycle costs and views long-range planning of energy cost minimization/ optimization as part of overall facility maintenance, management, and design. To sustain the savings over time, an EMP calls for hiring an energy manager/ coordinator and setting up an energy team. But who is this energy leader? Not simply the facility manager wearing yet another hat, but rather, a highly trained and, often, certified specialist with sufficient acumen and expertise to understand, handle, and maintain whatever new energy systems and practices are to be put in place. Willingness to identify (from within the organization) or hire an individual with the required qualifications and to define his or her responsibility as managing energy rather than managing the facility is a requisite indication of senior management's true commitment to energy master planning. The policy guidance needs to come from senior management levels to convey "buy-in" throughout the management structure of the organization.

Improving Business as Usual

Energy master planning is one way of creating a new norm of "business as usual," taking its cues from time-tested business management practices. For example, rather than carrying out upgrade projects with a defined start and end point, an energy manager uses processes that continue to turn up new sources of savings. This requires a level of creativity for identifying and capturing new opportunities. If a chiller replacement is needed in Building A, for instance, a better approach might be to expand the capacity of the existing central chilled water plant in Building B and run piping from Building B to Building A.

Without an EMP, the Purchasing department often buys equipment and energy, rather than the Facilities department. And Facilities is so busy handling emergencies that they have little interaction with other departments in the organization, let alone industry groups or other end users. An EMP avoids either departmental isolation and turf wars by including representatives from such departments on the energy team. The team consists of more than an energy manager, facilities personnel, and design and construction specialists. To be effective, it needs to incorporate representatives from every department in the organization impacted by energy. This may mean including purchasing, accounting, engineering, environmental affairs, maintenance, legal, health and safety, corporate relations, human resources and training, public relations and marketing, and members from the rank and file (hourly employees).

Energy master planning is a significant challenge. It often rejects the status quo and may question existing components of an organizational culture that do not support energy master planning. If you have always done something one way, you don't necessarily have to perpetuate what could be a costly mistake. For example, using outdated specs (e.g., calling for T12 lamps instead of T8) allows inefficiency to continue.

Origins of the American EMP Approach

A common set of energy master planning definitions and processes has taken root in English-speaking countries other than the U.S. In the United Kingdom, the energy master planning concept has been practiced for well over a decade and vigorously promoted by the government's Action Energy program. It has been so successful that Canada, New Zealand, and Australia have each adapted the process to their own conditions.

In the U.S., there are a few recent models cover some, but not all, aspects of energy master planning management system for energy (MSE) 2000 is a specialized quality improvement standard (like ISO 14000), developed by Georgia Tech. It's an American National Standards Institute (ANSI)-approved management system for energy that covers all sectors, not only buildings. Georgia Tech offers a certificate program to train energy professionals in this standard. The Association of Energy Engineers delivers the Developing an Energy Management Master Plan and Creating a Sustainable Energy Plan Workshops to both 'real' and 'virtual' end users. Live presentations and online seminars present an energy master planning approach with strong emphasis on integration with business strategy.

Steps to an Energy Master Plan

To fill this gap in the U.S., the Energy Master Planning Institute (EMPI) was established and has developed a set of steps that lay out the process for applying energy master planning to a commercial, institutional, or industrial facility (See Figure 1). This model, which builds on accepted international approaches, offers US organizations a broad and integrated business approach for managing energy that is both strategic and sustainable.

The steps presented in Figure 1 appear sequential, but built into the energy master planning process are series of feedback loops, evident in Figure 1. These should not be overlooked, as they guarantee the viability of the process and offer many points for input from internal and external stakeholders, from the Chief Executive Officer (CEO) to boiler room personnel, and from the local community to organizational peers across the country.

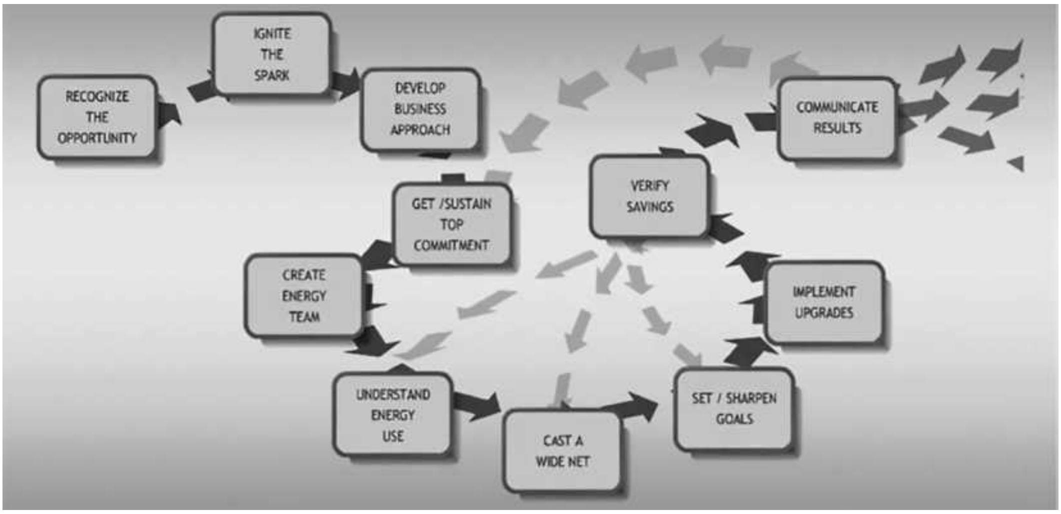


FIGURE 1 Energy Master Planning Institute (EMPI) steps to an energy master plan (EMP). Copyright 2004. Energy Master Planning Institute. All rights reserved.

Recognize the Opportunity

This is where the process starts—becoming aware of the facility's major energy-related opportunities and challenges. Whether it's the facility manager or the energy manager, whoever takes the initial action must define the opportunities and challenges succinctly so they can be clearly communicated. When the leader creates and implements an EMP, that plan can actually generate a revenue stream and gain recognition for the leader's contribution to the company's bottom line. By integrating energy concerns into the overall corporate business strategy, the energy budget will come to be seen as a potential profit center and source of opportunity—and not just an uncontrollable expense.

Ignite the Spark

Since top executives in an organization are the ultimate decision-makers, particularly concerning funding, it's critical to spark their attention early in the process. It's unlikely that the facility manager (or the energy manager, if there is one), has direct access to upper management, so the right individual to make the pitch to the CEO or Chief Financial Officer CFO must be identified. This could be someone along your management chain, or a consultant or board member or a senior officer in another department who can sell an idea at the top. To find the right person, it's important to understand the decision-making structure of your organization as well as the vision, mission, and long-term business plan.

Develop a Business Approach

To get access to senior executives and persuade them to listen to a new idea, you must speak to them in their own language—dollars/ft^2/year savings, not kWh. Since most executives have not been introduced to the bottom-line value of energy master planning, your task is to change management culture so executives no longer view energy as an uncontrollable expense. Part of the marketing message is that managing energy is no different from tracking, controlling, and accounting for the costs of raw materials, IT, personnel, safety, or the corporate fleet.

Obtain and Sustain Top Commitment

Without question, this step is the most difficult in the energy master planning process. Serious commitment from senior management means providing ongoing financial resources and personnel with appropriate credentials. A one-time memo of support is not effective. To secure top executive commitment, you must show how an EMP can support key business goals, such as growth, customer satisfaction, or a sharper competitive advantage. You might, for instance, explain that lower operating costs and increased energy efficiency can bring a higher level of occupant comfort, which, in turn, can mean a lower worker absentee rate—and, possibly, a greater employee retention rate. To gain top executive commitment, make sure that CEO and CFO see the dollar savings highlighted in the energy team's regular reports.

Top management commitment is the single most important goal in an effective and lasting EMP. Not only is it crucial to have this commitment, it must also be obvious to everyone throughout the organization. Top management should participate in the program start-up and continue to reinforce that commitment periodically with both words and actions. Such organizations as Walt Disney have achieved this top-level buy-in and gained significant and lasting bottom-line results from energy master planning.

Create an Energy Team

A middle manager in your organization may declare that energy master planning is only the responsibility of the Facilities department. However, when an energy team represents the company's broad interests, energy concerns can be successfully integrated into the overall business plan. Every department that's impacted by the organization's energy use should be invited to participate on the team. This includes

- Facilities
- Construction/engineering
- Purchasing
- Accounting
- Inventory
- Environmental, health, safety
- Legal
- Public affairs
- Property/asset management
- Leasing/real estate
- Risk management
- Security
- Financial service

This should result in the creation of an Energy Committee. Depending on the size of the organization, there may be separate Technical and Steering Energy Committees. In addition to representatives from the departments listed above it is important to include members of the rank and file (hourly employees or line workers). By including such folks in the planning process it allows not only those individuals, but their peers as well, to become aligned with the EMP objectives and particular initiatives during the early stages. Having these employees on the team helps the rest of the rank and file staff (who will be needed to carry out many of the activities) see the energy program as something other than 'just another management flavor of the month.' It also increases success, as feedback from these employees often helps address many of the nuts-and- bolts "bugs" in advance.

The energy team should be headed by an energy manager or coordinator. The energy manager needs to have a mix of technical, people, and communications skills, and must be enthusiastic. He or she should thoroughly understand the organization's operations and should report to someone as close to the top of the organization as possible, so he or she has the clout need to get things accomplished.

Understand the Organization's Energy Use

Conduct disciplined information analysis. Collect and use metered–not just billed–energy data, and analyze it with software designed to manage energy costs. "Fully understanding current energy use practices–the when, where and how much of energy consumption–in detailed qualitative terms is an essential precondition to energy management master planning."[2] For example, to determine exactly where the energy is going, specify and install sub-metering and a data acquisition system. Such tools will help you monitor and collect, as well as analyze, the meter data for electricity, fossil fuels, steam and condensate, and water. For loads of several hundred kW, it's advantageous to use interval meters that allow a close look at 15-minute interval or hour-by-hour use of energy, so you can see where peaks and valleys in usage occur. This kind of metering is also essential for internal billing. Besides establishing points of excessive usage, it can also help manage loads and pinpoint efficiency opportunities.

Cast a Wide Net

Casting a wide net means looking beyond the central plant for savings opportunities, such as lighting, office equipment, elevators, and localized process loads. It means making construction and equipment specs energy-conscious and creating ways to "enforce" such specs. For example, specifying T8 lamps for efficiency upgrades is not sufficient. The architecture, design, and construction staff should build T8s into their specs for non-energy-related upgrades, such as converting a library to training rooms. Casting a wide net also means revising inventory and purchasing practices to support energy efficiency. As in the example above, to ensure that the new installation continues to function properly, the purchasing specs need to list T8s, not T12s.

Set or Sharpen Goals

Once the organization's energy use is determined, strategic thinking linked to the long-term growth objectives will determine how much energy you should aim to save, where the savings should come from, and when those savings should occur. The energy manager along with the energy team, proposes a phase-one timetable with quantifiable goals to reduce energy use and operating and energy costs. Senior management, however, must mandate the goals and schedule or they carry no weight. Ideally, the CEO, CFO, or the Board issues a position calling for measurable reductions at the end of a 3-year fiscal cycle, based on current use: a percentage reduction in peak electrical demand across the entire facility, a reduction in annual kWh consumption of electricity, and a reduction in overall British Thermal Unit (BTU) for fuel per gross square foot. Lesser annual goals will help keep progress on track. A key is to set achievable goals, and then work to meet or beat them.

Implement Upgrades

Because energy managers generally have experience with efficiency upgrades, it is critical not to fall into old patterns of sporadic efforts and a piecemeal approach, with an eye on the quickest payback. To carry out an EMP, start the upgrade process with a comprehensive energy audit—not just a walk-through—to determine which upgrades will give the best results, not just the best rate-of-return. For example, in a commercial building operation, to avoid a rush upgrade for a new tenant, pre-audit all your buildings so you know what work needs to be done before that tenant signs their lease. Or, if you're considering recommissioning, look at it from the longer-term energy master planning perspective. Recommissioning may not be sufficient, as it means only the existing systems would operate more efficiently. But what if they need to be ripped out as part of the upgrade?

Don't fall into the common trap of focusing just on the high-profile, capital-intensive, projects. While a new chiller or micro-turbine cogeneration system may be a good photo opportunity, focusing on the less visible details can often provide the lion's share of savings through lower-cost measures that improve operations and maintenance (O&M) or maintenance and management practices.

Verify Savings

Once the upgrade process is underway, you must first validate the savings with a recognized technique such as the International Performance Measurement and Verification Protocol (IPMVP or MVP). However, instead of following the M&V protocol after the fact, build measurement and verification into the design—that is, install it as part of the upgrade—so you can identify the savings as soon as the device is turned on. Then, set up an energy accounting system that tracks usage and savings, thereby providing objective accountability. All too often energy management activities are not perceived as being of value because the energy team did not plan ahead and position themselves and theirs efforts for recognizable success. A North American floor coverings manufacturer reported how "in the past we would complete a project to find that we did not have the baseline or operational data to judge whether the project was successful."[3] How willing will management be to fund the next energy related project or initiative if the energy team cannot prove that prior efforts actually saved what was projected?

With an energy accounting system, everyone on the energy team will be working from the same data, and you can prepare regular progress reports, plus quarterly reports to the CFO and an annual report to the Board. This is critical: The energy team must account for the savings—or lack of savings-to senior management, using a feedback process so goals and targets can be reset if they aren't met. Use this process to verify energy use and to help identify any new opportunities for savings, and then to fine-tune your goals. If you fall short of your targets, you need to figure out why. Or perhaps you did meet your targets, but parts of the data were wrong. Perhaps the metering is off, or the energy accounting system needs readjustment. Regularly scheduled feedback from these reports will keep the energy master planning process up-to-date and realistic and will ensure accountability. One approach that has been successfully used in many organizations is the use of "energy report cards" and "intra-organizational listings/scorecards." Produced by the energy manager and sent out to all facilities/ operating groups/ departments on a monthly basis, these tools inform as well as motivate, based on the certainty that no one wants to be on the bottom of the list.

Communicate Results

Successful implementation of all stages of the EMP can be a useful vehicle for departmental—and personal—recognition. With a representative of the corporate PR office already on the energy team—enlist that individual's expertise for internal and external communications. Inform the organization's Board about the financial savings and improved asset value. Let the staff and community know about the environmental benefits. Apply for energy awards for national recognition in your sector. And use the good will generated by the success to keep the energy master planning process moving forward.

Tips for Success

From these steps to an EMP, two points emerge as the most critical to success.

The first is ensuring buy-in from the top, with a long term commitment and binding, formal statement of energy policy for the organization. Commitment to action also means that senior executives support the people in the middle—delegating authority to the energy manager or facility manager. If the EMP lacks commitment from the top down, no amount of effort by the energy manager will make it succeed. Once you've caught executives' attention, keep energy master planning on their radar screen by having the energy team build it into the organization's business plan. Such a commitment must also be apparent. Top management must continue to be seen (by all levels of staff) reaffirming their commitment, lest the EMP be perceived as just another short-term initiative of the organization.

The second critical component is line-management accountability, making specific individuals accountable for sustaining the savings. On the management side, executives need to mandate quantifiable goals and targets, strengthened by obligatory deadlines. Such requirements should even be built into job descriptions and evaluated as part of annual performance standards reviews. Likewise, incentives can be offered through these same personnel standards for individuals who meet their energy goals. Employee teams with day-to-day knowledge of the facility's operation can also be organized to identify additional opportunities for savings. At the upper level of the organization, corporate accountability for successful energy management can be communicated (and made public) through such mechanisms as the Global Reporting Initiative, especially in firms with sustainability principles that follow a "triple bottom-line."

Conclusion

Energy master planning is an effective and long-term shift in organizational cultures to address and adapt to the impact energy resources can have on their competitive posture and economic success. Energy is a universal raw material and is essential to the operation of almost every commercial, institutional, governmental, and even nonprofit organization. Significant changes to the basic structure of the energy supply chain and significant energy price volatility, driven by a wide range of political and physical events (e.g., climate change, weather), have made energy both a competitive opportunity and a risk requiring active long-term planning to manage. Energy master planning requires buy-in from top levels of an organization down through the rank and file, selection/hiring of a dedicated energy manager, and creation of an energy team with membership from across the organization. This team must be given the resources and top-level access to develop and implement long-term planning for procurement and operations throughout the organization that places a premium on lasting energy-use reductions and cost optimization, rather than "burst" efforts that provide quick and quickly forgotten energy management efforts.

Successful organizations do not exist on a quarterly basis. They plan for the long-term in all aspects of their operations. Experience has shown that organization-wide adoption of EMP is effective in optimizing energy costs and needs to ensure competitive posture and success in the long-term business reality in which organizations exist. This is the new bottom line for energy.

Acknowledgments

Permission has kindly been granted by the Energy Master Planning Institute for use of materials on its Web site and for materials published and presented by Coriolana Simon, founder of the EMPI.

References

1. Allen, P.J.; Kivler, W.B. *Walt Disney World's Utility Efficiency Awards and Environmental Circles of Excellence,* Proceedings of WEEC 1995; Association of Energy Engineers: Atlanta, GA, 1995.
2. Tripp, D.E.; Dixon, S. *Making Energy Management "Business as Usual": Identifying and Responding to the Organizational Barriers,* Proceedings of the 2003 World Energy Engineering Congress; Association of Energy Engineers: Atlanta, GA, 2003.
3. Key, G.T.; Benson, K.E. *Collins and Aikman Floorcoverings: A Strategic Approach to Energy Management,* Proceedings of the 2003 World Energy Engineering Congress; Association of Energy Engineers: Atlanta, GA, 2003.

31

Energy: Solid Waste Advanced Thermal Technology

Alex E.S. Green
and Andrew R.
Zimmerman

Solid Waste, Solid Fuels, and Their Properties

In 2011 the United States was heavily (~50%) reliant on foreign sources for its liquid fuels and somewhat (~10%) dependent upon imports for its gaseous fuels. Our country is now expending “blood and treasure” in its efforts to stabilize regions of the globe that supply these premium fuels. Yet the United States is well endowed with solid fuels in the form of coal, oil shale, and substantial quantities of renewable but wasted solids. As part of a continuing long search for alternatives to oil,[1–10] this entry is focused on converting our solid waste to energy by advanced thermal technologies (SWEATTs) while mitigating environmental and economic problems. Table 1 is a list of United States’ abundant supply of solid waste (SW) whose organic matter can be converted into gaseous and liquid fuels as well as charcoal. The value society places on a specific fuel or energy type is very sensitive to its physical form as indicated in Table 2[11] which gives prices of various forms of energy in the United States at the beginning of 2010. The large carbon dioxide neutral (neither net producing nor consuming CO_2) plant matter components in Table 1 can help in greenhouse heating mitigation. The great diversity of physical and chemical characteristics of fuel wastes (feedstock) in Table 1 implies that the world now needs “omnivorous feedstock converters” (OFCs) to change these solid fuels into much more usable liquid or gaseous fuels or better solid fuels. Figure 1 is a conceptual illustration of an OFC adapted from a number of prior papers[8–10] in which a SW pyrolyzer–gasifier–liquifier–carbonizer is coutilized with a natural gas-fired combined cycle (NGCC) system, as will be discussed below.

Table 3 shows major ranks of coals as well as of peat, wood, and cellulose and their ultimate and proximate analyses as measured by industry for over a century. The numbers listed in columns labeled C, H,

TABLE 1 Potential Sources of Useful Non-Conventional Fuels

Waste Type	MDTa
1. Agricultural residues	1000
2. Forest under-story and forestry residues	400
3. Hurricane debris	40
4. Construction and deconstruction debris	20
5. Refuse-derived fuels	10
6. Urban yard waste	20
7. Food-serving and food-processing waste	80
8. Used newspaper and paper towels	30
9. Used tires	60
10. Energy crops on underutilized lands	50
11. Ethanol production waste	20
12. Anaerobic digestion waste	10
13. Bio-oil production waste	10
14. Waste plastics	40
15. Infested trees (beetles, canker, spores)	20
16. Invasive species (cogon-grass, melaleuca, cat-tail)	50
17. Plastics mined when restoring landfills	30
18. Biosolids (dried sewage sludge)b	40
19. Poultry and pig farm wasteb	20
20. Water plant remediators (algae, hydrilla)b	10
21. Muck pumped to shoreb	10
22. Manure from cattle feed lots	10
23. Plants for phyto-remediation of toxic sites	10
24. Treated wood past its useful life	10
Total	2000

[a] MDT = million dry tonne.

[b] Denote water remediation-related items.

and O (wt% of carbon, hydrogen, and oxygen, respectively) essentially apply to ideal carbon, hydrogen, and oxygen (CHO) materials by correcting measurements to their dry-, ash-, sulfur-, and nitrogen-free (DASNF) form. Then $[C]+[H]+[O] = 100$ and any two of the three variables fixes the third. In this work we mostly focus on the variables [O] and [H] which then essentially specifies [C]. The column labeled higher heating values (HHV) gives typical HHV in millions of joules per kilogram (MJ/kg) as measured with standard bomb calorimeters after allowing for the minor components.

Figure 2a is mainly a plot of [H] (solid diamonds with values read on the left scale) vs. [O] on the top scale for 185 representative DANSF CHO materials taken from ultimate analysis data available in the technical literature. The trend can be represented by $[H] = 6\ (1 - \exp([O]/2))$. The bottom scales give conventional coal ranks, some potential names for the biomass region, and some names that might foster more friendly discussions between the coal and biomass sectors. This [H] vs. [O] coalification plot shows that apart from the anthracite region, all natural DANSF feed-stock have [H] values that are close to 6%. The [H] and [O] coordinates of the three main components of all plant matter are lignin-6.1, 32.6; cellulose-6.2, 49.4; and hemi-cellulose-6.7, 53.3. Materials present in SW can depart substantially above and below the coalification curve. For example, the coordinates of polyethylene and polypropylene are 14.2 and 0, respectively.

TABLE 2 Market and Energy Prices, December 2010[11]

Fossil Fuels	Market Price	$/MMBtu
Crude oil	$84.93 $/Barrel	$14.64
Gasoline	$2.865 $/Gallon	$22.92
Diesel Fuel	$3.116 $/Gallon	$24.21
Natural gas	$3.56 $/MMBtu	$3.56
Liquid Propane (Gulf)	$1.11 $/Gallon	$12.19
Heating oil	$2.084 $/Gallon	$15.10
Electricity retail, resid.	12.02 c/kWh	$35.23
Coal	$47.25 $/ton	$2.00
Liquid Fuels		
Ethanol (Iowa)	$2.42 $/Gallon	$31.78
biodiesel (Iowa)	$4.13 $/Gallon	$34.96
Soybean oil (Central IL)	51.68 c/Lb	$30.40
No 2, Yellow grease	$32.88 $/cwt	$21.35
Solid Fuels		
Fuel pellets	$206.60 $/Ton	$12.91
Shelled corn	$5.37 $/Bushel	$11.76
Compost	$25.00 $/cu. yard	$3.63
Wheat straw	$80.00 $/Ton	$5.41
Grass hay (lg md bale)	$50.00 $/Ton	$3.33
DDGS	$156.00 $/Ton	$8.30

Source: Adapted from Jenner.[11]

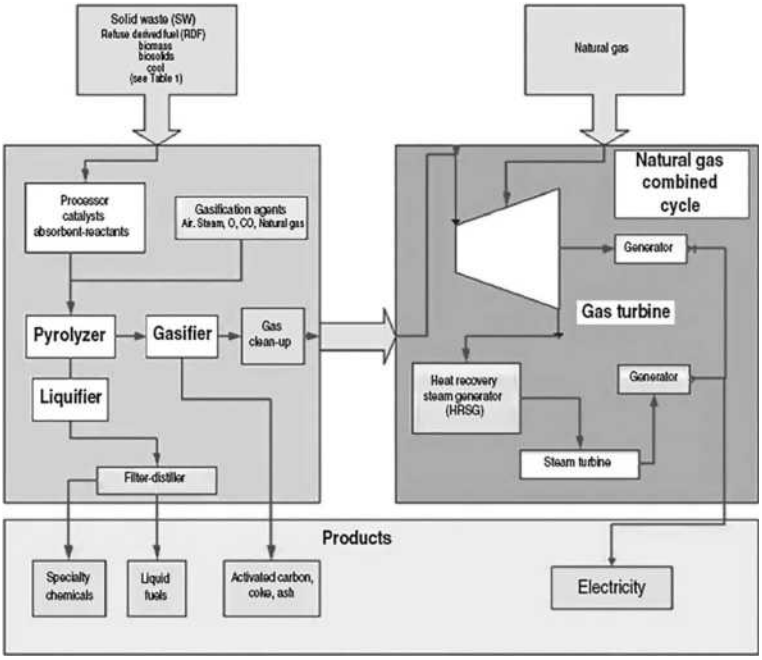


FIGURE 1 Diagram of the Omnivorous Feedstock Converter (OFC) illustrating the addition of a solid waste system to an existing NGCC plant to create an effective SWCC system.
Source: Pyrolysis in Waste to Energy Conversion (WEC).[10]

TABLE 3 Properties of Fuels along Nature's Coalification Path

	Ultimate Analysis			Proximate Analysis			Other Properties		
Name	C	H	O	HHV	VT	FC	Dens	E/vol	charR
Anthracite	94	3	3	36	7	93	1.6	58	1.5
Bituminous	85	5	10	35	33	67	1.4	49	5
Sub-Bituminous	75	5	20	30	51	49	1.2	36	16
Lignite	70	5	25	27	58	42	1	27	50
Peat	60	6	34	23	69	31	0.8	18	150
Wood	49	7	44	18	81	19	0.6	11	500
Cellulose	44	6	50	10	88	12	0.4	9	1600

C, H, and O are wt% of carbon, hydrogen, and oxygen, HHV = higher heating value (millions of joules per kilogram: MJ/kg), V_T = weight percent volatiles, FC = fixed carbon weight percent, Dens = g cm^{-3}, E/vol = relative energy density, charR = relative char reactivity.

The [C] vs. [O] data calculated with [C] = 100−[O]−[H] for DASNF feedstock are also shown in Figure 2A. When the smooth [H] vs. [O] formula is used one gets the smooth upper curve in relation to the data. This figure provides strong reasons for regarding peat and biomass simply as lower rank coals. The diagram suggests that coalification is a natural geophysical deoxygenating process. Much of this treatise on SWEATT will be devoted to attempting to bring some order to the confused literature on artificial pyrolysis, deoxygenating or carbonizing processes and their gaseous, liquid, and solid products. For many purposes, natural solid fuels could be ranked simply by [O] to replace the different ranking systems of various countries (a Tower of Babel!). For example, using 34-O for peat, called "turf' in Ireland, might help temper the "turfwars" in fuel sector competitions and in energy vs. environmental confrontations on the use of our available fuels.

HHV of various fuels measured with calorimeters are often reported along with proximate analyses. Representative values for the various coal ranks are given in the column labeled HHV in Table 3. The column labeled V_T gives representative "total volatiles," V_T, as determined by an American Standard Test Measurement Method. A solid sample is heated (pyrolyzed) in a platinum crucible at 950°C for 7 minutes. The weight percent loss due to the escaping volatiles is designated as the total volatile yield (V_T). The balance from 100% then represents the weight percent of the fixed carbon (FC) plus ash. The ash wt% is the weight percent remaining after combustion in full atmosphere at 750°C for 6 hours.

The columns of Table 3 labeled Dens and E/vol give the physical density (in g/cc) and relative energy density of the various natural solid fuels. These are important factors in determining handling and transportation costs. The column labeled charR gives some relative measures of the reactivity of the chars that are produced by the pyrolysis of these natural feedstock.

Figure 2b displays HHV data for the compilation of 185 materials after correction to DASNF cases. Most points within this scattered HHV data can be fit within a few percent by a two variable form of Dulong's formula:

$$\text{HHV in MJ/kg} = 34.9 - 0.453[\text{O}] + 0.829[\text{H}] \tag{1a}$$

Or

$$\text{HHV in MBtu/1b} = 15.00 - 0.194[\text{O}] + 0.356[\text{H}] \tag{1b}$$

The first form is simplified from the six-variable DuLong formula found by Channiwala and Parikh[12] who fit a large body of HHV measurements of biomass and other fuels. The smooth curve in Figure 2b shows the trend of the HHV vs. [O] when [H] = 6 (1 − exp([O]/2)) is used. When measured [H] values are

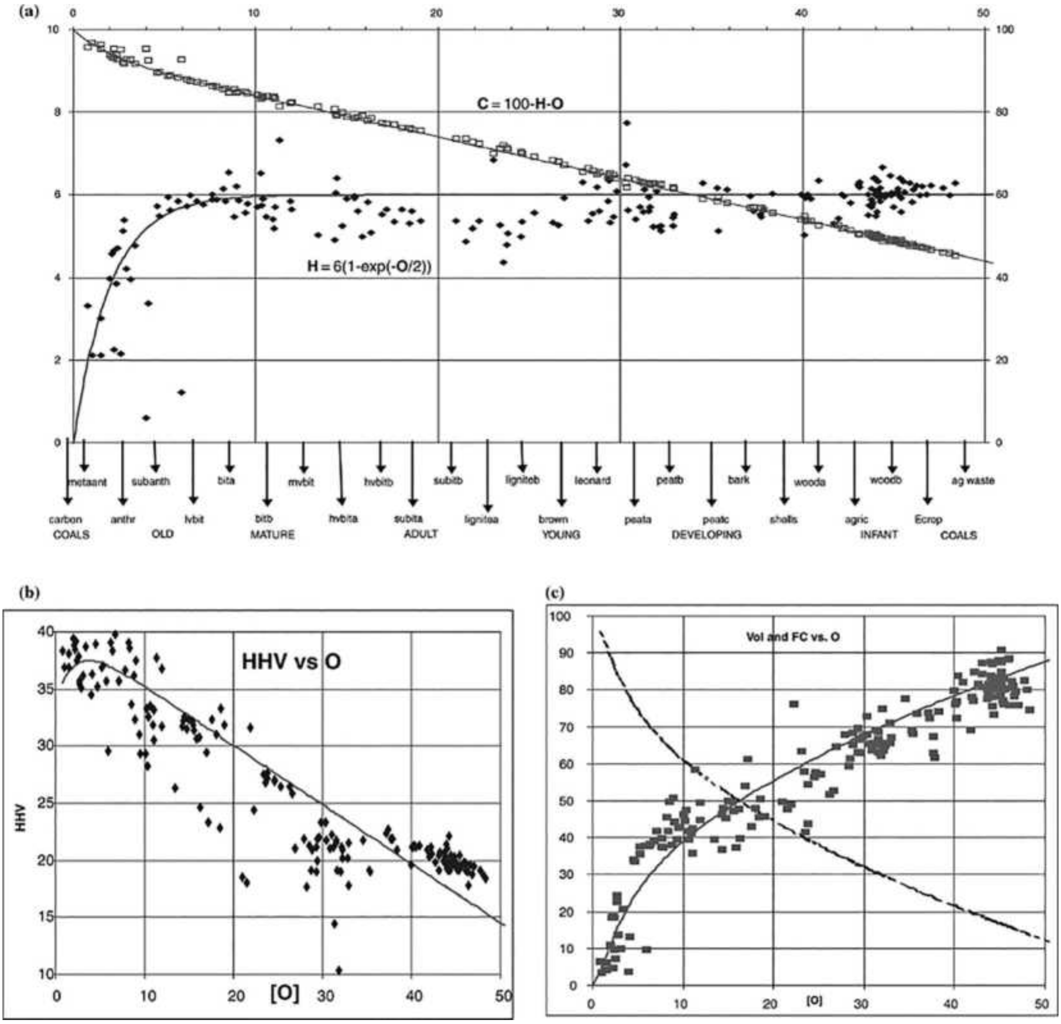


FIGURE 2 (a) Weight percentages of hydrogen [H, left] vs. [O, top] for 185 DANSF carbonaceous materials (diamonds) and additional solid waste materials (red circles). Classification labels are given at the bottom axis. (b) Higher heating values (HHV) of 185 carbonaceous materials (corrected to DANSF) vs. [O]. The smoothed curve represents: Eq. 1a when [H] = 6(1- exp([O]/2)) is used. (c) Total volatile weight percentages (left) vs. [O] for 185 DASNF carbonaceous materials (squares) from proximate analysis. The curve through the data points satisfies: $V_T = 62([H]/6)([O]/25)^{1/2}$. The analytic fixed carbon (FC) vs. [O] is shown as a dashed line.
Source: Green.[4]

used the HHV formula given fits within a few percent. From a HHV standpoint deoxygenating biomass by pyrolysis endows the char progressively with some properties of the higher ranks of coal except that chars tend to be more porous.

The general trends of total volatiles along nature's coalification curve can approximately be represented by the empirical formula $V_T = 62([H]/6)([O]/25)^{1/2}$. Note the rapidly increasing trend in V_T from low [O] materials to high [O] materials (Figure 2c). Because of the large production of volatiles by high [O] materials, pyrolysis of these materials is substantially equivalent to gasification. The [H] dimension is also important and small deviations of [H] from the smooth coalification path have a large impact on the volatile release.

The three diagrams in Figure 2 all indicate the importance of [O] in determining the fuel and carbonization or pyrolysis properties of organic materials. Coalification might be called nature's carbonization or deoxidization process whereas pyrolysis is an artificial process for carbonization or deoxidation of organic feedstock.

Global and U.S. Primary Energy Supplies

Figure 3a presents an overview of the world total primary energy supply (TPES) in 2004 (see International Energy Agency website). Among the major sources of energy, combustible renewables and waste (CRW, mostly biomass) need only be doubled to be together with coal and natural gas, and tripled to be competitive with petroleum. Note that the global CRW is currently about twice as large as nuclear. On the other hand, wind and solar must grow by factors of over 100 to become major global energy supplies.

This global TPES picture is not representative of the industrial world, particularly the United States today. Figure 3b shows the percentage subdivisions of the US TPES in 2007 when the total consumption

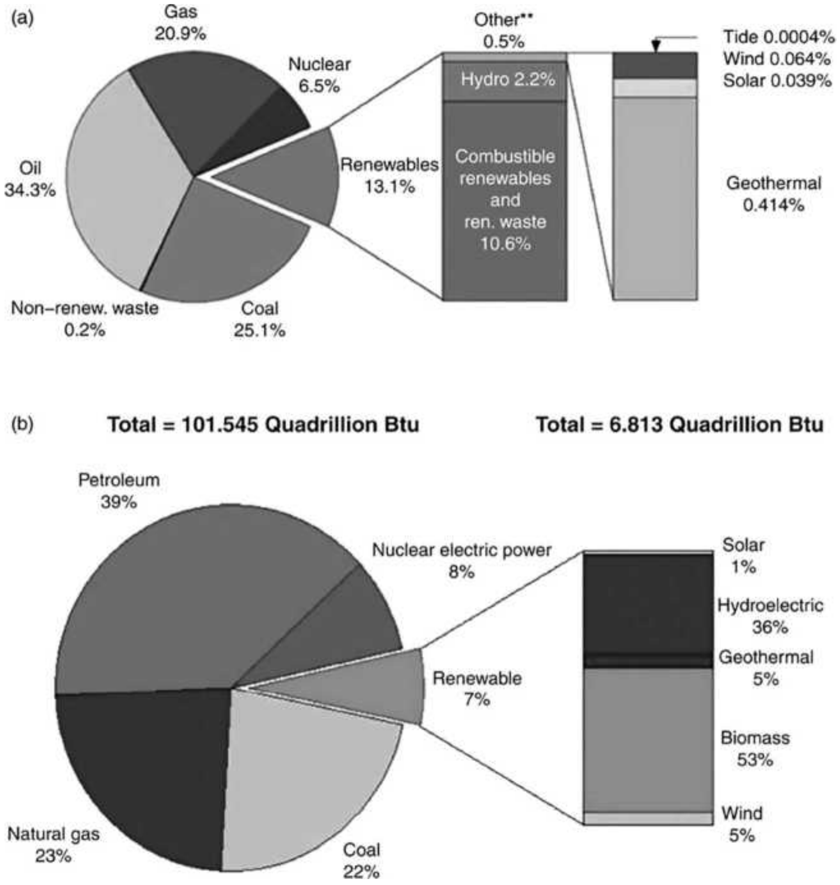


FIGURE 3 (a) Total primary energy supply (TPES) for the globe at 2004 (IEA Website). (b) TPES for the United States in 2007 (EIAWebsite). *TPES is calculated using the IEA conventions (physical energy content methodology). It includes international marine bunkers and excludes electricity/heat trade. The figures include both commercial and non-commercial energy. **Geothermal, solar, wind, tide/ wave/ocean. Totals in graph might not add up due to rounding.

Source: *IEA Energy* Statistics.

was over 101.5 quads, (quadrillion British thermal units (Btu)/annum or quads). It is seen that about 39% of our energy consumption is in the form of petroleum that is mainly consumed in our transportation sector. As Figure 3c illustrates renewables only constitute 7% of the U.S. TPES. The percentage subdivisions of these in 2007 are shown in the figure, and it is seen that biomass was 53% of the 7%. The major thrust of this work is that the solid wastes listed in Table 1, consisting mostly of biomass, were only a minor component (~3.5%) of the U.S. annual TPES, and could in the near term become a major component comparable to coal and natural gas, both still at about 23%. The more popular renewables, geothermal wind and solar, have much further to go than solid waste before becoming a major primary energy source in the United States. Since SWEATT is based on locally available solid waste, it would also create good non-exportable local industries and jobs while mitigating serious U.S. energy import and waste disposal problems. An Oak Ridge National Laboratory study[13] estimates the sustainable supply of the first few biomass categories in Table 1 at about 1.4 billion dry tons. The remaining categories should readily bring the total sustainable U.S. solid waste available to over 2 billion dry tons. Assuming a conservative HHV of 7500 Btu/lb, a simple calculation shows that with SWEATT technologies similar to those that are now in place in Japan, U.S. solid waste contribution to its primary energy supply could reach the 25% level.

Without a doubt, the biggest energy problem faced by the United States today is the need to find alternatives to oil.[1–3] In the 1970s and early 1980s, the United States focused heavily on alternatives to oil in the utility sector. The alternatives first were pulverized coal plants and in the late 1980s and 1990s, on NGCC systems. The U.S. and the globe have been slow in generating electricity by Solid Waste with Advanced Thermal Technologies (SWEATT) and flexible systems such as those illustrated in Figure 1.

It is important to differentiate secondary energy supplies (SESs) from the primary energy supplies (PESs) shown in Figure 3. Secondary energy supplies include steam, syngas, reactive chemicals, hydrogen, charges in batteries, fuel cells, and other energy sources that draw their energy from PESs. If a SES is converted to another type of energy, say mechanical energy via a steam turbine, the mechanical energy becomes a tertiary energy supply (TES). This TES can be converted to electrical energy using magnetic generators, in which case the electricity is a quaternary (QES) supply. In the case of electricity, the many conversions are usually justified since electricity can readily be distributed by wire and has so many uses as a source of energy for highly efficient electric motors, illumination systems, home appliances, computers, etc. Table 2 points to the high cost/value placed on electricity and on liquid fuels.

In many communities debates are underway as to whether increasing electricity needs should be met with the solid fuel coal, MSW, biomass via conventional steam and steam turbine generator systems, or via conversion to a gaseous fuel and using integrated gasifier combined cycle (IGCC) systems. Granting that the steam turbine route has had many advances over the last century, converting the solid fuel to gaseous fuel is increasingly being accepted as the ATT route of the future. The ATT route is not only driven by environmentally acceptable waste disposal needs and increased needs for electricity, but also by the need for liquid and gaseous fuels. A number of petroleum resource experts have recently advanced the date that the globe's supply of oil and natural gas will run out. The prices of oil and to a lesser extent natural gas now reflect this drawdown and are already high enough that conversion of organic matter in solid waste to liquid and gaseous fuels makes economic sense. We should recognize, however, that for the most part, cartels govern fuel prices not free markets. Thus we should not abandon alternative fuels efforts whenever cartels, for their interests, lower prices.

Advanced Thermal Technologies

The largest solid waste to energy systems in operation today are direct combustion municipal solid waste (MSW) incinerators with capacities in the range of 1000–3000 tons SW per day. In such mass burn systems, the organic constituents of the solid waste are combusted into the gaseous products CO_2 and H_2O. These have no fuel value but can be carriers of the heat of combustion as in coal and biomass boiler-furnace systems. Along with the flame radiation, these gases may be used to transfer heat to pressurized water to produced pressurized steam that drives a steam turbine-driven electric generator. The steam can also serve as a valuable SES to distribute heat for heating buildings, industrial processes, etc. The production and use of steam, along with the steam engine, launched the industrial age.

Instead of using the heat released to raise steam, in SWEATT systems the solid waste is first converted into gaseous or liquid fuels and, in pyrolysis systems, partly to char. The volatiles, gases, and vaporized liquids fuel then serve as a SES that can be used in efficient internal combustion engines (ICEs), combustion turbines or, in the future, in fuel cells, none of which can directly use solid fuels. Over the past century automotive and aircraft developments have pushed ICEs and gas turbines (GT) to very high levels of efficiency. Furthermore, with the use of modern high temperature GTs in NGCC systems, the heat of the exhaust gases can be used with a heat recovery steam generator (HRSG) to drive a steam turbine. Alternatively, the HRSG can provide steam for heating buildings or industrial applications of steam. These combined heat and power (CHP) system at this time make the most efficient use of the original solid fuel energy.

If one considers the United States' heavy dependence on foreign sources of liquid and gaseous fuels, the most challenging technical problem facing the United States today should be recognized as the development and implementation of efficient ways of converting our abundant domestic solid fuels into more useful liquid and gaseous fuels. In view of the diversity of feedstock represented in agricultural, municipal, or institutional solid waste, the United States and the world need an omnivorous feedstock converter such as is illustrated in Figure 1. Here, the right block represents a typical NGCC system, whereas the left block represents a conceptual Omnivorous Conversion System that can convert any organic material into a gaseous or liquid fuel.

We will first consider the gross nature of the output gas from biomass or cellulosic type material, the major organic components of most solid waste streams. Apart from minor constituents such as sulfur and nitrogen, the cellulosic feed types are complex combinations of carbon, hydrogen, and oxygen such as ($C_6H_{10}O_5$) that might serve as the representative cellulosic monomer.

ATT systems used to produce output can be divided into 1) air blown partial combustion (ABPC) gasifiers; 2) oxygen blown partial combustion (OBPC) gasifiers; and 3) pyrolysis (PYRO) systems. The three types of systems for converting waste into a gaseous fuel have many separate technical forms depending upon the detailed arrangements for applying heat to the incoming feed and the source of heat used to change the solid into a gas or liquid fuel. We use "producer gas" as a generic name for gases developed by partial combustion of the feedstock with air as in many traditional ABPC gasifiers that go back to Clayton's coal gasifier of 1694. We will use "syngas" for gases developed by partial combustion of the feedstock with oxygen as in OBPC gasifiers, which are mainly a development of the twentieth century. We will use "pyrogas" for gases produced by oxygen-free heating of the feedstock such as in indirectly heated (pyro)gasifiers. The objective is to replace the natural gas, that is, fossil fuel gas, that has a HHV, ~1000 Btu/cft = 1 MBtu/cft (with Btu $M = 1000$) with a biomass-generated fuel gas having similar energy and combustion qualities.

When an ABPC gasifier is used with cellulosic materials (cardboard, paper, wood chips, bagasse, etc.), the HHV of biomass producer gas is very low, 100–200 Btu/cft. Essentially, the useful product of partial combustion of biomass is CO that only has a HHV of 322 Btu/cft. Unfortunately, considerable CO_2 and H_2O are produced during partial combustion and together with the air-nitrogen these inerts substantially dilute the output gas. The "syngas" obtained from biomass with OBPC gasifiers is better, ~320 Btu/cft, since it is not diluted by the atmospheric nitrogen. However, because of the partial combustion it is still somewhat lower than the energy contained within the feedstock molecules. Additionally, the oxygen separator is a major capital cost component of an OBPC gasifier. With a Pyro system, the original cellulosic polymer is broken by the applied indirect heat to its monomers and then to the major pyro-products CO, CO_2, and H_2O as well as hundreds of hydrocarbons (HCs) and carbohydrates (HCOs), each with yields that depend upon the applied temperature, heating time, and particular processing arrangement. Cellulosic pyrogas can have heating values in excess of 400 Btu/cft.

Among the pyro-volatiles coming from pyrolysis systems are the paraffins (CH_4, C_2H_6, C_3H_8, …), olefins (C_2H_4, C_3H_6, …), acetylenes (C_2H_2, C_3H_3, …), and various carbohydrates, carbonyls, alcohols, ethers, aldehydes and phenols, and other oxygen-containing gaseous products. Attempting to find some patterns or regularities in the literature on products of pyrolysis from various natural and man-made fuels has been the goal of multiyear effort.[4–26] Table 4 is a list of the families of molecules that have been detected in pyrolysis volatiles and the rules that connect the family member, labeled by $j = 1, 2, 3, \ldots$ (see "The ASEM and Organization of Pyrolysis Products").

HC plastics such as polyethylene and polyolefins are heavily represented in many solid waste streams. Thus one might use (C_2H_4) as representative of the monomers in the plastic component of MSW or refuse-derived fuels (RDFs). Polyethylene pyrolysis products are dominated by C2–C4 olefins, acetylenes, and other HCs and at higher temperatures by H_2 as well as aromatics (Ar) and polynuclear aromatics (PNAs) identified in Table 4. On a per unit weight basis, all but H_2 have gross heating values in the range 18–23 MBtu/lb, similar to oil, whereas H_2 has a gross heating value of 61 MBtu/lb. On a per unit volume basis, polyethylene pyrolysis products have gross heating values ranging from 1 to 5 MBtu/cft whereas H_2 is 0.325 MBtu/ cft = 325 Btu/cft. Because natural gas is typically about MBtu/cft, we would expect the pyrogas from polyethylene to have a gross heating value comparable or greater than that of natural gas and much greater than cellulosic pyrogas.

In summary since cellulosic feedstock is already oxygenated compared with pure HC plastics, its pyrogas, syngas, and producer gas will all have considerably lower heating values than the corresponding gases from HC feedstock. From the viewpoint of maximizing the HHV of SW derived gas, Pyro gasification scores better than OBPC gasification, both of which score much better than ABPC gasification. Pyrolysis also leaves more of a solid residue in char-ash form than ABPC gasification or OBPC gasification. Figure 4 illustrates a typical pattern of evolution of the solid in an indirectly heated slow pyrolysis system.[27] For DASNF materials the asymptote of the solid (char) curve would represent the FC and the balance from 100% would represent the total volatiles. Figure 4 also shows a typical pattern of evolution of the tar and gas from wood feedstock in the pyrolysis of a small particle of wood when its temperature is raised at a slow rate such as 10°C/min. These curves are representative of results from the analytical semiempirical model (ASEM).

TABLE 4 Organization of Functional Groups by Families

Families	a	B	c
Paraffins	j	$2a + 2$	0
Olefins	$j+1$	$2a$	0
Acetylenes	$j+1$	$2a - 2$	0
Aromatics	$5+j$	$4+2j$	0
Polynuclear	$6+4j$	$6+2j$	0
Aldehydes	$j+1$	$2a$	1
Carbonyls	J	$2a$	1
Alcohols	J	$2a + 2$	1
Ethers	$j+1$	$2a + 2$	1
Phenols	$5+j$	$4+2j$	1
Formic acids	J	$2a$	2
Guaiacols	$6+j$	$6+2j$	2
Syringols 1	$7+j$	$8+2j$	3
Syringols 2	$8+j$	$10+2j$	4
Sugars 1	$4+j$	10	5
Sugars 2	$5+j$	$10+2j$	5

a, b, and c are the subscripts in $C_aH_bO_c$, $j = 1, 2, 3\ldots$

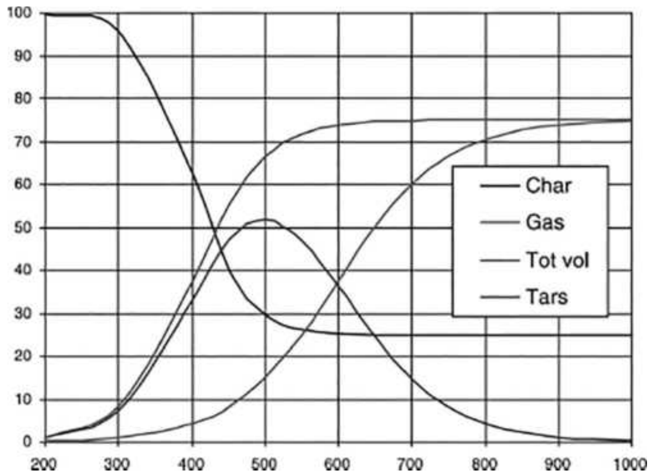


FIGURE 4 Typical pattern of evolution of char, tar, and gas from wood feedstock as temperature is raised at a slowly (10°C/min).

Studies of the evolution of chars, tars (volatiles condensable at standard temperature), gas (not condensable at standard temperature), and total volatiles have been carried out at heating rates (r) from 1°C/min to 1000°C/sec. Usually the heat rate is coupled to the temperature via a linear relationship such as $T = rt + \mathrm{T}_o$ (where t is time and T_o is an initial temperature). The dependence of these products upon temperature then changes dramatically from that shown in Figure 4 in ways that are difficult to track via standard kinetic modeling. However, they can be relatively easily represented via formulas used in the ASEM by letting the parameters of the model be simply dependent upon the heating rate. This approach is often dismissed as "just curve fitting" in academic circles where the search for models that depend upon fundamentals physical variables has become traditional. Unfortunately, in pyrolysis studies, because of the complexity of products released at various temperatures, this quest is still far from realization. Accordingly, it appears to some investigators that after the experimental assembly of reproducible scientific data (the first step of the scientific method), organizing the results in some robust analytical form (the usual second step of the scientific method) cannot only be useful for applications, but could help in achieving a fundamental model.

The ASEM and Organization of Pyrolysis Products

Proximate analyses of coal and biomass measured for over a century provide extensive data on total volatile content. However, quantitative data as to the molecular constituents in these volatiles have only been reported in recent years and a predictive method for identifying these molecules is still not available. This despite the fact that such knowledge could provide a more fundamental understanding of humankind's oldest technology (the use of fire). For control and application of a pyrolysis system it would be useful to have at least an engineering-type knowledge of the expected yields of the main products from various feedstock subjected to oxygen-free thermal treatment (pyrolysis).

In most attempts to describe the systematic of pyrolysis yields of organic materials such as coal and biomass, including the initial CCTL studies,[14–19] it has been customary to characterize the feedstock by its atomic ratios $y = \mathrm{H/C}$ and $x = \mathrm{O/C}$. In recent studies,[20–26] it has been found more advantageous to work with the weight percentages [C], [H], and [O] of the feedstock after correcting to DASNF conditions (i.e., pure CHO materials). Focusing on weight percentages appears to facilitate easier connections between the great complexity of compounds that evolve from pyrolysis and the gas, liquid (tar), and solid products of pyrolysis.

The ASEM is a phenomenological attempt to find some underlying order in the pyrolysis yields of any product $C_aH_bO_c$ vs. the [O] and [H] of the DASNF feedstock and the temperature (T) and time (t) of exposure. The ASEM was developed so as to be useful for a number of applications of pyrolysis.[18–26] Some progress has been made in including the time dimension but much more work remains on that front. When the time dimension is not an important factor, as in many cases of slow pyrolysis, the yield Y(T) as a function of temperature of each product for slow pyrolysis (or fast pyrolysis at a fixed time) is represented by

$$Y(T) = W\left[L(T:T_0,D)\right]^p\left[F(T:T_0,D)\right]^q \tag{2}$$

where

$$L(T:T_0,D) = \frac{1}{1+\exp\left((T-T_0)/D\right)} \tag{3}$$

and

$$F(T:T_0,D) = 1 - L(T) = \frac{1}{1+\exp\left((T-T_0)/D\right)} \tag{4}$$

Here $L(T)$ is the well-known logistic function that is often called the "learning curve." Its complement, $F(T) = 1 - L(T)$ thus might be called the "forgetting curve." For engineering applications this "curve fitting" approach provides a more robust and convenient means for organizing pyrolysis data than traditional methods that use conventional Arrhenius reaction rate formulas.[27] In the ASEM each product is assigned five parameters (W, T_0, D_0, p, q) to represent its yield vs. temperature profile. The objective has been to find how these parameters depend on the [H] and [O] of the feedstock and the *a, b, c* of the $C_aH_bO_c$ products for the data from particular types of pyrolyzers. Studies by Xu and Tomita (XT)[28] that gave data on 15 products from 17 coals at 6 temperatures have been particularly helpful in revealing trends of the parameters with [O] and [H]. In applying the ASEM to the CCTL data collection, the XT collection, and several other collections, a reasonable working formula was found for the yield of any *abc* product for any [O], [H] feedstock given by

$$Y(C_aH_bO_c) = W_{abc}z^{\alpha}h^{\beta}x^{\gamma}\ \left[L(T:T_0,D)\right]^p\left[F(T:T_0,D)\right]^q \tag{5}$$

where z = [C]/69, h = [H]/6, and x = [O]/25 and the parameters α, β, and γ. T_0, D, p, and q were found to have simple relationships to the feedstock and product defining parameters [H], [O], *a, b,* and c. The final ASEM formulas that fit the data could then be used to extrapolate or interpolate the XT results to any [H], [O] feedstock and temperature. Figure 5 gives an overview of the interpolated and extrapolated outputs $Y(T)$ for a selection of products for six representative feedstock along nature's coalification path.

Several hundreds, even thousands, of organic products of pyrolysis have been identified in the literature. Thus, to bring order from chaos will require some comprehensive organization of these products. Toward this goal, the ASEM approach groups products into the families as summarized in Table 4, which gives rules for the *a, b, cs* that connect these groups. This list can be subdivided into pure HCs, i.e., (C_aH_b), and the oxygenates (C_aH_bO, $C_aH_bO_2$, $C_aH_bO_3$, …, etc.). Isomers (groups with identical *a, b,* and *c*) can differ in detailed pyrolysis properties and hence parameters. We use $j = 1, 2, 3, \ldots$ to denote the first, second, third, etc., members of each group or the carbon number (n). In the most recent ASEM studies[20–26] of specific feedstock, pyrolysis formulas have been proposed and tested for the dependence of the W, T_0, D_0, p, and q parameters upon the carbon number of the product within each group. This makes it possible to compact a very large body of data with simple formulas and a table of parameters.

The case of polyethylene is an example of such a study. It is not shown on Figure 2a, as it is far removed from the coalification curve having the position [H] = 14.2 on the [O] = 0 axis. Without oxygen in the

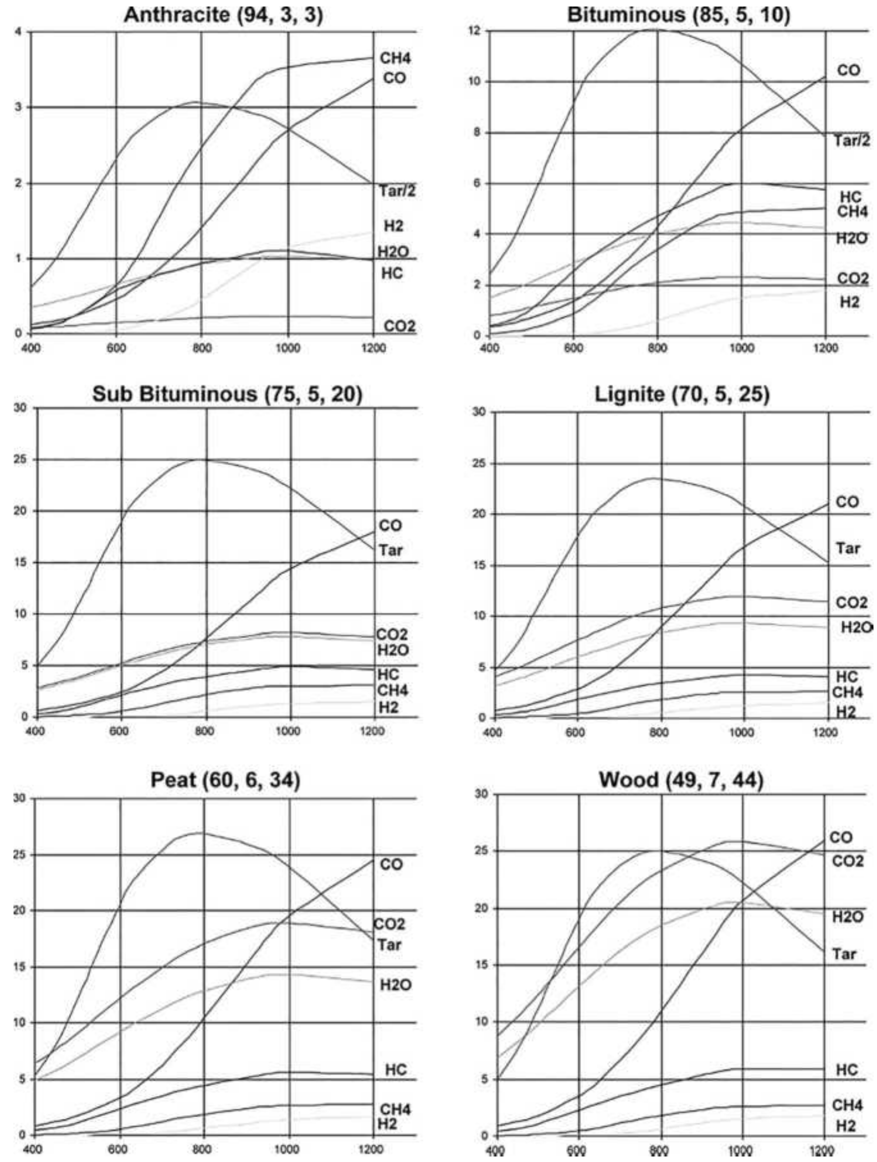


FIGURE 5 Weight percent yields (left axis) vs. temperature (in °C, bottom axis) of pyrolysis products of anthracite, bituminous coal, subbituminous coal, lignite, peat, and wood of ([C], [H], [O]) composition as shown. HC represents C2 and C3 gasses and aromatics.
Source: Green and Bell.[10]

feedstock, the pyrolysis products are much fewer and the ASEM is much simpler to use than with carbohydrates. Thus, only the first 6 rows of Table 4 are needed to cover the main functional groups involved in organizing the pyrolysis products of polyethylene. Figure 31.6 gives an ASEM-type summary of the product yields vs. temperature for polyethylene based on fits to the experimental data of Mastral et al.[29,30] at five temperatures that were constrained to approximately satisfy mass, [C], and [H] balances. Once the parameter systematic is identified, the ASEM representation can be used to estimate the pyrolysis product of polyethylene pyrolysis at any intermediate temperature or at reasonable extrapolated

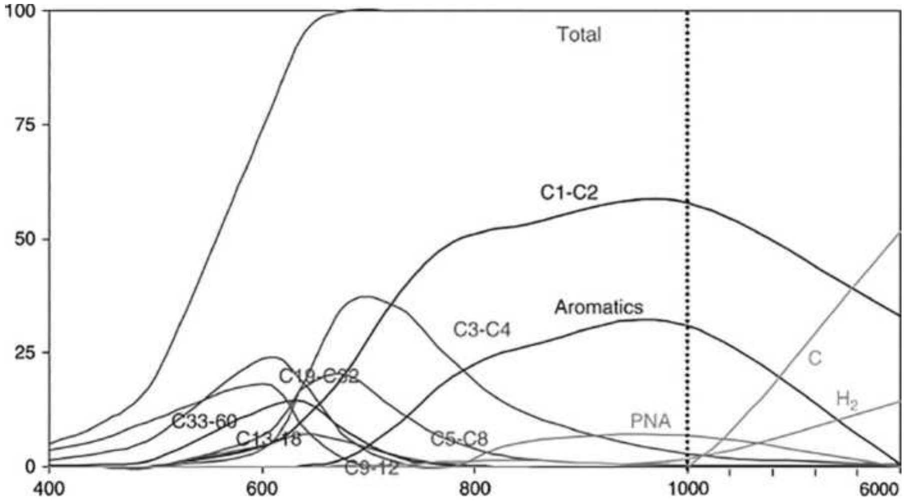


FIGURE 6 Weight percent yields (left axis) vs. temperature (in °C, bottom axis) of pyrolysis products of polyethylene in various hydrocarbon groups.
Source: Green and Sadrameli.[22]

temperatures. The experimental data was only available up to 850°C but the extrapolations to 1000°C were constrained in detail to conform to mass, [C], and [H] balances. Figure 6 also shows extrapolations to 6000°C that might be of interest if one goes to very high temperatures, for example, by plasma torch heating. Here we incorporate a conjecture that at the highest temperatures H_2 and C emerge among the products at the expense of the C1-C2 compounds as well as Ar and PNA components. While we have already found that an ASEM can begin to bring some order into pyrolysis yields, clearly there is a long way to go. When the time dimension is important, the overall search is for a reasonable function of seven variables [H], [O], a, b, c, T, and t. In comparison, Einstein's special theory of relativity only dealt with four variables x, y, z, and t.

Analytical Cost Estimation and SW-IGCC vs. NGCC

Before World War II (WWII) almost every town had its own gas works, mainly using coal as a feedstock. After WWII, cheap natural gas became available and became a major PES for home heating and cooking as well as for industrial purposes. In the 1980s, factory produced natural gas combined cycle (NGCC) systems became available and natural gas became a base load fuel source for many electric utilities. This hastened the drawdown of U.S. conventional domestic supplies and natural gas prices rose to high levels. However, natural gas prices have proven very volatile so that pursuing combinations of SWEATT with NGCC provides good flexibility in facing high energy price swings. For most biomass and plastic feedstock, pyrolysis is substantially equivalent to gasification.

Most comparative economic analyses use detailed life cycle analysis (LCA) or other forms of cost-benefit (C/B) approaches. However, it must be recognized that in recent years fuel costs, an important component of LCA, or C/B approaches have become so volatile that long-term projections based on assumed fuel cost can be grossly inaccurate. Fortunately, the economic feasibility of using a gasifier in front of a gas-fired system can be examined with simple arithmetic and algebra using an analytical cost estimation (ACE) method.[8–10] ACE takes advantage of the almost linear relationship between the cost of electricity (COE = Y) vs. cost of fuel (COF = X) observed in utility practice and in many LCA for many technologies, i.e.,

$$Y(X) = K + SX \tag{6}$$

In Eq. 6, Y is in ct/kWh (cents per kilowatt hour), X is in \$/MMBtu, and S is the slope of the $Y(X)$ line in ct/kWh/\$/ MMBtu or 10,000 Btu/kWh. S relates to the net plant heat rate (NPHR, see Chapter 37 in Stultz and Kitto[31]) via

$$S = \frac{\text{NPHR}}{10{,}000} \text{ or efficiency via } \mathrm{S} = \frac{34.12}{\text{Eff}} \tag{7}$$

A slope $S_{ng} = 0.7$ is now a reasonable assignment for a NGCC system reflecting the high efficiency of recent gas and steam turbines.

In Eq. 6 the parameter K is obviously the COE if the fuel comes to the utility without cost (i.e., $X = 0$). In previous studies,[8–10] $K_{ng} = 2$ was used as a reasonable zero fuel cost parameter for a 100-MW NGCC system.[32,33] This low number reflects the low capital costs of factory-produced gas and steam turbines in NGCC systems. In addition the contribution to the intercept K from operations and maintenance costs are reasonable. The K_{sw} for a solid waste-integrated gasification-combined cycle (SW-IGCC) system is higher than K_{ng} because the capital costs and operating cost must include the gasifier and gas cleanup system. The value of S_{sw} is also higher than S_{ng} because we must first make a SES producer gas, syngas, or pyrogas which involves conversion losses. A study of the literature[33–37] suggests that $S_{sw} = 1$ is a reasonable estimated slope for a SW-IGCC system. The X_{sw} for a SW-IGCC system that would compete with a NGCC system at various X_{ng} thus must satisfy

$$K_{\mathrm{SW}} + X_{\mathrm{SW}} S_{\mathrm{SW}} = K_{ng} + X_{ng} S_{ng} \tag{8}$$

It follows that the solid waste fuel cost X_{sw} that would enable a SWCC system to deliver electricity at the same cost as a NGCC system paying X_{ng} is given by

$$X_{\mathrm{SW}} = \frac{K_{ng} - K_{\mathrm{SW}}}{S_{\mathrm{SW}}} + X_{ng}\left(\frac{S_{ng}}{S_{\mathrm{SW}}}\right) \tag{9}$$

In the following, all X numbers are in \$/MMBtu and all Y and K numbers are in ct/kWh. Let us use Eq. 9 with $K_{kg} = 2$, $S_{ng} = 0.7$, $S_{sw} = 1$, and $K_{sw} = 4$ as reasonable estimates based on several SWCC analyses.[6–10] Then the first term in Eq. 9 is –2. Now when the $X_{ng} = 2$ to generate SWCC electricity at the same cost the solid waste provider must deliver the fuel at a negative price, i.e., pay the tipping fee of –0.7. However, if X_{ng} is near 6 as it has been several times between 2004 and 2010 the SWCC utility could pay up to 2.4 for the SW fuel. If the X_{ng} is at 12, the SWCC facility could pay 7.1 to the SW supplier. This X_{sw} price is much higher than that of coal whose delivered price (X_c) these days usually is in the 2 range. It is also much higher than a pulp and paper mill would pay for waste wood. This simple cost comparison is illustrated in Figure 7 that shows the opportunities for SWCC systems when natural gas prices are above say \$5/MMBtu. The results are slightly less favorable if the K_{sw} were higher say at $K = 5$. However, the conclusions that at high natural gas prices SWCC electricity becomes competitive with NGCC electricity would be similar. It is conceivable that Ksw could be held as low as 2 ct/kWh by retrofitting a NGCC system stranded by high natural gas prices. In this case the first term in Eq. 9 vanishes and the competitive $K_{sw} = (S_{ng}/S_{sw})X_{ng}$. This illustrates the main point that at high natural gas prices, with an ATT system, SW can be a valuable PES. Indeed, this simple algebraic–arithmetic exercise establishes the feasibility of a new paradigm in which solid waste (mostly biomass) can become potentially valuable marketable assets.

As described above, the values of K and S are the key factors in determining the COF_{sw} that can be used in a SW-integrated pyrolysis combine cycle (IPCC) to have the COE equal or less than the COE with a NGCC system at the available COF_{ng}. The ACE method can be extended to the use of SW or biomass with other technologies if we can identify the K and S for each technology. Where actual facilities

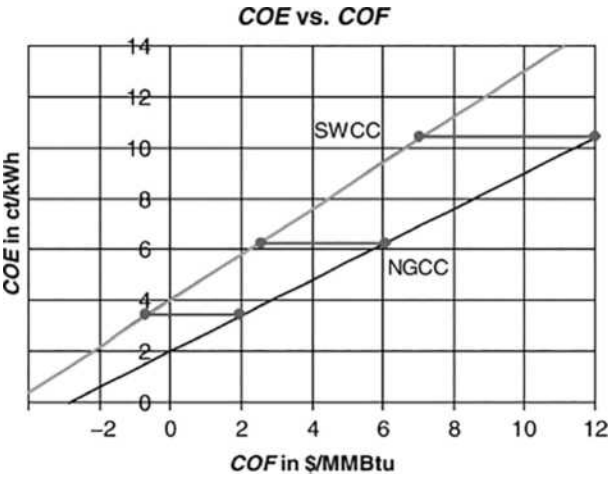


FIGURE 7 Cost of electricity (COE) vs. the cost of fuel (COF) for an solid waste combined cycle (SWCC) system and for a natural gas combined cycle (NGCC) system. COF comparisons of 2, 6, 12 dollar/MMBtu are indicated in red lines.
Source: Green and Bell.[10]

have been built and placed in operation, the K_s and S_s can be assigned on the basis of actual plant experience which is the case for many fossil fuel technologies. However, for new technologies proposed for the renewable age, some accepted form of C/B analysis is needed to provide COE vs. COF relationships. The ACE method cannot serve in this role. However, if a detailed C/B analysis or LCA is available at a particular power level, its *Y(X)* results can generally be cast into the Eq. 6 form which is generally more useful and transparent than tables of numbers. A major advantage of doing so is the possibility of then making reasonable *Y(X,P)* extrapolation to other power levels *(P)* on the basis of many years of economy of scale experience with a wide range of technologies.

The ACE method has been applied to reformulations of a large body of COE vs. COF calculations on biomass use presented in an Antares Group Inc. report (AGIR)[34] for a number of technologies. The technologies investigated in the AGIR were for systems in which 100 tons per day forest thinning were available in wild land–urban interface areas. This assumption limited the power level (*P*) for that technology quite severely. The technologies in the AGIR included a biomass-integrated gasifier combined cycle (B-IGCC) system, a B-IG simple cycle (B-IGSC) system, a B-IG internal combustion (B-IGIC) system, a biomass–gasification–coal co-firing B-IGCC system, a direct cofiring of biomass and coal in a coal steam boiler BCoSt, a direct use of biomass in a feed water heat recovery arrangement (FWHR), direct use of biomass in a Stoker fire boiler steam turbine (SFST) system, and direct firing in a CHP plant (CHP) with a steam market at $6/MMBtu. For each technology, it was possible to approximately represent the tabulated COE vs. COF results of their detailed economic analysis by Eq. 6 and to evaluate *K* and *S* for that technology at that power level.

The most interesting result of this ACE digest of the massive tables of the AGIR was that by slight extrapolations to higher power levels[9] the competitive results in several important cases were opposite to those for the power levels limited by the 100 tons biomass per day assumption.

Several other detailed economic COE vs. COF analyses have been used to refine ACE and generalize the ACE methodology. In particular, *K* has been broken into components $K = K_c + K_{\mathrm{om}} + K_{ne}$, where *c* stands for capital costs, om for operating and maintenance costs, and *en* for environmental costs. Establishing the magnitudes of these components for various technologies and power levels is still at the cutting edge of utility economic analyses, and there are large disagreements particularly on K_{en}. In one generalized component form of ACE (CACE) Eq. 6 is replaced by

$$Y(X,P)=K_{cr}\left(\frac{P_r}{P}\right)^{\alpha}+K_{omr}\left(\frac{P_r}{P}\right)^{\beta}+K_{er}\left(\frac{P_r}{P}\right)^{\gamma}+XS_r\left(\frac{P_r}{P}\right)^{\delta} \tag{10}$$

where K_{cr}, K_{omr}, K_{er}, and S_r are established on the basis of a detailed analysis at a reference power level P_r and α, β, γ and 5 are scaling parameters intended to reflect the tendency of per energy unit cost to go down as the power goes up (economy of scale). Table 5 lists CACE parameters extracted from a detailed analysis "Options for Meeting the Electrical Supply Needs of Gainesville" prepared by ICF Consulting.[35] Here the final COE is given in 2003 ct/kWh. The third and fourth cases NGCCc and NGCC-d have been added to better reflect the high volatility of natural gas prices that have ranged from \$1/MMBtu to \$16/MMBtu over the past 25 years.

The final column shows that at the reference power levels without the NGCCc case the IGCC scores the lowest COE as was concluded in the ICF report (ICFR). The value of the ACE analysis is that with a bit of algebra anyone can easily consider other fuel cost projections and other power levels (with the assigned values of α, β, γ and δ). Based on prior exploratory work and economy of scale investigations it was estimated that for costly field erected facilities $\alpha = \beta = 0.3$ are reasonable choices. However, with factory fabrication of gas and steam turbines these parameters might not follow the usual economy of scale pattern and that their α may be somewhat smaller in magnitude. Assigning a value for γ is a wide open question since environmental costs and methods of incorporating them into the COE are still highly debated issues.[36] Reasonable values for δ are also somewhat difficult to find. For NGCCs the author tentatively assigns close to zero or a very small value (~0.1) perhaps because the development of highly efficient aero-derivative turbines has proceeded on a wide range of power levels.

A Somewhat simper generalized formula for $Y(X,P)$ has been developed in the form [37]

$$Y(X,P)=\alpha C_{cr}\left\{\left(\frac{P_r}{P}\right)^{\gamma}\right\}\left(1+f_{om}\right)+SX\left(1+f_e\right) \tag{11}$$

Here, $\alpha C_{cr} = K_{cr}$ at a reference power level where C_{cr} is the specific capital cost for that facility in dollar per watt (\$/W or \$1000/kW) for that technology at the reference power level. Based on tabular data contained in the RA study a value $\alpha = 1.34$ was identified as the coefficient the approximately relates specific capital costs (in \$/W) to K_{cr} (in ct/kWh). This essentially is the COE when the fuel cost is free and OM can be ignored. As for economies of scale, based on an extensive literature survey. Green et al.[37] fond

TABLE 5 Analytical Cost Estimation (ACE) Results from ICFR for Five Technologies—Four Natural Gas Prices Are Assumed for the NGCC Technology

Technology	P_r	K_o	K_{om}	K_{en}	S_o	COF	COE
NGCC-a	220	0.598	0.234	−0.17	0.68	11.34	8.37
NGCC-b	220	0.598	0.234	−0.17	0.68	6.1	4.81
NGCC-c	220	0.598	0.234	−0.17	0.68	5	4.06
NGCC-d	220	0.598	0.234	−0.17	0.68	4	3.38
SCPC	800	1.491	0.299	0.714	0.93	1.91	4.28
CFB-CB	220	2.531	0.261	0.618	1.05	1.41	4.89
CFB-B	75	2.845	0.261	0.039	1.39	1.67	5.47
IGCC	220	2.2	0.196	0.407	0.86	1.41	4.02

COE = cost of electricity, COF = cost of fuel.

the different technologies had scaling parameters in the range $0.13 < \gamma < 0.33$. When no information is available one might use some intermediate gamma between these extremes or let $\gamma = 0$.

Typical dimensionless values for f_{om} can be identified for various technologies from the Antares report on 6 technologies, ICER on 5 technologies, and the RA study of 14 widely ranging technologies. In most cases they are less than unity and can be assigned within reasonable bounds on the basis of experience.

In Eq. 11 f_e is an added dimensionless "correction" to the delivered COF that reflects environmental costs not included in the cost charged by the utility but paid for by the public in other ways (reduced visibility, added coughs, higher cancer rates, etc.). The landmark RA[36] study that incorporated externalities into a levelized cost of energy analysis for 14 electric generating plants provided a basis for estimating the dimensionless externality correction f_e to the price of fuel. Table 6 translates RA's results for the 14 utilities using their minimum externality cost estimates into the generalized analytic cost estimation (GACE) analytical form of Eq. 11. As one sees these f_e are substantially larger than 1 for fossil fuel technologies but small (0.15) for a biomass system. Wind turbines, photo voltaic, and landfill gas systems do not have a fuel cost hence f_e cannot be assigned. However, these technologies are directly assigned externality C_{ex} by RA. The major uncertainty in the future COE is probably represented by the variable X, the externality parameter f_e, and possibly the power level of the facility. Having an explicit formula with these variables and parameters can help in bringing transparency to important policy decisions.

The GACE approach to reaching decisions in the face of large uncertainties might be viewed as application of the operations analysis methods used by one of the authors in WWII.[38] It will be interesting to compare this approach with recent European Union operations analysis effort for incorporating externalities in electricity-generating technology evaluations.[39]

Going back to the 2007 SWEATT study it mainly considered the competition between NG-fueled technologies and coal-steam-generated electricity. These included supercritical coal-burning units that reached efficiencies as high as 40%. However, when coal-steam turbine's expensive scrubber costs are included in the K_c and environmental costs in their K_{en} this technology did not compete compared with the IGCC. Thus a major conclusion of the 2007 SWEATT study was that the age of making gas has returned. At the same time the most favorable EEE position of natural gas among the fossil fuels is recognized.

TABLE 6 Roth-Ambs Externality Impacts on Cost of Electricity for 14 Technologies Using Low RA Estimates, Impacts, and Derived GACE Parameters

Technology	C/W	K_c	f_{om}	S	C_{of}	C_{ex}	f_e	COE	COE_e
Coal Boiler	1.80	2.81	0.36	0.995	1.06	4.45	4.20	4.86	9.29
Adv Fld Bed	2.20	3.52	0.49	0.975	1.04	2.86	2.75	6.26	9.05
IGCC (coal)	2.10	3.28	0.28	0.889	0.95	2.64	2.78	5.05	7.40
Oil Boiler	1.30	2.15	0.19	0.943	3.22	6.03	1.87	5.59	11.27
Gas Turb SC	0.70	10.1	0.12	1.15	3.47	4.62	1.33	15.28	20.59
Gas T Adv	0.40	0.82	0.51	1.09	3.29	4.45	1.35	4.83	9.68
NGCC	0.60	0.91	0.34	0.683	2.11	3.46	1.64	2.66	5.02
MSW Inc.	5.70	9.63	0.44	1.687	5.16	7.7	1.49	22.55	35.54
LFG	1.50	3.3	0.3	1.215	0	0.7		4.29	5.14
SOFC	1.60	2.42	2.71	0.758	2.29	2.75	1.20	10.71	12.79
Wind Turb	1.00	5.74	0.29	0	0	0.7		7.40	7.40
PV Utility	4.70	49.5	0.02	0	0	0.25		50.53	50.53
Hybrid solar	3.70	20.3	0.15	0.346	1.07	2.38	2.22	23.64	24.46
Biomass	2.40	3.54	0.73	1.431	2.75	0.41	0.15	10.07	10.65

COE_e are Roth-Amb's total *COE* with low externalities.

It should be noted that in an effort to minimize a major environmental externality, global warming, the U.S. Department of Energy is now investing a $4 billion plus effort in carbon capture and storage (CCS) technologies that first starts with oxygen-blown coal gasification. DOE funds will be matched by about $7 billion from the coal and utility industry. This effort could bring up to ten commercial demonstration projects online by 2016. The goal of the program is to provide the information needed to evaluate whether such CCS technologies are commercially deployable.

Mercury emission control has recently become mandatory for coal fired plants. Injecting activated carbon as a sorbent to capture flue gas mercury has shown the most promise as a near-term mercury control technology. The process is still in its early stages and its effectiveness under varied conditions (e.g., fuel properties, flue gas temperatures, and trace-gas constituents) is still being investigated.

Bioenergy and Biochar

During the first decade of the 21st century, the most widely pursued sources of renewable biofuels were fermentation of corn or corn stover leading to ethanol, anaerobic digestion of animal waste yielding methane, and compression of plant seeds to extract bio-oil. Since this Solid Waste to Energy by Advanced Thermal Technologies (SWEATT) work is focused on advanced thermal technologies (ATTs), we refer the reader to the literature on these non-thermal conversion methods. This section is largely devoted to the economic, environmental, and energy (EEE) impacts on SWEATT that the production of "pyro-char" or "black carbon" may have. These terms can be used to include many solid pyrolysis products that might serve as charcoal, biochar, or activated carbon. Charcoal and activated carbon are well documented in the technical literature. However, the literature on biochar is just developing. Biochar applications have recently inspired an International Biochar Initiative, a community of scientists and enthusiasts that envision large-scale conversion of waste biomass into biochar while generating energy at the same time.[40,41] This biochar can be applied to soils, both enhancing soil fertility and mitigating climate change by sequestering CO_2 drawn from the atmosphere.

Charcoal was used as early as 5000 BCE in the smelting of copper, and 2000 years later, it became commonplace in the smelting of iron and bronze. Charcoal was burned to produce temperatures in excess of 1000°C that are needed to produce these alloys. Other uses include blacksmith forges and household cooking, in which maximal heat with minimal smoke is desirable, and the filtering and removal of impurities such as in the spirit or sugar processing industries. Charcoal was commonly made in covered conical piles of wood, sometimes covered with earth, constructed so as to exclude air, thus attaining greater yields of charcoal. Its large-scale production is presumed to have led to widespread deforestation of Europe and Eastern North America in the 18th century, until coal supplanted charcoal as an industrial fuel.

In an extensive review, Antal and Gronli[42] have summarized knowledge of the production and properties of charcoal that has been accumulated over the past 38,000 years. They point out that biomass carbonization can be carried out, leading to high char yields (~30%) by the manipulation of pressure, moisture content, and gas flow involved in the process. The review also provides a good summary of measurements of the heat of pyrolysis from various plant feedstock that range from +0.7 MJ/kg (exothermic) to −0.3 MJ/kg (endothermic). When viewed in the light of the fact that the higher heating values (HHV) for most dry biomass are about 17 MJ/kg and pure carbon is 32 MJ/kg, it should be clear that, in any case, the heat cost of pyrolysis is small compared to the heat content of the feedstock. However, in various practical biomass pyrolysis arrangements, component system losses need to be kept as small as possible so that acceptable conversion efficiencies from the feedstock to the desired form of energy are achieved.

Conventional combustion technologies, or even advanced combustion systems, when applied to biomass/solid waste (SW), leave little carbon in the fly or bottom ash. Thus, it might be difficult to adapt these technologies to the useful production of pyro-char products. Air-blown partial combustion and oxygen-blown partial combustion systems might be adapted since they essentially take advantage of substoichiometric combustion to produce CO fuel rather than CO_2 and, depending upon the oxygen content of the feedstock, could leave a substantial carbon residue. The solid char residues produced during forest fires are an example of natural combustion under limited oxygen conditions. The charred woods and plastics remaining after building fires are further examples of limited oxygen combustion. An extensive technical literature is available on fire and fire protection technology that could be drawn upon if carbonization again becomes a widespread technology. Pyro-char or black carbon, when produced intentionally in partial combustion gasifiers or more efficiently in indirectly heated pyrolyzers, is now being referred to as "biochar." Some uses of biochar are quite ancient, while others are quite recent.[40–47]

It is only fairly recently that another ancient use of biochar has come to the wider attention of environmental scientists. *Terra preta* are small plots (20 ha, on average) of highly fertile Amazonian soils, enriched in organic carbon and nutrients, that are surrounded by Oxisols, typical of tropical soils, that are extremely depleted in nutrients and organic matter. Because *terra preta* are associated with high concentrations of charcoal and ceramic fragments, and can be dated to have formed between 800 BCE and 500 ACE, they are presumed to be anthropogenic, made either intentionally by some method of slash and char forestry for agricultural purposes or accidentally through the dumping of kitchen fire wastes over long time periods.[41] This discovery has inspired the current "biochar movement" and interest in identifying optimum methods of producing and using biochar. It is recognized that the optimum characteristics of biochar are still uncertain and are the subject of research.

The unique properties of some biochar, particularly its high adsorption capacity, can be attributed to its high surface-specific surface area (SA) as well as its surficial functional group content. Although charcoal can be "activated," that is, altered with physical or chemical treatment or by "carbonization," heating above 800°C, to produce extremely high SAs or oxidized surfaces, even non-activated biochars can possess some of these features. Biochar SA tends to increase with pyrolysis temperature but starting biomass type and pyrolysis atmosphere and duration of heating will also play a role. For example, Zimmerman[48] has reported N_2-BET SA of less than 13 m^2/g for a variety of biomass types pyrolyzed at 400°C (3 hr), including grasses as well as softwoods and hardwoods. At 525°C, SA ranged from 31 to 501 m^2/g, and at 650°C, 220–550 m^2/g. Between 800°C and 1000°C, SA of between 400 and 1000 m^2/g are commonly recorded by Downie et al.[44] These measurements, however, include only pores larger than a few nanometers (nm) in diameter. Surface present within micropores (pores smaller than about 1.5 nm in diameter), measured using CO_2 sorptometry, have yielded SA in the range of 160 to 650 m^2/g and have been found to be more strongly related to the ability of a biochar to sorb low molecular weight organic compounds and cations.[40,50]

One can envision a number of possible ways in which the sorbent properties of biochar could be utilized. First, much as activated carbons have been used for many centuries in a wide variety of industrial process that require the adsorption of noxious, odorous, or colored substances from gases or liquids, biochar could be used as a low-cost alternative, especially in circumstances where large volumes of material are required. Much like activated carbon,[51] biochar can be powdered to increase SA or granulated for use in fixed bed filtration systems. Although somewhat lower in SA and, thus, sorption capacity compared to activated carbons, its characteristics can be tuned via production conditions, for sorption of specific components. For example, biochar made from anaerobically digested bagasse has been shown to be a superior sorbent of metal including lead.[52] Its most cost-effective industrial use is likely to be in the areas of primary or secondary water treatment or in contaminant remediation as reactive media for surficial or subsurface permeable barriers such as trenches, wall barriers, funnels and gates, or landfill bottom linings. In all these cases, both the biochar C and the adsorbed C may be sequestered from the atmosphere and, thus, may be considered an additional C sink, or at least an avoided C source. With the 2010 EPA (Environmental Protection Agency) limitations on mercury emissions from coal plants and municipal waste incinerators, one might anticipate a large market increase for mercury-adsorbing activated or nonactivated carbons.

Much as black-carbon-enriched soil such as *terra preta* in the Amazon has been prized for centuries for its ability to produce sustained enhanced crop yields, it is presumed that biochar amendments to soil, if carried out properly, can increase soil fertility in both the United States and perhaps more critically in the third world where soil depletion is reaching critical levels. Some biochars have high cation exchange capacity, lending it the ability to adsorb and retain such essential plant nutrients as nitrate, ammonium, calcium, and potassium.[43,46] Biochars have also been shown to adsorb the critical anionic nutrient phosphate, though the chemical mechanism for this is unclear. Other positive agricultural effects may include better soil moisture retention and the encouragement of unique microbial populations that may be beneficial to plant growth.[53] Thus, while not yet shown on a large-scale basis, biochar amendment may reduce a farmer's costs for fertilizer and irrigation, while reducing runoff of environmentally damaging nutrients into surrounding surface waters and groundwater (cultural eutrophication) and reliance on inorganic fertilizers made using energy from fossil fuels (another CO_2 source).

A recent life-cycle assessment (LCA) study assessed the energy and carbon impacts of four biochar-cropping systems.[47] They found that, for late and early corn stover, switch grass, and yard waste as biomass feedstock sources, the net energy generated was +4116, +3044, +4899, and +4043 MJ t^{-1} dry feedstock, respectively. Most of the energy consumed was in either agrochemicals or feedstock drying, and most of the energy yield was in syngas heat. Net greenhouse gas emissions were negative for both stover types and yard waste, with the majority of the total reductions, 62%–66%, realized from C sequestration by the biochar. For switch grass, however, land use change and field emissions were high enough to drive net emissions to positive. The main conclusion of this LCA analysis was that the energy and carbon impact of small-scale use of pyrolysis systems using yard waste is the most economically favorable at this time. However, many numbers used in these calculations were, by necessity, broad estimates. Much more research is required to improve the inputs to these types of models.

Because of biochar's environmental stability, conversion of biomass to biochar represents a long-term transfer from a C pool rapidly cycling between biomass and the atmosphere to a pool held sequestered within soils or even aquatic sediments. Conversion of 1% of all biomass to biochar each year could reduce the atmosphere CO_2 by 10% in only 14 years (assuming 50% conversion efficiency and no biochar C degradation). These figures are certain not realistic, however. First, it has been shown that biochars degrade abiotically as well as microbially at rates ranging from C half-lives of a few 100 years (for lower-temperature chars, particularly those made from grasses) to 10^5years for higher-temperature chars with additional losses to be expected from leaching. Second, the amount of biomass that could be reasonably used as feedstock without using major quantities of fossil fuels in the process of gathering and transportation, and without endangering soil stocks, habitat, or human food resource security (i.e., without land-use conversion), is likely in the range of 2.27 Pg C $^{yr-1}$.[55] Aside from C sequestration and enhanced crop growth, further reduction in greenhouse gas concentration and associated climate change may be obtained via reductions in methane (CH_4) production associated with waste land filling and suppression of nitrous oxide (N_2O) production when biochar is added to soils[43,46] and energy extraction. Using this estimate of maximum sustainable feedstock generation and accounting for all possible benefits, biochar production could potentially offset a maximum of 12% of current anthropogenic CO_2-C equivalent emissions each year.[54] Another interesting finding of this study was that the greatest environmental benefits are to be had by the biochar approach in regions of infertile soils or where water resources are scarce. However, where soils are already fertile, and particularly in regions where coal emissions can be offset, bioenergy (see next section) may be a better approach.

It is hoped that production of biochars or application of biochar to soils may soon qualify as a "carbon offset" or be traded on the open market should a "C cap and trade" policy be implemented. The biochar concept has received formal political support in the U.S. and globally. The U.S. 2008 Farm Bill established the first federal-level policy in support of biochar production and utilization programs nationally and biochar has been included in the United Nations Framework Convention on Climate Change (UN-FCCC in Dec. 2009).

Bio-Liquid Fuels

Intensive use of liquid petroleum products, particularly diesel and gasoline, by automobiles, trucks, airplanes, trains, and ships in the 20th century drew down national and global reserves to the point that energy security has become a major concern of the United States, other industrial countries, and the globe in general. During this same period, CO_2, the major product of hydrocarbon combustion, has further increased from its preindustrial level of 280 ppm to its current level of 385 ppm. Global warming is now emerging as the biggest environmental problem of the 21st century and "What to do about CO_2" is the biggest environmental question.[56–59]

Producing liquid fuels from plants could potentially mitigate both security and environmental problems in countries that have land available that is not in food production. Plants use the sun's energy, the atmosphere's CO_2, and the soil's H_2O to make carbohydrates such as cellulosic matter and lignin. When biomass, or a converted form of it, is combusted (oxidized), the CO_2 is returned to the atmosphere as a part of a short-term cycle that can be considered *carbon neutral.* On the other hand, combustion of coal and petroleum fossil fuels adds CO_2 that had been extracted by plants from ancient atmospheres to today's atmosphere.

In contrast to solid and gaseous fuels, the convenience of energy storage and transfer makes liquid fuels much more useful in the transportation sector. Because ethanol, pyrolysis oil, vegetable oil, and biodiesel are biomass-derived liquid fuels that are closer to carbon neutral, they are now under rapidly increasing consideration as replacements of or supplements to conventional diesel and gasoline. The high monetary value of liquid fuels is illustrated in Table 2. Table 7 lists typical properties of these liquid fuels.

Technologies to convert plant simple sugars to ethanol go back to the beginnings of the wine, beer, and liquor industries thousands of years ago. In recent years, genetic manipulation[60] has led to microbes that can convert cellulose to ethanol as well. We refer the reader to the extensive biochemical literature for such recent developments. As compared to ethanol, vegetable oil, biodiesel and diesel in Table 7 have about twice the energy/volume, important in transportation applications.

The properties of pyrolysis oils vary over wide ranges depending upon the feedstock, the rate of heating, the temperature reached, catalysts used if any, the speed of quenching after the polymeric bonds are broken, and other specifics of the thermal processing. Fluidized bed systems with fast heat transfer followed by rapid quenching produce bio-oil yields ranging from 50% to 75% of feedstock weight, with pyro-gas and pyro-char representing most of the remainder material. Extensive R&D efforts are now underway to upgrade pyro-oils into more energy dense, water-free, and oxygen-reduced liquid fuels.[61]

Since the 1973 oil embargo, vegetable oils such as corn, soybean, canola, rapeseed, sunflower, palm, and coconut have been given serious consideration for liquid fuels.[62] In effect, this would be a return to what Rudolf Diesel used with his first compression ignition engine (CIE). Used vegetable oil from fast-food restaurants has become a favorite inexpensive source of such feedstock. The high viscosity of most vegetable oils presents CIE problems, but these problems can be overcome by suitable preheaters. Since the supply of used vegetable oil is limited, there has been rising interest in high-yield, non-food vegetable oils, such as *Jatropha, Camelina,* flax, and algae that can be grown on marginal lands.

TABLE 7 Some Key Properties of Biomass-Derived Liquid Fuels and Diesel

Fuel	C (wt%)	H (wt%)	O (wt%)	HV (MJ/kg)	Density (kg/m^3)	E/vol (GJ/m^3)	Visc. cs
Ethanol	52.2	13.0	34.8	22.6	790	17.9	1.1
Pyro-oil	~45	~8	~47	~18	~1000	~18	var.
Vegetable oil	74.5	10.6	10.8	40.4	906	36.6	46.7
Biodiesel	79.0	12.9	8.0	41.2	920	37.9	4.7
Diesel	87	13	0	45.3	852	38.6	3.2

The trend in bio-oil production is now towards converting vegetable oil to low-viscosity biodiesel by mixing it with an alcohol and a catalyst. In this esterification process, the oil's glycerin is replaced by the alcohol, making a mono-alkyl ester, which greatly reduces the viscosity and slightly increases the HHV. Biodiesel can be used in unmodified diesel engines as a sole fuel or in mixtures with diesel. Emissions of sulfur oxides and other regulated pollutants, apart from NO_x, are generally lower from biodiesel than from conventional diesel. Since biodiesel is derived from plants that were made with solar energy, it is considered approximately carbon neutral.

Several states are initiating non-food vegetable oil programs to meet transportation needs beginning with the fuel needs of agriculture and particularly the needs of the individual farmer. In the Pacific Northwest, canola (rapeseed) has[63] and is being studied but not yet adopted on a large scale because revenues from growing canola are generally somewhat lower than that from wheat. The differential is currently being subsidized in the form of a federal blender's credit of $1.00 per gallon to jump-start this new industry.

In Texas, flax, an annual plant and prolific biomass producer, is now under consideration as a source of vegetable oil for transportation applications. Flaxseed has long been used as the source of linseed oil, which is used as a component of many wood-finishing and other industrial products and as a nutritional supplement. In a description of recent developments by a Texas A&M researcher,[64] "It's kind of like [Texas] is coming full circle. Flax was grown on about 400,000 acres in the 1950s and Texas AgriLife Research at A&M had an active flax breeding program. Those varieties were known nationwide for having good cold tolerance. That's what we needed, a flax variety that was something you could plant in the fall, survive the winter, avoid late freezes, and produce seed in the spring. Now we're evaluating this as a possible biodiesel product or (one which) could be used in the vegetable oil industry."[64]

Florida is developing a bio-oil production program utilizing *Camelina sativa* planted on non-food-producing land.[65] When *Camelina* seed is cold pressed, 30% (by weight) oil is obtainable. The resulting pressed cake called meal has oil content between 10% and 12% (by weight) that can also be extracted with organic solvents, but this process is expensive. For biodiesel production to be economically competitive, it is essential to minimize production costs and maximize all potential revenue streams. The use of the meal as animal feed is one such stream. However, if non-food vegetable oil ever reaches its full transportation potential, the seed meal and plant residue would far exceed what could be consumed by existing cattle herds. With current budget deficits, the prospects for adequate subsidies in the future are not favorable. Thus, additional services or commodities must be developed to generate additional revenues.

In the spirit of this overall study of SWEATT, the use of on-site SWEATT systems to extract additional liquid fuel, energy, biochar, and specialty chemicals from the waste generated in a farmer's overall bio-oil production cycle warrants careful consideration. The meal produced by pressing the canola, flax, or *Camelina* seed and the plant stover could be dried and continuously fed into an on-site pyrolyzer yielding thermal energy, pyro-oil, biochar, and possibly activated carbon or other valuable products. The biochar then could be used together with local biosolids from wastewater sewage to restore or improve the farmer's soil. The pyro-oil could be collected and distilled off-site into additional transportation fuel and chemicals.[61] The thermal energy can be used on-site for heating, drying, hot water, and steam and other farm applications. Along with biochar and a productive outlet for sewage sludge (sometimes called biosolids) and its otherwise unusable water, such a comprehensive strategic program could provide cost-competitive renewable liquid fuel alternatives to diesel from imported petroleum.

The production of biodiesel in the United States increased from 75 million gallons in 2005 to 250 million gallons in 2006 and 450 million gallons in 2007, with an expected capacity of well over 1000 million gallons in the next few years.[66] This is still a small rate compared to over 60,000 million gallons of petroleum-based diesel consumed in the United States in 2009.

Sewage sludge represents a potentially large source of fatty acids for biodiesel production. After completion of the treatment cycle inside a sewage plant, these biosolids can be chemically processed to extract a biodiesel.[66,67] The waste residual solid and liquid can still serve for soil fertility enhancement thanks to its nitrogen, phosphate, and other mineral components.

The displacement of petroleum diesel can be further hastened and increased substantially with rendered animal fats, using thermal animal fat rendering technology, advanced greatly in 1811 by the French chemist Chevault. This technology has been further advanced by recent research and development.[68]

Recycling and SWEATT

Waste to energy is viewed by some groups as a threat to recycling, but the opposite is more likely the case. Municipal or institutional recycling programs in which the household sorts the SW for collective pickup can actually help maximize the return on waste components and minimize environmental problems. For example, if, at a given time, waste newspaper has no recycling value but, instead represents a disposal cost, preprocessing it for use as dry high-energy feedstock for SWEATT systems could be the optimum response. In effect, the marketplace could decide whether to recycle via the material recovery route or via an energy or fuel recovery route. Separation at the source generally requires simpler, less expensive preprocessing technologies than if all the waste sorting was carried out at the SWEATT plant. Sorting at the source also lends itself to a front-end pollution prevention program in which problem materials are separated and directed to hazardous waste facilities.[69,70]

Recycling with sorting at the source also facilitates application of beneficial feedstock-blending strategies. The analytical semi-empirical model (ASEM) study shows that polyethylene can serve as a rich source of energetic hydrogenic compounds. Thus, blending this plastic with biomass in a SWEATT pyrolysis system is expected to yield more energetic pyro-gas or pyro-liquid fuels and higher heat outputs than simple biomass. This has been demonstrated in high-temperature (1000°C) intermediate (between slow and fast) pyrolysis processing of Meals Ready to Eat (MRE) waste in which the heating value of the pyro-gas measured about 900 Btu/ft^3, quite close to the 980 Btu/ft^3 of natural gas. Food waste pyrolysis generally is expected to yield a low HHV pyro-gas (<150 Btu/ft^3). The high HHV of MRE pyrolysis is due to the large percentage of the ethylene monomer (C_2H_4) contributed by polyethylene plastic packaging. Plastics are used extensively in modern agriculture[71] and, after a growing season, present either a disposal problem or a good SWEATT feedstock blending opportunity. Alternatively, since plastics melt at relatively low temperatures, a sorting–recycling program together with very low temperature pyrolysis can relatively easily restore hydrocarbon plastics to liquid fuel forms.[72]

A Biomass Alliance with Natural Gas is a promising fuel blending strategy. Partial combustion gasifiers using biomass feedstock produce a low heating gas (~150 Btu/ ft^3) that will result in de-rating a natural gas-designed turbine-generator system. Coutilizing the biomass pyro-gas with natural gas can insure that the input energy requirement matches the output needs at least until the maximum rating of the generator is required. In a SW/biomass alliance with natural gas, an additional option becomes available when the SW comes from a recycling community. Then, the utility as a means of getting a richer gas to follow peak loads without calling upon the full use of natural gas might prepare and store high-energy plastics for increased use during times of high electricity demand. Another type of co-use strategy might be profitable if natural gas prices are low (≤$4/MMBtu) as they were at the end of 2011 and biodiesel is high (≥ $35/MMBtu). Then, the SWEATT system should be configured and operated and use the pyro gas yield supplemented by cheap natural gas to maximize the most valued pyro-liquid and pyro-chemical yields.[73] Many other opportunities for efficient co-use of domestic resources are described in the Proceedings of an International Conference on Co-utilization of Domestic Fuels.[6]

SWEATT: Summary and Conclusions

The acronym SWEATT was inspired by the senior author's (A.G.) recollection of Winston Churchill's historic call to arms in the darkest hours of World War II while admitting "I have nothing to offer but blood, toil, tears, and sweat." Today, the energy, environmental, economic, and security (EEES) problems of many nations are such as to bring to mind the problems of the World War II era and the need for bold imaginative solutions. The sustainable supplies of now wasted solids particularly in the United States that have an annual energy potential comparable to the total primary energy supply (TPES) contributed by coal as well as the TPES contributed by natural gas is a resource that should be a component of these solutions. Converting this SWEATT could multiply its current contribution to the U.S. TPES by a factor of about 10.

With low-energy density feedstock, minimizing transportations costs is very important. Thus, in addition to robust thermal technologies that can handle municipal solid waste (MSW), small on-site SWEATT systems have a particularly important role to play. Such systems should soon be available to harvest the energy and valuable chemicals in many of the SW categories listed in Table 1, most of which are greenhouse gas neutral and could gainfully be processed on-site, Table 1 does not list oil shale or tar sands in the United States that could substantially increase the available "solid waste" tonnage to address the U.S. need for liquid transportation fuels. A 2005 Rand study[74] shows that low temperature pyrolysis could be used to extract oils from domestic oil shale. The same can be said about hydraulic fracturing for the production of both oil and natural gas. Indeed this "fracking" might evolve as a big game changer if environmental concerns can be mitigated.

Marginal agricultural lands not in food production that could grow hardy, high-yield, fast-growing vegetable oil crops or can be improved to do so represent another form of waste that is gaining attention to solve energy security and climate change problems and particularly the need for biodiesel. Demonstration programs to date look favorable but the biodiesel price comes out higher than conventional diesel (see Table 2). Thus, it is important in such operations to generate additional revenue streams or to lower the costs of production. On-site SWEATT systems that convert the pressed seeds (meal) and plant residual (stover) into valuable pyro-oil, biochar, and thermal energy and reduce waste disposal costs could potentially make biodiesels more price competitive with diesel distilled from petroleum. It must be recognized, however, that imported petroleum is priced by cartels rather than a free market, Hence, countries poorly endowed with petroleum resources should not abandon biodiesel RD&D efforts or subsidies whenever this cartel to regain market share drastically lowers its price as in 1986.

Japan, a country with an outstanding sustainability record, has established the technical and environmental feasibility of large-scale conversion of SW to energy with thermal pyrolysis and gasification systems.[75] This should allay the concerns of environmentalists and risk-averse utility decision makers in other countries. A recent comprehensive report on the environmental performance of thermal conversion technologies throughout the world[76] identified more than 100 facilities that are using conversion technologies to convert MSW (mainly biomass) for energy production. The study used independently validated emissions data from operating facilities in five nations and found that pyrolysis and gasification facilities currently operating throughout the world with waste feedstocks meet each of their air quality emission limits. With few exceptions, most meet all of the current emission limits mandated in California, the United States, the European Union, and Japan. In the case of toxic air contaminants (dioxins/furans and mercury), every process evaluated met the most stringent emission standards worldwide. Facilities with advanced environmental controls are very likely to meet regulatory requirements in California. Thus, the report concludes that thermochemical conversion technologies possess unique characteristics that as a part of an integrated waste management system can generate useful energy and substantially reduce the amount of material that must be landfilled.

Finally, the following are our main conclusions:

The United States and most industrial nations are excessively reliant on imported oil for their liquid fuels and imported natural gas for their gaseous fuels.

The United States is well endowed with solid coal and oil shale, which can be converted into useful liquid and gaseous fuels by ATTs to mitigate its energy security problem, albeit not its problem of excessive CO_2 emissions.

The United States is also well endowed (~2 billion dry tons) with many forms of carbon-neutral forms of SW that can be converted to natural gas supplements, bio-liquid fuels, thermal energy, and pyro-char by on-site SWEATT systems.

ATTs are the fastest and most efficient SW to energy conversion methods. The high volatile content of biomass makes high-temperature pyrolysis of the biomass component of SW a direct form of gasification.

- If environmental and security externalities are included, renewable energy often compete with fossil fuel energy.

Approximate analytic representations of cost of electricity vs. the important controlling variables can be helpful in making SWEATT policy decisions in the face of highly volatile natural gas prices and uncertainties as to values assigned to externalities.

- Co-use of volatiles from the pyrolysis of SW with natural gas can be useful in overcoming a number of EEE problems.

SWEATT generally results in lower harmful emissions than traditional high-temperature waste to energy systems (incinerators).

- Given a free biomass source such as SW, SWEATT becomes cost competitive with NGCC energy production systems when NG fuel prices are above about $4/ MMBtu.
- Blending feedstock and including inexpensive catalysts can enhance production of energy and high-valued products.

Many areas of engineering research will be needed to optimize SWEATT systems.

The solid residuals of SWEATT, biochar, have potential value as soil amendments to boost fertility and as industrial and environmental adsorbents.

- The biochar dimension offers promise of achieving systems with carbon negativity. Thus, the economic profitability of pyrolysis SWEATT systems could be hastened if and when "C credits" or "C trading" becomes available.[40,41]
- Optimizing the production of pyro-char products and its characteristics for specific needs could enhance the economic and environmental benefits of SWEATT systems without greatly reducing its energy benefits.
- SWEATT can be used in parallel with other fuel-generating enterprises such as biodiesel production and presorted waste recycling operations to increase cost- competitiveness and maximize environmental benefit.
- At present, biochar might only deliver climate change mitigation benefits and be environmentally and financially viable as a distributed system using waste biomass, i.e., small-scale on-site SWEATT.[47]

Finally, the authors conclude that applications of SWEATT can play a significant role in the variety of solutions that will be needed to address today's EEES problems.

Acknowledgments

This work was partially supported by a grant from NSF- EAR #0819706, Geobiology and Low Temperature Geochemistry Program to A. Zimmerman. Alex Green would like to thank his many sponsors and collaborators in his quest for alternatives to oil that began in 1973 and in his work on On-Site SWEATT that began with Bruce Green in 1986. In this update effort, he acknowledges the help of Professor Nicholas Comerford, Dr. Deborah Green, and Julio Castro Vazquez.

Nomenclature and Acronyms Used in Chapter

ABPC	Air Blown Partial Combustion
ACE	Analytical Cost Estimation
AGIR	Antares Group Inc. Report
Ar	Aromatics
ASEM	Analytical Semi-Empirical Model
ATT	Advanced Thermal Technologies
BTU	British Thermal Units
CCTL	Clean Combustion Technology Lab
CHP	Combined Heat and Power
CIE	Compressed Ignition Engines
DANSF	Dry Ash, Nitrogen and Sulfur Free
EU	European Union
FC	Fixed Carbon
GACE	Generalized Analytic Cost Estimation
HHV	Higher Heating Values
HRSG	Heat Recovery Steam Generator
ICE	Internal Combustion Engines
IGCC	Integrated Gasifier Combined Cycle
MSW	Municipal Solid Waste
NGCC	Natural Gas-Fired Combined Cycle
NPHR	Net Plant Heat Rate
OBPC	Oxygen Blown Partial Combustion
OFC	Omnivorous Feedstock Converter
PES	Primary Energy Supplies
PNA	Polynuclear Aromatics
PYRO	Pyrolysis Systems
QES	Quaternary Energy Supply quads Quadrillion BTUs
RDF	Refuse Derived Fuels
SES	Secondary Energy Supplies
SW	Solid Waste
SWANG	Solid Waste Alliance with Natural Gas
SWEATT	Solid Waste to Energies by Advanced Thermal Technologies
TES	Tertiary Energy Supply
TPES	Total Primary Energy Supply
V_T	Volatiles
WEC	Waste to Energy Conversion

References

1. Green, A., Ed. *An Alternative to Oil, Burning Coal with Gas,* University of Presses of Florida: Gainesville, FL, 1981.
2. Green, A., et al. *Coal-Water-Gas, An All American Fuel for Oil Boilers,* Proceedings of the Eleventh International Conference on Slurry Technology, Hilton Head, SC, 1986.
3. Green, A., Ed. *Solid Fuel Conversion for the Transportation Sector,* FACT-Vol 12 ASME New York, NY. Proceedings of Special Session at International Joint Power Generation Conference San Diego, 1991.
4. Green, A. *A Green Alliance of Biomass and Coal (GABC),* Appendix F, National Coal Council Report May 2002: Proceedings of 27th Clearwater Conference, Clearwater, FL, March 2003, 2002.

5. Green, A.; Feng, J. A Green Alliance of Biomass and Natural Gas for a Utility Services Total Emission Reduction (GANGBUSTER), Final Report to School of Natural Resources and the Environment, 2003.
6. Green, A.; Hughes, E.; (EPRI), Kandiyoti, R.; (Imperial College London) conference organizers. Proceedings of the first international conference on co-utilization of domestic fuels. Int. J. Power Energy Syst. **2004**, *24,* 3.
7. Green, A.; Smith, W.; Hermansen-Baez, A.; Hodges, A.; Feng, J.; Rockwood, D.; Langholtz, M.; Najafi, F.; Toros, U.; Multidisciplinary Academic Demonstration of a Biomass Alliance with Natural Gas (MADBANG). Proceedings of the International Conference on Engineering Education, University of Florida, Conference Center, Gainesville, FL, Oct 2004.
8. Green, A.; Swansong, G.; Najafi, F. Co-utilization of Domestic Fuels Biomass Gas/Natural Gas, GT2004–54194, IGTI meeting in Vienna June 14–17, **2004**.
9. Green, A.; Feng, J. Assessment of Technologies for Biomass Conversion to Electricity at the Wildland Urban Interface Proceedings of ASME Turbo Expo 2005: Reno-Tahoe, 2005.
10. Green, A.; Bell, S. Pyrolysis in Waste to Energy Conversion (WEC), Proceedings of NAWTEC 14 Conference Tampa, FL, May, 2006.
11. Jenner, M. Biomass Rules. LLC Cited in Natural Gas Futures NYMEX Jan 10th, 2010.
12. Channiwala, S.; Parikh, P. A unified correlation for estimating HHV of solid liquid and gaseous fuels. Fuel **2002**, *81,* 1051–1064.
13. Perlack, R.; Stokes, B.; Erbach, D. Biomass as Feedstock for a Bioenergy and Bioproducts Industry: The Technical Feasibility of a Billion-Ton Annual Supply, Oak Ridge National Laboratory, ORNL/TM-2005/66, U.S. Department of Energy, 2005.
14. Green, A.; Peres, S.; Mullin, J.; Xue, H. *Co-gasification of Domestic Fuels*, Proceedings of IJPGC. Minneapolis, MN, ASME-NY, NY, 1996.
15. Green, A.; Zanardi, M.; Mullin, J. Phenomenological models of cellulose pyrolysis. Biomass Bioenergy **1997**, *13,* 15–24.
16. Green, A.; Zanardi, M. Cellulose pyrolysis and quantum chemistry. Int. J. Quantum Chem. **1998**, *66,* 219–227.
17. Green, A.; Mullin, J. Feedstock blending studies with laboratory indirectly heated gasifiers J. Eng. Gas Turbines Power; **1999**, *121,* 1–7.
18. Green, A.;Mullin, J.; Schaefer, G.; Chancy, N.A.; Zhang, W. *Life Support Applications of TCM-FC Technology,* 31st ICES Conference, Orlando, FL, Jul 2001.
19. Green, A.; Venkatachalam, P.; Sankar, M.S. Feedstock Blending of Domestic Fuels in Gasifier/Liquifiers, TURBO EXPO 2001, Amsterdam, GT, 2001.
20. Green, A.; Chaube, R.; Pyrolysis Systematics for Coutilization Applications. TURBO EXPO 2003, Jun 2003, Atlanta, GA, 2003.
21. Green, A.; Chaube, R.A. A unified biomass-coal pyrolysis reaction model. Int. J. Power Energy Syst. **2004**, *24*(3), 215–223.
22. Green, A.; Sadrameli, S. M. Analytical represntations of experimental polyethylene pyrolysis yields. J. Anal. Appl. Pyrolysis **2004**, 72, 329–335.
23. Sadrameli, S.; Green, A. Systematics and modeling representations of naphtha thermal cracking for olefin production. J. Anal. Appl. Pyrolysis, **2005**, *73,* 305–313.
24. Green, A.; Feng, J. Systematics of corn stover pyrolysis yields and comparisons of analytical and kinetic representations. J. Anal. Appl. Pyrolysis **2006**, *76,* 60–69.
25. Feng, J.; Green, A. Peat pyrolysis and the analytical semiemperical model. Energy Sources, Part A. **2007**, *29,* 1049–1059.
26. Feng, J.; YuHong, Q.; Green, A., Analytical model of corn cob pyroprobe-FTIR data. J. Biomass Bioenergy **2006**, *30,* 486–492.
27. Gaur, S.; Reed, T. *Thermal Data for Natural and Synthetic Fuels*; Marcel Dekker: New York, NY, 1998.

28. Xu, W.C.; Tomita, A. Effects of coal type on the flash pyrolysis of various coals. Fuel **1987b**, 66, 627–636.
29. Mastral, F.J.; Esperanza, E.; Garcia, P.; Juste, M. Pyrolysis of high-density polyethylene in a fluidised bed reactor. Influence of the temperature and residence time J. Anal. Appl. Pyrolysis **2002**, *63,* 1–15.
30. Mastral, F.J.; Esperanza, E.; Berruco, C.; Juste, M. Fluidized bed thermal degradation products of HDPE in an inert atmosphere and in air-nitrogen mixtures. J. Analy. Appl. Pyrolysis **2003**, 70, 1–17.
31. Stultz, S.; Kitto, J., Ed., *Steam* 40th Ed.; Babcock & Wilcox: Barberton, OH, 1992.
32. Liscinsky, D.; Robson, R.; Foyt, A.; Sangiovanni, J.; Tuthill, R.; Swanson, M. *Advanced Technology Biomass-Fueled Combined Cycle,* Proceedings of ASME Turbo Expo 2003, Power for Land, Sea and Air, Atlanta, GA, GT2003–38295, 2003.
33. Phillips, B.; Hassett, S., *Technical and Economic Evaluation of a 79 MWe (Emery) Biomass IGCC,* Gasification Technologies Conference, San Francisco, 2003.
34. Antares Group, Inc. Assessment of Power Production at Rural Utilities Using Forest Thinnings and Commercially Available Biomass Power Technologies. Landover, MD, Sept 2003.
35. ICF Consulting. *Options for Meeting the Electrical Supply Needs of Gainseville* Report March 6, 2006.
36. Roth, I.; Ambs, L. Incorporation externalities into a full cost approach to electric power generation life-cycle costing. Energy **2004**, l29(12–15), P2125–P2144.
37. Green, A.; Palacio, J.; Bell, S.; Avery, M. Sustainable Energy for Communities and Institutions, Proceedings of Sustainability Conference, University of Florida 2006.
38. Green, A. Operations Analysis and the 20th Air Force Slide Rules of WW II, J. Oughtred Soc. **2000**, 9, 212.
39. Mirasgedis, S.; Diakoulaki, D. Multicriteria analysis vs. externalities assessment for the comparative evaluation of electricity generation systems. Eur. J. Oper. Res. **1997**, 102(2), 364–379.
40. Lehmann, J.; Joseph, S. Eds., *Biochar for Environmental Management: Science and Technology.* Proceedings of Newcastle U.K. Conference, Earthscan Publishers Ltd., 2009.
41. Woods, W.; Teixeira, W.; Lehmann, J.; Steiner, C.; Winkler Prins, A.; Rebellato, L. Eds. *Amazonian Dark Earths: Wim Sombroek's Vision;* Springer Publishers: Dordrecht., 2009, 504 p.
42. Antal, M.J.; Gronli, M. The art, science, and technology of charcoal production. Ind. Eng. Chem. Res. **2003**, *42,* 1619–1640.
43. Falcão, N.; Comerford, N.; Lehmann, J. Determining nutrient bioavailability of amazonian dark earth soils—methodological challenges. In *Amazonian Dark Earths: Origin, Properties, and Management;* Lehmann, J., Kern, D., Glaser, B., Woods, W., Eds.; Kluwer Academic Publishers: The Netherlands, 2003, 255–270.
44. Downie, A.; Crosky, A.; Munroe, P. Physical properties of biochar. In *Biochar for Environmental Management: Science and Technology;* Lehmann, J.; Joseph, S., Eds.; Earth- scan: London, 2009; pp. 183–205.
45. Glaser, B. Prehistorically modified soils of central Amazonia: A model for sustainable agriculture in the twenty- first century. Phil. Trans. R. Soc. B Biol. Sci. **2007**, *362,* 187–196.
46. Lehmann, J. Bio-energy in the black. Front. Ecol. Environ. **2007**, 3, 381–387.
47. Roberts, K.; Gloy, B.A.; Joseph, S.; Scott, N.R.; Lehmann J. Life cycle assessment of biochar systems: Estimating the energetic, economic, and climate change potential. Environ. Sci. Technol. **2010**, *44*, 827–833.
48. Zimmerman, A. Abiotic and microbial oxidation of laboratory-produced black carbon (biochar). Environ. Sci. Tech- nol. **2010**, 44(4), 1295–1301, DOI: 10.1021/es903140c.
49. Kasozi, G.N.; Zimmerman, A.R.; Nkedi-Kizza, P.; Gao, B. Catechol and humic acid sorption onto a range of laboratory- produced black carbons (biochars). Environ. Sci. Technol. **2010**, 44, 6189–6195.

50. Mukherjee, A.; Zimmerman, A.R.; Harris, W. Surface chemistry variations among a series of laboratory-produced biochars. Geoderma. **2011**, *163,* 247–255.
51. Bansal, R.; Goyal, M. *Activated Carbon Adsorption;* Taylor and Francis, 2005; 497 pp.
52. Inyang, M.; Gao, B.; Pullammanappallil, P.; Ding, W.; Zimmerman, A.R.; Cao, X. Enhanced lead removal by biologically activated biochar from sugarcane bagasse. Separation Sci. Technol. **2011**, *46,* 1950–1956.
53. Khodadad, C.L.M.; Zimmerman, A.R.; Uthandi, S.; Foster, J.S. Changes in microbial community composition in soils amended with pyrogenic carbon. Soil Biol. Biochem. **2010**, *43,* 385–392.
54. Woolf, D.; Amonette, J.E.; Street-Perott, A.; Lehmann, J.; Joseph, S. Sustainable biochar to mitigate global climate change. Nat. Commun. **2010**, 1, 1–9.
55. Inyang, M.; Gao, B.; Pullammanappallil, P.; Ding, W.; Zimmerman, A.R. Biochar from anaerobically digested sugarcane bagasse. Bioresour. Technol. **2010**, *101,* 8868–8872.
56. Green, A. Transportation. In *Encyclopedia of Climate and Weather;* Schneider, S.H., Root, T.H. and Mastrandrea, M.D. Eds.; Oxford University Press: Oxford, N.Y., 2011, 184–187.
57. Green, A.; Schaefer, G. What to Do with CO_2. Proceedings Turbo Expo 2001, New Orleans paper IGTI 2001-GT-1.
58. Hansen, J.; Mki. Sato, P.; Kharecha, D.; Beerling, R.; Berner, V.; Masson-Delmotte, M.; Pagani, M.; Raymo, D.L.; Royer, and J.C. Zachos, 2008: Target atmospheric CO_2: Where should humanity aim? *Open Atmos. Sci. J,* 2, 217–231, doi:10.2174/1874282300802010217.
59. Princiotta, F. Global climate change and the mitigation challenge. J. Air Waste Manage. Assoc. **2009**, 39(10), 1194–1211.
60. Doran-Peterson, J.; Amruta Jangid, S. K.; Brandon, E.; De- Crescenzo-Henriksen, E.; Ingram, L.O. Simultaneous saccharification and fermentation and partial saccharification and co-fermentation of lignocellulosic biomass for ethanol production. Biofuels Methods Mol. Biol. **2009**, *381,* 263–280.
61. Holmgren, J.; Marinangeli, R.; Elliott, D.; Bain, R. Consider upgrading pyrolysis oils into renewable fuels. Hydrocarbon Process. **2008**, *Sept,* 95–103.
62. Peterson, C.; Ault, D. Technical overview of vegetable oil as a transportation fuel. ASME Fact **1991**, *12,* 45–54.
63. Brown, S.; Henry, C. Natural Selection Farms (Nov. 29, 2010), available at http://wwwkingcounty-gov/environment/wastewater/Biosolids/BiosolidsRecyclingProjects/ NaturalSelectionFarms.aspx (accessed Nov. 30, 2010).
64. Morgan, G. *Flax Finding New Life as Biodiesel Stock— Domestic Fuel.* Agra Life Communications April 4, 2008, College Station, Texas.
65. Vasden, B. *Commercial Biomass and Biofuel Feedstock Farming in Florida;* 2010; available at http://www.usqocom(accessed Dec. 1, 2010).
66. Kargbo, D.M. Biodiesel production from municipal sewage sludges. Energy Fuels **2010**, *24* (5), 2791–2794.
67. Dufreche, S.; Hernandez, R.; French, T.; Sparks, D.; Zappi, M.; Alley, E. Extraction of lipids from municipal wastewater plant. Microorganisms for production of biodiesel. J. Am. Oil Chem. Soc. **2007**, *84,* 181–187.
68. Dias J., et al. Production of biodiesel from acid waste lard. Bioresour. Technol. **2009**, *100,* 6355–6361.
69. Green, A., Ed. *Pollution Prevention and Medical Waste Incineration;* Van Nostrand Reinhold: New York, 1992
70. Wisconsin Department of Natural Resources Agricultural Plastics (Sept. 1, 2010) available at http://dnrwigov/org/aw/wm/recycle/newpages/agplastics.htm (accessed Nov. 30, 2010).
71. Smith, C. Plastic to Oil Fantastic (Apr. 4, 2009), available at http://www.rexresearch.com/ito/ito.htm (accessed Nov. 30, 2010).
72. Akinori Ito Invents a Machine Converting Waste Plastic into Oil (Aug. 23, 2010), available at http://www.neverthelessnation.com/2010/08/akinori-ito-invents-machine-converting.html (accessed Nov. 30, 2010).

73. Vispute, T.P.; Zhang, H.; Sanna, A.; Xiao, R.; Huber, G.W. Renewable chemical commodity feedstocks from integrated catalytic processing of pyrolysis oils. Science **2010**, 330(6008), 1222–1227.
74. Bartis, J.; LaTourrette, T., et al. *Oil shale development in the United States: Prospects and policy issues*; RAND Corp. ISE Division, NETL, US DoE: Pittsburgh, PA, 2005.
75. California Integrated Waste Management Board. Conversion Technologies Report to the Legislature; 2005, available at http://www.ciwmb.ca.gov/Organics/Conversion/ Events Agra Life Communications, College Station, Texas.
76. Bioenergy Producers Association. UC-R Issues Report on Thermal Conversion Technologies (Jan.19, 2010), available at http://www.bioenergyproducers.org/index.php/ucr-report(accessed Nov. 31, 2010).

32

Energy: Walls and Windows

Therese Stovall

Introduction

The building envelope protects you from the weather, separating the indoor air that you have paid to heat or cool from the outdoor air. The walls, roofs, and windows form the major surfaces of this envelope. Energy travels through these surfaces along many paths and in many forms, such as unintended air leakage through a wall or sunlight streaming through a window. We can improve the energy efficiency of the building envelope if we carefully consider all of these energy pathways.

Here, we are focusing on energy efficiency, but keep in mind that each part of the building envelope performs multiple jobs under challenging conditions. The roof keeps out rain, hail, and snow; bakes in the hot summer sun; freezes in the cold winter night; and must be sturdy enough to survive a workman's boots. The walls must repel the rain, hold up the roof, stop the wind, and provide a rigid support for windows and doors. The windows have to let in the light, allow ventilation when they are open, and keep out drafts when they are shut. Any change we make to the building envelope to conserve energy must account for these multiple functions and the complex interactions between the envelope components.

Resistance to heat transfer is often expressed as an *R*-value. For a complete description of R-values and their use, please refer to the DOE Insulation Fact Sheet (http://www.ornl.gov/roofs+walls/insulation).

Building Types

Buildings fall into two main classes: high-rise and low-rise. The high-rise buildings are typically custom engineered with structural steel frames. Low-rise buildings are typically divided into commercial, low-rise multifamily, and single-family buildings. The high-rise, commercial, and low-rise multifamily buildings are more likely to have low-slope (often mislabeled as "flat") roofs, whereas the single-family houses are more likely to have pitched or steep-sloped roofs.

Construction methods can be roughly grouped according to the portion of the assembly performed on site and the portion performed at a factory. With today's engineered wood products and premade trusses, few buildings are strictly built on site; but we still refer to a stick-built building as the one where the greatest part of the assembly takes part on the construction site. At the other end of the spectrum are manufactured buildings that can be moved from one site to another with relatively little effort. In the middle are factory-built modular buildings and panelized construction. A factory-built modular building typically includes one or more modules that are placed on a permanent foundation. In some modular buildings, the windows are installed after the modules have been installed. A panelized building is closer to the site-built model, but will have major wall, floor, or foundation sections prebuilt and delivered to the site.

Every building must conform to local building codes, which can limit the material choices or construction methods. Many local building codes now include energy conservation clauses or incorporate the Model Energy Code or the International Energy Conservation Code.[1]

Walls

The walls make up most of the exterior surface area of many buildings and therefore the energy transported through these surfaces is very important. Wall issues vary according to the building type. A framed building is constructed with a wood or metal skeleton that provides structural support to both the building and all the other wall components. A framed wall is characterized by numerous parallel heat paths and multiple layers of different materials. A non-framed building uses bulk material, such as masonry or adobe, to provide the structural support. This type of building is characterized by a more homogenous heat path and relatively few layers of materials.

For either type of building, the connections between the wall and the roof and between the wall and the foundation are important construction details from an energy conservation standpoint. These connections can provide unintended air passageways and may be overlooked in the overall insulation scheme. An air drywall approach is an effective way to limit air leakage from walls, roofs, and windows. Here, a rubber gasket is fitted along the perimeter of the window frame and along the exterior wall's base board and ceiling plate to compensate for openings that occur as the wood changes shape with time. The gasket, once placed, seals against the drywall gypsum board and makes an airtight barrier.

Whether the wall is built on a frame or constructed from masonry, there is a wide selection of exterior siding choices. These include brick; wood, fiber cement, or vinyl siding; or an exterior insulation finish system (EIFS). For all of these facades, repelling rain and wind is often more complex than it looks. For example, a brick wall looks like a solid surface, but the mortar joints provide capillary paths for moisture, especially when subjected to wind-driven rain. Similar pathways exist for other cladding materials. For this reason, air gaps are often provided behind the outermost wall layer. Depending on the vent/drain arrangement, this may provide a true pressure-equalized rain screen or a simple break in the capillary path. Old-fashioned wood lap siding is a good example of a simple rain screen, as shown in Figure 1. The small air gap behind each wood layer is well vented and drained to the outside, so that the air pressure within the air gap is equal to the air pressure outside the wall. This pressure equalization reduces the moisture moving into the air space, and therefore reduces the amount of moisture available to penetrate the rest of the wall.[2,3]

Proper moisture management is important for energy conservation for two reasons. First, unmanaged moisture must be removed by additional ventilation, which entails the energy load needed to heat and cool the additional air mass. Second, moist building materials will always have a higher thermal conductivity than dry materials. When wet, some insulation materials become matted and lose the greater part of their insulating value.

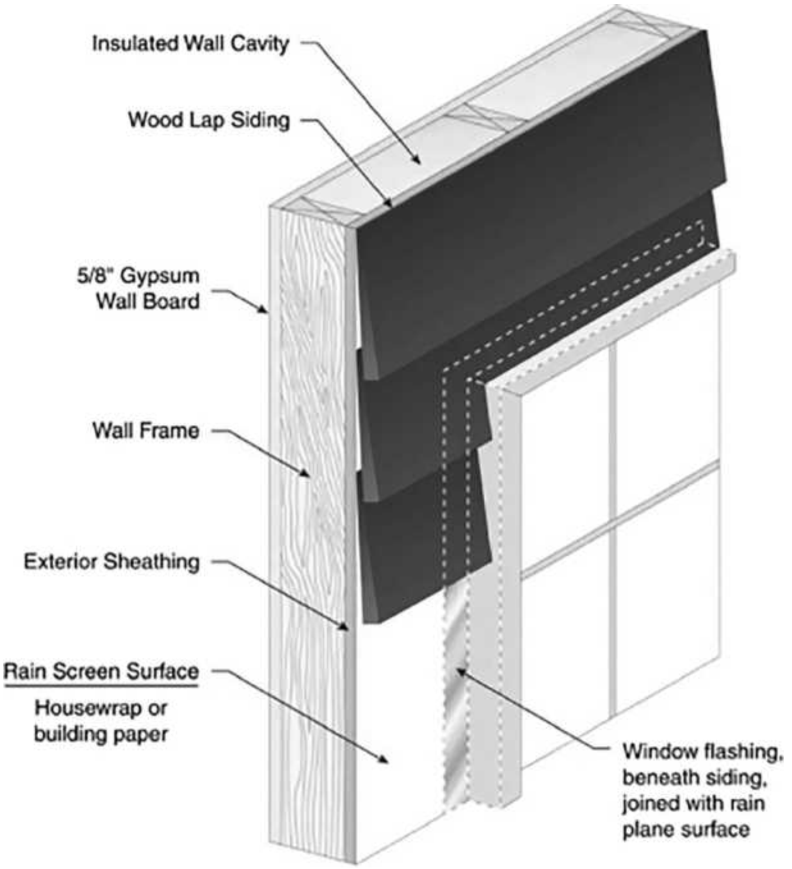


FIGURE 1 Simple rain screen within a typical wood lap siding wall.

Every wall system stores energy. This storage quality is called "thermal mass." The thermal mass can reduce the amount of heat lost or gained through the wall whenever the outdoor temperature varies above and below the indoor temperature on a daily basis. This occurs during the spring, summer, and fall for most of the United States, and during the winter as well for the southern regions. The relative benefit of thermal mass is therefore determined by both the wall properties and the climate. A wood-framed wall has very little thermal mass compared with a masonry wall. One study compared the energy consumed by a house with traditional wood-framed walls with a house constructed with masonry walls for six cities. Depending on the location and wall thickness, the more massive wall reduced the household energy use by an amount equal to increasing the traditional wall's thermal resistance by 10%–50%, with the savings greatest in Denver and Phoenix and least in Miami.[4]

Although we typically think of the exterior walls when we think about energy losses, the interior walls are also important. Air enters the wall cavity through a number of penetrations, some visible and some not. Holes made in the drywall to accommodate electrical outlets and plumbing connections also allow uncontrolled airflow. Other gaps are often present at the floor–wall and floor–ceiling connections. Therefore, it is important to provide a continuous top plate above every wall cavity. This top plate separates the interior wall cavity from the attic space, thus preventing a free flow of conditioned air from your house into the attic.

In addition to the energy used to heat and cool a building, energy is also embodied in the building materials and expended in the construction process. Among low-rise residential buildings, studies have shown that many of the building elements, including gypsum drywall, roofing materials, and carpeting, are common to all the wall types. But the wall type still makes a significant difference in the overall energy embodied in the building, with a wood-framed house containing about 15% less embodied energy than either a metal-framed or concrete house.[5,6]

Wood-Framed Walls

Wall construction methods and materials vary somewhat according to the local climate and natural resources, but most walls in the United States are made from wood framing with insulation between the studs, drywall on the inner surface, and exterior sheathing layer(s) (Figure 2). The wall studs used are either nominal 2 × 4, with a 3.5-in. cavity depth, or nominal 2 × 6, with a 5.5-in. cavity depth. Cavity insulation options for new construction include batts, a blown-in mixture of a foam binder and loose-fill insulation, blown-in loose-fill insulation secured by nets, or blown-in foam insulation.[8] In commercial buildings, high-density batts are sometimes used to provide both thermal insulation and acoustical buffering. In retrofit situations, professional installers can blow loose-fill insulation into the wall cavities by drilling a series of holes through the interior or exterior facade between each pair of adjacent studs.

Energy flows through the wood studs more easily than through the surrounding insulation. Most walls contain much more framing than you would think, as shown in Figure 3. In addition to the studs, there is additional wood framing around each window, around each door, at each building corner, where the wall sits on the foundation, where the roof sits on the wall, where an interior wall meets an

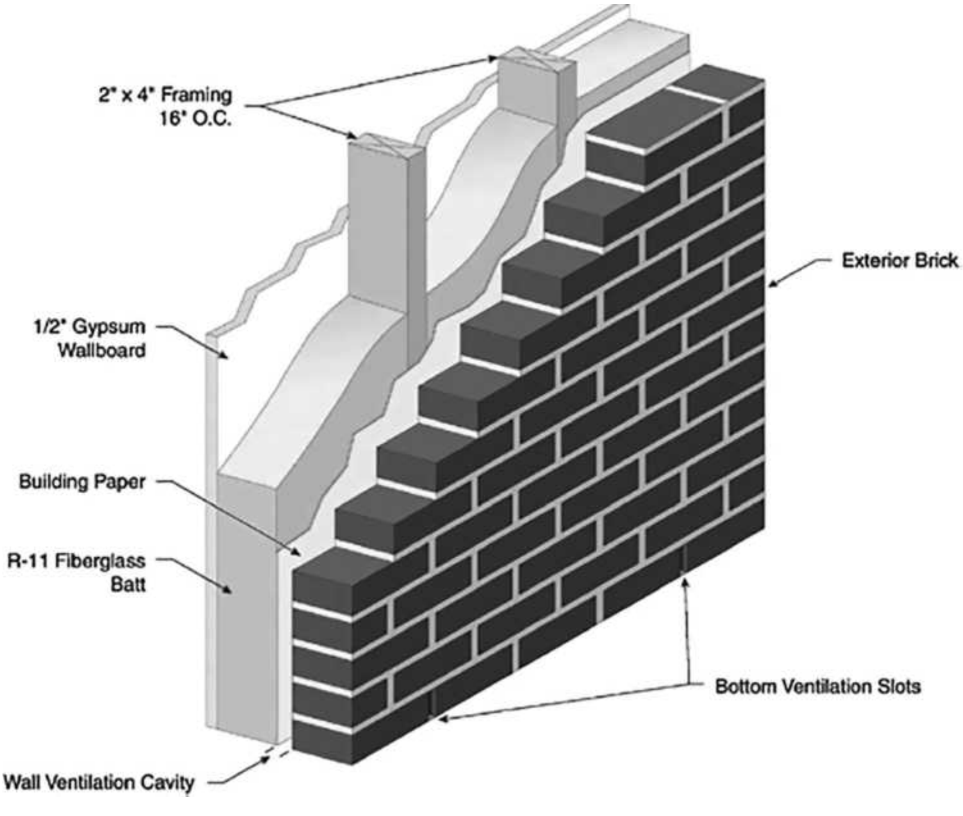


FIGURE 2 Typical wall structure for a wood-framed wall with brick cladding.
Source: ASHRAE Special Publications (see Christian et al.[7])

FIGURE 3 Framing lies behind a significant portion of the wall area in many houses.
Source: ASHRAE Special Publications (see Joint Center for Housing Studies of Harvard University, Remodeling, Measuring the Benefits of Home Remodeling[9]).

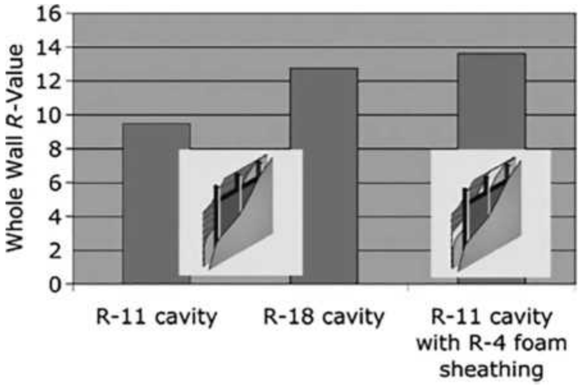


FIGURE 4 Whole wall *R*-values for wood-framed (nominal 2 × 4) walls on 16-in. centers with interior gypsum and exterior wood siding in a one-story ranch house.

exterior wall, and between floors in a multistory building. When you combine all of these thermal "short circuits," a wall (in a one-story ranch style house) filled with R11 insulation provides an overall performance of only R9–R10. If you replace that insulation with foam in the wall cavity, thereby increasing the cavity insulation from R11 up to about R18, you get an overall performance of R13 or an increase that is about half of the increased insulation value. One way to improve the thermal performance of a wall is therefore to place insulation between the studs and the exterior surface of the wall. For example, adding only 1 in. (R5) of foam sheathing to the R10 wall brings its overall R-value up to R14 (Figure 4).

Advanced framing techniques are available which reduce the amount of lumber required. These methods reduce the energy losses through the framing and allow more room for insulation. Many of these techniques also provide for improved air sealing.[10]

The wall sheathing provides a flat uniform surface to support the exterior air barrier, vapor retarder, and siding. If a wood product is used for the sheathing, it will also provide the structural stiffness needed at the building corners. If a foam insulation product is used instead, some form of additional bracing will be necessary at the corners. Sometimes a layer of foam is placed on top of a layer of wood product sheathing. This greatly improves the wall's thermal resistance because that layer covers the thermal short circuits provided by the wood frame. However, this configuration can require extra care during the finishing process, with longer nails needed to fasten siding materials. Also, specialty brick ties will be needed if a brick decor is selected for finishing the wall.

Steel-Framed Walls

Steel-framed walls share many of the characteristics of a wood-framed wall, but the steel components themselves have a very high thermal conductivity. Therefore, most steel-framed walls are built with a layer of foam sheathing to break that thermal pathway. There is also research underway to produce complex steel shapes to provide the structural support needed while providing a longer heat transfer pathway, and therefore greater thermal resistance.[11] Some metal-framed products also include an integral foam insulation element for the same reason. Connections between the walls and a steel-framed roof can be problematic from an energy point of view, especially if the steel framing extends out in the eaves. Such arrangements act like the fins on a heat exchanger and can cause excessive energy consumption if appropriate insulation arrangements are not included in the design. Steel-framed walls are most attractive where the heating loads are modest and where insect damage is more challenging.

Masonry Walls

Masonry walls can refer to walls with a masonry veneer, such as brick or stone, or to a wall where the masonry also provides the structural support, such as a poured concrete or concrete block wall. Such buildings are relatively resistant to the corrosive environment common near the ocean. Masonry also provides thermal mass that can both save energy and improve the interior comfort level in a hot climate with daily temperature swings. Aside from this thermal mass effect, masonry veneer walls share the same energy characteristics as other wood- or steel-framed walls.

Concrete Block and Poured Concrete

Full masonry walls are often used for foundations or basements. Whole houses built from masonry are popular in the warmer climates where termites and other insects are more populous. A full masonry wall can be built from concrete blocks or by pouring concrete into forms. For a load-bearing wall, steel reinforcement rods, or rebar, are used with either method to add strength. For a reinforced concrete block wall, reinforcing rods are placed vertically in the block cavities which are then filled with mortar. Steel wires or mesh are laid in the horizontal mortar joints to resist shear stress. In a poured concrete wall, the steel reinforcing rods are positioned both vertically and horizontally within the forms before the concrete is poured. Because these steel rods tend to be perpendicular to the heat transfer direction, they have little effect on the overall wall thermal resistance.

The thermal characteristics of a masonry wall vary, depending on the whether blocks or poured concrete is used and on the density of the concrete. But in general, the thermal resistance of the masonry portion of such walls will be very small, ranging from R1 to R2 for an 8-in. thick wall, even if the cores of a hollow concrete block are filled with perlite or vermiculite.[12] Foam board insulation, with a thermal resistance in the range of $5R$/in. is often used to increase the total thermal resistance of a masonry wall and can be placed on the inside and/or the outside surface. The foam board must be covered by some material with an appropriate fire rating, such as gypsum board. One study has shown that the thermal mass is more effective when the concrete is in good thermal contact with the interior building, i.e., when the insulation is placed on the outside of the wall.[13]

Autoclaved aerated concrete can be used in place of ordinary concrete for low-rise buildings. This type of concrete is much lighter than standard concrete, available in a variety of sizes and shapes, and may be reinforced. The autoclaved aerated concrete has a much higher R-value (1.25R/in.) than the standard concrete (0.05R/in.).[3] The overall R-value of a wall built with this material will depend on the shape and thickness of the concrete and on the thermal resistance of other wall components, such as air cavities. Several walls tested with aerated concrete (no facings applied), both in the traditional hollow concrete block and solid block forms, had R-values between 6 and 9.[14]

Precast Concrete

Walls can be made from reinforced concrete slabs that have been precast into the desired shape. Unless insulation is applied, these walls will have approximately the same thermal resistance as a site-poured concrete wall, i.e., about R1–R2 for an 8-in. thick wall. This construction method is more often used for larger buildings, such as apartment buildings or hospitals, but is also used for residential buildings. In the smaller buildings, the precast panels were first used for basement walls, but precast panels can be used for all exterior walls. The panels are precast and cured in the factory, thus avoiding weather limitations associated with pouring and curing concrete at the building site. The factory environment also allows the production of concrete that is stronger and more water resistant than site-poured concrete. Some of the panels are cast against foam insulation to improve the wall's R-value. In addition to the steel reinforcing, these walls can be produced with cavities for electrical wiring and rough openings. Some of the panels have been designed to provide the appearance of bricks and limestone, so that a new building will blend into an existing urban neighborhood. A crane is usually used to place the precast panels on top of a bed of crushed stone. The panels are then connected in place using weld joints or bolts and sealants. Installation for a typical residential unit can be completed in a single day.[15,16]

Insulated Concrete Forms

An insulated concrete form (ICF) wall is composed of a set of joined polystyrene or polyurethane forms that are filled with reinforced concrete, as shown in Figure 5. The joining methods vary from one manufacturer to another. Some have interlocking foam panels; others are linked with ties made from polypropylene or steel. The foam panels become a permanent part of the wall, providing a continuous layer of insulation on both the inner and outer wall surfaces. The concrete portion of the wall is reinforced with rebar positioned inside the forms before the concrete is poured to provide needed strength. The ICF walls can be covered on the outside with light-weight stucco, brick, or wood or vinyl siding. Gypsum board fastens onto the interior side. The attachment methods for these facing materials vary from one manufacturer to another. In the laboratory, the simple steady-state R-value of unfinished ICF walls varied from 12 to 18.[14] The time-varying effective energy performance of these walls is also determined by the temperature profile within the wall, which is improved by the walls' well-insulated thermal mass. Therefore, the ICF wall design provides both greater thermal mass and a higher thermal resistance than a typical 2 × 4 frame wall. Although the airtightness of any wall system tends to vary depending upon the expertise of the construction crew, an ICF wall is generally more airtight than a wood-framed wall.

A variant of the ICF wall uses forms made from a polystyrene–cement composite. The composite walls may have a greater thermal mass, but provide a significantly lower steady-state thermal resistance (about R8 for one 10-in. thick configuration).[14]

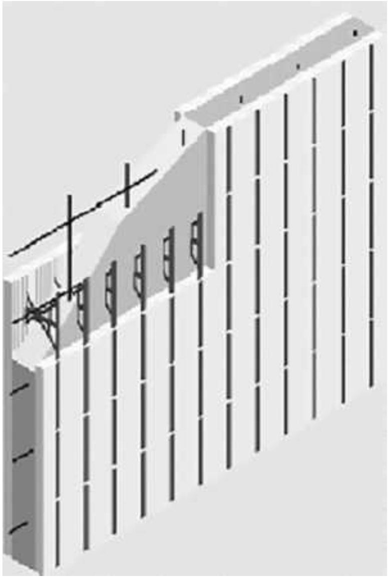


FIGURE 5 Insulated concrete form (ICF) wall.

Adobe

In parts of the southwest, adobe walls are popular because they absorb and store the daytime heat until it can be released to the cooler night air. Traditional adobe bricks are sun-dried, not fired, mixtures of clay, sand, gravel, water, and straw or grass. (For fired or stabilized bricks that are made to look like adobe, see the previous section on concrete masonry). The traditional production method produces a brick that swells and shrinks depending on its fluctuating water content. Adobe walls are relatively fragile and must be sealed with a protective covering.

The low compressive strength of the adobe bricks leads to the use of very thick (10–30 in.) walls that are seldom more than two stories high. These thick solid walls provide significant thermal mass. The thermal mass effect is especially important because the steady-state thermal resistance of these walls is not very great. A 14-in. thick wall would have a thermal resistance of from R2 to R10, depending on the density and water content of the wall. Insulation can of course be added to the interior or exterior face of the wall if covered with an appropriate coating.[17]

Exterior Insulation Finish Systems

The EIFS can be placed on a wood- or steel-framed wall or a masonry wall. In this system, a layer of polystyrene board insulation, one or more inches thick, is applied to the wall which is then covered with multiple coatings that produce the finished appearance of stucco. This wall system provides a continuous cover of insulation that breaks all the thermal short circuits associated with the framing materials and has the thermal advantage previously shown in Figure 4 for foam sheathing. The EIFS system has been and continues to be one of the most popular exterior claddings for commercial and institutional buildings. However, residential construction jobs often do not have the same high level of quality control and job oversight. This difference led to moisture-related problems in residential EIFS walls where moisture seeped into inadequately sealed window openings and became trapped within the walls.

Subsequent building failures led to the development of two classes of EIFS: barrier and drainable. Both classes are used in the commercial building class, but only the drainable system is allowed for residential construction in many locations. In the barrier class, the outer finish layer is designed to be the one and only weather-resistive barrier on the wall. In the drainable class, some form of spacer is placed between the polystyrene board insulation and a second weather-resistive barrier is located atop the wall's structural sheathing layer. This drain plane allows any moisture that seeps into the wall to drain safely out of the wall.[3]

Structural Insulated Panels (SIPS)

The SIPS walls are made by sandwiching foam insulation, typically 4–6 in. thick, between two sheets of a wood product, thus providing structural support, insulation, and exterior sheathing in a single panel. Each manufacturer specifies the proper method and materials to use when joining adjacent panels. Because the panels themselves have such a high thermal resistance, these joints are critical in maintaining a high thermal resistance for the whole wall. Walls have been measured with *R*-values of about R14 for a wall with a 3.5-in. thick foam core and about R22 for a wall with a 5.5-in. thick foam core.[14] The wall sections are relatively light and the exterior walls of a building can often be completed in a day. Two variations of the system include the use of metal sheets for the exterior skin and the use of alternative insulation materials.

Straw Bale Walls

Exterior walls can be made from stacked straw bales. The straw is a natural insulation material, but must be protected from the weather by an exterior surface, often a stucco-type finish applied over a wire mesh. Gypsum board can be used on the interior surface using a number of methods.[3] Experiments have shown that it is very important not to leave any air space between the straw and the surfacing material. Such air gaps work in concert with the hollow straw tubes to set up convection loops that may cut the overall thermal resistance from around R50 down to around R16.[18]

Windows

Transparent glass was first used for windows during Roman times. Technology has gradually advanced, improving the smoothness, strength, and clarity, and increasing the maximum size of manufactured glass. More recently, double-pane windows became popular as energy costs rose and the use of air conditioning became more common. By the mid-1990s, nearly 90% of all residential windows sold had two or more layers of glass.[19] The windows in older homes are being upgraded to double-pane windows as well; from 2000 to 2001, about nine million homeowners spent $15 billion on window and door replacements.[20] From 1990 to 2003, the replacement window market made up from 38% to 52% of the total window market.[21] However, in 2004, two-thirds of residential buildings and about half of the nonresidential building stock still had single-pane windows. The potential energy savings in this population is great, especially with the new selective coating techniques for glass.

Energy Transport

The primary function of windows is to let light and air into a building; so it is not surprising that the windows can be very challenging from an energy conservation point of view. Energy travels through a window by all three energy transfer phenomena, as shown in Figure 6.

The radiation portion of the energy transport includes short-wave radiation, or ultraviolet, that tends to fade the colors in fabric and paint; visible radiation, or light, that we desire; and long-wave radiation or heat. Long-wave radiation is a normal part of the solar spectrum. It helps to heat our buildings during the winter but increases our air conditioning load during the summer. Some values for the solar heat gain through a few prototypical windows are shown in Figure 7. Long-wave radiation also occurs between any two surfaces, traveling from the warmer surface to the cooler surface. Because the outside environment is warmer than the inside surfaces in the summer and the reverse in the winter, this

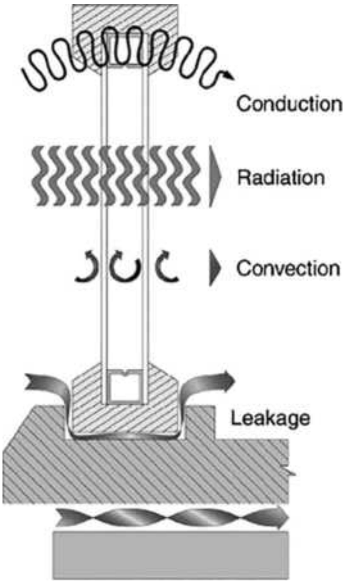


FIGURE 6 Energy traveling through windows.

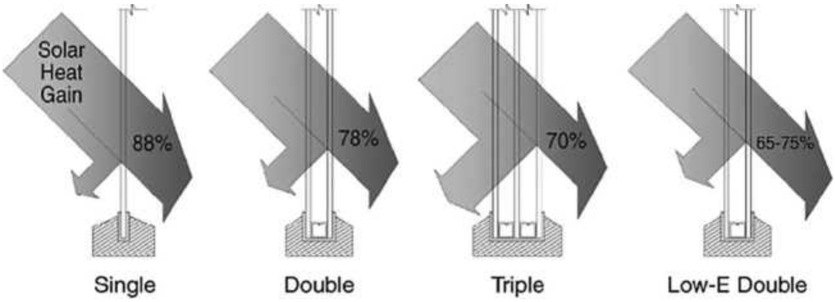


FIGURE 7 Solar heat gain through various windows.

long-wave radiation travels through the window and increases both our winter heating and summer cooling energy use. Newer windows have a “low-e” coating to reduce long-wave radiation and thus reduce the energy losses.

The conduction portion of the energy transport includes the heat that travels through the window frame and heat conducted through the glass pane(s) and through any gas between the panes. The energy that travels through the window frame and sashes is a complex function of the frame material(s), the shape (including any hollow cavities), and the exposed surface area. Each material used in windows has positive and negative qualities. In general, wood has a lower thermal conductivity than metal or plastic. However, wood is more likely to change in shape during its lifetime, so that sealing air leakage out of a wood window over a long time period may be more problematic. Also, wood is more susceptible to moisture damage and must be kept painted or varnished. Metal frames will conduct more heat than wood, but the exposed area can be reduced because the metal is a stronger material. Some of the best performance is found in windows that combine multiple materials. For example, a metal- or vinyl-clad window frame will require less maintenance than a wood window, but will conduct less heat than a solid metal or solid plastic frame. Some special gases, such as argon, have a lower thermal conductivity than air and are sometimes used to fill the gap between panes in multiple-pane windows.

The convection portion of the energy transport includes exterior air movement, or wind, across the glass surface; gas movement between the panes in a multilayer window; and air movement across the interior face of the window. The gas movement within a multilayer window is determined by the thickness of the gap, the height of the window, the gas temperature, and the temperature difference between the two panes of glass. Closed draperies or shades can reduce the air flow across the inside pane of glass.

The energy carried by air leakage falls into two major categories. As Figure 6 shows, some air leaks around the window frame itself. This leakage path should be sealed when the window is installed, although caulking around the frame of an existing window can also reduce the air leakage. Air also leaks through any moving joint in a window. These joints must be sealed by weather stripping or special gaskets that are built into the windows themselves.

Storm windows have been used for a long time and were very popular in northern climates before the introduction of multiple-pane windows. Storm windows require annual installation and storage and there are more glass surfaces to clean. Storm windows can be mounted on either the inside or outside of the window frame.[22] Tests have shown that the energy savings can be substantial when storm windows are added to an existing single-pane window. However, replacing an existing single-pane window with a modern double-pane window will save more energy than the addition of a storm window.[23,24]

Window Rating Systems

Considering the complexities of energy transport through windows, it can be difficult for consumers to compare one window with another. Fortunately, there are two important tools available to help with window selection. The National Fenestration Rating Council (NFRC) has developed a rating system that includes a standard label.[25] The U.S. Departments of Environmental Protection and Energy have cooperated in the production of an Energy Star label for windows.[26]

The NFRC label (Figure 8) shows four values: the *U*-factor, the solar heat gain coefficient (SHGC), the visible transmittance, and air leakage. Manufacturers may also choose to show a value for condensation resistance. The two most important factors used to rate the energy efficiency of a window are the

NFRC
National Fenestration Rating Council®
CERTIFIED

World's Best
Window Co.
Millennium 2000+
Vinyl-Clad Wood Frame
Double Glazing • Argon Fill • Low E
Product Type: **Vertical Slider**

ENERGY PERFORMANCE RATINGS

U-Factor (U.S./I-P)	Solar Heat Gain Coefficient
0.35	0.32

ADDITIONAL PERFORMANCE RATINGS

Visible Transmittance	Air Leakage (U.S./I-P)
0.51	0.2

Manufacturer stipulates that these ratings conform to applicable NFRC procedures for determining whole product performance. NFRC ratings are determined for a fixed set of environmental conditions and a specific product size. NFRC does not recommend any product and does not warrant the suitability of any product for any specific use. Consult manufacturer's literature for other product performance information.
www.nfrc.org

FIGURE 8 National Fenestration Rating Council (NFRC) label.

U-factor and the solar heat gain coefficient. The *U*-factor is the inverse of the *R*-value (the label that is quoted for insulation), so that a lower *U*-factor indicates a slower rate of heat transfer for any given temperature difference.[27] The NFRC *U*-factor ratings for windows sold nowadays range from 0.2 to 1.2.[28] The SHGC ranges from 0 to 1 and measures how much heat from the sunlight incident upon a window will enter the building. Windows with lower SGHC ratings do a good job of blocking this heat.

The Energy Star label is available to windows that have been certified and labeled by NFRC and that meet special standards. The standards vary among the four NFRC climatic regions because of the tradeoff between desirable winter heating and undesirable summer heating. The four regions used by the Energy Star program for windows are shown in Figure 9 and Table 1 shows the required *U*-factors and SHGCs for each region.[29]

FIGURE 9 Energy Star's four regions used for window rating program.

TABLE 1 Energy Star Window Criteria

Climate Zone	U-Factor	Solar Heat Gain Coefficient G-Factor (SHGC)	
Northern	≤0.35	Any	
North/Central	≤0.40	≤0.55	
South/Central	≤0.40	≤0.40	Prescriptive
	≤0.41	≤0.36	Equivalent performance (excluding CA)
	≤0.42	≤0.31	
	≤0.43	≤0.24	
Southern	≤0.65	≤0.40	Prescriptive
	≤0.66	≤0.39	Equivalent performance
	≤0.67		
	≤0.68	≤0.38	
	≤0.69	≤0.037	
	≤0.70		
	≤0.71	≤0.36	
	≤0.72	≤0.35	
	≤0.73		
	≤0.74	≤0.34	
	≤0.75	≤0.33	

Future Improvements

Researchers are working on a portfolio of window designs with automatic energy-saving features. Some forms of "smart" windows with switchable glazing have become commercially available, albeit at a relatively high cost. These windows vary the amount of light and heat transmitted based upon an electric current, which is usually programmed to respond to either the temperature or the amount of sunlight hitting the window. The electrochromic window uses a multilayer electrically conductive film where ions are moved from one layer to another by a short electrical signal. In one layer, the ions allow only 5% of the sunlight through the window; when the ions move to the other layer, 80% of the sunlight is transmitted. This system has the advantage that once the change from one state to another has been made, no electrical energy is required to maintain that state. Another switchable glazing, the suspended particle display (SPD) places a solution containing suspended particles between two glass panes. When an electrical charge is applied, the particles align and light is transmitted through the window. Without an electrical charge, the particles move about randomly, blocking up to 90% of the light. Other switchable windows are designed to provide privacy by changing from transparent to translucent, but are not effective at saving energy.[30,31] Another proposed window design includes sensors that automatically raise or lower a blind enclosed between two panes of glass based upon the outdoor temperature and solar radiation. This design admits solar radiation when it will help heat the house, but lowers the blind to block solar radiation when it will increase the air conditioning load.[32]

Conclusion

Walls and windows are often selected to achieve a desired appearance. Considering today's emphasis on energy conservation and overall sustainability, it is important to consider their thermal characteristics as well. Selecting more energy-efficient walls and windows will permit the building designer to specify a smaller heating and cooling system, so that often the total building cost is little more than that of a standard building. When you consider the reduced cost of heating and cooling the building during its lifetime, any added investment during construction is returned many times over.

Other Useful Guides

Graphic Guide to Frame Construction: Walls, by *Rob Thallon,* Published by Taunton Press, 2000, ISBN 1–56158-3537, # 070470, (http://www.taunton.com/finehomebuilding/pdf/Framing%20Walls.pdf) and http://www.taunton.com/finehomebuilding/pdf/Grading%20and%20Drainage.pdf.

References

1. Available at http://www.energycodes.gov/ (accessed February 12, 2007).
2. Ontario Association of Architects, *The Rain Screen Wall,* available at http://www.cmhc-schl.gc.ca/en/inpr/bude/himu/coedar/loader.ffm?url=/commonspot/security/getfilecfm&PageID=70139, other related publications from the Canadian Mortgage Housing C. available at http://www.-cmhc-schl.gc.ca/en/ (accessed February 12, 2007).
3. Toolbase Services, a consortium including the U.S. Department of Housing and Urban Development and the National Association of Home Builders (NAHB) provides a wealth of information about current and future building materials and housing systems, NAHB Research Center, Maryland, 2005, available at http://www.toolbase.org (accessed February 13, 2007).
4. Kosny, J.; Kossecka, E.; Desjarlais, A.O.; Christian, J.E. In *Dynamic Thermal Performance of Concrete and Masonry Walls,* Thermal Envelopes VII, Conference Proceedings, Clearwater Beach, FL, December 1998; ASHRAE Special Publications, 1998.

5. Lippke, B.; Wilson, J.; Perez-Garcia, J.; Bowyer, J.; Meil, J. CORRIM: life-cycle environmental performance of renewable building materials. Forest Prod. J. **2004**, *54* (6), 8–19, available at http://www.corrim.org/reports/pdfs/FPJ_Sept2004.pdf, (accessed February 13, 2007).
6. Brown, M.A.; Southworth, F.; Stovall, T.K. *Towards a Climate-Friendly BuildtEnvironment*; Pew Center on Global Climate Change: Arlington, VA, 2005; available at http://www.pewclimate.org/document.cfm?documentIDZ469 (accessed February 13, 2007).
7. Stovall, T. K., and Karagiozis, A., Airflow in the Ventilation Space Behind a Rain Screen Wall, Thermal Envelopes IX, Conference Proceedings, Clearwater Beach, FL, December 2004; ASHRAE Special Publications, 2004.
8. U.S. Department of Energy, Technology Fact Sheet: Wall Insulation, October, 2000, http://www.eere.energy.gov/buildings/info/documents/pdfs/26451.pdf (accessed February 13, 2007).
9. Kosny, J.; Mohiuddin, S. A. Interactive Internet-Based Building Envelope Materials Database for Whole-building Energy Simulation Programs, Thermal Envelopes IX, Conference Proceedings, Clearwater Beach, FL, December 2004; ASHRAE Special Publications, 2004.
10. Toolbase Services, a consortium including the U.S. Department of Housing and Urban Development and the National Association of Home Builders (NAHB) provides a wealth of information about current and future building materials and housing systems, NAHB Research Center, Maryland, 2005, found at www.toolbase.org.
11. Kosny, J. Steel Stud Walls; Breaking a Thermal Bridge. Home Energy July 2001, http://www.homeenergy.org/article_preview.php?id=249&article_title=Steel_Stud_Walls:_Breaking_the_Thermal_Bridge (accessed February 13, 2007).
12. ASHRAE Handbook of Fundamentals, 2005 Chapter 25.
13. Kosny, J.; Petrie, T.; Gawin, D.; Childs, P.; Desjarlais, A.; Christian, J. Thermal Mass—Energy Savings Potential in Residential Buildings, August 2001, http://www.ornl.gov/sci/roofsCwalls/research/detailed_papers/thermal/index.html (accessed February 13, 2007).
14. Advanced Wall Systems Hotbox Test R-value Database, Oak Ridge National Laboratory, http://www.ornl.gov/sci/roofsCwalls/AWT/Ref/TechHome.htm.
15. NAHB Research Center, ToolBase Services, PATH Technology Inventory, Precast Concrete Foundation and Wall Panels, Upper Marlboro, MD, 2005, http://www.toolbase.org/Technology-Inventory/Foundations/precastconcrete-panels (accessed February 13, 2007).
16. Precast/Prestressed Concrete Institute, Single Family Housing, http://www.pci.org/markets/projects/single.cfm (accessed February 13, 2007).
17. Thermal conductivity estimated considering the reported thermal conductivity of unfired clay bricks found at http://www.constructionresources.com/products/ and the measured thermal conductivity of a wall built from rammed earth bricks reported at http://www.ornl.gov/sci/roofsC-walls/AWT/Ref/TechHome.htm (accessed February 13, 2007).
18. Christian, J.E.; Desjarlais, A.O.; Stovall, T.K. In *Straw Bale Wall Hot Box Test Results and Analysis,* Conference Proceedings, Clearwater Beach, FL, December, 1999; Thermal Performance of the Exterior Envelopes of Buildings VII, 1999; 275–285.
19. Carmody, J.; Selkowitz, S.; Heschong, L. *Residential Windows, A Guide to New Technologies and Energy Performance;* W.W. Norton and Company: New York, 1996.
20. Joint Center for Housing Studies of Harvard University, *Remodeling, Measuring the Benefits of Home Remodeling;* Harvard University: Cambridge, MA, 2003.
21. U.S. Department of Energy, 2004 Buildings Energy Databook, available at http://buildingsdatabook.eren.doe.gov/default.asp
22. U.S. Department of Energy, A Consumer's Guide to Energy Efficiency and Renewable Energy, Storm Windows, available at http://www.eere.energy.gov/consumer/your_home/windows_doors_skylights/index.cfm/mytopicZ13490.
23. Klems, J.H. In *Measured Winter Performance of Storm Windows,* ASHRAE 2003 Meeting, Kansas City, MO, June 28–July 2, 2003; LBNL-51453.

24. Turrell, C. Storm windows Save Energy, Home Energy Magazine Online, July/August 2000, available at http://homeenergy.org/archive/hem.dis.anl.gov/eehem/00/000711.html.
25. NFRC. Various informative fact sheets regarding window issues. Available at http://www.nfrc.org/factsheets.aspx (accessed February 13, 2007).
26. Available at http://www.energystar.gov/index.cfm?cZwindows_doors.pr_windows.
27. The *U-factors* used by the NFRC and the Energy Star program have units of Btu/h-ft^2-°F.
28. Available at http://www.nfrc.org/label.aspx (accessed February 13, 2007).
29. Available at http://www.energystar.gov/index.cfm?cZwindows_doors.win_elig.
30. Fanjoy, R. Smart Windows Ready to Graduate, HGTV Pro. com, available at http://www.hgtvpro.com/hpro/dj_construction/article/0,2619,HPRO_20156_3679648,00.html (accessed February 13, 2007).
31. Toolbase Services, a consortium including the U.S. Department of Housing and Urban Development and the National Association of Home Builders (NAHB) provides a wealth of information about current and future building materials and housing systems, NAHB Research Center, Maryland, 2005, Available at http://www.toolbase.org.
32. Kohler, C.; Goudey, H.; Arasteh, D. A First-Generation Prototype Dynamic Residential Window, LBNL-56075, Lawrence Berkeley National Laboratory, Berkeley, CL, October 2004, available at http://www-library.lbl.gov/docs/LBNL/560/75/PDFZLBNL-56075.pdf (accessed February 13, 2007).

33

Energy: Waste Heat Recovery

Martin A. Mozzo, Jr.

Introduction

In many industrial and commercial energy applications, only a portion of the energy input is used in the process. The remainder of the useful energy is rejected to the environment. This rejected energy may potentially be recaptured as useful energy through waste heat recovery. Not all rejected energy can be recovered due to quality, usefulness in a host's load profile, and/or economical reasons that may make its recovery infeasible. This entry will serve as a guide to waste heat recovery in order to provide a framework for the energy engineer to develop waste heat recovery projects. It will discuss the concept of quality vs quantity, engineering concerns in waste heat recovery, sample calculations for waste heat recovery, and the types of waste heat recovery equipment that can be used.

Concept of Quality vs. Quantity

While at first glance, a waste heat recovery project may appear feasible, this may not always be the case. The concept of quality vs. quantity plays an important role. For the purposes of this entry, there are three classifications of waste heat. These are: (1) high-grade waste heat, generally 1000°F and above; (2) medium-grade waste heat, generally in the range of 400°F–1000°F; and (3) low-grade waste heat, generally below 400°F. Typically, the higher the grade of waste heat, the better the application for a successful and economical waste heat recovery project. This is referred to as the "quality" of the waste heat. It is better to have marginal amounts of high quality waste heat than large quantities of lower-grade waste heat.

There are numerous reasons why higher quality waste heat recovery provides a better application. One reason is that the waste heat stream has a much higher temperature than the recovery medium, resulting in higher temperature differences, and, thus, waste heat recovery equipment sizes can be smaller. This is a result of the higher heat transfer efficiency between the waste heat stream at a higher temperature to a lower temperature heat recovery medium. Second, the recovery medium can take on many forms, namely air, steam, or hot water, whichever is best suited for the host facility, whereas, with lower quality heat recovery, the form may be very limited, especially with low- grade waste heat. For example, if you have a waste heat stream that is at 400°F, steam is not a viable heat recovery medium, as the temperature difference will be too low. Finally, cost and savings become important issues in regard to the quality of the waste heat recovery system. Generally, the higher the quality of waste heat, the lower the capital costs for waste heat recovery equipment. This is because better heat transfer efficiencies will be realized with heat recovery equipment. Additionally, the higher the quality of the waste heat stream, the higher the savings usually are. Thus, with lower costs and higher savings, simple payback periods will be shortened, a result which is generally well- received by management.

A word of caution should be offered regarding the quality of waste heat. There are vendors who will reduce the grade of the available waste heat so that their equipment can be used in a heat recovery application. One such vendor manufactured only low temperature heat transfer equipment, which he was trying to sell as waste heat recovery equipment. His proposals included the dilution of high-grade waste heat streams by mixing the high temperature waste heat stream with ambient air in order to bring the temperature down to levels that his equipment could handle. His argument was that the total Btu content of the waste heat stream was not changed, just the temperature. Unfortunately, in reducing the temperature and quality of the waste heat stream, the size of the heat recovery equipment had to be increased to recover the waste energy, and the type of waste heat recovery medium was limited. This results in higher costs, lower savings, and longer-term payback periods.

There is one instance where the dilution of a waste heat stream may be justified. Steel is usually used in the manufacture of waste heat equipment, and there is an upper temperature limit for the waste stream where steel can be used. A substitution of other materials, for example, ceramics, can be made; however, the capital costs may become too prohibitive. In this instance, the dilution of the waste heat stream may be justified to bring the waste stream temperature down for the safe operation of the waste heat recovery equipment. It should, however, never be diluted simply to enable the use of a particular piece of heat recovery equipment. The primary reason is that by diluting the waste stream, energy savings may decrease and equipment costs will definitely increase; thus, payback periods will increase, perhaps significantly.

Engineering Considerations

There are several engineering factors that must be evaluated when considering and designing a waste heat recovery system. In terms of steps, these are: (1) quantifying the waste heat stream; (2) determining the value of the waste heat stream; (3) evaluating the best form of heat recovery for the host facility; (4) determining the host site heat load profile; (5) determining the grade of waste heat; (6) determining the cleanliness and quality of the waste stream; and (7) selecting the proper waste heat recovery equipment by considering size, location, and maintainability.

The first step that should be executed is to quantify the waste heat stream by determining how many Btu/h are in the waste stream. The equation to calculate this is as follows:

$$Q = M \times \text{specific heat} \times \text{delta temp} \tag{1}$$

where Q = total heat flow rate of waste stream in Btu/h; M = mass flow rate in Lb/h; specific heat = for air, 0.24 (Btu/Lb/°F); delta temp = $(T_{\text{upper}} - T_{\text{lower}})$ in °F.

The specific heat (Cp) changes as the temperature of the air rises. For example, at 1000°F, Cp is 0.26. Since this value is higher than that of Eq. 1, using the lower value, rather than correcting for temperature, provides a conservative calculation of the total heat rate in the waste stream.

The mass flow rate (M) is calculated as follows:

$$M = \rho \times V \tag{2}$$

where M = mass flow rate in Lb/h; ρ = density of the waste stream in Lb/ft^3; V = volumetric flow rate in ft^3/h (in standard cubic feet).

Note that the value of Q is not the total amount of waste heat that will be recovered, but, rather, the total amount of waste heat that is ideally available for recovery. Not all of this waste heat will be recovered, or even can be recovered. The total amount that will be recovered will be determined by numerous other factors, such as the cleanliness of the waste stream and the form of recovery (i.e., high pressure, superheated steam, saturated steam, or hot water). This step is necessary in determining if there are sufficient volumes available for waste heat recovery.

One common mistake committed when quantifying the amount of waste heat available is assuming values for flow and/or temperatures without making measurements. Too often, actual conditions vary from assumed conditions, a variation that can often cause disastrous results in a waste heat recovery project. The cost of making actual field measurements is a small price to pay to obtain reliable data.

After the quantity of the waste stream is correctly determined, step two is to find the dollar value of the waste heat stream to determine how much capital cost a potential project can bear, or if it even makes economical sense. In all waste heat recovery projects, the heat recovered will displace a medium, such as steam, which would have to be generated using another piece of equipment, such as a boiler. This equipment likewise has a related efficiency, and the heat output is always less than the heat input. To determine the dollar value of the waste heat stream, use Eqs. 3 and 4 below:

$$\text{Value} = Q \times \text{unit cost} \tag{3}$$

where value = the dollar value of the waste heat stream, per hour; Q = total heat flow rate of waste stream in Btu/h as calculated from Eq. 1; unit cost = unit cost of the waste stream in dollars/Btu.

$$\text{Unit cost} = \frac{\text{fuel cost}}{\text{efficiency}} \tag{4}$$

where unit cost = dollars/Btu; fuel cost = cost for fuel displaced in dollars/Btu; efficiency = efficiency of unused equipment. For example, a steam boiler @ 75%.

The third through fifth steps tend to overlap and be interrelated, and are discussed together. In order for waste heat recovery to be acceptable to a host facility, its use must be consistent with current energy usage at the facility. To best apply waste heat recovery, the engineer needs to examine current energy usage patterns, heat loads for the site, and the quality of the waste heat that is available. The load factor of the waste heat recovery stream is the first thing to determine. For example, the waste heat recovery stream may be exhaust air flows from a process furnace that is periodic in nature, rather than a continuous operation. In this case, the best potential would be to return the waste heat to the process in some manner that allows the usage to follow the waste heat stream generation. Examples include utilization in a drying operation or as preheated combustion air for the process furnace. In another situation, the process furnace may run continuously and be of a high enough grade to generate steam; however, if the host facility only uses steam for heating in the winter, the loads do not match and, thus, steam is not a good recovery medium. On the other hand, there may be a need for large quantities of hot water in the facility, and the loads may match up. A thorough evaluation of both the source and operation of the waste heat and potential uses to recover the waste heat must be performed.

The sixth step, determining the cleanliness and quality of the waste stream, is important. If the waste heat stream is the product of the combustion of a natural gas-fired or fuel oil-fired operation, then the gases should be relatively safe for waste heat recovery. The important issue here would be the condensation of acidic liquids in the products of combustion gas stream, especially if fuel oil is used. This condensation would happen if the temperature of the waste stream is allowed to drop to low temperatures, typically below 300°F. The waste stream could also contain a burn-off from the product itself, which could condense on the waste heat recovery equipment and cause blockages. Here, it is best to maintain the exit temperature significantly above the condensation temperature to avoid the formation of acids or other potential hazards from the waste stream.

If the waste heat stream contains the products of combustion and the products of the process, great care must be exercised when utilizing the waste heat. The selection and design of waste heat recovery equipment could lower the waste heat exhaust temperature below acceptable levels, resulting in condensation or blockage problems with the heat recovery equipment. When unsure of the condensation temperatures and the effect on the heat recovery equipment, simple tests must be conducted to obtain this information.

Some actual examples wherein the waste heat stream was dirty and created problems follow. The first was an air-to-air heat recuperator used in a process operation. The products of combustion gases were dirty, creating a buildup on the heat exchanger. A soot blower was used to remove the buildup on a periodic basis; however, this was not totally successful. During actual heat recovery operations, the soot blowers could not remove the buildup adequately. A solution to this problem occurred accidentally. The operation was a 24 h/five days a week, and by mistake one weekend, the soot blower was left on, while the waste heat recovery equipment was shut off. Because the buildup and the heat exchanger were composed of different materials, the heat exchanger and the buildup thermal contraction rates were different. As a result, the soot blower was able to break up the buildup on the heat exchanger surfaces. This method was then used to clean the heat exchanger each week. The efficiency of the heat exchanger did decrease somewhat during the week, though not significantly. During the weekend, the buildup was blasted off the heat exchanger surface, thus increasing the heat exchanger's efficiency at the beginning of another week.

The second example did not turn out as well. In this scenario, the quality of the waste heat stream was low, and the products of combustion, combined with product burn-off, created a dirty waste stream. As a result, a serious problem quickly developed. With the waste heat recovery system in operation, gas products in the waste heat stream condensed on the heat exchanger surfaces, and within a very short period of time, the heat exchanger was completely blocked. The only corrective action was to remove the waste heat recovery equipment from service and steam clean it, a very costly operation that eventually resulted in the failure of the project.

The seventh step in the design of a heat recovery system is the selection of the waste heat recovery equipment. Descriptions of several different types of equipment are given in another entry in this encyclopedia. Selection of the proper piece of equipment will be based on the quality of available waste heat; the recovery fluid that can be generated, i.e., steam, hot water, or air; the location of the equipment; and the maintenance capabilities of the host facility. While the first two criteria should be obvious, the last two are sometimes neglected.

The location where the equipment is to be placed is important. It obviously should be close to the waste heat stream, yet should not interfere with the other equipment involved in the generation of waste heat, for example, a process furnace. The equipment also must not be squeezed into an area where maintainability is an issue. Sufficient access for maintenance and overhaul must be included since these pieces of equipment will require periodic maintenance. Accessibility to perform maintenance is not just desirable, but mandatory.

Sample Calculations

This section describes some sample calculations that are used in the design of waste heat recovery systems. The first step is to calculate the amount of waste heat that is available for recovery. This is done using Eq. 1.

Example—Waste heat is available in the form of hot products of combustion from a natural gas-fired furnace. Assume that there are no contaminants in the waste heat stream. Measurements indicate that there are 14,000 available actual cubic feet per minute (ACFM) flowing from a process at 1000°F. The actual volumetric flow must first be converted to standard cubic feet per minute (SCFM) at 60°F, using the following equation:

$$\text{SCFM} = \text{ACFM} \times (T_{\text{absolute}} + 60)/(T_{\text{absolute}} + T_{\text{actual}}) \tag{5}$$

where SCFM = standard cubic feet per minute; ACFM = actual cubic feet per minute; T_{absolute} = 460°F; T_{actual} = actual gas temperature.

Using the data available here,

$$\text{SCFM} = 14{,}000 \times \left(460 + 60\right) / 460 + 1000\big) = 4{,}986\ ft^3/\text{min}$$

Since the product of combustion is essentially air, a good approximation of the density of the gases is similar to that of air at standard conditions, or $\rho = 0.074$ Lb/ft^3.

Using Eq. 2, we can calculate the mass flow rate for the waste stream as follows:

$$M = \rho \times V = 0.074 Lb/ft^3 \times 4{,}986 ft^3/\text{min} \times 60\text{min}/h = 22{,}139\ Lb/h$$

The next step is to calculate how much waste heat is available for transfer to the heat recovery system. The upper temperature has been given as 1000°F. The lower temperature will be determined by the medium used for heat recovery, that is, steam, hot water, or air, as well as the cleanliness and condensation temperature of the waste stream.

To continue the sample calculations, we will use three different applications of heat recovery mediums: 125 PSIG steam, the hot water required at 350°F, and pre-heated combustion air at whatever final temperature is available. Each of these applications is analyzed separately.

The first heat recovery application medium is saturated steam at 125 PSIG (140 PSIA). From steam tables, we find the enthalpy for saturated liquid (h_f) to be 324.82 Btu/Lb. and for saturated vapor (h_g) 1193.0 Btu/Lb. Thus, we need 868.18 Btu for every pound of steam (1193.0–324.82). The temperature of steam at this pressure is 353°F. For this temperature of steam, we can take the waste stream down to approximately 400°F without the risk of condensation (since condensation usually does not occur above about 325°F). Using Eq. 1, we can calculate the available waste heat as follows:

$$Q = 22{,}139 Lb/h \times 0.24\ Btu/Lb/^\circ F \times \left(1000 - 400\right) = 3{,}188{,}041\ \text{Btu/h} = 3.188041\text{MMB}tu/h$$

This waste heat stream's value can be calculated using Eqs. 3 and 4. Assuming we have natural gas priced at \$8.00/MMBtu in a steam boiler with an efficiency of 75%, the value becomes:

$$\text{Unit cost} = \$8.00/\text{MMBtu}/0.75 = \$10.67/\text{MMBtu}$$

$$\text{Value} = \$10.67/\text{MMBtu} \times 3.188{,}042\ \text{MMBtu/h} = \$34.00/\text{h}$$

To obtain steam mass flow, we must use Eq. 6 below:

$$M_s = \frac{Q}{\left(h_g - h_f\right)} = \frac{3{,}188{,}041\ \text{Btu/h}}{(1193.0 - 324.82)\text{Btu/Lb}} = 3{,}672\ Lb/h \tag{6}$$

The second heat recovery application medium will be hot water leaving the heat recovery equipment at 250°F, using the same waste stream. Water will be supplied to the heat recovery equipment at 200°F. With the temperatures involved, we can take the waste stream down to approximately 300°F, provided this low temperature does not cause any condensation issues. For the purposes of this example, we have determined that condensation occurs below 325°F. We will use 350°F as our lower limit. Using Eq. 1, we can calculate the available heat as:

$$Q = 22{,}139\ Lb/h \times 0.24\ Btu/Lb/^\circ F \times \left(1000 - 350\right)$$

$$= 3{,}453{,}711 Btu/h$$

In this case, we need to determine the mass flow rate, expressed as gallons per minute (GPM), of hot water we can heat from 200 to 250°F. Eq. 7 below provides the methodology:

$$Q' = 500 \times V \times \left(T_{\text{upper}} - T_{\text{lower}}\right) \tag{7}$$

where Q' = total heat recovered in Btu/h; V = volumetric flow rate of water in GPM. 500 is the constant used to convert GPM to Lb of water per hour.

To calculate the volume of water required for this system, we calculate:

$$V = \frac{Q'}{500 \times \left(T_{\text{upper}} - T_{\text{lower}}\right)}$$

$$= \frac{3{,}453{,}711\ Btu/h}{500 \times (250 - 200)} = 138\ \text{GPM}$$

Thus, our waste heat recovery system will generate approximately 138 GPM of hot water raised from 200 to 250°F.

The third heat recovery application medium is an air- to-air recuperation system. The waste stream is the same as in the previous example, providing 3,453,711 Btu/h (assuming that 350°F will be the lower limit that we can go to without condensation issues). The recovery medium will be air applied as pre-heated combustion air with an input temperature of 75°F. In this system, we need to provide a set volume of combustion air to the waste heat recovery equipment, so we are interested in finding the final temperature of the pre-heated combustion air.

Using equation

$$Q = M \times 0.24 \times \left(T_{\text{upper}} - T_{\text{lower}}\right)$$

$$3{,}453{,}711\ Btu/h = 50{,}000 Lb/h \times 0.24 Btu/Lb - {}^\circ F \times \left(T_{\text{upper}} - 75\right)$$

Solving for T_{upper} we get 362.8°F

Heat Recovery Equipment

There is a variety of equipment manufactured and/or sold as heat recovery equipment. The following is a list of some more common types of equipment.

1. Waste heat steam recovery—This piece of equipment can be either a water-tube or fire-tube boiler that uses the hot waste heat gas stream to heat boiler feed water to generate either low-pressure or high-pressure steam. Typically, this piece of equipment is called a heat recovery steam generator (HRSG). The unit can be a heat recovery-specific piece of equipment; however, some boiler manufacturers will sell their boilers without a burner package and call them HRSGs. This type of boiler can lose some of its effectiveness as a steam generator; however, it still is an acceptable piece of equipment. The reason for the decrease in recovery effectiveness is solely based on the possibility that the waste heat stream temperature will be somewhat lower than the burner flame temperature.
2. Recuperator—This piece of equipment is generally an air-to-air heat exchanger that transfers heat from the waste heat gas stream to air on the recovery side of the heat exchanger. This air can be used as pre-heated combustion air, or host make-up air in the facility.
3. Shell and tube heat exchanger—This piece of equipment consists of a bundle of tubes within a steel shell. It is usually used for water-to-water heat recovery. One of its uses may be to recover

heat from a boiler condensate blow down on the shell side to heat up boiler feed water on the tube side.

4. Fin-tube heat exchanger—This type of heat exchanger uses air (usually the waste heat stream) blowing across finned coils that contain water. A typical application for this type of heat exchanger is using boiler products of combustion gases to preheat boiler feed water. Plugging of the finned coils could be a problem if the fins are closely spaced, the exhaust gases are dirty, or the condensation temperature (of the gases) is approached.
5. Plate and frame heat exchanger—This piece of equipment consists of two frames sandwiching thin plates. It is usually used with fluids of relatively low temperature, which flow through alternate plates. Since the plates are thin, heat transfer is usually good; however, both fluid streams need to be relatively clean or the exchanger will plug up.
6. Heat wheels—These are typically used in low temperature applications, such as the exhaust of environmental (space) air coupled with the introduction of outside air. The two ducts, outside and exhaust air, must be adjacent to one another so that the wheel turns through both ducts. The wheel will turn at low revolutions per minute (RPM) collecting the waste heat or cooling, depending on what season it is, and exchange the heat or cooling to the outside air being introduced to the facility.
7. Heat pipes—This equipment consists of a pipe heat exchanger with the interior containing a coolant. The coolant is alternately vaporizing and condensing between an exhaust and outside air stream, exchanging cooling and heating between the two air streams. The exhaust air and outside air ducts must reside together in the same manner as a heat transfer wheel.
8. Run-around coils—These are similar to heat pipes in operation. They cool or heat exhaust air and outside air streams. Their typical construction is a finned water coil with air blown across the coil. The advantage of the run-around coil is that the exhaust air and outside air ducts can be physically separated. The disadvantage is that a pump, with pumping power, is required. Additionally, if the coils are subject to freezing conditions, then a water/glycol solution must be used. This solution will decrease the effectiveness of the heat transfer.

Summary

The purpose of this entry has been to provide the energy engineer with some general guidelines to applying waste heat recovery. There are numerous applications in the industrial, institutional, and commercial sectors where waste heat recovery can be used cost effectively. The best applications are situations wherein the waste heat stream is of high quality, produced for many hours of the year, preferably 24/7, with a heat load that matches the waste heat availability. Care must be taken by the engineer to ensure that the waste heat stream will not block or destroy the waste heat recovery equipment.

There are also numerous pieces of equipment that can be used for waste heat recovery. These include steam boilers, hot water boilers, recuperators, coil heat exchangers, shell and tube heat exchangers, plate and frame heat exchangers, heat transfer wheels, heat pipes, and run-around coil systems.

Bibliography

Thumann, A.; Metha, D.P. *Handbook of Energy Engineering*; The Fairmont Press, Inc.: Lilburn, GA, 1997.
Turner, W.C. *Energy Management Handbook,* 3rd Ed.; The Fairmont Press, Inc.: Lilburn, GA, 1997.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

34

Fuel Cells: Intermediate and High Temperature

Xianguo Li,
Gholamreza Karimi, and Kui Jiao

Introduction

The three major types of low-temperature fuel cells are discussed in a separate entry entitled "Fuel Cells: Low Temperature." This entry will elaborate on the intermediate- and high-temperature fuel cells including high-temperature proton exchange membrane fuel cells (HT-PEMFCs), phosphoric acid fuel cells (PAFCs), molten carbonate fuel cells (MCFCs), and solid oxide fuel cells (SOFCs). The HT-PEMFCs can be considered as the PEMFCs with membranes and other cell components that can tolerate temperatures above 100°C. The operating temperature range of HT-PEMFC is about between 100°C and 200°C. It is still at the laboratory research level mainly due to the durability issue of the membrane at high temperatures and is expected to be more suitable for stationary rather than automotive applications due to its slower start-up characteristics (higher operating temperature to be reached) than the conventional PEMFC. The PAFC is the most commercially developed fuel cell. It is being used in applications such as hospitals, hotels, offices, and schools. It can also be used in larger vehicles such as buses. The operating temperature range of PAFC is between 160°C and 220°C, and its efficiency is about 40%. The high-temperature MCFCs and SOFCs have advantages over conventional energy-generating systems in terms of reliability, fuel flexibility, modularity, low emission of NO_x and SO_x pollutants, and environmental friendliness. They also show relatively high tolerance to trace levels of impurities in the gas stream. In addition, due to their high operating temperatures typically in the range of 600°C to 1000°C, hydrocarbon fuels such as methane and natural gas can be reformed within the stack, eliminating the need for expensive external reformer systems. The high operating temperatures requires that most applications for these fuel cells be limited to large, stationary power plants. The high-quality heat produced can be used in cogenerated hybrid power systems such as space heating, industrial processing, or even in steam turbines to generate more electricity, improving the system efficiencies to high levels (up to 90%).

High-Temperature Proton Exchange Membrane Fuel Cells

The HT-PEMFCs with operating temperatures higher than 100°C have attracted growing interests in the past decade. By comparing with the conventional PEMFCs operating at around 80°C, the HT-PEMFCs with elevated operating temperatures feature faster electrochemical kinetics, simpler water management (presence of liquid water can be neglected), higher carbon monoxide tolerance, and easier cell cooling and waste heat recovery. The operating temperature range of HT-PEMFC is about between 100°C and 200°C.

Although HT-PEMFCs have many attractive features, technical challenges still remain and are mostly related to the PEM. The durability issue of the PEM at high temperatures is the reason that the HT-PEMFC remains at the laboratory research level. The conventional PEMs (e.g., Nafion membrane) widely used in PEMFCs suffer significant decrement in mechanical strength at the high operating temperature of HT-PEMFCs, and the much lower relative humidity (RH) in HT-PEMFCs than in conventional PEMFCs due to the significantly increased vapor saturation pressure with temperature also results in severe reduction of the proton conductivity of the conventional PEMs. Therefore, developing PEMs with high mechanical strength at temperatures higher than 100°C and with high proton conductivity in anhydrous environments becomes the major challenge, and most of the previous HT-PEMFC-related researches focused on this important issue.

Among the different high-temperature PEMs being developed, polybenzimidazole (PBI) membranes have been recognized as promising PEMs when doped with a strong oxo-acid (e.g., phosphoric acid or sulfuric acid) for HT-PEMFCs. At present, the phosphoric-acid-doped PBI membrane perhaps offers the best combination of durability and proton conductivity for HT-PEMFCs. The durability of HT-PEMFC with this type of PEM is over 6000 hr under continuous operation, and the durability with frequent start-stop cycles is still questionable. The durability issue and the previously mentioned slow start-up make the HT-PEMFC more favorable for stationary rather than automotive applications. Since most of the cell components are similar between PEMFC and HT- PEMFC, it is expected that the cost of HT-PEMFC is similar to conventional PEMFC.

Phosphoric Acid Fuel Cells

The PAFC is the most advanced type of fuel cells and is considered to be "technically mature" and ready for commercialization after nearly 30 years of RD&D and over half a billion dollars expenditure. Therefore, the PAFC has been referred to as the first-generation fuel cell technology. Unlike the alkaline fuel cell systems that were primarily developed for space applications, the PAFC was targeted initially for terrestrial applications with the carbon-dioxide- containing air as the oxidant gas and hydrocarbon-reformed gas as the fuel for electrochemical reactions and electric power generation.

The basic components of a PAFC are the electrodes consisting of finely dispersed platinum catalyst or carbon paper, silicon carbide matrix holding the phosphoric acid, and a bipolar graphite plate with flow channels for fuel and oxidant. The operating temperature ranges between 160°C and 220°C, and it can use either hydrogen or hydrogen produced from hydrocarbons (typically natural gas), ethanol, or methanol as the anodic reactant. In the case of hydrogen produced from a reformer with air as the anodic reactant, a temperature of 200°C and a pressure of as high as 8 atm are required for better performance. PAFCs are advantageous from a thermal management point of view. The rejection of waste heat and product water is very efficient in this system and the waste heat at about 200°C can be used efficiently for the endothermic steam-reforming reaction. The waste heat can also be used as a cogeneration for space heating and hot water supply.

However, the PAFC cannot tolerate the presence of carbon monoxide and H_2S, which are commonly present in the reformed fuels. These contaminants poison the catalyst and decrease its electrochemical catalytic activity. A major challenge for using natural gas reformed fuel, therefore, lies in the removal of carbon monoxide to a level of less than 200–300 ppm. Carbon monoxide tolerance is better at the operating temperature of above 180°C. However, removal of sulfur is still essential. Further, the PAFC has a lower performance, primarily due to the slow oxygen reaction rate at the cathode. Therefore, PAFC is typically operated at higher temperature (near 200°C) for better electrochemical reactivity and for smaller internal resistance, which is mainly due to the phosphoric acid electrolyte. As a result, PAFC exhibits the problems of both high- and low-temperature fuel cells, but possibly none of the advantages of either option.

The PAFC system is the most advanced fuel cell system for terrestrial applications. Its major use is in on-site integrated energy systems to provide electrical power such as in apartments, shopping centers, office buildings, hotels, and hospitals. These fuel cells are commercially available in the range from 24 V, 250 W portable units to 200 kW on-site generators. PAFC systems of 0.5–1.0 MW are being developed for use in stationary power plants of 1–11 MW capacity. The power density of PAFC system is about 200 mW/cm^2, and the power density for 36 kW brassboard PAFC fuel cell stack has been reported to be 0.12 kW/kg and 0.16 kW/L. One representing PAFC system is the PC-25 from the International Fuel Cells in Connecticut (the United States). It costs about \$3000/kW, while the conventional thermal power generation system costs only about \$1000/kW. In fact, present PAFC systems can offer reasonable durability and performance, and the cost is the primary issue that hinders the commercialization of PAFC.

Molten Carbonate Fuel Cells

Introduction

The MCFC is often referred to as the second-generation fuel cell because its commercialization is normally expected after the PAFC. It is believed that the development and technical maturity of the MCFC is about 5–7 yr behind the PAFC. At present, the MCFC has reached the early demonstration stage of precommercial stacks, marking the transition from fundamental and applied R&D towards product development. MCFCs are being targeted to operate on coal-derived fuel gases or natural gas. This contrasts with the PAFCs, as discussed earlier, which prefer natural gas as primary fuel.

The MCFC operates at higher temperature than all the fuel cells described so far. The operating temperature of the MCFC is generally around 600–700°C, typically 650°C. Such high temperature produces high-grade waste heat that is suitable for fuel processing, cogeneration, or combined cycle operation, leading to higher electric efficiency. It also yields the possibility of utilizing carbonaceous fuels (especially natural gas) directly, through internal reforming to produce the fuel (hydrogen) ultimately used by the fuel cell electrochemical reactions. This results in simpler MCFC systems (i.e., without external reforming or fuel processing subsystem), less parasitic load, and less cooling power requirements, hence higher overall system efficiency as well. The high operating temperature reduces voltage losses due to reduced activation and ohmic and mass transfer polarization. The activation polarization is reduced to such an extent that it does not require expensive catalysts as low-temperature fuel cells do, such as PAFCs and PEMFCs. It also offers great flexibility in the use of available fuels, say, through in situ reforming of fuels. It has been estimated that the MCFC can achieve an energy conversion efficiency of 52%–60% (from chemical energy to electrical energy) with internal reforming and natural gas as the primary fuel. Some studies have indicated that the MCFC efficiency of methane to electricity conversion is the highest attainable by any fuel cell or other single pass/simple cycle generation scheme.

Basic Operating Principle

A schematic of a MCFC is illustrated in Figure 1. A MCFC consists of two porous gas-diffusion electrodes (anode and cathode) and a carbonate electrolyte in liquid form. The electrochemical reaction occurring at the anode and the cathode is

$$\text{Cathode}: \frac{1}{2}O_2 + CO_2 + e^- = CO_3^{2-}$$

and the net cell reaction is

$$H_2 + \frac{1}{2}O_2 = H_2O$$

Beside the hydrogen oxidation reaction at the anode, other fuel gases such as carbon monoxide, methane, and higher hydrocarbons are also oxidized by conversion to hydrogen. Although direct electrochemical oxidation of carbon monoxide is possible, it occurs very slowly compared to that of hydrogen. Therefore, the oxidation of carbon monoxide is mainly via the water–gas shift reaction

$$CO + H_2O = CO_2 + H_2,$$

which, at the operation temperature of the MCFC, equilibrates very rapidly at catalysts such as nickel. Therefore, carbon monoxide becomes a fuel, instead of a contaminant as in the previously described low-temperature fuel cells. Direct electrochemical reaction of methane appears to be negligible. Hence, methane and other hydrocarbons must be steam reformed, which can be done either in a separate reformer (external reforming) or in the MCFC itself (the so-called internal reforming).

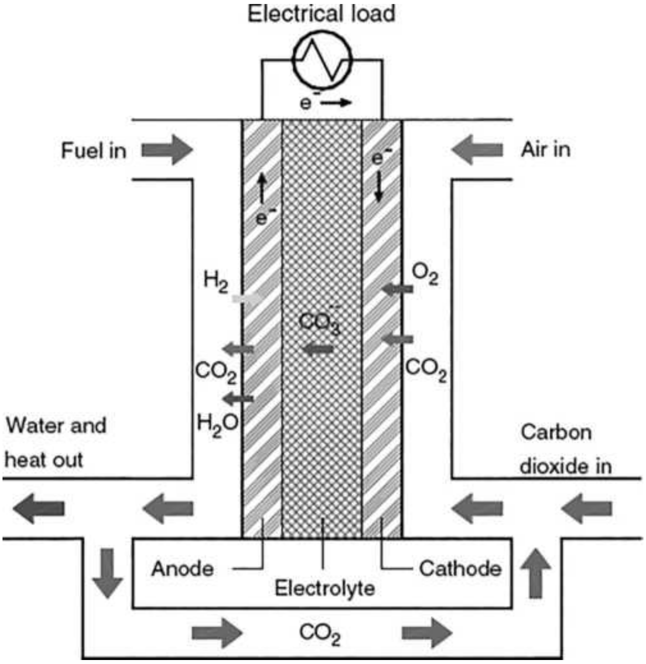


FIGURE 1 Schematic of a MCFC.

As a result, water and carbon dioxide are important components of the feed gases to the MCFCs. Water, produced by the main anode reaction, helps to shift the equilibrium reactions to produce more hydrogen for the anodic electrochemical reaction. Water must also be present in the feed gas, especially in low-Btu (i.e., high CO content) fuel mixtures, to avoid carbon deposition in the fuel gas flow channels supplying the cell, or even inside the cell itself. Carbon dioxide, from the fuel exhaust gas, is usually recycled to the cathode as it is required for the reduction of oxygen.

The MCFCs use a molten alkali carbonate mixture as the electrolyte, which is immobilized in a porous lithium aluminate matrix. The conducting species is carbonate ions. Lithiated nickel oxide is the material of the current choice for the cathode, and nickel, cobalt, and copper are currently used as anode materials, often in the form of powdered alloys and composites with oxides. As a porous metal structure, it is subject to sintering and creeping under the compressive force necessary for stack operation. Additives such as chromium or aluminum form dispersed oxides and thereby increase the long-term stability of the anode with respect to sintering and creeping. MCFCs normally have about 75%–80% fuel (hydrogen) utilization.

Acceptable Contamination Levels

MCFCs do not suffer from carbon monoxide poisoning, and in fact, they can utilize carbon monoxide in the anode gas as the fuel. However, they are extremely sensitive to the presence of sulfur (<1 ppm) in the reformed fuel (as hydrogen sulfide, H_2S) and oxidant gas stream (SO_2 in the recycled anode exhaust). The presence of HCl, HF, and HBr causes corrosion, while trace metals can spoil the electrodes. The presence of particulates of coal/fine ash in the reformed fuel can clog the gas passages.

Major Technological Problems

The main research efforts for the MCFCs are focused on increasing the lifetime and endurance and reducing the long-term performance decay. The main determining factors for the MCFC are electrolyte loss, cathode dissolution, electrode creepage and sintering, separator plate corrosion, and catalyst poisoning for internal reforming.

Electrolyte loss results in increased ohmic resistance and activation polarization, and it is the most important and continuously active factor in causing the long-term performance degradation. It is primarily a result of electrolyte consumption by the corrosion/dissolution processes of cell components, electric potential-driven electrolyte migration, and electrolyte vaporization. Electrolyte evaporation (usually Li_2CO_3 and/or K_2CO_3) occurs either directly as carbonate or indirectly as hydroxide.

The cathode consists of NiO, which slowly dissolves in the electrolyte during operation. It is then transported towards the anode and precipitates in the electrolyte matrix as Ni. These processes lead to a gradual degradation of cathode performance and the shorting of the electrolyte matrix. The time at which shorting occurs depends not only, via NiO solubility, on the CO_2 partial pressure and the cell temperature but also on the matrix structure, i.e., on the porosity, pore size, and, in particular, thickness of the matrix. Experience indicates that this cell shorting mechanism tends to limit stack life to about 30,000 hr under the atmospheric reference gas conditions, and much shorter for real operating conditions.

Electrode (especially anode) creepage and sintering (i.e., a coarsening and compression of electrode particles) result in increased ohmic resistance and electrode polarization. NiO cathodes have quite satisfactory sinter and creepage resistance. Creep resistance of electrodes has an important effect on maintaining low contact resistance of the cells and stacks. The corrosion of the separator plate depends on many factors, such as the substrate, possible protective layers, composition of the electrolyte, local potential and gas composition, and the oxidizing and reducing atmospheres at the cathode and anode, respectively. Poisoning of the reforming catalyst occurs for direct internal-reforming MCFCs. It is caused by the evaporation of electrolyte from the cell components and condensation on the catalyst, which is the coldest spot in the cell, and by liquid creep within the cell.

Technological Status

MCFC technology is in the first demonstration phase and under the product development with full-scale systems at the 250 kW to 2 MW range. The short-term goal is to reach a lifetime of 40,000 hr. It is estimated that the capital cost is about $1000–1600/kW for the MCFC power systems. The cost breakdown is, at full-scale production levels, about one-third for the stack, and two-thirds for the balance of the plant. It is also generally accepted that the cost of raw materials will constitute about 80% of total stack costs. Although substantial development efforts supported by fundamental research are still needed, the available knowledge and number of alternatives will probably make it possible to produce precommercial units in the earlier part of the coming decade at a capital cost of $2000–4000/kW. Precompetitive commercial units may be expected some years later by which time further cost reduction to full competitiveness will be guided by extensive operating experience and increased volume production.

Applications

The MCFC is being developed for their potential as baseload utility generators. However, their best application is in distributed power generation and cogeneration (i.e., for capacities less than 20 MW in size), and in this size range, MCFCs are 50% to 100% more efficient than turbines—the conventional power generator. Other applications have been foreseen, such as pipeline compressor stations, commercial buildings, and industrial sites in the near term and repowering applications in the longer term. Due to its high operation temperature, it only has very limited potential for transportation applications. This is because of its relatively low power density and long start-up times. However, it may be suitable as a powertrain for large surface ships and trains.

Solid Oxide Fuel Cells

Introduction

SOFCs have emerged as a serious alternative high- temperature fuel cell, and they have been often referred to as the third-generation fuel cell technology because their commercialization is expected after the PAFCs (the first generation) and MCFCs (the second generation).

SOFC is an all-solid-state power system, including the electrolyte, and it is operated at high temperature of around 1000°C for adequate ionic and electronic conductivity of various cell components. The all-solid-state cell composition makes the SOFC system simpler in concept, design, and construction; two-phase (gas–solid) contact for the reaction zone reduces corrosion and eliminates all the problems associated with the liquid electrolyte management. The high-temperature operation results in fast electrochemical kinetics (i.e., low activation polarization), without the need for noble metal catalysts. The fuel may be gaseous hydrogen, H_2/CO mixture, or hydrocarbons because the high-temperature operation makes the internal in situ reforming of hydrocarbons with water vapor possible. It is specially noticed that CO is no longer a contaminant; rather, it becomes a fuel in SOFCs. Even with external reforming, the SOFC fuel feedstock stream does not require the extensive steam reforming with shift conversion as it does for the low-temperature fuel cell systems. More important, the SOFC provides high-quality waste heat that can be utilized for cogeneration applications or combined cycle operation for additional electric power generation. The SOFC operating condition is also compatible with the coal gasification process, which makes the SOFC systems highly efficient when using coal as the primary fuel. It has been estimated that the chemical-to-electrical energy conversion efficiency is 50% to 60%, even though some estimates go as high as 70% to 80%. Also, nitrogen oxides are not produced and the amount of carbon dioxide released per kilowatt hour is around 50% less than for power sources based on combustion because of the high efficiency.

Basic Operating Principle

As mentioned earlier, both hydrogen and carbon monoxide can be oxidized in the SOFCs directly. Hence, if hydrogen or hydrogen-rich gas mixture is used as fuel, and oxygen (or air) is used as oxidant, the half cell reaction becomes (Figure 2)

$$\text{Anode}: H_2(g) + O^{2-} = H_2O(g) + 2e^-$$

$$\text{Cathode}: 2e^- + \frac{1}{2}O_2(g) = O^{2-}$$

$$H_2(g) + \frac{1}{2}O_2(g) = H_2O(g).$$

and the overall cell reaction becomes

Note that g represents gaseous phase.

However, if carbon monoxide is provided to the anode instead of hydrogen, the anode reaction becomes

$$\text{Anode}: CO(g) + O^{2-} = CO_2(g) + 2e^-$$

with the cathode reaction remaining the same, the cell reaction becomes

$$CO(g) + \frac{1}{2}O_2(g) = CO_2(g)$$

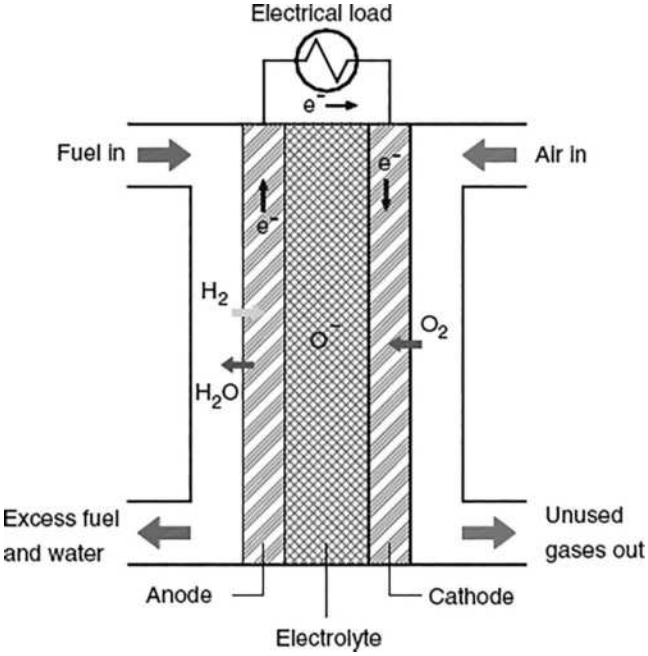


FIGURE 2 Schematic of SOFC.

If the fuel stream contains both hydrogen and carbon monoxide as is the case for hydrocarbon reformed gas mixture, especially from the gasification of coal, the oxidation of hydrogen and carbon monoxide occurs simultaneously at the anode, and the combined anode reaction becomes

$$\text{Anode}: \text{a H}_2(\text{g}) + \text{b CO} + (\text{a}+\text{b})\text{O}^{2-} = \text{a H}_2(\text{g}) + \text{b CO}_2 + 2\ (\text{a}+\text{b})\text{e}^-$$

Consequently, the corresponding cathode and overall cell reaction become

$$\text{Cathode}: \frac{1}{2}(\text{a}+\text{b})\text{O}_2 + 2(\text{a}+\text{b})e^- = (\text{a}+\text{b})\text{O}^{2-}$$

$$\text{Cell}: \frac{1}{2}(\text{a}+\text{b})\text{O}_2 + \text{a H}_2 + \text{b CO} = \text{a H}_2\text{O} + \text{b CO}_2$$

The solid electrolyte in SOFCs is usually yttria-stabilized zirconia (YSZ); thus, a high operating temperature of around 1000°C is required to ensure adequate ionic conductivity and low ohmic resistance. This is especially important because the cell open-circuit voltage is low, compared with low-temperature fuel cells, typically around 0.9–1 V under the typical working conditions of the SOFCs. The high-temperature operation of the SOFCs makes the activation polarization very small, resulting in the design operation in the range dominated by the ohmic polarization. The conventional material for the anode is nickel-YSZ-cermet, and the cathode is usually made of lanthanum-strontium-manganite. Metallic current collector plates of a high-temperature corrosion-resistant chromium-based alloy are typically used.

Acceptable Contamination Levels

Because of high temperature, the SOFCs can better tolerate impurities in the incoming fuel stream. They can operate equally well on dry or humidified hydrogen or carbon monoxide fuel or on mixtures of them. However, hydrogen sulfide (H_2S), hydrogen chloride (HCl), and ammonia (NH_3) are impurities typically found in coal gasified products, and each of these substances is potentially harmful to the performance of SOFCs. The main poisoning factor for SOFCs is H_2S. Though the sulfur tolerance level is approximately two orders of magnitude greater than other fuel cells, the level is below 80 ppm. However, studies have shown that the effect of hydrogen sulfide (H_2S) is reversible, meaning that the cell performance will recover if hydrogen sulfide is removed from the fuel stream or clean fuel is provided after the contaminant poison has occurred.

Major Technological Problems

The high-temperature operation of the SOFCs places stringent requirements on materials used for cell construction, and appropriate materials for cell components are very scarce. Therefore, the key technical challenges are the development of suitable materials and the fabrication techniques. Of the material requirements, the most important consideration is the matching of the thermal expansion coefficients of electrode materials with that of the electrolyte to prevent cracking or delamination of SOFC components either during high-temperature operation or heating/cooling cycles. One of the remedies for the thermal expansion mismatch is to increase the mechanical toughness of the cell materials by either developing new materials or doping the existing materials with SrO and CaO.

The electrode voltage losses are reduced when the electrode material possesses both ionic and electronic conductivities (the so-called mixed conduction), for which the electrochemical reactions occur throughout the entire surface of the electrode rather than only at the three-phase interface of, e.g., the cathode, the air (gas phase), and the electrolyte. Therefore, it is important for performance enhancement to develop mixed-conduction materials for both the cathode and anode, which have good thermal expansion match with the electrolyte used and good electrical conductivity to reduce the ohmic polarization that dominates the SOFC voltage losses.

Another focus of the current development is the intermediate-temperature SOFCs operating at around 800°C for better matching with the bottoming turbine cycles and lessening requirements for the cell component materials. Again, appropriate materials with adequate electrical conductivity are the key areas of the development effort, and thermal expansion matching among the cell components is still necessary.

Technological Status

There are three major configurations for SOFCs: tubular, flat plate, and monolithic. Even though SOFC technology is in the developmental stage, the tubular design has gone through development at Westinghouse Electric Corporation since the late 1950s and has been demonstrated at user sites in a complete operating fuel cell power unit of nominal 25 kW (40 kW maximum) capacity. The flat plate and the monolithic designs are at a much earlier development status typified by subscale, single cell, and short stack development (up to 40 cells). The present estimated capital cost is $1500/kW, but it is expected to be reduced with improvements in technology. Therefore, the SOFCs may become very competitive with the existing technology for electric power generation. However, it is believed that the SOFC technology is at least 5 to 10 years away from commercialization.

Applications

SOFCs are very attractive in electrical utility and industrial applications. The high operating temperature allows them to use hydrogen and carbon monoxide from natural gas steam reformers and coal gasification plants, a major advantage as far as fuel selection is concerned. SOFCs are being developed for the large (>10 MW, especially 100 to 300 MW) baseload stationary power plants with coal as the primary fuel. This is one of the most lucrative markets for this type of fuel cells.

A promising field for SOFCs is the decentralized power supply in the megawatt range, where the SOFC gains interest due to its capability to convert natural gas without external reforming. In the range of one to some tenths of a megawatt, the predicted benefits in electrical efficiency of SOFC-based power plants over conventional methods of electricity generation from natural gas can only be achieved by an internal-reforming SOFC. Thus, internal reforming is a major target of present worldwide SOFC development.

Concluding Remarks

A summary of the preceding discussion, including the operational characteristics and technological status of the intermediate- and high-temperature fuel cells, is given in Table 1. Note that the costs for PAFC, MCFC, and SOFC are for stationary power systems.

HT-PEMFC is still at the laboratory research level mainly due to the durability issue of the membrane at high temperatures and is expected to be more suitable for stationary rather than automotive applications due to its slower start-up characteristics (higher operating temperature to be reached) than the conventional PEMFC. PAFC is the most commercially developed fuel cell operating at intermediate temperatures. PAFCs are being used in applications such as hospitals, hotels, offices, and schools with relative high conversion efficiency. The high-temperature fuel cells like MCFCs and SOFCs may be most appropriate for cogeneration and combined cycle systems (with gas or steam turbine as the bottoming cycle). The MCFCs have the highest energy efficiency attainable from methane to electricity conversion in the size range of 250 kW to 20 MW, whereas the SOFCs are best suited for baseload utility application

TABLE 1 Operational Characteristics and Technological Status of Intermediate- and High-Temperature Fuel Cells

Type of Fuel Cells	Operating Temperature (°C)	Power Density (mW/cm^2), (present) Projected	Projected Rated Power Level (kW)	Fuel Efficiency	Lifetime Projected (hr)	Capital Cost Projected ($/kW)	Application Areas
HT-PEMFC	100–200	(600) >900	1–1000	45–70	>40,000	35	Transportation, space
PAFC	160–220	(200) 250	100–5000	55	>40,000	3000	Dispersed and distributed power
MCFC	600–700	(100) >200	1000–100,000	60–65	>40,000	1000	Distributed power generation
SOFC	800–1000	(240) 300	100–100,000	55–65	>40,000	1500	Baseload power generation

operating on coal-derived gases. It is estimated that the MCFC technology is about 5 to 10 years away from commercialization, and the SOFCs are probably years afterwards. In addition, due to the high operating temperatures and fuel reforming of MCFC and SOFC, the management of the exhaust gases (e.g., NO_x and SO_2) is important due to the environmental concerns.

Bibliography

1. Blomen, L.J.M.J.; Mugerwa, M.N. *Fuel Cell Systems;* Plenum Press: New York, 1993.
2. Larminie, J.; Dicks, A. *Fuel Cell Systems Explained,* 2nd Ed.; Wiley, Chichester, 2003.
3. Kordesch, K.; Simader, G. *Fuel Cells and Their Applications;* VCH Press: New York, 1996.
4. Appleby, A.J.; Foulkes, F.R. *Fuel Cell Handbook*; Van Nostrand Reinhold Company, New York, 1989.
5. Hoogers, G. *Fuel Cell Technology Handbook*; CRC Press, New York, 2002.
6. Vielstich, W.; Gasteiger, H.; Lamm, A. *Handbook of Fuel Cells—Fundamentals, Technology,* Applications, 4-volume set, Wiley, Chichester, 2003.
7. Li, X. *Principles of Fuel Cells*; Taylor and Francis, New York 2005.

35

Fuel Cells: Low Temperature

Xianguo Li
and Kui Jiao

Introduction

Fuel cell technology has been developed and improved dramatically in the past two decades, and this has once again captured public attention as well as the industry's attention concerning the prospect of fuel cells as practical power sources for terrestrial applications. Fuel cell is a highly energy-efficient and environmentally friendly technology for power generation that is also compatible with alternative fuels and renewable energy sources and carriers for sustainable development.

Fuel cell offers additional advantages for both mobile and stationary power generation applications. Fuel cell, as an electrochemical device, has no moving components except for peripheral equipment. As a result, its operation is very quiet, virtually without vibration and noise, thus capable of on-site cogeneration with no need of long distance power transmission lines that consume approximately 10% of electric energy delivered to consumers in North America. Its inherent modularity allows for simple construction and operation with possible applications for baseload electricity generation, dispersed, distributed, and portable power generation, because it may be made in any size from a few watts up to a megawatt-scale plant with equal efficiency. Its fast response to the changing load condition while maintaining high efficiency makes it ideally suited to load following applications. Its high efficiency represents less chemical, thermal, and carbon dioxide emissions for a given amount of power generation.

At present, fuel cell technology is being routinely used in many specific areas, notably in space explorations, where fuel cell operates on pure hydrogen and oxygen with over 70% efficiency and drinkable water as the only by-product. Fuel cell technology is also being widely used for terrestrial applications; impressive technical progress has been achieved in terms of higher power density and better durability as well as reduced capital and maintenance and operation cost, and is driving the development of competitively priced fuel-cell-based power generation systems with advanced features for terrestrial use, such as utility power plants and zero-emission vehicles. In light of decreasing fossil fuel reserves and increasing energy demands worldwide, fuel cell will probably become one of the major energy technologies with fiercest international competition in the 21st century.

The major terrestrial commercial applications of fuel cells are electric power generation in the utility industry and as a zero-emission powertrain in the transportation sector. For these practical applications, the efficiencies of fuel cells range somewhere from 40% to 65% based on the lower heating value (LHV) of hydrogen. Typically, the cell electric potential is only about 1 V across a single cell, and it decreases due to various loss mechanisms under the operational conditions. Thus, multiple cells are required to be connected together in electrical series in order to achieve a useful voltage for practical purposes, and these connected cells are often referred to as a fuel cell stack. A fuel cell system consists of one or multiple fuel cell stacks connected in series and/or parallel and the necessary auxiliaries whose composition depends on the type of fuel cells and the kind of primary fuels used. The major accessories include thermal management (or cooling) subsystem, fuel supply, storage and processing subsystem, and oxidant (typically air) supply and conditioning subsystem.

In light of the recent advancement in fuel cell technology and its potential impact on municipal utilities, the energy delivery marketplace, and transportation sector, an assessment of the current state of fuel cell technology has been conducted. The result of this study is described in the first and second parts of this report, which summarizes the state-of-the-art technology for the seven major types of fuel cells, including

- Alkaline fuel cells (AFCs)
- Proton exchange membrane fuel cells (PEMFCs)
- High-temperature PEMFCs (HT-PEMFCs)
- Direct methanol fuel cells (DMFCs)
- Phosphoric acid fuel cells (PAFCs)
- Molten carbonate fuel cells (MCFCs)
- Solid oxide fuel cells (SOFCs)

A critical parameter that determines the potential applications of each type of fuel cell is the operating temperature. For instance, AFCs, DMFCs, and PEMFCs have potential applications in transportation because they do not produce much heat (which otherwise would have to be eliminated by some cooling device) and, as a result of this, have a very short start-up period (a few minutes). These types of fuel cells are discussed in this entry.

On the other hand, HT-PEMFCs, PAFCs, MCFCs, and SOFCs producing high-temperature heat are more complex to run and are better fit for stationary applications like combined heat power generation or CHP. The fuel cells under this category will be discussed in a separate entry entitled “Fuel Cells: Intermediate and High Temperature.”

The study was based on a literature review and search of technical databases, and recent advancement of fuel cell technology has been identified and delineated for each of the seven major types of fuel cells listed above. The following sections describe each of the three low- temperature fuel cells. Their technological status is presented with respect to their basic operating principle, acceptable contamination levels, the state-of-the-art technology, major technical barriers to commercialization, their economics and suitability for utility, and transportation applications.

Alkaline Fuel Cells

AFCs give the best performance among all the fuel cell types under the same or similar operating conditions when running on pure hydrogen and oxygen. Hence, they are among the first fuel cells to have been studied and taken into development for practical applications, and they are the first type of fuel cells to have reached successful routine applications, mainly in space programs such as space shuttle missions in the United States and similar space exploration endeavors in China and Europe, where pure hydrogen and oxygen are used as reactants. Because of their success in space programs, AFCs are also the type of fuel cells on which probably the largest number of fuel cell development programs have begun in the world in an effort to bring them down to terrestrial applications, particularly in Europe, and almost all the AFC development programs have come to an end. In recent years, AFC with solid electrolyte membrane has attracted people's attention once again. Because with solid electrolyte membrane, AFC can be expected to offer similar or better power density and durability when compared with PEMFC. In addition to its most efficient energy conversion mechanism among all the fuel cell types, it is becoming a promising candidate for energy conversion in automotive applications.

The AFCs have the highest energy conversion efficiency among all types of fuel cells under the same operating condition if pure hydrogen and pure oxygen are used as reactants. That was one of the important reasons that AFC was selected for the U.S. space shuttle missions. The AFCs used in the shuttle missions are operated at about 200°C for better performance (i.e., high energy conversion efficiency of over 70% and high power density, which is critical for space applications), and the alkaline electrolyte is potassium hydroxide (KOH) solution immobilized in the asbestos matrix. As a result, the AFCs operate at high pressure in order to prevent the boiling and depletion of the liquid electrolyte. Consequently, these severe operating conditions of high temperature and high pressure dictate extremely strict requirement for cell component materials that must withstand the extreme corrosive oxidizing and reducing environment of the cathode and anode. To meet these requirements, precious metals such as platinum, gold, and silver are used for the construction of the electrodes, although these precious metals are not necessary for the electrochemical reactions leading to electric power generation. Each shuttle flight contains 36 kW AFC power system. Its purchase price is about $28.5 million, and it costs NASA additional $12–19 million annually for operation and maintenance. Although the manufacturer claims about 2400 hr of lifetime, NASA's experience indicates that the real lifetime is only about 1200 hr. With sufficient technology development, 10,000 hr are expected as the life potential (or upper limit) for the AFC system. This belief is based on the nature of the AFC systems and the data accumulated on both stacks and single cells.

The typical working temperature of AFC power systems aimed at commercial and terrestrial applications ranges from 20°C to 90°C, and the electrolyte is KOH solution (30%–35%) or solid electrolyte membrane. There are five different cell types investigated:

1. Cell with a free liquid electrolyte between two porous electrodes
2. ELOFLUX cell with liquid KOH in the pore systems.
3. Matrix cell where the electrolyte is fixed in the electrode matrix
4. The falling film cell
5. Cell with solid electrolyte membrane

Research groups are working on the development of technical AFC systems; the group name and the current status of the work are listed below:

- Siemens, Erlangen, Germany: stopped working on the AFCs
- VARTA AG, Kelkheim, Germany: terminated working on preparation of technical electrodes and technical fuel stacks in 1993
- GH, Kassel, Germany: stopped working on preparation of technical electrodes and technical fuel cell stacks in 1994
- ISET, Kassel, Germany: stopped working on technical AFC systems in 1994
- DLR-ITT (German Aerospace Research Establishment), Stuttgart: stopped working on the investigation of the degradation of technical electrodes, the development of new catalysts for AFC, and the theoretical simulation of stacks and systems in 1994
- ELENCO, Antwerpen, Belgium: stopped working on electrodes, stacks, and systems, as well as bus demonstration program (it went bankrupt in 1995)

- Royal Institute of Technology, Stockholm: working on the field of stationary fuel cells powered by biofuels
- Hoechst AG, Frankfurt: terminated working on AFC electrodes and stopped development of the falling film cell
- Technical University, Graz, Austria: planning an investigation of degradation effects
- Zevco with headquarters in London: demonstration of AFCs for mobile applications, particularly for bus with ELENCO's technology
- University of California, Riverside, California: development of solid electrolyte membrane for AFC

The following is a list of some selected technical applications and demonstration projects in Europe in recent years:

- Space applications: the space applications are the development of an AFC system for the European space shuttle HERMES (Siemens, ELENCO, VARTA/GHK) and the bipolar matrix AFC system Photon for space applications (Ural Electrochemical Integrated plant).
- Vehicles: applications of AFCs in vehicles are as follows:
  1. Forklift truck of VARTA AFC
  2. VW Van, 14 kW ELENCO AFC and battery
  3. VW van, 17.5 kW Siemens AFC
  4. Submarine, 100 kW Siemens AFC
  5. EUREKA-Bus, 80 kW ELENCO AFC and battery
- Decentralized energy supply: these supplies comprise the following:
  1. A meteorological station, 5 W, VARTA AFC for long-term operation
  2. A TV-transmitting installation Ruppertshain, 100 W VARTA AFC
  3. A mobile current supply unit for the Belgian Geological Service, 40 kW ELENCO AFC
- Energy storage: the solar hydrogen demonstration plant in Neunburg vorm Wald, 6.5 kW Siemens AFC, and the solar hydrogen system at Fachhochschule Wiesbaden, 1.2 kW ELENCO AFC, are typical examples. Many problems concerning AFCs are described in the literature. The most important problems are the following:
  - Preparation method of the electrodes: the electrodes consist of porous material that is covered with a layer of catalyst. Generally, it is very difficult to distribute the catalyst at the surface and to produce a defined pore system for the transportation of the reactants.
  - Costs of the electrode, stacks, and fuel cell systems: the preparation of electrodes with noble metal catalysts is very expensive. In general, the electrodes are manufactured in a small-scale production with high overhead costs.
  - Lifetime of the electrode/degradation: the electrolyte is very corrosive and the catalyst materials are sensitive to high polarization. Using nickel and silver as catalysts, in order to reduce the costs of the fuel cell, leads to a high degradation of these catalysts. This problem is expected to be solved or abated with solid electrolyte membrane.
  - Diaphragm made of asbestos: the diaphragm of low- temperature AFCs is made of asbestos. This material is hazardous for health, and in some countries, its use is even banned. Therefore, new diaphragms should be developed, but it is difficult to find a material with a similar behavior in alkaline electrolyte. This issue is solved when solid electrolyte membrane is used.
  - Carbon-dioxide-contaminated fuel and oxidant streams (carbonating of electrolyte and electrodes): the electrolyte intolerance of carbon dioxide is the most important disadvantage of air-breathing AFCs with reformate gases from primary fossil fuels. This problem is expected to be solved or abated with solid electrolyte membrane.

Other problems associated with the AFC power systems are the concern on the safety and reliability of AFC power systems, mainly related to AFCs with liquid electrolyte. For example, the liquid KOH electrolyte contained in an asbestos matrix can only withstand a 5 psi limit of pressure differential between the anode and cathode reactant gases. This dictates the need for sophisticated pinpoint pressure control during the operation including transient, start-up, and shutdown process. It is also a safety issue because of its greater likelihood of mixing of the reactants occurring in the AFC system with the possibility of a serious fire breaking out. In terms of general safety considerations, the use of the corrosive potassium hydroxide electrolyte in the AFCs represents the need for hazardous handling, and the handling of asbestos matrix poses potential hazard to one's health. With flowing reactant gases, the potential for the gradual loss of the liquid electrolyte, drying of the electrolyte matrix, reactant crossover of the matrix, and ensuing life-limiting reactant mixing (or actual AFC stack failure due to fire) is very real in the AFC system. Moreover, the liquid electrolytes, like potassium or sodium hydroxide, do not reject carbon dioxide; even the 300–350 ppm of carbon dioxide in the atmospheric air is not tolerated (carbon dioxide concentration in both cathode and anode gases must be less than 10–100 ppm by volume), while terrestrial applications almost invariably require the use of atmospheric air as oxidant due to technical and economic considerations. For municipal electric applications, hydrocarbon fuels, especially natural gas, are expected to be the primary fuel, and their reformation into hydrogen-rich gases invariably contains a significant amount of carbon dioxide, e.g., steam reforming of the natural gas results in the reformate gas consisting approximately of 80% hydrogen, 20% carbon dioxide, and a trace amount of other components such as carbon monoxide. Carbonaceous products of aging and corrosion shorten AFC life; they degrade the alkaline electrolyte. Whether originating as impurities in the gaseous reactants or from some fuel cell materials, oxides of carbon will chemically react with the alkaline electrolyte and produce irreversible decay, which will decrease performance and shorten life. Consequently, AFCs with liquid electrolyte have been restricted to specialized applications where pure hydrogen and oxygen are utilized.

The revival of this technology seems to have started in recent years, mainly due to the development of solid alkaline electrolyte membrane. For AFCs with solid electrolyte membrane, most of the above-mentioned problems can be solved or abated. The solid electrolyte membrane makes the AFC technology comparable with PEMFC in terms of safety, cost, power density, and durability. In addition, since AFC has the most efficient energy conversion mechanism among all the fuel cell types, it therefore also has higher possibility to use no-precious catalyst than PEMFC. Nevertheless, AFC with solid electrolyte membrane is still at its early stage of development; the cost, durability, and performance are expected to be further examined and improved, and the transport phenomena inside the cell need to be investigated as well.

Proton Exchange Membrane Fuel Cells

Introduction

The PEMFC is also called solid polymer (electrolyte) fuel cell. It is perhaps the most elegant of all fuel cell systems in terms of design and mode of operation. It was the first type of fuel cell that was put into practical application (in Gemini space missions from 1962 to 1966). It consists of a solid polymeric membrane acting as the electrolyte. The solid membrane is an excellent proton conductor, sandwiched between two platinum-catalyzed porous carbon electrodes. It has fast start capability and yields the highest output power density among all types of the fuel cells. Because of the solid membrane as the electrolyte, there is no corrosive fluid spillage hazard, and there is lower sensitivity to orientation. It has no volatile electrolyte and has minimal corrosion concerns. It has truly zero pollutant emissions with potable liquid product water when hydrogen is used as fuel. As a result, the PEMFC is particularly suited for vehicular power application, although it is also being considered for stationary power application, albeit to a lesser degree.

The proton-conducting polymer membrane belongs to a class of materials called ionomers or polyelectrolytes, which contain functional groups that will dissociate in the presence of water. The dissociation produces ions fixed to the polymer and simple counterions that can freely exchange with ions of the same sign from the solution. The current available polyelectrolytes have cation as the counterion. In the case of hydrogen, the cation is proton. Therefore, the membrane must be fully hydrated in order to have adequate ion conductivity. As a result, the fuel cell must be operated under conditions where the product water does not evaporate faster than it is produced, and the reactant gases, both hydrogen and oxygen, need to be humidified. Therefore, water and thermal management in the membrane become critical for efficient cell performance, are fairly complex, and require dynamic control to match the varying operating conditions of the fuel cell. Because of the limitation imposed by the membrane and problems with water balance, the operating temperature of PEMFCs is usually less than 100°C, typically at 80°C. This rather low operating temperature requires the use of noble metals as catalysts in both the anode and cathode side with generally higher catalyst loadings than those used in PAFCs.

Currently, the polymer electrolyte used is made of per-fluorinated sulfonic acid membrane, or it is essentially acid, though in solid polymeric form. Hence, PEMFCs are essentially acid electrolyte fuel cells, with its operational principle essentially the same as PAFCs. As a result, most of PEMFC design, material selection, component fabrication, etc., are similar to those of PAFCs. The only difference is the humidification of reactant gases dictated by the membrane performance. Reactant humidification is often achieved by a number of techniques, e.g., by passing gas stream through a water column, by using in-stack humidification section of cell and membrane arrangement, and by spraying water into the reactant streams. In the early stage of the PEMFC development, the membranes were based on polystyrene, but since 1968, a Teflon-based product named “Nafion” by DuPont is used. This offers high stability, high oxygen solubility, and high mechanical strength.

Basic Operating Principle

The schematic of a single PEMFC is illustrated in Figure 1. The PEMFC requires hydrogen gas as the fuel and oxygen (typically air) as the oxidant. The half-cell reactions are

$$\text{Anode}: H_2 = 2H^+ + 2e^-$$

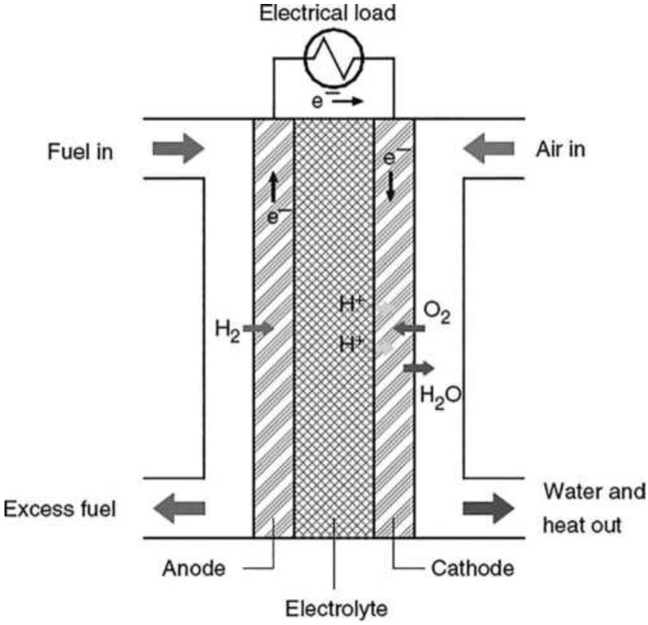


FIGURE 1 Schematic of a PEMFC.

$$\text{Cathode}: 1/2\,O_2 + 2H^+ + 2e^- = H_2O$$

and the overall cell reaction is

$$H_2 + 1/2\,O_2 = H_2O \; + \; \text{Heat Generated} + \text{Electric Energy}.$$

The current PEMFCs use perfluorinated sulfonic acid membrane as the proton-conducting electrolyte; carbon paper or cloth as the anode and cathode backing layers; and platinum or its alloys, often supported on carbon black, as the catalyst. The bipolar plate with the reactant gas flow fields is often made of graphite plate. The stoichiometry is around 1.1 to 1.2 for the fuel and 2 for the oxidant (oxygen). The PEMFCs usually operate at about 80°C and 1–8 atm pressure. The pressures, in general, are maintained equal on either side of the membrane. Operation at high pressure is necessary to attain high power densities, particularly when air is chosen as the cathodic reactant.

To prevent the membrane dryout leading to local hot spot (and crack) formation, performance degradation, and lifetime reduction, both fuel and oxidant streams are often fully humidified, and the operating temperature is limited by the saturation temperature of water corresponding to the operating pressure. The product liquid water formed at the cathode can dissolve in the electrolyte membrane or be removed from the cell by the excessive oxidant gas stream. The accumulation of liquid water in the cathode backing layer blocks the oxygen transfer to the catalytic sites, thus resulting in the phenomenon called "water flooding," causing performance reduction. Local hot and cold spots will cause the evaporation and condensation of water. Thus, an integrated approach to thermal and water management is critical to PEMFCs' operation and performance, and a proper design must be implemented.

Cold Start

In winter conditions, it is unavoidable for vehicles driving below the freezing point of water (0°C); therefore, for successful commercialization of PEMFC in automotive applications, rapid start-up from subzero temperatures must be achieved, which is referred to as "cold start" of PEMFC. The major problem of PEMFC cold start is that the product water freezes when the temperature inside the PEMFC is lower than the freezing point of water. If the catalyst layer is fully covered by ice before the cell temperature rises above the freezing point, the electrochemical reaction may be stopped due to the blockage of the reaction site. In addition, ice formation may also result in serious damage to the structure of the membrane electrode assembly. Most of the present PEMFCs employ various assisted cold start methods such as resistance heating using DC from batteries, coolant heating, hot air blowing, and catalytic hydrogen/oxygen reaction inside the PEMFC. The cold start performance can also be improved by design optimizations. Membrane thickness, ionomer content in catalyst layer, total heat capacity of the cell, and cell insulation are all important design factors to improve the cold start performance.

Acceptable Contamination Levels

As an acid electrolyte fuel cell operating at low temperature, the PEMFC is primarily vulnerable to carbon monoxide poisoning. Even a trace amount of CO drastically reduces the performance levels, although CO poisoning effect is reversible and does not cause permanent damages to the PEMFC system. Further, the performance reduction due to CO poisoning takes a long time (on the order of 2 hr) to reach steady state. This transient effect may have profound implication for transportation applications. Therefore, the PEMFC requires the use of a fuel virtually free of CO (must be less than a few parts per million). Also, high-quality water free of metal ions should be used for the cell cooling and reactant humidification to avoid the contamination of the membrane electrolyte. This requirement has a severe implication on the materials that can be used for cell components. On the other hand, carbon dioxide does not affect PEMFC operation and performance except through the effect of reactant dilution (the Nernst loss).

Major Technological Problems

For practical applications, PEMFC performance in terms of energy efficiency, power density (both size and weight), durability, and capital cost must be further improved. This can be accomplished by systematic research in the following:

i. New oxygen reduction electrocatalysts: This includes the development of non-precious catalyst and the reduction of precious metal platinum and its alloys loading from 4 mg/cm^2 to 0.4 mg/cm^2 or lower without affecting the long-term performance and the lifetime, as well as the development of CO-tolerant catalysts.
ii. New types of polymer electrolyte with higher oxygen solubility, thermal stability, long life, and low cost. A self-humidified membrane or a polymer without the need of humidification will be ideal for PEMFC operation and performance enhancement with significant simplification of system complexities and reduction of the cost.
iii. Profound changes in oxygen (air) diffusion electrode structure to minimize all transport-related losses. The minimization of all transport losses is the most promising direction for PEMFC performance improvement.
iv. Optimal thermal and water management throughout the individual cells and the whole stack to avoid local hot and dry spot formation and to avoid water flooding of the electrode.
v. Optimal design from single-cell component to system level to enhance the cold start performance.

Figure 2 illustrates the schematic of a PEMFC stack and its major components. In addition to the above issues, the development of low-cost lightweight materials for construction of reactant gas flow fields and bipolar plates is one of the major barriers to PEMFCs' large-scale commercialization. The successful solution to this problem will further increase the output power density, and it includes an optimal design of flow fields with the operating conditions, as well as an appropriate selection of materials and fabrication techniques. It has been reported that as much as over 20% improvement in the performance of PEMFC stacks can be obtained just by appropriate design of flow channels alone. The current leading technologies for bipolar plate design include the following: injection-molded carbon-polymer composites, injection-molded and carbonized amorphous carbon, assembled three-piece metallic, and stamped unitized metallic.

Technological Status

PEMFCs have achieved a high power density of over 2.2 kW/L, perhaps the highest among all types of the fuel cells currently under development. It is also projected that the power density may be further improved with unitized metallic (stainless steel) bipolar plates. The U.S. Department of Energy reported that the capital cost of 80 kW automotive fuel cell system costs in volume production (projected to 500,000 units per year) is $61/kW in 2009, and the future target is $35/kW. It is expected that PEMFC technology is about 5 to 10 years from commercialization, and precommercial demonstration for buses and passenger vehicles is under way with increasing intensity. The first demonstration for residential combined heat and power application has began at the end of 1999, and application of PEMFCs in powering portable and mobile electronics such as laptops has already been started.

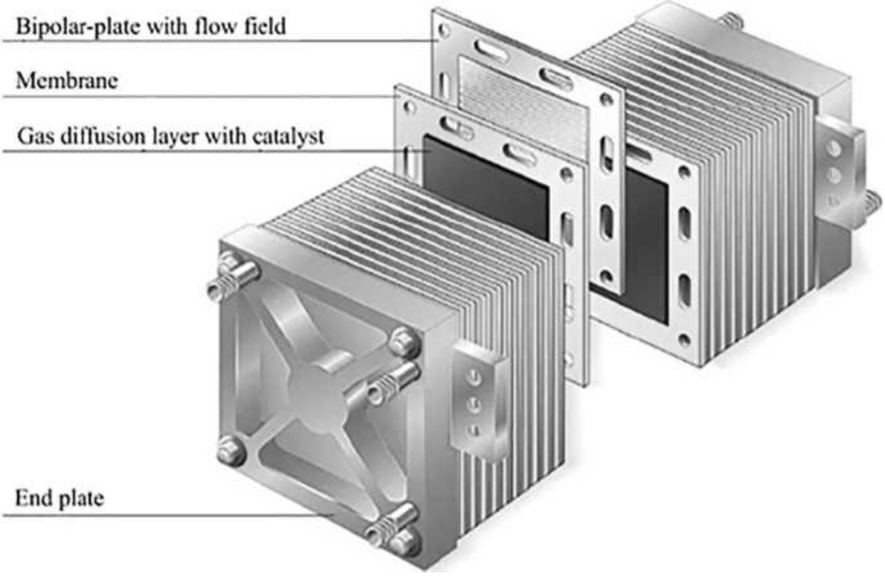


FIGURE 2 Schematic of a PEMFC stack and its components. (Source: Ballard Power Systems: http://www.ballard.com.)

Applications

PEMFCs have a high power density, a variable power output, and a short start-up time due to low operating temperature; the solid polymer electrolyte is virtually corrosion free and can withstand a large pressure differential (as high as 750 psi reported by NASA) between the anode and cathode reactant gas streams. Hence, PEMFCs are suitable for use in the transportation sector. Currently, they are considered the best choice for zero-emission vehicles as far as present-day available fuel cell technologies are concerned. Their high power density and small size make them primary candidates for light-duty vehicles, though they are also used for heavy-duty vehicles. For high-profile automobile application, pure hydrogen and air are used as reactants at the present. However, conventional gasoline and diesel engines are extremely cheap, estimated to cost about \$30–\$50/kW. Therefore, the cost of PEMFC systems must be lowered in order to be competitive with the conventional heat engines in the transportation arena.

For electricity generation from the hydrocarbon fuels, a reformer with carbon monoxide and sulfur cleaning is necessary. It is estimated that the cost of the reforming system is about the same as the fuel cell stack itself, which is also the same as the cost of other ancillary systems. Apart from the high cost, the optimal chemical-to-electric conversion efficiency is around 40%–45%, and the low operating temperature makes the utilization of the waste heat difficult, if possible at all; for the reforming of hydrocarbon fuels, cogeneration of heat and combined cycles can be applied. On the other hand, conventional thermal power plants with combined gas and steam turbines have energy efficiency approaching 60% with a very low capital cost of \$1000/kW. Therefore, the best possible application of the PEMFC systems interesting to utility industry is the use of PEMFCs in the size of tens to hundreds of kilowatts for remote region, as well as a possibility for residential combined heat and power application.

In addition, NASA has been conducting a feasibility study of using the PEMFC power systems for its space programs (mainly space shuttle missions) in place of its current three 12 kW AFC power modules. As discussed in the "Alkaline Fuel Cells" section, NASA is motivated by the extremely high cost, low lifetime, and maintenance difficulty associated with its current AFC systems.

Direct Methanol Fuel Cells

Introduction

All the fuel cells reviewed above for commercial applications require the use of gaseous hydrogen directly or liquid/solid hydrocarbon fuels, e.g., methanol, reformed to hydrogen as the fuel. Pure oxygen or oxygen in air is used as the oxidant. Hence, these fuel cells are often referred to as hydrogen-oxygen or hydrogen-air types of fuel cells. The use of gaseous hydrogen as a fuel presents a number of practical problems, such as storage system weight and volume as well as handling and safety issues especially for consumer and transportation applications. Although liquid hydrogen has the highest energy density, the liquefaction of hydrogen needs roughly one-third of the specific energy, and the thermal insulation required increases the volume of the reservoir significantly. The use of metal hydrides decreases the specific energy density, and the weight of the reservoir becomes excessive. The size and weight of a power system are extremely important for transportation applications, as they directly affect the fuel economy and vehicle capacity, although they are less critical for stationary applications. The low volumetric energy density of hydrogen also limits the distance between vehicle refueling.

Methanol as a fuel offers ease of handling and storage, and potential infrastructure capability for distribution. Methanol also has a higher theoretical energy density than hydrogen (~5 kWh/L compared with 2.6 kWh/L for liquid hydrogen). Easy refueling is another advantage for methanol. However, in the conventional hydrogen-air or hydrogen-oxygen fuel cells, a reformer is needed, which adds complexity and cost as well as production of undesirable pollutants such as carbon monoxide. The addition of a reformer also increases response time.

Therefore, direct oxidation of methanol is an attractive alternative in view of its simplicity from a system point of view. The DMFCs utilizing proton exchange membrane (PEM) have the capability of efficient heat removal and thermal control through the circulating liquid, as well as elimination of humidification required to avoid membrane dryout. These two characteristics have to be accounted for in the direct and indirect hydrogen systems that impact their volume and weight and, consequently, the output power density.

Basic Operating Principle

Figure 3 illustrates the basic operating principle of a DMFC. The DMFC allows the direct use of an aqueous, low-concentration (3%), liquid methanol solution as the fuel. Air is the oxidant. The methanol and water react directly in the anode chamber of the fuel cell to produce carbon dioxide and protons that permeate the PEM and react with the oxygen at the cathode. The half-cell reactions are as follows:

$$\text{Anode}: \mathrm{CH_3OH}(1)+\mathrm{H_2O}(1)=\mathrm{CO_2}(\mathrm{g})+6\mathrm{H}^+ +6\mathrm{e}^-$$

$$\text{Cathode}: 6\mathrm{H}^+ +6\mathrm{e}^- +3/2\mathrm{O_2}(g)=3\mathrm{H_2O}(1)$$

and the net cell reaction is

$$\mathrm{CH_3OH}(1)+3/2\mathrm{O_2}(g)=\mathrm{CO_2}(\mathrm{g})+2\mathrm{H_2O}(1).$$

Because the PEM (e.g., Nafion membrane) is used as the electrolyte, the cell operating temperature must be less than the water boiling temperature to prevent the dryout of the membrane. Typically, the operating temperature is around 90°C, and the operating pressure ranges from 1 atm to several atmospheres.

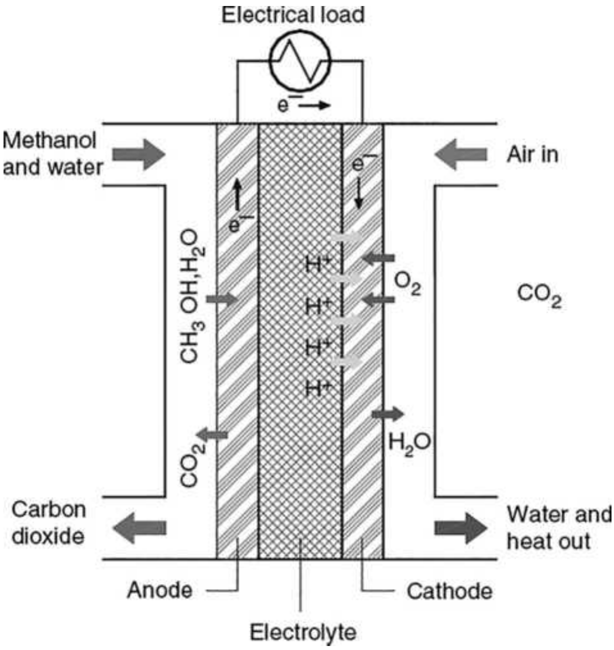


FIGURE 3 Schematic of DMFC.

Acceptable Contamination Levels

The system is extremely sensitive to carbon monoxide (CO) and hydrogen sulfide (H_2S). Carbon monoxide may exist as one of the reaction intermediaries and can poison the catalyst used. There are arguments whether CO is present in the anode during the reaction. Sulfur may be present if methanol is made of petroleum oils and needs to be removed.

Major Technological Problems

The PEM used in the DMFCs is the same as that employed in the PEMFCs. Since the electrolyte in DMFCs is essentially acid, expensive precious metals (typically platinum or its alloys) are used as the catalyst. However, the most serious problem is the so-called "methanol crossover." This phenomenon is caused by the electro-osmotic effect. When the protons migrate through the electrolyte membrane, a number of water molecules are dragged along with each proton, and because methanol is dissolved in liquid water on the anode side, methanol is dragged through the membrane electrolyte to reach the cathode side together with the protons and water. Fortunately, the methanol at the cathode is oxidized into carbon dioxide and water at the cathode catalyst sites, producing no safety hazards. However, the methanol oxidation in cathode does not produce useful electric energy. The development of a new membrane with low methanol crossover is a key to the success of DMFCs.

Such a low methanol crossover membrane has a number of advantages. First, it reduces the methanol crossover, enhancing fuel utilization and, hence, energy efficiency. Second, it reduces the amount of water produced at the cathode, leading to a lower activation and concentration polarization, thus allowing higher cell voltage at the same operating current. Third, it allows higher methanol concentration in the fuel stream, resulting in better performance.

Technological Status

Although methanol itself has simpler storage requirements than hydrogen and is simpler to make and transport, its electrochemical activity is much slower than that of hydrogen; that is, its oxidation rate is about four orders of magnitude smaller than that of hydrogen. Also, the conversion takes place at low temperature (about 80–90°C), and the contaminant problem is one of the serious issues.

The state-of-the-art performance is an energy conversion efficiency of 34% (from methanol to electricity) at 90°C using 20 psig air, that is, a cell voltage of 0.5 V (corresponding to a voltage efficiency of 42%) together with the methanol crossover accounting for 20% of the current produced (equivalent to a fuel efficiency of 80%). This 20% methanol crossover occurs when the fuel stream used is an aqueous solution containing only 3% methanol. It has been projected that with the better membrane developed by the University of Southern California (USC) and improvement of membrane electrode assembly, a cell voltage of 0.6 V can be achieved with only 5% methanol crossover. This is equivalent to 50% voltage efficiency and 95% fuel efficiency, resulting in an overall stack efficiency of 47% (from methanol to electricity). The DMFC system efficiency will be lower due to running the necessary auxiliary systems. Since the current DMFCs basically use the same cell components, materials, construction, and fabrication techniques as the PEMFCs, it is therefore expected that the system and component costs will be similar to that of the PEMFCs.

Applications

DMFCs have low power density due to the sluggish reaction kinetics of methanol, but high volumetric energy capacity since the volumetric energy density of methanol is high. These characteristics make DMFCs unlikely suitable for high power density applications such as powering large vehicles, but ideal for smaller applications such as laptops, cellular phones, digital cameras, battery chargers, or other electronic devices. Current DMFC systems are available with power outputs from 25 W to 5 kW with durations up to 100 hr between refueling operations. DMFCs therefore feature longer working period and shorter refueling/recharging time than conventional lithium batteries. In addition, since methanol can be made from agriculture products, the use of methanol is also compatible with renewable energy sources to allow for sustainable development.

Concluding Remarks

A summary of the preceding discussion, including the operational characteristics and technological status of the three major types of low-temperature fuel cells, is given in Table 1.

For stationary power generation in the utility industry, the economic fuels of today and in the near future are the fossil fuels. Although AFCs have the best performance when operating on pure hydrogen and oxygen, the liquid electrolyte's intolerance of carbon dioxide eliminates its role for utility

TABLE 1 Operational Characteristics and Technological Status of Various Low-Temperature Fuel Cells

Type of Fuel Cells	Operating Temperature (°C)	Power Density (mW/cm^2), (Present) Projected	Projected Rated Power Level (kW)	Fuel Efficiency	Lifetime Projected (hr)	Capital Cost Projected ($/kW)	Application Areas
AFC	<90	(100–400) >300	10–100	4070	>10,000	35	Space, transportation
PEMFC	50–90	(600) >900	1–1000	45–70	>40,000	35	Transportation, space
DMFC	90	(230)	0.025–5	34	>10,000	35	Portable electronic devices, transportation

applications. However, the development of solid electrolyte membrane for AFC lifts the curtain of revival of this technology. The cost of PEMFCs has been significantly reduced in recent years, and it is expected to be comparable with internal combustion engines. PEMFCs are believed to be the most promising candidate for transportation application because of their high power density, fast start-up, high efficiency, and easy and safe handling. DMFCs are most suitable for portable applications such as laptops, cellular phones, digital cameras, battery chargers, or other electronic devices, because they have low power density but high energy capacity.

Bibliography

1. Blomen, L.J.M.J.; Mugerwa, M.N. *Fuel Cell Systems;* Plenum Press: New York, 1993.
2. Larminie, J.; Dicks, A. *Fuel Cell Systems Explained,* 2nd Ed.; Wiley, Chichester, 2003.
3. Kordesch, K.; Simader, G. *Fuel Cells and Their Applications;* VCH Press: New York, 1996.
4. Appleby, A.J.; Foulkes, F.R. *Fuel Cell Handbook;* Van Nostrand Reinhold Company: New York, 1989.
5. Hoogers, G. *Fuel Cell Technology Handbook;* CRC Press: New York, 2002.
6. Vielstich, W.; Gasteiger, H.; Lamm, A. *Handbook of Fuel Cells—Fundamentals;* Technology, Applications, 4-volume set; Wiley, Chichester, 2003.
7. Li, X. *Principles of Fuel Cells;* Taylor and Francis: New York, 2005.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

36

Global Climate Change: Gasoline, Hybrid-Electric, and Hydrogen-Fueled Vehicles

Robert E. Uhrig

Introduction

It is theoretically possible for plug-in hybrid-electric light transportation vehicles to utilize electricity provided by electric utilities to displace almost 75% of the energy of gasoline used for light transportation vehicles (automobiles, SUVs, pickup trucks, minivans, etc.). It was also shown that replacing this gasoline energy with electricity would require 200–250 GW of new electrical generating capacity. Calculations show that about 930 GW of new electrical generating capacity would be required to produce hydrogen by electrolysis to replace all hydrocarbon fuels used in all U.S. transportation systems (including heavy trucks, aircraft, military vehicles, etc.). Hence, the choice of fuels for of these new electrical generating plants could have a large environmental impact if a significant fraction of gasoline used for light and heavy vehicle transportation was replaced by electricity or hydrogen.

Greenhouse Gases

Sunlight enters the atmosphere striking the earth, and it is reflected back towards space as infrared radiation (heat energy). Greenhouse gases absorb this infrared radiation, trapping the heat in the atmosphere. Until about 150 years ago, the amount of energy absorbed was about the same as the amount of energy radiated back into space. At that time, this equilibrium was upset, and the concentration of greenhouse gases—especially CO_2 (carbon dioxide)—began to increase. The concentration is continuing to increase at a rate that seems to parallel the increase in production of CO_2. During the same period, the average temperature of the atmosphere increased with subsequent climate changes in a manner that is thought by many to be a cause–effect relationship.[1]

The principal greenhouse gases are water vapor (60%–65%) and CO_2 (20%–25%). Water vapor stays in the atmosphere for a relatively short time—a matter of hours or days—whereas CO_2 has an average residence period of about a century. As a result, the amount of water vapor remains relatively constant related to the rate at which it is produced by weather phenomena, while CO_2 tends to accumulate as the amount emitted increases. All other greenhouse gases (10%–20%) such as methane, ozone, nitrous oxide, and carbon monoxide tend to have less effects on the atmosphere over time because of their small quantities or short residence times. Hence, the only greenhouse gas considered in this analysis is CO_2.

In contrast with the “well to wheels” approach that includes the emissions produced by processing the original fuel, this analysis starts with the fuel used to generate electricity or to produce hydrogen. This simplification does not materially change the relationship between the emissions of greenhouse gases by the various processes analyzed.

Models Used for Analyses

The models to quantitatively evaluate the greenhouse gas emissions, specifically carbon dioxide, when electrical energy or hydrogen is used to replace gasoline as an automotive fuel are those used in the author’s previous publications.[2,3] These models were based on information provided by or extrapolated from data provided by the Department of Energy (DOE) Energy Information Administration (EIA).[4]

These models utilize the following data:

1. There were 225 million light transportation vehicles in the United States in 2004, including 133 million automobiles and 92 million light truck-based vehicles (sport utility vehicles, vans, pickup trucks, passenger minivans, and delivery vans).
2. On any given day, 50% of these vehicles traveled less than 32.2 km (20 mi).
3. The average distance traveled by each vehicle in the United States was 19,749 km/year (12,264 mi/year). This distance traveled by each average vehicle is derived from the 9 million barrels of oil per day used to make gasoline for the 225 million light vehicles having an average fuel consumption of 8.51 km/l (20 mi/gal), the approximate average gasoline mileage reported by EIA.[4]

4. The quantitative index used for comparing electricity and hydrogen produced by different methods using different source fuels is kilograms of CO_2 per vehicle-year (pounds per vehicle-year) for the average distance traveled by each vehicle in the United States—19,749 km/year (12,264 mi/year). Actual emissions of specific vehicles may be greater or less than the values presented here depending upon whether the fuel consumption of the particular vehicle is greater or less than the average value of 8.51 km/l (20 mi/gal).

CO_2 Emissions for Gasoline-Fueled Light Vehicles

TerraPass, a vendor dealing in carbon dioxide emission credits, has developed a "carbon dioxide calculator" that gives the emission of carbon dioxide for virtually all American automobiles for the past 20 years.[5] Experimenting with this calculator shows that the quantity of carbon dioxide, the most important greenhouse gas from combustion of gasoline in internal combustion engines, emitted per unit of time is a direct function of the amount of gasoline used in that time. This calculator utilized the fact that the combustion of 1 gal of gasoline emits 2.35 kg of CO_2 per liter (19.56 lbs of CO_2 per gallon) of gasoline. Because the average gas mileage in our model is 8.51 km/l (20 mi/gal), the annual emission of carbon dioxide is

$$\frac{19749\text{ km/year}\times\text{kg CO}_2\text{ per liter}}{8.51\text{ km/liter}}=5458\text{ km/vehicle}-\text{year}$$

$$\left(11{,}997\ lb\ CO_2\ \text{ per vehicle-year}\right)$$

Clearly, the amount of CO_2 for smaller vehicles is less than for larger vehicles because the gasoline mileage is greater. However, the model for this analysis deals with average vehicles that achieve 8.51 km/l (20 mi/gal)—the reference performance used for all comparisons.

CO_2 Emissions for Micro, Mild, and Full Hybrid Vehicles

Because of the energy to propel all traditional types of hybrid vehicles is provided by gasoline, we will utilize the increased fuel mileages assigned in the author's earlier publication[3] and reduce the total CO_2 emissions for traditional hybrid-electric vehicles accordingly, as shown in Table 1.

These values are for hybrid vehicles of a size corresponding to a traditional vehicle that attains 8.51 km/l (20 mi/gal). A larger or smaller vehicle would have higher or lower emissions, respectively.

CO_2 Emissions for Plug-In Hybrid-Electric Vehicles

The model utilized in the author's hybrid article[3] has half the vehicles traveling 24.2 km (15 mi) per day on electricity, or 8816 km (5475 mi) per year. The other half of the vehicles travel 56.4 km (35 mi) per day, or 20,572 km (12,775 mi) per year on electricity. These assumptions result in an average of 14,694 km (9125 mi) per year on electricity for all vehicles.

TABLE 1 Emissions for Hybrid-Electric Vehicles

	Gasoline Mileage		CO_2 per Vehicle-Year	
Type Hybrid	km/l	mi/gal	Kg	lb
Micro hybrid	9.35	22	4,955	10,906
Mild hybrid	10.62	25	4,362	9,598
Full hybrid	12.32	29	3,761	8,274

The second half of these vehicles must also travel an additional 5055 km (3139 mi) per year as a full hybrid using gasoline at 12.32 km/l (29 mi/gal). Because the average total distance traveled by each car per year is 19,749 km (12,264 mi), the distance traveled using gasoline is 25.6% of the average distance each vehicle travels. The other 74.4% of this distance is traveled using electricity generated by a utility using nuclear fuels, fossil fuel, solar energy (wind, hydro, or photovoltaic systems), or renewable energy systems. Such a substitution of electricity for gasoline would save 6.7 million of the 9.0 million barrels of oil per day used in the United States for light vehicle transportation.

Electricity Generated Using Nuclear, Solar, and Renewable Fuels

Plug-in hybrid vehicles have two sources of CO_2—the CO_2 emitted due to operation in the full-hybrid mode and the CO_2 emitted in generating the electrical energy used for the rest of the time. Because nuclear, solar, and renewable fuels generate a net of zero CO_2, the only CO_2 generated is by operation in the full-hybrid mode for 25.6% of the distance traveled. Hence, the emission is 25.6% of the value in Table 1 for a full-hybrid, 964 kg (2122 lb) per vehicle-year.

Electricity Generated Using Fossil Fuels (Natural Gas, Oil, and Coal)

When the electricity is supplied by utilities using fuels that emit carbon dioxide during combustion, this CO_2 from generating electricity must be added to the CO_2 of the full-hybrid operation discussed above. Data on actual emissions from fossil power plants in the United States for 1999 (the last year for which complete data is available), provided by EPA and (DOE) (DOE-EIA 1999) show the emission rates in Table 2.

A reasonable average of 0.374 kWh/km (0.603 kWh/ mi) for the expenditure of electrical energy for hybrid vehicles that would be comparable to a vehicle getting 8.51 km/l (20 mi/gal) of gasoline using an internal combustion engine. Previously in this entry, it was shown that the average distance traveled by all plug-in hybrid vehicles while operating on electricity alone was 14,694 km/year (9125 mi/year). Hence, the electricity used is

$$\left(14{,}694\text{ km/year}\right) \times \left(0.374\text{ kWh/km}\right) = 5{,}496\text{ kWh/year}.$$

If we multiply this value by the amount of CO_2 emission per kilowatt-hour for the three fossil fuels given in Table 2 and add the 964 kg/vehicle-year (2122 lb/vehicle-year) of CO_2 for operation in the hybrid mode, we get the results for fossil fuels given in Table 3.

TABLE 2 Emission Rates for Fossil Fuels

Fuel to Generate Electricity	% of Total Generation	CO_2 Emission Rate kg/kwh	CO_2 Emission Rate lb/kwh
Coal	51.0	0.952	2.095
Oil	3.2	0.895	1.969
Natural gas	15.2	0.600	1.321
Renewables	0.6	0 (net)	0 (net)
Non-fossil[a]	30.0	0	0

[a] Nuclear, Solar, Wind, and Hydro.

Source: Energy Information Administration (see DOE–EIA[6]).

TABLE 3 Emissions for Fossil Fuels

	Emissions of CO_2	
Fuel to Generate Electricity	kg/Vehicle-Year	lb/Vehicle-Year
Coal	6,205	13,650
Oil	5,889	12,956
Natural gas	4,271	9,396

Source: Energy Information Administration (see DOE–EIA[6]).

Hydrogen-Fueled Vehicles

The emission of carbon dioxide from vehicles utilizing electricity generated with fuel cells operating on hydrogen is negligible except for the emissions from the various processes used to produce the hydrogen. These emissions of carbon dioxide from the production processes must be taken into account to give a valid comparison with emissions of greenhouse gases from the other vehicle configurations and their potential impact upon climate and weather modification.

Steam Methane Reforming

Some 95% of hydrogen used today is produced by steam methane reforming (SMR) from natural gas (~98% methane—CH_4). The two steps of SMR are

$$CO_4 + H_2O + \text{Heat}(206\text{ kJ/mol}) \overset{\text{Catalyst}}{\rightarrow} CO + 3H_2$$

$$(\text{Steam Reforming Reaction})$$

$$CO + H_2O \overset{\text{Catalyst}}{\rightarrow} CO_2 + H_2 + \text{Heat}(41\text{ kJ/mol})$$

$$(\text{Water–Gas Shift Reaction})$$

The hydrogen comes from the methane and the steam. The steam reforming reaction is endothermic with the required 206 kJ/mol heat energy normally produced by combustion of some of the methane. The water-gas shift reaction is exothermic, providing 41 kJ/mol heat energy that, if recovered, can reduce the amount of methane burned. Then the methane burned has to provide only the net 165 kJ/mol that represent ~17% of the total methane energy. About 3.3 mol of hydrogen (~83% of the theoretical maximum of 4 mol of hydrogen from CH_4) are produced for each mol of methane. A well-designed SMR plant will yield hydrogen having about 80% of the energy of the methane supplied. Unfortunately, SMR produces CO_2 in both the methane combustion and in the water–gas shift reaction.

Emissions of Vehicles Using Hydrogen Produced by SMR

In an earlier publication,[3] it was established that 0.228 kWh/km (0.367 kWh/mi) of mechanical energy at the tire- pavement interface corresponded to 8.51 km/l (20 mi/gal). If we use the following efficiencies:

- 70% efficiency for the electric motor drive (electricity to mechanical energy at the tire-pavement interface), the same used in the above reference,
- 60% efficiency (hydrogen to electricity) for the fuel cell, and
- 85% efficiency is distributing and dispensing the hydrogen to the fuel cell, the needed energy of the hydrogen input to the fuel cell is

$$\frac{0.228\ kWh/km}{0.70\times0.60\times0.85}$$
$$=0.638\ \text{kWh/km}\ (1.028\text{kWh / mi})$$

of hydrogen energy. Because the average total distance traveled per vehicle is 19,749 km/year (12,264 mi/year), the total kilowatt-hour per vehicle-year is

$$(19749\ \text{km / year})\times(0.638\ \text{kWh of H}_2\ \text{energy / km})$$

$$=12607\ \text{kWh of H}_2\ \text{energy / vehicleyear.}$$

The energy of hydrogen (lower heating value) is 119.9 MJ/kg (51,600 Btu/lb), so the amount of H_2 used per year is

$$\frac{12607\ \text{kWh per vehicle} - \text{year}}{119.9\ \text{MJ / kg}\times0.278\ \text{kWh / MJ}}$$
$$=379\text{kg H}_2\ \text{per vehicle} - \text{year}$$

$$(834\ lb\ H_2\ \text{per vehicle-year})$$

Goswami indicates that SMR produces a net of 0.43 mol of CO_2 for each mol of H_2 produced using SMR.[7] Because the molecular weights of CO_2 and H_2 are 44 and 2, respectively, the specific CO_2 emission can be calculated by

$$(0.43\ \text{mol}\ CO_2\ /\ mol\ H_2)\times\frac{(44\ gm\,/\,\text{mol}\ CO_2)}{(2\ gm\,/\,\text{mol}\ H_2)}$$
$$=9.46\ \text{gm CO}_2\ /\ \text{gm H}_2=9.46\ \text{kg CO}_2\ /\ \text{kg H}_2$$

and the total emission of CO_2 per vehicle-year is

$$(379\ \text{kg H}_2\text{per vehicle} - \text{year})$$

$$\times(9.46\ kg\ CO_2\ \text{per}\ kg\ H_2)$$

$$=3{,}585\ \text{kg CO}_2\ \text{per vehicle} - \text{year}$$

$$(=7{,}888\ \text{lb CO}_2\ \text{per vehicle} - \text{year}).$$

Emission for Vehicles Using Hydrogen Produced by SMR with Heat Provided by a High-Temperature Nuclear Reactor

Recent work in Japan has demonstrated the feasibility of substituting high-temperature heat from a gas-cooled nuclear reactor to replace the heat supplied by the combustion of methane. This increases the amount of hydrogen produced to 4 moles per mole of methane and eliminates the CO_2 produced by combustion of methane, but not the CO_2 produced by the water–gas shift reaction. The overall reaction of the steam reforming and water shift reactions is the production of 1 mole of CO_2 and 4 mol of hydrogen for each mole of CH_4. Hence, multiplying this ratio by the molecular weights and the amount of hydrogen used per year gives

$$\frac{1\text{ mol CO}_2}{4\text{ mol H}_2}\times\frac{44}{2}\times 379\text{ kg H}_2\text{ per vehicle}-\text{year}$$

$$=2{,}084\text{ kg CO}_2\text{ per vehicle}-\text{year}$$

$$\left(4{,}586\ lb\ CO_2\text{ per vehicle-year}\right).$$

Emissions for Vehicles Using H_2 Produced by Electrolysis

Conventional electrolysis of water to produce hydrogen is a well-developed technology and production units as large as 10 MWe are commercially available today. However, the typical overall efficiency of hydrogen production using electrolysis based on the thermal content of the generating plant fuel is about 25% today, consisting of two components—about 33% efficiency in converting fossil or nuclear fuel to electricity and about 75% in using electricity to separate water into hydrogen and oxygen. If a high-temperature gas-cooled reactor or a modern high-efficiency gas-fired combined cycle plant is used, the overall efficiency in producing hydrogen could approach 45%.

A leading manufacturer of electrolysis equipment indicated that 1 MW of electricity can generate 0.52 ton (1040 lb) of hydrogen per day or 0.473 kg H_2/kW day (1.04 lb H_2/kW day).[8] In the case of SMR, it was calculated that the hydrogen required per year for a vehicle using hydrogen to travel 12,260 mi/year was 379 kg of H_2/ vehicle-year (832 lb/vehicle-year). Hence, the amount of electricity required per year is

$$\frac{\left(379\text{ kg H}_2\text{ / year}\right)\times\left(24\text{ h / day}\right)}{0.473\text{ kgH}_2\text{ / kW day}}=19{,}230\text{ kWh / year.}$$

The carbon dioxide emitted for hydrogen-fueled vehicles is the carbon dioxide emitted in producing the electricity required to produce the hydrogen. The choices of fuel for generating electricity are coal, oil, natural gas, nuclear energy, and solar energy (wind, photovoltaics, and hydro). Hence, the emissions of carbon dioxide are the product of the kilowatt-hour per year and the appropriate emission per kilowatt-hour from Table 2 for the fuel used.

Emissions for Vehicles Using H_2 Produced by Electrolysis Using Solar, Nuclear, and Renewable Energy

Because none of these energy sources emit carbon dioxide when generating electricity, there are no emissions of greenhouse gases associated with these arrangements.

Emissions for Vehicles Using H_2 Produced by Electrolysis Using Fossil Energy (Coal, Oil, and Natural Gas)

The emissions of carbon dioxide for these fossil fuels are the products of the 19,230 kWh/year and the appropriate emission per kilowatt-hour for the fuel as given in Table 2. The results are shown in Table 4.

TABLE 4 Emissions for Fossil Electrolysis

Fuel	kg CO_2/Year	lb CO_2/Year
Natural gas	11,538	25,384
Oil	17,211	37,963
Coal	18,307	40,276

Emissions for Vehicles Using H_2 Produced by Thermochemical Processes

The overall efficiency of the thermochemical process of producing hydrogen, such as the sulfur–iodine process, approaches 50% at 900°C, while the overall efficiency of the electrolysis process is about 25%. Hence, for the same fuel, the carbon dioxide emissions will be half as much for thermochemical processes as for electrolysis.

Emissions for Vehicles Using Hydrogen Produced by Thermochemical Processes Using Solar, Nuclear, and Renewable Energy

Because none of these energy sources emit carbon dioxide when generating electricity, there are no emissions of carbon dioxide associated with these arrangements.

Emissions for Vehicles Using Hydrogen Produced by Thermochemical Processes Using Fossil Energy (Natural Gas, Oil, and Coal)

The emissions of carbon dioxide for these fossil fuels are half of those for electrolysis, as shown in Table 4. The results are shown in Table 5.

Results and Conclusions

The results of this analysis are presented in Tables 6 and 7, which provide emissions of carbon dioxide for the various hybrid and hydrogen-fueled vehicle arrangements for the reference case of 8.50 km/l (20 mi/gal) for the average vehicle being driven 19,749 km/year (12,264 mi/year).

TABLE 5 Emissions for Hydrogen Produced by Thermochemical Processes by Fossil Fuels

Fuel	kg CO_2year	lb CO_2year
Natural gas	5,769	12,692
Oil	8,605	18,932
Coal	9,153	20,138

TABLE 6 Comparison of Emissions of Carbon Dioxide for Hybrid Vehicles

Vehicle Emissions of Carbon Dioxide	kg CO_2 Vehicle-Year	lb CO_2 Vehicle-Year
Reference gasoline vehicle average 8.51 km/l (20 mi/gal)	5,500	12,000
Micro-hybrid vehicle 9.35 km/l (22 mi/gal)	4,950	10,900
Mild-hybrid vehicle 10.62 km/l (25 mi/gal)	4,350	9,600
Full-hybrid vehicle 12.32 km/l (29 mi/gal)	3,750	8,300
Plug-in hybrid vehicle		
Electricity generated with nuclear	950	2,100
Electricity generated with solar energy	950	2,100
Electricity generated with renewable	950	2,100
Electricity generated with natural gas	4,250	9,400
Electricity generated with oil	5,900	13,000
Electricity generated with coal	6,200	13,700

Vehicle travels 19,749 km (12,264 mi) per year.

TABLE 7 Comparison of Emissions of Carbon Dioxide for Hydrogen-Fueled Fuel Cell Vehicles

Vehicle Emissions of Carbon Dioxide	kg CO_2/Vehicle-Year	lb CO_2/Vehicle-Year
Reference gasoline vehicle utilized an average 8.50 km/l (20 mi/gal)	5,500	12,100
Vehicle uses hydrogen produced using SMR	3,600	7,900
Vehicle using hydrogen produced using SMR with heat supplied by		
High temperature gas cooled reactor	2,100	4,600
Vehicle using hydrogen produced using electrolysis with electricity		
Produced using natural gas	11,550	25,400
Produced using oil	17,200	37,900
Produced using coal	18,300	40,300
Produced using nuclear energy	0	0
Produced using solar energy (wind, hydro, and photovoltaic)	0	0
Produced using renewables	0	0
Vehicle using hydrogen generated using a thermochemical process with heat		
Generated using natural gas	5,750	12,700
Generated using oil	8,600	18,900
Generated using coal	9,150	20,100
Generated using nuclear energy	0	0
Generated using solar energy (wind, hydro, and photovoltaic)	0	0
Generated using renewables	0	0

Vehicle travels 19,749 km (12,264 mi) per year.

Plug-In Hybrid-Electric Vehicles

In Table 6, the vehicle emissions of CO_2 for all traditional hybrid-electric vehicles are inversely related to the gas mileage. The gasoline mileage for micro, mild, and full hybrids are assumed to be 9.35, 10.62, and 12.32 km/l (22, 25, and 29 mi/gal), respectively. These are reasonable overall values for the newer, larger hybrid vehicles that have larger gasoline engines and electric motors such as the Toyota Highlander and Honda Accord hybrids. The rounded emissions in Table 6 for electricity generated using oil and coal are greater than for the reference case of gasoline. However, the emission of a plug-in hybrid vehicle where the electricity is generated by nuclear or solar energy is less than 20% of the reference gasoline case. Hence, the only way to significantly improve the greenhouse gas situation if plug-in hybrids are implemented on a large scale is to use only nuclear energy, solar (wind, hydro, or photovoltaic) energy, or renewables to generate the needed electricity. Because hydro power is virtually out of the question due to limited sites and environmental concerns, renewables have large land requirements, and both photovoltaic and wind are intermittent in nature with availabilities of less than about 30%, nuclear power would appear to be the primary choice.

Hydrogen-Fueled Fuel Cell Vehicles

The vehicle emissions of CO_2 for hydrogen-fueled fuel-cell-driven vehicles are presented in Table 7, where the emissions have been rounded off. Because fuel cells do not emit CO_2, the only emissions are from the processes used to produce the hydrogen. The carbon dioxide emissions for hydrogen produced by steam methane reforming, electrolysis, and thermochemical methodologies are given.

Perhaps the most important observation is that, contrary to the widespread belief that hydrogen-fueled vehicles produce no greenhouse gases, the emissions for hydrogen- fueled vehicles vary widely depending upon the method used to generate the electricity or heat to produce the hydrogen. In the case of electrolysis, in which the electricity is generated using nuclear, solar, and renewable energy, the average annual emissions per vehicle-year are negligible. However, the emissions for fossil fuels (natural gas, oil, and coal) are about 11,500, 17,200, and 18,300 kg/year (25,300, 37,800, and 40,200 lb/year), respectively—almost three times those for plug-in hybrid vehicles. These are extremely large emissions compared to almost any other method investigated here.

The emissions for producing hydrogen for a hydrogen- fueled vehicles using SMR produces CO_2 by two methods: combustion of some of the methane to drive the reaction and the water–gas shift reaction. These reactions produce about 3600 kg (7900 lb) CO_2 per vehicle-year, about the same as a full-hybrid vehicle.

If the heat required to drive the reaction in steam methane reforming can be supplied by an outside source such as a high temperature gas cooled reactor, then the emission of CO_2 is reduced to about 2100 kg (4600 lb) per vehicle- year.

The conclusion is that using hydrogen in fuel cells to propel vehicles is a complex process with many steps (converting thermal energy to electricity to hydrogen to electricity to propelling the vehicle), each consuming considerable energy. Hence, the resulting CO_2 emissions when fossil fuels are used to generate the electricity for electrolysis or thermochemical processes are correspondingly high. Given the concerns about the influence of CO_2 on weather and climate, it seems compelling that fossil fuels not be used in producing electricity for electrolysis or thermochemical processes to produce hydrogen for transportation.

References

1. DOE–EIA, Energy Information Administration, *Greenhouse Gases, Climate Change, and Energy,* 2005; at http://www.eia.doe.gov/oiaf/1605/ggccebro/chapter1.html (accessed May 2006).
2. Uhrig, R.E. Engineering challenges of the hydrogen economy. The BENT of Tau Beta Pi, 2004; XCL (2), 10–19.
3. Uhrig, R.E. Using plug-in hybrid vehicles to drastically reduce petroleum-based fuel consumption and emissions. The BENT of Tau Beta Pi, 2005; XCLI (2), 13–19.
4. DOE–EIA, Energy Information Administration, Petroleum Flow 2002. *Annual Energy Review,* 2003; available at http://www.eia.doe.gov/emeulaer/diagram2.html (accessed January 2006).
5. TerraPass, *Carbon Dioxide Calculator*, Terrapass.com/ carboncalc.html (accessed January 2005).
6. DOE–EIA, Department of Energy, Energy Information Administration (DOE and EPA), *Carbon Dioxide Emissions from the Generation of Electric Power in the United States*, July 2000.
7. Goswami, D.Y. *Hydrogen Supply Technologies: Introduction to Issues,* 2005; available at http://www.ases.org/hydrogen_forum03/Goswami.pdf (accessed June 2005).
8. Stuart, S. *Nuclear-Electric Synergies: For Neat Hydrogen Production and Cleaner Fossil Reserves Harvesting;* ANS President's Special Session: Hydrogen Systems, ANS Meeting: Reno, NV, 2001.

37

Heat Pumps

Lu Aye

Introduction

This entry provides foundation knowledge about heat pumps for both professional and non-professional readers. A brief history of heat pumps, basic terms used in the heat pump industry and research and the fundamentals of heat pumps are presented. Working principles of the thermoelectric heat pump, the absorption heat pump, the gas compression heat pump and the vapor compression heat pump are explained by using schematic diagrams. Performance parameters of the heat pump systems—coefficient of performance (COP), energy efficiency ratio (EER), primary energy ratio (PER), and ambient energy fraction (AEF)—are also discussed.

Major advances have been made in heat pump technologies over the last 35 years. "Heat pump systems offer economical alternatives of recovering heat from different sources for use in various industrial, commercial and residential applications."[1] Heat pumps in their various and diverse forms allow one to harness, in a cost-effective manner, solar energy that has already been stored in the atmosphere and biosphere. To make a fair systematic comparison with other systems that provide the same thermal output, AEF may be used. The concept and the method of estimating the AEF are discussed. The relationships between the AEF, the heating coefficient of performance (HCOP) and PER are also presented. The short-term and long-term potential use of the heat pumps and corresponding greenhouse gas (GHG) emissions are discussed.

Brief History

Although most people are familiar with the common household refrigerator, the concept of using such a device to provide heating rather than cooling is less widely understood. The basic operational principle of this machine called the heat pump, however, was laid down in thermodynamic terms by Lord Kelvin, the first professor of natural philosophy at the University of Glasgow, Scotland, in the middle of the 19th century. Whereas the thermodynamic principle of the heat pump was found in 1852, it was not until 1855 that it was realized for producing heat by means of an open-cycle mechanical vapor recompression unit in Ebensee, Austria.[2] Much later, the closed vapor compression process was used for generating useful heat. After World War II, heat pump units for air-conditioning homes and individual rooms became common. Now, heat pump technology is well known as one of the energy conservation technologies.

Fundamentals

A heat pump is a thermodynamic system whose function is to heat, at the required temperature, with the aid of heat extracted from a source at lower temperature.[3] Heat pumps are devices designed to utilize low-temperature sources of energy that exist in atmospheric air, in lake or river water, and in the earth. These sources are referred to as "ambient energies." Heat pumps can also be operated using waste heat from commercial and industrial processes, thereby upgrading this to the required temperature level for some thermal process operations.[4]

The heat pump can be considered as a heat engine in reverse.[5] A heat engine removes heat from a high-temperature source and discharges heat to a low temperature and in doing so can deliver work (see Figure 1). Heat pumps capture heat energy from low-grade sources and deliver at a higher temperature. They require external work (W) to upgrade the heat absorbed (Q_L) (see Figure 2). Note that the heat energy flows from a low-temperature heat source to a higher application temperature. This is the

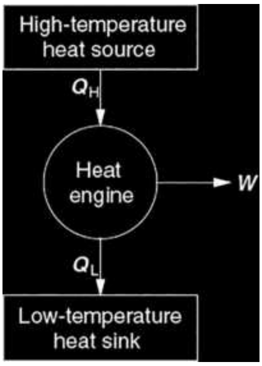


FIGURE 1 Thermodynamic model of a heat engine.

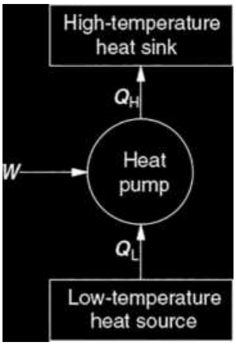


FIGURE 2 Energy flow of a heat pump.

TABLE 1 Commonly Used Heat Sources

Heat Source	Temperature Range (°C)
Ambient air	–10 to 15
Building exhaust air	15–25
Groundwater	4–10
Lake water	0–10
River water	0–10
Seawater	3–8
Rock	0–5
Ground	0–10
Wastewater and effluent	>10

reverse of the natural flow direction of heat from a higher temperature to a lower temperature. For ideal conditions without heat losses, the energy balance of a heat pump provides (Eq. 1):

$$Q_{\mathrm{H}} = Q_{\mathrm{L}} + W \tag{1}$$

where Q_H is the useful heat delivered, Q_L is the heat absorbed and W is the work input.

The useful heat delivered is always greater than the work input to the heat pump for ideal conditions.

The technical and economic performance of a heat pump is closely related to the characteristics of the heat source. Table 1 lists current commonly used heat sources. Ambient and exhaust air, soil and groundwater are practical heat sources for small heat pump systems, whereas sea/lake/river water, rock (geothermal) and wastewater are generally used for large heat pump systems.[6]

The fundamental theories of heat pumps and refrigerators are the same. In engineering practice, a distinction is made between heat pumps and refrigerators. The principal purpose of the latter device is to remove heat from a low-temperature heat source, and the purpose of the former device is to supply heat to a high-temperature sink. In many practical devices, the heating and cooling elements are interchangeable. The term *reverse cycle air conditioner* is frequently applied to heat pumps with interchangeable heating and cooling elements. In Japan and in the United States, reversible air-conditioning units are called heat pumps, and in Europe, the term *heat pump* is used for heating-only units.[2]

Carnot Coefficient of Performance (Theoretical Cop)

The heating performance of a heat pump is measured by the HCOP. The HCOP is the ratio of the quantity of heat transferred to the high-temperature sink to the quantity of energy driving the heat pump (see Figure 2).

$$\mathrm{HCOP} = \frac{Q_{\mathrm{H}}}{W} \tag{2}$$

In the Carnot ideal heat pump cycle, the heat is delivered isothermally at T_H and received isothermally at T_{L}. By using the laws of thermodynamics and definition of entropy, it can be shown that the Carnot HCOP is given by Eq. 3.

$$\mathrm{HCOP} = \frac{T_{\mathrm{H}}}{T_{\mathrm{H}} - T_{\mathrm{L}}} = \frac{T_{\mathrm{L}}}{T_{\mathrm{H}} - T_{\mathrm{L}}} + 1 \tag{3}$$

No practical heat pump constructed can have a better performance than this theoretical ideal HCOP. The best that our practical heat pump cycles can do is struggle towards achieving this ideal performance.[5]

Facts about Heating Relevant to Heat Pumps

It is argued that heat pumps are very energy efficient and therefore environmentally benign. The International Energy Agency (IEA) Heat Pump Centre provides the basic facts about heat supply and discusses the value of heat pumps.[7] The basic facts about heating explained are as follows:

- Direct combustion to generate heat is never the most efficient use of fuel.
- Heat pumps are more efficient because they use renewable energy in the form of low-temperature heat.
- If the fuel used by conventional boilers were redirected to supply power for electric heat pumps with HCOP of 3.5 to 4.5, about 35–50% less fuel would be needed, resulting in 35–50% less emissions.
- Around 50% savings are made when electric heat pumps with HCOP of 4.5 are driven by combined heat and power (CHP) or cogeneration systems.
- Whether a fossil fuel, nuclear energy or renewable energy is used to generate electricity, electric heat pumps make far better use of these resources than do resistance heaters.
- The fuel consumption, and consequently the emissions rate of an absorption or gas-engine heat pump is about 35–50% less than that of a conventional boiler.

Because heat pumps consume less primary energy than conventional heating systems, they are an important technology for reducing the unwanted air pollutants such as respirable particulate matters (PMs), carbon monoxide (CO), nitrogen oxides (NO_x), sulfur dioxide (SO_2), and carbon dioxide (CO_2) that harm the human environment.

Heat Pump Types

Five major types of heat pump may be identified in the literature.[8,9] These are the following:

- the magnetic heat pump
- the thermoelectric heat pump
- the absorption heat pump
- the gas compression heat pump
- the vapor compression heat pump

The Magnetic Heat Pump

The magnetic heat pump requires a solid magnetic material and it is based on the magnetocaloric effect. The magnetocaloric effect is the heating and cooling of a magnetic material in response to the application and removal of a sufficiently high external magnetic field.[10] The principle of the magnetic heat pump is shown in Figure 3.[11] With thermal Switch 1 off and thermal Switch 2 on, a magnetic field is applied to the solid magnetic material. The solid magnetic material is heated up above the heat sink temperature and heat will flow to the heat sink. Next, thermal switch 1 is on while thermal switch 2 remains on, and the magnetic field is partially removed; this cooled down the magnetic material to the heat source temperature. Thermal switch 2 is then off while the magnetic field is totally removed, completing the cooling of the magnetic material, and the magnetic material gains the heat from the heat source. Thermal switch 2 is then on and a magnetic field is applied to heat the magnetic material. Then, thermal switch 1 is off and the cycle is repeated to transfer heat from the low-temperature heat source to the higher-temperature heat sink.

The magnetic heat pumps do not use ozone-depleting refrigerants, and they have compact configuration, low noise, high efficiency, high stability, and longevity. They have great applicable prospects.[12,13]

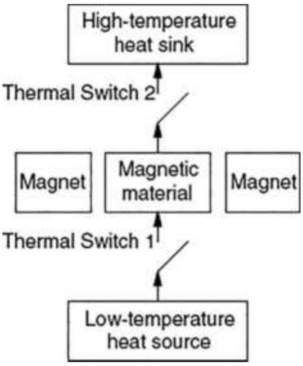


FIGURE 3 Schematic of a magnetic heat pump.

The Thermoelectric Heat Pump

The thermoelectric heat pump uses the Peltier effect. Two materials with different thermoelectric properties (*p* type and *n* type) are arranged to form hot and cold junctions and are connected in series with a direct current voltage source (Figure 4). These devices do not have moving parts and the cooling and heating elements may be reversed simply by changing the direction of the electric current. Current commercially available thermoelectric heat pumps have efficiencies that are well below those of vapor compression heat pumps. They are generally used in conditions where the solid-state nature (no moving parts, maintenance-free) outweighs efficiency, e.g., cooling of electronic components and small instruments.

The Absorption Heat Pump

An absorption cycle is a heat-activated thermal cycle. The absorption heat pump can be considered as a heat engine driving a heat pump (see Figure 5). All absorption cycles include at least three thermal exchanges with their surroundings (i.e., energy exchange at three different temperatures). For the system to operate, the generator temperature (T_g) must be greater than the condenser temperature (T_c) and the absorber temperature (T_a), must in turn be greater than the evaporator temperature (T_e).

The components of a simple single-effect absorption system are as follows:

- generator (desorber)
- condenser
- refrigerant expansion valve
- evaporator

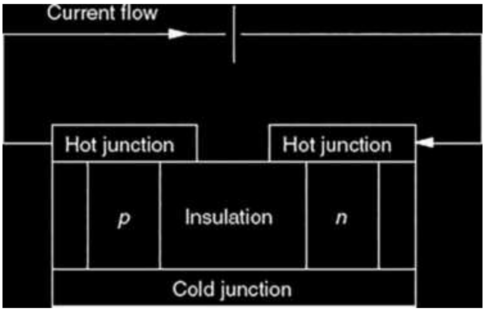


FIGURE 4 Schematic of a thermoelectric heat pump.

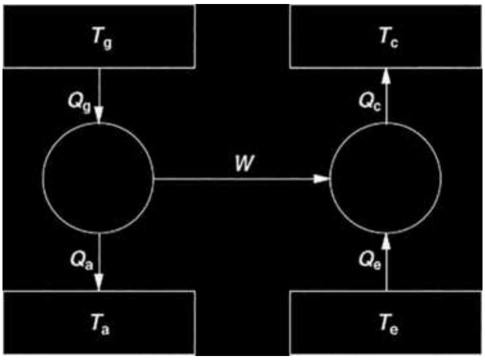


FIGURE 5 Thermodynamic model of an absorption heat pump.

- absorber
- pump
- solution heat exchanger
- solution pressure reducer
- connecting pipes

The absorption system utilizes a sorbent–refrigerant pair, lithium bromide and water, and water and ammonia, being the most widely used. A schematic of the continuous absorption heat pump system is shown in Figure 6. The major exchanges with its surroundings are thermal energy, plus a small amount of mechanical work at the pump to transport the working fluids.

A high-temperature heat source supplies heat to the generator to drive off the refrigerant vapor from solution. The condenser receives the refrigerant vapor from the generator. In the condenser, the refrigerant changes phase from vapor to liquid and heat is rejected at a medium temperature. The refrigerant expansion valve reduces the pressure of the refrigerant from the condensing pressure to

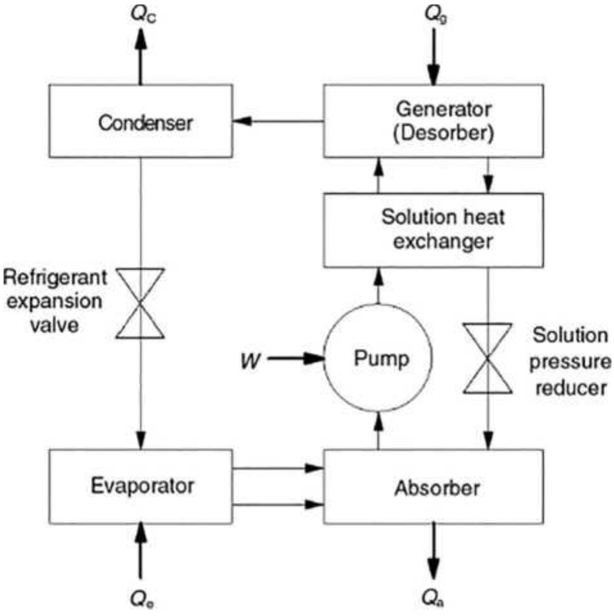


FIGURE 6 Simple absorption cycle.

the evaporating pressure. In the evaporator, the liquid refrigerant collects the low-temperature heat and changes to vapor phase. In other words, the working fluid (refrigerant) provides the cooling effect at the evaporator and the vaporized refrigerant enters the absorber. The strong-concentration sorbent absorbs the refrigerant vapor and becomes a weak solution. Note the heat rejected at the absorber. The low-pressure weak solution (refrigerant and sorbent) is pumped to the solution heat exchanger to exchange heat with the strong-concentration sorbent. The solution pressure reducer accepts the high-pressure strong solution from the generator and delivers low-pressure strong solution to the absorber. The generator receives the high-pressure weak solution from the solution heat exchanger and separates the refrigerant and sorbent.

The Gas Compression Heat Pump

Gas compression heat pumps use a gas as the working fluid. Phase changes do not occur during the cycle of operation. Major components of a gas compression heat pump system are as follows:

- gas compressor
- high-pressure heat exchanger
- gas expander (or turbine)
- drive motor
- low-pressure heat exchanger

A schematic of a gas compression heat pump is shown in Figure 7. The gas is drawn in to the compressor and compressed. The hot high-pressure gas delivers heat at the high-pressure heat exchanger. At the expander (turbine), the high-pressure gas does expansion work that assists to drive the compressor. In general, the external work input to the compressor-expander system is provided by a drive motor.

The Vapor Compression Heat Pump

Almost all heat pumps currently in operation are based on either a vapor compression or an absorption cycle.[14] Heat pumps and refrigerators of the vapor compression type utilize a working fluid that undergoes phase changes during operation. The phase changes occur principally during the heat collection and rejection stages of the cycles. The phase-changing processes occurring in the vapor compression heat pumps are accompanied by extremely favorable heat transfer conditions. Boiling agitation in evaporators and dropwise condensation in condensers make the heat transfer rates very high, which helps to improve the system efficiency. To carry out evaporation and condensation, any practical heat pump uses four separate components as illustrated in Figure 8:

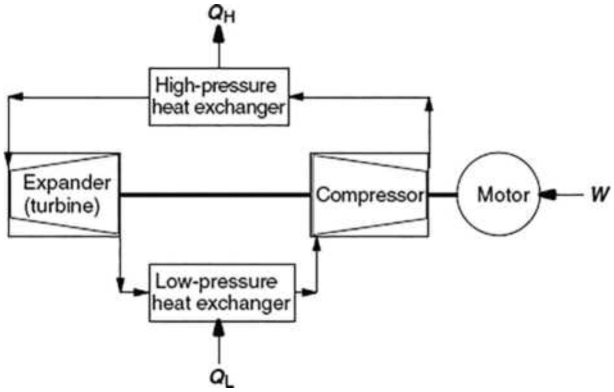


FIGURE 7 Schematic of a gas compression heat pump.

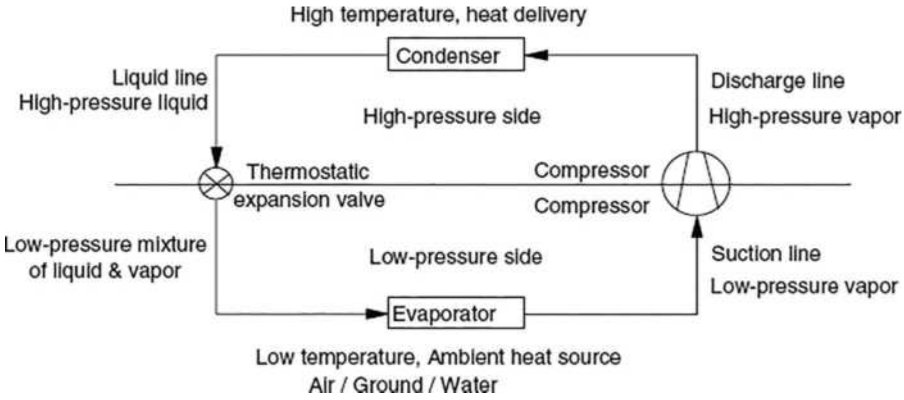


FIGURE 8 Schematic diagram of a vapor compression heat pump.

- compressor
- condenser
- expansion device
- evaporator

The compressor is the heart of the vapor compression heat pump system. By using mechanical power, the compressor increases the pressure of the working fluid vapor received from the evaporator and delivers it to the condenser. The working fluid condenses and provides useful heat at the condenser. The pressure of the liquid working fluid coming out of the condenser is reduced to the evaporating pressure by an expansion device. The low-grade heat is collected at the evaporator and the phase of the working fluid changes from liquid to vapor. The working fluid enters the compressor in vapor form to repeat the cyclic flow.[15]

Compressors used in practical units are generally of the positive displacement type: either reciprocating or rotary vanes. In practice, nearly all vapor compression heat pumps use thermostatic expansion (T-X) valves because of the availability of the valve to handle a wide range of operating conditions and because of the fact that the pressure reduction is essentially isenthalpic. These valves, which are relatively low-cost devices, control the liquid working fluid flow to the evaporator by sensing the superheat condition of the working fluid leaving the evaporator. Such control ensures that nearly all of the available evaporator surface is covered with a forced convection nucleate boiling film with consequential excellent heat transfer characteristics in the evaporation process.

Reversible circuit heat pumps for building cooling and heating are equipped with a four-way exchange valve that can be operated automatically or manually. This valve enables the normal evaporator to become the condenser and the normal condenser to become the evaporator so that the machine can be made to operate in reverse fashion. This arrangement is often incorrectly called reverse-cycle operation because of the end effect achieved—i.e., the interchange of roles for the cooling and heating parts of the circuit. It would be better to call this reverse-flow operation as the basic thermodynamic cycle remains the same for each mode of operation.

In general, vapor compression heat pumps are classified in many different ways. One method commonly used distinguishes among air, water and ground sources of low-temperature energy, and also between the high-temperature energy delivery medium (i.e., air or water). In such a classification, the most common types of heat pump units would include air-to-air units and water-to-water units. Solar energy could also be captured at the evaporator as a heat source. These types of heat pumps are called solar-assisted and solar-boosted heat pumps.[16]

Changes in condensing and evaporating temperatures affect the work input required at the compressor and consequently the COP. Decreasing the condenser or increasing the evaporator temperatures will

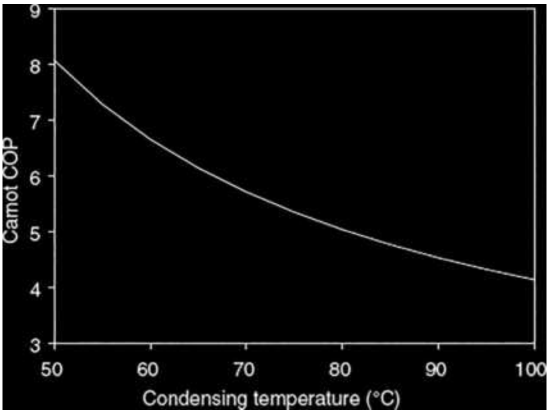


FIGURE 9 Effect of condensing temperature on ideal COP (evaporating temperature = 10°C).

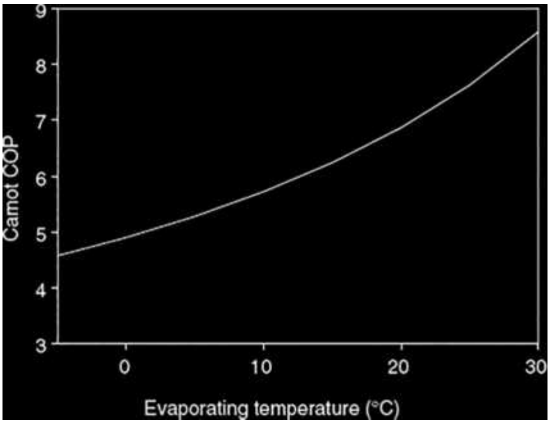


FIGURE 10 Effect of evaporating temperature on ideal COP (condensing temperature = 70°C).

decrease the compressor work and increase the COP. Figure 9 shows the effect of condensing temperature on COP for a fixed evaporator temperature. The effect of the evaporating temperature on the COP for a fixed condenser load temperature is shown in Figure 10. For simplicity, these curves have been obtained assuming operation on the ideal heat pump cycle—i.e., the Carnot cycle of two isothermal and two isentropic (reversible adiabatic) processes not allowing any practical deviations. Although the COP values for real machines will be substantially lower, the trend will be the same.

Performance Parameters

This section defines and explains some performance parameters that are used to compare various heat pumps for a particular application and also presents the relationships among them.

SCOP and EER

System coefficient of performance (SCOP) is defined as output heating capacity per unit of power input to the system (see Eq. 4).

$$\text{SCOP} = \frac{\text{Heating capacity (W)}}{\text{Power input (W)}} \tag{4}$$

SCOP is a dimensionless number that measures the performance of a heat pump. If the heating capacity is expressed in units other than watts, it is called heating energy efficiency ratio (HEER). Heat pumps sold in the United States are often stated in terms of HEER (see Eq. 5).

$$\text{HEER} = \frac{\text{Heating capacity (Btu/hr)}}{\text{Power input (W)}} \tag{5}$$

Seasonal heating energy efficiency ratio (SHEER) is the time average HEER value for the heating season, and it is also known as seasonal performance factor (SPF). In the equation described, HEER has a unit of "Btu/hr per W." It should be noted that 1 Btu/hr = 0.2928 W. Therefore, the relationship between HEER and SCOP can be expressed as in Eq. 6.

$$\text{SCOP} = \text{HEER} \times 0.2928 \tag{6}$$

Primary Energy Ratio

The HCOP provides a measure of the usefulness of the heat pump system in producing heat from work. It does not express the fact that energy available as work is normally more useful than energy available as heat. To assess different heat pump systems using compressor drives from different fuel or energy sources, the PER is applied. The PER takes into account not only the heat pump COP but also the efficiency of conversion of the primary fuel into the work that drives the compressor. PER is defined as in Eq. 7.[5] This can be also expressed as in Eq. 8.

$$\text{PER} = \frac{\text{Useful heat delivered by heat pump}}{\text{Primary energy consumed}} \tag{7}$$

$$\text{PER} = \frac{Q_{\text{H}}}{E_{\text{pe}}} = \frac{Q_{\text{H}}}{W} \cdot \frac{W}{E_{\text{pe}}} = \text{HCOP} \cdot \eta_{\text{pp}} \tag{8}$$

where Q_{H} is the load, E_{pe} is the primary energy used by the heat pump system, and η_{pp} is the power plant efficiency.

The drive energy of heat pumps is most commonly electricity. "Ideally a heat pump where free work is available should be contemplated e.g., wind or water power."[5] Consider an electric heat pump powered by a conventional power plant fueled by a non-renewable energy (see Figure 11). The power plant efficiency, η_{pp}, is up to 58% for oil- or gas-fired combined-cycle power plants currently available on the market. The PER is equal to the HCOP for direct power generation from renewable ambient energy sources such as solar and wind.[2] This concept is illustrated in Figure 12.

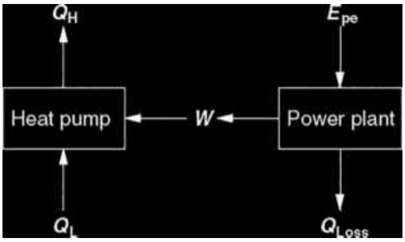


FIGURE 11 Energy flow of a conventional electric heat pump.

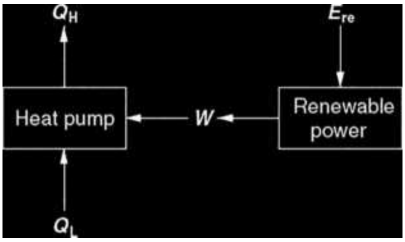


FIGURE 12 Energy flow of a renewable electric heat pump.

The amount of renewable or ambient energy spent (used up) to produce work for the heat pump is equal to the amount of work (i.e., $E_{re} = W$). For this case, by definition, $\eta_{pp} = 1$.[2] It should be noted that unlike the losses in a fossil fuel power plant, the unused ambient energy passing through the renewable power plant is still in the form of ambient energy and it is available to be used.

Ambient Energy Fraction

To make a fair systematic comparison of heat pumps and other systems that provide the same heat output, the term *AEF* was developed by Aye et al.[17] The term *solar fraction* is widely understood, accepted and used in the solar energy field. It is defined as the fractional reduction of purchased energy when a solar energy system is used.[18] It is the fraction of the load contributed by the solar energy, which can be calculated by Eq. 9.

$$f = \frac{L - E}{L} = \frac{L_s}{L} \tag{9}$$

where L is the load, E is the auxiliary energy supplied to the solar energy system, and L_s is the solar energy delivered.

Similar to the solar fraction of a solar system, the term AEF of a heat pump can be defined as the fraction of the load contributed by the ambient energy, which may be calculated as in Eq. 10.

$$\text{AEF} = \frac{Q_H - E_{pe}}{Q_H} = 1 - \frac{E_{pe}}{Q_H} = 1 - \frac{1}{\text{PER}} = 1 - \frac{1}{\text{HCOP} \times \eta_{pp}} \tag{10}$$

Figure 13 illustrates the relationship between the AEF and the PER.

Table 2 shows the AEFs of a heat pump, which has a HCOP of 3.5, for various electric power plants. It can be clearly seen from Table 2 that for the same heat pump, the AEF may vary from 1% to 71% depending on how electricity is generated for driving the heat pump compressor. The AEF for the renewable

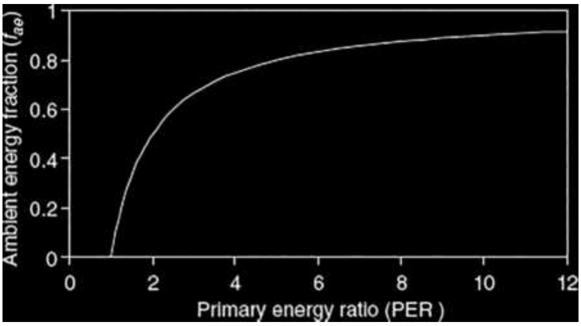


FIGURE 13 The AEF versus the PER.

TABLE 2 AEFs of a Heat Pump Powered by Typical Power Plants

Power Plant	Brown Coal	Nuclear	Black Coal	Gas-Fired Combined Cycle	Renewable
Power plant efficiency, η_{pp}	0.29a	0.33b	0.35a	0.58b	1.00
HCOP of heat pump	3.50	3.50	3.50	3.50	3.50
PER	1.02	1.16	1.23	2.03	3.50
AEF	0.01	0.13	0.18	0.51	0.71

[a] Typical Australian data.
Source: Taylor and Labson.[19]
[b] Typical European data.
Source: Gac et al.[3]

TABLE 3 Effect of HCOP on AEF of a Heat Pump Powered by a Renewable Source

HCOP of heat pump	1.50	2.50	3.50	4.50
Renewable power plant efficiency, η_{pp}	1.00	1.00	1.00	1.00
PER	1.50	2.50	3.50	4.50
AEF	0.33	0.60	0.71	0.78

electricity–driven heat pump is the highest (71%). The renewable energy used is only 29% (i.e., 100% – 71%) of the total thermal load. The energy use of the brown coal electricity-driven heat pump is 99% of the total thermal load (i.e., 0.99 MJ of brown coal energy is required for 1 MJ of thermal load). Table 3 shows the effect of heat pump COP on the AEF of a heat pump powered by a renewable source.

Heat Pump Applications

Heat pumps have been used for domestic, commercial, and industrial applications.

Domestic applications are as follows:

- provision of space heating
- provision of hot water
- swimming pool heating

Commercial and industrial applications are as follows:

- space heating
- water heating
- swimming pool heating
- drying and dehumidification
- evaporation and boiling
- desalination

Potential for the Use of Heat Pumps

In 2000, the primary energy used in buildings was 149 EJ, and the world total primary energy used was 387 EJ.[20] The building sector represented 38% of the total primary energy use and 34% of the total GHG emissions in 2000. Price et al.[20] reported that, in 2004, the emissions from the building sector, including the electricity consumed, were 8.6 Gt CO_2, 0.1 Gt $CO_{2\text{-e}}$ N_2O, 0.4 Gt $CO_{2\text{-e}}$ CH_4, and 1.5 Gt $CO_{2\text{-e}}$ halocarbons (including CFCs and HCFCs). These emissions arise from the following:

TABLE 4 GHG Saving Potential of Heat Pumps (Space Heating, Space Cooling, and Water Heating Applications) for the Building Sector

	2004	2010	2020	2030
GHG emissions from buildings (GtCO2-e)	6.5	7.5	9.6	12.4
Heat pump market share (%)	3	5	14	36
HCOP	3.5	3.7	4.1	4.5
Average power plant efficiency (%)	30	36	48	65
Global average PER	1.05	1.33	1.98	2.93
GHG saving (Mt CO2-e)	9	100	657	2911

- direct combustion of fossil fuels in residential and commercial buildings, amounting to 3.3 $GtCO_2$ (almost 1.7 $GtCO_2$ from combustion of oil, around 1.3 $GtCO_2$ from gas, and about 0.3 $GtCO_2$ from coal)[21]
- indirect or upstream CO_2 emissions from the demand of electricity and district heat were about 5.4 $GtCO_2$[21]
- combustion of biomass produces N_2O and CH_4 equivalent to 0.5 $GtCO_{2\text{-e}}$
- refrigerant or working fluid (halocarbons) leakages accounting for 1.5 $GtCO_{2\text{-e}}$

In 2004, about 60% of total GHG emissions generated by the building sector (6.5 $GtCO_{2\text{-e}}$ out of 10.7 $GtCO_{2\text{-e}}$) is estimated to be due to space heating, space cooling, and water heating. Heat pumps can meet these requirements in all types of buildings. Many low-temperature industrial heating requirements can also be met by heat pump technology. It is apparent that heat pumps have a large potential for saving energy and GHG emissions due to buildings. Energy and GHG saving potential of heat pumps can be estimated by using PER, one of the heat pump performance parameters presented. Table 4 shows the estimated GHG saving potential. The following assumptions were made for these estimations until 2030:

- annual GHG emissions growth rate for the building sector is 2.5%
- annual market growth rate for vapor compression heat pump is 10%
- HCOP improvement is 1% per annum
- annual global average power plant efficiency improvement is 3%

The GHG reduction potential of about 3 $GtCO_{2\text{-e}}$ in 2030 was estimated for the building sector alone. The IEA Heat Pump Centre reported a minimum of 0.2 $GtCO_{2\text{-e}}$ saving potential by industrial heat pumps in 1997, an estimation based on a study by project Annex 21. The total GHG reduction potential of 1.2 $GtCO_{2\text{-e}}$ was estimated in the year 2010 by IEA HPC. “This is one of the largest that a single technology can offer and this technology is already available in the market place.”[22]

Conclusions

The fundamentals of heat pumps have been presented together with the working principles of the thermoelectric heat pump, the absorption heat pump, the gas compression heat pump and the vapor compression heat pump. It should be noted that vapor compression heat pumps driven by electricity dominate the current market. Heat pumps are very energy-efficient and therefore environmentally benign compared to other available heating technologies. The technical and economic performance of a heat pump is closely related to the characteristics of the heat source.

Various performance parameters are available for comparing heat pumps; HCOP and HEER are the most widely used. PER and AEF can be used to compare various heat pump systems systematically and fairly. The AEF of an electric heat pump depends on the HCOP and the power plant efficiency based on the primary energy used. The AEF is highly dependent on the type of power plant used to generate the electricity that drives the heat pump compressor. Heat pumps driven by renewable electricity offer the possibility of reducing energy consumption significantly. In the future, the AEF may be used widely as a performance parameter of heat pumps since energy resource issues are becoming more important.

Acknowledgments

I wish to thank Emeritus Professor William W. S. Charters, former Dean of Engineering at the University of Melbourne, who introduced me to heat pump technology and for supporting my career and research in heat pumps.

References

1. Chua, K.J.; Chou, S.K.; Yang, W.M. Advances in heat pump systems: A review. Appl. Energy **2010**, *87* (12), 3611–3624.
2. Halozan, H.; Rieberer, R. Energy-efficient heating and cooling systems for buildings. IIR Bulletin, **2004**, *LXXXI* (2004–6), 6–22.
3. Gac, A.; Vrinat, G.; Blaise, J.-C.; Camous, J.-P.; Fleury, M. *Guide for the Design and Operation of Average and Large Capacity Electric Heat Pumps*; International Institute of Refrigeration: Paris, 1988.
4. McMullen, J.T.; Morgan, R. *Heat Pumps;* Adam Hilger Ltd: Bristol, U.K., 1981.
5. Reay, D.A.; Macmichael, D.B.A. *Heat Pumps Design and Applications: A Practical Handbook for Plant Managers, Engineers, Architects and Designers*; Pergamon Press: Oxford, 1979.
6. IEAHPC (International Energy Agency Heat Pump Centre). Heat Sources. Available at http://www.heatpumpcentre.org/About_heat_pumps/Heat_sources.asp (accessed September 2005).
7. IEAHPC (International Energy Agency Heat Pump Centre). How Heat Pumps Achieve Energy Savings and CO_2 Emissions Reduction—An Introduction. Available at http://www.heatpumpcentre.org/About_heat_pumps/Energy_and_CO_2.asp (accessed September 2005).
8. Brown, G.V. Magnetic heat pumping near room temperature. J. Appl. Phys. **1976**, *47* (7), 3673–3680.
9. Taylor, L.E. *The Design, Construction, and Evaluation of a Solar Boosted Heat Pump;* Master of Engineering Science thesis, Department of Mechanical Engineering, The University of Melbourne, 1978.
10. Phan, M-H.; Yu, S.-C. Review of the magnetocaloric effect in manganite materials. J. Magn. Magn. Mater. **2007**, *308* (2), 325–340.
11. Steyert, W.A. Stirling-cycle rotating magnetic refrigerators and heat engines for use near room temperature. J. Appl. Phys. **1978**, *49* (3), 1216–1226.
12. Yu, B.; Liu, M.; Egolf, P.W.; Kitanovski, A. A review of magnetic refrigerator and heat pump prototypes built before the year 2010. Int. J. Refrig. **2010**, *33* (6), 1029–1060.
13. Yu, B.F; Gao, Q.; Zhang, B.; Meng, X.Z.; Chen, Z. Review on research of room temperature magnetic refrigeration. Int. J. Refrig. **2003**, *26* (6), 622–636.
14. IEAHPC (International Energy Agency Heat Pump Centre). Heat Pump Technology. Available at http://www.heatpumpcentre.org/Aboutheatpumps/HP_technology.asp (accessed September 2005).
15. Aye, L.; Charters, W.W.S. Electrical and engine driven heat pumps for effective utilisation of renewable energy resources. Appl. Therm. Eng. **2003**, *23* (10), 1295–1300.
16. Aye, L.; Charters, W.W.S; Chaichana, C. Solar boosted heat pump. *Progress in Solar Energy Research;* Nova Science Publishers, Inc.: NY, 2006.

17. Aye, L.; Fuller, R.J.; Charters, W.W.S. Ambient energy fraction of a heat pump. In *Solar 2005, Renewable Energy for a Sustainable Future—A challenge for a post carbon world,* Proceedings of the 43rd Annual Conference of Australian and New Zealand Solar Energy Society, November 28–30, 2005, Dunedin, 2006.
18. Duffie, J.A.; Beckman W.A. *Solar Engineering of Thermal Processes,* 2nd Ed.; A Wiley-Interscience Publication: New York, 1991.
19. Taylor, M.; Labson, S. *Profiting from Cogeneration.* DPIE: Commonwealth Department of Primary Industries and Energy/Australian Cogeneration Association: Canberra, Australia, 1997.
20. Price, L.; de la Rue du Can, S.; Sinton, J.; Worrell, E.; Nan, Z.; Sathaye, J.; Levine, M. *Sectoral Trends in Global Energy Use and Greenhouse Gas Emissions*; Lawrence Berkeley National Laboratory: Berkeley, CA. LBNL-56144, 2006.
21. Stern, N.H. *The Economics of Climate Change: The Stern Review*; Cambridge University Press: Cambridge, 2007.
22. IEAHPC (International Energy Agency Heat Pump Centre). How Heat Pumps Achieve Energy Savings and CO_2 Emissions Reduction: An Introduction, 2010. Available at http://www.heatpumpcentre.org/en/aboutheatpumps/howheatpumpsachieveenergysavings/Sidor/default.aspx (accessed August 2010).

38

Hydroelectricity: Pumped Storage

Jill S. Tietjen

Introduction

The dynamics of electric utility system operation are such that integrating large amounts of intermittent or variable renewable energy resources (such as wind and solar) will require energy storage mechanisms. Wind and solar resources are weather dependent—they generate power when the wind blows or the sun shines—they are often unavailable at times when the electric utility's requirement for power is at or near peak levels. Wind generation is often at its peak output when the requirement for electricity is at or near its lowest levels. Pumped storage hydro is a commercially available technology that has been in use for more than 100 years to provide generation when it is most beneficial. The operational benefits of pumped storage hydro enhance the ability of the power system to accommodate renewable generation while meeting reliability requirements. These benefits will allow renewable generation to become a more significant portion of electric utility generation portfolios around the world.

This entry describes the dynamics of utility system operation. This is followed by information on wind and solar renewable energy resources and the historical experience with how and when the wind blows and the sun shines, enabling these resources to produce electricity. An overview of various energy storage mechanisms is provided, followed by a more in-depth discussion of pumped storage hydro. Information on how electric utilities use pumped storage hydro is followed by a discussion of how pumped storage hydro can be integrated with and provide benefits for renewable energy resources.

Hour	MW
1	54.19282
2	51.58654
3	49.39094
4	49.98949
5	52.04352
6	54.05509
7	60.46665
8	68.17764
9	74.70331
10	81.64417
11	90.04101
12	95.16006
13	98.23288
14	98.29861
15	100
16	98.22293
17	97.93742
18	94.14076
19	91.889
20	87.04675
21	80.462
22	75.42743
23	67.80997
24	59.64637

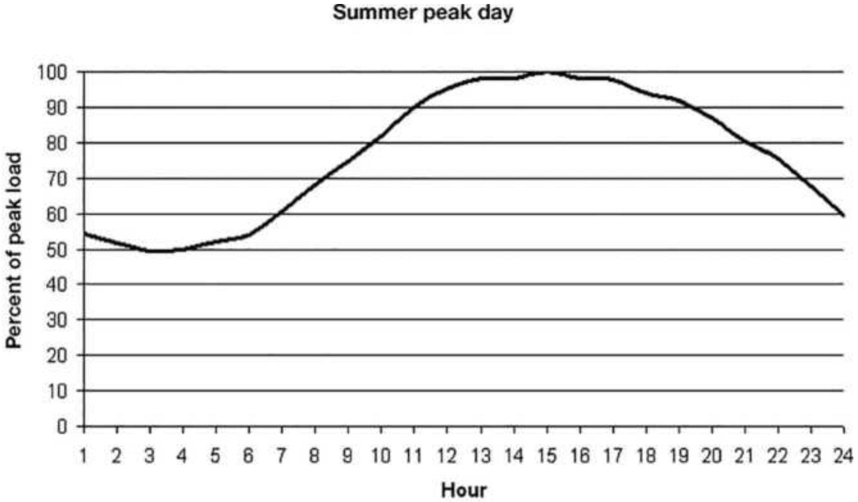


FIGURE 1 Electric customer demand—summer peak day.

Dynamics of Utility System Operation

Demand for electricity is constantly changing—with the hour of the day, the season—and it is dependent on customer behavior. Heating and cooling needs of consumers comprise a significant portion of the demand for electricity. Electric system operators balance the demand for electricity with its supply as electricity has to be used as soon as it has been generated. Generally, there is a greater demand for electricity in the summer (hot temperatures) and winter (cold temperatures) than in the more moderate temperatures associated with the spring and fall. In the aggregate, customers use more electricity during the day and less at night (see Figure 1 where 1 A.M. is hour 1, noon is hour 12, and midnight is Hour 24). Demand patterns change from minute to minute and from one hour to the next. Power system operators must manage the changes that occur with each of the demand fluctuations to balance the system and keep the lights on.[1,2]

Hour-by-hour changes in demand are normally managed by increasing or decreasing the set of resources that need to be operating to meet demand in that hour (unit commitment). Minute-by-minute fluctuations in demand are managed by directing the more flexible resources to increase or decrease their output. Every few seconds, system operators change the generation on specific generating units up or down to keep the system in balance.[1] By balancing fluctuations between supply and demand, system operators provide an ancillary service referred to as regulation. Some generating units, such as combustion turbines and conventional hydroelectric power, are particularly suited for providing regulation as they can operate across a wide range of generating levels. Baseload coal-fired units and nuclear units generally cannot operate with this level of flexibility. Highly variable generation resources such as wind and solar require a significant increase in the regulation capability of on-line generating resources in order to establish the required real-time balance between generation and demand.[3]

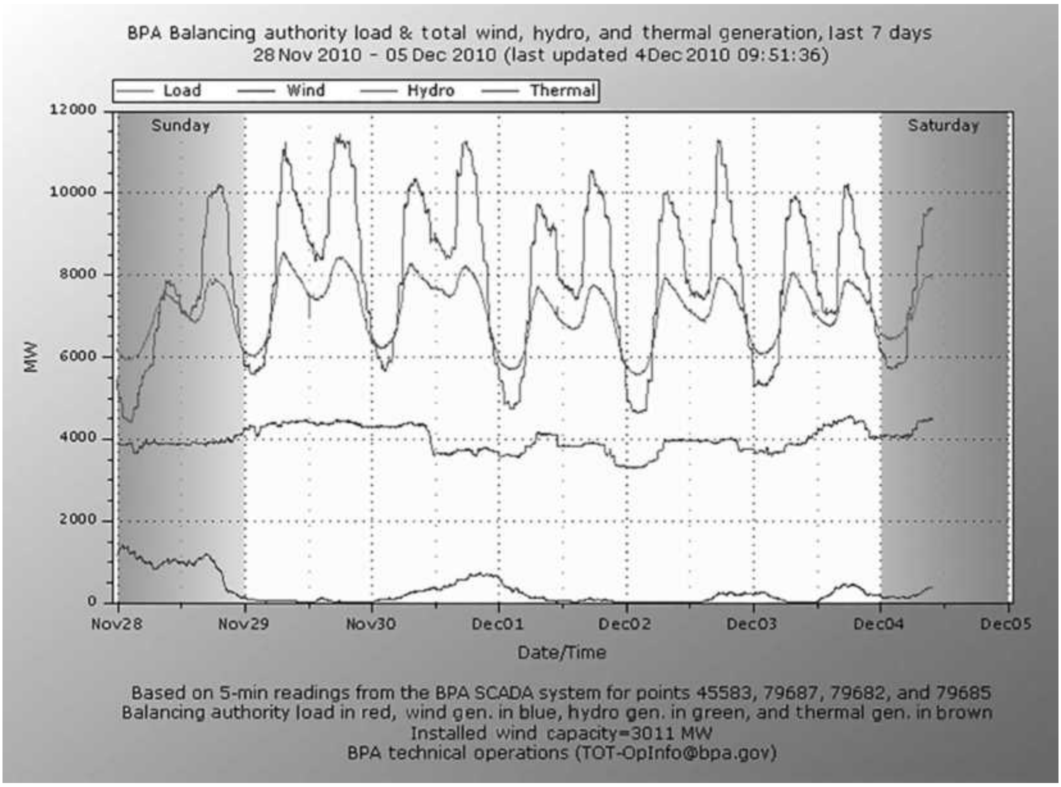
Wind and Solar Resources

Renewable energy resources other than hydroelectric are appearing in more and more utility generation portfolios in significant numbers. This trend is being driven by efforts to reduce carbon emissions and other greenhouse gases worldwide and by specific regulatory or other mandates to increase the percentage of capacity and the energy generated by these renewable resources. Due to advances in technology, tax incentives, and cost reductions, wind resources and solar resources are becoming ever more common.

However, these two types of technology present a major challenge to the operators of utility systems: they only generate electricity when the wind blows or the sun shines but electric load and electric generation must be balanced at each instant from on-line generation. The system cannot absorb or provide extra energy beyond the generating resources operating at any given time. Thus, when the wind dies down or the sun goes behind a cloud, the electric utility must have another generation resource ready and able to provide backup electricity almost instantaneously. Due to the variability of the electric generation from these resources, solar and wind resources are often referred to as intermittent resources. Solar resources generate electricity during the daylight hours. However, the sun is most intense around noon whereas the peak demand for electricity generally occurs in the evening hours in the summer and in a bimodal manner (early morning/evening) in the winter. These patterns mean that the peak availability of solar energy does not occur at the same time as the utility peak; the term used for this by the industry is *non-coincident.*

Wind resources tend to be much more non-coincident with a utility's peak demand than solar resources (see Figure 2).[4] In fact, wind energy resources often peak at night when the electricity demand is at its lowest levels.[5]

Energy storage technologies offer a way to continue the encouragement of the development of wind and solar resources to maximize the value of the energy they provide and to enable the electric utility system to operate reliably once those resources have been added.



BPA Balancing authority load & total wind, hydro, and thermal generation, last 7 days
28 Nov 2010 - 05 Dec 2010 (last updated 4Dec2010 09:51:36)

Based on 5-min readings from the BPA SCADA system for points 45583, 79687, 79682, and 79685
Balancing authority load in red, wind gen. in blue, hydro gen. in green, and thermal gen. in brown
Installed wind capacity=3011 MW
BPA technical operations (TOT-OpInfo@bpa.gov)

FIGURE 2 Bonneville power administration load and generation.
Source: Courtesy of Bonneville Power Administration.

Storage Technologies

Energy storage technologies are expected to play an important role in the integration of intermittent resources into the electricity grid worldwide. Energy storage involves a device that accepts electrical energy from the grid, converts it into an energy form suitable for storage, and then converts it back into electricity, which then, minus efficiency losses, is returned to the grid. Energy storage enhances the performance and economics of the power system, as it allows the capture and storage of renewable energy resources for later use, thus preserving resources. Energy storage can also be a valuable instrument for providing operational flexibility.[2,6]

Technologies that provide the ability to store intermittently generated power can smooth out the variations in the hour-by-hour, minute-by-minute, and second-by-second availability experienced by wind and solar resources. Energy storage can increase the usefulness of wind power by absorbing excess wind generated overnight and supplying that power to the grid during peak daytime hours. It can also help grid operators deal with the second-by-second variability of wind and solar by providing additional regulation service.[5]

Energy storage technologies also can help utilities provide the power quality and reliability required by increasingly complex and sensitive equipment. Energy storage devices improve system responsiveness, reliability, and flexibility, while reducing capital and operating costs for both suppliers and customers. Suppliers can use energy storage for transmission line stabilization, spinning reserve, and voltage control, which means customers would receive improved power quality and reliability.[7]

Many generating resources cannot be shut down during the off-peak period and must be kept in operation at minimum levels. For some utilities, a particular challenge is having more resources that must be kept operational during off-peak hours than the level of load demanded from customers. Energy storage technologies provide a sink for this minimum load energy. In addition, some energy storage technologies, such as pumped storage hydro, can actually provide load when needed to meet system minimum generation requirements.

In total, benefits of energy storage technologies include the following:[8]

- Enhancement to the value of intermittent renewable energy resources on the power grid by firming their energy.
- Improvement in power quality by providing ancillary services such as voltage regulation, spinning resources, and so forth.
- Ability to store low-value, excess energy when power supplies exceed demand until the energy can be economically used to meet load.
- Enhancement of the flexibility of the existing transmission grid.
- Relief of transmission congestion to defer capital expenditures on system upgrades.
- Conversion of less costly off-peak energy into higher- value on-peak power.
- Reduction of problems associated with minimum generation requirements.

The types of energy storage technologies in service and being evaluated worldwide include the following:[9–11]

- Batteries: sodium-sulfur (NaS), vanadium redox flow (VRB), lithium-ion, lead-acid, nickel-cadmium, nickel-metal hydride, zinc-bromine
- Pumped storage hydro
- Compressed Air Energy Storage (CAES)
- Electric double-layer capacitors
- Superconducting Magnetic Energy Storage (SMES)
- Flywheel systems
- Thermal

TABLE 1 Energy Storage Technology Development Status

Commercial	Precommercial	Demonstration Phase	Developmental
Pumped storage hydro		Electrochemical capacitor	Lithium ion (grid applications)
		Flywheel	
Flywheels (local power quality)	Flywheel (grid device)	Hydrogen loop	SMES (grid applications)
CAES		Zinc-bromine battery	
Lead-acid battery (distribution level)		Vanadium redox battery	
Nickel-cadmium battery (distribution level)			
NaS battery (distribution level)			

- Regenerative fuel cells
- Hydrogen

The commercialization status of each of the energy storage technologies is shown on Table 1.[12,13]

The technology that presents the largest scale for utility application (up to 1000 MW), is able to handle large quantities of energy, and is already in the most widespread use is pumped storage hydro.[14,15]

Pumped Storage Hydro

Pumped storage hydro is a form of hydroelectric power generation for electric utilities that incorporates an energy storage feature.[16] The water, the source of potential and kinetic energy, moves between two reservoirs—an upper and a lower one—with a significant vertical separation (see Figure 3).[17] Water is stored in the upper reservoir until such time as the utility determines it is economic to use it to produce electricity for the system. The water in the upper reservoir is stored gravitational energy.[18] When the water is released, the force of that water spins the blades of a turbine that connects to a generator that produces electricity.[19]

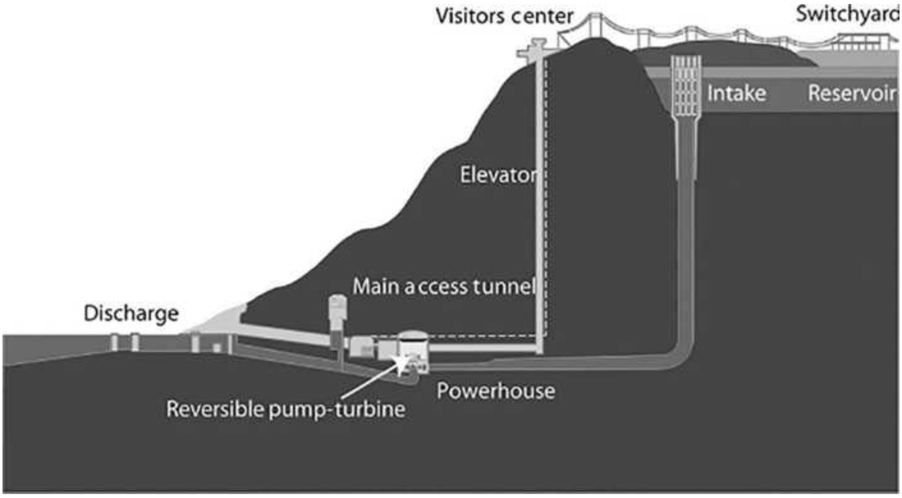


FIGURE 3 Schematic diagram of pumped storage hydro.
Source: Courtesy of the Tennessee Valley Authority.

Water is generally released from the upper reservoir to produce electricity during peak hours or during times of system need. After passing through the turbines, the water is discharged into the lower reservoir. At night and on weekends or during light load or off-peak hours, water is pumped up from the lower reservoir into the upper reservoir by the turbines that have now been reversed to work as electric-motor-driven pumps. Figure 4 shows the upper and lower reservoirs of the Rocky Mountain Hydroelectric Plant in Georgia, United States.

In the upper reservoir, the water is essentially stored energy. Water can be stored for a long time or a short time in the upper reservoir, depending on the needs of the utility (see Figure 5 for a view of the upper reservoir of the Taum Sauk Project in Missouri, United States). A vertical separation of at least 100 m (328 ft) is generally necessary to make a pumped hydro facility economic.[20] This height difference between the upper and lower reservoirs is called the head. The amount of potential energy in the water is directly proportional to the head, so the greater the height, the more energy that can be stored in a reservoir of a given size.[16,21]

FIGURE 4 Rocky mountain hydroelectric plant.
Source: Courtesy of Oglethorpe Power Corporation.

FIGURE 5 Taum Sauk Upper Reservoir.
Source: Courtesy of Ameren Missouri.

Pumped storage hydro is the most widespread energy storage system in use on power networks. Its main applications are for energy management, frequency control, and provision of reserve. Pumped hydro storage may make use of any generating resource on the system. It can be used to absorb temporary generating surpluses, to capture economic energy resources that would otherwise not be utilized, and to allow other generating resources to stay on-line by meeting their minimum generating requirements.[22]

Innovations in variable speed motors have helped these plants operate at partial capacity and greatly reduced equipment vibrations, increasing plant life.[23] Efficiency of any specific pumped storage facility, which is primarily dependent on the height between the upper and lower reservoirs, ranges from 70% to 85%. Many of the proposed pumped storage projects in the United States are off-stream from main stem rivers and either use existing reservoirs or are completely closed-loop projects. As such, the environmental impacts resulting from new pumped storage projects are significantly less than traditional large conventional hydroelectric power projects.[24]

Pumped storage hydro was first used in Italy and Switzerland in the 1890s to enable those utilities to store surplus nighttime output from run-of-river hydro stations for use in meeting their peak power demand requirements the following day. In 1929, the first major pumped storage hydroelectric plant, Rocky River, was built in the United States in New Milford, Connecticut. By 1933, reversible pump turbines with motor generators had become available. The turbines could operate both as turbine generators and, in reverse, as electric-motor-driven pumps.[24–26]

About 3% of total global generation capacity, over 127 GW, is pumped storage capacity. In 2010, the European Union had nearly 40 GW of pumped storage capacity in operation. In 2010, 42 projects provided more than 23 GW of pumped storage capacity in the United States. The largest pumped storage facility in the United States is in Bath County, Virginia, which has a capacity of 2100 MW. Pumped storage plants are characterized by long construction times and high capital expenditures.[27–30]

Technology Description

In conventional hydroelectric generation, hydraulic turbines rotate due to the force of moving water (its kinetic energy) as it flows from a higher to a lower elevation. This water can be flowing naturally in streams or rivers or it can be contained in man-made facilities such as canals, reservoirs, or pipelines. Dams raise the water level of a stream or river to a height sufficient to create an adequate head (height differential) for electricity generation.[31]

If the dam stops the flow of the river, then water pools behind the dam to form a reservoir or artificial lake. As hydroelectric generation is needed by the electric utility, the water is released to flow through the dam and powerhouse. In other cases, the dam is simply built across the river and the water in the river moves through the power plant or powerhouse inside the dam on its way downstream.[32] This latter type of hydroelectric facility is referred to as run-of-river.

In either case, as the water actually moves through the dam, the water pushes against the blades of a turbine causing the blades to turn. The turbine converts the energy in the form of falling water into rotating shaft power to turn a generator to produce electricity. The mechanical efficiency of hydroelectricity is high, about 95%. Availability of the generators is about 90%.[32–34]

The selection of the best turbine for any specific hydroelectric site is primarily dependent on the head (the vertical distance through which the water falls) and the water flow (measured as volume per unit of time) available. Generally, a high-head plant needs less water flow than a low- head plant to produce the same amount of electricity.[33,35]

The power available in a stream of water is

$$P = \eta^{*}\ \rho^{*}\ g^{*}\ h^{*}\dot{V},$$

Where η is turbine efficiency, ρ is power (joules per second or watts), g is acceleration of gravity (9.81 m/sec^2), h is head (meters, this is the difference in height between the inlet and outlet water surfaces), and $\dot{V}$ is flow rate (cubic meters per second).[36]

This equation can be roughly approximated as:

$$\text{POWER}(\text{kW}) = 5.9 \times \text{FLOW} \times \text{HEAD},$$

where FLOW is measured in cubic meters and HEAD is measured in meters.[37] In general terms, 1 gal/sec falling 100 ft can generate 1 kW of electrical energy.[38]

Hydroelectric power is generally found in mountainous areas where there are lakes and reservoirs and along rivers. Hydroelectric power currently provides about 10% of all of the electricity produced in the United States. Hydroelectricity provides about one-fifth of the world's electricity. Worldwide capacity of all hydroelectric capacity as of 2008 was approximately 850,000 MW.[32,38,39]

Producing electricity from hydropower is so economical because once the dam is built and the equipment has been installed, the flowing water has no cost. In addition, the dams are very robust structures and the equipment is relatively mechanically simple. Hydro plants are dependable and long lived, and their maintenance costs are low compared to most other forms of electricity generation, including fossil-fired and nuclear generation.[35] Traditional hydroelectric power once comprised almost 40% of U.S. electricity production. Due to land use issues in developed areas, the potential to build additional traditional hydroelectric facilities is limited and their relative contribution to the generation portfolio in the United States will continue to decrease. There is a significant effort in the United States to explore the utilization of existing dams and infrastructure to add power at dams that currently do not have associated electricity generation. At present, there are approximately 80,000 dams in the United States, of which about 2400 have powerhouses. Pumped storage hydro generation facilities capture many of the system benefits associated with traditional hydroelectric power.

Pumped storage hydro generation is a specific kind of hydroelectric generation requiring an upper and lower reservoir and special equipment that can both generate power as water flows downhill and then reverse and serve as a pump to move the water back uphill. This requirement for location in suitable topological areas limits the geographic flexibility of pumped storage hydro and, at least in Europe, means that most cost-effective sites have already been developed. Some existing conventional hydropower dams are, however, being upgraded with pumped storage capacity. In addition, seawater pumped storage facilities are being examined and a pilot project in Japan has been developed to research the feasibility of using seawater as opposed to freshwater.[40] Modular pumped storage systems, discussed in the next section, may provide additional siting opportunities.

Facility Description

A pumped storage hydropower plant typically has three major components: the upper reservoir, the lower reservoir, and the pumping/generating facilities (see Fig. 3). The water from the upper reservoir used for electricity generation is allowed into the intake shaft through the opening of the headgates. Water moves through the high-pressure shaft and steel-lined power tunnel until it reaches the turbines in the powerhouse.[16] The water turns the turbines that then drive the generators to produce electricity. The water then moves through the tailrace tunnel until it is discharged into the lower reservoir. Water discharge capacity in the upper reservoir can require from several hours to several days.[24]

Pumped storage facilities can be categorized as "pure" or "combined." Pure pumped storage plants, also referred to as closed loop, continually shift water between an upper and a lower reservoir.[29] Combined pumped storage plants also generate their own electricity like conventional hydroelectric plants through natural steam flow. Closed-loop pumped storage systems use closed water systems that are artificially created instead of natural waterways or watersheds and can be as large as traditional large hydroelectric power stations. The water for closed-loop storage usually is only put into the system when it begins operation, either from groundwater or possibly from municipal wastewater.[31,41] Old iron ore quarries and other types of old mines are being examined for use as reservoirs for closed-loop systems worldwide.[42,43]

The efficiency of pumped storage plants generally ranges from 70% to 85%. This means that 70–85% of the electrical energy used to pump the water into the upper reservoir is actually generated when water flows back down through the turbines to the lower reservoir. The losses of energy are primarily related to the mechanical efficiency losses during conversion from flowing water to electricity. Such losses are higher with older designed equipment.[41]

Today, there are three kinds of pumped storage hydro plants available. The most commonly used is a basic version utilizing single or synchronous speed machines that cannot be regulated in pump mode. The pumps can only be run at full capacity or not run at all. Some of these plants use a combined pump and turbine unit instead of two separate units. The compact size of the equipment is beneficial when the plant is built underground.[44]

The second type of pumped storage hydro facility is called "hydraulic short circuit," which allows more flexible operation. Pumps and turbines can be operated simultaneously. The pump can still only operate at full capacity. When large amounts of excess electricity are available, the turbine reduces capacity and more water is stored.[44]

The third type of plant uses variable speed pump technology in order to quickly adapt to fluctuating grid frequency. The operator is able to control how much power is consumed by the pumps. This type of pumped storage plant can experience a round-trip efficiency of 85%.[45]

Environmental Issues

Issues related to permitting of conventional hydroelectric and on-steam pumped storage hydroelectric facilities include factors relating to establishing both the reservoirs themselves and the operation of the facilities. Establishing the reservoirs may involve flooding of land currently in use for other purposes. During operation, the level of the water in both the upper and lower reservoirs fluctuates, which can lead to environmental concerns as well:[46]

- Water resource impacts—stream flows, reservoir surface area, groundwater recharge, water temperature, turbidity, oxygen content
- Biological impacts—displacement of terrestrial habitat, alteration of fish migration patterns, other impacts due to changes in water quality and quantity
- Potential damage to archaeological, cultural, or historic sites
- Visual quality changes
- Loss of scenic or wilderness resources
- Increased risks of landslides and erosion
- Navigation impacts
- Gain in recreational resources

These concerns are much less significant for closed-loop pumped storage plants that are not associated with natural waterways and watersheds. Usually, closed-loop pumped storage plants are specifically not located near existing rivers, lakes, streams, and other sensitive environmental areas to avoid the regulatory lag time and complexity associated with combined pumped storage hydroelectric facilities.[31,46]

In addition to closed-loop systems, underground powerhouses and lines are being used around the world to mitigate the environmental effects of pumped storage projects and reduce visual surface disturbances.[26]

TABLE 2 Countries with Pumped Storage Installed or under Construction

Argentina	India	Slovakia
Australia	Ireland	South Africa
Austria	Italy	South Korea
Belgium	Japan	Spain
Bulgaria	Lithuania	Sweden
Canada	Luxembourg	Switzerland
China	Norway	Taiwan
Croatia	Philippines	Ukraine
Czech Republic	Poland	United Kingdom
France	Portugal	United States
Germany	Russia	
Greece	Serbia	

Installed Facilities Worldwide

Pumped storage facilities are currently installed in many countries around the world (see Table 2) and numerous new facilities are under construction.[27,47]

Electric Utility Usage of Pumped Storage Hydro

Pumped storage facilities use power from available generating resources to pump water into the upper reservoir when it is cost-effective to do so (often at night). This allows the utility to have the water to use during higher load periods the next day to generate electricity at the pumped storage facility. Figs. 6 and 7

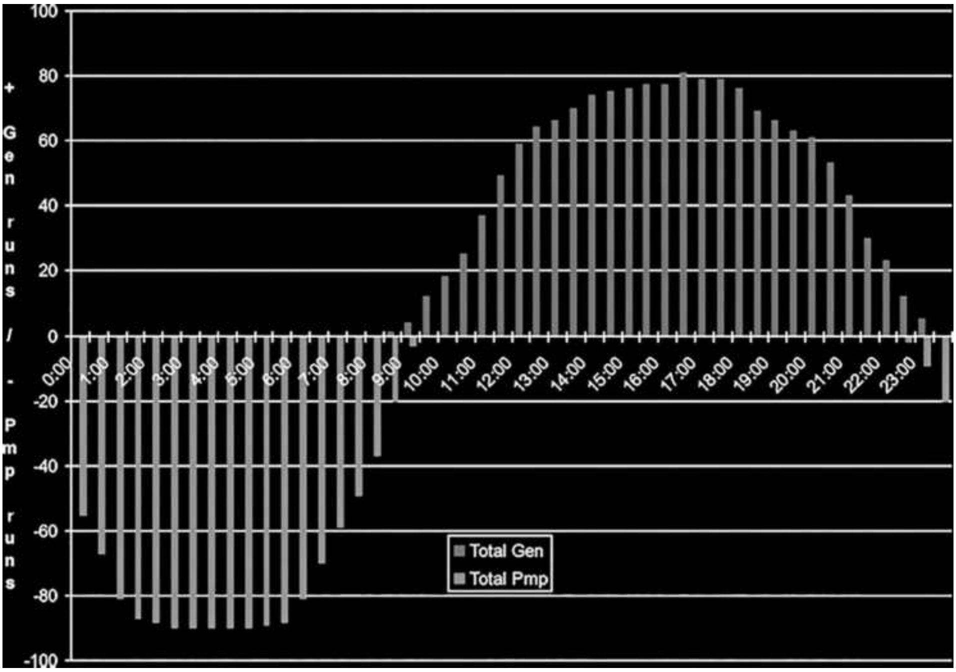


FIGURE 6 Pumped storage unit—typical summer operation.
Source: Courtesy of Oglethorpe Power Corporation.

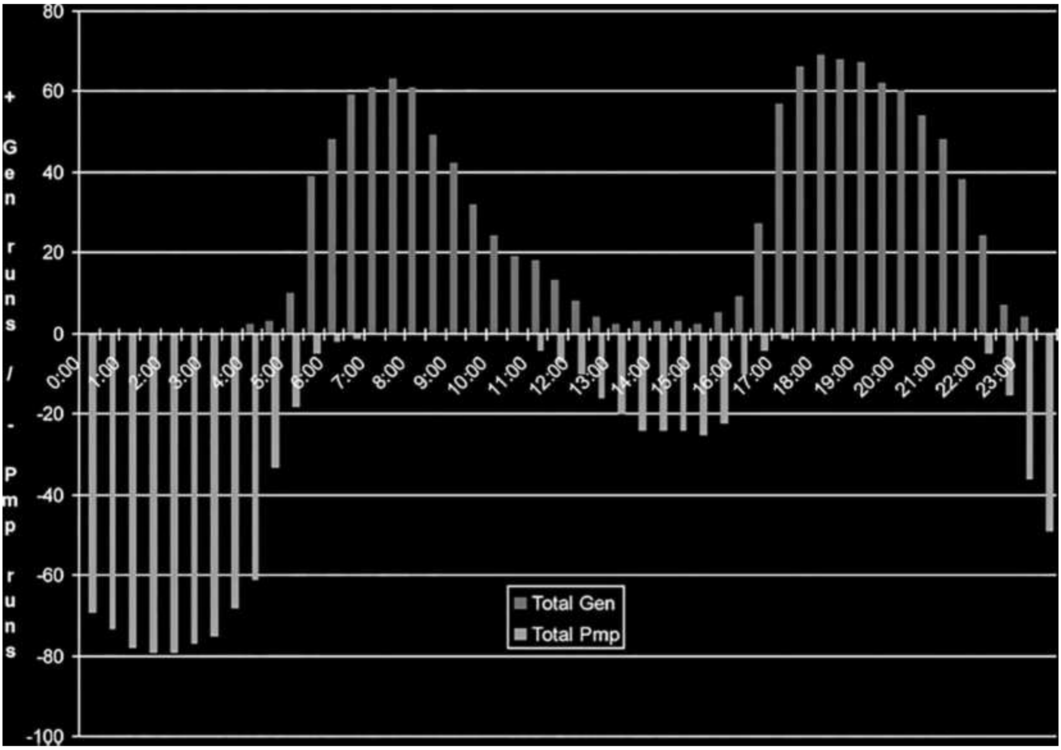


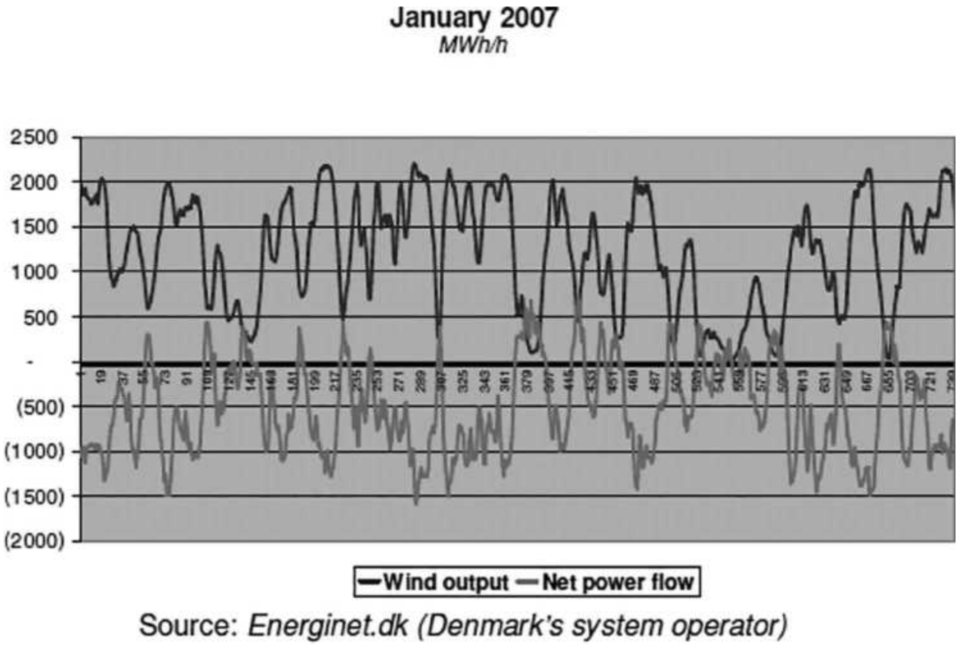
FIGURE 7 Pumped storage unit—typical winter operation.
Source: Courtesy of Oglethorpe Power Corporation.

show the typical operational patterns for pumped storage usage by a utility in each of the summer and winter periods. Power is generated by the pumped storage facility during the higher load hours across all seasons and pumping occurs in the off-peak hours and lower load hours.

Conventional hydroelectric power and pumped storage are actively being used as energy storage for Danish wind generation. Electricity from wind turbines in Denmark is stored in Norway's and Sweden's large, conventional storage reservoirs and then is sent back to Denmark during periods of low wind generation.[48–50] Pumped storage hydro projects in Germany are also critical to maintaining grid reliability and stable power flows within Denmark. Denmark has no native load balancing, so those balancing services are provided via its interconnections with Germany, Norway, and Sweden.[11] The curves in Fig. 8 demonstrate the non-coincident behavior of the wind generation versus the load demand from January 2007.[51]

How Pumped Storage Hydro Can Benefit Renewable Energy Resources

Pumped storage hydro is one of the best available energy storage mechanisms for use on electric utility systems. Pumped storage hydro allows utilities to integrate renewable energy resources such as wind and solar and helps to optimize the operation of the utility grid. Pumped storage hydro helps utilities achieve maximum efficiency for the power-generating system through load shifting. The ability for fast ramping provides a myriad of benefits from serving as spinning reserve to maintaining power system stability to providing voltage control.[52] Pumped storage hydro units can perform black starts and



Source: Energinet.dk (Denmark's system operator)

FIGURE 8 Western Denmark—wind output and net electricity flows.

deliver electricity within a few seconds. Modern pumped storage plants have the highest efficiency of all energy storage systems— 85%.[3,44]

Because of their ability to both pump and generate, pumped storage hydro units provide a load-shifting device for system operators. Pumped storage hydro can firm intermittent resources by absorbing excess generation at times of high output and low demand and releasing that stored energy during peak demand periods. Pumped storage hydro units can help systems avoid minimum generation problems and serve as critical backup facility during periods of excessive demand on the grid system.[3,30,53–56] Pumped storage hydro provides system operators a variety of ways to help balance portfolios from rapid starts and stops to generation in peak load hours.

Pumped storage hydro can move from idle position to full load very quickly, often in seconds, if so designed.[53,57]

This capability allows pumped storage units to provide system frequency control, regulation, and voltage control and to compensate promptly for the loss of other generating units or transmission lines. It allows pumped storage units to count for spinning reserve and standby reserve/ black start units. Typical ramping characteristics of modern pumped storage projects are as follows:[58]

• Shutdown to on-line	60–90 sec
• On-line to full-load generating	5–15 sec
• Spinning to full-load generating	5–15 sec
• Shutdown to normal pumping	6 min
• Spinning to normal pumping	60 sec

Rapid ramping is particularly desirable in the case of a unit becoming unavailable or forced out of service or on utility systems with high amounts of intermittent resources such as solar and wind. For example, the Dinorwig pumped storage facility in North Wales, United Kingdom, can go from 0 MW to full capacity of 1320 MW in 12 sec and can usually maintain this level until other generating units on the utility's system can be brought on-line. This makes pumped storage plants the "race cars" among power-generating facilities.[16,20,41,59] Pumped storage hydro projects can charge or discharge over a 2, 6, or even 12 hr period, as necessary for the utility. The Dinorwig station in North Wales, for example, has an energy storage capacity equivalent to about 5 hr of operation at full capacity. Most of Japan's modern pumped storage projects have energy storage capacity equivalent to 6 to 8 hr of operation at full capacity. The upper reservoirs for many of Japan's pumped storage projects are designed for long-term seasonal storage and drawdown.[26,53]

The dynamic, rapid response capabilities make pumped storage hydro highly valuable for keeping electrical grids stable and reliable with higher percentages of renewable resources, especially as pumped storage hydro can provide both real and reactive power.[59] The ability of pumped storage hydro to produce watts (real) and vars (reactive) power enables system operators to better ride out any system disturbances such as the loss of a major generating unit or transmission line or the significant instantaneous increase or decrease in available energy from renewable resources.

Pumped storage hydro contributes no greenhouse gas emissions and thus assists utilities, and countries, achieve a lower carbon footprint. Utility systems that utilize pumped storage units are also able to operate their thermal units more efficiently—with fewer stops and starts—also contributing to overall reduced emissions.[53,60]

Depending on its location, pumped storage hydro may help optimize transmission line loadings and defer the need for additional transmission assets. Pumped storage units near load centers can operate in condenser mode to generate or absorb reactive power as may be required for system voltage regulation.[53,60]

Conclusion

Energy storage devices in the form of pumped storage hydro facilities offer many benefits to utilities including the integration of the intermittent renewable resources—wind and solar. Due to the weather-dependent nature of these renewable resources and the fact that much wind generation is non-coincident with utility peak demands, pumped storage hydro provides a way for utilities to utilize the renewable resources more efficiently and more cost-effectively. Where topologically feasible, the reliability benefits associated with the rapid ramping capabilities of pumped storage hydro, in conjunction with the many other positive operational capabilities, will drive development of this energy storage technology worldwide for many years to come.

Acknowledgments

The author wishes to acknowledge Russ Schussler, Vice President, System Planning, Georgia Transmission Corporation, and William G. Tharp, Plant Manager, Rocky Mountain Hydroelectric Plan, Oglethorpe Power Corporation, for their assistance.

References

1. *Energy Storage in the New York Electricity Markets*; A New York Independent System Operator White Paper, March 2010; 2.
2. Lohiya, A.K. *Electrical Energy Storage—Large Scale*; August 2009; 1 available at http://www.scribd.com/doc/25835943/Pumped-Storage-Hydroelectricity (accessed December 2010).
3. Montero, F.P.; Pérez, J.J. Wind–hydro integration: Pumped storage to support wind. Hydro Rev. Worldwide **2009**, *17* (3).
4. Available at http://transmission.bpa.gov/Business/Operations/Wind/baltwg.aspx (accessed December 4, 2010).

5. *Energy Storage in the New York Electricity Markets*; A New York Independent System Operator White Paper, March 2010; 4.
6. Cailliau, M.; Foresti, M.; Villar, C.M. Winds of change. IEEE Power Energy Mag. **2010**, 59.
7. Schaber, C.; Mazza, P.; Hammerschlag, R. Utility-scale storage of renewable energy. Electr. J. **2004**, 21–22.
8. Gilbert, K.; Lee, H.; Manwaring, M.; MWH America, Inc. What's so hard about licensing a pumped storage project? Renewable Energy World North Am. **2010**, 76.
9. Inage, S.-I. *Prospects for Large-Scale Energy Storage in Decarbonised Power Grids;* International Energy Agency Working Paper, Paris, France, 2009.
10. Schaber, C.; Mazza, P.; Hammerschlag, R. Utility-scale storage of renewable energy. Electr. J. **2004**, 23–28.
11. Miller, R. Wind Integration Utilizing Pumped Storage. Platts 2nd Annual Power Storage, February 8–9, 2010.
12. Slowe, J. Emerging electricity storage technologies. Cogeneration and On-Site Power Production, **2008**; 59–71.
13. Delta Research Brief. Electric Storage Technologies—Time for Utility Engagement? available at http://www.deKa-ee.com (accessed June 2008).
14. Schaber, C.; Mazza, P.; Hammerschlag, R. Utility-scale storage of renewable energy. Electr. J. July 2004, 25.
15. Bandyk, M. PG&E explores 1,200-MW pumped water storage plant to back up renewables. SNL Energy Electric Utility Report, August 30, 2010; 35.
16. Pumped Storage Reservoirs—Storing Energy to Cope with Big Demands, available at home.clara.net/darvill/altenerg/pumped.htm (accessed March 2006).
17. Courtesy of the Tennessee Valley Authority, available at http://www.tva.gov/power/pumpstorart.htm (accessed December 22, 2010).
18. U.S. Department of Energy—Energy Efficiency and Renewable Energy, Distributed Energy Program. *Pumped Hydro*, available at http://www.eere.energy.gov/de/pumped_hydro.html (accessed March 2006).
19. Hydroelectric Power, available at http://www.tva.gov/power/hydro.htm (accessed March 2006).
20. Available at http://www.uaf.edu/energyin/webpage/webpage-withframes/pumpedhydro.text.htm (accessed January 2006).
21. Hydroelectricity, from Wikipedia, available at http://en.wikipedia.org/wiki/Hydroelectricity (accessed March 2006).
22. Special Report: Accommodating High Levels of Variable Generation; North American Electric Reliability Corporation, April 2009; 50.
23. Schaber, C.; Mazza, P.; Hammerschlag, R. Utility-scale storage of renewable energy. Electr. J. July 2004, 26.
24. Electricity Storage Association. Technologies and Applications, Technologies, Pumped Hydro Storage, available at http://www.electricitystorage.org/tech/technologies_tech-nologies_pumpedhydro.htm (accessed March 2006).
25. U.S. Bureau of Reclamation. *The History of Hydropower Development in the United States,* available at http://www.usbr.gov/power/edu/history.html (accessed April 2006).
26. Lohiya, A.K. *Electrical Energy Storage—Large Scale;* August 2009; 2, available at http://www.scribd.com/doc/25835943/Pumped-Storage-Hydroelectricity (accessed December 2010).
27. Ingram, E.A. Development Activity Snapshots. Hydro Rev. Worldwide **2009** *17,* (6).
28. Robb, D. Could CAES answer wind reliability concerns? Power **2010**, 59.
29. Gilbert, K.; Lee, H.; Manwaring, M.; MWH America, Inc. What's so hard about licensing a pumped storage project? Renewable Energy World North Am. **2010**; 78.
30. Miller, R. Opportunities in pumped storage hydropower. Power Eng. **2010**, 50.

31. California Energy Commission. *Hydroelectric Power in California,* available at http://www.energy.ca.gov/electricity/hydro.html (accessed April 2006).
32. Energy Story: Chapter 12—Hydro Power, available at http://www.energyquest.ca.gov/story/chapter12.html (accessed April 2006).
33. Micro Hydropower Basics: Turbines, available at http://www.microhydropower.net/turbines.html (accessed April 2006).
34. Blankinship, S. Hydroelectricity: The versatile renewable. Power Eng. **2009**, available at http://www.powergenworldwide.com/index/display/articledisplay/364609/articles/power-engineering/volume-113/issue-6/features/hydroelectricity-the-versatile-renewable.html (accessed November 2010).
35. Learn About Energy.org, "Part 11" Hydropower, available at http://www.learnaboutenergy.org/focus/part11.htm (accessed April 2006).
36. Water turbine, from Wikipedia, available at http://en.wikipedia.org/wiki/Water_turbine (accessed April 2006).
37. Hydro-Electric Power, available at http://www.ecology.com/archived-links/hydroelectric-energy/ (accessed April 2006).
38. Union of Concerned Scientists. *How Hydroelectric Energy Works*, available at http://www.ucsusa.org/clean_energy/re-newable_energy_basics/how-hydroelectric-energy-works_html (accessed April 2006).
39. *U.S. Energy Information Administration—Independent Statistics and Analysis*; International Energy Statistics, available at http://tonto.eia.doe.gov/cfapps/ipdbproject/IEDIndex3.cfm (accessed December 2010).
40. Geschler, T. *Survey of Energy Storage Options in Europe*; London Research International, March 2010; 5–6, available at http://www.londonresearchinternational.com/Pu13_energy_storage.pdf (accessed December 2010).
41. Pumped-storage hydroelectricity, from Wikipedia, available at http://en.wikipedia.org/wiki/Pumped-storage_hydroelectricity (accessed March 2006).
42. Fodstad; Audun, L. *Pumped Storage,* 13, available at http://www.risoe.dtu.dk/en/Conferences/Workshop_Sustainable_Energies/~/media/Risoe_dk/Conferences/Energyconf/Documents/Storage/lars_fodstad_tilladelse.ashx (accessed December 2010).
43. Mandel, J. Renewable Energy: DOE Promotes Pumped Hydro as Option for Grid Storage, Greenwire 10/15/2010, available at http://www.eenews.net/public/Greenwire/2010/10/15/10 (accessed December 2010).
44. Geschler, T. *Survey of Energy Storage Options in Europe,* London Research International, March 2010; 3, available at http://www.londonresearchinternational.com/Pu13_energy_storage.pdf (accessed December 2010).
45. Geschler, T. *Survey of Energy Storage Options in Europe,* London Research International, March 2010; 3–4, available at http://www.londonresearchinternational.com/Pu13_energy_storage.pdf (accessed December 2010).
46. Fodstad; Audun, L. *Pumped Storage,* 15, available at http://www.risoe.dtu.dk/en/Conferences/Workshop_Sustainable_Energies/~/media/Risoe_dk/Conferences/Energyconf/Documents/Storage/lars_fodstad_tilladelse.ashx (accessed December 2010).
47. List of pumped-storage hydroelectric power stations, available at http://en.wikipedia.org/wiki/List_of_pumped-storage_hydroelectric_power_stations (accessed December 2010).
48. van Loon, J. Energy: Where wind power is blowing away profits. Bloomberg Businessweek **2010**, 22.
49. Miller, R. Opportunities in Pumped Storage Hydropower. Power Eng. **2010**, 51.
50. Robb, D. Could CAES answer wind reliability concerns? Power **2010**, 58.
51. Wind Energy: The Case of Denmark, CEPOS—Center for Politiske Studier, September 2009; 11.

52. Lohiya, A.K. *Electrical Energy Storage—Large Scale,* August 2009; 4, available at http://www.scribd.com/doc/25835943/Pumped-Storage-Hydroelectricity (accessed December 2010).
53. Makansi, J. Bulk storage could optimize renewable energy. Power **2010**; 68.
54. Miller, R. Johnson, C.Y. Efforts turn to storage for renewable energy. The Boston Globe September 1, 2008, available at http://www.boston.com/news/science/articles/2008/09/01/efforts_turn_to_storage_for_renewable_energy/?page=1 and http://www.boston.com/news/science/articles/2008/09/01/efforts_turn_to_storage_for_renewable_energy/?page=12 (accessed February 2010).
55. Collinson, A. The costs and benefits of electrical energy storage, S684/003/99, as published in Renewable Energy Storage: Its Role in Renewable and Future Electricity Markets, IMechE Seminar Publication 2000–7, Professional Engineering Publishing Limited: London, 26.
56. Fodstad, L.A. *Pumped Storage,* 19, available at http://www.risoe.dtu.dk/en/Conferences/Workshop_Sustainable_Energies/~/media/Risoe_dk/Conferences/Energyconf/Documents/Storage/lars_fodstad_tilladelse.ashx (accessed December 2010).
57. Fodstad, L.A. *Pumped Storage,* 9, available at http://www.risoe.dtu.dk/en/Conferences/Workshop_Sustainable_Energies/~/media/Risoe_dk/Conferences/Energyconf/Documents/Storage/lars_fodstad_tilladelse.ashx (accessed December 2010).
58. Lohiya, A.K. *Electrical Energy Storage—Large Scale,* August 2009; 3–4, available at http://www.scribd.com/doc/25835943/Pumped-Storage-Hydroelectricity (accessed December 2010).
59. Vansant, C. Viewpoint: Pumped storage: An idea whose time has come...again. Hydro Rev. Worldwide **2008**, *16* (6).
60. Lohiya, A.K. *Electrical Energy Storage—Large Scale*; August 2009; 5, available at http://www.scribd.com/doc/25835943/Pumped-Storage-Hydroelectricity (accessed December 2010).

39

Integrated Energy Systems

Leslie A. Solmes
and Sven Erik
Jørgensen

Introduction

On August 6, 2003, the Board of Governors of the Electric Power Research Institute reported, "… there is a growing concern that the electricity sector's aging infrastructure, workforce, and institutions are losing touch with the needs and opportunities of the 21st Century. The investment gap—inadvertently reinforced by the regulatory uncertainty of electricity restructuring—is exacting a significant reliability cost that is seen as just the tip of the iceberg in terms of the electricity infrastructure's growing vulnerability to capacity, reliability, security, and service challenges."[1]

On August 14, 2003, the northeast United States experienced power failure that impacted 50 million consumers in the United States and Canada. The front page of the *Financial Times* stated, "Public to pay for power upgrade. Energy Secretary (Spencer Abraham) warns that much of the $50bn (billion) bill will be passed on to customers." The headline that followed on page 3 was, "Failure of the power came as little surprise to experts. Wake-up call was just latest alarm."[2] Energy supply investments are needed, but the public cost for energy should decline.

What does not get front-page headlines are the billions of dollars needed to upgrade unreliable, old, customer energy supply production and distribution systems like heat, air conditioning, and water. One example is APPA: The Association of Higher Education Facilities Officers. APPA reported in 1995 that the backlog of deferred maintenance for colleges and universities was $26 billion.[3] The major portion of this backlog was needed for utility systems upgrades. The problem has not gone away.

Investments in electricity generation and transmission systems are essential for social and economic well-being. Upgrade and expansion of customer production and distribution systems are equally essential and costly. These needs are not problems but opportunities—*huge investment opportunities.*

Electric utilities are large industrial customers who use massive amounts of energy—more than any other industry. The second law of thermodynamics states that every time one form of energy is converted into another form, some energy is lost. On average, two-thirds of the British thermal unit (BTU) value of the fuels used to produce electricity is thrown away as waste heat. More BTUs are lost in converting and transporting electricity to consumers. Consumers, in turn, convert electricity to heat energy to heat water, heat buildings, and cook. Consumers also burn fuels to produce heating, cooling, and process steam and distribute it to the end use. Every conversion and transportation event causes BTU loss.

The opportunity to save energy in producing and transporting electricity, heat, steam, and chilled water is enormous. However, it requires utilities and customers to plan and invest together. Investing in reliable electricity that is efficient, competitive, and environmentally friendly, and that offers the greatest financial benefits to utilities and their customers is directly related to how utilities and customers select, implement, and manage investments in electric, heat, and cooling capacity and dispatching of all forms of energy.

A comprehensive, integrated energy supply investment assessment at Illinois State University (ISU) offers a case study of the financial opportunities and energy savings for both ISU and the electric utility. Investment implementation roadblocks related to cultural practices and risk aversion are discussed. Through use of an integrated, energy supply system investment methodology adapted from development of independent power plants and a Web-based business plan software tool, incentives to overcome cultural inertia and manage life-cycle project risks are detailed. Examples of dynamic Web-based reports serve as a communication medium among the project team members and allow them to conduct real-time asset management. Finally, profits from electric utility and customer collaboration are highlighted.

Integrated Solution

Imagine if electric utilities could reduce the number of new plants and transmission facilities and instead invest in customer heating, cooling, process steam, or other heat energy needs with on-site electricity generation and load shifting technologies. Imagine that both utilities and consumers could benefit financially by integrating their energy supply investments and reducing the amount of energy consumed.

Culturally, electric utilities and customers have diametrically opposed business goals. Utilities want to sell the most energy at the highest price possible. Customers want to purchase the least amount of energy to reliably meet their operating requirements at the lowest cost. Investments in integrated energy supply systems not only create energy efficiency and emission reduction but also result in economic benefits to both utilities and consumers. Let us use a case study from ISU to demonstrate how to create valuable investment for both utilities and customers.

In 1999, ISU determined that they could fix their facilities operating budget at their 1997 level for 20 years by investing $37.56 million in utility and building systems infrastructure upgrades. Annual budget savings were conservatively estimated to be $3.45 million per year with debt service at $3 million. Net present value ranged from $8.7 to $13 million.

Five megawatts of on-site electricity generation was planned. Using the waste heat and load shifting technologies, electricity load was leveled and reduced by 3.2 MW. The electrical capacity addition to the grid was 8.2 MW. More than half of the waste heat from electrical generation became useful energy. Owing to ISU's level load, the utility was able to gain greater use of its existing power plants by supplying the remaining electricity required by ISU on a 24/7 basis.

History

For nearly 30 years, the development of energy business plans, using an integrated systems methodology with investments sized to meet the business goals of the end users, has demonstrated the ability to:

- Finance millions of dollars of infrastructure upgrades from operating budget savings
- Level electricity load requirements
- Reduce BTU consumption and unit costs
- Improve energy reliability and operating productivity

So why were so few of these opportunities implemented? Customers are not in the utility business. Instead, they are focused on investments that provide products and services. As a result, no energy savings or investment occurs.

More importantly, the problem has been, "How to communicate and manage risks?" How to insure risks associated with energy price escalation, building the project on time and budget, costs of financing and ownership, equipment performance, and changes in business goals can cause major headaches. What business manager in his or her right mind would risk his or her career on a performance-based investment that cannot be measured or managed?

At the same time, as customers were struggling to update their heat/steam and cooling systems, electric utility generation investments almost stopped. The risks associated with building central electricity generation plants on time and budget, the rising cost of financing, the uncertainty of the regulatory world, and competition from independent power producers strangled central power plant development.

On both sides of the meter, risks associated with energy systems investments caused inaction. Investing in efficient energy supply systems requires integration. Integrated investment in a dynamic marketplace demands that all stakeholders have the ability to measure, manage, communicate, and report project status at any point in time.

Stakeholders

An energy supply investment team or stakeholders include the following: chief executive; financial, budget, facilities, and utilities managers; engineering, procurement, construction, capital planning, maintenance, and information technology personnel; contractors; facility occupants; financiers; and investment partners. Every member of the team contributes to the assumptions that go into energy supply systems planning and implementation.

Once energy master plan assumptions are adopted, each team member now knows the complete game plan and his or her performance requirements. As the plan is implemented and measured over time, each team member is responsible for communicating changes that will occur. Instantaneously, the impact of these changes on the total project performance and each team member's area of responsibility must be calculated, compared with the original plan, and reported to all team members. Team members can, in turn, adjust and manage their area of responsibility.

Web-Based, Real-Time Communication and Reports

To conduct the investment communication necessary to manage energy investments, it is essential to conduct energy planning, investment, and asset management using a dynamic, Web-based tool. The tool should not only compare project investment options but also provide timely updates of business plan assumptions based on market changes so that all project team members can communicate and manage change. Businesses and organizations require adoption and commitment to an operating software system that is their master business planning and asset management standard. Web-based workflow connects all the elements of integrated energy systems needed to manage market and business interactivity and create the lowest energy unit cost at the highest reliability.

Using the ISU project, let us look at the kinds of reports that team members will need to see. Presented here are five examples: Executive Summary, Pro Forma, Cogeneration, Steam, and Chilled Water Operating Calculations. Note that although the ISU project includes cogeneration, each consumer is encouraged to assess all on-site generation options that apply to the facility/site situation.

Table 1 and Figure 1 are life-cycle examples of a preliminary engineering, implementation, and financial assessment of an infrastructure investment strategy for ISU. This project analysis was conservatively conducted and included all campus buildings and utility systems. In summary, the

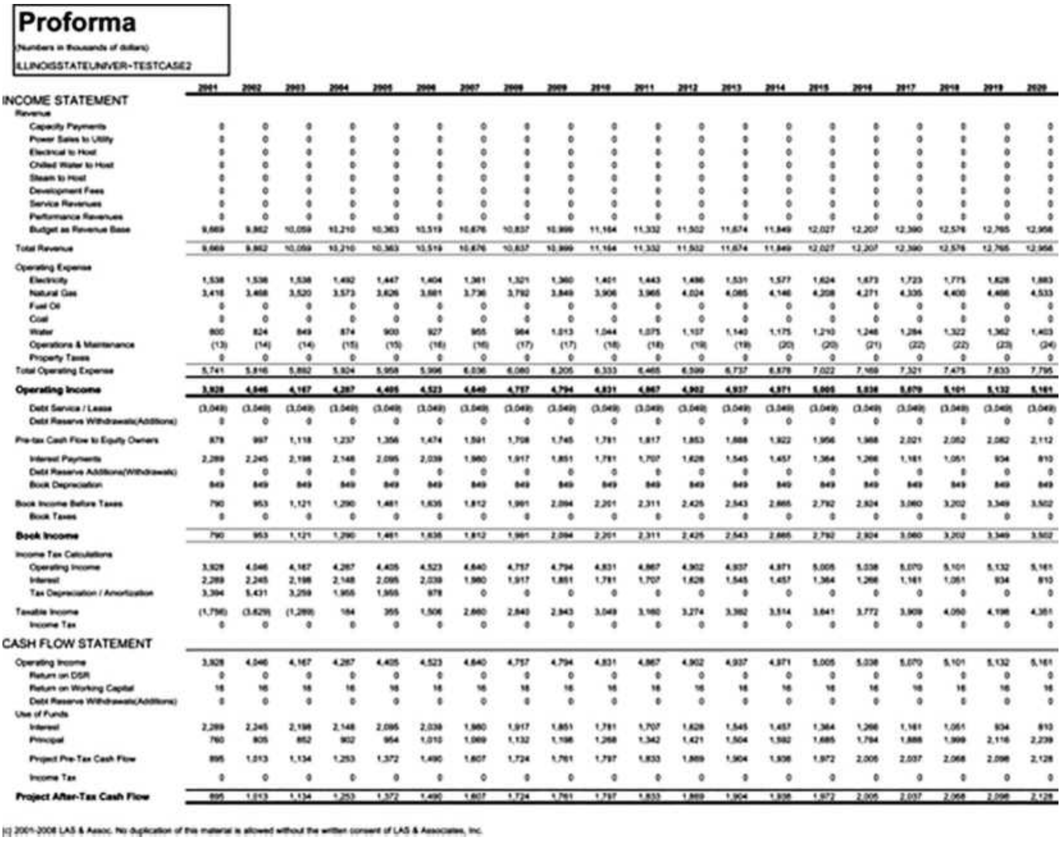
TABLE 1 ISU Executive Summary

Executive Summary			
Illinois State University - Test Case 3			
Current utility budget			$7,833,897
Energy cost savings			
Facility efficiency recommendations			$1,264,034
Electrical generation			$1,880,989
Steam production			$392,913
Chilled water production			$(89,549)
Water			$0
	Total utility budget savings	$3,448,386	
	Percent of current utility	**44.02%**	
Project costs			
Facility efficiency recommendation			Electrical generation $3,595,110
	$2,625,000		
Steam systems			$4,260,000
Chilled water systems			$10,707,100
Infrastructure			$7,100,000
Engineering/contingencies			$5,657,442
Additional project costs			$2,069,200
	Total EPC cost	$36,013,851	
	Annual debt service	$3,049,409	
	Energy budget savings: Debt service	**113.08%**	
Unit energy costs			
Average cost for purchased electricity			$0,046 /kWh
Average cost for cogenerated electricity			$0,043 /kWh
Average cost for steam			$6,994 /KLbs
Average cost for chilled water			$0,312 /ton-hr
Financial			
Capitalization		NPV	
Term debt	$36,013,851		
		Before Tax @ 8%	$12,927,353
Debt service reserve	$1,550,863	10%	$10,547,035

(*Continued*)

TABLE 1 (*Continued*) ISU Executive Summary

Executive Summary						
Illinois State University - Test Case 3						
Total Debt		$37,564,714		After Tax @ 12%		8%
	$8,718,363	$12,927,353				
Equity		$0			10%	$10,547,035
Total capital		$37,564,714			12%	$8,718,363
Proforma assumptions						
			Debt coverage			
Term	Interest Rate			20	5 85% A	Average 1.55
Equity	100%			Minimum	1.29	
Partner equity	Effective Tax Rate		0%	0.00%		
Interest on reserves	0.00%			Maximum	**1.69**	



Proforma

(Numbers in thousands of dollars)

ILLINOISSTATEUNIVER-TESTCASE2

(c) 2001-2008 LAS & Assoc. No duplication of this material is allowed without the written consent of LAS & Associates, Inc.

FIGURE 1 ISU pro forma.

university could fix its 1997 utility budget at $7.8 million for 20 years and finance $37.5 million in facility and utility systems upgrades. The cash flow benefit came from simultaneously installing highly efficiency equipment, efficiently sizing and dispatching equipment, providing for fuel switching, reducing the amount of BTUs needed to provide reliable service, and lowering the BTU unit costs needed to serve the university.

The investment reflects a capital investment, which included a 5 MW gas turbine, new boiler, boiler plant upgrades, a new boiler structure, steam and condensate lines, electric chillers, thermal energy storage system, new chiller structure, chilled water building loop and distribution system upgrades, lighting conversions, air-side HVAC upgrades, new equipment control system, system recommissioning, electrical distribution systems upgrades, a new substation, energy information management system, and $5 million in asbestos abatement. Total capital costs were estimated to be more than $28 million. Engineering, procurement, construction, and contingency costs were $5.6 million with additional costs for permitting, legal fees, start-up, and numerous other project costs estimated at more than $2 million. The financing term is 20 years with an interest rate of $5.85 and a debt service reserve of $1.5 million.

Note the area called *Unit energy costs*. Unit costs including capital, O&M, and energy are averaged by system including costs for purchase of electricity. Change in the technology solutions to any system impacts the BTU unit costs for the other systems. For example, if ISU were to purchase steam from another provider, the unit costs for electricity and chilled water would be impacted. This Executive Summary represents one plan option. To gain the greatest BTU value at the lowest combined BTU unit costs, numerous project options should be compared.

Figure 1 presents the annual cash flow of the master energy plan for the project life. Additional reports shown later detail the operating calculation assumptions associated with each of the project utility systems' financial reports. Cogeneration, steam, and chilled water operating calculation reports are shown in Tables 2–4.

Once a capital investment strategy is identified, decision-makers will want to see an assessment of what might happen if the assumptions change. Risks in an infrastructure investment largely include changes in assumptions related to energy growth and cost escalation, getting the project built on time and within budget, financing costs, and long-term operating and maintenance costs. Table 5 shows a range of cost impacts to the project if the initial assumptions were to change by −20%, −10%, or −5%. Notice that the project shown stays healthy at −5% but loses money at −10%. Its project analysis should analyze worse- and best-case scenarios and be prepared to modify the project or find a way to mitigate the risks.

An executive Summary, supplemented by systems operating assumptions, a life-cycle financial pro forma, and sensitivity analysis are essential to gain the approval from top management to proceed with the project. Although this assessment demonstrated that the opportunity for upgrading the university infrastructure was significant, the project was not implemented.

Planning Methodology

The OA methodology slide presented in Figure 2 is a summary of the myriad assumptions that are needed to create a life-cycle business plan for any facility.

Starting with the first step or lowest level of the OA methodology pyramid, the planning process begins with identifying the facility, its owner, and the business goals associated with the use of the facility over the life of the proposed debt. Infrastructure goals can range from environmental improvements, to business expansion, comfort needs, new operating modifications, financing requirements, and code compliance. For ISU, the university wanted to fix its energy budget and finance all infrastructure capital backlog.

Using a life-cycle approach, the growth assumptions should reflect the energy systems impacts associated with the facility use business goals. For ISU, these included growth of electricity, steam, and chilled water for 20 years on a year-by-year basis. Cost escalation estimates by year were also inputted individually for electricity, natural gas, other fuels, budget growth, O&M, and general inflation. Detail of base year budget and tax status is the last group of assumptions usually required to determine the business situation for the facility. Obviously, ISU is a tax-exempt entity.

TABLE 2 ISU Cogeneration Operating Calculations

Operational Calculations

Cogeneration Systems

Illinois State University - Test Case 3

Gas turbine/engine	Capacity	Fuel rate		Cap factor	Availability	O&M cost	Production (kWh)
Fuel Generator #1	5,000	56		92.00%	88.00%	$0.00	35,460,480
Steam turbines	Capacity	Inlet enthalpy	Outlet enthalpy	Cap factor	Availability	O&M cost	Production (kWh)
Customer annual usage		68,892,296 kWh			Retail rate		0.0000 $/kWh
Wheeling		0 kWh					
		Gas turbine/engine					**Steam Turbines**
Generation capacity		4,400 kW					0 kW
Turbine output		35,460,480 kWh/Yr.					0 kWh/Yr.
Annual fuel usage		397,157 MMBtu					0 MMBtu
Annual fuel cost		$1,247,074					$0
Annual O&M cost		$0					$0
Annual steam output		0 KLbs.					
Fuel chargeable to power		0 Btu/kWh					
							Annual
Cost breakdown:							
Fuel		0.0352 $/kWh					$1,247,074
O&M		0.0000 $/kWh					$0
Debt service & returns		0.0077 $/kWh					$271,297
Total		**0.0428** $/kWh					
Current avg. cost		0.0558 $/kWh					
Projected avg. Cost		0.0460 $/kWh					
Weighted avg. Cost		0.0444 $/kWh					
Customer Savings from Cogeneration:							$1,609,693
Saving Without Debt Service							$1,880,989

Determining the assumptions for ISU's business situation (the first step of the pyramid) begins to highlight the risks associated with a final business investment. Just estimating the costs for purchase of natural gas or electricity for the next 20 years is a major guess. As planning continues up the pyramid, the stability of the entire project can be compromised by changing assumptions at the base. Having the capability to automatically update and integrate the impacts of changes in cost and growth assumptions or any other baseline assumption greatly improves the team's ability to manage and report these impacts.

Step 2 of the pyramid, first offers the ability to determine the base year energy consumption, loads, and costs associated with purchasing and producing steam/heat, cooling, and electricity. Further data input is needed to describe each system's equipment, age, operating characteristics, maintenance requirements, and other data to calculate each system's BTU unit costs.

At this stage, a status quo business situation pro forma should be run to determine the life-cycle costs of what the business entity will spend if no investments are made. The planner should also chart the facility energy load requirements for electricity, steam/heat, and chilled water energy over time when compared with the equipment capability for each system to provide reliable service. For example,

TABLE 3 ISU Steam Operating Calculations

Operational Calculations						
Steam System						
Illinois State University - Test Case 3						
Assumptions						
Cogen steam output				0 KLbs.		
Fuel used	From GT			0 MMBtu		
	Supplemental			0 MMBtu		
Cost				$0		
Projected steam requirements				532,619 KLbs. per Year		
Steam required beyond cogen supply				**532,619** KLbs. per Year		
Projected steam quality		*Enthalpy*	*Temperature*	*Pressure*		
	High Pressure Intermediate Pressure 1,203		1,189 Btu/Lb.	Btu/Lb. 0 400	deg. F	deg. F
	0	125 psig	psig			
	Low Pressure	203 Btu/Lb.	200 deg. F	50 psig		
	Hot Water	Condensate 0	5 Btu/Lb.	Btu/Lb. 0 85	deg. F	deg. F
	0	220 psig	psig			
Steam Production:	**Capacity**	**Fuel**	**Efficiency**	**Pressure**	**Production**	
Standard boiler#10	35,000	Primary	81%	Intermediate	54,622	
Standard boiler #6	85,000	Primary	78%	Intermediate	81,933	
Standard boiler#7	85,000	Primary	78%	Intermediate	191,178	
Standard boiler#8	85,000	Primary	81%	Intermediate	136,556	
Standard boiler #9	65,000	Primary	81%	Intermediate	81,933	
Summary						
Boiler plant:						
Total steam capacity	355,000 Lbs./Hr.			Operation & Maintenance 1.50/KLbs.		
Annual steam output	546,222 KLbs.			MMBTU		
Consumption	815,501					
Cost	$2,560,673					
Cost breakdown:						
Boiler plant				**Annual**		
O&M		Fuel 1.50		4.69 /KLbs. /kLbs.		
$819,333	$2,560,673					
Debt service & returns		0.81 /KLbs.	$440,276			
		6.99 /KLbs.				
Steam from Cogen				**Annual**		
Fuel		0.00 /KLbs.		$0		
O&M	Debt Service & Returns 0.00			0.00 /kLbs. /kLbs.		
$103,351		$0				
		0.00 /KLbs.				
	Weighted average cost per klbs			$6.99		
	Current cost per Klbs			$6.91		
Customer savings from new steam production				**$(150,714)**		
Saving without debt service				$392,913		

TABLE 4 ISU Chilled Water Operating Calculations

Operational Calculations				
Chilled Water System				
Illinois State University - Test Case 3				
Assumptions				
Chiller	**Type**	**Capacity**	**COP**	**Production**
Existing units	Electric	1,500	5.10	1,577,513
New 1000(2)	Steam	2,000	0.90	3,155,025
New 2000(2)	Electric	5,000	6.00	5,784,213
				10,516,751
Projected avg. electrical cost	$0.0444 $/kWh			
Projected fuel cost	$3.14 $/MMBTU			
(Delivered to burner tip)				
Operation & maintenance	$0.01 /ton			
Summary				
Total chilled water capacity			8,500 tons	
Annual production			10,516,751 ton-hrs.	
		Electric	*Gas*	*Steam*
Consumption		42,766,369 kWh	0 MMBTU	42,664 MMBtu
Cost		$1,897,116	$0	$167,254
Cost breakdown:			**Annual**	
Fuel		$0.1963 /ton-hr.	$2,064,370	
O&M		$0.0100/ton-hr.	$105,168	
Debt services & returns		$0.1052 /ton-hr.	$1,106,590	
		$0.3115 /ton-hr.		
	Equal to:	$25.96 /MMBTU		
Current cost per ton of chilled water			$0,177 **/ton-hr.**	
Saving on total new chilled water production				$(1,196,140)
Saving without debt service				$(89,549)

TABLE 5 Sensitivity Analysis

Project cost	–20%	–10%	–5%	estimate
	$51.9 M	$47.6 M	$45.4 M	$43.3 M
Interest rate	–20%	–15%	10%	
	7.02%	6.73%	6.44%	5.85%
Savings	–20%	–15%	10%	
	$2.3 M	$2.5M	$2.6M	$2.9M
Electrical cost	–15%	–10%	–5%	
	4.270	4.090	3.900¢	3.720¢
Natural gas cost	–15%	–10%	–5%	
	$3.45	$3.30	$3.15	$3.00
12% NPV	–$8.5 M	–$2.6 M	$2.3 M	$6.5 M

Figure 3 shows the baseline demand (the diamond line) to exceed the central plant capacity (the triangle line) within a few years.

Step 3 begins data input to come up with integrated energy supply and production systems technology solutions to reliably and efficiently meet business energy needs over time. First, before addressing the supply purchase, production, and distribution technologies, investments are needed to make the end

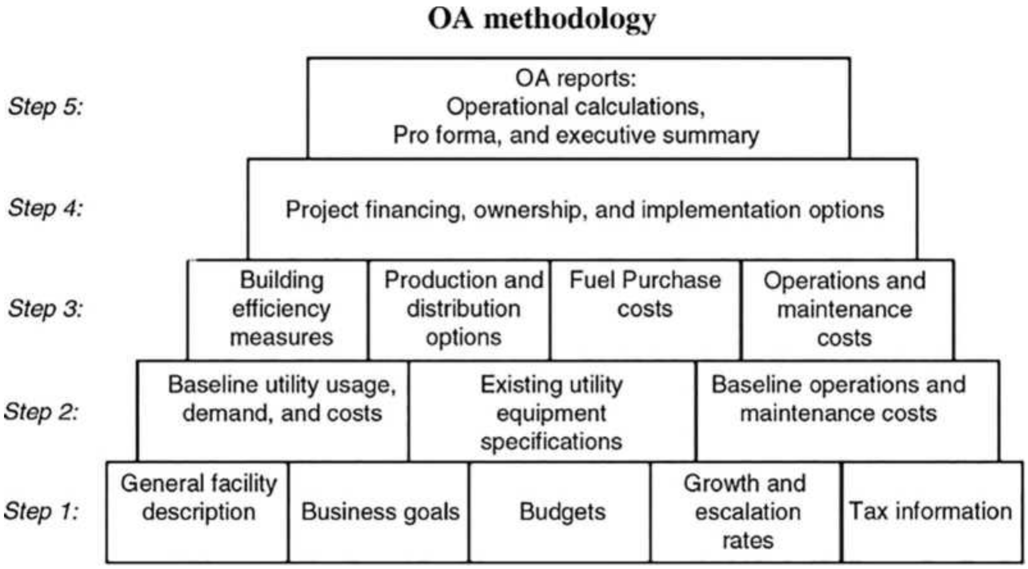
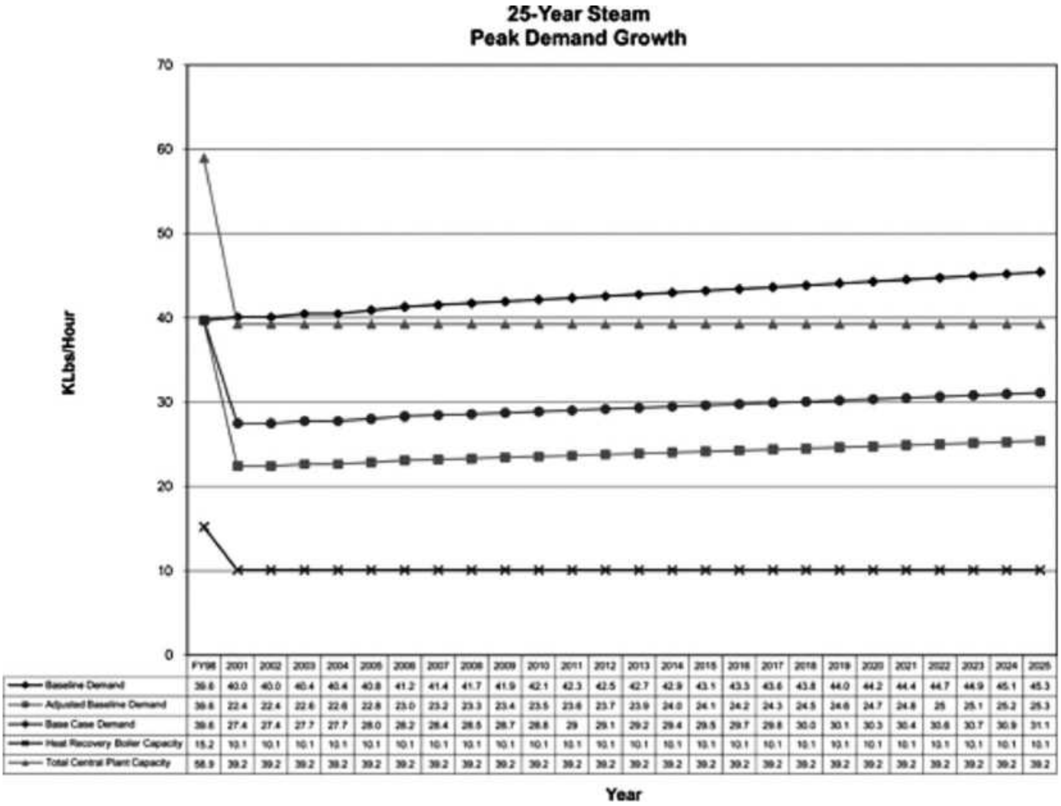


FIGURE 2 OA methodology.



	FY98	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
—◆— Baseline Demand	39.6	40.0	40.0	40.4	40.4	40.8	41.2	41.4	41.7	41.9	42.1	42.3	42.5
—■— Adjusted Baseline Demand	39.6	22.4	22.4	22.6	22.6	22.8	23.0	23.2	23.3	23.4	23.5	23.6	23.7
—◆— Base Case Demand	39.6	27.4	27.4	27.7	27.7	28.0	28.2	28.4	28.5	28.7	28.8	29	29.1
—■— Heat Recovery Boiler Capacity	15.2	10.1	10.1	10.1	10.1	10.1	10.1	10.1	10.1	10.1	10.1	10.1	10.1
—▲— Total Central Plant Capacity	58.9	39.2	39.2	39.2	39.2	39.2	39.2	39.2	39.2	39.2	39.2	39.2	39.2

	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025
—◆— Baseline Demand	42.7	42.9	43.1	43.3	43.6	43.8	44.0	44.2	44.4	44.7	44.9	45.1	45.3
—■— Adjusted Baseline Demand	23.9	24.0	24.1	24.2	24.3	24.5	24.6	24.7	24.8	25	25.1	25.2	25.3
—◆— Base Case Demand	29.2	29.4	29.5	29.7	29.8	30.0	30.1	30.3	30.4	30.6	30.7	30.9	31.1
—■— Heat Recovery Boiler Capacity	10.1	10.1	10.1	10.1	10.1	10.1	10.1	10.1	10.1	10.1	10.1	10.1	10.1
—▲— Total Central Plant Capacity	39.2	39.2	39.2	39.2	39.2	39.2	39.2	39.2	39.2	39.2	39.2	39.2	39.2

FIGURE 3 Twenty-five-year steam peak demand.

use reliable and efficient. Lighting, controls, commissioning, air-side HVAC improvements, and other end-use thermal and electrical improvements need to be identified. For example, if lighting retrofits are implemented, the load demand for heating and cooling will change. A new peak demand line or adjusted baseline (square line) is added to the peak demand growth chart. Now, the planner is ready to select and assess the benefits of new energy supply systems' equipment solutions. The Base case demand (circle line) shows the reduction in steam energy or thousand pounds per hour demand that resulted from the improvement in equipment efficiency. The old baseline assumptions are modified to incorporate the new project technology solution assumptions.

Before running a new pro forma, the planner must input assumptions for how the project will be implemented (Step 4). What are the schedule; financing structure; costs for engineering, procurement, construction, contingency, permitting, start-up, and project management; legal fees; and the myriad other costs associated with implementing the project? Who will own the project? How will it be financed? What are the annual principal and interest payments? Are there tax benefits? All these assumptions are inputted so a full life-cycle cost pro forma can be run to compare the new business plan with the status quo or baseline. ISU Executive Summary and Pro Forma show the comparison of ISU's new business plan with its status quo.

We have now reached the top of the pyramid. Step 5 is the business plan. Look at the assumptions that enter into each business investment and asset management decision. The very bottom of the pyramid requires the planner to begin by guessing the cost escalation for primary fuels, electricity, labor, maintenance, inflation, and budgets over the life of equipment. Debt terms could be 20 years. Now, factor in the business growth planned for that same period of time and try to project the increases or decrease in business growth for electricity, heat, and cooling energy. Planners have not begun to deal with equipment selection and performance or the assumptions associated with implementation to get a project financed, built on time and within budget, and other investment assumptions that impact investment risks.

Weakness in each level of assumptions could make the entire investment crash. The dynamic nature of changing energy prices coupled with business changes is already enough to discourage anyone from pursuing integrated energy investments by using the old-fashioned paper plan approach. Assumptions in a Web-based plan enable constant updates and integrated, real-time impact reporting.

Asset Management

Let us put all those assumptions into a Web model and automate the update of critical assumptions like time-of-day energy pricing, equipment performance, and load changes. In addition, let us approach the investment in a manner that allows the user to turn on and turn off equipment or conduct fuel switching based on unit costs. Further, let us integrate price signals into equipment controls to automate the activity. Let us look at what new capital planning does to the investment. Let us evaluate proposals from potential business partners to determine the value of laying of risk and measuring results.

Software abounds for the following:

- Metering, accounting, billing, tracking, profiling, and verification
- Equipment energy management, measurement, optimization, and automated dispatching
- Integrated energy engineering analysis, calculating technology applications for energy and water consumption, load, and cost savings
- Enterprise energy management for energy systems analysis and optimization
- Conversion of hand-drawn prints to an AutoCAD database
- Work order management

The software listed can be accessed, reported, and analyzed for timely decision-making using Web-based information management software.

What does not abound are software systems that provide for developing, managing, and reporting integrated energy supply investments. Nobody ties supply side energy efficiency investment altogether into a financial pro forma, yet all the energy information software described above can be used to update and manage the investment.

Let us create a communication and reporting platform where the entire business investment team can view real time reports. Reports show the financial pro forma and operating calculation for each and all facilities that need to be reviewed. The reports can be delivered from anywhere, disaggregated at any level of granularity, evaluated compared with any time sequence, rearranged, and put back together.

All team members are required to work together to create each project's assumptions and present the project for management approval. This includes risk assessment and mitigation options. Each team member will know his or her job requirements associated with each project and can be evaluated based on meeting his or her portion of the investment plan performance. In addition, each team member is responsible for communicating market and business changes in the business plan assumptions with the rest of the team.

Now let us connect and integrate consumer energy information systems to electric utility supply systems. What if the utility was able to coordinate dispatching of its generation and transmission options with consumer energy supply systems dispatching? The optimum use of all investments in energy supply systems can result. Not only does efficiency and emissions control occur, but also fewer new electric central plants and transportation investments will be needed. Everyone gains greater productivity through modernization, efficiency, and communication. Less energy is consumed.

Risk Management

For decades, independent power producers have set up limited liability corporations (LLCs) and financed generation project based on the strength of performance contracts for sale of energy, purchase of primary fuels and standby power, financing, engineering/procurement/construction (EPC), and operation/maintenance.

All of these risks can be mitigated contractually. The question is, What is it worth to have business relationships or partners who will take the risks compared to the facility owner assuming the risk? The fun part of having a Web- based business plan is that it is easy to assess the value of laying off risks onto other business entities. Comparison of proposals can be integrated into the Web-based energy plan and compared with the cost to the facility owner. During the course of EPC, the plan can track the actual costs and schedule. Constant monitoring and updating of the Web plan strengthen the ability of both owner and contractor to make decisions and communicate project development.

Risks associated with purchase of fuels or electricity are often most difficult for building owners to control. In addition, regulatory changes and changes in business goals can significantly impact long-term investments. With the economic and political strength of electric utilities, the fact that building and operating energy supply systems are their core competency, and the direct links electric companies have to fuel supply opportunities, both consumers and utilities can benefit from partnering.

Energy service companies, worldwide, are in business to earn profits by taking these risks. However, supply system investments often are not integrated. Methods of asset management and reporting are varied and often subject to dispute.

Profits

Profits to electric companies from partnering with consumers can be expansive. For example, institutions especially often have EPC rules, and situations where the project cost for EPC can be very high—as much as a 100% increase in capital costs. This situation leaves a lot of room for having a business arrangement where another company is paid a profit to build the project and demonstrate equipment performance.

Profits can result:

- By adding value to customer energy systems investments and sharing the savings
- Through implementation of EPC contracts
- From operating and maintaining customer energy supply systems
- From sale of primary fuels
- From sale of heat, chilled water, water, and wastewater
- By investing in customer infrastructure
- From tax benefits
- By gaining emissions credits
- Through access to customer information to dispatch integrated systems
- From advanced growth management information
- From firm power sales contracts
- Through optimization of current supply systems' fixed capital
- From reduced financing costs

In turn, the customer is able to upgrade his or her infrastructure, improve the reliability of energy systems, control facility operating budget, and gain an experienced partner who can manage the short- and long-term investment risks. Emissions are reduced, and productivity is increased, often with no customer debt on his or her balance sheet.

Any and all of these benefits are possible depending on the specific site and partnering options. Another paper could be written describing the structure and value of integrated, energy supply systems business opportunities. The benefits to all suppliers and customers include the ability to:

- Finance millions of dollars of infrastructure upgrades from operating budget savings
- Level electricity load requirements
- Reduce BTU consumption and unit costs
- Improve energy reliability and operating productivity
- Communicate and manage energy supply systems risks

Conclusion: "Big Collaboration, It Is Now and the Subject Is Energy" [4]

Thomas Friedman in his book, *The World Is Flat,* recognizes, "One of the unintended consequences of the flat world is that it puts different societies and cultures in much greater direct contact with one another."[51] Promoting electric utility and customer partnering seeks a change in different cultures. Friedman's vision of a radical new approach to energy use and conservation is not just among countries to jointly develop clean alternative energies.[61] Part of the radical new approach is closer to home—be- tween electric utilities and their customers. Inherent in the solution to energy resource efficiency is a radical new approach to earning profits from value-added. To achieve this, new relationships must be developed. Trust is built from sharing a commitment to save energy and communicate and manage risk.

Traditionally, energy supplier generation, transmission, consumer production and distribution systems, and building end use have been dealt with separately—an approach that has proven to be costly and inefficient. Rather than losing money and energy through the traditional approach, right now, we can build systems as integrated investments, save money, make money, and stop wasting energy.

Inherent to integrated energy investments is commitment to using a methodology and dynamic Web-based tool on which to plan, finance, implement, and sustain joint energy supplier and customer economic and environmental opportunities.

A Note on Sources

Energy gained discussions of energy infrastructure investment practices and solutions rely largely on experience and information drawn from publicly available sources: public-sector studies, Web sites, industry groups, professional associations, and U.S. government sources. The ISU data resulted from a study that my staff and I conducted with the help of the University staff. Extensive reference information can be found at the following Web sites: http://www.acee.org, http://www.aeecenter.org, http://www.appa.org, http://www.chpa.co.uk, http://www.cogeneration.net, http://www.eei.org, http://www.eia.doe.gov, http://www.ieadhc.org, http://www.epri.com, http://www.naruc.org, and http://www.naesco.org. Numerous other Web sites exist.

Addendum

Questions

It has been asked, "Would it be possible to give an estimation of how much energy could be saved from investment and use of integrated energy systems?" Putting a simple number to answer this question is not easy and is complicated by the many ways that experts define and measure energy. The answer varies based on two primary factors: 1) the age, location, pollution costs, and inefficiency of current energy systems and 2) our level of commitment to invest in and operate highly efficient, reliable, clean energy supply systems.

Our goals are to gain the greatest amount of energy output from every unit of energy mined, transported, and converted to electricity, heat, and cooling and to size production and distributions systems to meet reliable, efficient end-use facilities and equipment. (The complexity of this topic cannot be fully explained in this short narrative. Further elaboration and definition of terms can be found in my textbook, *Energy Efficiency: Real Time Energy Infrastructure Investment and Risk Management.*) The goal also includes how to dispatch and operate the energy commodities listed above to achieve efficient, reliable, and nonpolluting water and wastewater systems investments at the lowest production and distribution unit costs.

Although it is important to answer to "how much energy could be saved?," the more important questions are as follows:

- How do we implement and earn profits from integrated resource efficiency investments?
- How do we evaluate alternative fuels; technologies; operating and maintenance options; engineering, procurement, and construction progress; financing and ownership structures; life cycle operation and maintenance maximization; and regulatory and technology changes; and follow market demand changes and numerous other risks to close deals and earn the best returns on investments over time?

Savings Potential

In addition to the coal, natural gas, and petroleum used to produce electricity and heat, the U.S. Department of Energy, Energy Information Administration, reports in its World Consumption of Primary Energy by Energy Type and Selected Country Groups, Table 1, the "net" generation of electricity from numerous renewable fuels and nuclear. The efficiency or net output of useful electricity compared to the total energy value contained in the primary fuel during the 20th century is reported to have a sustained average in the United States of about 33%. The amount of energy consumed is embedded in the conversion factors to change the primary fuel to secondary commodities like electricity and heat. Each primary fuel has energy content. When a primary fuel is converted to electric energy in the United States, on average, two-thirds of the original energy content is thrown away.

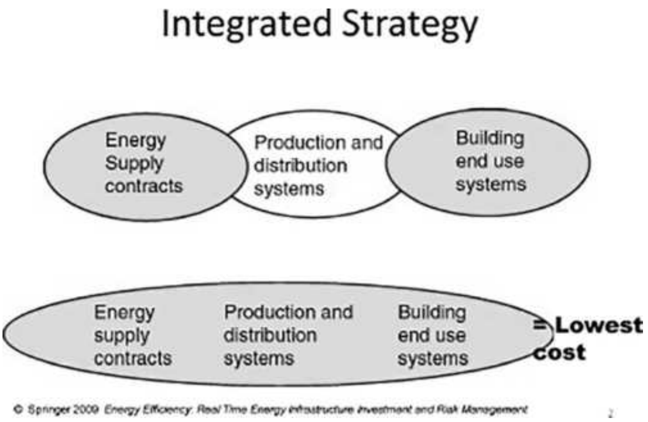


FIGURE A1 Integrated strategy.

Further energy losses also result from decisions about selection of sites to build electric, heat, and cooling production and transportation systems as well as energy use efficiencies, choices of equipment, time of use, fuel switching, storage, and other infrastructure investment efficiency options.

Investment in efficient energy systems requires a comprehensive, integrated approach solving for gaining the greatest output and reliability from every energy unit consumed at the lowest energy unit cost. Figure A1—Integrated Strategy (Solmes, Leslie, *Energy Efficiency: Real Time Energy Infrastructure Investment and Risk Management*, Springer, 2009, p. 20)—first reflects the traditional way of investing in energy systems as separate parts: primary fuels; production and distribution of electricity, heat, and cooling systems; and building end-use systems.

In order to create the greatest production and transmission/distribution efficiency, investment analyses begin with updating end-use facilities and equipment based on defined business goals and projections of end-use growth and cost escalation for primary fuels, labor, emissions, budgets, and other long-term cost considerations that will impact the returns of investment over its life cycle. An assessment of end-use efficiency and system upgrades will result in adjusting baselines for electricity, heat, and cooling requirements. Selection and sizing of new production and distribution equipment will be affected and often reduced.

Add up the energy and load savings that result from comprehensive, integrated energy systems' investments that include the following:

1. Updating of all aged, inefficient, unreliable, and polluting end-use facilities and equipment—*savings of 10%–50%*. A July 2009 McKinsey report on energy efficiency projects that there exists an investment opportunity of $520 billion that would save $1.5 trillion through 2020, reduce U.S. consumption by 23%, and more than offset U.S. expected growth in energy use (http://green.blogs.nytimes.com/2009/07/29/mckinsey-report-cites-12-trillion-in-potential-savings- from-energy-efficiency/). Success stories abound.
2. Sizing production and transmission/distribution systems to meet new, efficient end-use loads—*savings based on load reductions from end-use investments, often 20%–30%*. End-use efficiency will change the size and number of electric, heat, and cooling infrastructure, thus lowering the capital costs for replacement and/or new capacity additions and gaining greater optimization and efficiency of new production equipment and reducing the need for distribution systems expansions. For example, just one building retrofit at a major California university showed a 30% reduction in chilled water requirements due to end-use efficiency savings. The capital cost, even with 15% redundancy, was lowered by 22%.

3. Sizing on-site generation of electricity to meet the thermal load resulting in the greatest energy output from every unit burned—*savings from doubling energy output of primary fuel to produce electricity and dramatic reduction in consumption of primary fuels to produce heat energy could be 40%.* The U.S. Clean Heat and Power Association reports that combined heat and power investment get up to 2–3 times the useful energy products from the fuel and can effectively cut fuel costs by up to two-thirds (http://www.uschpa.org/14a/pages/index.cfm?pageid=3297).
4. Using recycled heat energy to provide heat and cooling also reduces the amount of new equipment needed—*same as above including reduction in new equipment needed to produce heat energy.*
5. Leveling the electrical load by using recycled heat energy for cooling—*savings in electric consumption and need for investments in electric peaking capacity.* Optimal equipment sizing for electricity production usually means that some purchase of electricity from the grid will be needed. However, use of waste heat for heating and air conditioning will result in flattening the electric load and reducing the need for building electric generation capacity to meet peak loads. It will also enable a high percentage of supplemental electricity to be purchased 24/7 thus improving the efficiency of the energy supplier's generation station. (Solmes 2009)
6. Incorporating multiple production equipment solutions to efficiently address shoulder months where heating and cooling vary greatest—*savings from turning off equipment operation during shoulder months and operating equipment at optimum efficiency.* For example, multiple small boilers can be run at full load during low heat demand compared to operating a large boiler for limited hours or partially loaded.
7. Incorporating district heating and cooling systems efficiencies—*it can result in as much as 50% energy savings in industrial, campus, military, hospital, dense city, and similar environments.* Placing boilers and chillers in each building means buying equipment sized to meet the higher heating and cooling loads for each building. Most of the year, this equipment runs partially loaded. A central plant with multiple generation units can run equipment to meet the aggregate load of multiple buildings, thus optimizing equipment performance efficiency and needing to invest in less equipment.
8. Increasing load and efficiency of electricity purchases from utilities, thus improving existing generation efficiency and capital utilization—*improved 24/7 use of electric generation capacity.* See (5).
9. Reducing investment requirements and losses associated with electricity transport systems—*savings from elimination of electric transportation losses and reduction in grid requirements.*
10. Improving load requirements on equipment and systems used to provide primary fuels—*savings from 12-month purchase and/or use of primary fuels.* Greater contracting cost leverage can result for longer-term contracts.

No doubt, more efficiency items can be listed as a part of integrated energy systems investments. Factor in the energy and capital efficiency that can be gained by the combination of all 10 items. Now, include the efficiency that results from financing a single larger project rather than numerous small ones. Add the savings that result from emissions reductions and reduced need to build new electricity generation and transmission systems. This cascading of energy and capital savings by integrating energy systems investments can fund the investment debt as well as earn significant returns on investment.

Thomas Casten in his book *Turning Off the Heat: Why America Must Double Energy Efficiency to Save Money and Reduce Global Warming* provides an answer to "how much could be saved from integrated energy systems investment?"—at least half of current consumption. The opportunity is that due to age and inefficiency, use of the integrated approach will result in funding investments, greater productivity, less pollution, more jobs, and a stronger economy.

New Energy Supply Systems Should Reduce Costs to Consumers

The savings reflected by McKinsey is appropriately titled "end-use" energy savings. Although some waste heat recovery is reflected in the table, the more than half a billion dollars in energy infrastructure investment does not address the investment opportunity that exists in investing in highly efficient, integrated electricity, heat, and cooling production and distribution systems. When Casten wrote about doubling our energy efficiency, he was citing the doubling effect of committing to supply-side resource efficiency—conversion of primary fuels to electricity, heat, and cooling production and distribution combined with efficient end use. This formula will result in meeting the goals of resource efficiency at the lowest cost to the consumer, doubling emissions reductions and tremendous opportunities to earn profits from value-added.

Tracking Investment Returns and Managing Risks

Numerous international measurement and verification guidelines and protocols have been developed and adopted to determine the energy savings from energy efficiency measures and renewable fuels. The oversight of this effort is now done by a non-profit organization called the Energy Valuation Organization or EVO. According to EVO's website, "EVO is the only non-profit organization in the world solely dedicated to creating measurement and verification (M&V) tools to allow efficiency to flourish. Our vision is a global market that properly values the efficiency resource, enabling and assisting the optimal investment in these opportunities."

The limitation to the work accomplished to date is that Web-based energy systems business investment plans, real time tracking, and calculation of electric, heat, and cooling unit costs as a means to achieve the lowest combined energy unit cost are not available. A software application that communicates, tracks, and reports all the assumptions that go into creating a baseline, comparing the benefits of a combination of efficient investment options and investment strategies, and automating the update of these assumptions over the term of the debt does not exist.

To gain the greatest value from integrated energy investments, stakeholders need to communicate and see the financial impact of market and business changes in real time. Even going through the investment stages of baseline development to preliminary assessment to detailed engineering, procurement, and construction entails very big financial risks if continuous automated tracking and data comparison using internet technology are not instituted to address critical time factors for decision making. An energy infrastructure investment and risk management software application that provides the answers to achieving resource efficiency is the means to successfully accomplish the efficiency investments described above and to encourage electric utilities, service companies, consumers, and investors to earn returns from integrated efficiency investments. An initiative to create this application standard is how to answer the question, "How much energy could be saved from investment and use of integrated energy systems?"

References

1. Electric Power Research Institute for the Board of Directors, Ed. Electricity Sector Framework for the Future, Summary Report; EPRI: Palo Alto, CA, 2003; 2—August 6.
2. ed.;—Financial Times August 18, **2003**, 1 and 3.
3. APPA, CRDM: *Capital Renewal/Deferred Maintenance,* http://www.appa.org.
4. Friedman, T.L., Ed. *The World Is Flat: A Brief History of the Twenty-First Century;* 2005; 413.
5. Friedman, T.L., Ed. *The World Is Flat: A Brief History of the Twenty-First Century;* 2005; 391.
6. Friedman, T.L., Ed. *The World Is Flat: A Brief History of the Twenty-First Century;* 2005; 412–413.

Bibliography

1. Diamond, J., Ed. *Collapse: How Societies Choose to Fail or Succeed;* The Penguin Group: New York, 2005.
2. Friedman, T.L., Ed. *The World Is Flat: A Brief History of the Twenty-First Century*; Farrar, Straus and Giroux: New York, 2005.
3. Greenspan, A., Ed. *The Age of Turbulence: Adventures in a New World;* The Penguin Press: New York, 2007.
4. Vaitheeswaran, V.V., Ed. *Power to the People: How the Coming Energy Revolution Will Transform an Industry, Change Our Lives, and Maybe Even Save the Planet*; Farrar, Straus and Giroux: New York, 2003.

40

Bioreactors for Waste Gas Treatment

Sarina J. Ergas

- Low solubility compounds.
- Cometabolism of chlorinated organic compounds.
- Compounds requiring specialized microbial populations or conditions.
- Indoor air applications.

Introduction

Successful biofiltration applications have been limited to control of relatively soluble volatile organic compounds (VOCs) at low loading rates. Biofiltration is also of limited use for compounds that produce acidic or toxic metabolites or are degraded via cometabolism. To overcome the limitations described above, gas-phase biological treatment systems must be developed which (1) incorporate mass transfer of gas-phase pollutants across a media with high specific surface area and low diffusion length, (2) incorporate high biomass concentrations to maintain high biodegradation rates, (3) provide a method for wasting biomass to prevent clogging, and (4) provide a method for addition of pH buffers, nutrients, cometabolites, and/or other amendments to support the microbial population and neutralize acidic metabolites.

Hollow fiber membrane bioreactor (HFMB) systems have been under investigation which meet the above requirements and have been shown to achieve high VOC removal efficiencies in small reactor volumes. A schematic of a typical HFMB system is shown in Figure 1. The hollow fiber membranes serve as a support for the microbial population and provide a large surface area for VOC and oxygen mass transfer. Waste gases containing VOCs are passed through the lumen of the hollow fibers. Soluble compounds in the gas phase are transferred through the membrane pores and partition into

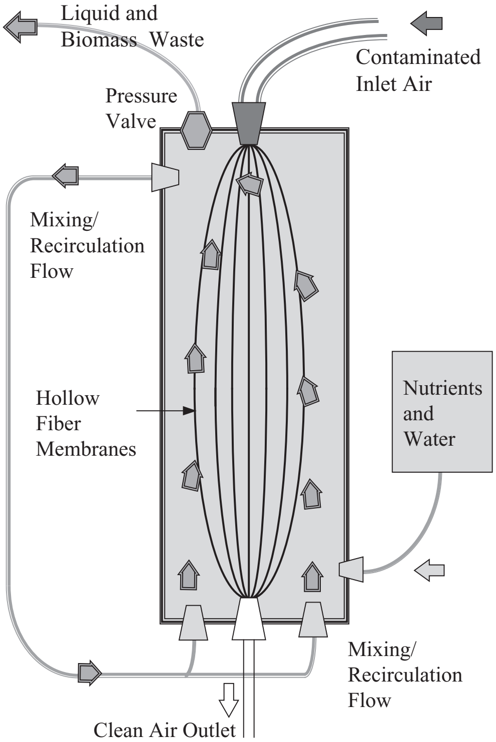


FIGURE 1 Schematic of a hollow fiber membrane bioreactor for biological air pollution control. (From Shumway, 1997.)

a VOC-degrading biofilm surrounded by a circulating nutrient media. Compounds in the biofilm are available for biodegradation. Potential advantages of HFMBs for waste gas treatment include

- Ability to continuously remove biomass to prevent clogging at high VOC loading rates.
- Ability to remove degradation by-products and add pH buffers, nutrients, and/or co-substrates to support the microbial population and neutralize acidic metabolites.
- Ability to treat low solubility compounds such as nitric oxide (NO) and methane.
- Separation of the microbial process from the gas being treated. This may be useful in indoor air applications where carryover of biomass into the ambient air is a concern.
- Optimal humidification of the gas stream.
- Low pressure drop.
- Gas and liquid flow rates can be varied independently without flooding or foaming.
- Modular design, no moving parts.

Disadvantages of HFMBs for waste gas treatment include high capital costs and that the technology has not been demonstrated at full scale.

Membrane Bioreactor Applications for Pollution Control

Hollow fiber membrane bioreactors have been used in a number of pollution control applications including

- Separation and retention of biosolids.
- Bubble-free aeration of bioreactors.

- Providing hydrogen and other low solubility gases to bioreactors.
- Extractive membrane bioreactors for controlled transfer of pollutants from industrial wastewaters.
- Biological air pollution control.

Membrane bioreactors are most often used for separation of biomass. The ultrafiltration of microfiltration membranes is used as a substitute for sedimentation. These systems are used to produce high-quality effluent and/or provide high solid retention times without washout of biomass. These applications are well documented in the literature and are outside the scope of this discussion (Stephenson et al., 2000).

Gas-to-liquid transfer membrane systems were first introduced for use in blood oxygenation and have been studied extensively for this application (Lund et al., 1996). Pollution control research into membrane aeration HFMBs has been focused on the enhancement of oxygen transfer for in high oxygen demanding applications such as the treatment of high-strength industrial wastewaters (Yamagiwa et al., 1998) and nitrification (Brindle and Stephenson, 1996; Brindle et al., 1998). These systems have also been used for bioreactors degrading VOCs, since the membranes allow for reactor aeration without stripping volatile compounds to the atmosphere (Pressman et al., 1999). Due to the low solubility of oxygen, oxygen transfer often limits aerobic degradation of high oxygen demanding wastewaters. Many facilities switch from atmospheric air to pure oxygen to enhance oxygen mass transfer; however, conventional oxygenation systems have high power requirements and are incompatible with attached growth reactors (Stephenson et al., 2000). Hollow fiber membranes provide a very high surface area for transfer of oxygen directly to the biofilm. Hydrophobic hollow fibers have been developed with sealed hydrophilic ends that enable 100% utilization efficiency in gas-to-liquid mass transfer applications (Ahmed and Semmens, 1992a,b).

A novel application of HFMB gas-transfer technology, hydrogenotrophic denitrification of drinking water, has been investigated by several researchers (Gantzer, 1995; Lee and Rittmann, 2000; Ergas and Reuss, 2001). A number of common genera of bacteria can use hydrogen as an electron donor and nitrate as an electron acceptor under anoxic conditions. These organisms have been shown to denitrify nitrate contaminated drinking water to acceptable levels. Disadvantages of hydrogenotrophic denitrification include lower denitrification rates and the difficulty in dissolving sufficient quantities of hydrogen into the water due to its low solubility. Use of HFMBs has been shown to support high biomass densities and improve hydrogen mass transfer rates while preventing the waste of excess hydrogen and avoiding accumulation of explosive gasses in a confined space.

Extractive membrane bioreactors have been used to biologically treat industrial wastewaters in the presence of high concentration of acids, bases, and salts that can inhibit degradation (Livingston, 1994; Livingston et al., 1998). Extractive membrane bioreactors utilize dense silicone membranes that selectively extract organic pollutants such as chloroethanes, chlorobenzenes, and toluene, from polar and ionic compounds. A VOC-degrading biofilm grows on the surface of the membranes creating a driving force for mass transfer. Due to the selectivity of the membranes, the biofilm is isolated from the harsh conditions in the wastewater to ensure high biodegradation rates.

Membrane Fundamentals

A membrane separates two distinct bulk phases of a system while allowing the transport of compounds from one phase to the other. In waste gas treatment applications, gases are most often blown through the lumen (inside) of tubes made from membrane materials. Pollutants from the gas phase diffuse through the membranes to a liquid phase on the shell side (outside) of the membranes. The membranes also serve as a support for the microbial population. Once in the liquid, compounds are biodegraded creating a concentration gradient, which serves as a driving force for mass transfer.

Membranes are available in a wide variety of materials, porosities, and pore sizes. For successful application, membrane materials must strike a balance between reasonable mechanical strength, high permeability, and high selectivity (Stephenson et al., 2000). For microporous membranes, high selectivity requires a membrane material with a narrow range of pore sizes, and high permeability requires a membrane material with a high porosity.

Dense and porous membranes are fundamentally different from each other. Dense membranes rely on physical–chemical interactions between the permeating compounds and the membrane materials. The mass transfer rate through a dense membrane depends on the solubility and the diffusivity of the permeating compound in the dense matrix (Reij et al., 1998). Dense membranes are limited to polymeric materials, such as latex and silicone, and they can be operated at high gas pressures, and are resistant to chemical and mechanical abrasion (Fitch et al., 2000; Stephenson et al., 2000). They have also been shown to be more resistant to biofouling than porous membranes (Coté et al., 1988, 1989), possibly because the hydrophobic nature of silicone resists attachment of microorganisms.

Microporous hydrophobic membranes are most often used in gas-transfer applications because they provide a high gas permeability, while not allowing transport of water across the membrane. The membrane pores remain gas filled and compounds transfer from the gas stream through the membrane pores by gaseous diffusion. At excess liquid side pressures above the critical pressure, ΔP_{cr}, water enters the pores of the membranes, dramatically decreasing mass transfer rates (Ergas and Reuss, 2001). Gas side pressures greater than the bubble point result in bubble formation in the liquid phase (Semmens et al., 1999). Thus, over the excess pressure range of 0 to ΔP_{cr}, the gas/liquid interface is immobilized at the mouth of the membrane pore on the liquid side.

Microporous hydrophobic membrane materials include polytetrafluorethylene (PTFE), polypropylene, Teflon™, Gore-Tex™ (PTFE/nylon), and other composites. Hydrophobic microporous membranes coated with an extremely thin layer of silicone have also been investigated (Sikar, 1992). The thin silicone layer increases mass transfer resistance but also decrease biofouling. Microporous hydrophobic membranes are available with pore diameters between 0.1 and 1.0 μm (Stephenson et al., 2000). The membranes are manufactured as small-diameter (200–400 μm ID) hollow fiber bundles that provide surface area-to-volume ratios as high as 30–100 cm^{-1} (Sirkar, 1992). This is an order of magnitude greater than equivalent sized packed towers.

Gas-to-liquid transfer can also be carried out using microporous hydrophilic membranes such as polysulfone and cellulose membranes. In these applications, gas pressure must be higher than liquid side pressures. The gas/liquid interface is thus immobilized at the pore mouth on the gas side of the membrane.

Control of biomass thickness has been shown to be a key operational consideration in continuously operated HFMBs for aeration of wastewaters (Stephenson et al., 2000) and biological waste gas treatment (Ergas and McGrath, 1997). Decreased HFMB performance has been observed in bioreactors after the development of a thick biofilm due to substrate mass transfer limitations, membrane fiber plugging, decreased biomass activity, and/or metabolite accumulation (Brindle and Stevenson, 1996; Freitas dos Santos et al., 1997). Hollow fiber bundles also tend to clump together when biofilm growth is high, resulting in fiber tangling and reduction of available membrane surface area. Several operational strategies have been used to maintain film thickness at an optimum level including the use of crossflow membrane configurations (Ahmed and Semmens, 1996) and periodic shearing of biomass from the membranes using high liquid velocities combined with scouring with gas bubbles (Pankhania et al., 1994; Dolasa and Ergas, 2000).

Research Overview

A summary of the laboratory-scale investigations of HFMBs for air pollution control is shown in Table 1. In an early study by Hartmans et al. (1992), a HFMB was used to control air emissions of toluene and dichloromethane. Mass transfer coefficients were determined for a number of different membrane materials. Using the experimentally determined mass transfer coefficients, dichloromethane removal was simulated. Results of simulations suggested that a significantly lower reactor volume would be required

TABLE 1 Summary of HFMB for Waste Gas Treatment Research

Reactor Type	Compound	Membrane Materials	Fiber ID (mm)	Pore Size (μm)	Membrane area (m^2)	References
–	Toluene, dichloromethane	Polypropylene	–	0.10	0.004	Hartmans et al., 1992
Flat sheet	Propene	Polypropylene	NA	0.1	0.0040	Reij et al., 1995
Hollow fiber	Propene	Polypropylene	1.8	0.2	0.10	Reij and Hartsmans, 1996
Hollow fiber	Toluene	Polysulfone	1.1	–	0.028	Parvatiyar et al., 1996a
Hollow fiber	Trichloroethene	Polysulfone	1.1	–	0.028	Parvatiyar et al., 1996b
Flat sheet	Propene	Polypropylene	NA	0.1	0.0040	Reij et al., 1997
Hollow fiber	Toluene	Polyethylene	0.28	–	0.23	Ergas and McGrath, 1997
Hollow fiber	Toluene	Polypropylene	0.20	0.05	0.37	Ergas et al., 1999
Hollow fiber	Toluene/TCE	Polypropylene	0.20	0.05	0.37	Dolasa and Ergas, 2000
Hollow fiber	Ammonia	Polyolefin multilayer	0.20	–	0.063	Keskiner and Ergas, 2001
Hollow fiber	Benzene	Latex rubber	9.5	NA	0.012	Fitch et al., 2000
Hollow fiber	Benzene	Silicone rubber	9.5	NA	0.012	Fitch et al., 2000
Hollow fiber	Benzene	Polypropylene	0.20	0.2	0.30	Fitch et al., 2000
Hollow fiber	Butanol	Polysulfone	1.1, 1.9, 2.7	0.05	0.030, 0.022, 0.013	Fitch et al., 2000
Hollow fiber	Trichloroethene	Polypropylene	0.24	0.03	0.70	Pressman et al., 2000

for a HFMB than for a biotrickling filter. Greater than 95% removal of toluene and dichloromethane were observed in experiments with a flat sheet membrane bioreactor.

Parvatiyar et al. (1996a) used a two module-in-series polysulfone HFMB to investigate toluene removal from a contaminated airstream. Toluene removal reached 84% with a 16-second gas residence time and an inlet concentration of 600 ppm. A similar experimental system was used by the authors to study degradation of trichloroethylene (TCE) (Parvatiyar et al., 1996b). The biofilm was initially acclimated to toluene and then gradually weaned from a toluene/TCE mixture to 100% TCE. A 30% TCE removal efficiency was achieved with a 36-second gas residence time.

Fitch et al. (2000) compared mass transfer and biodegradation rates for benzene and butanol contaminated gases in HFMBs that utilized dense (latex and silicone), microporous hydrophobic (polypropylene), and microporous hydrophilic (polysulfone) membranes. The highest overall pressure drops were observed with the polypropylene membranes due to the smaller diameter of these fibers compared with the other fibers. Significant sorption of benzene was observed in initial tests with the polysulfone membranes; therefore, butanol was used in subsequent experiments with these membranes. Removal efficiencies for butanol of up to 99% were obtained at an inlet concentration of 200 ppm. Dense high permeability latex and silicone membranes were found to have high benzene mass flux rates, possibly because the greater solubility of benzene in the polymers than in air creates a greater effective concentration gradient than observed in the air phase. The low total surface area of the dense membrane tubes limited overall removal; however, and the polypropylene membrane unit was the most effective on the basis of removal per total unit volume of reactor.

In my own laboratory, we have conducted a number of studies using toluene as a model VOC. In our first set of experiments (Ergas and McGrath, 1997), a laboratory-scale HFMB was constructed and operated with toluene at varying loading rates. The gasses passed from an inlet manifold to a membrane distributor made from 336 polypropylene hollow fibers (280 μm ID, 63% porosity, and an active fiber length of 1.1 m). A plot of removal efficiency vs. gas flow rate over the experimental period is shown in Figure 2. Toluene removal efficiencies of greater than 97% were achieved with an inlet toluene concentration of 100 ppm and gas flow rates less than 1.0 L/min (1.4 second residence time). When the gas flow rate was increased above 1.2 L/min (1.1-second residence time), a significant decrease in removal

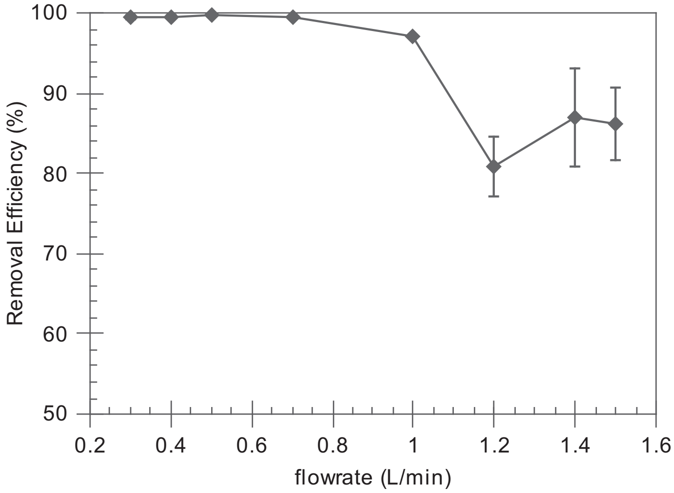


FIGURE 2 Removal efficiency in a HFMB gas flow rate. (From Ergas and McGrath, 1997.) Removal efficiency was determined after an acclimation period at each flow rate. The inlet toluene concentration was maintained at 100 ppm_v, and the liquid recirculation rate was 4 L/min.

efficiency was observed. Removal efficiency was found to decrease over the four-month operational period due to clogging of the bioreactor with microbial biomass, possibly due to the growth of nitrifying bacteria.

In subsequent experiments (Ergas et al., 1999), we investigated the effects of toluene loading rate, gas residence time, and liquid-phase turbulence on toluene removal in a laboratory-scale HFMB. Nitrate was used as a nitrogen source to discourage the growth of nitrifying bacteria. Initial acclimation of the microbial culture to toluene occurred over a period of nine days, after which a 70% removal efficiency was achieved at an inlet toluene concentration of 200 ppm and a gas residence time of 1.8 seconds (elimination capacity of 20 $\mathrm{g/m^3/min}$). At higher toluene loading rates, a maximum elimination capacity of 42 $\mathrm{g/m^3/min}$ was observed. Liquid-phase recirculation rate had no effect on toluene removal in the HFMB.

We have also investigated the use of a HFMB to control ammonia using nitrifying bacteria (Keskiner and Ergas, 2001). The reactor utilized polyolefin multilayer membrane bundles consisting of 200 fibers, with a length of 50 cm, an inner diameter of 200 μm, and a porosity of 42%. Greater than 92% removal efficiency was obtained at an inlet ammonia concentration of 60 ppm and a gas residence time of less than 0.4 seconds. Ammonia mass transfer rates were found to increase at higher recirculation rates. Biomass adhesion and accumulation was not a problem in this system, possibly due to the slow growth rates of nitrifying bacteria. Experiments are currently being conducted with this reactor using nitrifying bacteria to remove nitric oxide from combustion gas streams.

Treatment of Low Solubility Compounds

In biological air pollution control systems, compounds must partition from the gas phase into the moist biofilm before they can be degraded. Pollutant concentrations at the gas/biofilm interface can be described by Henry's law:

$$S_L = C_g / H \tag{1}$$

where H is the Henry's law coefficient, C_g is the gas-phase concentration, and S_L is the liquid phase concentration. Conventional biofilters have therefore been limited to the control of relatively soluble compounds. A number of environmentally relevant compounds such as nitric oxide (NO), hexane, and methane are biodegradable but have high Henry's law constants. For these compounds, mass transfer from the gas phase to the biofilm limits removal unless very large reactor volumes are used. Due to their higher mass transfer rates, HFMBs may make biological treatment of these compounds more economically feasible.

A number of studies have been carried out using propene as a model VOC because its low solubility makes it difficult to remove in conventional biofilters (air/water partition coefficient at 25°C of 8.6). A flat sheet microporous polypropylene HFMB inoculated with *Xanthobacter Py2* was investigated by Reij et al. (1995). After five days with an inlet propene concentration of 2,300 ppm, the biofilm acclimated and 58% propene removal was maintained for the duration of the 30-day test. Because of propene's poor solubility, all mass transfer resistance was found to be in the liquid phase. For more soluble compounds, the authors determined that membrane-phase resistance could approach the same order of magnitude as liquid-phase resistance. A similar reactor system and bacterial culture was investigated by Reij et al. (1997) to degrade propene at concentrations varying from 10 to 1,000 ppm. Once a biofilm was established, propene flux to the membranes was stable, even at low concentrations (9–30 ppm) when mass transfer limitations should be greatest.

Reij and Hartsmans (1996) investigated a HFMB that utilized 40 polypropylene hollow fibers with a length of 500 mm, an inner diameter of 1.8 mm and a pore size of 0.2 μm. Propene was again used as a model compound due to its low solubility. Maximum propene removal rates were 70–110 g/m^3/h. A gas residence time of 80 seconds was required for 95% removal at an inlet propene concentration of 480 ppm. The reactor was changed from ammonia as a nitrogen source to nitrate to discourage the growth of nitrifying bacteria. Increasing the shell side velocity was found to alleviate clogging of the fibers with biomass. A gradual decrease in propene degradation was observed, possibly due to aging of the biofilm.

Cometabolism

A number of chlorinated organic compounds, such as TCE, can only be degraded aerobically under cometabolic conditions. Cometabolism is defined as the transformation of a compound by a microorganism that is unable to use the substrate as a source of energy or as an essential nutrient element. A second substrate (primary substrate) is used to support growth and induce the enzymes necessary to cometabolize the target compound. The primary substrates often inhibit metabolism of the target compound due to enzyme competition (Alvarez-Cohen and McCarty, 1991). In plug-flow reactors, such as biofilters, the induced bacteria are located near the inlet of the bioreactor where the primary substrate is readily degraded. In these areas, competitive inhibition occurs between the two compounds. In the remaining sections, the bacteria do not get sufficiently induced with the low primary substrate concentrations remaining to cometabolize the target compound. Microbial activity can also be inhibited by the toxicity of the target compound or its metabolites. The result is low removal efficiencies for TCE in conventional biofilters (Speitel and McLay, 1993).

Several researchers have studied HFMBs for cometabolism in TCE in wastewaters. Aziz et al. (1995) investigated a HFMB with an external semi-batch reactor for TCE cometabolism. Wastewater contaminated with TCE flowed inside the hollow fibers, with a liquid media containing an active methanotrophic culture circulating around the fibers. At residence times of 5–9 minutes in the fiber lumen, TCE conversions of 80–95% were observed. Pressman et al. (1999) presented a follow-up study with a similar system and methanol as the primary substrate. Over 93% of the transferred TCE was biodegraded.

Pressman et al. (2000) used a HFMB for cometabolism of TCE-contaminated gases. A pure culture of *Methylosinus trichosporium* OB3b PP358 was grown in a continuous-flow chemostat and circulated through the fiber lumen of a HFMB, while TCE-contaminated air was circulated on the shell side of the reactor. *M. trichosporium* OB3b PP358 is a methanotrophic bacterium that has the ability to rapidly cometabolize chlorinated solvents when grown on either methane or methanol. Between 54% and 84% TCE transfer was observed, and 92–96% of the transferred TCE was cometabolized at gas residence times of 1.6–5.0 minutes. Biomass clogging did not occur in this system, possibly because the biomass was pumped through the lumen rather than the shell side of the fibers.

In my own laboratory, we have investigated TCE cometabolism in HFMBs using toluene as a primary substrate (Dolasa and Ergas, 2000). A TCE cometabolizing culture enriched from a wastewater seed was inoculated into the HFMB. Initially, toluene was supplied to the reactor to build a sufficient biomass density on the fibers. After steady-state toluene removal was achieved, TCE was added to gas phase of the reactor. Toluene was added in three different configurations: (1) as a mixture with TCE in the gas phase, (2) by pulsing into the gas phase, or (3) to the liquid phase. Addition of a toluene/TCE mixture through the fibers resulted in an initial decrease in toluene removal followed by complete recovery within 5 days and a maximum TCE removal efficiency of approximately 30%. Pulsing of toluene and TCE through the membranes did not result in significant TCE removal. Adding TCE in the gas phase and toluene in the liquid phase resulted in a maximum TCE removal efficiency of 23%; however, results were highly variable and appeared to be related to liquid-phase toluene utilization rates and biofilm thickness.

Theoretical Models

A number of researchers have presented mathematical models of mass transfer and biodegradation of substrates by biofilms growing on the surfaces of gas-transfer membranes (Livingston, 1993; Essila et al., 2000). The general biofilm model reported here was developed in collaboration with Dr. Mark W. Fitch of the University of Missouri, Rolla, and tested using experimental data from HFMB studies conducted using toluene as a model VOC (Ergas et al., 1999). The model uses an inert surface to establish a boundary condition, with substrate entering the biofilm from the gas/liquid interface. The model was derived for a single lumen and related to the total removal by the number of fibers. A conceptual model of this system is shown in Figure 3. Model assumptions include steady-state operation; Monod biodegradation kinetics; and constant biomass density, ρ_b. Because concentration varies both axially and radially, no analytical solution exists for a single lumen. Therefore, the lumen is divided along the axis into n sections, each with an axial length, Δz. The influent gas concentration to the nth section is the concentration exiting the previous section, C_{n-1}. The concentration exiting the nth section, C_n, is equal to the influent less the removal in the section due to the mass transfer and biodegradation.

Membrane Mass Transfer

The flux, J_n, of substrate through the membrane can be expressed as

$$J_n = k_m A_m \left(C_{n,m} - C_{n-1}\right) \tag{2}$$

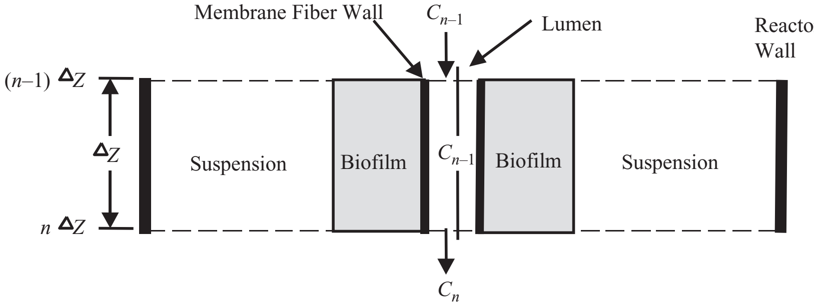


FIGURE 3 Conceptual model of the phases of the hollow fiber membrane reactor. (From Ergas et al., 1999.)

where k_m is the membrane mass transfer coefficient, A_m is the area of the membrane in the section, and $C_{n,m}$ is the gas-phase concentration on the outer face of the membrane. A number of authors have assumed that gas and membrane resistances are negligible compared to liquid-phase resistance (Ergas and McGrath, 1997; Yang and Cussler, 1986; Coté et al., 1989) and therefore $C_{n,m}$ is approximately equal to the concentration in the gas stream, C_{n-1}. Since the membrane is surrounded by biofilm, the liquid-phase concentration at the biofilm interface, $S_{n,0}$, can be related to C_{n-1} using Henry's law:

$$S_{n,0} = C_{n-1}/H \tag{3}$$

Equation 3 sets the inner surface (left-hand) boundary condition for the biofilm model.

Suspension Mass Transfer and Degradation

The suspension (liquid volume) was treated as a continuous-flow stirred tank reactor (CFSTR). The mass flux of the substrate, J_b, from the biofilm to the liquid can be described by

$$J_b = k_L A_B \left(S_{n,i} - S_L \right) \tag{4}$$

where k_L is the liquid mass transfer coefficient, A_b is the outer surface area of the biofilm, $S_{n,i}$ is the VOC concentration at the outer surface of the biofilm, and S_L is the bulk liquid VOC concentration. Assuming Monod biodegradation kinetics, a mass balance on the liquid volume yields:

$$\frac{\rho_L}{Y} \frac{\mu_{\max} S_L}{K_S + S_L} V = QS_o - QS_L + J_b \tag{5}$$

where V is the liquid volume, ρ_L is the biomass density in the liquid, $\mu_{\max}$ is the maximum specific growth rate, Y is the yield coefficient, K_S is the half saturation coefficient, Q is the liquid flow rate, and S_O is the influent VOC concentration to the suspension. Substituting Eq. (4) into Eq. (5) and solving for $S_{n,i}$ yields

$$S_{n,i} = \left(1 + \frac{V \rho_L \mu_{\max}}{Y k_L A_b (K_s + S_L)} \right) S - \frac{Q}{k_L A_b} (S_O - S_L) \tag{6}$$

This relationship sets the exterior surface (right-hand) boundary condition for the biofilm model.

Biofilm Mass Transfer and Degradation

Assuming no advection in the biofilm, steady-state ($\partial S/\partial t = 0$), no concentration gradient in the z direction, and Monod substrate utilization kinetics, the continuity equation for the biofilm in cylindrical coordinates (r, z, and θ) is

$$D_S \frac{1}{r} \left(\frac{1}{r} \frac{\partial^2 S}{\partial r^2} + \frac{\partial S}{\partial r} \right) = -\frac{\rho_b}{Y} \frac{\mu_{\max} S}{K_S + S} \tag{7}$$

where S is the substrate concentration in the biofilm and D_S is the VOC diffusion coefficient in the biofilm. There is no analytical solution for Eq. (7); therefore, a numerical solution was generated.

To fit the models, the biofilm density was varied until the predicted removal matched the observed removal (Ergas et al., 1999). The biofilm biomass density was calibrated at 29,000 mg/L, slightly higher than that reported by Characklis and Marshall (1990). The numerical model predicted the observed trend that shell side liquid flow rate had little effect on observed removal once a biofilm was established. The model slightly underpredicted the effect that substrate loading rate had on removal in the system.

The only case of poor prediction by the model was the effect of gas residence time on removal. Observed removals were higher than predicted removals for these runs. If the pore space of the membranes became water-filled during the course of experimentation, as discussed above, an added resistance to mass transfer would be expected, and mass transfer to the biofilm could become dependent upon the gas flow rate (gas-to-liquid transfer). Sensitivity analysis indicated that removal was a strong function of the biofilm phase biomass density and also of the biofilm diffusion coefficient, with diffusion rates below $(10)^{-9}\,m^2/s$ resulting in decreased removal rates.

Conclusions

Hollow fiber membrane bioreactors are a promising technology for the treatment of biodegradable gas phase pollutants. It has been shown to be effective for aerobic degradation of a range of compounds including ammonia, benzene, butanol, dichloromethane, propene, TCE, and toluene. Advantages of HFMBs include high mass transfer rates, low pressure drops, and small reactor volume requirements. In addition, the ability to separate the microbial population from the gases being treated allows for independent optimization of each phase of the system. In the opinion of this author, the most promising areas of air pollution research for HFMBs include

- Low solubility compounds.
- Cometabolism of chlorinated organic compounds.
- Compounds requiring specialized microbial populations or conditions.
- Indoor air applications.

The greatest disadvantages for the technology are high capital costs and that it has yet to be demonstrated at full scale. Due to their modular design; however, HFMBs should be relatively easy to scale up. EnviroGen, Inc. and the Medical University of South Carolina have developed a prototype with funding from the Department of Energy that they hope to field test in 2001 (Togna, 2000). Long-term studies have been conducted with membrane aeration bioreactors (Stephenson et al., 2000), which are similar in concept and operation. Membrane reactors have also been in use for several decades for blood oxygenation (Sikar, 1992) and separation and retention of biosolids (Brindle and Stephenson, 1996). In common with conventional biofilters, HFMBs have problems with excess biofilm growth and long-term stability of VOC-degrading biofilms. It is also unknown what effect long periods of association of membrane materials and biomass will have on mass transfer rates and mechanical strength in these systems.

Acknowledgments

I would like to thank my students Ayesha Dolasa, Carolyn Gendron, Yenner Keskiner, Michael McGrath, Fereshteh Mehmandoust, Andreas Reuss, and Leslee Shumway for their work on these projects. This material is based on work supported by the National Science Foundation (NSF). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF.

References

Ahmed, T., Semmens, M.J. 1992a. Use of sealed end hollow fibers for bubbleless membrane aeration: experimental studies. *J. Membr. Sci.*, 69:1–10.

Ahmed, T., Semmens, M.J. 1992b. The use of independently sealed microporous hollow fiber membranes for oxygenation of water: model development. *J. Membr. Sci.*, 69:11–20.

Ahmed, T., Semmens, M.J. 1996. The use of transverse hollow fibers for bubbleless membrane aeration. *Water Res.*, 30(2):440–446.

Alvarez-Cohen, L., McCarty, P. 1991. Effects of toxicity, aeration, and reductant supply on trichloroethylene transformation by a mixed methanotrophic culture. *Appl. Envir. Microbiol.*, 57:228–235.

Aziz, C.E., Fitch, M.W., Linquist, L.K., Pressman, J.G., Georgiou, G., Speitel, G.E. 1995. Methanotrophic biodegradation of trichloroethylene in a hollow fiber membrane bioreactor. *Envir. Sci. Technol.*, 47:2574–2583.

Brindle, K., Stephenson, P. 1996. Mini review: the application of membrane biological reactors for the treatment of wastewaters. *Biotechnol. Bioengrg.*, 49(6):601–610.

Brindle, K., Stevenson, P. 1996. Nitrification in a bubbleless oxygen mass transfer membrane bioreactor. *Wat. Sci. Tech.*, 34(9):261–267.

Brindle, K., Stevenson, P., Semmens, M.J. 1998. Nitrification and oxygen utilization in a membrane aeration bioreactor. *J. Membr. Sci.*, 144:197–209.

Characklis, W.G., Marshall, K.C. 1990. *Biofilms.* Wiley, New York.

Coté, P., Bersillion, J.L., Huyard, A. 1989. Bubble-free aeration using membranes: mass transfer analysis. *J. Membr. Sci.*, 47:91–106.

Coté, P., Bersillon, J.L., Huyard, A., Faup, G. 1988. Bubble-free aeration using membranes: process analysis. *J. Water Pollution Control Fed.*, 60:1986–1992.

Dolasa, A.R., Ergas, S.J. 2000. Membrane bioreactor for cometabolism of trichloroethene air emissions. *J. Environ. Engr. ASCE.*, 126(10):969–973.

Ergas, S.J., McGrath, M.S. 1997. Membrane bioreactor for control of volatile organic compound emissions. *J. Envir. Engrg., ASCE*, 123:593–598.

Ergas, S.J., Reuss, A. 2001. Hydrogenotrophic denitrification of drinking water using a hollow fiber membrane bioreactor. *J. Water Supply: Res. Technol.-Aqua*, 50(3):161.

Ergas, S.J., Shumway, L., Fitch, M.W., Neeman J.J. 1999. Membrane processes for biological treatment of contaminated airstreams. *Biotechnol. Bioengrg.*, 63:431–441.

Essila, N.J., Semmens, M.J., Voller, V.R. 2000. Modeling biofilms on gas permeable supports: concentration and activity profiles. *J. Environ. Engr. ASCE.*, 126(3):250–257.

Fitch, M., Sauer, S., Zhang, B. 2000. Membrane biofilters: material choices and diurnal loading effects. *Proceedings USC-TRG Conference on Biofiltration*, University of Southern California, October 19–20, 83–90.

Freitas dos Santos, L.M., Pavasant, P., Strachan, L.F., Pistikopoulos, E.N., Livingston, A.G. 1997. Membrane attached biofilms for waste treatment – Fundamentals and applications. *Pure Appl. Chem.*, 69(11):2459–2469.

Gantzer, C. J. 1995. Membrane dissolution of hydrogen for biological nitrate removal. *68th Annual Conference and Exhibition, Water Environment Federation, Research Symposium*, Miami Beach, FL, 49–60.

Hartmans, S., Leenen, E.J.T.M., Voskuilen, G.T.H. 1992. Membrane bioreactor with porous hydrophobic membranes for waste-gas treatment. In *Biotechniques for Air Pollution Abatement and Odour Control Policies*; A.J. Dragt and J. van Ham, Eds. Elsevier Science Pub, B.V., Amsterdam.

Keskiner, Y., Ergas, S.J. 2001. Control of ammonia and NOx emissions using a nitrifying membrane bioreactor. *Proceedings 94th Annual Meeting of the Air & Waste Management Association*, Orlando, FL.

Lee, K.C., Rittmann, B.E. 2000. A novel hollow-fibre membrane biofilm reactor for autohydrogenotrophic denitrification of drinking water. *Water Sci. Technol.*, 41(4–5):219–226.

Livingston, A.G. 1993. A novel membrane bioreactor for detoxifying industrial wastewater: I. Biodegradation of phenol in a synthetically concocted wastewater. *Biotechnol. Bioengrg.*, 41(10):915–926.

Livingston, A.G. 1994. Extractive membrane bioreactors – a new process technology for detoxifying chemical-industry wastewaters. *J. Chemical Technol. Biotechnol.,* 60(2):117–124.

Livingston, A.G., Arcangeli, J.P., Boam, A.T., Zhang, S.F., Marangon, M., dos Santos, L.M.F. 1998. Extractive membrane bioreactors for detoxification of chemical industry wastes: process development. *J. Membr. Sci.*, 151(1):29–44.

Lund, L.W., Federspiel, W.J., Hattler, B.G. 1996. Gas permeability of hollow fiber membranes in a gas-liquid system. *J. Membr. Sci.*, 117:207–219.

Pankhania, M., Stephenson, T, Semmens, M.J. 1994. Hollow fibre bioreactor for wastewater treatment using bubbleless membrane aeration. *Water Res.*, 28(10):2233–2236.

Parvatiyar, M.G., Govind, R., Bishop, D.F. 1996a. Biodegradation of toluene in a membrane biofilter. *J. Membr. Sci.,* 119:17–24.

Parvatiyar, M.G., Govind, R., Bishop, D.F. 1996b. Treatment of trichloroethylene (TCE) in a membrane biofilter. *Biotechnol. Bioengr.*, 50:57–64.

Pressman, J.G., Georgiou, G, Speitel, G.E. 1999. Demonstration of efficient trichloroethylene biodegradation in a hollow-fiber membrane bioreactor. *Biotechnol. Bioengrg.*, 62:681–692.

Pressman, J.G., Georgiou, G., Speitel, G.E. 2000. A hollow-fiber membrane bioreactor for the removal of trichloroethylene from the vapor phase. *Biotechnol. Bioengr.,* 68(5):548–556.

Reij, M. W., de Gooijer, K.D., de Bont, J.A.M., Hartmans, S. 1995. Membrane bioreactor with a porous hydrophobic membrane as a gas-liquid contactor for waste gas treatment. *Biotechnol. Bioengr.,* 45:107–115.

Reij, M.W., Hamann, E.K., Hartmans, S. 1997. Biofiltration of air containing low concentrations of propene using a membrane bioreactor. *Biotechnol. Prog.,* 13:380–386.

Reij, M.W. and Hartsmans, S. 1996. Propene removal from synthetic waste gas using a hollow fiber membrane bioreactor. *J. Appl. Microbiol. Biotechnol.*, 45:730–736.

Reij, W.M., Keurentjes, J.T.F., Hartmans, S. 1998. Membrane bioreactors for waste gas treatment. *J. Biotechnol.,* 59:155–167.

Semmens, M.J., Gulliver, J.S., Anderson, A. 1999. An analysis of bubble formation using microporous hollow fiber membranes. *Water Environ. Res.*, 71(3):307–315

Shumway, L. 1997. A membrane process for biological treatment of contaminated airstreams. M.S. Thesis, Department of Civil and Environmental Engineering, University of Massachusetts, Amherst.

Sirkar, K.K. 1992. Other new membrane processes. In *Membrane Handbook*; W.S.W. Ho and K.K. Sirkar, Eds. Chapman & Hall, New York.

Speitel, G.E.; McLay, D.S. 1993. Biofilm reactors for treatment of gas streams containing chlorinated solvents. *J. Envir. Engrg., ASCE*, 119, 658–678.

Stephenson, T., Judd. S., Jefferson, B., Brindle, K. 2000. *Membrane Bioreactors for Wastewater Treatment*. IWA Publishing, London.

Togna, P. 2000. Personal communication. EnviroGen Inc. Lawrenceville NJ.

Yamagiwa, K., Yoshida, M., Ito, A., Ohkawa, A. 1998. A new oxygen supply method for simultaneous organic carbon removal and nitrification by a one-stage biofilm process. *Wat. Sci. Tech.*, 37(4–5):117–124.

Yang, M-C, Cussler, E.L. 1986. Designing hollow fiber contactors. *AIChE J.*, 32, 1910.

41

Review of Fine-Scale Air Quality Modeling for Carbon and Health Co-Benefits Assessments in Cities

Andrew Fang and
Anu Ramaswami

Introduction

Globally, cities exist as centers of population, economic activity, and innovation. Currently, cities around the world are grappling with serious air quality challenges that require action from local and national policymakers to protect the well-being of urban residents. Many cities around the world have annual air pollution levels exceeding the WHO standards (Health Effects Institute, 2019), and globally, this excess ambient air pollution contributed to 2.9 million premature mortalities in 2017 (Stanaway et al., 2018). As policymakers attempt to reduce urban air pollution exposure, they are challenged by air pollution emission sources which exist outside the city (and which they may not be able to manage) and inequities in intracity air pollution levels due to pollution "hotspots".

Cities are also important centers of climate change action as evidenced by the over 9,000 cities that have committed to the Global Covenant of Mayors for Climate and Energy and various organizations (e.g., ICLEI and C40) which support cities in their efforts to mitigate and adapt to climate change. Because cities are embedded within larger infrastructure and trade networks, various tools have been developed to assess current community-wide emissions, project future greenhouse gas (GHG) emissions, and assess the impact of climate mitigation policies.

Because cities are concurrently trying to manage carbon and air pollution emissions, urban environmental policies should take into account the co-benefits of carbon mitigation, particularly as they relate to reducing local air pollution levels. Recent literature (Mayrhofer and Gupta, 2016; Nemet, Holloway, and Meier, 2010) has documented the challenges of quantifying co-benefits, but the health benefits of reducing air pollution are too large to ignore. The health benefits of the reduction in air pollutants (PM2.5) from both clean energy and carbon mitigation policies may outweigh the costs of carbon mitigation in the United States (Fowlie, 2018) and globally (Markandya et al., 2018). Further, we know these benefits will be heterogeneous (Muller and Mendelsohn, 2009; Goodkind et al., 2019), and we know they will be concentrated in cities (West et al., 2013). The heterogeneity and size of these health benefits means that they also have significant equity implications (Anderson et al., 2018; Cushing et al., 2018). It is therefore important to develop accurate estimates of these benefits at the urban scale, so that the size and distribution of health co-benefits of carbon mitigation are accurately reflected in environmental policies.

For cities grappling with air quality challenges, the first question is how can we reduce annual average air pollution levels and the number of acute pollution episodes? The answer to this question varies from city-to-city depending on the economic structure of the city and various physical and meteorological characteristics. In particular, cities must identify which sectors (industrial, transportation, residential, agriculture, etc.) contribute to their local air quality challenges and whether the sources of pollution are located within or outside the city. Once policies are identified that will prove beneficial to reducing exposure to local air pollution, the next question becomes how do we make sure these benefits are distributed fairly? Potential approaches to distributing these health benefits equitably are complicated by economic (e.g., Where are marginal benefits highest? Where are marginal costs lowest?) and political concerns (e.g., Which neighborhoods historically have high levels of pollution? Does a given policy reduce this burden?). Policymakers seek to address both sets of questions as they develop strategies for urban air quality management; therefore, co-benefit models need to account for emissions inside and outside the city, while assessing changes in air quality within the city.

Separately, cities have also been developing carbon emission inventories for local climate action planning and for reporting to the Global Covenant of Mayors in Climate and Energy. Carbon accounting of community-wide emissions for cities has become more standardized over the past few years as cities have begun to report emissions through the Global Protocol for Community-Scale Greenhouse Gas Emission Inventories (GPC) (Fong et al., 2014). Community-wide carbon emissions are reported across multiple spatial scales grouped by three categories:

- Scope 1 emissions: Emissions that physically occur within the city
- Scope 2 emissions: Energy-related emissions from the use of electricity, steam, and/or heating/cooling supplied by grids which may or may not cross city boundaries
- Scope 3 emissions: Emissions that physically occur outside the city but are driven by activities taking place within the city's boundaries.

Historically, cities have used production-based approaches to quantify Scope 1 and 2 emissions. Production-based approaches have historically been ill-defined, but many have used the term in reference to territorial emissions or emissions that are occurring geographically within the city boundary (Chen et al., 2019). Recent studies (Lin et al., 2015) have re-defined the term to account for the goods and services imported into the city that are used by industries/manufacturers to better account for the impact of industrial symbiosis and other interventions improving the efficiency of urban-industrial supply chains. But cities may account for different Scope 3 emissions depending on the methodology they are following. For example, many cities have used a consumption-based approaches (C40, 2018; Lin et al., 2017) which account for Scope 1, 2, and 3 emissions due to the production of goods and services that are consumed by residents of the city, but exclude emissions from exported goods and services that are produced by the city and consumed elsewhere. The production-based approach likely underestimates emissions in cities with high consumption activity (such as cities with tourist-based economies), while the consumption-based approach has limited application in cities with high industrial or manufacturing activity (Chavez and Ramaswami, 2013). None of these approaches is perfect, but all seek to better account for carbon emissions associated with city activity that are outside the geographic boundary, and it is therefore critical that there is transparency in the methods used to quantify these emissions.

For cities using consumption-based approaches, there may be a spatial mismatch between the emissions accounted for in the carbon emissions inventory and the local/regional emissions required for air pollution modeling.

Many models and methods have been developed to manage urban air pollution and carbon emissions, but few models have been able to address the nexus of carbon, air pollution, and health co-benefits within cities. Assessing co-benefits at the city scale requires sophisticated models that are able to account for the transboundary nature of the supply chains supporting city activity and the potential for non-local emissions to impact local air quality, while also having the spatial resolution to accurately assess intracity air pollution concentrations. This review assesses the state-of-the-science tools that have been developed to assess co-benefits at the city scale.

Overview of City-Relevant Co-Benefit Models

Models which are able to assess carbon, air pollution, and health co-benefits are summarized below with a focus on nested models. This list is not meant to be exhaustive but illustrates how carbon and air pollution modeling methods can be coupled. Table 1 demonstrates how the authors' view the scope and spatial scale of models that have the ability to assess co-benefits of carbon and air pollution emission reductions. The first type of models develop carbon and air pollution co-pollutant inventories but assign air pollution/health benefits directly to each unit of emission without using air pollution and transport modeling to assess changes in air pollution concentration and exposure (e.g., UrbanFootprint, 2013). A second type of model (e.g., AERMOD) utilizes similar community-scale emission inventories to mechanistically estimate how changes in local emissions impact local air pollution concentration and exposure levels based upon non-linear dynamics of air pollution transport.

Nested models, a third type of model, are most applicable to addressing city-scale co-benefits because they acknowledge the fact that city activity induces emissions outside the territorial boundary of the city. While the location of carbon emissions is not as important due to its global impacts, the location of air pollution emissions is critical to determining to associated impacts on air quality and pollution exposure. For example, SIM-Air (Guttikunda, Nishadh, and Jawahar, 2019; Guttikunda and Jawahar, 2012), GAINS (Wagner et al., 2018), COBRA (USEPA, 2018), and AP3 (National Research Council, 2010) are well-known models which utilize gridded emission inventories and regional estimates of air pollution transport and formation to predict changes in air quality/pollution exposure resulting from changes in emissions. Models like the Carbon Footprinting and Air Pollution Dispersion (CFAD) (Ramaswami et al., 2017) utilize community-wide carbon and air pollution inventories coupled with air pollution models (e.g., AERMOD) to distinguish between how changes in local and regional emissions may impact urban air quality and exposure. Lastly, nested-reduced form models, such as InMap (Tessum et al., 2019; Hill et al., 2019), have the potential to couple carbon and air pollution inventories at multiple scales (national, state, city) with reduced form air pollution models in order to estimate changing air quality and exposure at multiple scales resulting from changes in consumption of goods and services.

Air Pollution Models

In order to assess carbon and air pollution co-benefits, air pollution models are required to connect spatially explicit carbon and air pollutant inventories with air pollution formation and transport in the atmosphere. This is critical because of the non-linear relationship between air pollution emissions and air pollution concentrations (e.g., each unit of air pollutant emission may have a varied effect on concentration). Air pollution concentrations then become the unit of analysis for air quality standards, public health, and exposure estimates. The relationship between emissions and concentration may be governed by physical characteristics of the pollution source (e.g., stack height and toxicity of pollutants), non-linear dynamics of physical processes governing pollution in the atmosphere (e.g., transport, reaction, and deposition), and geographic characteristics of the exposure site (e.g., topography and wind patterns).

TABLE 1 Boundary of Carbon Emission Accounting and Air Pollution Models for City-Relevant Co-Benefit Models

		Boundary of Carbon Emission Accounting		
		Geographic (community-wide, Scope 1+2)	Geographic (community-wide, Scope 1+2+3)	Consumption-based (Scope 1+2+3)
Boundary of Air Pollution Modeling	Air pollution emission inventories	Urban Footprint		
	Community-scale air quality models	AERMOD		
	Nested local-regional-national models (source–receptor)	GAINS, SIM-Air, COBRA, AP3		
	Nested local-regional-national models (dispersion)		CFAD	
	Nested local-regional-national models (reduced form)			InMap

Note: Co-benefit models with the ability to couple across spatial scales are organized based upon the geographic boundaries of the carbon emission accounting and air pollution exposure.

To accurately assess the impact of these factors, air pollution models are needed to determine the dynamics of each unit of emissions once they enter the atmosphere. For city-level co-benefits modeling, these tools need to have the potential utilize urban and regional emissions inventories while evaluating air pollution concentration at a fine-grained spatial resolution so that intercity and intracity air pollution/health benefits can be estimated. This means that the air pollution models that are most applicable at the city scale need to balance data availability and accessibility with the computational complexity necessary to model air pollution transport and formation.

Given the uncertainty in modeling and the need to estimate air quality across varying spatial and temporal resolution, various methods exist for estimating air quality. Monitors are generally used to track air quality over time and to ensure compliance with air quality regulations; they also provide data validation for modeling results. Land use regression techniques use the historical data from monitors in comparison with land use variables (e.g., density) to predict air pollution levels with high spatial resolution. Recently, aerosol optical dispersion techniques (Di et al., 2017; van Donkelaar et al., 2016) have advanced considerably, using satellite data with increasingly high spatial resolution to estimate air pollution concentrations over time in locations lacking monitoring data.

While these methods are extremely sophisticated, they do not connect emissions reductions directly to concentration reductions. Therefore, these techniques are not able to directly inform policy design and implementation but are useful for empirical and historical air quality studies; for example, many studies have explored how health outcomes change before and after a policy change (e.g., Ebenstein et al., 2017). The models described below estimate air pollution concentrations and changes in exposure based on spatially resolved emissions inventories. This is critical to developing prospective scenarios that determine a priori the benefits of policies and interventions which reduce air pollution.

The challenge for air pollution models is to balance the complexity of modeling atmospheric processes that induce primary and secondary formation of pollutants with the spatial and temporal resolution necessary to answer policy-relevant questions, for example, where are high air pollution concentrations located or how can cities improve region-wide air quality? Urban air pollution concentration will be a function of the emissions (local and non-local) and the meteorological conditions (wind speed, wind direction, mixing height, etc.) that govern pollution formation at any given time. In practice, modelers must make the choice between model processing time and spatial resolution (Gilmore et al., 2019; Tessum, Hill, and Marshall, 2017). This tradeoff is very important for determining which models can assess the benefits of policy-induced emissions reduction at the city scale.

Application of Fine-Scale Air Pollution Models to Co-Benefits in Cities

Air pollution and transport models are tailored to specific spatial and temporal scales, which means they can estimate air pollutant concentrations across multiple spatial scales ranging from global to between buildings. Figure 1 illustrates the spatial dimensions of carbon and air pollution emission inventories compared to the scale at which air pollution models operate. The discussion below will focus on those tools that have potential to be applied to urban-scale co-benefits evaluation and policy scenarios.

Models that are able to address co-benefits at the urban scale need to be designed to utilize emissions data that distinguishes between local and non-local emissions, estimate changes in air pollution exposure within cities, and be customizable to policies that are specific to the jurisdiction in question. Not all air pollution models will be well suited to evaluate co-benefits at the city-wide scale, because many may be designed for alternate purposes such as regulatory compliance. The air pollution models which are most relevant need to differentiate between in-boundary and transboundary emissions so that urban and non-urban pollution control strategies can be evaluated. They must also account for the variation of inter- and intra-city air pollution exposure, so that the distribution of air pollution and health benefits can be assessed to determine which areas and populations will benefit most from air pollution control policies being considered. To systematically evaluate the air pollution and transport models that have the ability to estimate city-scale co-benefits, the following air pollution model characteristics need to be considered:

1. Primary/Secondary PM2.5 formation – Primary PM2.5 drives the local emissions component of urban PM2.5 concentrations because it consists of the particles directly emitted during energy combustion, industrial, and agricultural processes. In contrast, secondary PM2.5 is formed in the atmosphere by reactions between SOx, NOx, NH_3, etc. that may occur over larger provincial and regional areas.

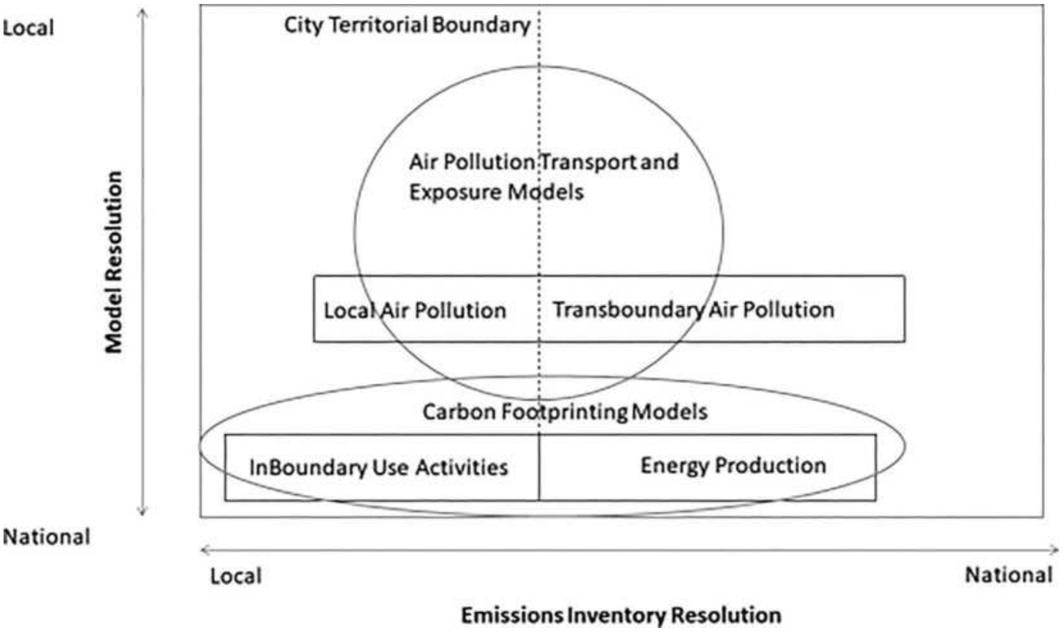


FIGURE 1 Spatial Resolution of Carbon Footprinting and Air Pollution Models – Geographic Scale of Carbon/Air Pollution Models shown as black ovals and geographic scale of emissions inventories shown as black boxes. Although carbon and air pollution emissions are aligned at the city scale, carbon and air pollution models evaluate outcomes at different scales because air pollution damages occur on a local-regional scale while climate damages occur on a global scale.

2. Spatial Resolution – Model inputs (emission inventories) and outputs (concentration estimates) will be set to a standardized or variable grid ranging from 0.1 to 50 km. The spatial resolution of these grids has implications for data availability and the model's ability to assess air pollution exposure within the city
3. Ground-Truthing and Relevance – Some models are comparable to real-time monitored air pollution levels, while others are intended to predict changes in air pollution levels based on changes in emissions. Recently, a number of reduced form tools (AP3, EASIUR, InMap) have been developed to quantify marginal benefits in pollution reduction. Critically, these reduced form tools do not claim to be more accurate than previous models, nor is their intention to be used for regulatory compliance. Instead, they focus on developing assessments of changes in pollution exposure due to policies that reduce local, regional, or national emissions.
4. Application – Certain air pollution models may be better suited for different policy-relevant questions. Models may be tailored to assess the health impacts of changes in emissions and exposure to air pollution, while others may be better suited for the identification of concentrated areas of pollution.

Although a variety of air pollution models exist, the following section is focused on the subset of models that can be readily applied to city-scale co-benefits. It is critical to note that the most state-of-the-art models may not be widely used by local policymakers due to data, computational, and technical capacity limitations. Reduced form models have been developed to make air pollution models more accessible and more policy-relevant by reducing computational effort and allowing for broader evaluation tailored to location-specific policies/interventions. The following three sections evaluate three types of air pollution models using the criteria above to determine which air pollution and transport models are best suited for different questions surrounding co-benefits at the urban scale.

Dispersion Models

- *Description*: Gaussian plume dispersion models (e.g., AERMOD) estimate air pollution levels that are downwind of individual sources or source groups. They are useful for predicting pollution impacts within cities due to nearby sources but are not recommended for predictions of long-range pollution transport (>50 km) (Cimorelli et al., 2005; USEPA, 2015).
- *PM2.5 Formation*: Primary PM2.5 only; assumes that increased concentration in cities is mainly due to local sources.
- *Spatial Resolution*: Fine-scale resolution (<1 km) enables assessment of intracity air pollution exposure.
- *Application*: Used for regulatory purposes to determine near-source impacts of large emitters, but generally cannot estimate secondary formation of PM2.5 or predict long-range pollution transport.
- *Co-Benefits Examples*: Applied at the local to regional scale, can be nested with carbon footprinting models (e.g., CFAD) (Ramaswami et al., 2017).

Chemical Transport Models

- *Description*: Chemical transport models (CTMs) are three-dimensional mechanistic models that predict ambient concentrations of pollutants using mass balance principles accounting for emissions, transport, dispersion by winds, chemical transformations, and atmospheric removal processes. CTMs are the most scientifically sophisticated and rigorous tools available for linking emissions to ambient concentrations (e.g., CMAQ, WRF-Chem) (Byun and Ching, 1999; Grell et al., 2005).
- *PM2.5 Formation*: Primary and Secondary PM2.5 including gridded emission inventory of all PM2.5 precursors.
- *Spatial Resolution*: Coarse (>4 km) due to model complexity, computational processing time, and availability of nationwide fine-scale emissions inventory data (US-based models utilize the triennial National Emissions Inventory).

TABLE 2 Summary of Air Pollution Models for Urban Application

	Increasing Complexity		
	Reduced Form Models	Air Pollution Dispersion	Chemical Transport Models
Example	InMap, AP3	AERMOD	CMAQ, WRF-Chem
Contribution	Less computationally intensive, predictions of marginal damages	High spatial resolution, intracity variability (identification of pollution hotspots)	State-of-the-art physical and chemical transport process model, most complete spatial/temporal coverage
Limitations	Limited temporal coverage, limited comparability to observed concentrations	Only primary PM2.5, not recommended for long-range transport	Computationally intensive, coarse spatial scale`

Note: In order of increasing computational complexity, this table provides examples, contribution, and limitations of air pollution models that can be applied to urban co-benefits.

- *Application*: Can be compared to real-time air pollution monitoring and used for regulatory impact assessment, but limited spatial and temporal resolution.
- *Co-Benefits Examples*: Applied at the National or Regional Scale (Thompson et al., 2016; Zhang et al., 2016).

Reduced Form Models

- *Description*: To improve the availability and accessibility of state-of-the-science CTM air quality modeling and cost estimates, the air quality research community has recently created a new set of simplified models, known as reduced form or reduced-complexity models, to estimate the marginal social costs of air pollution in terms of monetized health damages per pollutant (Gilmore et al., 2019). These models estimate marginal social costs with nested local-regional-national emission inventories.
- *PM2.5 Formation*: Variable – AP3 (Primary PM2.5 only), InMap (Primary and Secondary PM2.5 Formation).
- *Spatial Resolution*: InMap (1–48 km) but coarser than dispersion models.
- *Application*: Able to model prospective policy scenarios and sensitivity to model/scenario uncertainties due to lower computational intensity. These models are meant to estimate marginal air quality and health benefits based upon the changes in air quality that can be attributed to changes in emissions; may be limited in their comparability to real-time pollution data.
- *Co-Benefits Examples*: Applied at multiple scales – national (Markandya et al., 2018), regional (Muller and Mendelsohn, 2009), and local (Goodkind et al., 2019; Paolella et al., 2018) (Table 2).

In the US context, air pollution dispersion and CTMs have historically been used by state and national regulators to ensure regulatory compliance and perform regulatory impact assessments. These models on their own have robust applications to urban air quality management strategies, but are not necessarily aligned with local analyses that seek to address carbon, air pollution, and health co-benefits. Dispersion and reduced form models have already been integrated with city-scale carbon accounting methods due to their ability to incorporate nested emission inventories at multiple spatial scales. Local policymakers and practitioners should use these tools or tools with similar functionality to quantify the magnitude and distribution of carbon, air pollution, and health co-benefits

Conclusion

Reducing air pollution has significant benefits both globally and locally. Environmental policies that are designed and implemented at the local-regional scale must take into account both the magnitude and distribution of these health benefits because of their implications on cost-benefit analyses and developing political support for local environmental policies. Models developed to date have been limited in their ability to measure urban carbon and air pollution co-benefits. First, it has been challenging to acquire data on activity and emissions delineated between local sources and regional/national sources of pollution (>50 km from city). Second, the complexity of state-of-the-art models leads to limitations in terms of the software/equipment, computational time, and technical capacity needed to apply these tools towards the development of local air pollution and carbon policy. Further, these state-of-the-art models have been designed for state/national level assessments and may not have the spatial resolution necessary to estimate urban air pollution/health benefits.

A variety of commercially available models have been developed to assess how carbon emissions will respond to changes in land use, technology, and policy/behavior changes at the urban scale. Certain models also attempt to measure air pollution benefits of these changes using linear marginal benefits estimates (UrbanFootprint, 2013) or neural networks to predict air quality changes based on weather, traffic, and air pollution sensor data (Siemens, 2019). While these estimates may have some predictive value, these models have limited ability to inform the design and implementation of local policies that impact air pollution. Models that inform policy should estimate both the change in criteria air pollutant emissions *and* the changes in local air pollution concentrations due to the non-linear relationship between air pollution emissions reduction and air pollution exposure reduction. This is critical because of the heterogeneity that exists in the marginal benefits of reducing one unit of air pollution and because these benefits will be higher within cities but distributed unequally throughout the city. These marginal benefits need to be estimated at high spatial resolution (e.g., sub-county scale) because there may be an order of magnitude difference within counties (Goodkind et al., 2019).

The recent development of reduced-complexity models is a promising alternative, recognizing the importance of estimating the marginal benefits of reductions in air pollution with high spatial resolution and limited computational complexity (Gilmore et al., 2019). These models combine the ability to estimate the impact of both local/long-range pollution sources on air quality levels within cities with the granularity to estimate variations in air pollution exposure at the sub-county level. As these models develop and become used in discussions around equity (Tessum et al., 2019; Tessum, Hill, and Marshall, 2017), it is important to assess the relationship between mechanistic and empirical models.

The spatial resolution of relevant co-benefits models is a critical variable to assess, particularly when considering equity implications of the distribution of air pollution/health benefits (Paolella et al., 2018). Because carbon and air pollution emission inventories also take place across nested spatial scales (city, region, national), models addressing co-benefits must also have the ability to differentiate between urban and non-urban emission reductions. Models which can couple spatially explicit emission inventories with regional air pollution transport methods are best suited for assessing co-benefits at the city scale. If these models can combine regional air pollution transport with high spatial resolution at the sub-county scale, then they have the ability to inform environmental policy design so that the health benefits of reducing air pollution are both optimal and distributed more equitably.

Given the uncertainties in underlying local emissions, exposure, and demographic data, policymakers must acknowledge two key truths. First, not all emissions reductions are created equal; improving air quality across urban areas requires an understanding of the implications of emitter location and physical characteristics of local and non-local emitters. Second, the marginal health benefits of reducing local air pollution will not be distributed uniformly across the city, so policy needs to be designed to account for the presence of concentrated areas of pollution, particularly when addressing multi-pollutant policies (e.g., policies and interventions which reduce multiple pollutants, such as carbon and criteria air pollutant emissions). The complex nature of air pollution and atmospheric transport has resulted in multiple models that may be able to serve different co-benefits-related questions. Moving forward, it is critical that the modeling community develops tools that are suited to the questions that policymakers are asking so that environmental policy design is being informed by the best available science. These tools need to have the ability to utilize local and regional emissions data across multiple spatial scales spanning the city boundary, while also measuring air pollution exposure variation within the city. The existing suite of tools is not perfect but can be used to evaluate the magnitude and distribution of carbon, air pollution, and health co-benefits in cities. Policymakers can use these tools for prospective analyses to determine which carbon and air pollution policy options may have the most health benefits at the city scale and then utilize other empirical methods involving satellite-derived and air quality monitor data to evaluate how these policies change air quality and health outcomes over time.

References

Anderson, C. M., Kissel, K. A., Field, C. B., and K. J. Mach. 2018. "Climate Change Mitigation, Air Pollution, and Environmental Justice in California." *Environmental Science & Technology* 52 (18): 10829–38. https://doi.org/10.1021/acs.est.8b00908.

Byun, D. W., and J. K. S. Ching. 1999. "Science Algorithms of the EPA Models-3 Community Multiscale Air Quality (CMAQ) Modeling System." *United States Environmental Protection Agency* 44 (6): 1765–78. https://nepis.epa.gov/Exe/ZyPDF.cgi/30003R9Y.PDF?Dockey=30003R9Y.PDF.

C40. 2018. "Consumption-Based GHG Emissions of C40 Cities." https://www.c40.org/researches/consumption-based-emissions.

Chavez, A., and A. Ramaswami. 2013. "Articulating a Trans-Boundary Infrastructure Supply Chain Greenhouse Gas Emission Footprint for Cities: Mathematical Relationships and Policy Relevance." Energy Policy 54: 376–84. https://doi.org/10.1016/j.enpol.2012.10.037.

Chen, G., Shan, Y., Hu, Y., Tong, K., Wiedmann, T., Ramaswami, A., Guan, D., Shi, L. and Y. Wang. 2019. "Review on City-Level Carbon Accounting." *Environmental Science & Technology* 53 (10): 5545–58. https://doi.org/10.1021/acs.est.8b07071.

Cimorelli, A. J., Perry, S. G., Venkatram, A., Weil, J. C., Paine, R. J., Wilson, R. B., … R. W. Brode. 2005. "AERMOD: A Dispersion Model for Industrial Source Applications. Part I: General Model Formulation and Boundary Layer Characterization." *Journal of Applied Meteorology* 44 (5): 682–93. https://doi.org/10.1175/JAM2227.1.

Cushing, L., Blaustein-Rejto, D., Wander, M., Pastor, M., Sadd, J., Zhu, A., and R. Morello-Frosch. 2018. "Carbon Trading, Co-Pollutants, and Environmental Equity: Evidence from California's Cap-and-Trade Program (2011–2015)." *PLOS Medicine* 15 (7): e1002604. https://doi.org/10.1371/journal.pmed.1002604.

Di, Q., Wang, Y., Zanobetti, A., Wang, Y., Koutrakis, P., Choirat, C., Dominici, F., and J.D. Schwartz. 2017. "Air Pollution and Mortality in the Medicare Population." *New England Journal of Medicine* 376 (26): 2513–22. https://doi.org/10.1056/NEJMoa1702747.

Ebenstein, A., Fan, M., Greenstone, M., He, G., and M. Zhou. 2017. "New Evidence on the Impact of Sustained Exposure to Air Pollution on Life Expectancy from China's Huai River Policy." *Proceedings of the National Academy of Sciences* 114 (39): 10384–89. https://doi.org/10.1073/pnas.1616784114.

Fong, W. K., Sotos, M., Doust, M., Schultz, S., Marques, A., and C. Deng-Beck. 2014. "Global Protocol for Community-Scale Greenhouse Gas Emission Inventories." http://www.iclei.org/fileadmin/user_upload/ICLEI_WS/Documents/Climate/GPC_12-8-14_1_.pdf.

Fowlie, M. 2018. "Air Pollution Co-Benefits Matter." Energy Institute Blog. 2018. https://energyathaas.wordpress.com/2018/11/13/air-quality-matters/.

Gilmore, E. A., Heo, J., Muller, N. Z., Tessum, C. W., Hill, J. D., Marshall, J. D., and P.J. Adams. 2019. "An Inter-Comparison of Air Quality Social Cost Estimates from Reduced-Complexity Models." Environmental Research Letters. http://iopscience.iop.org/10.1088/1748-9326/ab1ab5.

Goodkind, A. L., Tessum, C. W., Coggins, J. S., Hill, J. D., and J. D. Marshall. 2019. "Fine-Scale Damage Estimates of Particulate Matter Air Pollution Reveal Opportunities for Location-Specific Mitigation of Emissions." *Proceedings of the National Academy of Sciences* 116 (18): 8775–80. https://doi.org/10.1073/pnas.1816102116.

Grell, G. A., Peckham, S. E., Schmitz, R., McKeen, S. A., Frost, G., Skamarock, W. C., and B. Eder. 2005. "Fully Coupled 'Online' Chemistry within the WRF Model." Atmospheric Environment 39 (37): 6957–75. https://doi.org/10.1016/j.atmosenv.2005.04.027.

Guttikunda, S. K., and P. Jawahar. 2012. “Application of SIM-Air Modeling Tools to Assess Air Quality in Indian Cities.” Atmospheric Environment 62: 551–61. https://doi.org/10.1016/j.atmosenv.2012.08.074.

Guttikunda, S. K., Nishadh, K. A., and P. Jawahar. 2019. “Air Pollution Knowledge Assessments (APnA) for 20 Indian Cities.” *Urban Climate* 27: 124–41. https://doi.org/10.1016/j.uclim.2018.11.005.

Health Effects Institute. 2019. “State of Global Air 2019.” http://www.stateofglobalair.org.

Hill, J. D., Goodkind, A., Tessum, C. W., Thakrar, S., Tilman, D., Polasky, S., … J.D. Marshall. 2019. “Air-Quality-Related Health Damages of Maize.” *Nature Sustainability* 2 (5): 397–403. https://doi.org/10.1038/s41893-019-0261–y.

Lin, J., Hu, Y., Cui, S., Kang, J., and A. Ramaswami. 2015. “Tracking Urban Carbon Footprints from Production and Consumption Perspectives.” *Environmental Research Letters* 10 (5): 1–12. https://doi.org/10.1088/1748–9326/10/5/054001.

Lin, J., Hu, Y., Zhao, X., Shi, L. and J. Kang. 2017. “Developing a City-Centric Global Multiregional Input-Output Model (CCG-MRIO) to Evaluate Urban Carbon Footprints.” Energy Policy 108: 460–66. https://doi.org/10.1016/j.enpol.2017.06.008.

Markandya, A., Sampedro, J., Smith, S. J., Van Dingenen, R., … M. González-Eguino. 2018. “Health Co-Benefits from Air Pollution and Mitigation Costs of the Paris Agreement: A Modelling Study.” *The Lancet Planetary Health* 2 (3): e126–33. https://doi.org/10.1016/S2542–5196(18)30029–9.

Mayrhofer, J. P., and J. Gupta. 2016. “The Science and Politics of Co-Benefits in Climate Policy.” Environmental Science and Policy 57: 22–30. https://doi.org/10.1016/j.envsci.2015.11.005.

Muller, N. Z., and R. Mendelsohn. 2009. “Efficient Pollution Regulation: Getting the Prices Right.” *American Economic Review* 99 (5): 1714–39. https://doi.org/10.1257/aer.99.5.1714.

National Research Council. 2010. Hidden Costs of Energy: Unpriced Consequences of Energy Production and Use – Appendix C. Environmental Forum. Vol. 27. Washington, DC: National Academies Press. https://doi.org/10.17226/12794.

Nemet, G. F., Holloway, T., and P. Meier. 2010. “Implications of Incorporating Air-Quality Co-Benefits into Climate Change Policymaking.” *Environmental Research Letters* 5 (1): 014007. https://doi.org/10.1088/1748–9326/5/1/014007.

Paolella, D. A., Tessum, C. W., Adams, P. J., Apte, J. S., Chambliss, S., Hill, J., … J. D. Marshall. 2018. “Effect of Model Spatial Resolution on Estimates of Fine Particulate Matter Exposure and Exposure Disparities in the United States.” *Environmental Science & Technology Letters* 5 (7): 436–41. https://doi.org/10.1021/acs.estlett.8b00279.

Ramaswami, A., Tong, K., Fang, A., Lal, R. M., Nagpure, A. S., Li, Y., … S. Wang. 2017. “Urban Cross-Sector Actions for Carbon Mitigation with Local Health Co-Benefits in China.” Nature Climate Change 7 (10): 736–42. http://dx.doi.org/10.1038/nclimate3373.

Siemens. 2019. “City Air Management Tool.” 2019. https://new.siemens.com/global/en/company/topic-areas/intelligent-infrastructure/city-performance-tool.html.

Stanaway, J. D., Afshin, A., Gakidou, E., Lim, S. S., Abate, D., Abate, K. H., … C. J. L. Murray. 2018. “Global, Regional, and National Comparative Risk Assessment of 84 Behavioural, Environmental and Occupational, and Metabolic Risks or Clusters of Risks for 195 Countries and Territories, 1990–2013: A Systematic Analysis for the Global Burden of Disease Study 2017.” *The Lancet* 392 (10159): 1923–94. https://doi.org/10.1016/S0140-6736(18)32225-6.

Tessum, C. W., Apte, J. S., Goodkind, A. L., Muller, N. Z., Mullins, K. A., Paolella, D. A., … J. D. Hill. 2019. “Inequity in Consumption of Goods and Services Adds to Racial–Ethnic Disparities in Air Pollution Exposure.” *Proceedings of the National Academy of Sciences* 116 (13): 6001–6. https://doi.org/10.1073/PNAS.1818859116.

Tessum, C. W., Hill, J. D., and J. D. Marshall. 2017. “InMAP: A Model for Air Pollution Interventions.” *PLoS ONE* 12 (4): e0176131. https://doi.org/10.1371/journal.pone.0176131.

Thompson, T. M., Rausch, S., Saari, R. K., and N. E. Selin. 2016. "Air Quality Co-Benefits of Subnational Carbon Policies." *Journal of the Air & Waste Management Association* 66 (10): 988–1002. https://doi.org/10.1080/10962247.2016.1192071.

UrbanFootprint. 2013. "Technical Documentation Model Version 1.0." https://urbanfootprint.com/.

USEPA. 2015. Revision to the Guideline on Air Quality Models: Enhancements to the AERMOD Dispersion Modeling System and Incorporation of Approaches to Address Ozone and Fine Particulate Matter. https://www3.epa.gov/ttn/scram/11thmodconf/9930-11-OAR_AppendixW_Proposal.pdf.

USEPA. 2018. "User's Manual for the Co-Benefits Risk Assessment (COBRA) Screening Model." https://www.epa.gov/statelocalenergy/users-manual-co-benefits-risk-assessment-cobra-screening-model.

van Donkelaar, A., Martin, R. V., Brauer, M., Hsu, N. C., Kahn, R. A., Levy, R. C., … D.M. Winker. 2016. "Global Estimates of Fine Particulate Matter Using a Combined Geophysical-Statistical Method with Information from Satellites, Models, and Monitors." *Environmental Science & Technology* 50 (7): 3762–72. https://doi.org/10.1021/acs.est.5b05833.

Wagner, F., Borken-Kleefeld, J., Kiesewetter, G., Klimont, Z., Nguyen, B., Rafaj, P., and W. Schopp. 2018. "The GAINS PMEH-Methodology – Version 2.0."

West, J. J., Smith, S. J., Silva, R. A., Naik, V., Zhang, Y., Fry, M. M., Anenberg, S., … J.-F. Lamarque. 2013. "Co-Benefits of Mitigating Global Greenhouse Gas Emissions for Future Air Quality and Human Health." *Nature Clim. Change* 3 (10): 885–89. https://doi.org/10.1038/NCLIMATE2009.

Zhang, Y., Bowden, J. H., Adelman, Z., Naik, V., Horowitz, L. W., Smith, S. J., and J. J. West. 2016. "Co-Benefits of Global and Regional Greenhouse Gas Mitigation for US Air Quality in 2050." *Atmospheric Chemistry and Physics*. https://doi.org/10.5194/acp-16-9533-2016.

42

Thermal Energy: Solar Technologies

Muhammad Asif
and Tariq Muneer

Introduction

The sun is a source of energy on Earth, and in a sense, it is a source of life, as energy is the most important commodity for life besides providing light and heat to the planet. The reaction between the sun's energy and Earth's atmosphere determines weather patterns and rainfall, and Earth's tilt towards the sun creates the seasons. Its role in photosynthesis helps plants to grow and in biodegradation complete the natural cycle of ecosystems. The sun also sources several other forms of energy on the planet: wind power depends on the sun's impact on atmospheric movement as it creates wind patterns; through photosynthesis, sun contributes to bioenergy (wood and other organic materials); and fossil fuels indirectly owe their creation millions of years ago to solar energy.[1]

Solar energy is one of the most promising renewable technologies. It is abundant, inexhaustible, environmentally friendly, and widely available. Solar energy has the potential not only to play a very important role in providing most of the heating, cooling, and electricity needs of the world, but also to solve global environmental problems. Solar energy can be exploited through solar thermal and solar photovoltaic (SPV) routes for various applications. While SPV technology enables direct conversion of sunlight into electricity through semiconductor devices called solar cells, solar thermal technologies utilize the heat energy from the sun for a wide range of purposes.

Solar thermal technologies are quite diverse in terms of their operational characteristics and applications—they include fairly simple technologies such as solar space heating and solar cooking as well as complex and sophisticated ones like solar air conditioning and solar thermal power generation. Solar thermal technologies have also a broad bandwidth in terms of their economical standing. Solar water heating and solar space heating, for example, are very cost-effective and are regarded among the most economical renewable energy technologies, while high-temperature technologies such as solar thermal power generation and solar air conditioning are on the higher economic bandwidth. Solar thermal technologies on the basis of their working temperature can be classified into the following three types:

- Low-temperature technologies (working temperature <70°C)—solar space heating, solar pond, solar water heating, and solar crop drying.
- Medium-temperature technologies (70°C< working temperature <200°C)—solar distillation, solar cooling, and solar cooking.
- High-temperature technologies (working temperature >200°C)—solar thermal power generation technologies such as parabolic trough, solar tower, and parabolic dish.

In the coming sections, this entry provides an overview of these technologies, briefly highlighting their technology fundamental and operating principles.

Solar Space Heating

Solar energy can be used to accomplish heat for comfort in buildings. The application, referred to as solar space heating, is used to optimize the reduction of auxiliary energy consumption in such a way that minimum overall cost is obtained. In combination with conventional heating equipment, solar heating provides the same levels of comfort, temperature stability, and reliability as conventional systems. A building that includes some arrangement to admit, absorb, store, and distribute solar energy as an integral part is also referred to as a solar house. A solar space-heating system can consist of a passive system, an active system, or a hybrid of the two.

Passive Space Heating

In passive space heating, the buildings are designed or modified so that they independently capture, store, and distribute solar heat throughout the building without using any electrical or mechanical equipment. Inherently flexible passive solar design principles typically accrue energy benefits with low maintenance risks over the life of the building. The design does not need to be complex, but it does involve knowledge of solar geometry, window technology, and local climate. Passive solar heating techniques generally fall into three categories: direct solar gain, indirect solar gain, and isolated solar gain, as shown in Figure 1.

Direct Solar Gain Design

Direct gain passive designs typically have large windows with predominately equatorial aspects. In this design, solar radiation directly penetrates and is stored in the building's inherent thermal mass, in materials such as concrete, stone floor slabs, or masonry partitions that hold and slowly release heat.

Indirect Solar Gain Design

In indirect solar passive design, a glazed heat collector, also referred to as Trombe wall, collects and stores solar radiation during the day. A Trombe wall consists of an 8–16″ thick masonry wall coated with a dark, heat-absorbing material and covered by a single or double layer of glass, placed from about

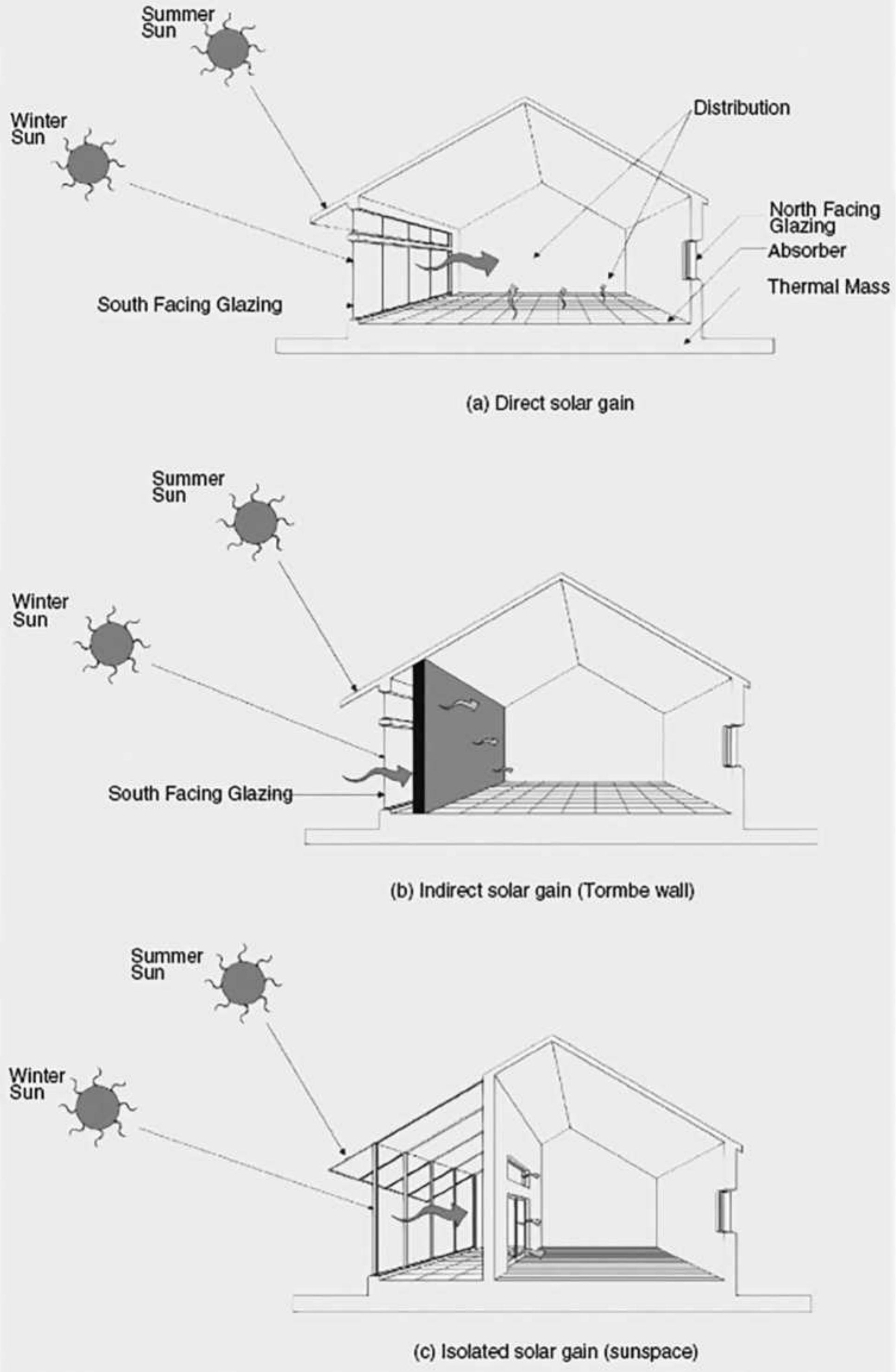


FIGURE 1 Passive solar space-heating principles. a-c captions adequately describe the relevant techniques, same is the case with Figure 6.

¾″ to 6″ away from the masonry wall. Heat from the sun is stored in the air space between the glass and the dark material, and conducted slowly to the interior of the building through the masonry through the conduction and convection mechanisms.

Isolated Solar Gain Design

In isolated solar passive systems, an extra highly glazed unheated room—a sun-space or conservatory—is added to the south side of the house. Solar gains always make sun-spaces warmer than the outside air, and this reduces heat losses from the house and warms any ventilation air that passes through the sun-space. When solar gains raise the sun-space above house temperature, the heat collected can be let into the house by opening communicating doors and windows.[2]

Active Space Heating

In active space heating of buildings, additional electrical and mechanical equipment is incorporated to circulate solar heated water or air. The main components of an active system are the heat collectors, storage tanks or pebble bed storage, heat exchangers, heat emitters, fans/pumps, connecting pipes or ducts, and controls. Active solar heating systems can be designed to provide the same levels of control of condition in the heated spaces as conventional systems. With indoor temperature essentially fixed at or little above a minimum, load estimations can be done by conventional methods. Passively heated buildings in many cases are not controlled within the same narrow temperature ranges.[3]

Hybrid Solar Space Heating

Solar space-heating system can also be of hybrid nature, combining both the passive and active modes. For example, in a hybrid system, a roof-space collector accomplishes passive collection of solar energy that can be actively distributed in the house using a fan and associated ductwork.

Solar Water Heating

Water heating is an essential feature of energy requirements in industrial and commercial sectors in general and in domestic sector in particular. A solar water heater consists of two main elements—the collector and the water storage tank, which, respectively, have the functions of absorbing solar radiation and transferring it to the water, and storing the water for usage. The collectors in solar water heaters can be broadly classified into two categories—flat plate and evacuated tube. A flat plate collector consists of an absorber plate that absorbs solar energy, while a glazing above it is used to reduce convective heat loss. An evacuated tube collector consists of tubes with vacuum maintained between the tubes and glazing for better protection against convective heat loss.

Solar water heaters come in three main types: thermosyphon, built-in-storage, and forced circulation. There are two operating principles for solar water heaters: passive system, which relies on natural circulation of water (such as thermosyphon, built-in-storage types); and active system, which uses an external element such as an electric pump to circulate the water (such as the forced circulation type). Another criterion that distinguishes solar water heaters is the way they transfer heat to water. Again, there are two types: direct system, in which the collector itself transfers heat to water; and indirect system, in which a heat-transfer fluid, circulating in collector in a closed loop, transfers heat to water through a heat exchanger. Figure 2 shows an indirect active solar water heater.

The efficiency of a solar water heater depends upon its design and the available solar radiation. In this entry, solar water heating has been classified as a low-temperature thermal technology because most of its application is in residential sector where it operates at a temperature of ≤70°C. Also, in industrial applications, solar water heaters are used as preheaters and to hold supply water at almost the

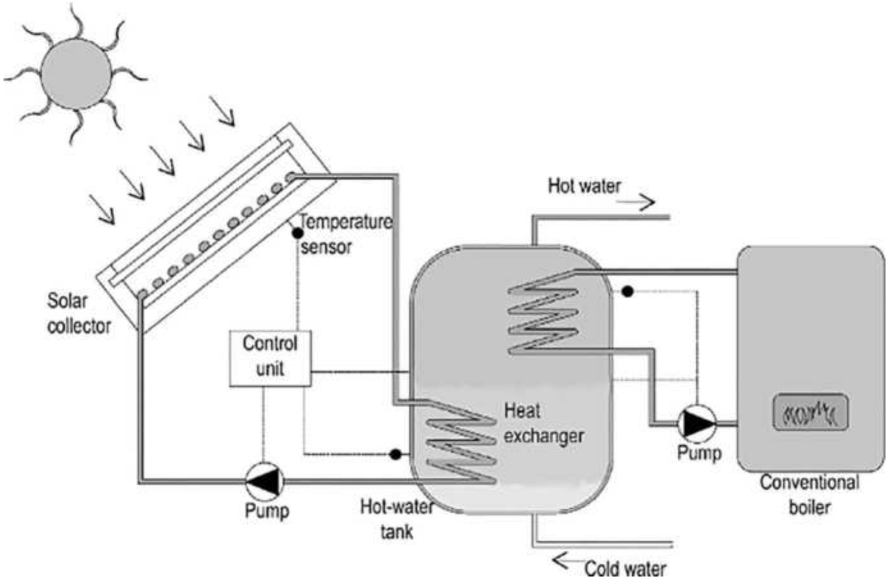


FIGURE 2 Indirect active solar water heating system.

same temperature for further heating by conventional means. Solar water heaters are a cost-effective technology—the payback period of solar water heaters can be as low as 3 years while having a service life of more than 20 years. Owing to its technical and economical viability across the world, solar water heating is one of the most established and efficient application of solar energy. Among solar thermal technologies, solar water heating holds the greatest market share and the highest market growth rate.

Solar Ponds

Solar ponds are naturally occurring salt gradient lakes that collect and store solar energy. A solar pond contains salt water with increasing concentrations of salt, hence the density of the solution. When solar radiation is absorbed, the density gradient prevents heat in the lower layers from moving upward by convection and leaving the pond. This results in an increased temperature at the bottom of the pond and a near atmospheric temperature at the top of the pond. The phenomenon of solar ponds was first discovered in 1902 by von Kalecsinsky, who reported that the Medve Lake in Transylvania, containing nearly saturated NaCl solution at a few meters depth with almost fresh water at its surface, had a bottom temperature of 70°C.

A solar pond has three distinctive zones. The top layer is the surface zone that has a low salt content and is at atmospheric temperature. It is also called the upper convective zone (UCZ), as shown in Figure 3. The bottom layer has a very high salt content and is at a high temperature, 70°C–90°C. This is the zone that collects and stores solar energy in the form of heat, and it is called the lower convective zone (LCZ). There is an intermediate insulating zone with a salt gradient. It establishes a density gradient that prevents heat exchange by natural convection, and hence, it is called the nonconvective zone (NCZ). In this zone, salt content increases with depth, creating salinity.

Solar ponds can be broadly classified into two main types: nonconvective and convective. In nonconvective solar ponds, the heat loss to environment is reduced by suppressing natural convection normally by using salt stratification. While in convective ponds, heat loss to environment is reduced by covering the pond surface with a transparent material. The heat trapped in the solar ponds can be used for many different purposes, such as industrial process heating, the heating of buildings, desalination, and to drive a turbine for generating electricity.

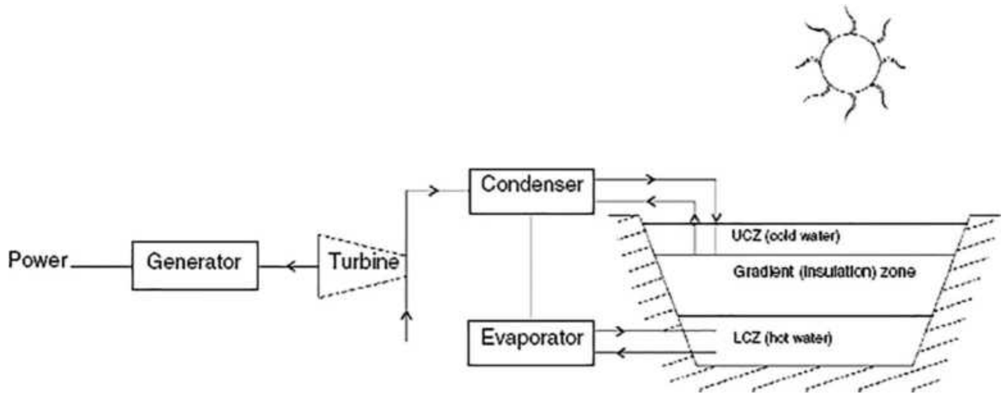


FIGURE 3 Schematic of solar pond power generation system.

The first artificial solar pond was developed in Israel in 1958. Since then, many countries such as Australia, the United States, China, India, Iran, Italy, and Mexico have constructed solar ponds, mostly for research and development purposes. During the last decade, significant success in operational practices and applications of solar pond technologies has been achieved.[4]

Solar Crop Drying

Drying is the oldest technique used to preserve food. Until around the end of the 18th century when canning was developed, drying was virtually the only method of food preservation. Solar energy is the main driving force that utilizes warm air to dry food. In drying, the moisture from the food is reduced to a certain level—as low as 5–25% depending on the type of food-to prevent decay and spoilage in an environment free of contaminations such as dust and insects. Successful drying depends on[5]

- Enough heat to draw out moisture, without cooking the food
- Dry air to absorb the released moisture
- Adequate air circulation to carry off the moisture.

Solar drying can be carried out in open air under the sun by simply spreading the material on a clean surface or in particularly designed solar dryers. Solar dryers, however, exhibit many advantages over open air drying. First, solar dryers are more efficient because they require lesser drying time and area. Second, the product is protected from rain, insects, animals, and dust, which may contain fecal material. Third, faster drying reduces the likelihood of mold growth. Fourth, higher drying temperatures mean that more complete drying is possible, and this may allow much longer storage times (only if rehumidification is prevented in storage). Finally, more complex types of solar driers allow some control over drying rates. Solar dryers can be made in many different designs depending upon various factors, i.e., the type of produce, scale of operation, and local economical and environmental conditions. In terms of their operational mode, solar dryers can be broadly classified into two main types, active and passive dryers, which can both be further subclassified into direct (in which the produce is directly heated from sun) and indirect types (in which the produce is not directly exposed to sun).

Almost all types of food—for example, vegetables, fruits, milk, herbs, spices, meat, and fish—can be dried by solar energy. The advantages of solar food drying are numerous. Dried foods, for example, are tasty, nutritious, lightweight, easy to prepare, and easy to store and use. The energy input is less than what is needed to freeze or can, and the storage space is minimal compared with that needed for canning jars and freezer containers.

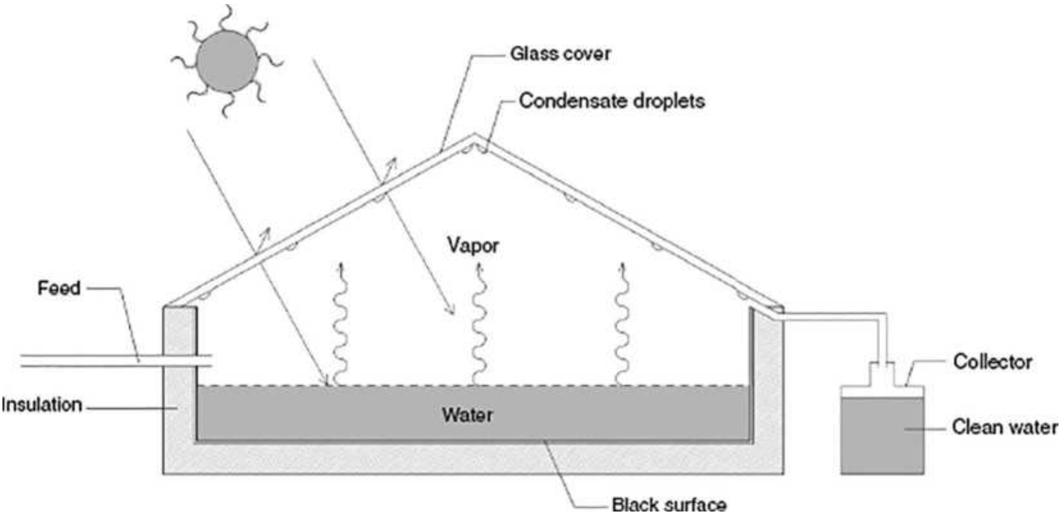


FIGURE 4 Schematic diagrams of a solar still.

Solar Distillation

Solar distillation is a process that utilizes solar energy to purify water through evaporation and condensation processes. The process is also referred to as water desalination when solar energy is used to purify water from saline water. Solar water distillation is a solar technology with a very long history. Installations were built over 2,000 years ago, although they were to produce salt rather than drinking water. Documented use of solar stills (the distillation unit) began in the 16th century. An early large-scale solar still was built in 1872 that spread over an area of 4,600 m^2 capable of producing 23,000 L of drinking water for a mining community in Chile. Mass production occurred for the first time during the World War II when 200,000 inflatable plastic stills were made to be kept in life-crafts for the US Navy.[6] In addition to their use in obtaining drinking water, solar stills are also suitable for the production of distilled water if there is appreciable demand for it in industry, laboratories, and medical facilities or to fill lead acid batteries.

Solar stills come in different designs; however, the main features of operation are the same for all of them. In its simple form, water can be placed in an airtight basin that has a sloped transparent cover normally made of glass or plastics, although glass is preferred for its high transparency. The basin is coated with a black lining to maximize absorption of solar radiation. The incident solar radiation is transmitted through the glass cover and is absorbed as heat by the black surface in contact with the water to be distilled. The water is thus heated and gives off water vapor. The vapor condenses on the glass cover, which is at a lower temperature because it is in contact with the ambient air, and runs down into a tray where it is fed to a storage tank, as shown in Figure 4. The economic viability of solar stills is determined to a critical degree by the design, the construction, the materials employed, and the local market conditions.

Solar Cooking

A solar cooker or solar oven harnesses solar energy to cook food. The solar cooker was first developed by a Swiss scientist Horace de Saussure in 1767.[7] Solar cookers are now being used in many countries across the world, especially in remote areas of poor countries. Solar cookers accomplish free cooking with environment friendliness as they only capitalize solar energy. Solar cooking can be very helpful in reducing the deforestation and pollution that originate from consumption of wood, and animal and agricultural residues for cooking in remote areas that lack access to electricity and gas. Solar cookers are capable of performing various types of cooking phenomena, i.e., frying, baking, and boiling. The maximum achievable temperature depends on the intensity of the available solar radiation and the design and size of the solar cooker. Solar cookers come in a wide range of designs, which can be categorized under the following three major types.

Solar Box Cookers

A solar box cooker consists of an insulated box with a transparent top and a reflective lid. It is designed to capture solar radiation and make use of the greenhouse effect to cause heat to accumulate inside. The top is removable to allow food pots to be placed inside. Temperatures in a typical box cooker can reach above 200°C, but the temperatures achieved obviously depend on the size and design parameters of the cooker and the location of use.

Solar Panel Cookers

The solar panel cooker is the simplest solar cooker, and it consists of multiple simple reflectors arranged to focus solar radiation onto a covered black pot enclosed in a clear heat-resistant plastic bag or other transparent enclosure, such as glass bowel.

Solar Parabolic Cookers

Parabolic solar cookers, also called concentrated cookers, consist of a concave disk that focuses the light onto the bottom of a pot that is arranged at the focal length of the disk, as shown in Figure 5. These are the most efficient types of solar cookers.

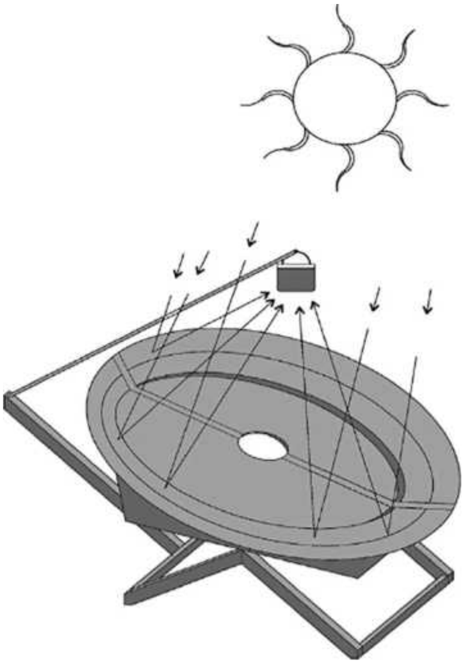


FIGURE 5 Parabolic type of solar cooker.

TABLE 1 Overview of Processes for Thermally Powered Cooling and Air Conditioning

Solar Thermal Cooling System Design	Adsorption Refrigeration	Absorption Refrigeration	Desiccant Air Conditioning
Solar collector	Vacuum tube collector, flat plate collector	Vacuum tube collector	Flat plate collector, solar air collector
Coolant circulation process	Closed refrigerant circulation systems	Closed refrigerant circulation systems	Open refrigerant circulation systems (in contact with the atmosphere)
Process basic principle	Cold water production	Cold water production	Air dehumidification and evaporative cooling
Sorbent type	Solid	Liquid	Solid
Refrigerant/sorbent	Water–silica gel ammonia–salt	Water–water–lithium bromide, ammonia–water	Water–silica gel water–lithium chloride–cellulose
Typical operating temp.	60°C–95°C	80°C–110°C (one step) 130°C–160°C (two step)	45°C–95°C

Solar Cooling/Air Conditioning

Solar thermal energy can be used for cooling and dehumidification. Collectors play a critical role in extracting the energy from solar radiation to operate the cooling device. The collectors used in solar thermal cooling could be of various types, such as low-temperature flat plates and high-temperature evacuated tubes and concentrators. The basic principle behind solar thermal cooling is the thermochemical process of sorption—a liquid or gaseous substance is either attached to a solid, porous material (adsorption) or taken in by a liquid or solid material (absorption). The heat-transfer fluid is heated in the solar collectors to a temperature well above ambient and used to power a cooling device—a type of heat-actuated pump. The heat-transfer fluid may be air, water, or another fluid; it can also be stored in a hot state for use during times of no sunshine. Heat extracted by the cooling device from the conditioned space and from the solar energy source is rejected to the environment using ambient air or water from a cooling tower.[8]

The solar thermal cooling process can be broadly classified under open cycle systems and closed cycle systems. Open cycle systems are those in which the refrigerant is in direct contact with atmosphere and is discarded from the system after providing the cooling effect and new refrigerant is supplied in an open-ended loop. In closed systems, on the other hand, the refrigerant is not in direct contact with the atmospheric air. Open and closed cycle systems can further be distinguished according to the type of sorbent used, which can be in a liquid or a solid form. The three main designs of solar thermal cooling technologies that have gained the most attraction include solar adsorption, solar absorption, and solar desiccant. The key features of these designs are provided in Table 1.[9]

Solar Thermal Power Generation

Solar thermal power generation systems start with capturing heat from solar radiation. Direct solar radiation can be concentrated and collected by a range of concentrating solar power technologies to provide medium- to high-temperature heat. This heat then operates a conventional power cycle—for example, through a steam turbine or a Stirling engine to generate electricity. Solar thermal power plants can be designed for solar-only or hybrid operation, where some fossil fuel is used in case of lower radiation intensity to secure reliable peak-load supply. Five distinct solar thermal power generation concepts are available:

- Solar pond
- Solar chimney

TABLE 2 Characteristics of Typical Concentrated Solar Collectors

Solar Collector Technology	Typical Operating Temperature (°C)	Concentration Ratio	Tracking	Maximum Conversion Efficiency (Carnot) (%)
Solar Fresnel reflector technology	260–400	8–80	One-axis	56
Parabolic trough collectors	260–400	8–80	One-axis	56
Heliostat field + central receiver	500–800	600–1000	Two-axis	73
Paraboloidal dish concentrators	500–1200	800–8000	Two-axis	80

- Solar parabolic trough
- Solar central receiver or solar tower
- Solar parabolic dish.

Solar pond and solar chimney are nonconcentrated types of technology. In this section, the three concentrated types of solar thermal technologies—solar parabolic trough, solar central receiver, and solar parabolic dish—are discussed, as they have received the greater degree of attention over the years due to their favorable technical and commercial characteristics. These technologies can be used to generate electricity for a variety of applications, ranging from remote power systems as small as a few kilowatts (kW) up to grid-connected applications of 200–350 megawatts (MW) or more.[10]

Solar thermal power generation systems have three essential elements needed to produce electricity: a concentrator (to collect and focus solar radiation), a receiver (to convert concentrated solar radiation into heat), and an engine cycle (to generate electricity). Some systems also involve a transport or storage system. Solar collectors have a crucial role to play in the whole system and can be mainly classified into two types: concentrating and nonconcentrating. They are further categorized on the basis of their concentrator optical properties and the operating temperature that can be obtained at the receiver. Most of the techniques for generating electricity from heat need high temperatures to achieve reasonable efficiencies. Concentrating systems are hence used to produce higher temperatures. Table 2 shows the operational characteristics of concentrated collectors.[11]

Parabolic Trough

The parabolic trough systems consist of large curved mirrors or troughs that concentrate sunlight by a factor of 80 or more onto thermally efficient receiver tubes placed in the trough's focal line, as shown in Figure 6a. A thermal transfer fluid, such as synthetic thermal oil, is circulated in the tubes at focal length. Heated to approximately 400°C by the concentrated sun's rays, this oil is then pumped through a series of heat exchangers to produce superheated steam.[12] The steam is converted to electrical energy in a conventional steam turbine generator, which can either be part of a conventional steam cycle or integrated into a combined steam and gas turbine cycle, as shown in Figure 7. Parabolic trough power plants are the only type of solar thermal power plant technology with existing commercial operating systems.

It is also possible to produce superheated steam directly using solar collectors. This makes the thermal oil unnecessary and also reduces costs because the relatively expensive thermo oil and the heat exchangers are no longer needed. However, direct solar steam generation is still in the prototype stage.

Central Receiver or Solar Tower

In solar thermal tower power plants, hundreds or even thousands of heliostats (large individually tracking mirrors) are used to concentrate sunlight onto a central receiver mounted at the top of a tower, as indicated in Figure 6b. A heat-transfer medium in this central receiver absorbs the highly concentrated radiation reflected by the heliostats and converts it into thermal energy to be used for the subsequent

(a) Parabolic trough

(b) Central receiver or solar tower

(c) Parabolic dish (Stirling engine)

FIGURE 6 Solar thermal power generation technologies.

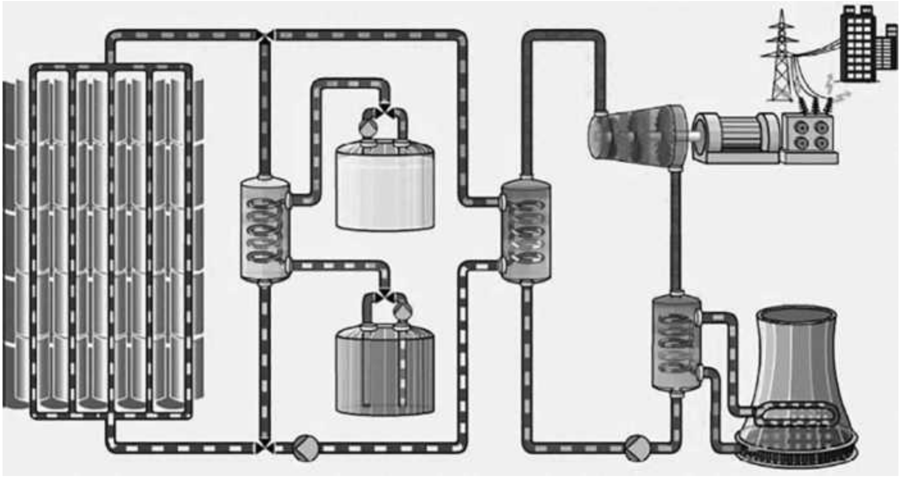


FIGURE 7 Schematic of solar parabolic trough power plant.

generation of superheated steam for turbine operation. To date, the heat-transfer media demonstrated include water or steam, molten salts, liquid sodium, and air. If pressurized gas or air is used at very high temperatures of about 1000°C or more as the heat-transfer medium, it can even be used to directly replace natural gas in a gas turbine, thus making use of the excellent cycle of modern gas and steam combined cycles.

Parabolic Dish

A parabolic dish system uses a parabolic concave mirror to concentrate sunlight onto a receiver located at the focal point of the mirror, as highlighted in Figure 6c. The concentrated beam radiation is absorbed into the receiver to heat a fluid or gas (air) to approximately 750°C. This fluid or gas is then used to generate electricity in a small piston, Stirling engine, or a microturbine attached to the receiver. These systems stand alone, and they are normally used to generate electricity in the kilowatts range.[13]

Solar Thermal Technologies—Market Growth and Trends

Solar thermal technologies, like other renewables, are experiencing a rapid growth. Between 2007 and 2017, for example, the global installed capacity of solar water heaters increased from 145 GW_{th} to 472 GW_{th}. The year 2017 saw an addition of 35 GW_{th}. China alone accounts for over 71% of the world's total solar water heating installed capacity. Other leading countries include Turkey, India, Brazil, the USA, and Germany. In terms of collector type, while glazed flat plate collectors are the preferred choice around the world, in China and India glazed evacuated tube collectors account for over two-thirds of the market. Of the new installations in 2017, vacuum tube collectors had a share of 73%. Flat plate collectors and unglazed collectors accounted for respective shares of 23% and with 4%.[14]

Over the past decade, concentrated solar power (CSP) has experienced a rapid growth —between 2007 and 2017 the global installed capacity increased from 0.45 to 4.9 GW, with around 2 GW of projects being under construction. Spain has emerged as the global leader over this period installing 2.3 GW of CSP projects. Over 80% of the world's total installed capacity is in Spain and the USA, the two countries having respective figures of 2.3 and 1.7 GW. In terms of the ongoing developments, China is having 20 projects—including parabolic trough, tower, and Fresnel facilities—at various stages of construction, with a combined capacity of 1 GW. The Middle East and North Africa (MENA) region, being rich in the direct solar radiation, has shown a significant interest in the technology. United Arab Emirates (UAE), having already installed 100 MW of CSP, has awarded tender for 700 MW (600 MW parabolic basin complex and a 100 MW solar tower) of projects. Morocco is having two projects with total capacity of 350 MW at advanced stages of completion. Projects are also underway in Kuwait, Israel, and Saudi Arabia with respective capacities of 50, 121, and 93 MW. Projects in Chile and Australia of respective capacity 110 and 100 MW are also under construction. Driven by factors like competition and technology cost reduction, the year 2017 has also witnessed record-low tariffs: AUD78/MWh in Australia, USD50/MWh in Chile, and USD73/MWh in UAE.[14]

Solar thermal technologies are regarded to be playing an important role in the future energy scenarios. A scenario of what could be achieved by the year 2025 was prepared by Greenpeace International, the European Solar Thermal Industry Association, and International Energy Agency (IEA) SolarPACES projects. It suggested that by 2025, the total installed capacity of solar thermal power around the world will reach over 36 GW. It is also projected that by 2040, more than 5% of the world's electricity demand may be satisfied by solar thermal power.[12]

Conclusions

Solar thermal technologies operate by converting solar radiation into heat, which can be either directly utilized in various applications such as solar space heating, solar water heating, and solar air conditioning, or can be transformed into electricity to serve any purpose similar to conventional electricity. The key element in all solar thermal technologies is the collector, whose function is to gather the heat of solar radiation. Collectors normally come in three different types: flat plate, evacuated tube, and concentrated, and they operate in a wide range of temperatures, i.e., from less than 50°C to more than 1200°C. Solar thermal technologies normally operate in passive or active modes. Different types of solar thermal technologies are gaining huge attention across the world depending upon their technical and economic viability. Solar thermal power generation is also expected to grow at a healthy rate in coming years, as it is projected that by 2040, more than 5% of the world's electricity demand could be satisfied by solar thermal power.

References

1. Asif, M. *Energy Crisis in Pakistan: Origins, Challenges and Sustainable Solutions*; Karachi: Oxford University Press, 2011.
2. Muneer, T. Solar energy. KEMPS Engineering Year Book; John H Stephens (Ed.) Miller Freeman: New York, 2000.
3. Duffie, J.; Beckman, W. Solar Engineering of Thermal Processes, 2nd Ed.; Wiley: New York, 1991.
4. Akbarzadeh, A.; Andrews, J.; Golding, P. Solar pond technology: A review and future directions. Advances in Solar Energy; Yogi Goswami (Ed.) Earthscan: London, 2005.
5. Whitfield, D.E. Solar drying. *International Conference on Solar Cooking*, South Africa, November 26–29, 2000.
6. Solar distillation. Technical Brief, Intermediate Technology Development Group. http://www.itdg.org/?idZtechnical_briefs (accessed on March 2006).
7. History of Solar Cooking, https://www.solarcooker-at-cantinawest.com/solarcooking-history.html (accessed on 15 May, 2000).
8. Grossman, G. Solar cooling, dehumidification, and airconditioning, solar thermal power generation. In Encyclopedia of Energy, Cutler Cleveland (Ed.) Vol. 5; Elsevier: Amsterdam, 2004.
9. Asif, M., and Muneer, T. Thermal Energy: Solar Technologies, Encyclopaedia of Environmental Management; Taylor & Francis: New York, 2013; Vol. IV, 2526-2535.
10. Concentrating Solar Power: Energy from Mirrors, report Produced by NREL 2001. Available at: http://www.nrel.gov/docs/fy01osti/28751.pdf.

11. Luzzi, A.; Lovegrove, K. Solar thermal power generation. In Encyclopedia of Energy, Vol. 5; Cutler Cleveland (Ed.) Elsevier: Amsterdam, 2004.
12. Concentrated solar thermal power—now! Report by European Solar Thermal Industry Association and Greenpeace, 2005 http://www.greenpeace.org/raw/content/international/press/reports/Concentrated-Solar-Thermal-Power.pdf.
13. Quaschning, V. Solar thermal power plants. Renewable Energy World, June 2003.
14. Renewables 2018 Global Status Report, Renewable Energy Policy Network for the 21st Century, Paris.

VI

PRO: Basic Environmental Processes

43

Acid Rain

Umesh Kulshrestha

Introduction

Any form of precipitation (rain, snow, or hail) having high acidity is known as acid rain. The term "acid rain" was first used by Robert Angus Smith in his book *Air and Rain: The Beginnings of a Chemical Climatology,* published in England in 1872.[1] He had chemically analyzed the rainwater near Manchester and observed three types of rain composition—"that with carbonate of ammonia in the fields of distance, that with sulfate of ammonia in the suburbs and that with sulfuric acid or acid sulfate, in the town."

In broader perspectives, acid rain refers to wet deposition (rain, snow, hail, cloud water, fog, dew, or sleet) and dry deposition (absorption of SO_2, NO_x, other acidic gases and particles) of acidic compounds. High acidity is generally caused by higher levels of sulfuric and nitric acids. These acids are contributed by their precursor gases (SO_2 and NO_2), which are emitted by natural as well as anthropogenic sources. Natural sources include volcanoes, vegetation decay, various biological processes on the land, and oceans, while major anthropogenic sources of these gases are fossil fuel combustion and smelting of metal ores. In regions of North America, the rates of anthropogenic emissions of these two gases have gone up to 100 times more than the natural rates, adding to higher atmospheric acidity.[2] Acid rain has caused severe damage in Europe, North America, parts of China and Japan through acidification of lakes and other water bodies, decline of forests, acidification of soils, and corrosion of building materials.

How Acid Rain Happens

Pure water (H_2O) is neutral in nature, having a pH value of 7. Any aqueous solution having a pH higher than 7 is said to be alkaline, while one having a pH lower than 7 is known as an acidic solution. Rainwater in remote and unpolluted atmospheres has a slightly acidic pH of around 5.6 due to the presence of carbonic acid formed at equilibrium due to dissolution of atmospheric carbon dioxide in cloud water:[3]

$$CO_2 + H_2O \leftrightarrow H_2CO_3 \tag{1}$$

In water, carbonic acid is dissociated, forming bicarbonate ion:

$$H_2CO_3 \leftrightarrow HCO_3 - + H^+ \tag{2}$$

Rainwater pH is further depressed to about 5.2 in unpolluted regions by organic acids. However, anthropogenic acid rain arises due to oxidation of SO_2 and NO_2 in the atmosphere to form sulfuric and nitric acids. There are a number of probable reactions for the oxidation of these gases involving both homogeneous and heterogeneous oxidation.[4] The gas-phase oxidation of these gases is initiated by reaction with hydroxyl radicals:

$$SO_2 + OH \rightarrow HOSO_2 \tag{3}$$

$$HOSO_2 + O_2 \rightarrow HO_2 + SO_3 \tag{4}$$

$$SO_3 + H_2O \rightarrow H_2SO_4 \tag{5}$$

Homogeneous aqueous-phase oxidation of SO_2 takes place by its dissolution and dissociation in water, forming equilibrium similar to CO_2:

$$SO_2 + H_2O \leftrightarrow SO_2H_2O \tag{6}$$

$$SO_2.H_2O \leftrightarrow HSO_3 - + H^+ \tag{7}$$

$$HSO_3 - \leftrightarrow SO_3{}^{2-} + H^+ \tag{8}$$

Gas-phase oxidation of NO_2 is faster than SO_2 by one order of magnitude:

$$NO_2 + OH \rightarrow HNO_3 \tag{9}$$

In addition, significant formation of nitric acid takes place through ozone and NO_3 radical reactions.

During daytime, NO_3 radical is formed as follows:

$$NO_2 + O_3 \rightarrow NO_3 + O_2 \tag{10}$$

NO_3 radical so formed reacts with NO_2 at nighttime, finally resulting in the formation of HNO_3:

$$NO_3 + NO_2 \leftrightarrow N_2O_5 \tag{11}$$

$$N_2O_5 + H_2O \rightarrow 2HNO_3 \tag{12}$$

At ambient levels of NO, its aqueous-phase oxidation is very slow due to its low solubility in water and also the dependence on NO_2 concentrations. It can be faster at higher NO_2 levels. However, the reaction follows the path

$$2NO_2 + H_2O \leftrightarrow 2H^+ + NO_3 - + NO_2 - \qquad (13)$$

Heterogeneous oxidation of SO_2 and NO_2 involves gas- particle reactions. In the liquid phase, SO_2 is rapidly converted into sulfate by H_2O_2. SO_2 is also converted into sulfate on freshly emitted soot particles, but subsequently, the rate of oxidation is retarded due to saturation of soot particle surface. Preferable oxidation of SO_2 onto soil dust particles is reported in dusty regions where formation of calcium sulfate takes place instead of free sulfuric acid. NO_2 is also oxidized onto particles—for example, it reacts with NaCl of sea salt, forming $NaNO_3$ on the surface. However, over time, the surface is saturated, and the rate of oxidation becomes lower.

History of Acid Rain

The major cause of acid rain is the increased combustion of fossil fuels, which has been practiced at larger scale after the industrial revolution. The presence of sulfur compounds in the air of Sweden and England was realized in the 18th century. In fact, Robert Boyle, in 1692, mentioned in his book *A General History of the Air* the "nitrous and salino-sulphureous spirits" in the air.[4] The term "acid rain" was first used in 1872 by Robert Smith, who discovered acid rain in 1852 in the area surrounding Manchester. He referred to this term in a treatise on the chemistry of rain published in England in 1872. Robert Smith mentioned various factors such as coal combustion and the amount of rain affecting the precipitation. Unfortunately, this wonderful publication was overlooked until it was revisited and critiqued by Gorham in 1981.[1]

Acid rain attracted attention of the scientific community and society when Odén[6,7] reported that large-scale acidification of surface waters in Sweden could be attributed to pollution from the United Kingdom and central Europe. The worst-hit areas of acid rain were Scandinavia and Central and Southern Germany. In Europe, the rain pH was observed to be as low as 3.97 in Germany. Drastic loss of fish population was seen in lakes of Sweden and parts of southwest Norway. European data show that most of the acidity was intensified during 1955–1970, with a sudden increase in the mid-1960s. In parts of Germany and other European countries, forest damage and loss of needles from pine and spruce trees were noticed due to acid rain. Acid smog killed almost 4000 people in London in 1952.

Sources of Acidity

As mentioned earlier, in natural conditions, atmospheric CO_2 when dissolved in water forms carbonic acid (H_2CO_3), which brings down the pH of water. Other gases such as SO_2 and NO_2 also form acids, viz., sulfuric (H_2SO_4) and nitric acid (HNO3), respectively. The main cause of acid rain is excess contribution of H_2SO_4 and HNO_3 in precipitation due to anthropogenic sources, especially through combustion of coal and petroleum. Sulfur is present in significant amounts in fossil fuel (coal and petroleum), which is the major source of SO_2. Oxidation of SO_2 is accelerated by higher concentrations of H_2O_2 and O_3 found in polluted air. Martin and Barber[8] noticed that acidity of precipitation at several sites in England was the highest during spring, when O_3 concentrations were higher. During past century, huge consumption of fossil fuel in North America and Europe resulted in high SO_2 emissions.[9,10]

Natural Acidity Contributed by Organic Acids

Apart from sulfuric and nitric acids, acidity in rainwater is also contributed by organic acids. Formic acid (HCOOH) and acetic acid (CH3COOH) are the major species reported in rainwater, contributing around two-thirds of the total acidity at remote sites.[11] Generally, formic acid is found to dominate over acetic acid. A relatively higher contribution of organic acids is observed at tropical sites than at temperate ones. These organic acids are produced in gas phase by the oxidation of isoprene and terpenes emitted by the vegetation,[12,13] which are then scavenged by the rain. These are also formed through aqueous-phase oxidation of aldehydes. Sometimes, these acids are emitted by soils.[14] It is to be noted that organic acids may be important for pH in cloud and rainwater in some areas, but their contribution to acidification of soils and surface waters is small because these are quickly consumed by microorganisms.

Spread and Monitoring of Acid Rain

Considering the degree of damage caused by acid rain, several efforts are made by European countries to monitor and control it. The European Air Chemistry Network was started in the early 1950s by Stockholm University in collaboration with the Swedish University of Agricultural Sciences. Both institutes served as centers for the network for the chemical analysis of samples. Originally, the purpose of this network was to study the depositions of plant nutrients to forest and agriculture systems. Under this network, continuous data related to chemical composition of precipitation have been available since 1955.[15] Later on, this network became part of the Swedish National Monitoring Programme. A Norwegian program called SNSF, "Acid Precipitation: Effects on Forests and Fish" was run up to 1980.[16] Immediately after the Stockholm conference in 1972, the European Organization for Economic Cooperation and Development (OECD), in 1978, established a network to monitor long-distance transport of pollutants and the impacts of European countries on their neighbors, known as the Cooperative Program for Monitoring and Evaluation of Long-Range Transmission of Air Pollutants in Europe (EMEP). Further, in 1983, the Convention on Long-Range Transboundary Air Pollution (CLRTAP) was signed by more than 30 countries, including the United States, Canada, and the European Union, to deal with transboundary air pollution.

The discovery of acid rain in Europe attracted attention of scientific community in North America too. Odén's study was followed up by the United Nations (UN).[17] This also led to the first international conference on acid rain in Columbus, Ohio, United States, in 1975. Later on, under the Acid Precipitation Act of 1980, U.S. Congress formed a national network called National Acid Precipitation Assessment Program (NAPAP), which supported the expansion of National Atmospheric Deposition Program (NADP) to monitor the trends in long-term precipitation chemistry and deposition. Further, the NADP was changed to the NADP National Trends Network, which has around 250 monitoring sites.

Similar to Europe and United States, Canada also experienced acid rains. Environment Canada has developed its program called the Canadian Air and Precipitation Monitoring Network (CAPMoN) to monitor the regional patterns and trends of atmospheric pollutants. including acid rain, smog, particulate matter, etc., Canada started the Canadian Network for Sampling Precipitation in 1978, which was renamed as the Air and Precipitation Network (APN). CAP- MoN is the new name of APN (changed in 1983).

Later on, the spread of acid rain was also noticed in East Asia. After successful implementation of CLRTAP in Europe, the UN Conference on Environmental Development adopted to continue and share the experience gained from acid rain programs in Europe and North America and established the Acid Deposition Monitoring Network in East Asia (EANET) in 1993. This network includes Japan, Russia, China, the Republic of Korea, Mongolia, Thailand, Singapore, Cambodia, Lao People's Democratic Republic, Myanmar, Vietnam, the Philippines, Malaysia, and Indonesia.

Measurements through the long-term acid rain network have not been carried out extensively in other parts of the world such as India, Africa, and Latin America. However, programs such as the Composition of Atmospheric Aerosols and Precipitation in India and Nepal and the Composition of Asian Deposition (CAD), as part of the Regional Air Pollution in Developing Countries (RAPIDC) program funded by the Swedish International Development Cooperation Agency), were very effective in providing a summarized picture of the acid rain scenario in the Indian region.[18,19] The RAPIDC program was coordinated by the Stockholm Environment Institute, which facilitated international cooperation on air pollution issues to develop relevant knowledge to support decision making in Asia and Africa. The CAD program was a part of the International Global Atmospheric Chemistry/Deposition of Biogeochemically Important Trace Species (IGAC/DEBITS) activities of the International Geosphere-Biosphere Programme, which focused on good-quality measurements at rural sites in Asia to produce high-quality data so as to understand the Asian wet deposition scenario. In the African region, the IGAC/DEBITS-Africa program has its network of 10 stations for the measurement of wet and dry depositions at selected sites.

In Australia, acid rain studies have been carried out under the Commonwealth Scientific and Industrial Research Organisation (CSIRO) network of sites.[20] Globally, the 1989 initiative of the World Meteorological Organization, under the Global Atmospheric Watch (GAW), is carrying out precipitation measurements at around 80 stations in different countries. Earlier, GAW used to be known as the Background Air Pollution Monitoring Network (BAPMoN).

Regional Acidity of Precipitation

United States

As described by Gibson,[10] northeastern United States and southeastern Canada were the most affected areas by acid rain. Likens[21] was the first who evaluated the 1955–1956 and 1972–1973 data and found a significant increase in acidity in northeastern United States and southeastern Canada during the two decades. He also noticed a significant increase in the spread of acid rain in the areas of southeastern and Midwestern United States.[22]

Canada

Long-term measurements of acid deposition in Canada showed that other than local sources, acidity was also contributed by the long-range transport of oxides of sulfur and nitrogen from sources located southerly in the United States. According to Environment Canada, more than half of the acid deposition in eastern Canada is originated from the United States. Studies by the APN showed higher acidity in southern Ontario, having a pH of around 4.2.[23] Regionally representative sites Long Point and Chalk River experienced that, most of the time, wind parcels came from southerly source areas. These sites are the receptor sites to the major sources in Ontario and the lower Great Lakes region. Estimates of the year 1995 showed that 3.5–4.2 Tg per year of SO_2 was transported from the United States to Canada. Acid deposition in Canada can be reduced by the joint measures of the United States and Canada. Collaborative efforts in this direction are already in progress.

Europe

Areas affected by acid rain in Europe include northern and western Europe, southeast England, Germany, the Netherlands, and parts of Denmark. Table 1 shows a drastic decrease of precipitation pH in western Europe during the 1950s and 1970s.[24]

TABLE 1 Change in pH of Precipitation in Western Europe during the 1950s and 1970s

Country	pH in 1950s	pH in 1970s
Southern Norway	5.0–5.5	4.7
Northern Sweden	5.5–6.0	4.3
Southern Sweden	5.5–6.0	4.3
Southeast England	4.5–5.0	4.2

Source: Environmental Resources Ltd. Pearce.[24]

OECD 1977 estimates indicate that anthropogenic emissions of sulfur in Europe increased by 50% during 1955–1970. However, later on (1972–1982), many European countries, viz., the United Kingdom, West Germany, the Netherlands, Sweden, Norway, and Denmark reduced their sulfur emissions. These reduction measures improved the situation in Europe.[15]

Asia

In the Asian region, much of the acid rain problem prevails in East Asia, covering China, Japan, North Korea, and Thailand. Among these, China is the biggest polluter. According to estimates, China's sulfur emissions will triple between 1990 and 2020.[25] After the United States and Europe, China is the biggest consumer of fossil fuel. Rapid increase in SO_2 and NO_x emissions from 2000 onward is a major reason for the spread of acid rain in China.[26] The area most affected by acid rain in China is south of the Yangtze River, where average pH is recorded to be less than 4.5.

Acidity levels in precipitation in Japan show seasonality. During the summer season, most of the acidity is observed to be due to local sources of sulfur oxides, whereas during the winter season, increased level of acidity is due to long-range transport from the Asian continent, which results in higher acidity of precipitation at the sites in western Japan.[27] Similarly, Thailand also experiences acid rain. Around 50% of rain events are reported acidic due to high concentration of sulfate and nitrate. Of these oxides, 70%–80% are contributed by Thai sources.[28]

In the Indian subcontinent, precipitation is reported to have relatively higher pH (>5.6).[19,29] The pH of rainwater at some of the continental sites in India is as high as 8.3 (Table 2) due to interference of soil dust (rich in calcium carbonate) suspended in the atmosphere. Abundance of soil dust in air is a

TABLE 2 Average pH of Rainwater at Various Sites in India

Site	Nature of Site	pH	Reference
Calcutta	Urban	6.8	Das[31]
Nainital	High altitude	6.2	Hegde et al.[32]
Iqbalpur	Rural	7.1	Jain et al.[33]
Mumbai (Colaba)	Urban	5.9	Khemani et al.[34]
Darjeeling	High altitude semiurban	6.4	Kulshrestha[35]
Haflong	High altitude rural	7.3	Kulshrestha[35]
Delhi	Urban	5.7	Kulshrestha et al.[36]
Hyderabad	Urban	6.4	Kulshrestha et al.[37]
Jorhat	Rural	5.8	Kulshrestha et al.[38]
Hudegadde	Reserve forest	6.0	Kulshrestha et al.[38]
Agra (Dayalbagh)	Semiurban	7.1	Kumar et al.[39]
Malikadevi	Remote	6.4	Mahadevan et al.[40]
Allahabad[a]	Urban	7.1	Mukhopadhyay et al.[41]
Jodhpur[a]	Rural	8.3	Mukhopadhyay et al.[41]
Kodaikanal[a]	Rural	6.1	Mukhopadhyay et al.[41]
Mohanbari[a]	Rural	6.4	Mukhopadhyay et al.[41]
Nagpur[a]	Rural	6.3	Mukhopadhyay et al.[41]
Srinagar[a]	Rural	7.0	Mukhopadhyay et al.[41]
Pune	Urban	6.3	Pillai et al.[42]
Sinhagad	High altitude rural	6.2	Pillai et al.[42]
Silent Valley	Reserve forest	5.3	Rao et al.[43]
Agra (Tajganj)	Semiurban	7.0	Saxena et al.[44]
Indian Ocean (during Jan–Mar)	Northern and central	Below 5.6	Kulshrestha et al.[45]

[a] BAPMoN sites.

TABLE 3 Average pH and Major Ions (μeq/L) in Precipitation at Banizaumbou during 1994–2005

Parameter	Value
pH	6.05
SO_4^{2-}	9.4
NO^{3-}	11.6
Ca^{2+}	27.3
NH^{4+}	18.1

Source: Galy-Lacaux et al.[47]

common feature of the Indian atmospheric environment. The pH of most soils of India is very high as compared with the pH of soils in acidified regions of the world. Generally, in India, the pH of rainwater is the mirror image of the pH of soil in the region. The acidity generated by the oxidation of gases like SO_2 and NO_x is buffered by soil-derived particles. Acidity of SO_2 is buffered by $CaCO_3$ of soil dust forming calcium sulfate, which is removed by below-cloud scavenging (Figure 1). Due to this, the spread of acid rain at continental sites in India is controlled by the continuous suspension of loose soil during prevailing dry weather conditions. However, a bigger number of hot spots of higher wet deposition of non-sea salt sulfate (nss SO_4) are reported in urban and industrial areas than in rural areas. Several of these larger hot spots lie in the Indo-Gangetic region (Figure 2).

Although rainwater pH higher than 5.6 is more frequently recorded in India, sometimes, occurrence of acid rain (pH <5.6) is also reported. Figure 3 shows the frequency of acid rain reported from various sites in India.[46] Rainwater is noticed to be acidic in India if any of following applies:

1. Rain continues for a long time, washing off soil dust from the atmosphere. A similar situation prevails over the Indian Ocean, where soil dust interference is at a minimum.
2. In the areas where a large part of ground is covered with vegetation.
3. In the areas where soil itself is acidic (northeast, east, and southwest India).
4. Near heavy sources of SO_2 (e.g., thermal power plants).

In the African region, the pH of precipitation is reported to be nearly similar to that in the Indian region. Longterm data (1994–2005) showed that the acidity of precipitation at Banizaumbou, a regional

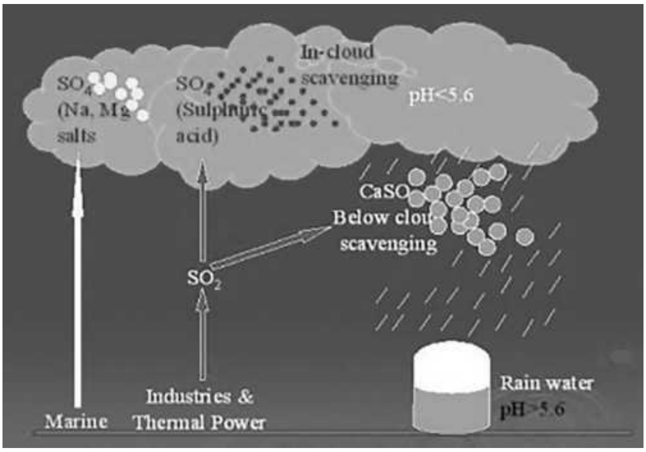


FIGURE 1 Schematic diagram showing alkaline rains by removal of soil dust during below-cloud scavenging process in India.
Source: Kulshrestha.[30]

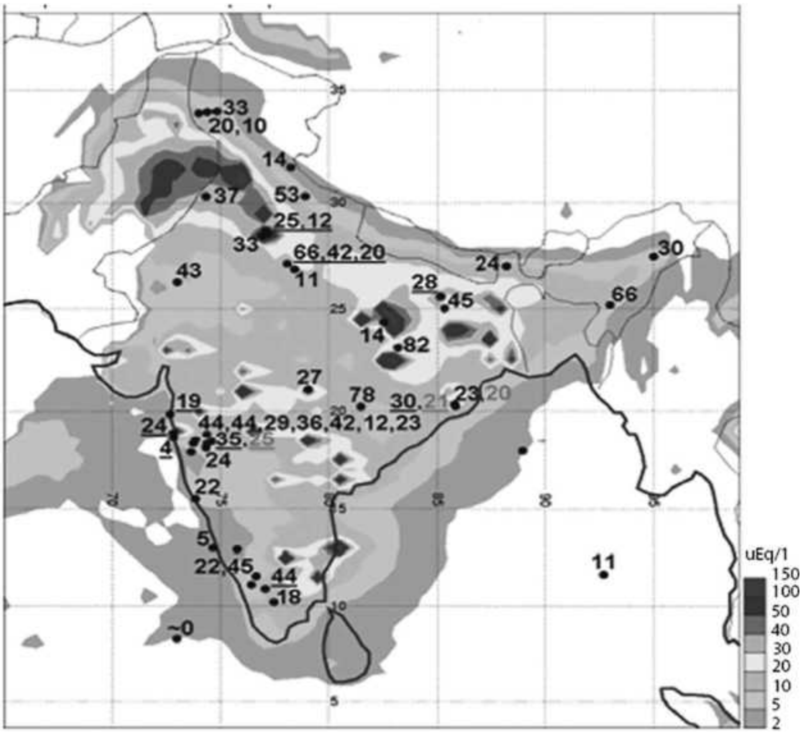


FIGURE 2 Concentration of nss SO_4 in rainwater. Data from measurements at rural and suburban (underlined) sites obtained with bulk (black) and wet (red) collectors only (scaled to year 2000) compared with the concentration field obtained with the Multiscale Atmospheric Transport and Chemistry Model (MATCH) for the year 2000.
Source: Kulshrestha et al.[19]

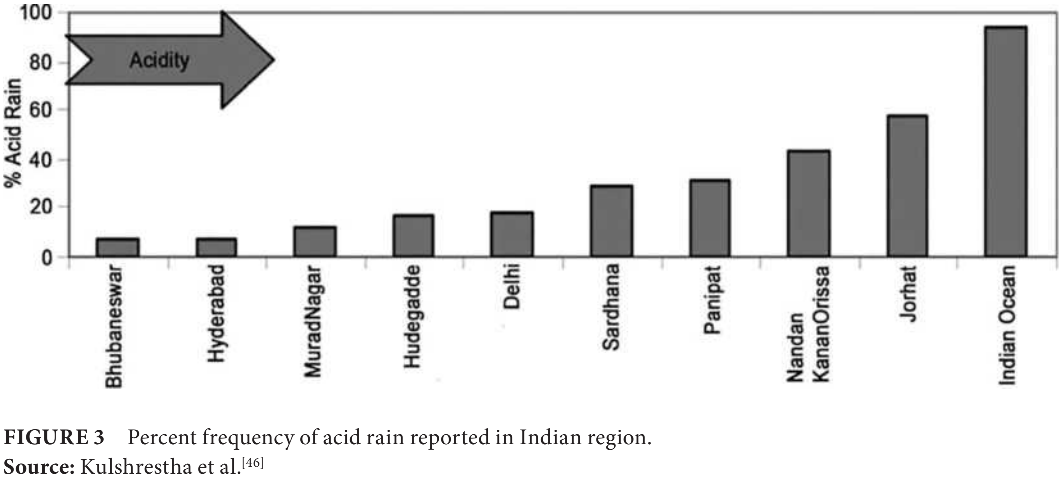


FIGURE 3 Percent frequency of acid rain reported in Indian region.
Source: Kulshrestha et al.[46]

representative site in the semiarid savanna region, has high interference of soil dust, resulting in higher pH.[47] In addition, neutralization by ammonium ion is also partly responsible for elevated pH in Africa.

At a glance, the model based global distribution of pH of rain water is shown in Figure 4.[48] The distribution of pH shows high acidity in Europe, eastern North America, and East Asia. To some extent, acidity is seen in the west coast of South America and Africa also. High acidity in these areas is due to free acidity

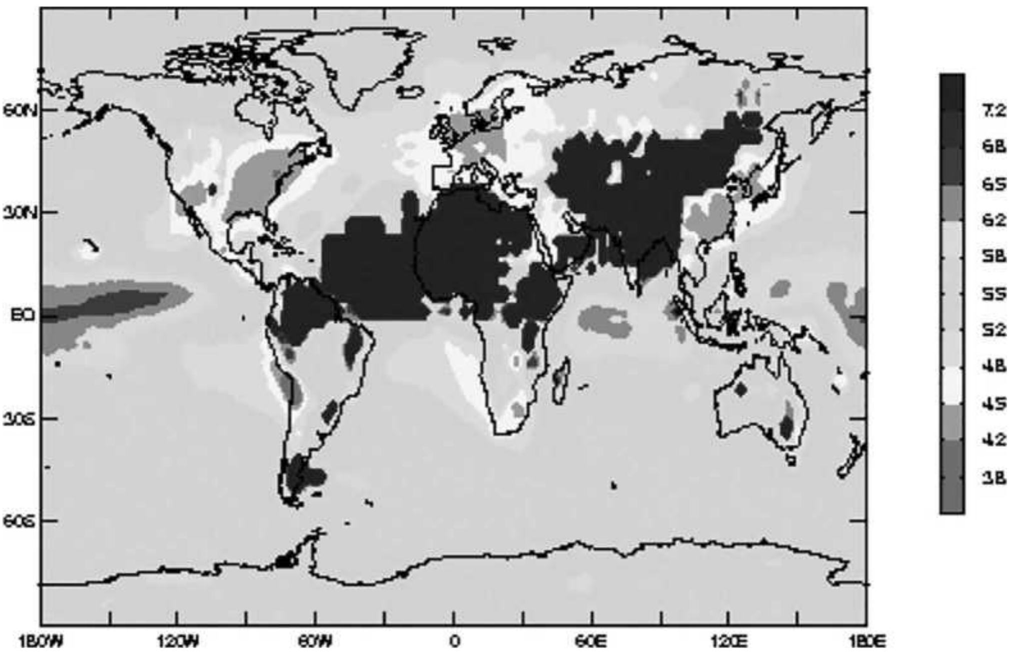


FIGURE 4 Estimated global distribution of pH of precipitation.
Source: Rodhe et al.[48]

contributed by sulfuric acid. However, in areas such as north of South America, northern Africa, South Asia, and part of China, the acidity is lower which is due to neutralization by ammonia and soil dust.

Trends in Acidity

United States and Europe

After 1980, significant reduction in SO_2 emissions has improved the situation of acidity in Europe.[49] Similarly, in Canada and the United States, effective steps of reduction in SO_2 emissions have contributed to improved acidity levels. In the United States, the average reduction in SO_2 emissions from 1980 to 2008 was around 54%, as shown in Figure 5. Overall, the reduction measures have resulted in a decrease in H^+ in precipitation over these regions. Trends of acidity in the United States show a significant improvement after implementation of the 1995 Clean Air Act Amendment, which forced them to reduce SO2 emissions. Figure 6 is an example of trends of acidity of precipitation in North Carolina during 1985–2005,[52] which shows around 50% reduction in H^+ during two decades. In Europe, sulfate concentrations increased by approximately 50% between the 1950s and the late 1960s but have been declining since the mid-1970s. In Sweden and Norway, an average of 20% reduction in SO_4 levels has been achieved since the 1970s,[15] followed by higher reductions in more recent decades.

Asia

China

In the developed countries, efforts have been made to control SO_2 emissions, but in the developing countries, SO_2 emissions continue to be high. In Asia, China is the biggest SO_2 emitter. Total Chinese SO_2 emission increased from 21.7 Tg to 33.2 Tg (53% increase) from 2000 to 2006, showing an annual growth rate of 7.3% per year.[53]

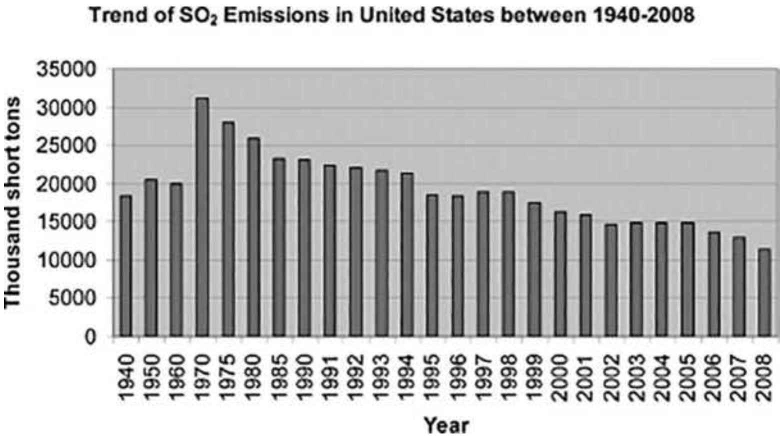


FIGURE 5 Trend of SO_2 Emissions in the United States from 1940 to 2008.
Source: Adapted from the USEPA Web site[50] and Stensland.[51]

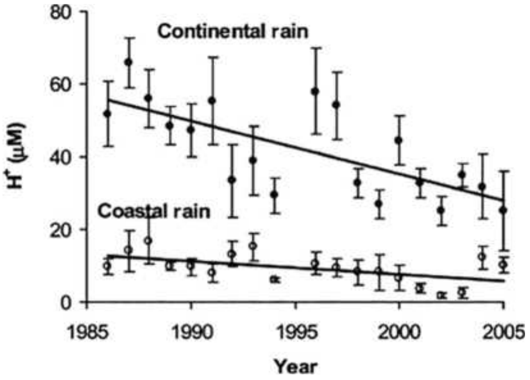


FIGURE 6 Annual volume-weighted average H^+ concentration in Wilmington, North Carolina, continental (filled circles) and coastal (open circles) precipitation from 1985 to 2005.
Source: Willey et al.[52]

Long-term acid precipitation observations show that the temporal and spatial distribution of rain acidity in China has changed remarkably since 2000.[26] Future estimates of acidification potential, using a dynamic soil acidification model, indicate that sensitive soils in south China and Southeast Asia may reach a critical threshold within a few decades.[54] Model-based estimates (Figures 7a and b) indicate higher levels of total depositions of S-SO_x and N-N_r in East Asia.

India

SO_2 emissions in India are relatively less, but the rate of increase is almost doubled from 1985 to 2005. In 2000, Indian SO_2 emission was estimated to be 4.26 Mt.[53] Precipitation studies in India lack long-term measurements. Most of the studies were carried out by individual scientists/groups. A few sites under the GAW network are in operation. In addition, one study from Pune[56] reports that during 1984–2002, there was significant increase in SO_4 and NO_3 concentrations (Figure 8), which resulted in decrease in pH of rainwater from 6.9 in 1982 to 6.5 in 2002. These changes are due to increase in industrial and vehicular activities in the region. Another long-term network study in a rural area of Nandankanan (Orissa state in east India) reported 57% frequency of acid rain events during 1997–1998,[57] which was reduced

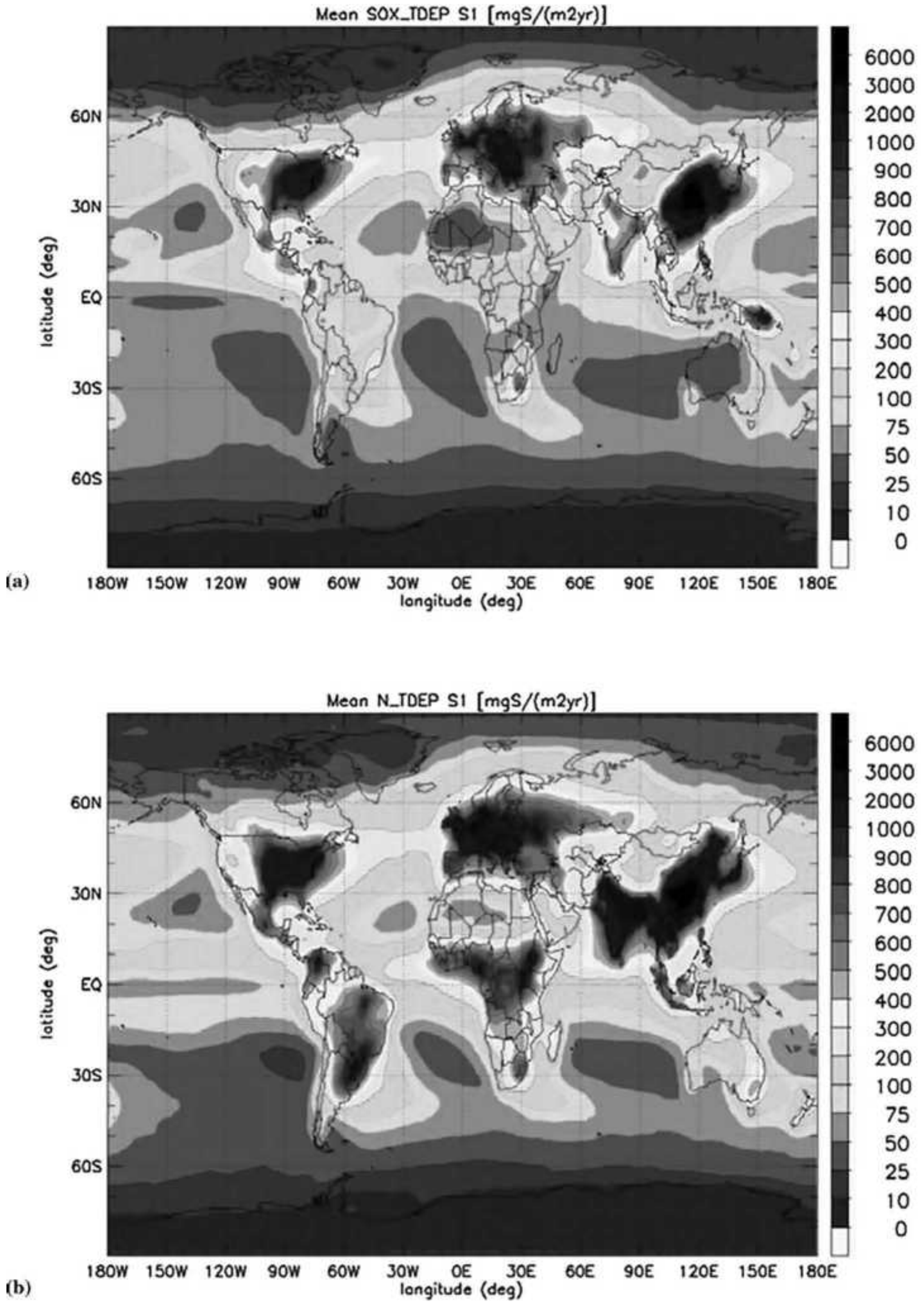


FIGURE 7 Estimated total deposition of (a) S (mg m^{-2} yr^{-1}) of SO_x and (b) N (mg m^{-2} yr^{-1}) of reactive nitrogen for year 2000.
Source: Dentener et al.[55]

to 40% during 2005–2007.[58] SO_4/Na ratios were also reduced drastically from 1.58 in 1997–1998 to 0.519 during 2005–2007. In a review compiled by Kulshrestha and coworkers,[19] it is reported that most of the Indian precipitation measurements lack quality assurance (QA) and quality control (QC) in sampling, storage, and analysis of samples. Hence, this region really needs quality-controlled measurements of

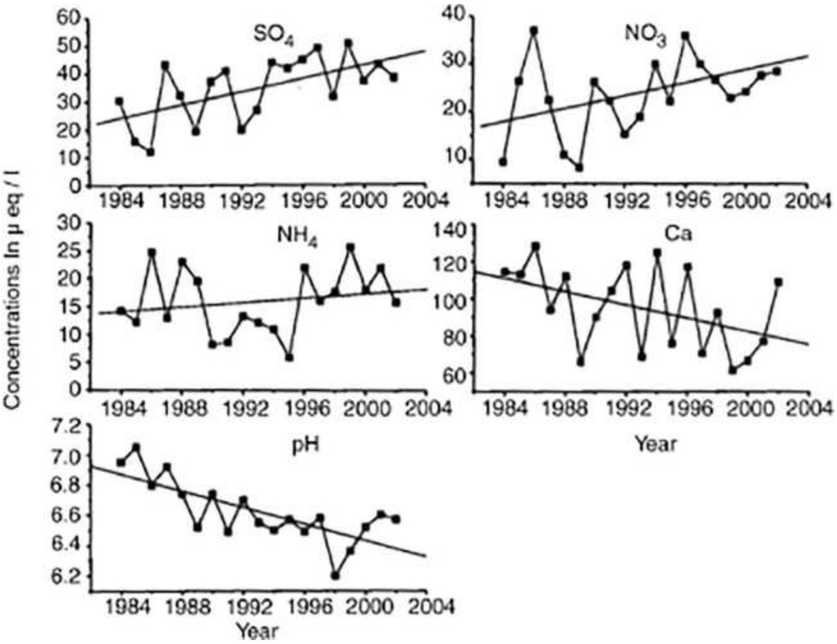


FIGURE 8 Trends of pH, SO_4, NO_3, Ca, and NH_4 in rainwater at Pune during 1984–2000.
Source: Rao et al.[56]

wet and dry depositions at a few selected sites in order to get an idea about the trends of acidity with the growing emissions of oxides of S and N.

Acidification of Oceans

Oceans are the biggest sinks for atmospheric CO_2. Increasing emissions of CO_2 due to anthropogenic sources will lead to ocean acidification through excess dissolution of CO_2 in seawater. Since the industrial revolution, the acidity of the ocean has increased by 30% (from a pH of 8.2 to 8.1). Future projections show that under a business-as-usual scenario, surface ocean pH will be lowered by 0.4 pH units by the end of the century. Acidification of the ocean affects the nitrification process, which further affects marine biota. Apart from CO_2 rise, acid deposition can add to the acidification of oceans. Precipitation having very low pH contributes a significant amount of hydrogen ions in seawater, which in the long term may alter the pH of seawater. Results from Indian Ocean Experiment showed the pH of rainwater to be between 3.8 and 5.6 over the Indian Ocean.[45] The acidic nature of rainwater over the Indian ocean is due to insignificant influence of soil dust and the dominance of anthropogenic sulfate contributed by long-range transport.[59,60] This aspect needs to be investigated in the future in order to protect the marine ecosystem.

Global Sensitivity toward Acidification

Global precipitation acidity and mapping of soil sensitivity to acid deposition suggests three main problematic areas. These are North America, Europe, and southern China (Figure 9), where acid rain control is necessary. Already, in North America and Europe, steps have been taken to reduce SO_2 emissions. Other parts of the world also need to take appropriate steps to reduce sulfur emissions.[62]

FIGURE 9 Global sensitivity toward acidification.
Source: Kuylenstierna et al.[63]

Global Scenario: Future Projections through Modeling

Recently, the Regional Air Pollution Information and Simulation model has been developed by International Institute for Applied Systems Analysis (IIASA) as a tool for the integrated assessment of alternative strategies to reduce acid deposition in Europe and Asia.[63] Dentener and coworkers[55] have attempted simulation of the global future scenario (up to 2030) of deposition of oxides of nitrogen and sulfur by using 23 atmospheric chemistry transport models. The stud focused mainly upon three emission scenarios: 1) current legislation (CLE); 2) case of the maximum emission reductions (MFR); and 3) pessimistic IPCC SRES A2 scenario. The model output showed a good agreement with observations in Europe and North America primarily because of quality-controlled measurements reported from these regions. The study suggested that in the future, deposition fluxes are going to be controlled mainly by the changes in emissions, with atmospheric chemistry and climate having a very limited role.

Regional Comparison of Precipitation Scenario

Sulfate in the Atmosphere

Normally, free acidity is contributed by the acids of SO_2 and NO_2. In case of free acidity of H_2SO_4, pH value decreases with increasing concentration of SO_4 ions. A comparison of pH and SO4 in rainwater at different sites in the United States, Sweden, and India is shown in Figs. 10–12. The United States and Sweden are examples of developed and acidified countries, whereas India is a developing country having high pH of rainwater. From Figures. 10 and 11, pH decreases with increase of SO_4 concentration, while at Indian sites (Figure 12), even at higher SO_4 levels, higher pH of rainwater is observed. This indicates that in the United States and Sweden, the SO_4 is present as H_2SO_4, which gives free H^+ in rain, but in India, it is present in a different form that does not contribute free H^+ in rainwater. The Indian dusty atmosphere is rich in $CaCO_3$, which allows SO_2 to form calcium sulfate, due to which pH of rainwater in India and other dusty regions is observed to be higher. In the Indian region, the possible

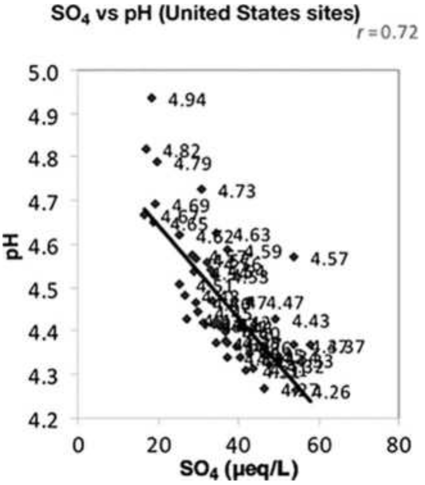


FIGURE 10 Variation of SO_4 and pH in rainwater at the sites in the United States.
Source: Trends in Precipitation Chemistry in the United States, 1983–94: An Analysis of the Effects in 1995 of Phase I of the Clean Air Act Amendments of 1990, Title IV.[61]

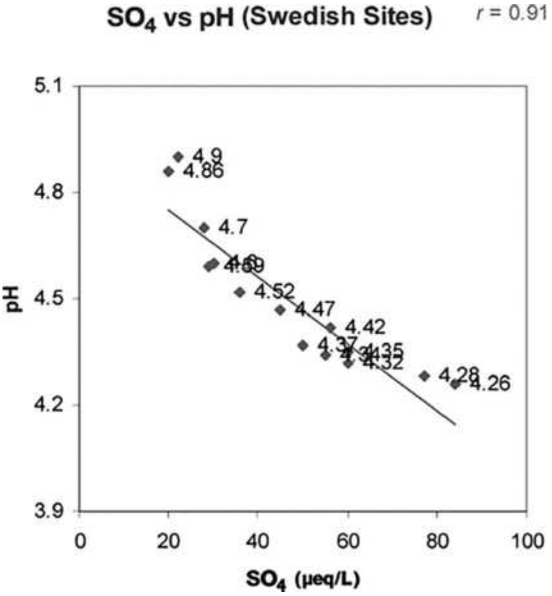


FIGURE 11 Variation of SO_4 and pH in rainwater at Swedish sites.
Source: Granat.[65]

mechanism of SO_2 removal is the adsorption of SO_2 onto the $CaCO_3$-dominated dust particles forming calcium sulfate.[62]

$$CaCO_3 + SO_2 + 1/2O_2 + 2H_2O \rightarrow CaSO_4.2H_2O + CO_2 \quad (14)$$

A comparison of typical composition of precipitation in an acidified region and a dusty region has been reported by Rodhe and coworkers,[48] establishing such differences very clearly.

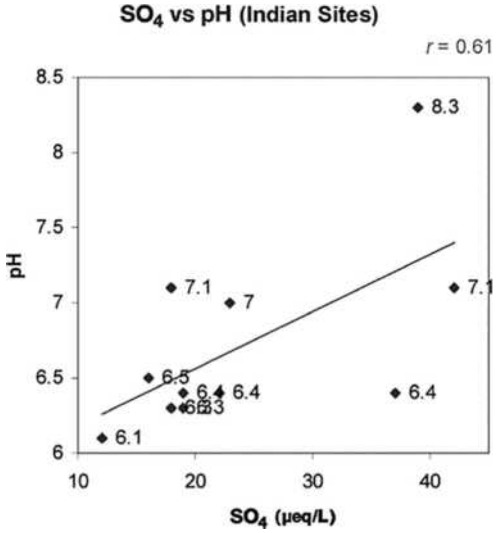


FIGURE 12 Variation of SO_4 with pH at Indian sites.
Source: Granat.[65]

Control of Acid Rain

Acid Rain Control Policy of the United States

Due to pressure from the public, the federal government of the United States adopted the Clean Air Act in 1970. Under this act, emission standards were set for SO_2 and NO_x, and states were directed to compliance with the National Ambient Air Quality Standards. Congress formed a 10 years program, the NAPAP, and mandated it to conduct scientific, technological, and economic study of the acid rain.

In 1990, under Clean Air Act Amendments, the National Deposition Control Program was implemented. This legislation was to control adverse effects of acidic deposition through reductions in emissions of SO_2 and NO_x. It was targeted to achieve a 50% reduction in annual SO_2 emissions by the year 2000. In 2001, SO_2 emissions from utilities subject to the provisions of the acid rain program were 39% below their 1980 levels, and total emissions from all sources were 50% less than their 1980 levels.[67] These implementation steps will help in environment protection, in particular, to check deterioration of historic buildings, to reduce fine particulate matter (sulfates, nitrates) and ground-level ozone (smog), and to improve public health.[68]

Since Canada is affected by transboundary pollution, an air quality accord was signed between the United States and Canada in 1991 under the framework of the UN Economic Commission for Europe. This bilateral accord facilitates the United States' and Canada's meeting their emissions targets for SO_2 and NO_x and putting coordinated efforts into atmospheric modeling and monitoring the effects of transboundary air pollution.[69]

European Policy to Control Acid Rain

Among European countries, most of the scientific research on acid rain effects was conducted in Norway and Sweden. In the beginning, acidification of lakes through transboundary pollution was the primary issue in Scandinavian countries. In 1972, at the UN Conference on the Human Environment in Stockholm, Sweden's case study on the effects of long-range transport of sulfur compounds was presented, which emphasized the need for international agreement to reduce damage from acid deposition. Soon after the Stockholm conference, in 1978, OECD initiated EMEP to monitor long-range pollution.

The 1985 Helsinki Protocol was the first binding commitment on the reduction of sulfur emissions or their transboundary fluxes by at least 30%. Later, a group of 12 countries decided to sign a declaration to reduce by 30% NO_x emissions by 1998 as compared with 1986 (base year). The latest agreement is the 1999 Gothenburg Protocol, which deals with SO_2, NO_x, NH_3, and non-methane volatile organic compounds (VOCs), aiming to mitigate the problem of acidification, eutrophication, and ground-level ozone. The Gothenburg Protocol targets the reduction of Europe's sulfur emissions by at least 63%, NO_x emissions by 41%, VOC emissions by 40%, and NH_3 emissions by 17%by the year 2010 from their 1990 levels.[70]

Effects of Acid Rain

Acid rain is very harmful to the environment as it damages many living and non-living things over a period of time. Acid rain affects both terrestrial as well as aquatic life. There are many inevitable impacts of acid rain, which affect natural as well as anthropogenic environments.

Effects on Aquatic Ecosystem

The aquatic ecosystem is visibly affected by acid rain as it directly falls into water bodies like lakes, streams, and rivers. Also, the extra acidic rainwater from other terrestrial places like forests and roads flows into nearby water bodies. Although the acidic effect of the rain may get diluted after it is mixed into the water bodies, it may lower the average pH of the aquatic system over the period, and in case the water body has low base cation supply or buffering capacity, it can become acidic faster. Charles and Norton[71] have reviewed the situation in lakes in United States and Canada and found that around the 1920s–1950s onward, weakly buffered lakes in some regions became more acidic. There are estimated to be around 50,000 lakes in the United States and Canada with a pH below 5.3. Out of these, hundreds of lakes have very low pH and are unable to support aquatic life, eliminating many existing insect and fish species. At pH lower than 5, the life of aquatic animals is threatened as they are unable to absorb oxygen from water. At pH lower than 4.8, fishes, frogs, and aquatic insects experience increased mortality. However, constructive steps, such as SO_2 emission reduction and liming of lakes, have significant potential for reversibility.[72,73]

Effects on Vegetation and Soil

In addition to the aquatic ecosystem, acid rain significantly affects trees, plants, forests, and other vegetation. Acid rain can damage the leaves and stems of the trees and affect their growth by getting absorbed through roots via soil. Acid rain reacts with leaves and stem wax coatings and allows acidic water to enter the leaves, thereby damaging the trees and plants. Experimental studies have established that acidic deposition causes some physiological effects in plants.[74] Especially at high altitude, the forests are surrounded by clouds carrying acidic water. The moisture of the clouds passing through the forests leads to severe damaging effects on the forests. In addition to individual effects on trees and forests, there are also effects on the soil, which contains the necessary nutrients and microorganisms for the healthy growth of the trees, plants, forests, and other vegetation. Virtual effects of acids rain have been observed all over the world, especially in Europe and eastern North America.[75,76] Hedin and coworkers[77] have reported evidence of steep decline of base cations in precipitation, which is based on long-term quality-controlled measurements, in Europe and North America. These measurements support that decline in base cations in precipitation might have resulted in increased sensitivity of a weakly buffered ecosystem, affecting forests and vegetation. Nitrification of ammonia (NH_3) and ammonium (NH_4^+) leads to the acidification of soil, which also adds to nutrient leaching:

$$NH_3 + O_2 \rightarrow NO_2 - +3H^+ + 2e- \tag{15}$$

$$NO_2 - +H_2O \rightarrow NO_3 - +2H^+ + 2e- \tag{16}$$

Effects on Buildings and Monuments

Acid rain can damage buildings and historic monuments as well by reacting with their paints and construction material. The damaged walls of the buildings and monuments leave a rough surface along with the moisture, which is a favorable place for the growth of microorganisms. Acid rain can affect sculpture and architecture adversely by corrosion. Acid rain can even corrode railway tracks, paints of cars, and joints of bridges and flyovers. A BERG report[78] gives more details of evidence of damage to stone and other materials by acid deposition. Under CLRTAP, the International Cooperative Programme on Effects on Materials, Including Historic and Cultural Monuments, has been set up, with 39 test sites, three of which are in the United States and Canada. According to a report, the corrosion rates of carbon steel, paint on steel, limestone, and bronze have decreased to about 60% from 1987 values in Europe due to decrease in SO_2 levels.[79]

Conclusion

Acid rain, which was a problem of North America and Europe, is now spreading in East Asia due to increased emissions of SO_2 from fossil fuel combustion. Experiences of the global community show that acid rain is spread through transboundary pollution. Scandinavia, Canada, and Japan are such examples. Acid rain damages water bodies, forests, vegetation, buildings, human health, etc., costing a huge loss to the economy. Although appreciable steps are taken by the United States, Canada, and Europe to control SO_2 and NO_x emissions, the considerable increase in SO_2 emission rates in China is of concern in the Asian region. During the past two decades, Chinese SO_2 emissions increased tremendously. Data show that spread of acid rain is very much controlled by suspended atmospheric soil dust in Indian and African regions. At a glance, the following are concluded:

1. Decreasing trends of SO_2 emissions in Europe and North America will be helpful in improving the pH of precipitation in coming decades.
2. Increasing trends of acidity in China in Asia may result in more acid rains in the region.
3. Systematic monitoring networks are established in North America, Europe, and East Asia to monitor acid deposition.
4. There is a strong need for long-term studies on acid deposition, including wet and dry depositions at selected sites in South Asia, Africa, and South America, through extensive networking.

References

1. Cowling, E.B. Acid precipitation in historical perspectives. Environ. Sci. Technol. **1982**, *16,* 110A.
2. Galloway, J.N. Acidification of the world: Natural and anthropogenic. Water Air Soil Pollut. **2001**, *130,* 17–24.
3. Charlson, R.J.; Rodhe, H. Factors controlling the acidity of natural rainwater. Nature **1982**, *295,* 683–685.
4. Seinfeld, J.; Pandis, S. *Atmospheric Chemistry and Physics,* 2nd Ed.; Wiley: New Jersey, 2006.
5. Brimblecombe, P. Interest in air pollution among early fellows of the Royal Society. Notes Rec. R. Soc. **1978**, *32,* 123.
6. Odén, S. Dagens Nyheter (Sweden), October 24, 1967.
7. Odén, S. The acidification of air precipitation and its consequences in natural environment. Ecology Committee Bulletin No. 1; Swedish Sciences Research Council, Stockholm. Translation Consultants Ltd.: Arlington, VA, 1968.
8. Martin, A.; Barber, F.R. Acid gases and acid in rain monitored for over 5 years in rural east-central England. Atmos. Environ. **1984**, *18,* 1715–1724.

9. Vermeulen, A.J. The acidic precipitation phenomenon. A study of this phenomenon and of a relationship between the acid content of precipitation and the emission of sulphur dioxide and nitrogen oxides in the Netherlands. In *Polluted Rain;* Toribara, T.Y., Miller, M.W., Morrow, P.E., Eds.; Plenum Press: New York, 1979; 7–60.
10. Gibson, J.H. Evaluation of wet chemical deposition in north America. In *Deposition Both Wet and Dry*, Acid Precipitation Series; Hicks, B.B., Ed.; Butterworth Publishers: London, 1984; Vol. 4, 1–13.
11. Keene, W.C.; Galloway, J.N.; Holden, J.D. Measurement of weak organic acidity in precipitation from remote areas of the world. J. Geophys. Res. **1983**, 88, 5122.
12. Jacob, D.J.; Wofsy, S.C. Photochemistry of biogenic emissions over the Amazon forest. J. Geophys. Res. **1988**, *93,* 1477–1486.
13. Chameides, W.L.; Davis, D.D. Aqueous phase source for formic acid in clouds. Nature **1983**, *304,* 427–429.
14. Sanhueza, E.; Andreae, M.O. Emission of formic and acetic acids from tropical Savanna soils. Geophys. Res. Lett. **1991**, *18,* 1707–1710.
15. Rodhe, H.; Granat, L. An evaluation of sulfate in European precipitation 1955–1982. Atmos. Environ. **1984**, *18*, 2627–263.
16. Overrein, L.; Seip, H.M.; Tollan, A. *Acid Precipitation— Effects on Forest and Fish,* Final report of the SNSF project 1972–1980. FR 19/80; Norwegian Institute for Water Research: Oslo, Norway, 1980.
17. Bolin, B., Ed. The impact on the environment of sulphur in air and precipitation. Sweden's National Report to the United Nations Conference on the Human Environment; Air Pollution Across Boundaries; Norstadt, Stockholm, 1971.
18. Parashar, D.C.; Granat, L.; Kulshrestha, U.C.; Pillai, A.G.; Naik, M.S.; Momim, G.A.; Prakasa Rao, P.S.; Safai, P.D.; Khemani, L.T.; Naqvi, S.W.A.; Narverkar, P.V.; Thapa, K.B.; Rodhe, H. Chemical composition of precipitation in India and Nepal. A preliminary report on an Indo-Swedish project on atmospheric chemistry. Report CM 90; IMI, Stockholm University: Sweden, 1996.
19. Kulshrestha, U.C.; Granat, L.; Engardt, M.; Rodhe, H. Review of precipitation chemistry studies in India—A search for regional patterns. Atmos. Environ. **2005**, *39,* 7403–7419.
20. Ayers, G.P.; Gillet, R.W. Acidification in Australia. In *Acidification in Tropical Countries,* SCOPE 36; Rodhe, H., Herrera, R., Eds.; John Wiley and Sons: Chichester, England, 1988; 347–400.
21. Likens, G.E.; Acid precipitation. Chem. Eng. News, November 22, 1976; 29–43.
22. Likens, G.E.; Butler, T.J. Recent acidification of precipitation in North America. Atmos. Environ. **1981**, *15,* 1103–1109.
23. Barrie, L.A.; Anlauf, K.; Wiebe, H.A.; Fellin, P. Acidic pollutants in air and precipitation at selected rural locations in Canada. In *Deposition of Both Wet and Dry*, Acid Precipitation Series; Hicks, B.B., Ed.; Butterworth Publishers: Boston, 1984; Vol. 4, 15–36.
24. Environmental Resources Ltd Pearce. *Acid Rain—A Review of the Phenomenon in the EEC and Europe;* Graham and Trotman: London, 1983.
25. Streets, D.G.; Carmichael, G.R.; Amann, M.; Arndt, R.L. Energy consumption and acid deposition in northeast Asia. Ambio **1999**, *28,* 135–143.
26. Jie, T.; Xiao, X.; Bin, B.; Jin, A.; Feng, W.S. Trends of the precipitation acidity over China during 1992–2006. Chin. Sci. Bull. **2010**, 55, 1800–1807. doi: 10.1007/s11434-009-3618-1.
27. Hara, H.; Akimoto, H. National level variations in precipitation chemistry in Japan. In Proceedings of the International Conference on Regional Environment and Climate Changes in East Asia, Taipei. November 30–December 3, 1993.
28. Garivaita, H.; Yoshizumib, K.; Morknoya, D.; Chanatorna, D.; Meepola, J.; Mark-Maia, A. Characterization of wet deposition in suburban area of Bangkok, Thailand. In Proceedings of CAD Workshop, Hyderabad, India, Nov–Dec 2006.
29. Khemani, L.T.; Momin, G.A.; Prakash Rao, P.S.; Safai, P.D.; Singh, G.; Kapoor R.K. Spread of acid rain over India. Atmos. Environ. **1989**, *23,* 757–762.

30. Kulshrestha, U.C. Air quality assessment through atmospheric depositions: A comparison of measurements and model calculations for India. Indian J. Environ. Manage. **2007**, *34,* 51–55.
31. Das, D.K. Chemistry of monsoon rains over Calcutta, West Bengal. Mausam **1988**, *39,* 75–82.
32. Hegde, P.; Kulshrestha, U.C.; Dumka, U.C.; Naza, M.; Pant, P. Chemical characteristics of rainwater and atmospheric aerosols at an elevated site in Himalayan Ranges in India. In Proceedings of CAD Workshop, Hyderabad, India, Nov–Dec 2006.
33. Jain, M.; Kulshrestha, U.C.; Sarkar, A.K.; Parashar, D.C. Influence of crustal aerosols on wet deposition at urban and rural sites in India. Atmos. Environ. **2000**, *34,* 5129–5137.
34. Khemani, L.T.; Momin, G.A.; Rao P.S.P.; Pillai, A.G.; Safai, P.D.; Mohan, K.; Rao, M.G. Atmospheric pollutants and their influence on acidification of rain water at an industrial location on the west coast of India. Atmos. Environ. **1994**, *28,* 3145–3154.
35. Kulshrestha, U.C. Chemistry of atmospheric depositions in India. In Proceedings of Symposium Science at High Altitudes, Darjiling, May 7–11, 1997; 1998. 187–194.
36. Kulshrestha, U.C.; Sarkar, A.K.; Srivastava, S.S.; Parashar, D.C. Investigation into atmospheric deposition through precipitation studies at New Delhi (India), Atmos. Environ. **1996**, *30,* 4149–4154.
37. Kulshrestha, U.C.; Kulshrestha, M.J.; Sekar, R.; Sastry, G.S.R.; Vairamani, M. Chemical characteristics of rain water at an urban site of south-central India. Atmos. Environ. **2003**, 37, 3019–3026.
38. Kulshrestha, M.J.; Reddy, L.A.K.; Satyanarayana, J.; Duarah, R.; Rao, P.G.; Kulshrestha, U.C. Chemical characteristics of rain water at two rural sites of NE and SW India. In Proceedings of CAD Workshop, Hyderabad, India, Nov–Dec 2006.
39. Kumar, N.; Kulshrestha, U.C.; Saxena, A.; Kumari, K.M.; Srivastava, S.S. Formate and acetate in monsoon rain water of Agra. J. Geophys. Res. **1993**, *98,* D3, 5135.
40. Mahadevan, T.N.; Negi, B.S.; Meenakshi, V. Measurements of elemental composition of aerosol matter and precipitation from a remote background site in India. Atmos. Environ. **1989**, *23,* 869–874.
41. Mukhopadhyay, B.; Datar, S.V.; Srivastava, H.N. Precipitation chemistry over the Indian region. Mausam **1992**, *43,* 249–258.
42. Pillai, A.G.; Naik, M.S.; Momin, G.A.; Rao, P.D.; Safai, P.D.; Ali, K.; Rodhe, H.; Granat, L. Studies of wet deposition and dustfall at Pune, India. Water Air Soil Pollut. **2001**, *130,* 475–480.
43. Rao, P.S.P.; Momin, G.A.; Safai, P.D.; Pillai, A.G.; Khemani, L.T. Rain water and throughfall chemistry in the Silent Valley forest in South India. Atmos. Environ. **1995**, *29*, 2025–2029.
44. Saxena, A.; Sharma, S.; Kulshrestha, U.C.; Srivastava, S.S. Factors affecting alkaline nature of rain water in Agra (India). Environ. Pollut. **1991**, *74,* 129–138.
45. Kulshrestha, U.C.; Jain, M.; Mandal, T.R.; Gupta, P.K.; Sarkar, A.K.; Parashar, D.C. Measurements of acid rain over Indian Ocean and surface measurements of atmospheric aerosols at New Delhi during INDOEX pre-campaigns. Curr. Sci. **1999**, *76,* 968–972.
46. Kulshrestha, U.C.; Rodhe, H. Precipitation chemistry metadata from Asia. Presented in the 2nd Steering Committee Meeting of CAD programme, Bangkok, Thailand, Nov 26–27, 2007.
47. Galy-Lacaux, C.; Laouali, D.; Descroix, L.; Gobron, N.; Liousse, C. Long term precipitation chemistry and wet deposition in a remote dry savanna site in Africa (Niger) Atmos. Chem. Phys. **2009**, 9, 1579–1595.
48. Rodhe, H.; Dentener, F.; Schulz, M. The global distribution of acidifying wet deposition. Environ. Sci. Technol. **2002**, *36,* 4382–4388.
49. Zhu, Q. Trends in SO_2 emissions. Report of IEA Clean Coal Center, October 2006. PF 06–09, London.
50. USEPA Web site, available at http://www.epa.gov/air/emissions/so2.htm. (accessed March 2, 2011).
51. Stensland, G.J. Precipitation chemistry trends in the northern United States. In *Polluted Rain;* Toribara, T.Y., Miller, M.W., Morrow, P.E., Eds.; Plenum Press: New York, 1979; 87–108.
52. Willey, J.; Kiber, R.; Avery, G.B., Jr. Changing chemical composition of precipitation in Wilmington, North Carolina, U.S.A.: Implications for the continental U.S.A. Environ. Sci. Technol. **2006**, *40,* 5675–5680.

53. Lu, Z.; Streets, D.G.; Zhang, Q.; Wang, S.; Carmichael, G.R.; Cheng, Y.F.; Wei, C.; Chin, M.; Diehl, T.; Tan, Q. Sulfur dioxide emissions in China and sulfur trends in East Asia since 2000. Atmos. Chem. Phys. Discuss. **2010**, *10,* 8657–8715.
54. Hicks, K.; Kuylenstierna, J.; Owen, A.; Rodhe, H.; Seip, H.; Dentener, F. Assessing the time development of acidification damage in Asian soils at regional scale. In Proceedings of CAD Workshop, Hyderabad, India, Nov–Dec 2006.
55. Dentener, F.; Drevet, J.; Lamarque, J.F.; Bey, I.; Eickhout, B.; Fiore, A.M.; Hauglustaine, D.; Horowitz, L.W.; Krol, M.; Kulshrestha, U.C.; Lawrence, M.; Galy-Lacaux, C.; Rast, S.; Shindell, D.; Stevenson, D.; Van Noije, T.; Atherton, C.; Bell, N.; Bergman, D.; Butler, T.; Cofala, J.; Collins, B.; Doherty, R.; Ellingsen, K.; Galloway, J.; Gauss, M.; Montanaro, V.; Müller, J.F.; Pitari, G.; Rodriguez, J.; Sanderson, M.; Solmon, F.; Strahan, S.; Schultz, M.; Sudo, K.; Szopa, S.; Wild, O. Nitrogen and sulfur deposition on regional and global scales: A multimodel evaluation. Global Biogeochem. Cycles **2006**; *20* (4), GB4003, doi:10.1029/2005GB002672.
56. Rao, P.S.P.; Safai, P.D.; Momin, G.A.; Ali, K.; Chate, D. M.; Praveen, P.S.; Tiwari, S. Precipitation chemistry at different locations in India. In Proceedings of CAD Workshop, Hyderabad, India, Nov–Dec 2006.
57. Das, R.; Das, S.N.; Misra, V.N.; Chemical composition of rainwater and dust fall at Bhubaneswar in the east coast of India. Atmos. Environ. **2005**, *34,* 5908–5916.
58. Das, N.; Das, R.; Chaudhury, G.R.; Das, S.N. Chemical composition of precipitation at background level. Atmos. Res. **2010**, *95,* 108–113.
59. Kulshrestha, U.C.; Jain, M.; Sekar, R.; Vairamani, M.; Sarkar, A.K.; Parashar, D.C. Chemical characteristics and source apportionment of aerosols over Indian Ocean during INDOEX-1999. Curr. Sci. **2001**, *80,* 180–185.
60. Granat, L.; Norman, M.; Leck, C.; Kulshrestha, U.C.; Rodhe, H. Wet scavenging of sulfur and other compounds during INDOEX. J. Geophys. Res. **2002**, *107,* D19, 8025, doi:1011029/2001JD000499.
61. Trends in Precipitation Chemistry in the United States, 198394: An Analysis of the Effects in 1995 of Phase I of the Clean Air Act Ammendments of 1990, Title IV, http://pvbs.usgs.gov/acidrain/index.html#tables.
62. Kulshrestha, M.J.; Kulshrestha, U.C.; Parashar, D.C.; Vairamani, M. Estimation of SO_4 contribution by dry deposition of SO_2 onto the dust particles in India. Atmos. Environ. **2003**, 37, 3057–3063.
63. Kuylenstierna, J.C.I.; Rodhe, H.; Cinderby, S.; Hicks, K. Acidification in developing countries: Ecosystem sensitivity and the critical load approach on a global scale. Ambio **2001**, *30,* 20–28.
64. Alcamo, J.; Shaw, R.; Hordijk, L., Eds. *The RAINS Model of Acidification. Science and Strategies in Europe.* Kluwer Academic Publishers: Dordrecht, Netherlands, 1990.
65. Granat, L. Luft-och nederbordskemiska stationsnatet inom PMK. Report 3942, Meteorological Institute at Stockholm University: Stockholm, 1990.
66. Hara, H. Temporal variation of wet deposition in the EANET region during 2000–2004. In Proceedings of CAD Workshop, Hyderabad, India, Nov–Dec 2006.
67. U.S. Environmental Protection Agency (USEPA). *Acid Rain Program, 2001 Progress report;* U.S. Environmental Protection Agency: Washington, DC, available at http://www.epa.gov/airmarkets/cmprpt/arp01/2001report.pdf.
68. U.S. Environmental Protection Agency (USEPA). *Acid Rain Program: Overview, Environmental Benefits*, available at http://www.epa.gov/airmarkets/arp/overview.html. Retrieved on March 2, 2011.
69. U.S. Environmental Protection Agency (USEPA). Canada- United States agreement, 1994 Progress report, EPA/430/ R-94/013; Washington, DC. Retrieved on March 2, 2011.
70. Menz, F.C.; Seip, F.M. Acid rain in Europe and the United States: An update. Environ. Sci. Policy **2004**, *7,* 253–265.
71. Charles, D.F.; Norton, S.A. Paleolimnological evidence for trends in the atmospheric deposition of acids and heavy metals. In *Atmospheric Deposition: Historic Trends and Spatial Patterns*; Norton, S.A., Ed.; National Academy Press: Washington, DC, USA, 1985; 86–105.

72. Battarbee, R.W.; Flower, R.J.; Stevenson, A.C.; Harriman, R.; Appleby, P.G. Diatom and chemical evidence for reversibility of acidification of Scottish Lochs. Nature (London) **1988**, *332,* 530–532.
73. Jenkins, A.; Whitehead, P.G.; Cosby, B.J.; Birks, H.J.B. Modelling long term acidification: A comparison with diatom reconstructions and the implications for reversibility. Philos. Trans. R. Soc. London **1990**, *B327,* 435–440.
74. Evans, L.S.; Lewin, K.F.; Parri, M.J.; Cunningham, E.A Productivity of field-grown soybeans exposed to simulated acid rain. New Phytol. **1983**, *93,* 377–388.
75. Rehfuess, K.E. On the impact of acid precipitation in forest ecosystems. Forstwiss. Centralbl. **1981**, *100,* 363–370.
76. Cowling, E.B. Regional declines on forests in Europe and North America: The possible role of airborne chemicals. In *Aerosols, Research Risk Assessment and Control Strategies;* Lewis: Chelsea, MI, USA, 1986; 855–864.
77. Hedin, L.O.; Granata, L.; Likens, G.E.; Buishand, T.A.; Galloway, J.N.; Butler, T.J.; Rodhe, H. Steep decline in atmospheric base cations in regions of Europe and north America. Nature **1994**, *367,* 351–367.
78. BERG, Building Effects Review Group. *The Effects of Acid Deposition on Building Materials in U.K.;* Department of Environment, HMSO: London, 1989.
79. Kucera, V. *Atmospheric Corrosion in Urban Air Pollution in Asia and Africa: The Approach of the RAPIDC Programme;* Kuylenstierna, J., Hicks, K., Eds.; Stockholm Environment Institute (SEI): Stockholm, Sweden, 2002.

44

Acid Rain: Nitrogen Deposition

George F. Vance

Introduction

Air pollution has occurred naturally since the formation of the Earth's atmosphere; however, the industrial era has resulted in human activities greatly contributing to global atmospheric pollution.[1,2] One of the more highly publicized and controversial aspects of atmospheric pollution is that of acidic deposition. Acidic deposition includes rainfall, acidic fogs, mists, snowmelt, gases, and dry particulate matter.[3] The primary origin of acidic deposition is the emission of sulfur dioxide (SO_2) and nitrogen oxides (NO_x) from fossil fuel combustion; electric power generating plants contribute approximately two-thirds of the SO_2 emissions and one-third of the NO_x emissions.[4]

Acidic materials can be transported long distances, some as much as hundreds of kilometers. For example, 30%–40% of the S deposition in the northeastern U.S. originates in industrial midwestern U.S. states.[5] After years of debate, U.S. and Canada have agreed to develop strategies that reduce acidic compounds originating from their countries.[5,6] In Europe, the small size of many countries means that emissions in one industrialized area can readily affect forests, lakes, and cities in another country. For example, approximately 17%of the acidic deposition falling on Norway originated in Britain and 20% in Sweden came from eastern Europe.[5]

The U.S. EPA National Acid Precipitation Assessment Program (NAPAP) conducted intensive research during the 1980s and 1990s that resulted in the "Acidic Deposition: State of the Science and Technology" that was mandated by the Acid Precipitation Act of 1980.[6] NAPAP Reports to Congress have been developed in accordance with the 1990 amendment to the 1970 Clean Air Act and present the expected benefits of the Acid Deposition Control Program,[6,7] http://www.nnic.noaa.gov/CENR/NAPAP/. Mandates include an annual 10 million ton or approximately 40% reduction in point-source SO_2 emissions below 1980 levels, with national emissions limit caps of 8.95 million tons from electric utility and 5.6 million tons from point-source industrial emissions. A reduction in NO_x of about 2 million tons from 1980 levels has also been set as a goal; however, while NO_x has been on the decline since 1980, projections estimate a rise in NO_x emissions after the year 2000. In 1980, the U.S. levels of SO_2 and NO_x emissions were 25.7 and 23.0 million tons, respectively.

Acidic deposition can impact buildings, sculptures, and monuments that are constructed using weatherable materials like limestone, marble, bronze, and galvanized steel,[7,8] http://www.nnic.noaa.gov/CENR/NAPAP/. While acid soil conditions are known to influence the growth of plants, agricultural impacts related to acidic deposition are of less concern due to the buffering capacity of these types of ecosystems.[2,5] When acidic substances are deposited in natural ecosystems, a number of adverse environmental effects are believed to occur, including damage to vegetation, particularly forests, and changes in soil and surface water chemistry.[9,10]

Sources and Distribution

Typical sources of acidic deposition include coal- and oil-burning electric power plants, automobiles, and large industrial operations (e.g., smelters). Once S and N gases enter the earth's atmosphere they react very rapidly with moisture in the air to form sulfuric (H_2SO_4) and nitric (HNO_3) acids.[2,3] The pH of natural rainfall in equilibrium with atmospheric CO2 is about 5.6; however, the pH of rainfall is less than 4.5 in many industrialized areas. The nature of acidic deposition is controlled largely by the geographic distribution of the sources of SO_2 and NO_x (Figure 1). In the midwestern and northeastern U.S., H_2SO_4 is the main source of acidity in precipitation because of the coal-burning electric utilities.[2] In the western U.S., HNO_3 is of more concern because utilities and industry burn coal with low S contents and populated areas are high sources of NO_x.[2]

Emissions of SO_2 and NO_x increased in the 20th century due to the accelerated industrialization in developed countries and antiquated processing practices in some undeveloped countries. However, there is some uncertainty as to the actual means by which acidic deposition affects our environment,[11,12] http://nadp.sws.uiuc.edu/isopleths/maps1999/. Chemical and biological evidence, however, indicates that atmospheric deposition of H_2SO_4 caused some New England lakes to decrease in alkalinity.[13,14] Many scientists are reluctant to over-generalize cause and effect relationships in an extremely complex environmental problem. Although, the National Acid Deposition Assessment Program has concluded there were definite consequences due to acidic deposition that warrant

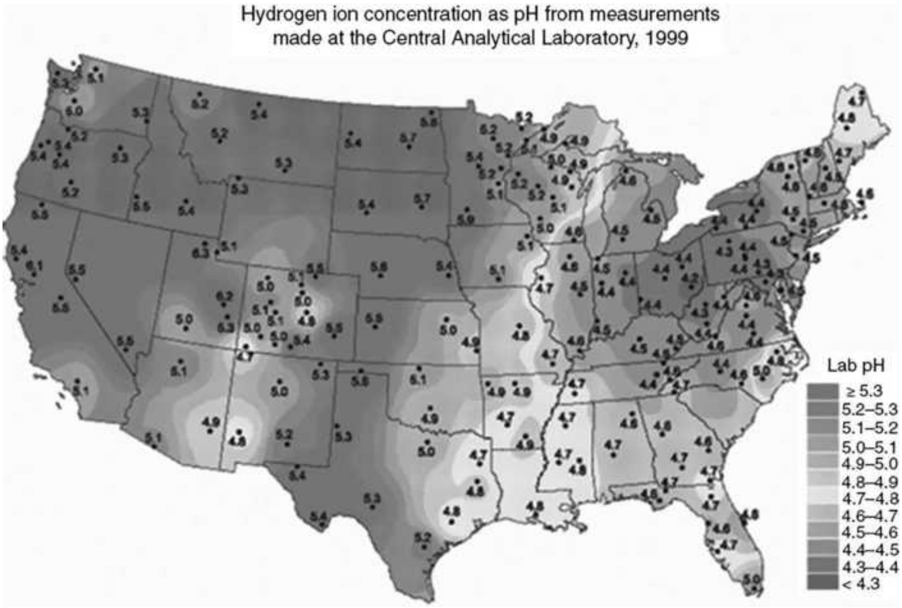


FIGURE 1 Acidic deposition across the U.S. during 1999.
Source: National Atmospheric Deposition Program/National Trends Network http://nadp.sw.uluc.edu.[11]

remediation[6,7] http://www.nnic.noaa.gov/CENR/NAPAP/. Since 1995, when the 1990 Clean Air Act Amendment's Title IV reduction in acidic deposition was implemented, SO_2 and NO_x emissions have, respectively, decreased and remained constant during the late 1990s.[4]

Both H_2SO_4 and HNO_3 are important components of acidic deposition, with volatile organic compounds and inorganic carbon also components of acidic deposition- related emissions. Pure water has a pH of 7.0, natural rainfall about 5.6, and severely acidic deposition less than 4.0. Uncontaminated rainwater should be pH 5.6 due to CO_2 chemistry and the formation of carbonic acid. The pH of most soils ranges from 3.0 to 8.0.[2] When acids are added to soils or waters, the decrease in pH that occurs depends greatly on the system's buffering capacity, the ability of a system to maintain its present pH by neutralizing added acidity. Clays, organic matter, oxides of Al and Fe, and Ca and Mg carbonates (limestones) are the components responsible for pH buffering in most soils. Acidic deposition, therefore, will have a greater impact on sandy, low organic matter soils than those higher in clay, organic matter, and carbonates. In fresh waters, the primary buffering mechanism is the reaction of dissolved bicarbonate ions with H^+ according to the following equation:

$$H^+ + HCO_3^- = H_2O + CO_2 \tag{1}$$

Human Health Effects

Few direct human health problems have been attributed to acidic deposition. Long-term exposure to acidic deposition precursor pollutants such as ozone (O_3) and NO_x, which are respiratory irritants, can cause pulmonary edema.[5,6] Sulfur dioxide (SO_2) is also a known respiratory irritant, but is generally absorbed high in the respiratory tract.

Indirect human health effects due to acidic deposition are more important. Concerns center around contaminated drinking water supplies and consumption of fish that contain potential toxic metal levels. With increasing acidity (e.g., lower pH levels), metals such as mercury, aluminum, cadmium, lead, zinc, and copper become more bioavailable.[2] The greatest human health impact is due to the consumption of fish that bioaccumulate mercury; freshwater pike and trout have been shown to contain the highest average concentrations of mercury.[5,15] Therefore, the most susceptible individuals are those who live in an industrial area, have respiratory problems, drink water from a cistern, and consume a significant amount of freshwater fish.

A long-term urban concern is the possible impact of acidic deposition on surface-derived drinking water. Many municipalities make extensive use of lead and copper piping, which raises the question concerning human health effects related to the slow dissolution of some metals (lead, copper, zinc) from older plumbing materials when exposed to more acidic waters. Although metal toxicities due to acidic deposition impacts on drinking waters are rare, reductions in S and N fine particles expected by 2010 based on Clean Air Act Amendments will result in annual public health benefits valued at $50 billion with reduced mortality, hospital admissions and emergency room visits.[16]

Structural Impacts

Different types of materials and cultural resources can be impacted by air pollutants. Although the actual corrosion rates for most metals have decreased since the 1930s, data from three U.S. sites indicate that acidic deposition may account for 31%–78% of the dissolution of galvanized steel and copper,[7,8] http://www.nnic.noaa.gov/CENR/NAPAP/. In urban or industrial settings, increases in atmospheric acidity can dissolve carbonates (e.g., limestone, marble) in buildings and other structures. Deterioration of stone products by acidic deposition is caused by: 1) erosion and dissolution of materials and surface details; 2) alterations (blackening of stone surfaces); and 3) spalling (cracking and spalling of stone surfaces due to accumulations of alternation crusts.[8] Painted surfaces can be discolored or etched, and there may also be degradation of organic binders in paints.[8]

Ecosystem Impacts

It is important to examine the nature of acidity in soil, vegetation, and aquatic environments. Damage from acidification is often not directly due to the presence of excessive H^+, but is caused by changes in other elements. Examples include increased solubilization of metal ions such as Al^{3+} and some trace elements (e.g., Mn^{2+}, Pb^{2+}) that can be toxic to plants and animals, more rapid losses of basic cations (e.g., Ca^{2+}, Mg^{2+}), and the creation of unfavorable soil and aquatic environments for different fauna and flora.

Soils

Soil acidification is a natural process that occurs when precipitation exceeds evapotranspiration.[2] "Natural" rainfall is acidic (pH of ~ 5.6) and continuously adds a weak acid (H_2CO_3) to soils. This acidification results in a gradual leaching of basic cations (Ca^{2+} and Mg^{2+}) from the uppermost soil horizons, leaving Al^{3+} as the dominant cation that can react with water to produce H^+. Most of the acidity in soils between pH 4.0 and 7.5 is due to the hydrolysis of Al^{3+},[17,18] http://www.epa.gov/airmarkets/acidrain/effects/index.html. Other acidifying processes include plant and microbial respiration that produces CO_2, mineralization and nitrification of organic N, and the oxidation of FeS_2 in soils disturbed by mining or drainage.[2] In extremely acidic soils ($pH < 4.0$), strong acids such as H_2SO_4 are a major component.

The degree of accelerated acidification depends both upon the buffering capacity of the soil and the use of the soil. Many of the areas subjected to the greatest amount of acidic deposition are also areas where considerable natural acidification occurs.[19] Forested soils in the northeastern U.S. are developed on highly acidic, sandy parent materials that have undergone tremendous changes in land use in the past 200 years. However, clear-cutting and burning by the first European settlers have been almost completely reversed and many areas are now totally reforested.[5] Soil organic matter that accumulated over time represents a natural source of acidity and buffering. Similarly, greater leaching or depletion of basic cations by plant uptake in increasingly reforested areas balances the significant inputs of these same cations in precipitation.[20,21] Acidic deposition affects forest soils more than agricultural or urban soils because the latter are routinely limed to neutralize acidity. Although it is possible to lime forest soils, which is done frequently in some European countries, the logistics and cost often preclude this except in areas severely impacted by acidic deposition.[5]

Excessively acidic soils are undesirable for several reasons. Direct phytotoxicity from soluble Al^{3+} or Mn^{2+} can occur and seriously injure plant roots, reduce plant growth, and increase plant susceptibility to pathogens.[21] The relationship between Al^{3+} toxicity and soil pH is complicated by the fact that in certain situations organic matter can form complexes with Al^{3+} that reduce its harmful effects on plants.[18] Acid soils are usually less fertile because of a lack of important basic cations such as K^+, Ca^{2+}, and Mg^{2+}. Leguminous plants may fix less N_2 under very acidic conditions due to reduced rhizobial activity and greater soil adsorption of Mo by clays and Al and Fe oxides.[2] Mineralization of N, P, and S can also be reduced because of the lower metabolic activity of bacteria. Many plants and microorganisms have adapted to very acidic conditions (e.g., $pH < 5.0$). Examples include ornamentals such as azaleas and rhododendrons and food crops such as cassava, tea, blueberries, and potatoes.[5,22] In fact, considerable efforts in plant breeding and biotechnology are directed towards developing Al- and Mn-tolerant plants that can survive in highly acidic soils.

Agricultural Ecosystems

Acidic deposition contains N and S that are important plant nutrients. Therefore, foliar applications of acidic deposition at critical growth stages can be beneficial to plant development and reproduction. Generally, controlled experiments require the simulated acid rain to be pH 3.5 or less in order to produce injury to certain plants.[22] The amount of acidity needed to damage some plants is 100 times greater than natural rainfall. Crops that respond negatively in simulated acid rain studies include garden beets, broccoli, carrots, mustard greens, radishes, and pinto beans, with different effects for some cultivars. Positive responses to acid rain have been identified with alfalfa, tomato, green pepper, strawberry, corn, lettuce, and some pasture grass crops.

Agricultural lands are maintained at pH levels that are optimal for crop production. In most cases the ideal pH is around pH 6.0–7.0; however, pH levels of organic soils are usually maintained at closer to pH 5.0. Because agricultural soils are generally well buffered, the amount of acidity derived from atmospheric inputs is not sufficient to significantly alter the overall soil pH.[2] Nitrogen and S soil inputs from acidic deposition are beneficial, and with the reduction in S atmospheric levels mandated by 1990 amendments to the Clean Air Act, the S fertilizer market has grown. The amount of N added to agricultural ecosystems as acidic deposition is rather insignificant in relation to the 100–300 kg N/ha/yr required of most agricultural crops.

Forest Ecosystems

Perhaps the most publicized issue related to acidic deposition has been widespread forest decline. For example, in Europe estimates suggest that as much as 35% of all forests have been affected.[23] Similarly, in the U.S. many important forest ranges such as the Adirondacks of New York, the Green Mountains of Vermont, and the Great Smoky Mountains in North Carolina have experienced sustained decreases in tree growth for several decades.[6] Conclusive evidence that forest decline or dieback is caused solely be acidic deposition is lacking and complicated by interactions with other environmental or biotic factors. However, NAPAP research[6] has confirmed that acidic deposition has contributed to a decline in high-elevation red spruce in the northeastern U.S. In addition, nitrogen saturation of forest ecosystems from atmospheric N deposition is believed to result in increased plant growth, which in turn increases water and nutrient use followed by deficiencies that can cause chlorosis and premature needle-drop as well as increased leaching of base cations from the soil.[24]

Acidic deposition on leaves may enter directly through plant stomates.[1,22] If the deposition is sufficiently acidic (pH ~ 3.0), damage can also occur to the waxy cuticle, increasing the potential for direct injury of exposed leaf mesophyll cells. Foliar lesions are one of the most common symptoms. Gaseous compounds such as SO_2 and SO_3 present in acidic mists or fogs can also enter leaves through the stomates, form H_2SO_4 upon reaction with H_2O in the cytoplasm, and disrupt many metabolic processes. Leaf and needle necrosis occurs when plants are exposed to high levels of SO_2 gas, possibly due to collapsed epidermal cells, eroded cuticles, loss of chloroplast integrity and decreased chlorophyll content, loosening of fibers in cell walls and reduced cell membrane integrity, and changes in osmotic potential that cause a decrease in cell turgor.

Root diseases may also increase in excessively acidic soils. In addition to the damages caused by exposure to H_2SO_4 and HNO_3, roots can be directly injured or their growth rates impaired by increased concentrations of soluble Al^{3+} and Mn^{2+} in the rhizosphere,[2,25] http://nadp.sws.uiuc.edu. Changes in the amount and composition of these exudates can then alter the activity and population diversity of soil-borne pathogens. The general tendency associated with increased root exudation is an enhancement in microbial populations due to an additional supply of carbon (energy). Chronic acidification can also alter nutrient availability and uptake patterns.[8,22]

Long-term studies in New England suggest acidic deposition has caused significant plant and soil leaching of base cations,[1,21] resulting in decreased growth of red spruce trees in the White Mountains.[6] With reduction in about 80% of the airborne base cations, mainly Ca^{2+} but also Mg^{2+}, from 1950 levels, researchers suggest forest growth has slowed because soils are not capable of weathering at a rate that can replenish essential nutrients. In Germany, acidic deposition was implicated in the loss of soil Mg^{2+} as an accompanying cation associated with the downward leaching of SO_4^{2-}, which ultimately resulted in forest decline.[2] Several European countries have used helicopters to fertilize and lime forests.

Aquatic Ecosystems

Ecological damage to aquatic systems has occurred from acidic deposition. As with forests, a number of interrelated factors associated with acidic deposition are responsible for undesirable changes. Acidification of aquatic ecosystems is not new. Studies of lake sediments suggest that increased acidification began in the mid-1800s, although the process has clearly accelerated since the 1940s.[15] Current studies indicate there is significant S mineralization in forest soils impacted by acidic deposition and that the SO_4^{2-} levels in adjacent streams remain high, even though there has been a decrease in the amount of atmospheric-S deposition.[24]

Geology, soil properties, and land use are the main determinants of the effect of acidic deposition on aquatic chemistry and biota. Lakes and streams located in areas with calcareous geology resist acidification more than those in granitic and gneiss materials.[16] Soils developed from calcareous parent materials are generally deeper and more buffered than thin, acidic soils common to granitic areas.[2] Land management decisions also affect freshwater acidity. Forested watersheds tend to contribute more acidity than those dominated by meadows, pastures, and agronomic ecosystems.[8,14,20] Trees and other vegetation in forests are known to "scavenge" acidic compounds in fogs, mists, and atmospheric particulates. These acidic compounds are later deposited in forest soils when rainfall leaches forest vegetation surfaces. Rainfall below forest canopies (e.g., throughfall) is usually more acidic than ambient precipitation. Silvicultural operations that disturb soils in forests can increase acidity by stimulating the oxidization of organic N and S, and reduced S compounds such as FeS_2.[2]

A number of ecological problems arise when aquatic ecosystems are acidified below pH 5.0, and particularly below pH 4.0. Decreases in biodiversity and primary productivity of phytoplankton, zooplankton, and benthic invertebrates commonly occur.[15,16] Decreased rates of biological decomposition of organic matter have occasionally been reported, which can then lead to a reduced supply of nutrients.[20] Microbial communities may also change, with fungi predominating over bacteria. Proposed mechanisms to explain these ecological changes center around physiological stresses caused by exposure of biota to higher concentrations of Al^{3+}, Mn^{2+}, and H^+ and lower amounts of available Ca^{2+}.[15] One specific mechanism suggested involves the disruption of ion uptake and the ability of aquatic plants to regulate Na^+, K^+, and Ca^{2+} export and import from cells.

Acidic deposition is associated with declining aquatic vertebrate populations in acidified lakes and, under conditions of extreme acidity, of fish kills. In general, if the water pH remains above 5.0, few problems are observed; from pH 4.0 to 5.0 many fish are affected, and below pH 3.5 few fish can survive.[23] The major cause of fish kill is due to the direct toxic effect of Al^{3+}, which interferes with the role Ca^{2+} plays in maintaining gill permeability and respiration. Calcium has been shown to mitigate the effects of Al^{3+}, but in many acidic lakes the Ca^{2+} levels are inadequate to overcome Al^{3+} toxicity. Low pH values also disrupt the Na^+ status of blood plasma in fish. Under very acidic conditions, H^+ influx into gill membrane cells both stimulates excessive efflux of Na^+ and reduces influx of Na^+ into the cells. Excessive loss of Na^+ can cause mortality. Other indirect effects include reduced rates of reproduction, high rates of mortality early in life or in reproductive phases of adults, and migration of adults away from acidic areas.[16] Amphibians are affected in much the same manner as fish, although they are somewhat less sensitive to Al^{3+} toxicity. Birds and small mammals often have lower populations and lower reproductive rates in areas adjacent to acidified aquatic ecosystems. This may be due to a shortage of food due to smaller fish and insect populations or to physiological stresses caused by consuming organisms with high Al^{3+} concentrations.

Reducing Acidic Deposition Effects

Damage caused by acidic deposition will be difficult and extremely expensive to correct, which will depend on our ability to reduce S and N emissions. For example, society may have to burn less fossil fuel, use cleaner energy sources and/or design more efficient "scrubbers" to reduce S and N gas entering our atmosphere. Despite the firm conviction of most nations to reduce acidic deposition, it appears that the staggering costs of such actions will delay implementation of this approach for many years. The 1990 amendments to the Clean Air Act are expected to reduce acid-producing air pollutants from electric power plants. The 1990 amendments established emission allowances based on a utilities' historical fuel use and SO_2 emissions, with each allowance representing 1 ton of SO_2 that can bought, sold or banked for future use,[4,6,7] http://www.nnic.noaa.gov/CENR/NAPAP/. Short-term remedial actions for acidic deposition are available and have been successful in some ecosystems. Liming of lakes and some forests (also fertilization with trace elements and Mg^{2+}) has been practiced in European counties for over 50 years.[16,23] Hundreds of Swedish and Norwegian lakes have been successfully limed in the past 25 years. Lakes with short mean residence times for water retention may need annual or biannual liming; others may need to be limed every 5–10 years. Because vegetation in some forested ecosystems has adapted to acidic soils, liming (or over-liming) may result in an unpredictable and undesirable redistribution of plant species.

References

1. Smith, W.H. Acid rain. In *The Wiley Encyclopedia of Environmental Pollution and Cleanup*; Meyers, R.A., Dittrick, D.K., Eds.; Wiley: New York, 1999; 9–15.
2. Pierzynski, G.M.; Sims, J.T.; Vance, G.F. *Soils and Environmental Quality;* CRC Press: Boca Raton, FL, 2000; 459 pp.
3. Wolff, G.T. Air pollution. In *The Wiley Encyclopedia of Environmental Pollution and Cleanup;* Meyers, R.A., Dittrick, D.K., Eds.; Wiley: New York, 1999; 48–65.
4. U.S. Environmental protection agency. In *Progress Report on the EPA Acid Rain Program;* EPA-430-R-99-011; U.S. Government Printing Office: Washington, DC, 1999; 20 pp.
5. Forster, B.A. *The Acid Rain Debate: Science and Special Interests in Policy Formation;* Iowa State University Press: Ames, IA, 1993.
6. *National Acid Precipitation Assessment Program Task Force Report,* National Acid Precipitation Assessment Program 1992 Report to Congress; U.S. Government Printing Office: Pittsburgh, PA, 1992; 130 pp.
7. *National Science and Technology Council,* National Acid Precipitation Assessment Program Biennial Report to Congress: An Integrated Assessment; 1998 (accessed July 2001).
8. Charles, D.F., Ed. The acidic deposition phenomenon and its effects: critical assessment review papers. In *Effects Sciences;* EPA-600/8–83-016B; U.S. Environmental Protection Agency: Washington, DC, 1984; Vol. 2.
9. McKinney, M.L.; Schoch, R.M. *Environmental Science: Systems and Solutions*; Jones and Bartlett Publishers: Sudbury, MA, 1998.
10. United Nations, World band and World resources institute. In *World Resources: People and Ecosystems—The Fraying Web of Life*; Elsevier: New York, 2000.
11. *National Atmospheric Deposition Program (NRSP-3)/National Trends Network*. Isopleth Maps. NADP Program Office, Illinois State Water Survey, 2204 Griffith Dr., Champaign, IL, 61820, 2000 (accessed July 2001).
12. Council on environmental quality. In *Environmental Quality, 18th and 19th Annual Reports;* U.S. Government Printing Office: Washington, DC, 1989.
13. Charles, D.F., Ed.; *Acid Rain Research: Do We Have Enough Answers*? Proceedings of a Speciality Conference. Studies in Environmental Science #64, Elsevier: New York, 1995.
14. Kamari, J. *Impact Models to Assess Regional Acidification;* Kluwer Academic Publishers: London, 1990.
15. Charles, D.F., Ed. *Acidic Deposition and Aquatic Ecosystems;* Springer-Verlag: New York, 1991.
16. Mason, B.J. *Acid Rain: Its Causes and Effects on Inland Waters;* Oxford University Press: New York, 1992.

17. U.S. environmental protection agency. Effects of acid rain: human health. In *EPA Environmental Issues Website*. Update June 26, 2001 (accessed July 2001).
18. Marion, G.M.; Hendricks, D.M.; Dutt, G.R.; Fuller, W.H. Aluminum and silica solubility in soils. Soil Science 1976, 121, 76–82.
19. Kennedy, I.R. *Acid Soil and Acid Rain*; Wiley: New York, 1992.
20. Reuss, J.O.; Johnson, D.W. *Acid Deposition and the Acidification of Soils and Waters*; Springer-Verlag: New York, 1986.
21. Likens, G.E.; Driscoll, C.T.; Buso, D.C. Long-term effects of acid rain: response and recovery of a forest ecosystem. Science 1996, 272, 244–246.
22. Linthurst, R.A. *Direct and Indirect Effects of Acidic Deposition on Vegetation*; Butterworth Publishers: Stoneham, MA, 1984.
23. Bush, M.B. *Ecology of a Changing Planet;* Prentice- Hall: Upper Saddle River, NJ, 1997.
24. Alawell, C.; Mitchell, M.J.; Likens, G.E.; Krouse, H.R. Sources of stream sulfate at the hubbard brook experimental forest: long-term analyses using stable isotopes. Biogeochemistry 1999, 44, 281–299.
25. National Atmospheric Deposition Program. *Nitrogen in the Nation's Rain*. NADP Brochure 2000–01a (accessed July 2001).

45

Carbon Sequestration

Nathan E. Hultman

Introduction

The increasing likelihood of human-caused changes in climate could lead to undesirable impacts on ecosystems, economies, and human health and well-being. These potential impacts have prompted extensive assessment of options to reduce the magnitude and rate of future climate changes. Since climate changes are derived ultimately from increases in the concentrations of greenhouse gases (GHGs) in the atmosphere, such options must target either (a) reductions in the rate of inflow of GHGs to the atmosphere or (b) the removal of GHGs from the atmosphere once they have been emitted. Carbon sequestration refers to techniques from both categories that result in the storage of carbon that would otherwise be in the atmosphere as CO_2.

CO_2 is often targeted among the other GHGs because it constitutes the vast majority of GHG emissions by mass and accounts for three-fifths of the total anthropogenic contribution to climate change. Human emissions of CO_2 come primarily from fossil fuel combustion and cement production (80%), and land-use change (20%) that results in the loss of carbon from biomass or soil.

The rate of inflow of GHGs to the atmosphere can be reduced by a number of complementary options. For CO_2, mitigation options aim to displace carbon emissions by preventing the oxidation of biological or fossil carbon. These options include switching to lower-carbon fossil fuels, renewable energy, or nuclear power; using energy more efficiently; and reducing the rate of deforestation and land-use change. On the other hand, sequestration options that reduce emissions involve the capture and storage of carbon before it is released into the atmosphere.

CO_2 can also be removed directly from the atmosphere. While the idea of a large-scale, economically competitive method of technologically "scrubbing" CO_2 from the atmosphere is enticing, such technology currently does not exist. Policy has therefore focused on the biological process of carbon absorption through photosynthesis, either through expanding forested lands or, perhaps, enhancing photosynthesis in the oceans. This entry describes both the technological and biological approaches to carbon sequestration.

Technological Sequestration: Carbon Capture and Storage

The technological process of sequestering CO_2 requires two steps: first, the CO_2 must be separated from the industrial process that would otherwise emit it into the atmosphere; and second, the CO_2 must be stored in a reservoir that will contain it for a reasonable length of time. This process is therefore often referred to as carbon capture and storage (CCS) to distinguish it from the biological carbon sequestration that is described later.

Sources of Carbon

The best sites for CCS are defined by the efficiency of the capture technique, the cost of transport and sequestration, and the quantity of carbon available. The large capital requirements for CCS also dictate that large, fixed industrial sites provide the best opportunities. Therefore, although fossil-fueled transportation represents about 20% of current global CO_2 emissions, this sector presents no direct options for CCS at this time. The industrial sector, on the other hand, produces approximately 60% of current CO_2 emissions; most of these emissions come from large point sources which are ideal for CCS, such as power stations, oil refineries, petrochemical and gas reprocessing plants, and steel and cement works.[1]

Separation and Capture

Carbon capture requires an industrial source of CO_2; different industrial processes create streams with different CO2 concentrations. The technologies applied to capture the CO_2 will therefore vary according to the specific capture process.[2–4] Capture techniques can target one of three sources:

- Post-combustion flue gases
- Pre-combustion capture from gasification from power generation
- Streams of highly pure CO_2 from various industrial processes

Post-Combustion Capture

Conventional combustion of fossil fuels in air produces CO_2 streams with concentrations ranging from about 4 to 14% by volume. The low concentration of CO_2 in flue gas means that compressing and storing it would be uneconomical; therefore, the CO_2 needs to be concentrated before storage. Currently, the favored process for this task is chemical absorption, also known as chemical solvent scrubbing. Cooled and filtered flue gas is fed into an absorption vessel with a chemical solvent that absorbs the CO_2. The most common solvent for this process is monoethanolamine (MEA). The CO_2-rich solvent is then passed to another reaction vessel called a stripper column. It is then heated with steam to reverse the process, thus regenerating the solvent and releasing a stream of CO_2 with a purity greater than 90%.

Scrubbing with MEA and other amine solvents imposes large costs in energy consumption in the regeneration process; it requires large amounts of solvents since they degrade rapidly; and it imposes high equipment costs since the solvents are corrosive in the presence of O_2. Thus, until solvents are improved in these areas, flue gas separation by this method will remain relatively costly: just the steam and electric load from a coal power plant can increase coal consumption by 40% per net kWh_e. Estimates of the financial and efficiency costs from current technology vary. Plant efficiency is estimated to drop from over 40% to a range between 24 and 37%.[2,5,6] For the least efficient systems, carbon would cost up to \$70/t CO_2 and result in an 80% increase in the cost of electricity.[5] Other studies estimate an increase in the cost of electricity of 25%–75% for natural gas combined cycle and Integrated Gasification Combined Cycle (IGCC), and of 60%–115% for pulverized coal.[4] A small number of facilities currently practice flue gas separation with chemical absorption, using the captured CO_2 for urea production, foam blowing, carbonated beverages, and dry ice production.

In addition, several developments may improve the efficiency of chemical absorption.

Several other processes have been proposed for flue-gas separation. Adsorption techniques use solids with high surface areas, such as activated carbon and zeolites, to capture CO_2. When the materials become saturated, they can be regenerated (releasing CO_2) by lowering pressure, raising temperature, or applying a low-voltage electric current. A membrane can be used to concentrate CO_2, but since a single pass through a membrane cannot achieve a great change in concentration, this process requires multiple passes or multiple membranes. An alternative use for membranes is to use them to increase the efficiency of the chemical absorption. In this case, a membrane separating the flue gas from the absorption solvent allows a greater surface area for the reaction, thus reducing the size and energy requirements of the absorption and stripper columns. *Cryogenic* techniques separate CO_2 from other gases by condensing or freezing it. This process requires significant energy inputs and the removal of water vapor before freezing.

One of the main limitations to flue-gas separation is the low pressure and concentration of CO_2 in the exhaust. An entirely different approach to post-combustion capture is to dramatically increase the concentration of CO_2 in the stream by burning the fuel in highly enriched oxygen rather than air. This process, called oxyfuel combustion, produces streams of CO_2 with a purity greater than 90%. The resulting flue gas will also contain some H_2O that can be condensed and removed, and the remaining high-purity CO_2 can be compressed for storage. Though significantly simpler on the exhaust side, this approach requires a high concentration of oxygen for the intake air. While this process alone may consume 15% of a plant's electric output, the separated N_2, Ar, and other trace gases also can be sold to offset some of the cost. Oxyfuel systems can be retrofitted onto existing boilers and furnaces.

Pre-Combustion Capture

Another approach involves removing the carbon from fossil fuels before combustion. First, the fuel is decomposed in the absence of oxygen to form a hydrogen-rich fuel called synthesis gas. Currently, this process of gasification is already in use in ammonia production and several commercial power plants fed by coal and petroleum byproducts; these plants can use lower-purity fuels and the energy costs of generating synthesis gas are offset by the higher combustion efficiencies of gas turbines; such plants are called IGCC plants. Natural gas can be transformed directly by reacting it with steam, producing H_2 and CO_2. While the principle of gasification is the same for all carbonaceous fuels, oil and coal require intermediate steps to purify the synthesis fuel and convert the byproduct CO into CO_2.

Gasification results in synthesis gas that contains 35%–60% CO_2 (by volume) at high pressure (over 20 bar). While current installations feed this resulting mixture into the gas turbines, the CO_2 can also be separated from the gas before combustion. The higher pressure and concentration give a CO_2 partial pressure of up to 50 times greater than in the post-combustion capture of flue gases, which enables another type of separation technique of physical solvent scrubbing. This technique is well known from ammonia production and involves the binding of CO_2 to solvents that release CO_2 in the stripper under lower pressure. Solvents in this category include cold methanol, polyethelene glycol, propylene carbonate, and sulpholane. The resulting separated CO_2 is, however, near atmospheric pressure and requires compression before storage (some CO_2 can be recovered at elevated pressures, which reduces the compression requirement). With current technologies, the total cost of capture for IGCC is estimated to be greater than \$25 per ton of CO_2; plant efficiency is reduced from 43 to 37%, which raises the cost of electricity by over 25%.[5]

Pre-combustion capture techniques are noteworthy not only for their ability to remove CO_2 from fossil fuels for combustion in turbines, but also because the resulting synthesis gas is primarily H_2. They therefore could be an important element of a hydrogen-mediated energy system that favors the higher efficiency reactions of fuel cells over traditional combustion.[7]

Industrial CO_2 Capture

Many industrial processes release streams of CO_2 that are currently vented into the atmosphere. These streams, currently viewed as simple waste in an economically viable process, could therefore provide capture opportunities. Depending on the purity of the waste stream, these could be among the most economical options for CCS. In particular, natural gas processing, ethanol and hydrogen production, and cement manufacturing produce highly concentrated streams of CO_2. Not surprisingly, the first large-scale carbon sequestration program was run from a previously vented stream of CO_2 from the Sleipner gas-processing platform off the Norwegian coast.

Storage of Captured CO_2

Relatively small amounts of captured CO_2 might be re-used in other industrial processes such as beverage carbonation, mineral carbonates, or commodity materials such as ethanol or paraffins. Yet most captured CO_2 will not be re-used and must be stored in a reservoir. The two main routes for storing captured CO_2 are to inject it into geologic formations or into the ocean. However, all reservoirs have some rate of leakage and this rate is often not well known in advance. While the expected length of storage time is important (with targets usually in the 100–1000 year range), we must therefore also be reasonably confident that the reservoir will not leak more quickly than expected, and have appropriate measures to monitor the reservoir over time. Moreover, transporting CO_2 between the point of capture and the point of storage adds to the overall cost of CCS, so the selection of a storage site must account for this distance as well.

Geologic Sequestration

Geologic reservoirs—in the form of depleted oil and gas reservoirs, unmineable coal seams, and saline formations— comprise one of the primary sinks for captured CO_2. Estimates of total storage capacity in geologic reservoirs could be up to 500% of total emissions to 2050 (Table 1).

Captured CO_2 can be injected into depleted oil and gas reservoirs, or can be used as a means to enhance oil recovery from reservoirs nearing depletion. Because they held their deposits for millions of years before extraction, these reservoirs are expected to provide reliable storage for CO_2. Storage in depleted reservoirs has been practiced for years for a mixture of petroleum mining waste gases called "acid gas."

A petroleum reservoir is never emptied of all its oil; rather, extracting additional oil just becomes too costly to justify at market rates. An economically attractive possibility is therefore using captured

TABLE 1 CO_2 Reservoirs. Carbon Dioxide Storage Capacity Estimates. *E* Is Defined as the Total Global CO_2 Emissions from the Years 2000–2050 in IPCC's Business-as-Usual Scenario IS92A. Capacity Estimates Such as These Are Rough Guidelines Only and Actual Utilization Will Depend on Carbon Economics

	Storage Capacity	
Reservoir Type	Billion Tonnes CO_2	% of E
Coal basins	170	8%
Depleted oil reservoirs	120	6%
Gas basins	700	37%
Saline formations		
Terrestrial		276%
Off-shore		192%
Total	10.490	517%

Source: Dooley and Friedman.[8]

CO_2 to simultaneously increase the yield from a reservoir as it is pumped into the reservoir for storage. This process is called enhanced oil recovery. Standard oil recovery yields only about 30%–40% of the original petroleum stock. Drilling companies have years of experience with using compressed CO_2, a hydrocarbon solvent, to obtain an additional 10%–15% of the petroleum stock. Thus, captured CO_2 can be used to provide a direct economic benefit along with its placement in a reservoir. This benefit can be used to offset capture costs.

Coal deposits that are not economically viable because of their geologic characteristics provide another storage option. CO_2 pumped into these unmineable coal seams will adsorb onto the coal surface. Moreover, since the coal surface prefers to adsorb CO_2 to methane, injecting CO_2 into coal seams will liberate any coal bed methane (CBM) that can then be extracted and sold. This enhanced methane recovery is currently used in U.S. methane production, accounting for about 8% in 2002. Such recovery can be used to offset capture costs. One potential problem with this method is that the coal, as it adsorbs CO_2, tends to swell slightly. This swelling closes pore spaces and thus decreases rock permeability, which restricts both the reservoir for incoming CO_2 and the ability to extract additional CBM.

Saline formations are layers of porous sedimentary rock (e.g., sandstone) saturated with saltwater, and exist both under land and under the ocean. These layers offer potentially large storage capacity representing several hundred years' worth of CO_2 storage. However, experience with such formations is much more limited and thus the uncertainty about their long-term viability remains high. Moreover, unlike EOR or CBM recovery with CO_2, injecting CO_2 into saline formations produces no other commodity or benefit that can offset the cost. On the other hand, their high capacity and relative ubiquity makes them attractive options in some cases. Statoil's Sleipner project, for example, uses a saline aquifer for storage.

Research and experimentation with saline formations is still in early stages. To achieve the largest storage capacities, CO_2 must be injected below 800 m depth, where it will remain in a liquid or supercritical dense phase (supercritical point at 31°C, 71 bar). At these conditions, CO_2 will be buoyant (a density of approximately 600–800 kg/m^3) and will tend to move upward. The saline formations must therefore either be capped by a less porous layer or geologic trap to prevent leakage of the CO_2 and eventual decompression.[9] Over time, the injected CO_2 will dissolve into the brine and this mixture will tend to sink within the aquifer. Also, some saline formations exist in rock that contains Ca-, Mg-, and Fe-containing silicates that can form solid carbonates with the injected CO_2. The resulting storage as rock is highly reliable, though it may also hinder further injection by closing pore spaces. Legal questions may arise when saline formations, which are often geographically extensive, cross national boundaries or onto marine commons.

Ocean Direct Injection

As an alternative to geologic storage, captured CO_2 could be injected directly into the ocean at either intermediate or deep levels. The oceans have a very large potential for storing CO_2, equivalent to that of saline aquifers (~10^3 Gt). While the ocean's surface is close to equilibrium with atmospheric carbon dioxide concentrations, the deep ocean is not because the turnover time of the oceans is much slower (~5000 years) than the observed increases in atmospheric CO_2. Since the ocean will eventually absorb much of the atmospheric perturbation, injecting captured CO_2 into the oceans can therefore be seen as simply bypassing the atmospheric step and avoiding the associated climate consequences. Yet little is known about the process or effects—either ecological or geophysical—of introducing large quantities of CO_2 into oceanic water.

At intermediate depths (between 500 and 3000 m), CO_2 exists as a slightly buoyant liquid. At these depths, a stream of CO_2 could be injected via a pipe affixed either to ship or shore. The CO_2 would form a droplet plume, and these droplets would slowly dissolve into the seawater, disappearing completely before reaching the surface. Depressed pH values are expected to exist for tens of km downcurrent of

TABLE 2 Additional Costs to Power Generation from CCS. Approximate Capture and Storage Costs for Different Approaches to Power Plant Sequestration

Fossil Type	Cost of CCS ¢ per kWh
Natural gas combined cycle	1–2
Pulverized coal	2–3
Coal IGCC	2–4

Source: Herzog and Golomb,[4] National Energy Technology Laboratory,[5] Dooley et al.,[11] and Freund and Davison.[12]

the injection site, though changing the rate of injection can moderate the degree of perturbation. In addition, pulverized limestone could be added to the injected CO_2 to buffer the acidity.

Below 3000 m, CO_2 becomes denser than seawater and would descend to the seafloor and pool there. Unlike intermediate injection, therefore, this method does not lead to immediate CO_2 dissolution in oceanic water; rather, the CO_2 is expected to dissolve into the ocean at a rate of about 0.1 m/y. Deep injection thus minimizes the rate of leakage to the surface, but could still have severe impacts on bottom-dwelling sea life.

The primary obstacles to oceanic sequestration are not technical but relate rather to this question of environmental impacts.[10] Oceanic carbon storage might affect marine ecosystems through the direct effects of a lower environmental pH; dissolution of carbonates on fauna with calcareous structures and microflora in calcareous sediments; impurities such as sulfur oxides, nitrogen oxides, and metals in the captured CO_2; smothering effects (deep injection only); and changes in speciation of metals and ammonia due to changes in pH. Few of these possibilities have been studied in sufficient detail to allow an informed risk assessment. In addition, the legality of dumping large quantities of CO_2 into the open ocean remains murky.

Overall Costs of CCS

The costs of CCS can be measured either as a cost per tonne of CO_2, or, for power generation, a change in the cost of electricity (Table 2). The total cost depends on the cost of capture, transport, and storage. Capture cost is mainly a function of parasitic energy losses and the capital cost of equipment. Transport cost depends on distance and terrain. Storage costs vary depending on the reservoir but are currently a few dollars per tonne of CO_2. The variety of approaches to CCS and the early stages of development make precise estimates of cost difficult, but current technology spans about \$25–\$85/t CO_2.

Biological Sequestration: Enhancing Natural Carbon Sinks

The previous sections have described processes by which CO_2 could be technologically captured and then stored. Photosynthesis provides an alternate route to capture and store carbon. Enhancing this biological process is therefore an alternative method of achieving lower atmospheric CO_2 concentrations by absorbing it directly from the air.

Terrestrial Carbon Sinks

Carbon sequestration in terrestrial ecosystems involves enhancing the natural sinks for carbon fixed in photosynthesis. This occurs by expanding the extent of ecosystems with a higher steady-state density of carbon per unit of land area. For example, because mature forest ecosystems contain more carbon per hectare than grasslands, expanding forested areas will result in higher terrestrial carbon storage.

Another approach is to encourage the additional storage of carbon in agricultural soils. The essential element in any successful sink enhancement program is to ensure that the fixed carbon remains in pools with long lives.

Afforestation involves planting trees on unforested or deforested land.[13,14] The most likely regions for forest carbon sequestration are Central and South America and Southeast Asia because of relatively high forest growth rates, available land, and inexpensive labor. However, the translation of forestry activities into a policy framework is complex. Monitoring the carbon changes in a forest is difficult over large areas, as it requires not only a survey of the canopy and understory, but also an estimate of the below-ground biomass and soil carbon. Some groups have voiced concern over the potential for disruption of social structures in targeted regions.

Soil carbon sequestration involves increasing soil carbon stocks through changes in agriculture, forestry, and other land use practices. These practices include mulch farming, conservation tillage, agroforestry and diverse cropping, cover crops, and nutrient management that integrates manure, compost, and improved grazing. Such practices, which offer the lowest-cost carbon sequestration, can have other positive effects such as soil and water conservation, improved soil structure, and enhanced soil fauna diversity. Rates of soil carbon sequestration depend on the soil type and local climate, and can be up to 1000 kg of carbon per hectare per year. Management practices can enhance sequestration for 20–50 years, and sequestration rates taper off toward maturity as the soil carbon pool becomes saturated. Widespread application of recommended management practices could offset 0.4 to 1.2 GtC/y, or 5%–15% of current global emissions.[15]

If sinks projects are to receive carbon credits under emissions trading schemes like that in the Kyoto Protocol, they must demonstrate that the project sequestered more carbon than a hypothetical baseline or business-as-usual case. They must also ensure that the carbon will remain in place for a reasonable length of time, and guard against simply displacing the baseline activity to a new location.

Ocean Fertilization

Vast regions of the open ocean have very little photosynthetic activity, though sunlight and major nutrients are abundant. In these regions, phytoplankton are often deprived of trace nutrients such as iron. Seeding the ocean surface with iron, therefore, might produce large phytoplankton blooms that absorb CO_2. As the plankton die, they will slowly sink to the bottom of the ocean, acting to transport the fixed carbon to a permanent burial in the seafloor. While some experimental evidence indicates this process may work on a limited scale, little is known about the ecosystem effects and potential size of the reservoir.[16]

Prospects for Carbon Sequestration

Carbon sequestration techniques—both technological and biological—are elements of a portfolio of options for addressing climate change. Current approaches hold some promise for tapping into the geologic, biologic, and oceanic potential for storing carbon. The costs of some approaches, especially the improved management of agricultural and forest lands, are moderate (Table 3). Yet these opportunities are not infinite and additional options will be necessary to address rising global emissions. Thus, the higher costs of current technological approaches are likely to drop with increasing deployment and changing market rates for carbon.

Possible developments include advanced CO_2 capture techniques focusing on membranes, ionic (organic salt) liquids, and microporous metal organic frameworks. Several alternative, but still experimental, sequestration approaches have also been suggested. Mineralization could convert CO_2 to stable minerals. This approach seeks, therefore, to hasten what in nature is a slow but exothermic weathering process that operates on common minerals like olivine, forsterite, or serpentines (e.g., through selected sonic frequencies). It is possible that CO_2 could be injected in sub-seafloor carbonates. Chemical looping

TABLE 3 Costs of Carbon Sequestration. Estimates for Sequestration Costs Vary Widely. Future Costs Will Depend on Rates of Technological Change

	Cost
Sequestration Technique	\$ per T CO_2
Carbon capture and storage	26–84
Tree planting and agroforestry	10–210
Soil carbon sequestration	6–24

Source: Herzog and Golomb,[4] National Energy Technology Laboratory,[5] Williams[7] Dooley et al.[11] Freund and Davison,[12] Van Kooten,[13] and Richards and Stokes.[14]

describes a method for combusting fuels with oxygen delivered by a redox agent instead of by air or purified oxygen; it promises high efficiencies of energy conversion and a highly enriched CO_2 exhaust stream. Research also continues on microbial CO_2 conversion in which strains of microbes might be created to metabolize CO_2 to produce saleable commodities (succinic, malic, and fumeric acids). In addition, the nascent science of monitoring and verifying the storage of CO_2 will be an important element toward improving technical performance and public acceptance of sequestration techniques.

Acknowledgments

The author thanks the anonymous reviewer for helpful comments on an earlier draft.

References

1. Gale, J. Overview of CO_2 Emission Sources, Potential, Transport, and Geographical Distribution of Storage Possibilities. In *Proceedings of IPCC workshop on carbon dioxide capture and storage* Regina, Canada Nov 18–21, 2002 15–29. http://arch.rivm.nl/env/int/ipcc/pages_media/ ccs2002.html (accessed April 2005). A revised and updated version of the papers presented at the Regina workshop is now available. See, Metz, B.; Davidson, O.; de Coninck, H.; Loos, M.; Meyer, L., Special Report on Carbon Dioxide Capture and Storage. Intergovernmental Panel on Climate Change: Geneva, http://www.ipcc.ch. (accessed May 2006), 2006.
2. Thambimuthu, K.; Davison, J.; Gupta, M. CO_2 Capture and Reuse, In *Proceedings of IPCC workshop on carbon dioxide capture and storage,* Regina, Canada, Nov 18–21, 2002 31–52 http://arch.rivm.nl/env/int/ipcc/pages_media/ccs2002.html (accessed April 2005).
3. Gottlicher, G. *The Energetics of Carbon Dioxide Capture in Power Plants;* U.S. Department of Energy Office of Fossil Energy Washington, DC, 2004; 193
4. Herzog, H.; Golomb, D. Carbon capture and storage from fossil fuel use *Encyclopedia of Energy* Cleveland, C.J. Ed.; Elsevier Science: New York, 2004; 277–287.
5. National Energy Technology Laboratory *Carbon Sequestration: Technology Roadmap and Program Plan* U.S. Department of Energy Washington, DC, 2004.
6. Gibbins, J.R.; Crane, R.I.; Lambropoulos, D.; Booth, C.; Roberts, C.A.; Lord, M. Maximizing the Effectiveness of Post Combustion CO_2 Capture Systems In *Proceedings of the 7th International Conference on Greenhouse Gas Abatement,* Vancouver, Canada, Sept 5–9, 2004 Document E2-2. http://www.ghgt7.ca (accessed June 2004).
7. Williams, R.H. Decarbonized Fossil Energy Carriers and Their Energy Technology Competitors In *Proceedings of IPCC workshop on carbon dioxide capture and storage* Regina, Canada, November, 2002; 119–135, http://arch.rivm.nl/env/int/ipcc/pages_media/ccs2002.html (accessed April 2005).
8. Dooley, J.J.; Friedman, S.J. *A Regionally Disaggregated Global Accounting of CO_2 Storage Capacity: Data and* Assumptions Battelle National Laboratory Washington DC, 2004; 15.

9. Kârsted, O. Geological Storage, Including Costs and Risks, in Saline Aquifers In *Proceedings of IPCC workshop on carbon dioxide capture and storage* Regina, Canada 2002 Nov 1–-21 53–60 http://arch.rivm.nl/env/int/ipcc/pages_media/ccs2002.html (accessed April 2005).
10. Johnston, P.; Santillo, D. Carbon Capture and Sequestration: Potential Environmental Impacts, In *Proceedings of IPCC workshop on carbon dioxide capture and storage*, Regina, Canada, Nov 18–21, 2002; 95–110. http://arch.rivm.nl/env/int/ipcc/pages_media/ccs2002.html (accessed April 2005).
11. Dooley, J.J.; Edmonds, J.A.; Dahowski, R.T.; Wise, M.A. Modeling Carbon Capture and Storage Technologies in Energy and Economic Models. In *Proceedings of IPCC workshop on carbon dioxide capture and storage,* Regina, Canada, Nov 18–21, 2002; 161–172. http://arch.rivm.nl/env/int/ipcc/pages_media/ccs2002.html (accessed April 2005).
12. Freund, P.; Davison, J. General Overview of Costs. In *Proceedings of IPCC workshop on carbon dioxide capture and storage,* Regina, Canada, Nov 18–21, 2002 79–93. http:// arch.rivm.nl/env/int/ipcc/pages_media/ccs2002.html (accessed April 2005).
13. Van Kooten, G.C.; Eagle, A.J.; Manley, J.; Smolak, T. How costly are carbon offsets? A meta-analysis of carbon forest sinks. Environ. Sci. Policy **2004,** 7, 239–251.
14. Richards, K.R.; Stokes, C. A review of forest carbon sequestration cost studies: a dozen years of research. Climatic Change **2004,** *68,* 1–48.
15. Lal, R. Soil carbon sequestration impacts on global climate change and food security. Science **2004,** *304,* 1623–1627.
16. Buesseler, K.O.; Andrews, J.E.; Pike, S.M.; Charette, M.A. The effects of iron fertilization on carbon sequestration in the southern ocean. Science **2004,** *304,* 414–417.

46

Energy Conservation

Ibrahim Dincer
and Adnan Midili

Introduction

Civilization began when people found out how to use fire extensively. They burned wood and obtained sufficiently high temperatures for melting metals, extracting chemicals, and converting heat into mechanical power, as well as for cooking and heating. During burning, the carbon in wood combines with O_2 to form carbon dioxide (CO_2), which then is absorbed by plants and converted back to carbon for use as a fuel again. Because wood was unable to meet the fuel demand, the Industrial Revolution began with the use of fossil fuels (e.g., oil, coal, and gas). Using such fuels has increased the CO_2 concentration in the air, leading to the beginning of global warming. Despite several warnings in the past about the risks of greenhouse- gas emissions, significant actions to reduce environmental pollution were not taken, and now many researchers have concluded that global warming is occurring. During the past two decades, the public has became more aware, and researchers and policymakers have focused on this and related issues by considering energy, the environment, and sustainable development.

Energy is considered to be a key catalyst in the generation of wealth and also a significant component in social, industrial, technological, economic, and sustainable development. This makes energy resources and their use extremely significant for every country. In fact, abundant and affordable energy is one of the great boons of modern industrial civilization and the basis of our living standard. It makes people's lives brighter, safer, more comfortable, and more mobile, depending on their energy demand and consumption. In recent years, however, energy use and associated greenhouse-gas emissions and their potential effects on the global climate change have been of worldwide concern.

Problems with energy utilization are related not only to global warming, but also to such environmental concerns as air pollution, acid rain, and stratospheric ozone depletion. These issues must be taken into consideration simultaneously if humanity is to achieve a bright energy future with minimal environmental impact. Because all energy resources lead to some environmental impact, it is reasonable to suggest that some (not all) of these concerns can be overcome in part through energy conservation efforts.

Energy conservation is a key element of energy policy and appears to be one of the most effective ways to improve end-use energy efficiency, and to reduce energy consumption and greenhouse-gas emissions in various sectors (industrial, residential, transportation, etc.). This is why many countries have recently started developing aggressive energy conservation programs to reduce the energy intensity of the their infrastructures, make businesses more competitive, and allow consumers to save money and to live more comfortably. In general, energy conservation programs aim to reduce the need for new generation or transmission capacity, to save energy, and to improve the environment. Furthermore, energy conservation is vital for sustainable development and should be implemented by all possible means, despite the fact that it has its own limitations. This is required not only for us, but for the next generation as well.

Considering these important contributions, the energy conservation phenomenon should be discussed in a comprehensive perspective. Therefore, the main objective of this entry is to present and discuss the world's primary energy consumption and production; major environmental problems; potential solutions to these issues; practical energy conservation aspects; research and development (R&D) in energy conservation, energy conservation, and sustainable development; energy conservation implementation plans; energy conservation measurements; and life-cycle costing (LCC) as an excellent tool in energy conservation. In this regard, this contribution aims to:

- Help explain main concepts and issues about energy conservation
- Develop relations between energy conservation and sustainability
- Encourage energy conservation strategies and policies
- Provide energy conservation methodologies
- Discuss relations between energy conservation and environmental impact
- Present some illustrative examples to state the importance of energy conservation and its practical benefits

In summary, this book contribution highlights the current environmental issues and potential solutions to these issues; identifies the main steps for implementing energy conservation programs and the main barriers to such implementations; and provides assessments for energy conservation potentials for countries, as well as various practical and environmental aspects of energy conservation.

World Energy Resources: Production and Consumption

World energy consumption and production are very important for energy conservation in the future. Economic activity and investment patterns in the global energy sector are still centered on fossil fuels, and fossil-fuel industries and energy-intensive industries generally have been skeptical about warnings of global warming and, in particular, about policies to combat it. The increase of energy consumption and energy demand indicates our dependence on fossil fuels. If the increase of fossil-fuel utilization continues in this manner, it is likely that the world will be affected by many problems due to fossil fuels. It follows from basic scientific laws that increasing amounts of CO_2 and other greenhouse gases will affect the global climate. The informed debate is not about the existence of such effects, but about their magnitudes and seriousness. At present, the concentration of CO_2 is approximately 30% higher than its preindustrial level, and scientists have already been able to observe a discernible human influence on the global climate.[1]

In the past, fossil fuels were a major alternative for overcoming world energy problems. Fossil fuels cannot continue indefinitely as the principal energy sources, however, due to the rapid increase of world energy demand and energy consumption. The utilization distribution of fossil-fuel types has changed significantly over the past 80 years. In 1925, 80% of the required energy was supplied from coal, whereas in the past few decades, 45% came from petroleum, 25% from natural gas, and 30% from coal. Due to world population growth and the advance of technologies that depend on fossil fuels, reserves of those fuels eventually will not be able to meet energy demand. Energy experts point out that reserves are less than 40 years for petroleum, 60 years for natural gas, and 250 years for coal.[2] Thus, fossil-fuel costs are likely to increase in the near future. This will allow the use of renewable energy sources such as solar, wind, and hydrogen. As an example, the actual data[3,4] and projections of world energy production and consumption from 1980 to 2030 are displayed in the following figures, and the curve equations for world energy production and consumption are derived as shown Table 1.

As presented in Figures 1 and 2, and in Table 2, the quantities of world primary energy production and consumption are expected to reach 14,499.2 and 13,466.5 Mtoe, respectively, by 2030. World population is now over six billion, double that of 40 years ago, and it is likely to double again by the middle of

TABLE 1 World Energy Production and Consumption Models through Statistical Analysis

Energy	Production (Mtoe)	Correlation Coefficient Coefficient	Consumption (Mtoe)	Correlation Coefficient
World primary	= 148.70 × Year − 287,369	0.998	= 139.62 × Year − 269,953	0.998
World oil	= 44.47 × Year − 85,374	0.997	= 42.18×Year − 80,840	0.998
World coal	= 20.05 × Year − 37,748	0.946	= 23.18× Year − 43,973	0.968
World NG	= 45.73 × Year − 89,257	0.999	= 46.27 × Year − 90,347	0.999

Note: Mtoe, million tons of oil equivalent.

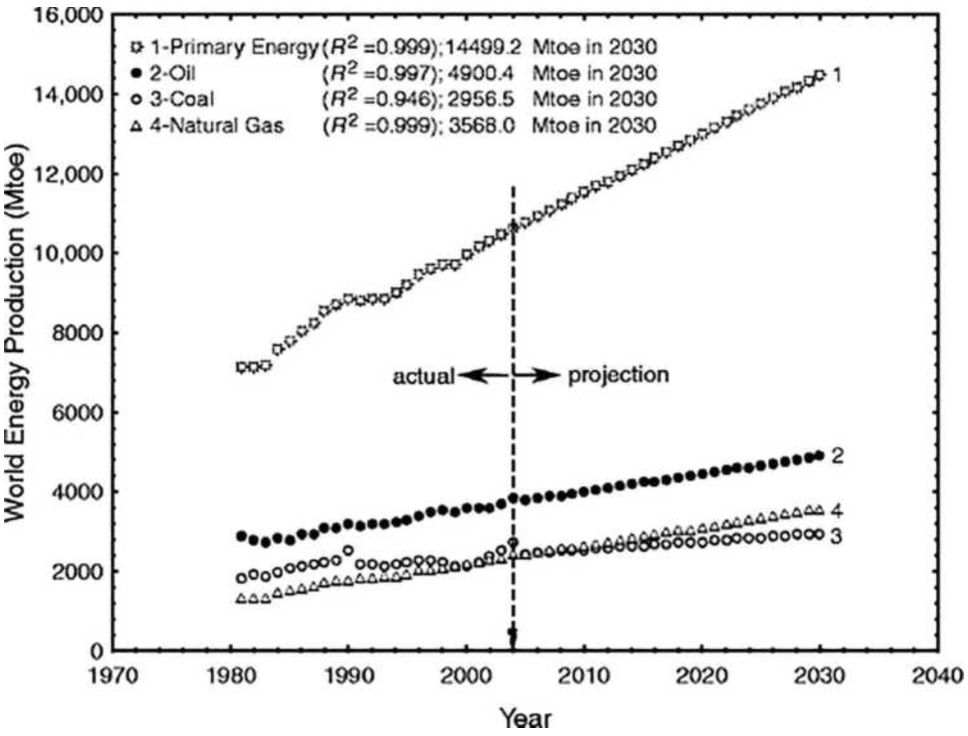


FIGURE 1 Variation of actual data taken from BP[3] and IEE,[4] and projections of annual world energy production. Mtoe, million tons of oil equivalent.
Source: BP[3] and IEE.[4]

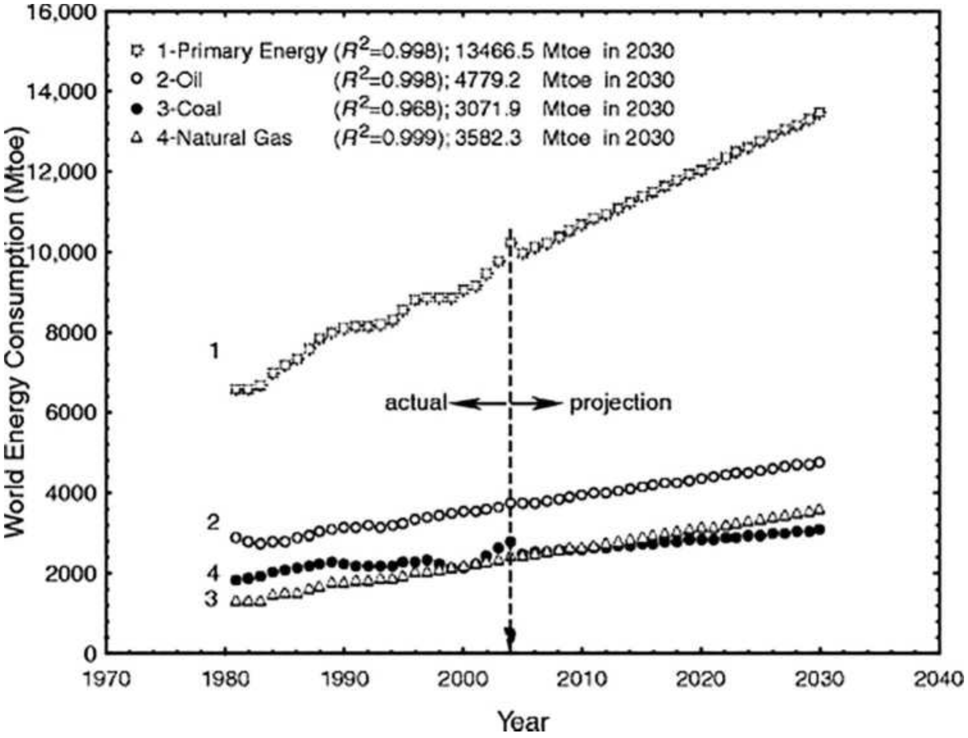


FIGURE 2 Variation of actual data taken from BP[3] and IEE,[4] and projections of annual world energy consumption. Mtoe, million tons of oil equivalent.
Source: BP[3] and IEE.[4]

TABLE 2 Some Extracted Values of World Primary and Fossil Energy Production and Consumption

Year	Primary Energy Production (Mtoe)	Primary Energy Consumption (Mtoe)	Oil Prod. (Mtoe)	Oil Cons. (Mtoe)	Coal Prod. (Mtoe)	Coal Cons. (Mtoe)	NG Prod. (Mtoe)	NG Cons. (Mtoe)
1994	8,996.9	8,310.1	3237.1	3204.4	2178.1	2185.5	1891.2	1876.7
2000	9,981.9	9,079.8	3614.0	3538.7	2112.4	2148.1	2189.9	2194.5
2006	10,930.3	10,115.7	3833.1	3767.0	2475.3	2515.7	2470.5	2471.8
2012	11,822.5	10,953.4	4099.9	4020.0	2595.6	2654.8	2744.9	2749.5
2018	12,714.7	11,791.1	4366.7	4273.1	2715.9	2793.8	3019.2	3027.1
2024	13,606.9	12,128.8	4633.5	4526.1	2836.2	2932.9	3293.6	3304.7
2030	14,499.2	13,466.5	4900.4	4779.2	2956.5	3071.9	3568.0	3582.3

Note: Mtoe, million tons of oil equivalent.

the 21st century. The world's population is expected to rise to about seven billion by 2010. Even if birth rates fall so that the world population becomes stable by 2030, the population still would be about ten billion. The data presented in Figures 1 and 2 are expected to cover current energy needs provided that the population remains constant. Because the population is expected to increase dramatically, however, conventional energy resource shortages are likely to occur, due to insufficient fossil-fuel resources. Therefore, energy conservation will become increasingly important to compensate for shortages of conventional resources.

Major Environmental Problems

One of the most important targets of modern industrial civilizations is to supply sustainable energy sources and to develop the basis of living standards based on these energy sources, as well as implementing energy conservation measures. In fact, affordable and abundant sustainable energy makes our lives brighter, safer, more comfortable, and more mobile because most industrialized and developing societies use various types of energy. Billions of people in undeveloped countries, however, still have limited access to energy. India's per-capita consumption of electricity, for example, is one-twentieth that of the United States. Hundreds of millions of Indians live "off the grid"- that is, without electricity-and cow dung is still a major fuel for household cooking. This continuing reliance on such preindustrial energy sources is also one of the major causes of environmental degradation.[5]

After many decades of using fossil fuels as a main energy source, significant environmental effects of fossil fuels became apparent. The essential pollutants were from greenhouse gases (e.g., CO_2, SO_2, and NO_2). Fossil fuels are used for many applications, including industry, residential, and commercial sectors. Increasing fossil-fuel utilization in transportation vehicles such as automobiles, ships, aircrafts, and spacecrafts has led to increasing pollution. Gas, particulate matter, and dust clouds in the atmosphere absorb a significant portion of the solar radiation directed at Earth and cause a decrease in the oxygen available for the living things. The threat of global warming has been attributed to fossil fuels.[2] In addition, the risk and reality of environmental degradation have become more apparent. Growing evidence of environmental problems is due to a combination of factors.

During the past two decades, environmental degradation has grown dramatically because of the sheer increase of world population, energy consumption, and industrial activities. Throughout the 1970s, most environmental analysis and legal control instruments concentrated on conventional pollutants such as SO_2, NO_x, particulates, and CO. Recently, environmental concern has extended to the control of micro or hazardous air pollutants, which are usually toxic chemical substances and harmful in small doses, as well as to that of globally significant pollutants such as CO_2. Aside from advances in environmental engineering science, developments in industrial processes and structures have led to new environmental problems.[6,7] In the energy sector, for example, major shifts to the road transport of industrial goods and to individual travel by cars has led to an increase in road traffic and, hence, to a shift in attention paid to the effects and sources of NO_x and to the emissions of volatile organic compounds (VOC). In fact, problems with energy supply and use are related not only to global warming, but also to such environmental concerns as air pollution, ozone depletion, forest destruction, and emission of radioactive substances. These issues must be taken into consideration simultaneously if humanity is to achieve a bright energy future with minimal environmental impact. Much evidence exists to suggest that the future will be negatively impacted if humans keep degrading the environment. Therefore, there is an intimate connection among energy conservation, the environment, and sustainable development. A society seeking sustainable development ideally must utilize only energy resources that cause no environmental impact (e.g., that release no emissions to the environment). Because all energy resources lead to some environmental impact, however, it is reasonable to suggest that some (not all) of the concerns regarding the limitations imposed on sustainable development by environmental emissions and their negative impacts can be overcome in part through energy conservation. A strong relation clearly exists between energy conservation and environmental impact, because for the same services or products, less resource utilization and pollution normally are associated with higher-efficiency processes.[8]

Table 3 summarizes the major environmental problems such as acid rain, stratospheric ozone depletion, and global climate change (greenhouse effect)-and their main sources and effects.

As shown in Figure 3, the world total CO_2 production is estimated to be 18,313.13 million tons in 1980, 25,586.7 million tons in 2006, 27,356.43 million tons in 2012, and 29,716.1 million tons in 2020 whereas fossil-fuel consumption is found to be 6092.2 million tons in 1980, 8754.5 million tons in 2006,

TABLE 3 Major Environmental Issues and Their Consequences

Issues	Description	Main Sources	Main Effects
Acid precipitation	Transportation and deposition of acids produced by fossil-fuel combustion (e.g., industrial boilers, transportation vehicles) over great distances through the atmosphere via precipitation on the earth on ecosystems	Emissions of SO_2, NO_x and volatile organic compounds (VOCs) (e.g., residential heating and industrial energy use account for 80% of SO_2 emissions)	Acidification of lakes, streams and ground waters, resulting in damage to fish and aquatic life; damage to forests and agricultural crops; and deterioration of materials, e.g., buildings, structures
Stratospheric ozone depletion	Distortion and regional depletion of stratospheric ozone layer though energy activities (e.g., refrigeration, fertilizers)	Emissions of CFCs, halons (chlorinated and brominated organic compounds) and N_2O (e.g. fossil fuel and biomass combustion account for 65%–75% of N_2O emissions)	Increased levels of damaging ultraviolet radiation reaching the ground, causing increased rates of skin cancer, eye damage and other harm to many biological species
Greenhouse effect	A rise in the earth's temperature as a result of the greenhouse gases	Emissions of carbon dioxide (CO_2), CH_4, CFCs, halons, N_2O, ozone and peroxyacetylnitrate (e.g., CO_2 releases from fossil fuel combustion (w50% from CO_2), CH_4 emissions from increased human activity)	Increased the earth's surface temperature about 0.68C over the last century and as a consequence risen sea level about 20 cm (in the next century by another 28C–48C and a rise between 30 and 60 cm); resulting in flooding of coastal settlements, a displacement of fertile zones for agriculture and food production toward higher latitudes, and a decreasing availability of fresh water for irrigation and other essential uses

Source: Dincer,[9] Dincer,[10] and Dincer.[11]

9424.3 million tons in 2012, and 10,317.2 million tons in 2020. These values show that the CO_2 production will probably increase if we continue utilizing fossil fuel. Therefore, it is suggested that certain energy conversion strategies and technologies should be put into practice immediately to reduce future environmental problems.

The climate technology initiative (CTI) is a cooperative effort by 23 Organization for Economic Cooperation and Development (OECD)/International Energy Agency (IEA) member countries and the European Commission to support the objectives of the united nations framework convention on climate change (UNFCCC). The CTI was launched at the 1995 Berlin Conference of the Parties to the UNFCCC. The CTI seeks to ensure that technologies to address climate change are available and can be deployed efficiently. The CTI includes activities directed at the achievement of seven broad objectives:

- To facilitate cooperative and voluntary actions among governments, quasigovernments, and private entities to help cost-effective technology diffusion and reduce the barriers to an enhanced use of climate-friendly technologies
- To promote the development of technology aspects of national plans and programs prepared under the UNFCCC
- To establish and strengthen the networks among renewable and energy efficiency centers in different regions
- To improve access to and enhance markets for emerging technologies
- To provide appropriate recognition of climate-friendly technologies through the creation of international technology awards

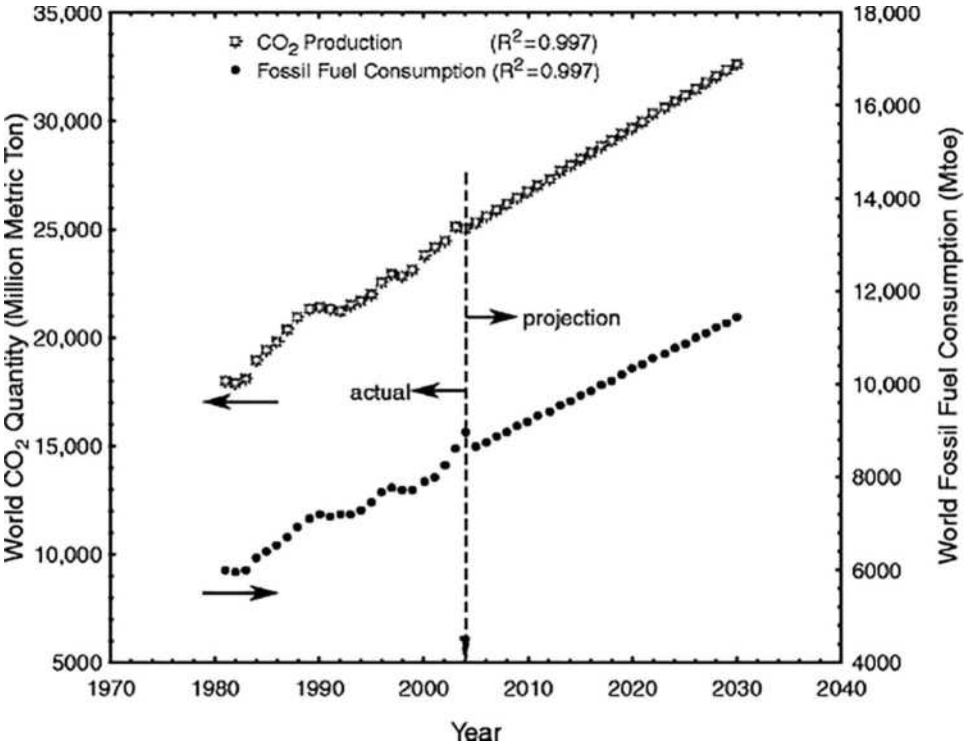


FIGURE 3 Variation of world total fossil-fuel consumption and CO_2 production; actual data from BP[3] and projections. Mtoe, million tons of oil equivalent.
Source: BP.[3]

- To strengthen international collaboration on short-, medium-, and long-term research; development and demonstration; and systematic evaluation of technology options
- To assess the feasibility of developing longer-term technologies to capture, remove, or dispose of greenhouse gases; to produce hydrogen from fossil fuels; and to strengthen relevant basic and applied research

Potential Solutions to Environmental Issues

Although there are a large number of practical solutions to environmental problems, three potential solutions are given priority, as follows[11]:

- Energy conservation technologies (efficient energy utilization)
- Renewable energy technologies
- Cleaner technologies

In these technologies, we pay special attention to energy conservation technologies and their practical aspects and environmental impacts. Each of these technologies is of great importance, and requires careful treatment and program development. In this work, we deal with energy conservation technologies and strategies in depth. Considering the above priorities to environmental solutions, the important technologies shown in Figure 4 should be put into practice.

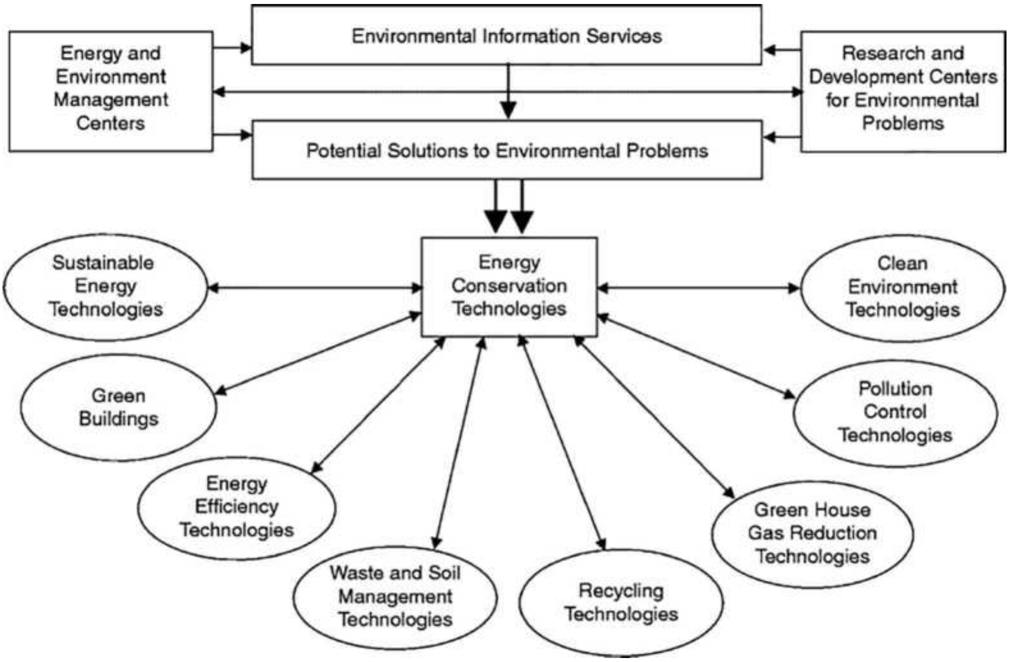


FIGURE 4 Linkages between possible environmental and energy conservation technologies.

Practical Energy Conservation Aspects

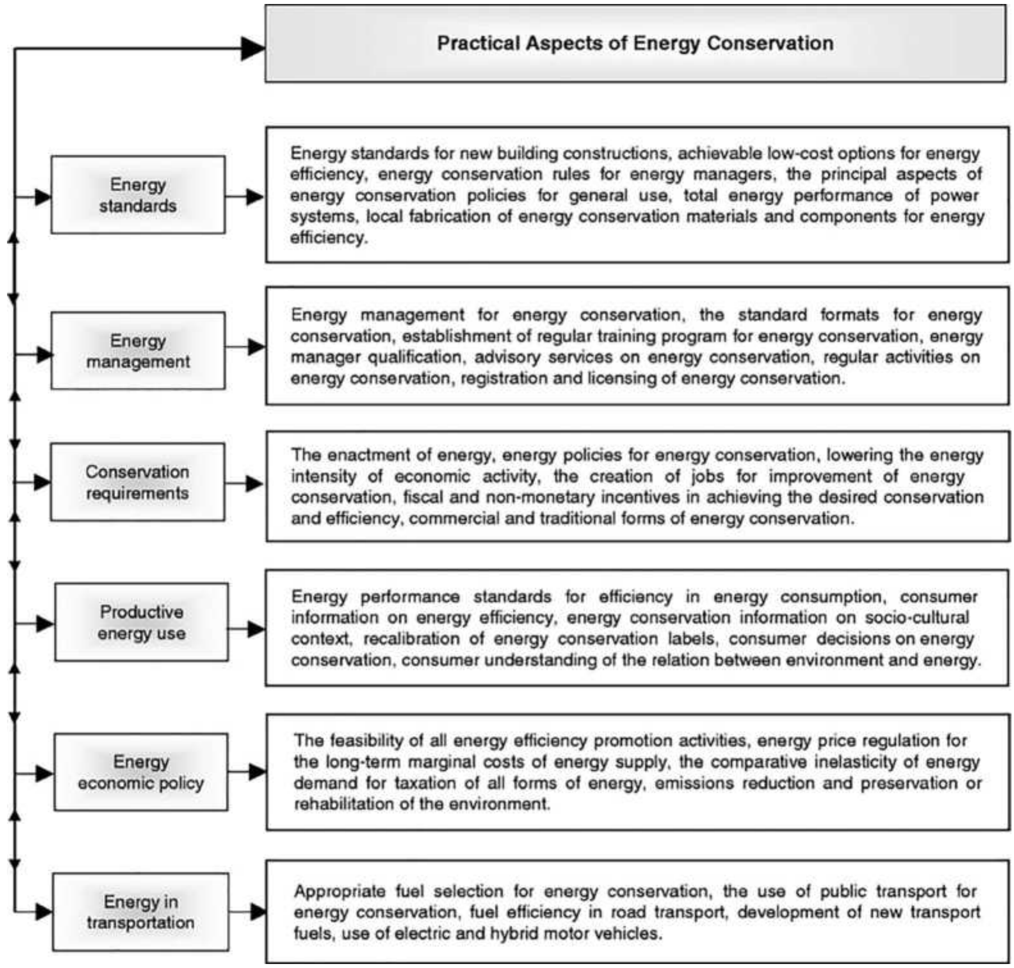
The energy-saving result of efficiency improvements is often called energy conservation. The terms efficiency and conservation contrast with curtailment, which decreases output (e.g., turning down the thermostat) or services (e.g., driving less) to curb energy use. That is, energy curtailment occurs when saving energy causes a reduction in services or sacrifice of comfort. Curtailment is often employed as an emergency measure. Energy efficiency is increased when an energy conversion device-such as a household appliance, automobile engine, or steam turbine-undergoes a technical change that enables it to provide the same service (lighting, heating, motor drive, etc.) while using less energy. Energy efficiency is often viewed as a resource option like coal, oil, or natural gas. In contrast to supply options, however, the downward pressure on energy prices created by energy efficiency comes from demand reductions instead of increased supply. As a result, energy efficiency can reduce resource use and environmental impacts;[12]

The quality of a country's energy supply and demand systems is increasingly evaluated today in terms of its environmental sustainability. Fossil-fuel resources will not last indefinitely, and the most convenient, versatile, and inexpensive of them have substantially been used up. The future role of nuclear energy is uncertain, and global environmental concerns call for immediate action. OECD countries account for almost 50% of total world energy consumption: Current use of oil per person averages 4.5 bbl a year worldwide, ranging from 24 bbl in the United States and 12 bbl in western Europe to less than 1 bbl in sub-Saharan Africa. More than 80% of worldwide CO_2 emissions originate in the OECD area. It is clear, then, that OECD countries should play a crucial role in indicating a sustainable pattern and in implementing innovative strategies.[11]

From an economic as well as an environmental perspective, energy conservation holds even greater promise than renewable energy, at least in the near-term future. Energy conservation is indisputably beneficial to the environment, as a unit of energy not consumed equates to a unit of resources saved and a unit of pollution not generated.

Furthermore, some technical limitations on energy conservation are associated with the laws of physics and thermodynamics. Other technical limitations are imposed by practical technical constraints related to the real-world devices that are used. The minimum amount of fuel theoretically needed to produce a specified quantity of electricity, for example, could be determined by considering a Carnot (ideal) heat engine. However, more than this theoretical minimum fuel may be needed due to practical technical matters such as the maximum temperatures and pressures that structures and materials in the power plant can withstand.

As environmental concerns such as pollution, ozone depletion, and global climate change became major issues in the 1980s, interest developed in the link between energy utilization and the environment. Since then, there has been increasing attention to this linkage. Many scientists and engineers suggest that the impact of energy-resource utilization on the environment is best addressed by considering exergy. The exergy of a quantity of energy or a substance is a measure of the usefulness or quality of the energy or substance, or a measure of its potential to cause change. Exergy appears to be an effective measure of the potential of a substance to impact the environment. In practice, the authors feel that a thorough understanding of exergy and of how exergy analysis can provide insights into the efficiency and performance of energy systems is required for the engineer or scientist working in the area of energy systems and the environment.[8] Considering the above explanations, the general aspects of energy conservation can be summarized as shown in Figure 5.



Energy standards for new building constructions, achievable low-cost options for energy efficiency, energy conservation rules for energy managers, the principal aspects of energy conservation policies for general use, total energy performance of power systems, local fabrication of energy conservation materials and components for energy efficiency.

Energy management for energy conservation, the standard formats for energy conservation, establishment of regular training program for energy conservation, energy manager qualification, advisory services on energy conservation, regular activities on energy conservation, registration and licensing of energy conservation.

The enactment of energy, energy policies for energy conservation, lowering the energy intensity of economic activity, the creation of jobs for improvement of energy conservation, fiscal and non-monetary incentives in achieving the desired conservation and efficiency, commercial and traditional forms of energy conservation.

Energy performance standards for efficiency in energy consumption, consumer information on energy efficiency, energy conservation information on socio-cultural context, recalibration of energy conservation labels, consumer decisions on energy conservation, consumer understanding of the relation between environment and energy.

The feasibility of all energy efficiency promotion activities, energy price regulation for the long-term marginal costs of energy supply, the comparative inelasticity of energy demand for taxation of all forms of energy, emissions reduction and preservation or rehabilitation of the environment.

Appropriate fuel selection for energy conservation, the use of public transport for energy conservation, fuel efficiency in road transport, development of new transport fuels, use of electric and hybrid motor vehicles.

FIGURE 5 A flow chart of practical energy conservation aspects.

Research and Development Status on Energy Conservation

Now we look at R&D expenditures in energy conservation to assess the importance attached to energy conservation in the long range. The share of energy R&D expenditures going into energy conservation, for example, has grown greatly since 1976, from 5.1% in 1976 to 40.1% in 1990 and 68.5% in 2002.[11] This indicates that within energy R&D, research on energy conservation is increasing in importance. When R&D expenditures on energy conservation are compared with expenditures for research leading to protection of the environment in the 2000s, the largest share was spent on environment research. In fact, it is not easy to interpret the current trends in R&D expenditures, because energy conservation is now part of every discipline from engineering to economics. A marked trend has been observed since the mid-1970s, in that expenditures for energy conservation research have grown significantly, both in absolute terms and as a share of total energy R&D. These expenditures also grew more rapidly than those for environmental protection research, surpassing it in the early 1980s. Therefore, if R&D expenditures reflect long-term concern, there seems to be relatively more importance attached to energy conservation as compared with environmental protection.

In addition to the general trends discussed above, consider the industrial sector and how it has tackled energy conservation.

The private sector clearly has an important role to play in providing finance that could be used for energy efficiency investments. In fact, governments can adjust their spending priorities in aid plans and through official support provided to their exporters, but they can influence the vast potential pool of private-sector finance only indirectly. Many of the most important measures to attract foreign investors include reforming macroeconomic policy frameworks, energy market structures and pricing, and banking; creating debt recovery programs; strengthening the commercial and legal framework for investment; and setting up judicial institutions and enforcement mechanisms. These are difficult tasks that often involve lengthy political processes.

Thus, the following important factors, which are adopted from a literature work[13] can contribute to improving energy conservation in real life. Figure 6 presents the improvement factors of energy conservation.

Energy Conservation and Sustainable Development

Energy conservation is vital for sustainable development and should be implemented by all possible means, despite the fact that it has its own limitations. This is required not only for us, but for the next generation as well.

A secure supply of energy resources is generally considered a necessity but not a sufficient requirement for development within a society. Furthermore, sustainable development demands a sustainable supply of energy resources that, in the long term, is readily and sustainably available at reasonable cost and can be utilized for all required tasks without causing negative societal impact. Supplies of such energy resources as fossil fuels (coal, oil, and natural gas) and uranium are generally acknowledged to be finite. Other energy sources (such as sunlight, wind, and falling water) are generally considered to be renewable and, therefore, sustainable over the relatively long term. Wastes (convertible to useful energy forms through, for example, waste- to-energy incineration facilities) and biomass fuels usually also are viewed as being sustainable energy sources. In general, the implications of these statements are numerous and depend on how the term *sustainable* is defined.[14]

Energy resources and their utilization are intimately related to sustainable development. For societies to attain or try to attain sustainable development, much effort must be devoted not only to discovering sustainable energy resources, but also to increasing the energy efficiencies of processes utilizing these resources. Under these circumstances, increasing the efficiency of energy-utilizing devices is important. Due to increased awareness of the benefits of efficiency improvements, many institutes and agencies have started working along these lines. Many energy conservation and efficiency improvement programs have been developed and are being developed to reduce present levels of energy consumption.

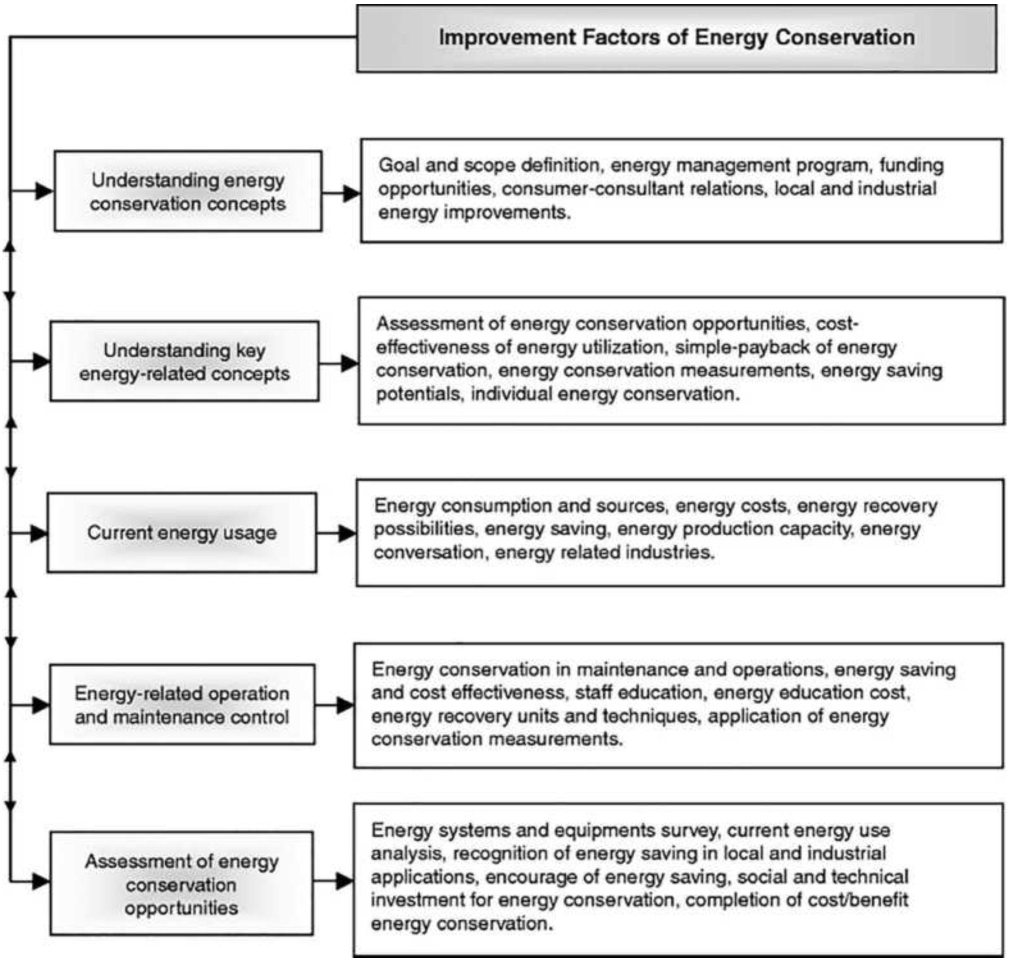


FIGURE 6 Improvement factors of energy conservation.

To implement these programs in a beneficial manner, an understanding is required of the patterns of "energy carrier" consumption-for example, the type of energy carrier used, factors that influence consumption, and types of end uses.[15]

Environmental concerns are an important factor in sustainable development. For a variety of reasons, activities that continually degrade the environment are not sustainable over time-that is, the cumulative impact on the environment of such activities often leads over time to a variety of health, ecological, and other problems. A large portion of the environmental impact in a society is associated with its utilization of energy resources. Ideally, a society seeking sustainable development utilizes only energy resources that cause no environmental impact (e.g., that release no emissions to the environment). Because all energy resources lead to some environmental impact, however, it is reasonable to suggest that some (not all) of the concerns regarding the limitations imposed on sustainable development by environmental emissions and their negative impacts can be overcome in part through increased energy efficiency. Clearly, a strong relationship exists between energy efficiency and environmental impact, because for the same services or products, less resource utilization and pollution normally are associated with increased energy efficiency.

Here, we look at renewable energy resources and compare them with energy conservation. Although not all renewable energy resources are inherently clean, there is such a diversity of choices that a shift to renewables carried out in the context of sustainable development could provide a far cleaner system than would be feasible by tightening controls on conventional energy. Furthermore, being by nature site-specific, they favor power system decentralization and locally applicable solutions more or less independently of the national network. It enables citizens to perceive positive and negative externalities of energy consumption. Consequently, the small scale of the equipment often makes the time required from initial design to operation short, providing greater adaptability in responding to unpredictable growth and/or changes in energy demand.

The exploitation of renewable energy resources and technologies is a key component of sustainable development.[11] There are three significant reasons for it:

- They have much less environmental impact compared with other sources of energy, because there are no energy sources with zero environmental impact. Such a variety of choices is available in practice that a shift to renewables could provide a far cleaner energy system than would be feasible by tightening controls on conventional energy.
- Renewable energy resources cannot be depleted, unlike fossil-fuel and uranium resources. If used wisely in appropriate and efficient applications, they can provide reliable and sustainable supply energy almost indefinitely. By contrast, fossil-fuel and uranium resources are finite and can be diminished by extraction and consumption.
- They favor power system decentralization and locally applicable solutions more or less independently of the national network, thus enhancing the flexibility of the system and the economic power supply to small, isolated settlements. That is why many different renewable energy technologies are potentially available for use in urban areas.

Taking into consideration these important reasons, the relationship between energy conservation and sustainability is finally presented as shown in Figure 7.

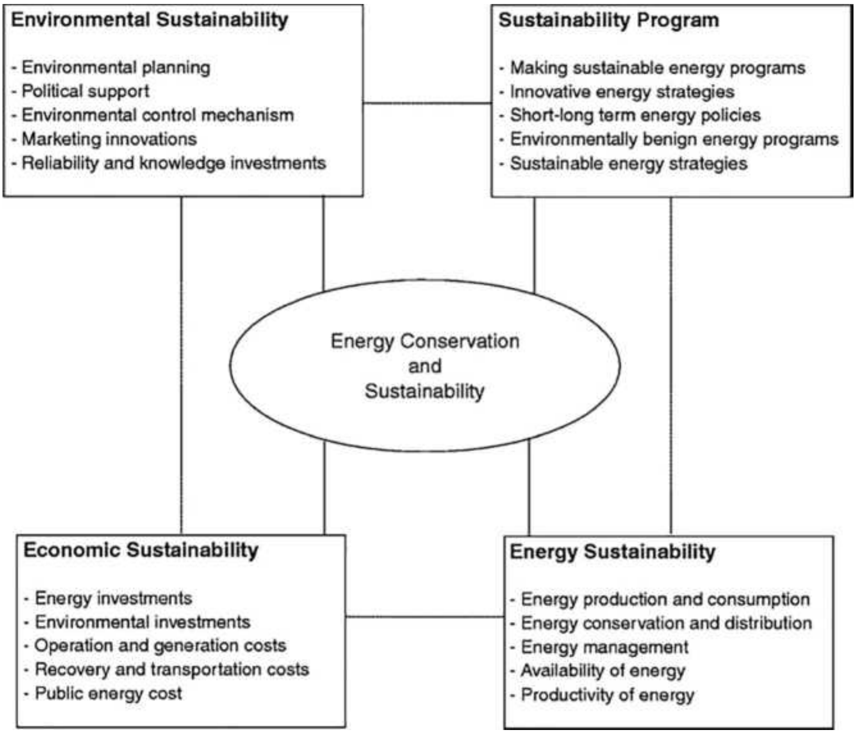


FIGURE 7 Linkages between energy conservation and sustainable development.

Energy Conservation Implementation Plan

The following basic steps are the key points in implementing an energy conservation strategy plan[11]:

1. *Defining the main goals.* It is a systematic way to identify clear goals, leading to a simple goal-setting process. It is one of the crucial concerns and follows an organized framework to define goals, decide priorities, and identify the resources needed to meet those goals.
2. *Identifying the community goals.* It is a significant step to identify priorities and links among energy, energy conservation, the environment, and other primary local issues. Here, it is also important to identify the institutional and financial instruments.
3. *Performing an environmental scan.* The main objective in this step is to develop a clear picture of the community to identify the critical energy-use areas, the size and shape of the resource-related problems facing the city and its electrical and gas utilities, the organizational mechanisms, and the base data for evaluating the program's progress.
4. *Increasing public awareness.* Governments can increase other customers' awareness and acceptance of energy conservation programs by entering into performance contracts for government activities. They can also publicize the results of these programs and projects. In this regard, international workshops to share experiences on the operation would help overcome the initial barrier of unfamiliarity in countries.
5. *Performing information analysis.* This step carries out a wide range of telephone, fax, email, and Internet interviews with local and international financial institutions, project developers, and bilateral aid agencies to capture new initiatives, lessons learned, and viewpoints on problems and potential solutions.
6. *Building community support.* This step covers the participation and support of local industries and communities, and the understanding the nature of conflicts and barriers between given goals and local actors; improving information flows; activating education and advice surfaces; identifying institutional barriers; and involving a broad spectrum of citizen and government agencies, referring to the participation and support of local industrial and public communities.
7. *Analyzing information.* This step includes defining available options and comparing the possible options with various factors (e.g., program implementation costs, funding availability, utility capital deferral, potential for energy efficiency, compatibility with community goals, and environmental benefits).
8. *Adopting policies and strategies.* Priority projects need to be identified through a number of approaches that are best for the community. The decision-making process should evaluate the cost of the options in terms of savings in energy costs; generation of business and tax revenue; and the number of jobs created, as well as their contribution to energy sustainability and their benefit to other community and environmental goals.
9. *Developing the plan.* When a draft plan has been adopted, it is important for the community to review it and comment on it. The public consultation process may vary, but the aim should be a high level of agreement.
10. *Implementing new action programs.* This step involves deciding which programs to concentrate on, with long-term aims being preferred over short-term aims. The option that has the greatest impact should be focused on, and all details should be defined, no matter how difficult the task seems. Financial resources needed to implement the program should be identified.
11. *Evaluating the success.* The final stage is evaluating and assessing how well the plan performed, which helps identify its strengths and weaknesses and to determine who is benefiting from it.

Energy Conservation Measures

For energy conservation measures, the information about the measure's applicability, cost range, maintenance issues, and additional points should be presented. Energy conservation involves efficiency improvements, formulation of pricing policies, good "housekeeping practices," and load management strategies, among other measures. A significant reduction in consumer energy costs can occur if conservation measures are adopted appropriately. The payback period for many conservation programs is less than 2 years.

In spite of the potentially significant benefits of such programs to the economy and their proven successes in several countries, conservation programs have not yet been undertaken on a significant scale in many developed and developing countries. Some reasons for this lack of energy conservation programs relate to the following factors:

- Technical (e.g., lack of availability, reliability, and knowledge of efficient technologies)
- Institutional (e.g., lack of appropriate technical input, financial support, and proper program design and monitoring expertise)
- Financial (e.g., lack of explicit financing mechanisms)
- Managerial (e.g., inappropriate program management practices and staff training)
- Pricing policy (e.g., inappropriate pricing of electricity and other energy commodities)
- Information diffusion (e.g., lack of appropriate information)

Reduced energy consumption through conservation programs can benefit not only consumers and utilities, but society as well. In particular, reduced energy consumption generally leads to reduced emissions of greenhouse gases and other pollutants into the environment.

Accelerated gains in energy efficiency in energy production and use, including those in the transportation sector, can help reduce emissions and promote energy security. Although there is a large technical potential for increased energy efficiency, there exist significant social and economic barriers to its achievement. Priority should be given to market forces in effecting efficiency gains. Reliance on market forces alone, however, is unlikely to overcome these barriers. For this reason, innovative and bold approaches are required by governments, in cooperation with industry, to realize the opportunities for energy efficiency improvements, and to accelerate the deployment of new and more efficient technologies.

Here, we look at energy conservation measures, which may be classified in six elements:

- Sectoral measures
- Energy conservation through systematic use of unused energy
- Energy conservation by changing social behavior
- International cooperation to promote energy conservation to counteract global warming
- Enhancing international and government-industry- university cooperation in developing technologies for energy conservation
- Promoting diffusion of information through publicity and education

The emphasis is on sectoral energy conservation. Table 4 presents some examples of such sectoral energy conservation measures. After determining which energy conservation measures are applicable, you should read the description of each of the applicable energy conservation measures. Information about the savings that can be expected from the measure, maintenance issues related to the measure, and other items to consider is provided for each energy conservation measure.

To evaluate the energy conservation measures, the following parameters should be taken into consideration[13]:

- *Cost estimation*. The first step is to estimate the cost of purchasing and installing the energy conservation measure. Cost estimates should be made for the entire development rather than for a single piece of equipment (e.g., obtain the cost for installing storm windows for an entire

TABLE 4 Sectoral Energy Conservation Measures

Sector	Measures
Industrial	Strengthening of financial and tax measures to enhance adoption and improvement of energy saving technologies through energy conservation equipment investments Re-use of waste energy in factories and/or in surrounding areas Enhancing recycling that reduces primary energy inputs such are iron scraps and used papers, and devising measures to facilitate recycling of manufactured products Retraining of energy managers and diffusion of new energy saving technologies through them Creating database on energy conservation technologies to facilitate diffusion of information
Residential and Commercial	Revising insulation standards provided in the energy conservation law, and introducing financial measures to enhance adoption of better insulation Development of better insulation materials and techniques Developing' energy conservation' model homes and total energy use systems for homes Revising or adopting energy conservation standards for home and office appliances Developing more energy saving appliances Revising guidelines for managing energy use in buildings, and strengthening advisory services to improve energy management in buildings
Transportation	Because 80% of energy consumption of the sector is by automobiles, further improvement in reducing fuel consumption by automobiles is necessary together with improvement in transportation system to facilitate and reduce traffic flow Diffusion of information about energy efficient driving Adopting financial measures to enhance the use of energy saving transportation equipment such as wind powered boats

Source: Adapted from Energy Conservation Policies and Technologies in Japan: A Survey.[16]

development or building, rather than the cost of one storm window). If you are planning to implement the energy conservation measure without the help of an outside contractor, you can obtain cost estimates by calling a vendor or distributor of the product. If, on the other hand, you will be using a contractor to install or implement the energy conservation measure, the contractor should provide estimates that include all labor costs and contract margins.

- *Data survey.* In this step, the questions on fuel consumption and cost should be listed for more than one possible fuel type (e.g., gas, oil, electric, or propane). The appropriate data for each fuel type should be selected and used accordingly for the cost estimation of each fuel.
- *Energy savings.* The amount of energy or fuel used should be estimated.
- *Cost savings.* This step determines the level of savings.
- *Payback period.* The last step in the cost/benefit analysis estimates the simple payback period. The payback period is found by dividing the cost of the measure by the annual cost savings.

Life-Cycle Costing

The term LCC for a project or product is quite broad and encompasses all those techniques that take into account both initial costs and future costs and benefits (savings) of a system or product over some period of time. The techniques differ, however, in their applications, which depend on various purposes of systems or products. Life-cycle costing is sometimes called a cradle-to-grave analysis. A life-cycle cost analysis calculates the cost of a system or product over its entire life span. Life-cycle costing is a process to determine the sum of all the costs associated with an asset or part thereof, including acquisition, installation, operation, maintenance, refurbishment, and disposal costs. Therefore, it is pivotal to the asset management process.

From the energy conservation point of view, LCC appears to be a potential tool in deciding which system or product is more cost effective and more energy efficient. It can provide information about how to evaluate options concerning design, sites, materials, etc., how to select the best energy conservation feature among various options; how much investment should be made in a single energy conservation feature; and which is the most desirable combination of various energy conservation features.

A choice can be made among various options of the energy conservation measure that produces maximum savings in the form of reduction in the life-cycle costs. A choice can be made between double-glazed and triple-glazed windows, for example. Similarly, a life-cycle cost comparison can be made

TABLE 5 An Example of Life-Cycle Costing (LCC) Analysis

Cost of Purchasing Bulbs	Incandescent	Compact Fluorescent
Lifetime of one bulb (hours)	1,000	10,000
Bulb price ($)	0.5	6.0
Number of bulbs for lighting 10,000 h	10	1
Cost for bulbs ($)	$10 \times 05 = 5.0$	$1 \times 6 = 6$
Energy cost		
Equivalent wattage (W)	75	12
Watt-hours (Wh) required for lighting for 10,000 h	$75 \times 10{,}000 = 750{,}000$ Wh = 750 kWh	$12 \times 10{,}000 = 120{,}000$ Wh = 120 kWh
Cost at 0.05 per kWh	750 kWh × \$0.05 = \$37.5	120 kWh × \$0.05 = \$6
Total cost ($)	$5 + 37.5 = 42.5$	$6 + 6 = 12$

between a solar heating system and a conventional heating system. The one that maximizes the life-cycle costs of providing a given level of comfort should be chosen. The application of such techniques to energy conservation is related to determining the optimum level of the chosen energy conservation measure. Sometimes, energy conservation measures involve the combination of several features. The best combination can be determined by evaluating the net LCC effects associated with successively increasing amounts of other energy conservation measures. The best combination is found by substituting the choices until each is used to the level at which its additional contribution to energy cost reduction per additional dollar is equal to that for all the other options.

Illustrative Example

Here, we present an illustrative example on LCC to highlight its importance from the energy conservation point of view. This example is a simple LCC analysis of lighting for both incandescent bulbs and compact fluorescent bulbs, comparing their life-cycle costs as detailed in Table 5. We know that incandescents are less expensive (95% to heat and 5% to usable light) and that compact fluorescent bulbs are more expensive but much more energy efficient. So the question is which type of lighting comes out on top in an LCC analysis.

This example clearly shows that LCC analysis helps in energy conservation and that we should make it part of our daily lives.

Conclusion

Energy conservation is a key element in sectoral (e.g., residential, industrial, and commercial) energy utilization and is vital for sustainable development. It should be implemented by all possible means, despite the fact that it sometime has its own limitations. This is required not only for us, but for the next generation as well. A secure supply of energy resources is generally considered a necessary but not a sufficient requirement for development within a society. Furthermore, sustainable development demands a sustainable supply of energy resources that, in the long term, is readily and sustainably available at reasonable cost and can be utilized for all required tasks without causing negative societal impact.

An enhanced understanding of the environmental problems relating to energy conservation presents a high-priority need and an urgent challenge, both to allow the problems to be addressed and to ensure that the solutions are beneficial for the economy and the energy systems.

All policies should be sound and make sense in global terms-that is, become an integral part of the international process of energy system adaptation that will recognize the very strong linkage existing between energy requirements and emissions of pollutants (environmental impact).

In summary this study discusses the current environmental issues and potential solutions to these issues; identifies the main steps for implementing energy conservation programs and the main barriers to such implementations; and provides assessments for energy conservation potentials for countries, as well as various practical and environmental aspects of energy conservation.

References

1. Azar, C.; Rodhe, H. Targets for stabilization of atmospheric CO_2. Science **1997**, *276,* 1818–1819.
2. Midilli, A.; Ay, M.; Dincer, I.; Rosen, M.A. On hydrogen and hydrogen energy strategies: I: Current status and needs. Renew. Sust. Energy Rev. **2005**, *9* (3), 255–271.
3. BP. Workbook 2005. *Statistical Review—Full Report of World Energy,* British Petroleum, 2005, Available at http://www.bp.com/centres/energy (accessed July 25, 2005).
4. IEE. World Energy Outlook 2002. Head of Publication Services, International Energy, 2002, Available at http://www.worldenergyoutlook.org/weo/pubs/we02002/WEO21.pdf.
5. Kazman, S. Global *Warming and Energy Policy;* CEI; 67–86, 2003. Available at http://www.cei.org.
6. Dincer, I. Energy and environmental impacts: Present and future perspectives. Energy Sources **1998**, *20* (4/5), 427–453.
7. Dincer, I. *Renewable Energy, Environment and Sustainable Development.* Proceedings of the World Renewable Energy Congress; September 20–25, 1998; Florence, Italy; 2559–2562.
8. Rosen, M.A.; Dincer, I. On exergy and environmental impact. Int. J. Energy Res. **1997**, *21* (7), 643–654.
9. Dincer, I. Renewable energy and sustainable development: A crucial review. Renew. Sust. Energy Rev. **2000**, *4* (2), 157–175.
10. Dincer, I. *Practical and Environmental Aspects of Energy Conservation Technologies.* Proceedings of Workshop on Energy Conservation in Industrial Applications; Dhahran, Saudi Arabia, February 12–14, 2000; 321–332.
11. Dincer, I. On energy conservation policies and implementation practices. Int. J. Energy Res. **2003**, *27* (7), 687–702.
12. Sissine, F. Energy efficiency: Budget, oil conservation, and electricity conservation issues. *CRS Issue Brief for Congress.* IB10020; 2005.
13. Nolden, S.; Morse, D.; Hebert, S. *Energy Conservation for Housing: A Workbook,* Contract-DU100C000018374; Abt Associates, Inc.: Cambridge, MA, 1998.
14. Dincer, I.; Rosen, M.A. A worldwide perspective on energy, environment and sustainable development. Int. J. Energy Res. **1998**, *22* (15), 1305–1321.
15. Painuly, J.P.; Reddy, B.S. Electricity conservation programs: Barriers to their implications. Energy Sources **1996**, *18,* 257–267.
16. Anon. *Energy Conservation Policies and Technologies in Japan: A Survey*, OECD/GD(94)32, Paris, 1994.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

47

Energy Conservation: Lean Manufacturing

Bohdan W. Oppenheim

Introduction

At the time of this writing (2005), the world is experiencing strong contradictory global trends of diminishing conventional energy resources and rapidly increasing global demands for these resources, resulting in substantial upwards pressures in energy prices. Because the energy used by industry represents a significant fraction of the overall national energy use, equal to 33% in the United States in the year 2005, a major national effort is underway to conserve industrial energy.[1] The rising energy prices place escalating demands on industrial plants to reduce energy consumption without reducing production or sales, but by increasing energy density.

Optimization of industrial hardware and its uses, including motors and drives, lights, heating, ventilation and cooling equipment, fuel-burning equipment, and buildings, are well understood, have been practiced for years,[2] and are important in practice. However, they offer only limited energy conservation opportunities, rarely exceeding a few percent of the preoptimization levels. In contrast, the impact of productivity on energy use and energy density offers dramatically higher savings opportunities in energy and in other costs. In the extreme case, when transforming a factory from the traditional "process village" batch-and- queue system to the state-of-the-art, so-called Lean system, the savings in energy can reach 50% or more.

The best organization of production known at this time is called Lean, developed at Toyota in Japan.[3] It is the flow of value-added work through all processes required to convert raw materials to the finished products with minimum waste. Major elements of Lean organization include: steady single-piece flow with minimum inventories and no idle states or backflow; flexible production with flexible equipment and operators and flexible floor layouts ready to execute the order of any size profitably and just-in-time; reliable and robust supplies of raw materials; minimized downtime due to excellent preventive maintenance and quick setups; first-pass quality; clean, uncluttered, and well-organized work space; optimized work procedures; and, most importantly, an excellent workforce–well trained, motivated, team-based and unified for the common goals of having market success, communicating efficiently, and being well-managed. The Lean organization of production is now well understood among productivity professionals, but it is not yet popular among the lower tier suppliers in the United States. Its implementation would save energy and benefit the suppliers in becoming more competitive.

The engineering knowledge of energy conservation by equipment improvements is well understood and can be quantified with engineering accuracy for practically any type of industrial equipment.[2] In contrast, industrial productivity is strongly influenced by intangible and complex human factors such as management, work organization, learning and training, communications, culture, and motivation. These work aspects are difficult to quantify in factory environments. For this reason, the accuracy of productivity gains and the related energy savings are typically much less accurate than the energy savings computed from equipment optimization. Simple quantitative models with a conservative bias are therefore recommended as tools for energy management in plants. This entry includes some examples. They are presented in the form of energy savings or energy cost savings that would result from implementing a given productivity improvement, or eliminating a given productivity waste, or as simple metrics measuring energy density.

It is remarkable that in most cases, these types of energy savings occur as a natural byproduct of productivity improvements, without the need for a direct effort centered on energy. Thus, the management should focus on productivity improvements. In a traditional non-Lean plant intending to transform to Lean production, the first step should be to acquire the knowledge of the Lean system. It is easily available from industrial courses and workshops, books,[3,4] and video training materials.[6] The next step should be the actual transformation of production to Lean. Most of the related energy savings will then occur automatically. Implementation of individual productivity elements such as machine setup time reduction will yield some energy savings, but the result will not be as comprehensive as those yielded by the comprehensive implementation of Lean production.

Traditional vs. Lean Production

The traditional organization of production still used frequently in most factories tends to suffer from the following characteristics:

- Supplier selection is based on minimum cost, resulting in a poor level of mutual trust and partnership, the need for receiving inspection, and often large inventories of raw materials (RM).
- Work-in-progress (WIP) is moving in large batches from process village to process village and staged in idle status in queues in front of each machine, while the machine moves one piece at a time. This work organization is given the nickname "batch-and-queue" (BAQ).[3]
- Finished goods (FG) are scheduled to complex forecasts rather than customer orders, resulting in large inventories.
- The floor is divided into "process villages" populated with large, complex, and fast similar machines selected for minimum unit cost.
- Minimum or no information is displayed at workstations, and the workers produce quotas.
- Work leveling is lacking, which results in a random mix of bottlenecks and idle processes.
- Unscheduled downtime of equipment occurs frequently.
- Quality problems with defects, rework, returns, and customer complaints are frequent.
- Quality assurance in the form of 100% final inspections attempts to compensate for poor production quality.
- The floor space is cluttered, which makes moving around and finding items difficult.
- The workforce has minimum or no training and single skills.
- The management tends to be authoritarian.
- A language barrier exists between the workers and management.
- There is a culture of high-stress troubleshooting rather than creative trouble prevention.

In such plants, the waste of materials, labor, time, space, and energy can be as much as 50%–90%.[3]

The Lean production method developed primarily at Toyota in Japan under the name Just-In-Time (JIT), and generalized in the seminal work[3] is the opposite of the traditional production in almost all respects, as follows:

- Raw materials are bought from reliable supplier– partners and delivered JIT in the amount needed, at the price agreed, and with the consistently perfect quality that obviates incoming inspection.
- Single-piece flow (SPF) of WIP is steadily moving at a common takt time (Takt time is the common rhythm time of the pieces moving from workstation to workstation on the production line. It is the amount of time spent on EACH operation. It precisely synchronizes the rate of all production operations to the rate of sales JIT.), from the first to the last process.
- The FG are produced to actual customer orders JIT resulting in minimum inventories.
- The floor is divided into flexible production lines with small simple machines on casters that can be pushed into position and setup in minutes.
- The labor is multiskilled, well motivated and well trained in optimized procedures.
- Quality and production status are displayed on large visible boards at each workstation, making the entire production transparent for all to see.
- Preventive maintenance assures no unscheduled downtime of equipment.
- All process operators are trained in in-line quality checks and variability reduction.
- No final inspection is needed, except for occasional sampled checks of FG.
- Defects, rework, returns, and customer complaints are practically eliminated.
- The floor space is clean and uncluttered.
- The workforce is trained in company culture and commonality of the plant mission, customer needs, workmanship, and quality.
- The culture promotes teamwork, multiple job skills, supportive mentoring management, and company loyalty.
- The management promotes trouble prevention and “stopping the line” at the first sign of imperfection so that no bad pieces flow downstream.

According to Womack et al. the transformation from traditional to Lean production can reduce overall cost, inventory, defects, lead times by 90%, and space by 50%, and vastly increase plant competitiveness, customer satisfaction, and workforce morale. The resultant energy savings can be equally dramatic. Liker,[4] contains interviews with industry leaders who have succeeded in this transformation.

Impact on Energy

The impact of productivity on plant energy falls into the following two broad categories:

1. Productivity improvements that save infrastructure energy. These improvements reduce the energy consumed by all plant support systems, which tend to be energized regardless of the actual production activities, such as lights, space cooling and heating devices, cooling towers, combustion equipment (boilers, molten metal furnaces), air compressors, forklift battery chargers, conveyors, etc. To the first approximation, the infrastructure energy is reduced in proportion to the production time reductions, which can be huge in the Lean system. In order to perform more detailed estimates of the infrastructure energy savings, the management would have to conduct detailed energy accounting and understand how much energy is used by each support system under different production conditions. This knowledge is rarely available; therefore the former simplistic approach, combined with conservative estimates, offer useful tools.
2. Process energy savings. In this category, the energy savings of process equipment are obtained by improving the process productivity. Examples include the reduction of unscheduled machine downtime or setup time and the elimination of process variability, defects, rework, scrap, excessive labor time, etc.

Single Piece Flow (SPF)

Changing the traditional BAQ production to Lean production is by far the most effective productivity transformation a plant can undertake, creating dramatic savings in the overall throughput time, cost, quality, and energy. The example shown in Figure 1 compares just one aspect of the transformation—a reduction of batch size from five to one, i.e., the SPF. In both cases, four processes of equal one-minute takt time are assumed. The benefits of the SPF alone are dramatic, as follows:

1. In BAQ, the batch is completed in 20 min and in SPF in only 8 min, a 60% reduction.
2. In BAQ, only one machine at a time produces value, while three others are idle. If the idle machines remain energized, as is the case, e.g., with injection molding, three of the four machines (75%) would be wasting energy, and doing it for 16 min each, adding up to 64 min of machine energy

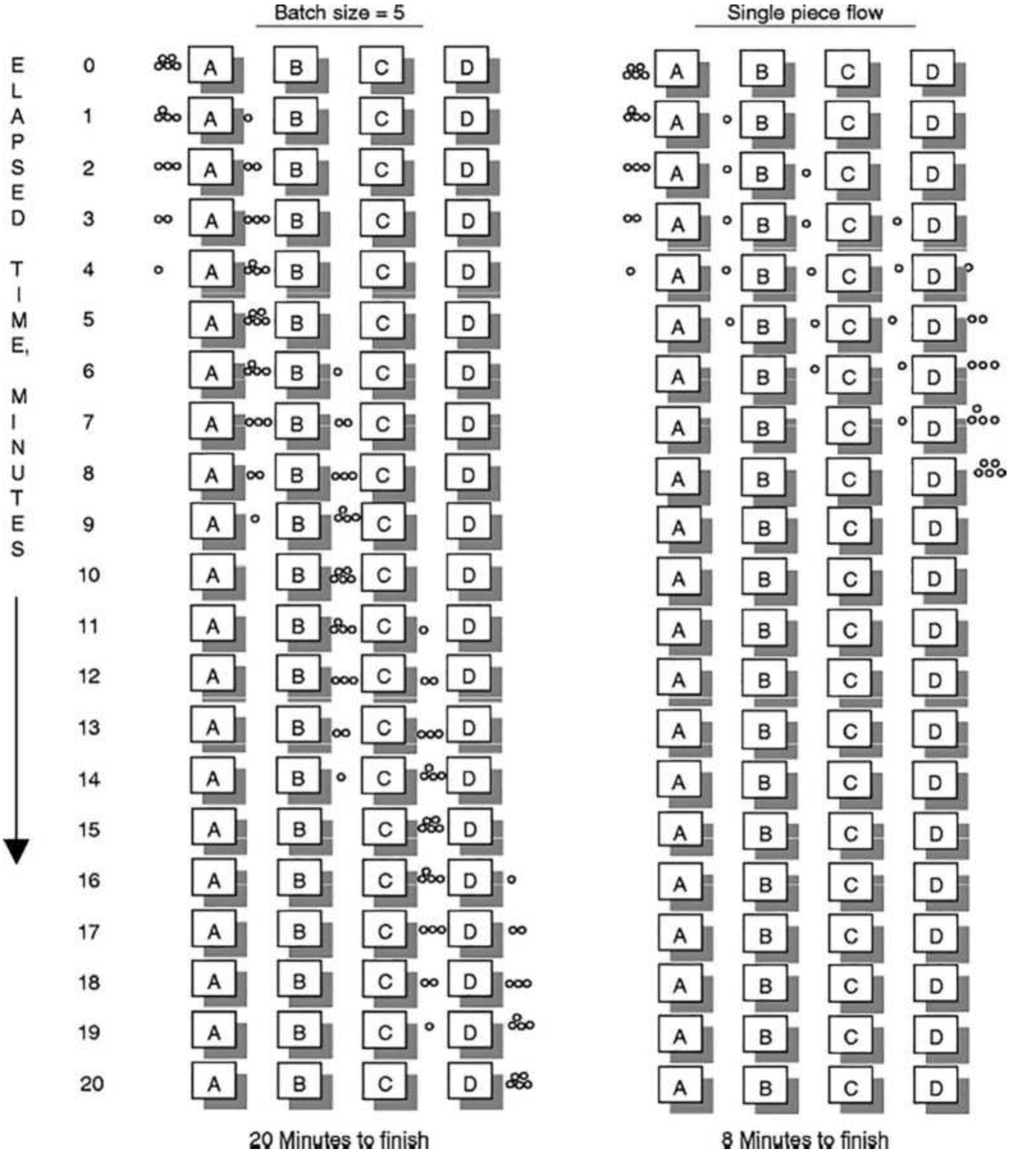


FIGURE 1 BAQ with batch size of five vs SPF.

wasted. In the SPF system, no machine energy is wasted as no machine would be idle, except for the lead and tail of each process of 4 min, adding up to 16 min of machine energy wasted, a savings of 75% from BAQ.

3. Reducing the batch throughput time by 60% reduces the infrastructure energy by the same amount, assuming the production is completed faster and the plant is de-energized. Alternatively, the freed 60% time and energy could be used for additional production and profits.
4. An important additional benefit is that in SPF, a defect can be detected on the first specimen–as soon as it reaches the next process, while in the BAQ, the entire batch may be wasted before the defect is discovered and a corrective action undertaken, with the energy used for making the batch wasted.

This simple example clearly illustrates the dramatic impact of SPF on both overall productivity and energy consumption. Typically, as the factories transform to the Lean system, their sales, production, and profits increase simultaneously and the energy used decreases. A convenient metric to track the overall benefit is the gross energy density, ED_1 or ED_2:

$$\mathrm{ED}_1 = \frac{\mathrm{EC}_\mathrm{T}}{P} \tag{1a}$$

$$\mathrm{ED}_2 = \frac{\mathrm{EC}_\mathrm{T}}{\mathrm{AC}} \tag{1b}$$

where EC_T is the overall annual cost of energy in the plant, P is the number of products produced per year and AC is the total annual costs (sales minus net profit). ED_1 should be used if similar products are made most of the time, and ED_2 should be used if the plant has a wide menu of dissimilar products. The ED ratios will decrease as progress is made from BAQ to SPF. If the volume of production remains constant during the transformation, energy savings and energy cost savings alone may be more convenient metrics to track plant energy efficiency.

Inventory Reduction

All inventories, whether in RM, WIP, or FG, beyond the immediate safety buffers, are detrimental. Inventory means that company capital is “frozen” on the floor; cutting into the cash flow; wasting labor for inventory control, storage, and security; wasting infrastructure energy for lights, forklift energy, and possible cooling or heating of the inventory spaces if the goods require temperature or humidity control; wasting space and the associated lease/mortgage fees and taxes; and becoming scrap if not sold (a frequent waste in large inventories). Inventory and inventory space reductions lead to infrastructure energy savings. Process energy can also be saved by not making the FG that end up in inventory, cannot be sold, and become scrap. Womack and Jones[3] and Liker[4] contain case studies for, among others, inventory reductions. A convenient nondimensional metric to track the overall impact of all inventories on energy savings is

$$\mathrm{EC}_T \times \frac{I_T}{AC} \tag{2}$$

where I_T is the number of inventory turns per year.

Workmanship, Training, and Quality Assurance

In the ideal Lean system, the processes, equipment, procedures, and training are perfected to the degree that guarantees consistent and robust production with predictable effort, timing, quality, and cost; with no variability, defects, or rework, and with maximum ergonomics and safety. This is accomplished by a

TABLE 1 Energy Waste from Poor Workmanship

A plant with \$20,000,000 in sales and \$2,000,000 in profits spends \$1,000,000 on energy per year. The typical order requires 10 processes of roughly equal energy consumption. The production equipment consumes 60% and the supportive infrastructure consumes 40% of the plant energy. Sequential process #5 has the defect rate of 10%. In order to compensate for the defects, the first 5 processes must produce 10% extra pieces. The annual waste of energy cost (and the energy cost savings, if the defective process is fixed) is then:

(\$1,000,000/yr)(5/10 processes) (60% process energy) (10% defect rate) = \$30,000/yr (3)

The additional production time of 10% waste not only the cost of the process energy computed in (3) but also the infrastructure energy cost of:

(\$1,000,000/yr) (40% infrastructure energy) (10% defects) = \$40,000/yr (4)

Such delays also extend the promised delivery time and reduce customer satisfaction and factory competitiveness. Adding (1) and (2) together. (not counting the direct productivity losses), the wasted energy cost alone of \$70,000/yr represents 3.5% of the annual profits and 7% in annual energy costs. Based on the author's experience,[5] these numbers are not infrequent in industry. Fixing the productivity of process #5 would eliminate these wastes.

consistent long-term strategy of continuous improvement of all the above elements, including intensive initial training of the workforce and subsequent retraining in new procedures. A procedure must be developed for each process until it is robust and predictable and optimized for minimum overall cost, required quality, maximum ergonomics, and safety. Process operators must be trained in the procedures as well as in the process quality assurance, and they must be empowered to stop the process and take corrective action or call for help if unable to avoid a defect. Management culture must be supportive for such activities. Any departure from this ideal leads to costly penalties in quality, rework, delays, overtime or contract penalties, crew frustrations, and customer dissatisfaction. These, in turn, have negative impacts on energy as follows:

1. Defects require rework, which requires additional energy to remake or repair the part. The best metric to use here is the energy or energy cost per part used in the given defective process multiplied by the number of bad parts produced per year.
2. Variability in the process time or delays caused by defects mean that the production takes more time and more infrastructure and process energy for the same amount of value work and profits when compared with the ideal nonvariable process. Table 1 illustrates cases (1) and (2).
3. Defective processes usually require a massive final inspection to sort out the good products. Finding the finished goods defective is the most inefficient means of quality assurance because often the entire batch must then be remade, consuming the associated energy. The inspection space, labor, and energy represent a direct waste and should be replaced with in-line quality assurance (ILQA) that detects the first bad piece (Governmental, medical, etc., orders usually require a 100% final inspection. In the Lean system, this is performed as a formality because everybody in the plant knows that all pieces will be perfect because all imperfections have been removed in real time before the inspection process.) and immediately undertakes a corrective action. Typically, the ILQA can be implemented in few days of operators' training and has the simple payback period measured in days or weeks.[5]

Overage Reduction

Many a plant compensates for its notorious defects by routinely scheduling production in excess of what the customer orders. Some minimum overage is usually justified for machine setups, adjustments, and QA samples. In a Lean plant this rarely exceeds a fraction of one percent. In a traditional plant, the value of 5%–15% is not infrequent. A 5% overage means that the plant spends 105% of the necessary costs. If the profit margin is 5%, the overage alone may consume the entire profit. The overall energy waste (and the opportunity to save energy) is simply proportional to the overage amount. Overage is one of the most wasteful ways of compensating for defective processes. The best remedy is to simply identify the defective process with ILQA, find the root cause (typically the lack of training, excessive work quotas, or bad process or material), and repair it.

Unintentional overage can also be destructive to profits and energy use. Example: A worker is asked to cut only a few small pieces from a large sheet of metal, but instead he cuts the entire sheet, thinking, "my machine is already setup and soon they will ask me to cut the rest of the sheet anyway, so I may as well do it now." The excessive pieces then move through all processes, unknowingly to the management, consuming energy, labor and fixed costs, to end up as excessive FG inventory and, in the worst case, find no buyer and end up as scrap. Uncontrolled and careless overage can easily consume all profits, and, of course, waste energy proportionately to the overage amount.

Downtime

Equipment downtime and idleness may occur due to scheduled maintenance, unscheduled breakdowns, machine setups, and poor process scheduling. The downtime may cause proportional loss of both profits and energy. The downtime may have fourfold impact on energy use, as follows:

1. When a process stops for whatever reason during an active production shift, the plant infrastructure continues to use energy and loosing money, as in Equation 4. A good plant manager should understand what fraction of the infrastructure energy is wasted during the specific equipment downtime. With this knowledge, the energy waste can be estimated as being proportional to the downtime.
2. Some machines continue using energy during maintenance, repair, or setup in proportion to the downtime (e.g., the crucible holding molten metal for a die casting machine remains heated by natural gas while the machine is being setup or repaired). Reducing the setup time or eliminating the repair time saves the gas energy in direct proportion to the downtime saved. In order to calculate energy savings in such situations, it is necessary to understand the energy consumption by the equipment per unit of time multiplied by the downtime reduction.
3. When a particular machine is down, additional equipment upstream or downstream of that machine may also be forced into an idle status but remain energized, thus wasting energy. In an ideal single-piece flow, the entire production line (As in the saying "In Lean either everything works or nothing works") will stop. In order to estimate the energy-saving opportunity from reducing this cumulative downtime, the energy manager must understand which equipment is idled by the downtime of a given machine and how much energy it uses per unit time while being idle.
4. Lastly, energized equipment should be well managed. A high-powered machine may be left energized for hours at a time when not scheduled for production. A good practice is to assign each of these machines to an operator who will have the duty of turning the machine off when not needed for a longer time, if practical, and to turn it back on just in time to be ready for production exactly when it is needed.

Preventive maintenance and setup time reduction have a particularly critical impact on both productivity and related energy use, as follows:

Preventive Maintenance

Practical and routine preventive maintenance should be done during the hours free of scheduled production (e.g., during night shifts, on weekends, or during layover periods). The maintenance should be preventive rather than reactive (The term "preventive" tends to be replaced with "productive" in modern industrial parlance). Well-managed "total" preventive[6] maintenance involves not only oiling and checking the machines per schedule but also ongoing training of the mechanics; developing a comprehensive database containing information on the particular use and needs of various machines; preparing a schedule of part replacement and keeping inventory of frequently used spare parts; and a

TABLE 2 Energy Savings from Setup Time Reduction

A plant operates on two shifts, 260 days per year, performing on average of 20 two-hour setups per day on their electrically heated injection molding machines. Each machine consumes 20 kW when idle bur energized. By a focused continuous improvement system and training, the crew reduces the routine setup time 0.5 h, with few, if any expenses for additional hardware, thus saving:

(260 days/yr) (20 setups/day) (1.5 h saved/setup) = 7800 machine h/yr.

The resultant process energy saved will be:

(7800 h/yr) (20 kW) = 156,000 kWh/yr (5)

In addition, infrastructure energy will be saved because of the reduced downtime. Using the data from Example 1, if the work is done in two shifts for 260 days per year (4160 h/yr.), the plant infrastructure uses 40% of the plant energy, and each machine consumes 2% of the plant infrastructure energy during the set up, the additional energy cost savings due to the setup time reduction will be:

(7,800 hr/yr) (0.02) (0.04) ($1,000,000)/(4160 h/yr) = #15,000 (6)

well-managed ordering system for other parts, including vendor data so that when a part is needed it can be ordered immediately and shipped using the fastest possible means. Industry leaders have demonstrated that affordable preventive maintenance can reduce the unscheduled downtime and associated energy waste to zero. This should be the practical goal of well-run factories.

Setups

Modern market trends push industry towards shorter series and smaller orders, requiring, in turn, more and shorter setups. Industry leaders have perfected routine setups to take no more than a few minutes. In poorly managed plants, routine setups can take as long as several hours. In all competitive modern plants, serious efforts should be devoted to setup time reductions. The effort includes both training and hardware improvements. The training alone, with only minimal additional equipment (such as carts), can yield dramatic setup time reductions (i.e., from hours to minutes). Further gains may require a change of the mounting and adjustment hardware and instrumentation. Some companies organize competitions between teams for developing robust procedures for the setup time reductions. In a plant performing many setups, the opportunity for energy savings may be significant, both in the process and infrastructure energy, as shown in Table 2.

Flexibility

Production flexibility, also called agility, is an important characteristic of competitive plants. A flexible plant prefers small machines (if possible, on casters) that are easy to roll into position and plug into adjustable quick-connect electrical and air lines and that are easy to setup and maintain over the large fixed machines selected with large batches and small unit costs in mind (such machines are called "monuments" in Womack and Jones[3]). Such an ideal plant will also have trained a flexible workforce in multiple skills, including quality assurance skills. This flexibility allows for the setup of new production lines in hours or even minutes, optimizing the flow and floor layout in response to short orders, and delivers the orders JIT. The energy may be saved in two important ways, as follows:

- Small machines processing one piece at a time use only as much energy as needed. In contrast, when excessively large automated machines are used, the typical management choice is between using small batches JIT, thus wasting the large machine energy, or staging the batches for the large machine, which optimizes machine utilization at the expense of throughput time, production flow, production planning effort, and the related infrastructure energy.
- Small machines are conducive to flexible cellular work layout, where 2–4 machines involved in the sequential processing of WIP are arranged into a U-shaped cell with 1–3 workers serving all processes in the cell in sequence, and the last process being quality assurance. This layout can be made very compact—occupying a much smaller footprint in the plant compared to traditional

"process village" plants, roughly a reduction of 50%[3,4]—and is strongly preferred by workers because it saves walking and integrates well the work steps. Such a layout also saves forklift effort and energy and infrastructure energy due to the reduction of the footprint.

Other Productivity Elements

The complete list of productivity elements is beyond the scope of this entry, and all elements have some leverage on energy use and conservation. In the remaining space, only the few most important remaining aspects are mentioned, with their leverage on energy. Descriptive details can be found in Ohno[7] and numerous other texts on Lean production.

- *Visual factory:* Modern factories place an increasing importance on making the entire production as transparent as possible in order to make any problem visible to all, which is motivational for immediate corrective actions and continuous improvements. Ideally, each process should have a white board displaying the short-term data, such as the current production status (quantity completed vs required); the rate of defects or rejects and their causes; control charts and information about the machine condition or maintenance needs; and a brief list and explanation of any issues, all frequently updated. The board should also display long-term information such as process capability history, quality trends, operator training, etc. Such information is most helpful in the optimization of, among other things, process time and quality, which leads to energy savings, as discussed above.
- "*Andon*" *signals:* The term refers to the visual signals (lights, flags, markers, etc.,) displaying the process condition, as follows: "green=all OK," "yellow=minor problem being corrected," and "red=high alarm, stopped production, and immediate assistance needed." The signals are very useful in identifying the trouble-free and troubled processes, which is conducive to focusing the aid resources to the right places in real time, fixing problems immediately and not allowing defects to flow downstream on the line. These features, in turn, reduce defects, rework, delays, and wasted costs, which improve overall productivity and save energy, as described above. It is also useful to display the estimated downtime (Toyota and other modern plants have large centrally located Andon boards that display the Andon signal, the workstation number, and the estimated downtime.). Knowing the forecasted downtime frees other workers to perform their pending tasks which have waited for such an opportunity rather than wait idle. This leads to better utilization of the plant resources, including infrastructure energy.
- "*5Ss*": The term comes from five Japanese words that begin with the "s" sound and loosely translate into English as: sorting, simplification, sweeping, standardization, and self-discipline (many other translations of the words are popular in industry); and describes a simple but powerful workplace organization method.[8] The underlying principle of the method is that only the items needed for the immediate task (parts, containers, tools, instructions, materials) are kept at hand where they are needed at the moment, and everything else is kept in easily accessible and well-organized storage in perfect order, easy to locate without searching, and in just the right quantities. All items have their designated place, clearly labeled with signs, labels, part numbers, and possibly bar codes. The minimum and maximum levels of inventory of small parts are predefined and are based on actual consumption rather than the "just-in-case" philosophy. The parts, tools, and materials needed for the next shift of production are prepared by a person in charge of the storage during the previous shift and delivered to the workstation before the shift starts. The floor is uncluttered and marked with designated spaces for all equipment. The entire factory is spotlessly clean and uncluttered. Walls are empty except for the visual boards. In consequence of these changes, the searching for parts, tools, and instructions which can represent a significant waste of labor and time is reduced, and this, in turn, saves energy. Secondary effects are also important. In a well-organized place, fewer mistakes are made; fewer wrong parts are used; less inspection is

FIGURE 2 In this messy plant, the workers waste close to 20% of their time looking for items and scavenging for parts and tools, also wasting the plant energy.

needed; quality, throughput time, and customer satisfaction are increased; and costs and energy are decreased. Figure 2 illustrates a fragment of a messy factory, where the average worker was estimated to waste 20% of his shift time looking for and scavenging for parts and tools. This percentage multiplied by the number of workers yields a significant amount of wasted production time, also wasting plant energy in the same proportion. Sorting, cleaning, and organizing the workplace is one of the simplest and most powerful starting points on the way to improved productivity and energy savings.

Conclusion

Large savings in energy are possible as an inherent byproduct of improving productivity. The state-of-the-art Lean productivity method can yield dramatic improvements in productivity. In the extreme case of converting from the traditional batch-and-queue and "process village" manufacturing system to Lean production, overall costs, lead times, and inventories can be reduced by as much as 50%–90%, floor space and energy by 50%, and energy density can be improved by 50%. The amount of energy that can be saved by productivity improvements often radically exceeds the savings from equipment optimization alone, thus providing a strong incentive to include productivity improvements in energy-reduction efforts.

Productivity strongly depends on human factors such as management, learning, and training, communications, culture, teamwork, etc., which are difficult to quantify, making accurate estimates of the cost, schedule, and quality benefits from various productivity improvements and the related energy savings difficult to estimate with engineering accuracy. For this reason, simple metrics and models are recommended, and some examples have been presented. If applied conservatively, they can become useful tools for energy management in a plant. The prerequisite knowledge includes an understanding of Lean Flow and its various productivity elements and a good accounting of energy use in the plant, including the knowledge of the energy used by individual machines and processes both when in productive use and in the idle but energized state, as well as the energy elements used by the infrastructure (various light combinations, air- compressors, cooling and heating devices, combusting systems, conveyers, forklifts, etc.). In the times of ferocious global competition and rising energy prices, every industrial plant should make every effort to improve both productivity and energy use.

Acknowledgments

This work is a result of the studies of energy conservation using the Lean productivity method performed by the Industrial Assessment Center funded by the U.S. Department of Energy at Loyola Marymount University. The author is grateful to Mr. Rudolf Marloth, Assistant Director of the Center, for his help with various energy estimates included herein and his insightful comments, to the Center students for their enthusiastic work, and to his son Peter W. Oppenheim for his diligent editing.

References

1. U.S. Department of Energy, Energy Efficiency and Renewable Energy, available at http://www.eere.energy.gov/industry/ (accessed on December 2005).
2. U.S. Department of Energy, available at http://eereweb.ee.doe.gov/industry/bestpractices/plant_assessments.html, (accessed on December 2005).
3. Womack, P.J.; Jones, D.T. *Lean Thinking.* 2nd Ed.; Lean Enterprise Institute: Boston, 2005; (http://www.lean.org), ISBN: 0–7432-4927-5.
4. Liker, J. *Becoming Lean, Inside Stories of U.S. Manufacturers;* Productivity Inc.: Portland, OR, 1998; service@produc tivityinc.com.
5. Oppenheim, B.W. *Selected Assessment Recommendations,* Industrial Assessment Center, Loyola Marymount University: Los Angeles, (boppenheim@lmu.edu), unpublished 2004.
6. *Setup Reduction for Just-in-Time,* Video/CD, Society of Manufacturing Engineers, Product ID: VT90PUB2, available at http://www.sme.org/cgi-bin/get-item.pl?VT392&2&SME&1990, (accessed on December 2005).
7. Ohno, T. *Toyota Production System: Beyond Large Scale Production;* Productivity Press: New York, 1988; info@pro ductivityinc.com.
8. Hiroyuki, H. *Five Pillars of the Visual Workplace, the Sourcebook for 5S Implementation;* Productivity Press: New York, 1995; info@productivityinc.com.

48

Global Climate Change: Carbon Sequestration

Sherwood Idso
and Keith E. Idso

Introduction

Concomitant with mankind's growing numbers and the progression of the Industrial Revolution, there has been a significant increase in the burning of fossil fuels (coal, gas, and oil) over the past 200 years, the carbon dioxide emissions from which have led to ever-increasing concentrations of atmospheric CO_2. This "large-scale geophysical experiment," to borrow the words of two of the phenomenon's early investigators,[1] is still ongoing and expected to continue throughout the current century. Furthermore, this enriching of the air with CO_2 is looked upon with great concern, because CO_2 is an important greenhouse gas, the augmentation of which is believed by many to have the potential to produce significant global warming. Therefore, and because of perceived serious consequences, such as the melting of polar ice, rising sea levels, coastal flooding, and more frequent and intense droughts, floods, and storms,[2] a concerted effort is underway to slow the rate at which CO_2 accumulates in the atmosphere, with the goal of stabilizing its concentration at a level that would prevent dangerous anthropogenic interference with the planet's climate system.

Removing Carbon from the Air

Role of Man

There are only two ways to significantly increase the natural flux of carbon from the atmosphere to the biosphere within the time frame required for effective ameliorative action if the ongoing rise in the air's CO_2 content is indeed a bona fide global warming threat: 1) increase the rate of vegetative CO_2 assimilation (photosynthesis) per unit leaf area and/or 2) increase the total plant population of the globe, i.e., leaf area per unit land area. Additionally, these things must be done without increasing the rate at which carbon is lost from the soil.

Man can do certain things to promote both of these phenomena while meeting the latter requirement as well. He can, for example, increase the rate of CO_2 assimilation per unit leaf area in agro-ecosystems by supplying additional nutrients and water to his crops. As has recently been noted, however, there are significant carbon costs associated with the production and application of fertilizers, as well as the transport of irrigation water; and factoring the CO_2 emissions of these activities into the equation often results in little net CO_2 removal from the atmosphere via these intensified agricultural interventions.[3]

Man can also draw more CO_2 out of the air by increasing the acreage of land devoted to growing crops, but this approach simultaneously releases great stores of soil carbon built up over prior centuries. When the plow exposes buried organic matter and it is oxidized, for example, prodigious amounts of CO_2 are produced and released to the atmosphere. But if a transition to less intensive tillage is made on fields that have a long history of conventional management and have thus been largely depleted of carbon, there is a good opportunity for nature to rebuild previously lost stores of soil organic matter.[4]

This approach to carbon sequestration is doubly beneficial for it results in a net removal of CO_2 from the atmosphere at the same time that it enhances a whole host of beneficial soil properties.[4,5] Also, abandoned farmlands will gradually replenish their carbon stores, both above- and below-ground, as native vegetation gradually reestablishes itself upon them. And, of course, the process can be hastened and made even more effective if trees are planted on such lands. Even without trees, it has been estimated that agricultural "best management practices" that employ conservation tillage techniques have the potential to boost the current U.S. farm and rangeland soil carbon sequestration rate of 20 million metric tons of carbon per year to fully 200 million metric tons per year,[6] which is approximately 13% of the country's yearly carbon emissions.[7]

Commercial forests also offer excellent opportunities for CO_2 removal from the air for considerable periods of time, especially when harvested wood is used to produce products that have long lifetimes. In addition, since some species of trees, such as many of those found in tropical rainforests,[8] can live in excess of a thousand years, CO_2 can be removed from the atmosphere and sequestered within their tissues—if man protects the trees from logging—until either long after the Age of Fossil Fuels has run its course or until significant changes in energy systems have reduced our dependence on fossil fuels and the CO_2 content of the air has returned to a level no longer considered problematic. Furthermore, carbon transferred to the soil beneath the trees via root exudation and turnover has the potential to remain sequestered even longer.

Role of Nature

The fact that the biosphere has maintained itself over the eons in the face of a vast array of environmental perturbations (albeit with significant modifications) suggests that earth's plant life has great resiliency and may even be able to exert a restraining influence on climate change.

A particularly important negative feedback of this type is the biosphere's ability to intensify its rate of carbon sequestration in the face of rising atmospheric CO_2 concentrations, as this phenomenon slows the rate of rise of the air's CO_2 content and thereby reduces the degree of intensification of the atmosphere's greenhouse effect. This particular climate-moderating influence of atmospheric CO_2 enrichment was first described in quantitative terms by Idso.[9,10] It begins when the aerial fertilization effect produced by the rising CO_2 content of the atmosphere elicits an increase in plant CO_2 assimilation rate per unit leaf area and when the concomitant plant water use efficiency-enhancing effect of the elevated CO_2 leads to an increase in the total plant population of the globe, due to the ability of more water-use-efficient plants to live and successfully reproduce in areas where it was formerly too dry for them to survive. In fact, these two effects are so powerful, they may actually be able to stabilize the CO_2 content of the atmosphere sometime during the current century, but only if anthropogenic CO_2 emission rates do not rise by an inordinate amount in the interim.[9,10] At the very least, together with the things man can do, they have the potential to "buy time" until other less-CO_2-emitting technologies become available.[11]

Keeping Carbon in the Soil

As more carbon is added to soils via CO_2-enhanced root growth, turnover and exudation, as well as from CO_2- induced increases in leaf litter and other decaying plant parts, the trick of significantly augmenting soil carbon sequestration is to keep at least the same percentage of this carbon in the soil as has historically been the case and to do so in the face of potential global warming.

A number of studies have addressed various aspects of this subject in recent years, with most of them finding that atmospheric CO_2 enrichment has little to no significant effect on plant litter decomposition rates. Furthermore, in nearly all of the cases where elevated CO_2 was observed to impact this phenomenon, the extra CO_2 was found to actually slow the rate of plant decomposition.[12] Much the same results have been obtained when analogous studies have used temperature as the independent variable. Warming has had either no effect on CO_2 evolution from the soil, or it has led to an actual decrease in CO_2 loss to the atmosphere.[13] Hence, the balance of evidence obtained from these studies suggests that the same—or a greater—percentage of plant material produced in a world of elevated atmospheric CO_2 concentration (and possibly higher mean air temperature) would indeed be retained in the soils of the terrestrial biosphere.

Even more compelling are the results of experiments where scientists have made direct measurements of changes in soil carbon storage under conditions of elevated atmospheric CO_2. Nearly every such study has observed increases in soil organic matter. In a Free-Air CO_2 Enrichment (FACE) experiment where portions of a cotton field were exposed to a 50% increase in atmospheric CO_2, for example, Leavitt et al.[14] found that 10% of the organic carbon present in the soil below the CO_2-enriched plants at the conclusion of the three-year experiment came from the extra CO_2 supplied to the FACE plants. In addition, some of the stored carbon had made its way into a very recalcitrant portion of the soil organic matter that had an average soil residence time of 2200 years.

Here, too, most experiments indicate that concomitant increases in temperature do not negate the increased carbon storage produced by atmospheric CO_2 enrichment. In a two-year study of perennial ryegrass grown at ambient and twice-ambient atmospheric CO_2 concentrations, as well as ambient and ambient+3°C temperature levels, for example, Casella and Soussana[15] determined that the elevated CO_2 increased soil carbon storage by 32% and 96% at low and high levels of soil nitrogen supply, respectively, "with no significant increased temperature effect." Hence, as in the case of studies of plant decomposition rates, the balance of evidence obtained from these studies also suggests that the same—or a greater—percentage of plant material produced in a world of elevated atmospheric CO_2 concentration (and possibly higher mean air temperature) would indeed be retained in the soils of the terrestrial biosphere.

Conclusions

As the air's CO_2 content continues to rise, there will almost certainly be a significant upward trend in the yearly production of terrestrial vegetative biomass, due to the growth-enhancing aerial fertilization effect of atmospheric CO_2 enrichment and the concomitant CO_2-induced increase in plant water use efficiency that enables plants to grow where it is currently too dry for them. Experimental evidence further suggests that at least the same percent- age—but in all likelihood more—of this yearly-increasing mass of plant tissue will be sequestered in earth's soils. Consequently, it is almost impossible to conclude that the carbon sequestering prowess of the planet will not be greatly enhanced in the years ahead, even without any overt actions on the part of man. Hence, if the nations of the earth were to implement even a modicum of carbon- conserving measures—such as 1)using minimum tillage techniques wherever possible in agricultural settings; 2) allowing abandoned agricultural land to revert to its natural vegetative state; 3) allowing stands of trees that can grow to very old age to actually do so; and 4) employing wise forestry practices to produce wood for making products that have long lifetimes—it is possible that

TABLE 1 Potential Rates of Carbon Sequestration (Kilograms Carbon per Hectare per Year) due to Land Management Practices That Could Be Employed for This Purpose

Improved rangeland management	50 to 150
Improved pastureland management	
Commercial fertilizer applications	100 to 200
Manure applications	200 to 500
Use of improved plant species	100 to 300
Improved grazing management	300 to 1300
Nitrogen fertilization of mountain meadows	100 to 200
Restoration of eroded soils	50 to 200
Restoration of mined lands	1000 to 3000
Conversion of cropland to pasture	400 to 1200
Conversion of cropland to natural vegetation	600 to 900
Conversion from conventional to conservation tillage	
No till	500
Mulch till	500
Ridge till	500

Source: Adapted from data reported by Follett, R.F.; Kimble, J.M.; Lal, R. The *Potential of U.S. Grazing Lands to Sequester Carbon and Mitigate the Greenhouse Effect;* Lewis Publishers, Boca Raton, FL, 2001; 1–442, and by Lal et al.[4]

the antiwarming feedback produced by the subsequent removal of CO_2 from the atmosphere would be sufficient to keep the risk of potential greenhouse gas-induced global warming at an acceptable level. Estimates of the carbon-sequestering power of some of these "best management practices" are given in Table 1.

References

1. Revelle, R.; Suess, H.E. Carbon dioxide exchange between atmosphere and ocean and the question of an increase of atmospheric CO_2 during the past decades. Tellus **1957**, *9,* 18–27.
2. Intergovernmental panel on climate change. In *Climate Change 2001: The Scientific Basis, Summary for Policy Makers and Technical Summary of the Working Group I Report*; Cambridge University Press: Cambridge, U.K., 2001; 1–98.
3. Schlesinger, W.H. Carbon sequestration in soils: some cautions amidst optimism. Agric. Ecosys. Environ. **2000**, *82,* 121–127.
4. Lal, R.; Kimble, J.M.; Follett, R.F.; Cole, C.V. *The Potential for U.S. Cropland to Sequester Carbon and Mitigate the Greenhouse Effect*; Sleeping Bear Press: Chelsea, MI, 1998; 1–128.
5. Idso, S.B. *Carbon Dioxide and Global Change: Earth in Transition*; IBR Press: Tempe, AZ, 1989; 1–292.
6. Jawson, M.D.; Shafer, S.R. Carbon credits on the Chicago board of trade? Agric. Res. **2001**, *49* (2), 2.
7. Comis, D.; Becker, H.; Stelljes, K.B. Depositing carbon in the bank. Agric. Res. **2001**, *49* (2), 4–7.
8. Chambers, J.Q.; Higuchi, N.; Schimel, J.P. Ancient trees in Amazonia. Nature **1998**, *391,* 135–136.
9. Idso, S.B. The aerial fertilization effect of CO_2 and its implications for global carbon cycling and maximum greenhouse warming. Bull. Amer. Meteorol. Soc. **1991**, *72,* 962–965.
10. Idso, S.B. Reply to comments of L.D. Danny Harvey, Bert Bolin, and P. Lehmann. Bull. Amer. Meteorol. Soc. **1991**, *72*, 1910–1914.
11. Izaurralde, R.C.; Rosenberg, N.J.; Lal, R. Mitigation of climatic change by soil carbon sequestration: issues of science, monitoring, and degraded lands. Adv. Agron. **2001**, *70,* 1–75.

12. Nitschelm, J.J.; Luscher, A.; Hartwig, U.A.; van Kessel, C. Using stable isotopes to determine soil carbon input differences under ambient and elevated atmospheric CO_2 conditions. Global Change Biol. **1997**, *3*, 411–416.
13. van Ginkel, J.H.; Whitmore, A.P.; Gorissen, A. *Lolium perenne* grasslands may function as a sink for atmospheric carbon dioxide. J. Environ. Quality **1999**, *28,* 1580–1584.
14. Leavitt, S.W.; Paul, E.A.; Kimball, B.A.; Hendrey, G.R.; Mauney, J.R.; Rauschkolb, R.; Rogers, H.; Lewin, K.F.; Nagy, J.; Pinter, P.J., Jr.; Johnson, H.B., Jr. Carbon isotope dynamics of free-air CO_2-enriched cotton and soils. Agric. For. Meteorol. **1994**, *70,* 87–101.
15. Casella, E.; Soussana, J.-F. Long-term effects of CO_2 enrichment and temperature increase on the carbon balance of a temperate grass sward. J. Exp. Bot. **1997**, *48,* 1309–1321.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

49

Global Climate Change: Earth System Response

Amanda Staudt and
Nathan E. Hultman

Introduction

For the past century, Earth's climate has been changing due to human activities. Observations show that Earth's surface warmed by approximately 0.6°C (1.1°F) on average in the 20th century. Much of this warming has been attributed to increasing abundances of greenhouse gases emitted to the atmosphere by human activities, although it is difficult to quantify this contribution against the backdrop of natural variability and climate-forcing uncertainties. Atmospheric abundances of the major anthropogenic greenhouse gases (carbon dioxide; methane; nitrous oxide; halocarbons manufactured by humans, such as chlorofluorocarbons; and tropospheric ozone) reached their highest recorded levels at the end of the 20th century, and all but methane have continued to rise. Major causes of this rise have been fossil fuel use, agriculture, and land-use change.

The emerging impacts of climate change on natural systems include melting glaciers and ice caps, the rising sea level, extended growing seasons, changes in precipitation regimes, and changes in the geographical distributions of plant and animal species. Additional impacts, to which it may be difficult for human and natural systems to adapt, could arise from events whose triggers are poorly understood. Human-induced global warming will continue during the 21st century and beyond, because many parts of the Earth's system respond slowly to changes in greenhouse gas levels and because altering established energy-use practices is difficult. Uncertainties remain about the magnitude and the impacts of future climate change, largely due to gaps in understanding of climate science and the socioeconomic drivers of climate change.

The Climate System and the Natural Greenhouse Effect

While climate conventionally has been defined as the long-term statistics of the weather (e.g., temperature, cloudiness, precipitation), improved understanding of the atmosphere's interactions with the oceans, the cryosphere (ice-covered regions of the world), and the terrestrial and marine biospheres has led scientists to expand the definition of climate to encompass the oceanic and terrestrial spheres as well as chemical components of the atmosphere (Figure 1). Physical processes within the atmosphere are influenced by ocean circulation, the reflectivity of Earth's surface, the chemical composition of the atmosphere, and vegetation patterns, among other factors.

The Sun provides almost all of Earth's energy. Solar radiation intercepted by Earth first encounters the atmosphere, which allows most of it to pass to Earth's surface. The intensity of radiation at the surface depends on the amount of incident radiation and on the orientation of the surface with respect to that radiation. The surface either reflects or absorbs this incoming radiation. Different surfaces reflect different amounts of sunlight. The fraction of solar energy reflected is defined as a surface's albedo. Albedos range from about 10% for open water, dark soil, and asphalt to about 80% for fresh snow.[1] Earth's average albedo is about 31%.

Although we do not normally think of it as a radiative body, Earth—like all bodies with a nonzero temperature—emits electromagnetic radiation. For Earth's temperature, most of this radiation is in the form of infrared light. In the absence of an atmosphere, all the radiation emitted by Earth would escape to space. The balance of incoming solar radiation and outgoing infrared radiation would result in a global-mean temperature for Earth of 255 K (–18°C/0°F). However, some molecules in Earth's atmosphere absorb some of this outgoing infrared light, thereby increasing their temperature. These greenhouse gases in the atmosphere emit some energy back toward Earth, warming Earth's surface. This natural

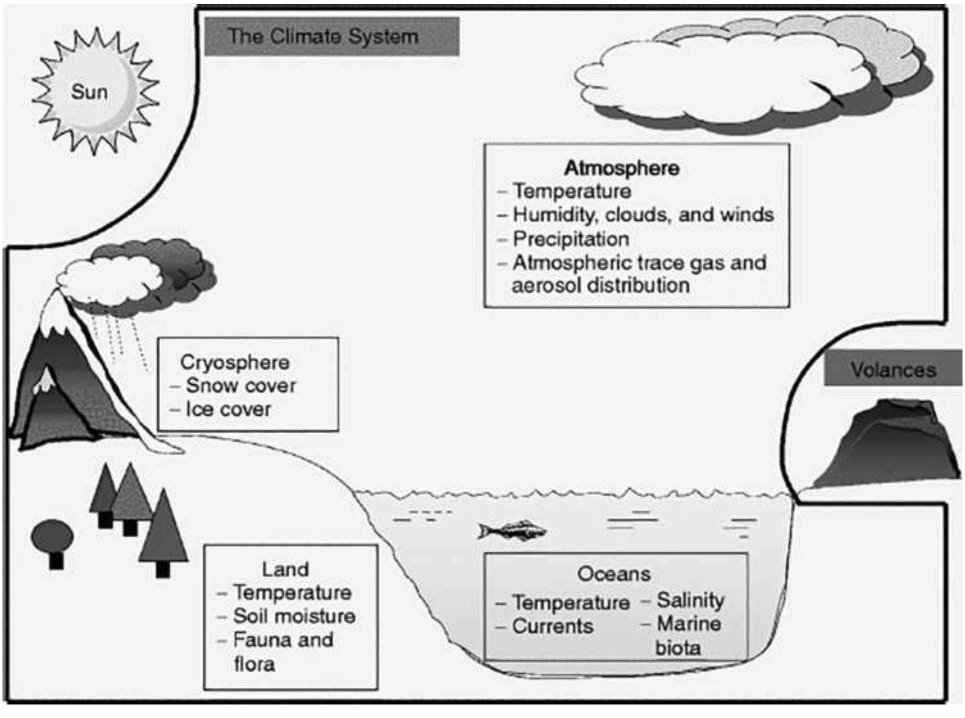


FIGURE 1 The climate system.
Source: National Academies Press (see NRC[5]).

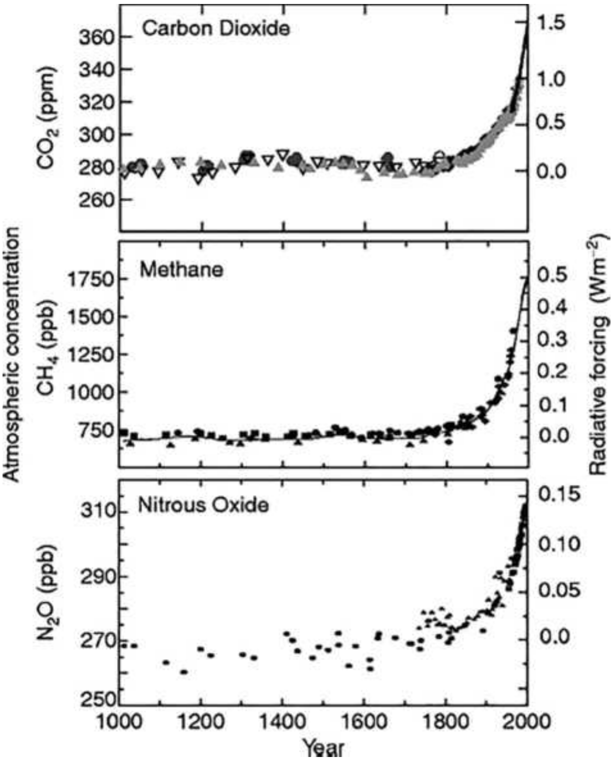


FIGURE 2 Concentrations of major greenhouse gases retrieved from gas bubbles trapped in ice cores from Antarctica and Greenland.
Source: Intergovernmental Panel on Climate Change (see IPCC[10]).

greenhouse effect, which is present in the absence of human activities, raises the global-mean surface temperature from 255 K to a comfortable 288 K (or about 15°C/59°F).[2]

Greenhouse gases that are present naturally in the atmosphere include water vapor (H_2O), carbon dioxide (CO_2), methane (CH_4), nitrous oxide (N_2O), and ozone (O_3). The most common greenhouse gas by quantity and the one exerting the greatest influence on the climate is water vapor; however, because water has a very short lifetime in the atmosphere (~1 week), any human perturbation will dissipate quickly. In most cases, the "greenhouse effect" or "climate change" refers not to this natural phenomenon but to additional, anthropogenic enhancements to the atmosphere's capacity to trap heat. Much higher concentrations of CO_2, CH_4, and N_2O have been observed in the past century than were naturally present for the past 1000 years (Figure 2) and likely much longer.[3] Earth's surface is warmer now on average than it was at any time during the past 400 years, and it is likely warmer now than it was at any time in the past 2000 years.[4]

Climate Forcings and Feedbacks

Factors that affect climate change are usefully separated into forcings and feedbacks. Climate forcings are energy imbalances imposed on the climate system either externally or by human activities.[5] Examples include human- caused emissions of greenhouse gases, as discussed in the preceding section, as well as changes in solar energy input; volcanic emissions; deliberate land modification; or anthropogenic emissions of aerosols, which can absorb and scatter radiation. Climate forcings can be either direct or indirect. Direct radiative forcings are simple changes to the drivers of Earth's radiative balance. For

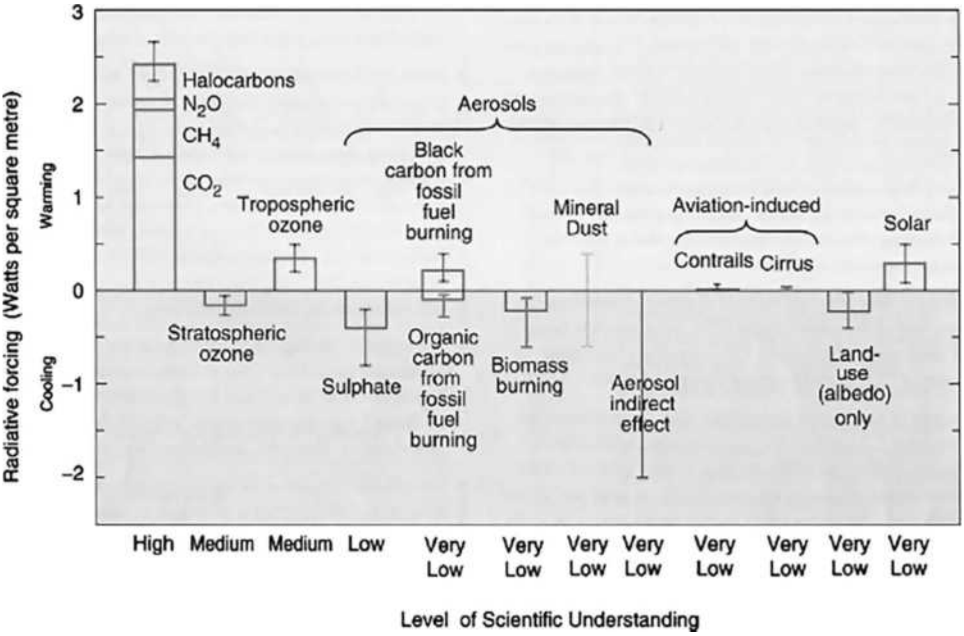


FIGURE 3 Estimated radiative forcings since preindustrial times for Earth and the troposphere system. The height of the rectangular bar denotes a central or best estimate of the forcing, while each vertical line is an estimate of the uncertainty range associated with the forcing, guided by the spread in the published record and physical understanding, and with no statistical connotation. Each forcing agent is associated with a level of scientific understanding, which is based on an assessment of the nature of assumptions involved, the uncertainties prevailing about the processes that govern the forcing, and the resulting confidence in the numerical values of the estimate. On the vertical axis, the direction of expected surface temperature change due to each radiative forcing is indicated by the labels "warming" and "cooling."
Source: Intergovernmental Panel on Climate Change (see IPCC[10]).

example, added CO_2 absorbs and emits infrared radiation. Indirect radiative forcings create a radiative imbalance by first altering climate system components that lead to consequent changes in radiative fluxes; an example is the effect of aerosols on the precipitation efficiency of clouds. Figure 3 provides a summary of the estimated contribution from major climate forcings. Additional information about specific climate forcings is provided in the discussion below.

Climate feedbacks are internal climate processes that amplify or dampen the climate response to an initial forcing.[6] An example is the increase in atmospheric water vapor that is triggered by an initial warming due to rising CO_2 concentrations, which then acts to amplify the warming through the greenhouse properties of water vapor (Figure 4). Other climate feedbacks involve snow and ice cover, biogeochemistry, clouds, and ocean circulation. Some of the uncertainty about how the climate will change in the future stems from unresolved research questions on climate change feedbacks.

Natural Climate Forcings: Solar and Volcanic Variability

Variations in the Sun's activity and in Earth's orbital parameters cause natural forcing of climate. Radiometers on various spacecraft have been measuring the total solar irradiance since the late 1970s. There is an 11-year cycle in total solar irradiance of peak-to-peak amplitude ~1 W m^{-2} (0.1%) in the past three cycles. Allowing for reflection of 30% of this incident energy (Earth's albedo) and

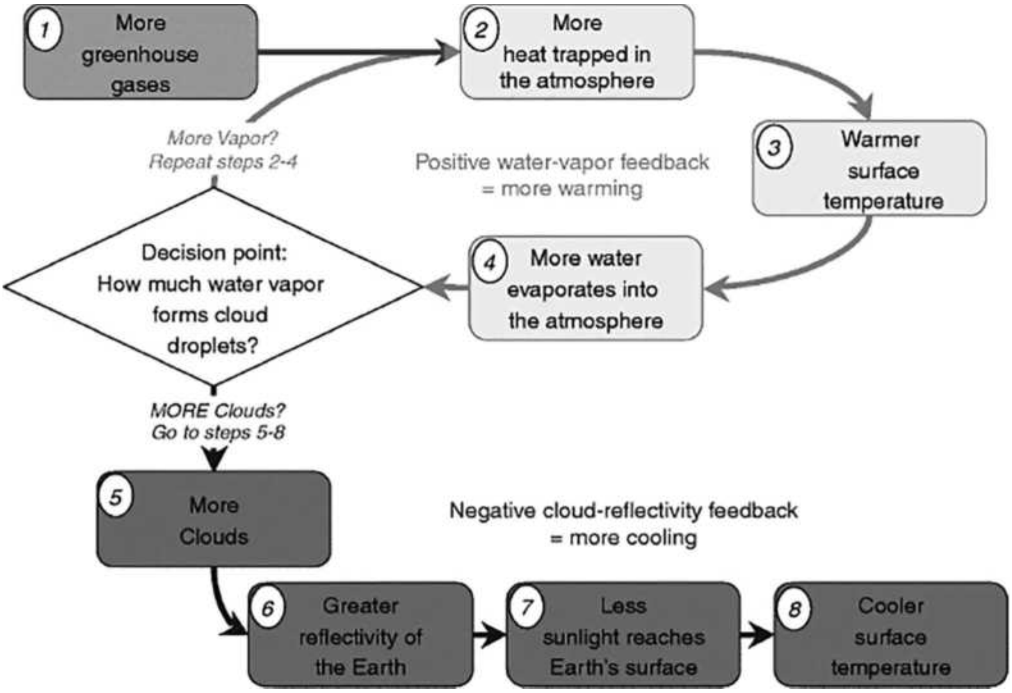


FIGURE 4 This schematic illustrates just two out of the dozens of climate feedbacks identified by scientists. The warming created by emitting more greenhouse gases leads to evaporation of water into the atmosphere. But water itself is a greenhouse gas and can cause even more warming via positive water-vapor feedback. On the other hand, if the water vapor leads to more clouds, the result could be to counteract some of the warming because clouds can reflect incoming sunlight back to space. This chain of events is negative cloud- reflectivity feedback. Trying to understand whether water vapor will create more clouds and what kinds of clouds water vapor will create is a major research objective right now. The answer depends on weather patterns, where the evaporation takes place, and the amount of small soot particles suspended in the air.

averaging over the globe, the corresponding climate forcing is of order 0.2 W m^{-2}, although recent analyses have found little secular trend in solar irradiance over the past 30 years.[7] Knowledge of solar irradiance variations prior to 1979 is less certain, as it relies upon models of how sunspot and facular influences relate to solar irradiance observed since then. These models are used to extrapolate variations back to about 1610, when telescopes were first used to monitor sunspots. The amount of energy Earth receives from the Sun also depends on Earth's distance from the Sun, which does not remain constant. The eccentricity of Earth's orbit (currently 0.0167) and the tilt of its axis relative to the orbital plane result in continual changes to the amount and distribution of energy Earth receives. In modern times, this variation is ±3.5% during the year, with maximum energy and minimum distance in January.

Volcanic forcing has been the dominant source of natural global radiative forcing over the past millennium. Emissions from volcanic eruptions have multiple effects on climate, as listed in Table 1.[8] The greater prevalence of explosive volcanic activity during both the early and the late 20th century and the dearth of eruptions from 1915 to 1960 represent a significant natural radiative forcing of 20th century climate.[9] Similarly, longer-term volcanic radiative forcing has been associated with a significant long-term forced cooling from 1000 to 1900, resulting from a general increase in explosive volcanic activity in later centuries.

TABLE 1 Effects of Large Explosive Volcanoes on Weather and Climate

Effect and Mechanism	Begins	Duration
Reduction of diurnal cycle	Immediately	1–4 days
	Blockage of shortwave and emission of longwave radiation	
Reduced tropical precipitation	1–3 months	3–6 months
	Blockage of shortwave radiation, reduced evaporation	
Summer cooling of Northern Hemisphere tropics and subtropics	1–3 months	1–2 years
	Blockage of shortwave radiation	
Reduced Sahel precipitation	1–3 months	1–2 years
	Blockage of shortwave radiation, reduced land temperature, reduced evaporation	
Stratospheric warming	1–3 months	1–2 years
	Stratospheric absorption of shortwave and longwave radiation	
Winter warming of Northern Hemisphere continents	6–18 months	1 or 2 winters
	Stratospheric absorption of shortwave and longwave radiation, dynamics	
Global cooling	Immediately	1–3 years
	Blockage of shortwave radiation	
Global cooling from multiple eruptions	Immediately	Up to decades
	Blockage of shortwave radiation	
Ozone depletion, enhanced UV radiation	1 day	1–2 years
	Dilution, heterogeneous chemistry on aerosols	

Source: Reproduced/modified by permission of the American Geophysical Union (see Robock[8]).

Greenhouse Gas Forcing

The role of greenhouse gases in the climate system is well understood by scientists because instruments can accurately measure the abundances of these gases in the atmosphere and their radiative properties. The concentrations of CO_2, CH_4, $N_2$0, various halocarbons, and O_3 have increased substantially since preindustrial times, and they are the greatest contributors to total anthropogenic radiative forcing.[10] Many of these greenhouse gases are emitted primarily as a byproduct of fossil fuel combustion.

For a given gas, the total amount of heat-trapping ability depends on the efficiency of heat trapping for a given unit of gas (i.e., radiative forcing), the number of units present in the atmosphere, and the average length of time a given unit spends in the atmosphere. While these three components are enough to characterize a single gas, the large number of gases has prompted the development of an index called the global warming potential (GWP), which represents the relative impact of a particular greenhouse gas on the atmosphere's radiative balance.[10] See Table 2 for some GWP calculations. As the standard reference gas, CO_2 has a GWP of 1, by definition. Over a time horizon of 100 years, CH_4 and N_2O have GWPs of 23 and 296, respectively. In other words, 1 additional kg of CH_4 in the atmosphere absorbs as much radiation as 23 additional kg of CO_2. However, these numbers change if the time horizon shifts.[10] By allowing greenhouse gases to be compared directly, GWPs enable policies that can reduce total climate impact by addressing the least-cost abatement options first.[11,12]

TABLE 2 Radiative Forcing Characteristics of Some Major Greenhouse Gases

	Contribution to Direct Radiative Forcing				Global Warming Potential (GWP) for Different Time Horizons		
	Wm2	%	Concentration in 1998	Lifetime (Years)	20 Years	100 Years	500 Years
Carbon dioxide (CO_2)	1.46	60	365 ppm	5–200a	1	1	1
Methane (CH_4)	0.48	20	1745 ppb	12.0	62	23	7
Nitrous oxide (N_2O)	0.15	6	314 ppb	114	275	296	156
Halocarbons and related compounds	0.34	14	—	0.3–3200	40–15100	12–22200	4–16300

Source: Intergovernmental Panel on Climate Change (see IPCC[10]).

Total direct radiative forcing uncertainty is approximately 10%. The abbreviations parts per million (ppm) and parts per billion (ppb) refer to the ratio of greenhouse gas molecules to molecules of dry air.

[a] No single lifetime can be defined for CO2 because it is affected by multiple removal processes with different uptake rates.

Atmospheric Aerosol Forcing

Aerosols are small particles or liquid droplets suspended in the atmosphere. Aerosols both scatter and absorb radiation, representing a direct radiative forcing. Scattering generally dominates (except for black carbon particles) so that the net effect is of cooling. The average global mean of aerosol-direct forcing from fossil fuel combustion and biomass burning is in the range of -0.2 to -2.0 W m^{-2}.[10] This large range results from uncertainties in aerosol sources, composition, and properties used in different models. Recent advances in modeling and measurements have provided important constraints on the direct effect of aerosols on radiation.[13–15] Aerosols have several indirect effects on climate, all arising from their interaction with clouds— particularly from their roles as cloud condensation nuclei (CCN) and ice nuclei (Table 3).

TABLE 3 Overview of the Different Aerosol Indirect Effects Associated with Clouds

Effect	Cloud Type	Description	Sign of Top of the Atmosphere Radiative Forcing
First indirect aerosol effect (cloud albedo or Twomey effect)	All clouds	For the same cloud water or ice content more but smaller cloud particles reflect more solar radiation	Negative
Second indirect aerosol effect (cloud lifetime or Albrecht effect)	All clouds	Smaller cloud particles decrease the precipitation efficiency, thereby prolonging cloud lifetime	Negative
Semidirect effect	All clouds	Absorption of solar radiation by soot leads to evaporation of cloud particles	Positive
Glaciation indirect effect	Mixed-phase clouds	An increase in ice nuclei increases the precipitation efficiency	Positive
Thermodynamic effect	Mixed-phase clouds	Smaller cloud droplets inhibit freezing, causing super cooled droplets to extend to colder temperatures	Unknown
Surface energy budget effect	All clouds	The aerosol-induced increase in cloud optical thickness decreases the amount of solar radiation reaching the surface, changing the surface energy budget	Negative

Source: National Academies Press (see NRC[5]).

Land-Use Change Forcing

Land-use changes include irrigation, urbanization, deforestation, desertification, reforestation, the grazing of domestic animals, and dryland farming. Each of these alterations in landscape produces changes in radiative forcing, both directly and indirectly.[16,17] Direct effects include the change of albedo and emissivity resulting from the different types of land covers. For example, the development of agriculture in tropical regions typically results in an increase of albedo from a low value of forest canopies (0.05–0.15) to a higher value of agricultural fields, such as pasture (0.15–0.20). The Intergovernmental Panel on Climate Change (IPCC)[10] reports the globally averaged forcing due to albedo change alone as -0.25 ± 0.25 W m^{-2}. Significant uncertainties remain in estimating the effect of land-use change on albedo because of the complexity of land surfaces (e.g., the type of vegetation, phenology, density of coverage, soil color).

Indirect effects of land-cover change on the net radiation include a variety of processes related to (1) the ability of the land cover to use the radiation absorbed at the ground surface for evaporation, transpiration, and sensible heat fluxes (the impact on these heat fluxes caused by changes in land cover is sometimes referred to as thermodynamic forcing); (2) the exchange of greenhouse and other trace gases between the surface and the atmosphere; (3) the emission of aerosols (e.g., from dust); and (4) the distribution and melting of snow and ice.[5] These effects are not yet well characterized or quantified.

Evidence of Human-Induced Climate Change

Because we do not have a “control Earth” against which to compare the effects of our current changing atmosphere, incontrovertibly linking human activities and observed climate change is difficult. Scientists therefore rely on multiple, overlapping evidence of changes and then compare observed patterns of change with what our scientific understanding indicates should happen under anthropogenic climate change. This two-stage concept of discovering changes in climate and linking them to human activity is called detection and attribution. Evidence used to detect climate change is summarized in this section, and the use of climate models for attribution is discussed in the following section.

One piece of evidence of global warming is an increase in surface temperature since the 1900s, with particularly rapid increases since the late 1970s (Figure 5). This dataset caused some controversy when researchers discovered that readings taken near the surface of Earth with thermometers appeared to be higher than readings of the lower atmosphere taken by satellites from above. Subsequent studies concluded that the warming trend in global-mean surface temperature observations during the past 30 years is undoubtedly real and is substantially greater than the average rate of warming during the 20th century.[18] Satellite-and balloon-based observations of middle-troposphere temperatures, after several revisions of the data, now compare reasonably with one another and with observations from surface stations, although some uncertainties remain.[19,20]

The ocean, which represents the largest reservoir of heat in the climate system, has warmed by about 0.118°C (0.212°F), averaged over the layer extending from the surface down to 700 m, from 1955 to 2003 (Figure 6).[21] Approximately 84% of the total heating of Earth’s system (oceans, atmosphere, continents, and cryosphere) over the past 40 years has gone into warming the oceans. Recent studies have shown that the observed heat storage in the oceans is what would be expected by a human-enhanced greenhouse effect. Indeed, increased ocean heat content accounts for most of the planetary energy imbalance (i.e., when Earth absorbs more energy from the Sun than it emits back to space) simulated by climate models.[22]

Changes in several other climate indicators have been observed over the past decades, providing a growing body of evidence consistent with a human impact on the climate. For example, reductions

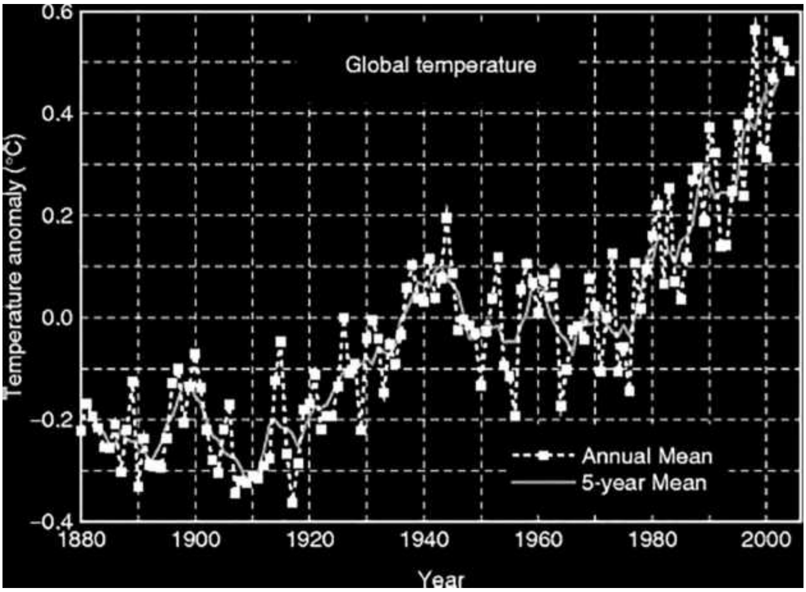


FIGURE 5 Global annual-mean surface air temperature change derived from the meteorological station network. Data and plots are available from the Goddard Institute for Space Sciences (GISS) at http://data.giss.nasa.gov/gistemp/graphs.

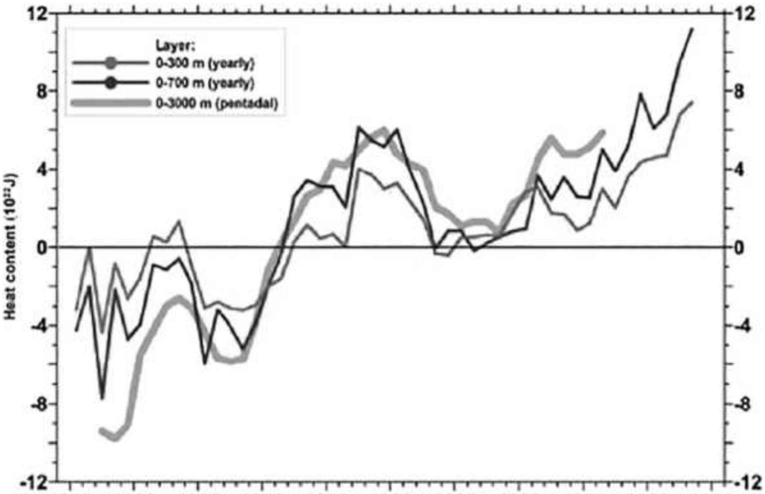


FIGURE 6 Time series of (i) yearly ocean heat content (10^{22} J) for the 0–300 m and 0–700 m layers and (ii) 5-year running averages for 1955–1959 through 1994–1998 for the 0–3000 m layer.
Source: Reproduced/modified by permission of the American Geophysical Union (see Levitus[21])).

in snow and ice cover are one important indicator.[10] Satellite observations indicate that snow cover has decreased by about 10% since the 1960s, while spring and summer sea-ice extent in the Northern Hemisphere has decreased by about 10%–15% since the 1950s. The shrinking of mountain glaciers in many nonpolar regions has also been observed during the 20th century.

Attribution of Observed Climate Change to Human Influence

An important question in global climate change is to what extent the observed changes are caused by the emissions of greenhouse gases and other human activities. Climate models are used to study how the climate operates today, how it may have functioned differently in the past, and how it may evolve in response to forcings. Built using our best scientific knowledge of atmospheric, oceanic, terrestrial, and cryospheric processes, climate models and their components are extensively tested against the full suite of observations of current and past climate to verify that they simulate a realistic version of the climate. Discrepancies between models and observations provide indications that we need to improve understanding of physical processes, model representations of the processes, or in some cases the observations themselves. Hence, climate models contain our accumulated wisdom about the underlying scientific processes and can be no better than our observations of the system and our understanding of the climate.

Figure 7 shows how scientists have used climate models to make the case that human activities have perturbed the climate since preindustrial times. In this experiment, the model is run with three different sets of climate forcings: (a) natural only, (b) anthropogenic only, and (c) natural and anthropogenic. When the natural or anthropogenic forcings are employed separately, the model is unable to reproduce the global-mean variation of temperature anomalies over the simulated time period. Only when both sets of forcings are used does the model capture the nature of the variations, providing evidence that human activities have caused a significant fraction of warming in the past 150 years.[10]

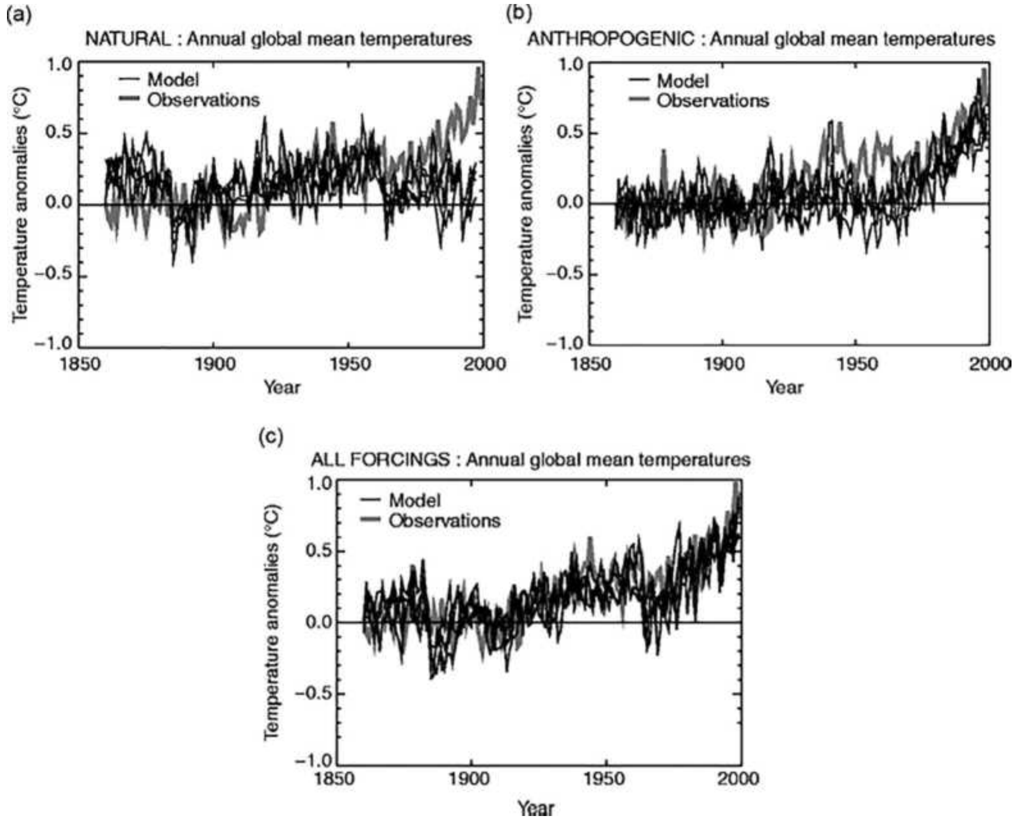


FIGURE 7 Climate model results with (a) solar and volcanic forcings only; (b) anthropogenic forcings only; and (c) all forcings, both natural and anthropogenic.

Source: Intergovernmental Panel on Climate Change (see IPCC[10]).

Projections for Future Climate Change

The IPCC has concluded that by 2100, global surface temperatures will likely be from 1.4 to 5.8°C (2.5°F–10.4°F) above 1990 levels (Figure 8) and that the combined effects of ice melting and seawater expansion from ocean warming will cause the global-mean sea level to rise by between 0.1 and 0.9 m.[10] Uncertainties remain about the magnitude and impacts of future climate change, largely due to gaps in understanding of climate science and the difficulty of predicting societal choices.

Climate changes in the coming century will not be uniformly distributed; some regions will experience more warming than others. There will be winners and losers from the impacts of climate change, even within a single region, but globally, the losses are expected to far outweigh the benefits. A changed climate will increase the likelihood of extreme heat and drought events.[23] High latitudes and polar regions are expected to see comparatively greater increases in average temperatures than lower latitudes, resulting in melting of permafrost and sea ice, which will result in additional costs for residents and in disruption to wildlife and ecosystems.[24] Precipitation changes, which are of great importance to agriculture, may have even more regional variability that is hard to predict.

Finally, several elements of Earth's system seem to be vulnerable to rapid destabilization. For example, the West Antarctic ice sheet and the Greenland ice sheet may be more prone to rapid melting than

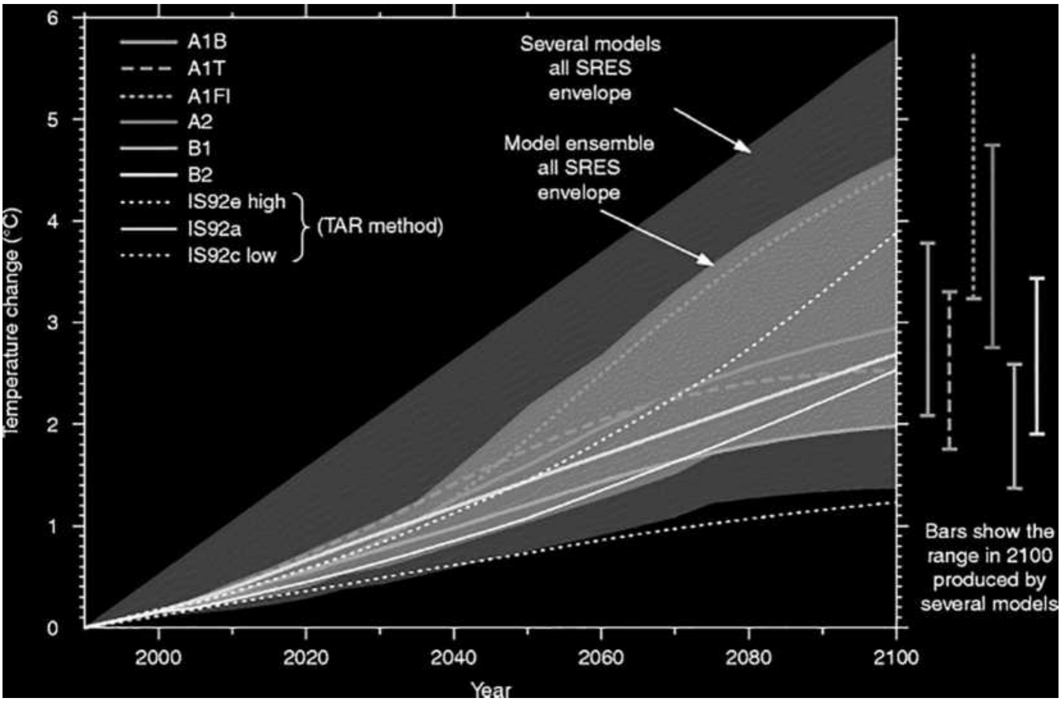


FIGURE 8 Climate models are often used to simulate possible future climates to help inform decisions about policy responses to potential climate changes. This figure shows the range of plausible global-mean temperature change over the next 100 years, simulated by a collection of models. The spread in 2100 temperatures from 1.4 to 5.8°C reflects two factors: (1) Each model was run multiple times using different scenarios (indicated by different line textures) for future climate forcings, and (2) each model makes different assumptions about how the climate responds to those forcings. The scenarios range from those that assume continued acceleration of greenhouse gas emissions to those that assume more moderate growth or leveling off of emissions rates. SRES refers to the collection of scenarios presented in the Special Report on Emissions Scenarios. TAR refers to the IPCC Third Assessment Report.

Source: Intergovernmental Panel on Climate Change (see IPCC[10]).

previously thought, and the loss of either of these would result in a large sea-level rise greater than 5 m. Moreover, the stability of the oceanic circulation that brings heat to Northern Europe has also been questioned. Because of feedback processes and the large uncertainty in system sensitivity, these outcomes are not easy to model and are usually not included in the gradual climate change projections quoted above. Nevertheless, they are nontrivial threats and represent active areas of current research.

Unfortunately, the regions that will be most severely affected are often the regions that are the least able to adapt. Bangladesh, one of the poorest nations in the world, is projected to lose 17.5% of its land if sea level rises about 1 m (40 in.), displacing tens of thousands of people.[10] Several islands throughout the South Pacific and Indian Oceans will be at similar risk for increased flooding and vulnerability to storm surges. Although wetland and coastal areas of many developed nations—including the United States—are also threatened, wealthy countries may be more able to adapt to sea-level rise and threats to agriculture. Solutions could include limiting or changing construction codes in coastal zones and developing new agricultural technologies.

Conclusion

Research conducted to understand how the climate system may be changing—and in turn affecting other natural systems and human society—has led to significant advancement in scientific understanding, but many questions remain. Society faces increasing pressure to decide how best to respond to a changing climate and associated global and regional changes.

One way to address global climate change is to take steps to reduce the amount of greenhouse gases in the atmosphere. Because CO_2 and other greenhouse gases can remain in the atmosphere for many decades, the climate-change impacts from concentrations today will likely continue throughout the 21st century and beyond. Failure to implement significant reductions in net greenhouse gas emissions now will make the job much harder in the future—both in terms of stabilizing CO_2 concentrations and in terms of experiencing more significant impacts. While no single solution can eliminate all future warming, many potentially cost-effective technological options could contribute to stabilizing greenhouse gas concentrations. These options range from personal choices such as driving less to national choices such as regulating emissions or seeking technologies to remove greenhouse gases from the atmosphere to international choices such as sharing energy technologies.

At the same time, it will be necessary to seek ways to adapt to the potential impacts of climate change. Climate is becoming increasingly important to public and private decision-making in various fields such as emergency management, water-quality assurance, insurance, irrigation and power production, and construction. For example, developing practical, “no regrets” strategies that could be used to reduce economic and ecological systems’ vulnerabilities to change could provide benefits whether a significant climate change ultimately occurs or not. No-regrets measures could include low-cost steps to improve climate forecasting; to slow biodiversity loss; to improve water, land, and air quality; and to make institutions—such as the health care enterprise, financial markets, and transportation systems—more resilient to major disruptions.

References

1. Houghton, J. *Global Warming: The Complete Briefing,* 3rd Ed.; Cambridge University Press: Cambridge, 2004; 351.
2. Houghton, J. *The Physics of Atmospheres,* 3rd Ed.; Cambridge University Press: Cambridge, U.K., 2002.
3. Petit, J.R.; Jouzel, J.; Raynaud, D.; et al. Climate and atmospheric history of the past 420,000 years from the Vostok ice core, Antarctica. Nature, **1999**, *399,* 429–436.
4. NRC (National Research Council). *Surface Temperature Reconstructions for the Last 2000 Years;* National Academies Press: Washington, DC, 2006.

5. NRC (National Research Council). *Radiative Forcing of Climate Change: Expanding the Concept and Addressing Uncertainties*; National Academies Press: Washington, DC, 2005.
6. NRC (National Research Council). *Understanding Climate Change Feedbacks*; National Academies Press: Washington, DC, 2003.
7. Lean, J.; Rottman, G.; Harder, J.; et al. SORCE contributions to new understanding of global change and solar variability. Solar Phy. **2005**, *230,* 27–53.
8. Robock, A. Volcanic eruptions and climate. Rev. Geophys. **2000**, *38,* 191–219.
9. Crowley, T.J. Causes of climate change over the past 1000 years. Science **2000**, *289*, 270–277.
10. IPCC (Intergovernmental Panel on Climate Change). In *Climate Change 2001: The Scientific Basis;* Houghton, J.T., Ding, Y., Griggs, D.J., Eds.; Contribution of Working Group 1 to the Third Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press: Cambridge, U.K., 2001.
11. Prinn, R.; Jacoby, H.; Sokolov, A.; et al. Integrated global system model for climate policy assessment: Feedbacks and sensitivity studies. Climatic Change **1999**, *41* (3–4), 469–546.
12. Schneider, S.H.; Goulder, L.H. Achieving low-cost emissions targets. Nature **1997**, *389* (6646), 13–14.
13. Ramanathan, V.; Crutzen, P.J.; Kiehl, J.T.; et al. Aerosols, climate and the hydrological cycle. Science **2001**, *294,* 2119–2124.
14. Russell, L.M.; Seinfeld, J.H.; Flagan, R.C.; et al. Aerosol dynamics in ship tracks. J. Geophys. Res. Atmos. **1999**, *104* (D24), 31077–31095.
15. Conant, W.C.; Seinfeld, J.H.; Wang, J.; et al. A model for the radiative forcing during ACE-Asia derived from CIRPAS Twin Otter and R/V Ronald H. Brown data and comparison with observations. J. Geophys. Res. **2003**, *108* (D23), 8661.
16. Pitman, A.J. Review: The evolution of, and revolution in, land surface schemes designed for climate models. Int. J. Climatol. **2003**, *23,* 479–510.
17. Kabat, P., Claussen, M., Dirmeyer, P.A. et al., Eds.; *Vegetation, Water, Humans and the Climate—A New Perspective on an Interactive System*, Berlin, 2004.
18. NRC (National Research Council) *Reconciling Observations of Global Temperature Change*; National Academy Press: Washington, DC, 2000.
19. Mears, C.A.; Wentz, F.J. The effect of diurnal correction on satellite-derived lower tropospheric temperature. Science **2005**, *309,* 1548–1551.
20. Sherwood, S.C.; Lanzante, J.R.; Meyer, C.L. Radiosonde daytime biases and late-20th century warming. Science **2005**, *309,* 1556–1559.
21. Levitus, S.; Antonov, J.; Boyer, T. Warming of the world ocean, 1955–2003. Geophys. Res. Lett. **2005**, *32* (L02604).
22. Hansen, J.; Nazarenko, L.; Ruedy, R.; et al. Earth's energy imbalance: Confirmation and implications. Science **2005**, *308,* 1431–1435.
23. Meehl, G.A.; Tebaldi, C. More intense, more frequent, and longer lasting heat waves in the 21st century. Science **2004**, *305,* 994–997.
24. ACIA (Arctic Climate Impact Assessment) *Impacts of a Warming Arctic: Arctic Climate Impact Assessment*; Cambridge University Press Cambridge, U.K., 2004.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

50

Global Climate Change: Gas Fluxes

Pascal Boeckx
and Oswald Van
Cleemput

Sampling Techniques, Sample Handling, and Analysis

Chamber Techniques

Flux chambers are simple inverted containers, which form an enclosure for gases emitted from the soil surface.[1] Both closed (static) and open (dynamic) chambers can be used. Advantages and disadvantages of chamber techniques are listed in Table 1.

TABLE 1 Advantages and Disadvantages of the Closed Chamber, Open Chamber, and Micrometeorological Methods to Measure Gas Fluxes from Soils

Method	Advantage	Disadvantage
Closed chamber	Simple and low cost	Labor intensive
	Multiple gases can be sampled	Small area is covered
	Small fluxes can be measured	Only a short-term emission event is monitored (1–2 h)
	Manual and automated gas sampling can be used	Disturbance of the emitting surface upon installation
		Altered conditions of temperature and soil atmosphere exchange
		Different functioning of plants in the chamber
Open chamber	Relatively simple	Small area is covered
	Environmental condition close to uncovered field	Disturbance of the emitting surface upon installation

(*Continued*)

TABLE 1 (*Continued*) Advantages and Disadvantages of the Closed Chamber, Open Chamber, and Micrometeorological Methods to Measure Gas Fluxes from Soils

Method	Advantage	Disadvantage
	Continuous long-term monitoring possible	Pressure deficits can cause artificially high fluxes
		Automated sampling is required
Micrometeorological	Useful for diurnal and seasonal variations	Expensive and sophisticated instrumentation needed
	Large areas can be monitored (aggregate flux)	Dependence on a uniform, large surface and constant atmospheric conditions
	Minimal disturbance of the emitting surface	

Closed and Open Chambers

Emissions of CO_2, CH_4, and N_2O are very variable, both spatially and temporally. It has not been determined whether the size of the flux chambers has an influence on this variability. Nevertheless, at least six chambers should be used per campaign. Flux chambers (cylindrical, or square or rectangular and box-like) can either be installed as a complete assembly, eventually for a short period and then removed until the next sampling occasion (Figure 1a), or they can exist out of a basal part, which is installed for the entire duration of the experiment with a gas-tight chamber attached to it for short periods (Figure 1b). This last variant is often used in flooded systems (e.g., paddy soils). The normal procedure for installation is to make a slit in the soil with a metal cutting edge, the same size and shape as the collar of the chamber, and to insert the chamber collar for about 3 cm into the slit.

When a closed chamber is fixed in place, gas samples can be taken from the headspace at different time intervals. The change in concentration in the chamber over time is used to calculate the gas flux. The calculation of the flux goes through a linear regression analysis of the gas concentration increase with time (corrected for eventual temperature changes) and a calculation of the chamber volume and area. Typical expressions of fluxes are g GHG $ha^{-1}d^{-1}$. A minimum of three measurements should be made to check the linearity of concentration increase in the chamber. A non-linear increase could indicate an inadequate sealing of the chamber or an important increase of pressure in the chamber due to temperature increases.[2] However, venting of the closed chambers can create large errors. The chamber cover should be removed once the final sample has been taken to minimize disturbance of the environmental conditions of the area covered by the chamber.

Open chambers can be used as well (Figure 2). In open chambers outside air flows into the chamber via an inlet and is forced to flow over the enclosed soil surface before leaving the chamber via an outlet. The concentration of the respective gas is measured at the in- and outlet sides. The gas flux from the soil surface can be calculated from the concentration difference between the in- and outlet, gas flow rate, and volume and area covered by the chamber.

Manual or Automated Sampling and Analysis

Using closed chambers, the headspace can be sampled by syringe. The gas samples are transferred to the laboratory into sealed containers (evacuated vials fitted with rubber septa) for analysis. To obtain a representative sample of the headspace, a diffusive mixing over 30 min is adequate, unless a mixing fan is used. Gas samples should be taken from the headspace immediately after sealing and at equal time intervals, thereafter, during which, gas concentrations increase linearly.

With automatic sampling, manual gas sampling from the chamber is replaced by a gas flow system providing a periodic sample transfer to a detector. The basic elements of an automatic system consist of closed chambers (equipped with lids that open and close automatically) or open chambers, gas flow

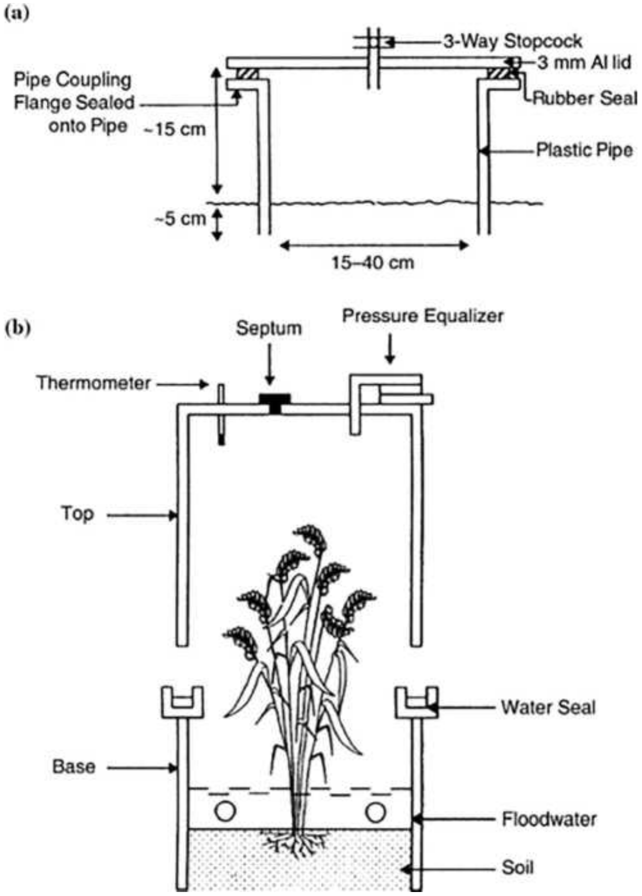


FIGURE 1 (a) Typical closed chamber used for greenhouse gas flux measurements from aerobic soils. **Source:** Adapted from Manual on Measurement of Methane and Nitrous Oxide Emissions from Agriculture.[2] (b) Schematic drawing of closed chamber used for greenhouse gas flux measurements in flooded soil.
Source: Lindau.[3]

FIGURE 2 Field set-up of an open chamber with removable lid to measure NO fluxes from forest soils: In the front, the outlet with the air sampling tube; in the back, the air inlet with the air sampling tube and the NO analyzer with data acquisition system.

systems (tubing and pump); a sampling unit; an analytical unit (detector); a time controller; and a data acquisition system. Automatic sampling is more expensive and it needs to be done in the vicinity of the analysis device. However, this technique is helpful when extensive data sets need to be collected over longer periods of time. It increases the reliability of the emission data obtained, because the number of manipulations is reduced. In most cases open chambers are sampled automatically.

Analytical Aspects

Gas Chromatography

Gas samples are injected into the gas chromatograph (GC) either manually or through the use of a sample loop (automated). Depending on the type of gas to be analyzed, specific GC settings and columns are used. Carbon dioxide is detected using a thermal conductivity detector (TCD). Methane is detected using a flame ionization detector (FID). Nitrous oxide is detected using a ^{63}Ni electron capture detector (ECD). For analysis of N_2O, care must be taken to remove both CO_2 and water vapor. In experiments with acetylene, to block the last step in the denitrification process, one may need to bypass the ECD with the acetylene in the sample after it exits from the chromatographic column. Acetylene alters both the sensitivity and stability of some ECDs.[2]

Photo-Acoustic-Infrared Detector (PAID)

The measuring principle is based on photo-acoustic detection of the absorption of infrared light. This means that any gas that absorbs infrared light of a specific wavelength can be measured. The most common PAID is the *Brüel and Kjær multi-gas monitor.* The PAID is equipped with a pump, which draws air from the flux chamber into an analysis cell inside the gas monitor. This cell can be sealed hermetically. Light from an infrared source is pulsated by a mechanical chopper and then passes through one of the optical filters of the filter carousel. The filter produces an infrared wavelength, which is selectively absorbed by the gas being monitored. Absorption of infrared light by the gas in the closed analysis cell causes the temperature to increase. Because the infrared light is pulsating, the gas temperature increases and decreases. In the closed cell this results in an increase and decrease of the pressure, which can be measured via 2 microphones. This acoustic signal is proportional to the concentration of the gas monitored. The *Brüel and Kjær multi-gas monitor* allows the analysis of 5 different gases and water vapor from 1 sample in approximately 120 sec. The device has the ability to compensate for temperature fluctuations and water vapor interference. Using an automated sampling system, careful calibration of the multi-gas monitor and the sampling unit is required.[4]

Micrometeorological Methods

The basic concept of micrometeorological methods for measuring trace gas fluxes to or from the soil surface is that gas transport is accomplished by the eddying motion of the atmosphere which displaces parcels of air from one level to another. Transport of a gas through the free atmosphere is provided by turbulent diffusion and convection in which the displacement of individual eddies is the basic transport process. Micrometeorological methods are based on the assumption that the flux to or from the soil surface is identical to the vertical flux measured at the reference level some distance above the surface. Therefore, a flat and homogeneous terrain is needed. The flux measured at the reference level provides than the average flux over the upwind area (fetch), provided that sampling point at the reference level is in the height range in which the vertical flux is constant with height (fetch$\approx$100H, where H$=$height of the sampling point). In the simplest of the micrometeorological methods, the flux may be measured by sensing the concentrations and velocities of components of the turbulence. [5,6]

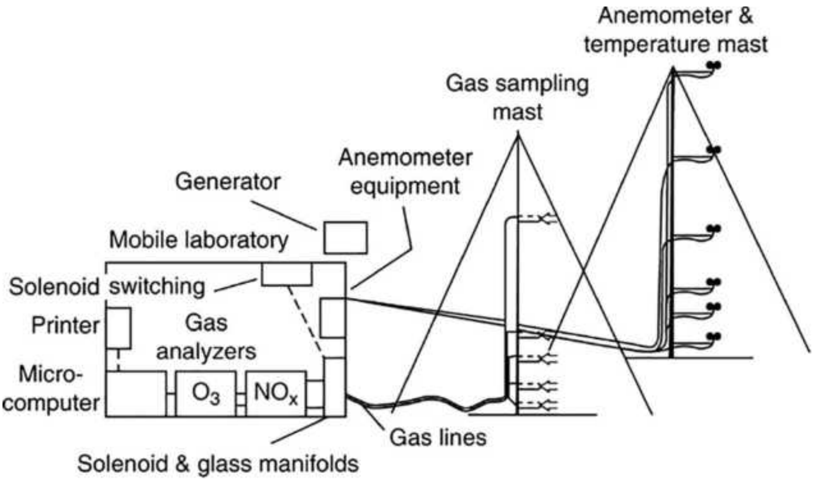


FIGURE 3 Typical instrument for flux gradient measurements of greenhouse gas fluxes in the field.
Source: Adapted from Denmead.[5]

Two general micrometeorological techniques are used to measure trace gas fluxes: the eddy correlation and flux- gradient technique (Figure 3). Application of both approaches is limited to situations in which the air analyzed has passed over a homogeneous exchange surface for a long distance so that profiles of gas concentration in the air are in equilibrium with the local rates of exchange. These methods also require that horizontal concentration gradients are negligible. Eddy correlation methods require a fast response detector. The tuneable diode laser (TDL) technique is based on infrared absorption spectrometry, whereby the absorption depends upon path length, line strength, and absorber concentration. Liquid nitrogen temperature diodes are commercially available to cover the infrared spectrum from about 2–10 μm, the region where most trace gases have absorption spectra.[5,6] Advantages and disadvantages of micrometeorological methods are listed in Table 1.

Nonisotopic Tracer Methods

Tracer methods involve the release of an inert tracer gas, most commonly sulphur hexafluoride (SF_6), from an emitting surface. The tracer gas is released at a known rate in a pattern similar to the release pattern of the GHG, perpendicular to the direction of the prevailing wind. This method can be applied when a definite plume of the GHG can be readily detected in the ambient environment. Under these conditions the plume of the dispersed emission is located based on analyses of upwind and downwind air samples. The flux rate is computed, using the ratio of the plume concentration of the tracer and the GHG and the known release rate of the tracer. The advantage of this technique is that aggregate gas emissions can be collected from heterogeneous areas, such as landfills, circumventing the problem of spatial heterogeneity. However, the high costs, dependence on meteorological conditions, and the potential for interfering sources limit its application.[2]

Ultra-Large Chambers with Long-Path Infrared Spectrometers

Infrared absorption spectrometers are available that can give an average value for the gas concentration over distances of tens or hundreds of meters. They are useful for measurements of average emissions from a whole experimental plot, by covering the plot temporarily with a large canopy to act as a chamber and retain the gas emitted from the soil. Two systems are available: 1) the Fourier Transform Infrared (FTIR) spectrometer with a mirror system, which allows multiple reflections and thus a total path of up to 1km that is capable of measuring GHG concentration changes down to a fraction of 1 ppb; and 2) a simpler, less-sensitive IR spectrometer with the capacity to detect a concentration change of about 25 ppbv of N_2O and 10 ppbv of CH_4.[2]

References

1. Hutchinson, G.L.; Livingstone, G.P. Use of chamber techniques to measure trace gas fluxes. In *Agricultural Ecosystems on Trace Gases and Global Change*; ASA Special Publication 55; ASA: Madison, 1993; 63–78.
2. Manual on Measurement of Methane and Nitrous Oxide Emissions from Agriculture; IAEA-TECDOC-674; IAEA: Vienna, 1992; 91 pp.
3. Lindau, C.W.; Bollich, P.K.; De Laune, R.D.; Patrick, W.H., Jr.; Law, V.J. Effect of urea fertilizer and environmental factors on CH_4 emissions from a Louisiana USA rice field. Plant and Soil **1991**, *136,* 195–203.
4. De Visscher, A.; Goossens, A.; Van Cleemput, A. Calibration of a multipoint sampling system connected with a photoacoustic detector. Intern. J. Environ. Anal. Chem. **2000**, *76,* 115–133.
5. Denmead, O.T. Micrometeorological methods for measuring gaseous losses of nitrogen in the field. In *Gaseous Loss of Nitrogen from Plant-Soil Systems;* Freney, J.R., Simpson, J.R., Eds.; Nyhoff Junk Publ.: The Hague, 1983; 133–157.
6. Fowler, D.; Duyzer, J. Micrometeorological techniques for the measurement of trace gas exchange. In *Exchange of Trace Gases Between Terrestrial Ecosystems and the Atmosphere;* Andreae, M.O., Schimel, D.S., Eds.; John Wiley and Sons: Chichester, 1989; 189–207.

Index

Page numbers followed by f and t indicate figures and tables, respectively

SECOND EDITION

ENVIRONMENTAL MANAGEMENT HANDBOOK

VOLUME VI

Managing Human and Social Systems

edited by

Brian D. Fath

Sven E. Jørgensen

CRC Press
Taylor & Francis Group

Managing Human and Social Systems

Environmental Management Handbook, Second Edition

Edited by
Brian D. Fath and Sven E. Jørgensen

Volume 1
Managing Global Resources and Universal Processes

Volume 2
Managing Biological and Ecological Systems

Volume 3
Managing Soils and Terrestrial Systems

Volume 4
Managing Water Resources and Hydrological Systems

Volume 5
Managing Air Quality and Energy Systems

Volume 6
Managing Human and Social Systems

Managing Human and Social Systems

Second Edition

Edited by
Brian D. Fath and Sven E. Jørgensen

Assistant to Editor
Megan Cole

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

Cover photo: Znojmo, Czech Republic, B. Fath

Second edition published 2021
by CRC Press
6000 Broken Sound Parkway NW, Suite 300, Boca Raton, FL 33487-2742

and by CRC Press
2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

First edition published by CRC Press 2013

CRC Press is an imprint of Taylor & Francis Group, LLC

ISBN: 978-1-138-34268-2 (hbk)
ISBN: 978-1-003-05351-4 (ebk)

Typeset in Minion
by codeMantra

Contents

SECTION I APC: Anthropogenic Chemicals and Activities

SECTION II COV: Comparative Overviews of Important Topics for Environmental Management

SECTION III CSS: Case Studies of Environmental Management

SECTION IV DIA: Diagnostic Tools: Monitoring, Ecological Modeling, Ecological Indicators, and Ecological Services

SECTION V ELE: Focuses on the Use of Legislation or Policy to Address Environmental Problems

SECTION VI ENT: Environmental Management Using Environmental Technologies

SECTION VII PRO: Basic Environmental Processes

Preface

Given the current state of the world as compiled in the massive Millennium Ecosystem Assessment Report, humans have changed ecosystems more rapidly and extensively during the past 50 years than in any other time in human history. These are unprecedented changes that need certain action. As a result, it is imperative that we have a good scientific understanding of how these systems function and good strategies on how to manage them.

In a very practical way, this multivolume *Environmental Management Handbook* provides a comprehensive reference to demonstrate the key processes and provisions for enhancing environmental management. The experience, evidence, methods, and models relevant for studying environmental management are presented here in six stand-alone thematic volumes, as follows:

VOLUME 1 – Managing Global Resources and Universal Processes
VOLUME 2 – Managing Biological and Ecological Systems
VOLUME 3 – Managing Soils and Terrestrial Systems
VOLUME 4 – Managing Water Resources and Hydrological Systems
VOLUME 5 – Managing Air Quality and Energy Systems
VOLUME 6 – Managing Human and Social Systems

In this manner, the handbook introduces in the first volume the general concepts and processes used in environmental management. The next four volumes deal with each of the four spheres of nature (biosphere, geosphere, hydrosphere, and atmosphere). The last volume ties the material together in its application to human and social systems. These are very important chapters for a wide spectrum of students and professionals to understand and implement environmental management. In particular, the features include the following:

- The first handbook that demonstrates the key processes and provisions for enhancing environmental management.
- Addresses new and cutting-edge topics on ecosystem services, resilience, sustainability, food–energy–water nexus, socio-ecological systems, etc.
- Provides an excellent basic knowledge on environmental systems, explains how these systems function, and gives strategies on how to manage them.
- Written by an outstanding group of environmental experts.

Since the handbook covers such a wide range of materials from basic processes, to tools, technologies, case studies, and legislative actions, each handbook entry is further classified into the following categories:

APC: Anthropogenic chemicals: The chapters cover human-manufactured chemicals and activities
COV: Indicates that the chapters give comparative overviews of important topics for environmental management

CSS: The chapters give a case study of a particular environmental management example
DIA: Means that the chapters are about diagnostic tools: monitoring, ecological modeling, ecological indicators, and ecological services
ELE: Focuses on the use of legislation or policy to address environmental problems
ENT: Addresses environmental management using environmental technologies
NEC: Natural elements and chemicals: The chapters cover basic elements and chemicals found in nature
PRO: The chapters cover basic environmental processes.

Volume 6, *Managing Human and Social Systems*, applies the cumulative knowledge of environmental science and systems specifically into managing human and social systems. There are over 50 entries covering a wide area from environmental legislation and policy to human health, economics, sustainable development, and green technologies. New entries are included to cover environmental accounting, limits to growth, and urban agriculture. Case studies investigate the impact of cell tower placement, health consequences of pesticides in developing countries, and the promise of community-based monitoring. This culminating volume gives guidance for effective environmental management and a glimpse into future challenges and opportunities.

Brian D. Fath
Brno, Czech Republic
December 2019

Editors

Brian D. Fath is Professor in the Department of Biological Sciences at Towson University (Maryland, USA) and Senior Research Scholar at the International Institute for Applied Systems Analysis (Laxenburg, Austria). He has published over 180 research papers, reports, and book chapters on environmental systems modeling, specifically in the areas of network analysis, urban metabolism, and sustainability. He has co-authored the books *A New Ecology: Systems Perspective* (2020), *Foundations for Sustainability: A Coherent Framework of Life–Environment Relations* (2019), and *Flourishing within Limits to Growth: Following Nature's Way* (2015). He is also Editor-in-Chief for the journal Ecological Modelling and Co-Editor-in-Chief for *Current Research in Environmental Sustainability*. He was the 2016 recipient of the Prigogine Medal for outstanding work in systems ecology and twice a Fulbright Distinguished Chair (Parthenope University, Naples, Italy in 2012 and Masaryk University, Czech Republic in 2019). In addition, he has served as Secretary General of the International Society for Ecological Modelling, Co-Chair of the Ecosystem Dynamics Focus Research Group in the Community Surface Modeling Dynamics System, and member and past Chair of Baltimore County Commission on Environmental Quality.

Sven E. Jørgensen (1934–2016) was Professor of environmental chemistry at Copenhagen University. He received a doctorate of engineering in environmental technology and a doctorate of science in ecological modeling. He was an honorable doctor of science at Coimbra University (Portugal) and Dar es Salaam (Tanzania). He was Editor-in-Chief of *Ecological Modelling* from the journal inception in 1975 until 2009. He was Editor-in-Chief for the *Encyclopedia of Environmental Management* (2013) and *Encyclopedia of Ecology* (2008). In 2004, he was awarded the Stockholm Water Prize and the Prigogine Medal. He was awarded the Einstein Professorship by the Chinese Academy of Sciences in 2005. In 2007, he received the Pascal Medal and was elected a member of the European Academy of Sciences. He has published over 350 papers and has edited or written over 70 books. He gave popular and well-received lectures and courses in ecological modeling, ecosystem theory, and ecological engineering worldwide.

Contributors

Diana Aga
Department of Chemistry
State University of New York at Buffalo
Buffalo, New York

Jirapat Ananpattarachai
Center of Excellence for Environmental Research and Innovation
Faculty of Engineering
Naresuan University
Phitsanulok, Thailand

Massimo Antoninetti
Institute for Electromagnetic Sensing of the Environment (IREA)
National Research Council of Italy (CNR)
Milan, Italy

Seungyun Baik
Department of Chemistry
State University of New York at Buffalo
Buffalo, New York

Simone Bastianoni
Ecodynamics Group
Department of Chemistry
University of Siena
Siena, Italy

Anders Baun
Department of Environmental Engineering
Technical University of Denmark
Kongens Lyngby, Denmark

Richard W. Bell
School of Environmental Science
Murdoch University
Perth, Western Australia, Australia

Sanford V. Berg
Director of Water Studies
Public Utility Research Center
University of Florida
Gainesville, Florida

W.E.H. Blum
Institute of Soil Research
University of Natural Resources and Life Sciences
Vienna, Austria

Ben Boer
School of Law
University of Sydney
Sydney, Australia

J.D. Booker
Environmental Compliance Department
Consolidated Nuclear Security LLC
Amarillo, Texas

Elvira Buonocore
Laboratory of Ecodynamics and Sustainable Development
Department of Science and Technology
Parthenope University of Naples
Napoli, Italy
and
CoNISMa
Piazzale Flaminio, Rome, Italy

Ni-Bin Chang
Department of Civil, Environmental, and Construction Engineering
University of Central Florida
Orlando, Florida

Guangnan Chen
Faculty of Engineering and Surveying
University of Southern Queensland
Toowoomba, Australia

Angelique Chettiparamb
School of Real Estate and Planning
University of Reading
Reading, United Kingdom

Richard Cowell
School of Geography and Planning
Cardiff University
Cardiff, United Kingdom

Gemma Cranston
Global Footprint Network
Geneva, Switzerland

Christina D. DiFonzo
Department of Entomology
Michigan State University
East Lansing, Michigan

Barbara Dinham
Eurolink Center
Pesticide Action Network UK
London, United Kingdom

George Ekström
Swedish National Chemicals Inspectorate (KEMI)
Solna, Sweden

Brian D. Fath
Department of Biological Sciences
Towson University
Towson, Maryland
and
Advanced Systems Analysis Program
International Institute for Applied System Analysis
Laxenburg, Austria

Natalia Fath
Department of Geography and Environmental Planning
Towson University
Towson, Maryland

Maria R. Finckh
Department of Ecological Plant Protection
University of Kassel
Witzenhausen, Germany

Pier Paolo Franzese
Laboratory of Ecodynamics and Sustainable Development
Department of Science and Technology
Parthenope University of Naples
Napoli, Italy
and
CoNISMa
Piazzale Flaminio, Rome, Italy

W. Friesl-Hanl
Environmental Analysis
Environment Agency Austria
and
Institute of Soil Research
University of Natural Resources and Life Sciences
Vienna, Austria

Alessandro Galli
Global Footprint Network
Geneva, Switzerland

M.H. Gerzabek
Institute of Soil Research
University of Natural Resources and Life Sciences
Vienna, Austria

Timothy S. Goebel
Agricultural Research Service (USDA-ARS)
U.S. Department of Agriculture
Lubbock, Texas

Khara D. Grieger
Department of Environmental Engineering
Technical University of Denmark
Kongens Lyngby, Denmark

Lisa Guan
School of Chemistry, Physics and Mechanical Engineering
Science and Engineering Faculty
Queensland University of Technology
Brisbane, Australia

Denis Hamilton
Queensland Department of Primary Industries
Brisbane, Australia

Ian Hannam
Center for Natural Resources
Department of Infrastructure
Planning and Natural Resources
Sydney, Australia

Steffen Foss Hansen
Department of Environmental Engineering
Technical University of Denmark
Kongens Lyngby, Denmark

Kelsey Hart
College of Veterinary Medicine
University of Georgia
Athens, Georgia

James G. Hewlett
Energy Information Administration
U.S. Department of Energy
Washington, District of Columbia

Rusty T. Hodapp
Energy and Transportation Management
Dallas/Fort Worth International Airport Board
Dallas/Fort Worth Airport, Texas

Tim Jackson
University of Surrey
United Kingdom

Ike Jeon
Department of Animal Science and Industry
Kansas State University
Manhattan, Kansas

Sven E. Jørgensen
Institute A, Section of Environmental Chemistry
Copenhagen University
Copenhagen, Denmark

Puangrat Kajitvichyanukul
Center of Excellence for Environmental Research and Innovation
Faculty of Engineering
Naresuan University
Phitsanulok, Thailand

Robert J. Lascano
Agricultural Research Service (USDA-ARS)
U.S. Department of Agriculture
Lubbock, Texas

Leslie London
Occupational and Environmental Health Research Unit
University of Cape Town
Observatory, South Africa

Graça Martinho
Department of Environmental Sciences and Engineering
Faculty of Sciences and Technology
New University of Lisbon
Caparica, Portugal

Thomas J. Mbise
Tanzania Association of Public Occupational and Environmental Health Experts
Dar-es-Salaam, Tanzania

Alexandra Navrotsky
Department of Chemical Engineering and Materials Science
University of California, Davis
Davis, California, U.S.A.

Elena Neri
Ecodynamics Group
Department of Chemistry
University of Siena
Siena, Italy

Aiwerasia V.F. Ngowi
Tanzania Association of Public Occupational and Environmental Health Experts
and
Department of Environmental and Occupational Health
Muhimbili University of Health and Allied Sciences (MUHAS)
Dar-es-Salaam, Tanzania

Valentina Niccolucci
Ecodynamics Group
Department of Chemistry
University of Siena
Siena, Italy

Egide Nizeyimana
Department of Agronomy
and
Environmental Resources Research Institute
Pennsylvania State University
University Park, Pennsylvania

Jacob Opadeyi
Department of Surveying and Land Information
Faculty of Engineering
University of the West Indies
St. Augustine, Trinidad and Tobago

Margareta Palmborg
Swedish Poisons Information Center
Stockholm, Sweden

Mark A. Peterson
Sustainable Success LLC
Clementon, New Jersey

David Pimentel
Department of Entomology
Cornell University
Ithaca, New York

Ana Pires
Department of Environmental Sciences and Engineering
Faculty of Sciences and Technology
New University of Lisbon
Caparica, Portugal

Wendell A. Porter
Department of Agricultural and Biological Engineering
University of Florida
Gainesville, Florida

Larama M.B. Rongo
Muhimbili University of Health and Allied Sciences
Dar-es-Salaam, Tanzania

Stephen A. Roosa
Energy Systems Group, Inc.
Louisville, Kentucky

Giovanni F. Russo
Laboratory of Ecodynamics and Sustainable Development
Department of Science and Technology
Parthenope University of Naples
Napoli, Italy
and
CoNISMa
Piazzale Flaminio, Rome, Italy

Alka Sapat
School of Public Administration
Florida Atlantic University
Boca Raton, Florida

Jan Henrik Schmidt
Department of Ecological Plant Protection
University of Kassel
Witzenhausen, Germany

Joshua Steinfeld
Old Dominion University
Norfolk, Virginia

Evamarie Straube
University Professor Emerita
Rostock, Germany

Sebastian Straube
Professor and Division Director
Division of Preventive Medicine
Department of Medicine
University of Alberta
Alberta, Canada

Praful Suchak
Sneha Plastics Pvt. Ltd.
Suchak's Consultancy Services
Mumbai, India

Abhishek Tiwary
School of Engineering and Sustainable Development
De Montfort University
United Kingdom

Peter A. Victor
Professor Emeritus
York University
Toronto, Canada

Nemi Vora
Biological Systems and Engineering Division
Lawrence Berkeley National Lab
Berkeley, California

Mathis Wackernagel
Global Footprint Network
Oakland, California

Apichon Watcharenwong
School of Environmental Engineering
Suranaree University of Technology
Nakhon Ratchasima, Thailand

W.W. Wenzel
Institute of Soil Research
University of Natural Resources and Life Sciences
Vienna, Austria

Catharina Wesseling
Central American Institute for Studies on Toxic Substances (IRET)
National University
Heredia, Costa Rica

Wanpen Wirojanagud
Department of Environmental Engineering
Faculty of Engineering
KhonKaen University
Khon Kaen, Thailand
and
Center of Excellence on Hazardous Substance Management
National Centers of Excellence (PERDO)
Bangkok, Thailand

I

APC: Anthropogenic Chemicals and Activities

1

Food: Pesticide Contamination

Denis Hamilton

Introduction

How much pesticide residue did I eat today?

No more than necessary, and less than would be detrimental to your health.

Government authorities must be able to support the answer to the consumer's question with scientific data and valid scientific studies.[1] The "no more residues than necessary" concept originates from the principle of good agricultural practice, which implies that the desired effect (pest control) will be achieved without leaving more residues than necessary in the food.

Before Registration

Pesticide residue evaluation and risk assessment prior to registration are summarized in Figure 1.[2,3]

Risks to the environment and to the user are also evaluated but are not considered further under the present topic—*food contamination with pesticide residues.*

Metabolism studies on a pesticide in crops and farm animals identify the nature of the residue. The residue may consist of a parent compound or metabolites or a mixture. In some cases, different pesticides produce the same metabolites; in other cases, the metabolite of one pesticide is another pesticide. Some crops genetically modified for herbicide resistance achieve their resistance by metabolizing the herbicide to a derivative with no herbicidal activity.

The acceptable daily intake (ADI) of a chemical is the daily intake, expressed on a body-weight basis, which, during an entire lifetime, appears to be without appreciable risk to the health of the consumer on the basis of all the known facts at the time. The ADI is based on animal feeding studies that find the daily dose over a lifetime resulting in no observable adverse effect on the most sensitive animal species tested. Then, a margin of safety (safety factor, commonly 100) is applied to allow for extrapolation from animals to humans and the variability in responses between average and highly sensitive humans.

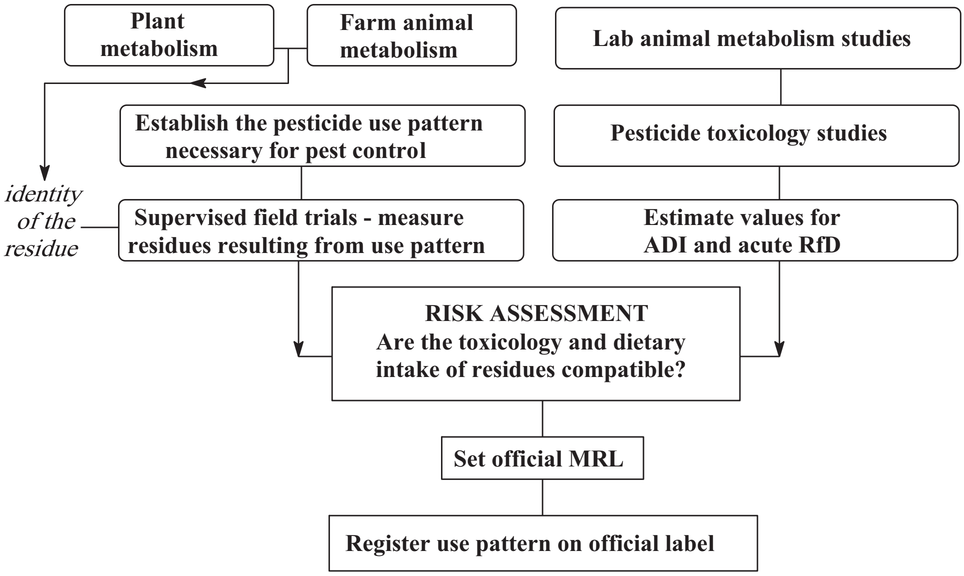


FIGURE 1 Risk assessment process before registration for pesticide residues in food. ADI: acceptable daily intake; Acute RfD: acute reference dose; MRL: maximum residue limit or tolerance.
Source: Hamilton DJ, Food contamination with pesticide residues, in *Encyclopedia of Pest Management*, 2002, p 287.

The acute reference dose (acute RfD or ARfD) of a chemical is an estimate of the amount normally expressed on a body-weight basis, which can be ingested in a period of 24 hours or less without appreciable health risk to the consumer on the basis of all known facts at the time of the evaluation. The acute RfD is also based on the results of animal dosing studies with a suitable safety factor.

The maximum residue limit (MRL), synonymous with "tolerance," is the maximum concentration of pesticide residue legally permitted in or on food commodities. The MRL usually applies to the commodity of trade, which may or may not be the same as the edible portion. For a fruit such as apples, it is the same, while for bananas, the MRL applies to the whole banana, but only the pulp is eaten. An MRL provides a division between food that is legally acceptable or unacceptable. Foods derived from commodities complying with the relevant MRLs are intended to be toxicologically acceptable, but the MRL is not a dividing line between safe and unsafe.

Supervised residue trials on animal feed commodities and livestock feeding studies with pesticide residues generate the information required to support MRLs for meat, milk, and eggs.

Risk assessment tells us whether or not the amounts of residue are likely to be safe for consumers.[4,5] We estimate dietary intake (also referred to as "dietary exposure") of pesticide residues by multiplying the level of residue in the food ready for consumption by the amount of the food consumed. For chronic risk assessment, we compare the sum for all foods of expected long-term average intake with the ADI for the pesticide. For acute risk assessment, we compare possible intake from high consumption of a food, in a period of 24 hours or less, with the acute RfD.

After Registration

The design of monitoring studies for residues in food commodities depends on the purpose: random survey of food consignments (surveillance), targeted enforcement sampling where a residue problem is suspected, export monitoring to meet trade requirements, and total diet studies.

Government authorities regularly survey agricultural and animal products for levels of pesticide residues. If the label directions were based on reliable and representative field trials and if users are faithfully following label directions, then residues will be within the legal MRLs. Most surveys have demonstrated a high level of compliance.

Total diet studies identify which pesticides and measure in what quantities people are actually consuming. Food purchased in the marketplace is prepared by peeling and cooking as in the normal household and is then subjected to residue analysis. Amounts of foods consumed are known from specially designed food surveys for subpopulations such as adult males and females, children, toddlers, and infants, as well as for ethnic groups and regions or localities. Dietary intakes for populations and subpopulations are calculated from the diets and the residue levels found by analysis. Commonly, total diet studies demonstrate intakes much less than the ADI.

Food Processing

Food processing usually reduces pesticide residue levels because of the washing or cleaning, peeling, milling, juicing, cooking, or baking. Residue levels may increase in some processed commodities because the residue tends more to one fraction than another.[6] For example, residues on the surface of a wheat grain will find their way into the bran fraction with little in the flour. Residues of oil-soluble pesticides will find their way mainly into the vegetable oil fraction from an oilseed such as soybean.

In particular cases, a food process can change the nature of the residue. For example, ethylenebisdithiocarbamate fungicides are converted, on cooking, to ethylenethiourea, which is more toxic than the parent pesticide. Fortunately, ethylenebisdithiocarbamates are essentially surface residues, and their levels can substantially be reduced by thorough washing before a cooking or blanching step.

Trade Issues

MRL values derived from good agricultural practice are, by their nature, local. A pesticide is used in the best way within local cultural practices to control a specific pest, and the rate of pesticide disappearance depends on local environmental conditions. Comparisons among countries of national MRLs and tolerances will frequently reveal substantial differences. Table 1 shows the range of MRLs for ethephon in 17 countries for each of four commodities.

TABLE 1 National Ethephon MRLs and Tolerances (mg/kg) in 1999[7]

	Peppers	Tomatoes	Pineapples	Grapes
Argentina	2	2		
Australia		2	2	10
Brazil		1.5	0.5	
Canada		2		
France				0.05
India		2	2	
Ireland	3	3		
Italy		3		0.05 Wine grapes 3 Table grapes
Korea		3	1	2
Netherlands	3	3		
New Zealand		1		
Poland		3		
Portugal	3	3		
South Africa			1	5
Taiwan		2	2	2
United Kingdom	3	3		
United States	30	2	2	2

The differences pose problems for international trade in food commodities. The importing country may reject shipments of food that do not comply with its national MRLs. It is attractive for some lobby groups and some governments to use national differences in MRLs as a barrier to trade.

Where no MRL or tolerance has been set for a pesticide on a food, some national governments apply a "zero tolerance," that is, the MRL is assumed to be zero unless otherwise stated. The reason no MRL is set could simply be that the pest problem does not occur or that the crop is not produced locally; for example, cold temperate countries do not produce pineapples, so there will be no local uses or local MRLs.

The Codex Alimentarius Commission was established in 1961 to implement the FAO/WHO Food Standards Program. A purpose of the program is to protect the health of consumers and to ensure fair practices in the food trade. The Codex Committee on Pesticide Residues (CCPR) has the responsibility to establish Codex MRLs for food commodities in international trade.

CCPR relies on the data supplied by member governments and has established many MRLs. The methods of data evaluation in Codex are very similar to the methods in countries with regulatory control of pesticide use; Codex draws on the expertise of scientists from such countries around the world. Member government acceptance of Codex MRLs for food commodities in international trade is reducing the incidence of trade barriers based on national MRLs.

Developing countries have sometimes suffered pesticide residue trade difficulties because a lack of resources has made it difficult for them to monitor their exports effectively to ensure compliance with the importing country MRL requirements.

Analytical Methods for Pesticide Residues

Analytical methods for pesticide residues in food typically rely on gas-liquid chromatography (GLC) or high-performance liquid chromatography (HPLC) in the final measurement step following extraction from the sample and a sequence of clean-up steps.[2,8] Multiresidue methods include many residues in one procedure for the sake of economy. Monitoring usually requires the detection and quantitative measurement of residue levels down to concentrations of around 0.01–0.05 mg/kg. Laboratories must validate their procedures down to the required level, that is, prove that the procedures can identify and measure with a specified precision residues down to a required "limit of quantification" (LOQ).

The LOQ is important in the interpretation of monitoring data. An analytical result reported as "less than LOQ" or sometimes as "no detectable residue" could possibly mean no residue or a residue at a level too low for the method.

Not all pesticide residues are amenable to inclusion in multiresidue methods; they may need separate analysis, which becomes expensive. Reports of monitoring data should state explicitly which residues would have been detected if present above stated LOQs.

Reliable high-quality data are essential for correct interpretation during registration, investigation, and control of residues in food. Highly skilled analysts using good laboratory practices, standard procedures, and other measures are generating valid data to support those requirements.

Future

The science of risk assessment will be further developed. Food safety and food security will continue to be important for consumer, government, and industry.[9] Trade issues will continue to be problematic with specific incidents of residues in foods arising from time to time. National governments will develop strategic approaches to deal with trade issues related to pesticide residues. Knowledgeable people in government and industry and experienced workers in functioning laboratories will be needed to support those strategic approaches. Exporters will need to monitor residues in a high percentage of their exports to meet the requirements of their customers. We might expect more developments with biopesticides. Relevant impurities in biopesticides are more likely to be biological than chemical, posing new challenges for analytical and test methods.

References

1. Frehse, H. ed. 1991. Pesticide Chemistry, Advances in International Research, Development, and Legislation, *Proceedings of the Seventh International Congress of Pesticide Chemistry (IUPAC), Hamburg, 1990*. VCH Verlagsgesellschaft mbH: Weinheim, Germany; 361–601.
2. FAO. 2016. Submission and Evaluation of Pesticide Residues Data for the Estimation of Maximum Residue Levels in Food and Feed; third edition. *FAO Plant Production and Protection Paper*, 225:1–286.
3. Dishburger, H.J., Ballantine, L.G., McCarthy, J., Murphy, J. and Tweedy, B. G., eds. 1991. *Pesticide Residues and Food Safety: A Harvest of Viewpoints. ACS Symposium Series 446*, American Chemical Society: Washington, DC, 1–348.
4. Hamilton, D.J., Holland, P.T., Ohlin, B. et al. 1997. Optimum use of available residue data in the estimation of dietary intake of pesticides. *Pure Appl. Chem*. 69:1373–1410.
5. WHO. 2009. Principles and Methods for the Risk Assessment of Chemicals in Food, IPCS. *Chapter 6, Dietary Exposure Assessment of Chemicals in Food*. Environmental Health Criteria 240.
6. Holland, P.T., Hamilton, D., Ohlin, B. and Skidmore, M.W. 1994. Effects of storage and processing on pesticide residues in plant products. *Pure Appl. Chem*. 66:335–356.
7. FAO. 2000. Pesticide Residues in Food. Ethephon. Evaluations 1999. FAO Plant Production and Protection Paper, 157:210.
8. Ambrus, Á. 1999. Quality of Residue Data. In *Pesticide Chemistry and Bioscience: The Food-Environment Challenge*, Brooks, G.T. and Roberts, T.R., eds., 339–360. The Royal Society of Chemistry: Cambridge.
9. Ambrus, Á. and Hamilton, D. 2017. Chapter 12, Future Directions. In *Food Safety Assessment of Pesticide Residues*, Ambrus, Á. and Hamilton, D., eds., 507–510. World Scientific Press: Singapore.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

2

Human Health: Consumer Concerns to Pesticides

George Ekström and
Margareta Palmborg

Introduction

A recent European survey of risk perception and food safety showed, in line with previous research findings, that consumers tend to worry most about risks caused by external factors over which they have little or no control. Consequently, consumers appear to be less worried about risks possibly associated with their own behavior or practices. Physicians and scientists are the most trusted information sources with regard to serious food risks, followed by public authorities and mass media. Economic operators (food manufacturers, farmers, and retailers) are cited as being among the least trusted.[1]

Interviews conducted with over 1,000 consumers in a survey done in 2001 by the British Co-op Group showed that consumers were concerned about the effects of pesticides. Consumers who took part in this survey, when prompted with a series of questions, were concerned that pesticides are harmful to wildlife, leave residues in food, pollute water courses, are harmful to growing children, are harmful to the respondents themselves, and damage the health of farm workers.[2,3] According to a personal communication with David Pimentel, in the United States, the Food and Drug Administration has reported that 97% of people prefer foods without pesticides.

Particular causes of consumer concern are the potential for 'cocktail' effects from multiple residues (see Table 1),[4–6] and the fact that children may exceed health- related acute reference doses even at legally acceptable residue levels (see Table 2).[6] In Australia, the Food Standards code contains provisions for an additional, overall limit for pesticides belonging to the same chemical group (see Table 3).[7]

TABLE 1 Multiple Residues Found in a Single Sample of Pears 2004

Pesticides Found	Residue Level (mg/kg)	Maximum Residue Limit (mg/kg)	Residue Level in % of Maximum Residue Limit
Dithiocarbamates	0.305	3	10
Chlorpropham	0.155	0.05	310
Azinphosmethyl	0.084	0.5	17
Procymidone	0.071	1	7
Dichlofluanid	0.060	5	1
Chlorpyriphos	0.059	0.5	12
Bromopropylate	0.055	0.05	110
Cyprodinil	0.022	—	—
		Combined total residues	466

Source: Swedish food residue monitoring report (see Andersson and Jansson[6]).

TABLE 2 Food Residues Potentially Leading to Short Time Intake in Excess of the Acute Reference Dose (ARfD) for Toddlers 2004

Pesticide	Food Commodity	Highest Residue Found (mg/kg)	Maximum Residue Limit (mg/kg)	ARfD (mg/kg Body weight)	Intake, % of ARfD for Toddlers
Dicrotophos	Chinese broccoli	4.14	—	0.0017	1,763
Lambda-cyhalothrin	Lettuce	0.92	1	0.0075	106
Oxamyl	Cucumber	0.42	—	0.009	135
Endosulfan	Melon	0.21	0.3	0.02	110
Monocrotophos	Zuccini	0.14	—	0.002	381
Aldicarb	Potatoes	0.035	0.5	0.003	122

Source: Swedish food residue monitoring report (see Andersson and Jansson[6]).

TABLE 3 Approaches to the Limitation of Organophosphorus Pesticide Residues in Food in Australia and by the British Co-operative Group, Respectively

Group Tolerance in Australia	Group Tolerance in Australia and Co-op Zero Tolerance	Co-op Zero Tolerance
Azamethiphos, azinphos-ethyl, azinphos-methyl, coumaphos, demeton, diazinon, dichlorvos, dimethoate, disulfoton, dithianon, ethion, famphur, fenchlorphos, fenitrothion, fenthion, formothion, maldison,[a] methamidophos, methidathion, mevinphos, naphtalophos,[b] parathion-methyl, phosmet, pirimiphos- ethyl, pirimiphos-methyl, pyrazophos, sulprophos, temephos, tetrachlorvinphos, thiometon, *S.S.S*-tributylphosphorotrithioate, trichlorfon, vamidothion	Ethoprophos, fenamiphos, omethoate, phorate, prothiofos	Cadusafos, chlorfenvinphos, demeton-*S*-methyl, phosphamidon, tebupirimfos, terbufos

Source: Maximum residue limits—Chemical groups[7] and The Co-operative Group.[11]

[a] ISO common name is malathion.

[b] WHO INN, no ISO common name available.

Pests and Pesticide Safety in Homes and Gardens

Consumers use a range of pesticides in their homes and gardens:

- *Herbicides* against weeds in vegetables, moss in turf, brush, etc.
- *Fungicides* against mold, mildew, etc.

- *Insecticides* against aphids, greenflies, ants, wasps, pests on potted plants, moths, pantry pests, cockroaches, flies, etc.
- *Rodenticides* against moles, rats, mice, voles, etc.
- *Repellents* against mosquitoes, black flies, ticks, game, and pests on dogs, cats, and horses
- *Wood preservatives* against rot on timber or furniture

For the general public, ingestion is the most common route of pesticide exposure. Accidental, single, high- level exposures can lead to acute pesticide poisoning, often in children, and may result from mistakenly swallowed pesticides stored in unlocked cabinets or in unmarked bottles or containers. With regard to long-term, low-level exposure of the general public, the main route of exposure, ingestion through food and drinking water, is followed by inhalation through air or dust. This exposure results in an unknown number of people with diverse chronic health effects.[8]

A comparison of pesticide poisoning cases in 1984, 1994, and 2004, performed by the Swedish Poisons Information Centre, showed that there was an increase in the overall number of human cases related to pesticide exposure—from 493 in 1984, and 774 in 1994, to 1,071 in 2004.

The proportion of pesticide-related inquiries, however, remained constant at 3%.

Most incidents were due to accidental exposure at home. Ingestion was the most frequent route of exposure, followed by inhalation. Data from Swedish hospitals reported to the Poisons Information Centre showed the same pattern, that is, an increase in the proportion of cases related to accidental exposure at home. Children were involved in about 60% of all cases in 1984 as well as in 1994 and 2004. With the exception of 'superwarfarins' found in some rodenticides, most pesticides involved in incidents at home were of low toxicity and present at low concentrations in the formulated products. The few severe cases are mainly intentional poisonings.[9]

Thirty percent of the total number of inquiries to the Centre in 2004 was due to children's ingestion of insecticides intended for control of ants, containing low concentrations of borax, organophosphorus compounds or pyrethroids. No symptoms were recorded. Therefore, the considerable number of inquiries may reflect anxiety about pesticide exposure among the Swedish population.

Food Residues

Pesticide Residues in Foods from Organic, Integrated and Conventional Production[6]

In the Swedish monitoring program for 2004, no residues were detectable in 57% of the samples. Residues at or below maximum residue limits (MRLs) were found in 39% of the samples. 3.5% of all samples contained residues above the MRLs. Of foods from organic production, 4%–5% (import and domestic, respectively) contained detectable residues. Foods from integrated production were free from detectable residues in 91% of domestic produce and 50% of imported product. Foods from conventional production contained no detectable residues in 83% of domestic foods and 46% of imported foods. Residues below the MRLs were found in foods from all three production categories. Residues above the MRLs were found only in imported products from conventional production. No residues were found in any of the 92 samples of foods intended for infants and young children.

Although produced without pesticides, organically produced foods sometimes contain residues. The reason for this may be unintentional mix-up of foods from different sources (organic and conventional), environmental contamination of soils and plants, or fraud. The organic foods that contained pesticides in the United States have been shown to come mostly from soils treated many years ago with DDT or arsenical compounds, according to a personal communication with David Pimentel.

In 11 food commodities (22 samples) from ten countries, residues of ten different pesticides were found at levels 10–37 times the MRL. Multiple residues were found in 492 samples of which 279 samples with two residues, 127 samples with three residues, 54 samples with four residues, 25 samples with five residues, five samples with six residues, and two samples with eight residues (see Table 1 and Figure 1).

FIGURE 1 European Union organic logo. (Available at http://europa.eu.int/comm/agriculture/qual/organic/logo/index_en.htm.)

Towards Residue-Reduced Food Crops

Government Action Plans

The British Food Standards Agency has recognized that while levels of pesticide residues typically found in food are not normally a food safety concern, consumer preference is for food that does not contain residues. Sixty-eight percent of consumers consider that reducing residue levels further than the current level is important. As a result, the Agency has developed an action plan for pesticide residue minimization with a goal of enabling consumers to make informed choices, and promoting best practice within the food industry. The overall action plan includes, among other things, development of crop specific action plans to achieve pesticide residue minimization for five priority crops: apples, pears, potatoes, tomatoes, and cereal grains.[10]

Retailer Initiatives[11,12]

Retailers may employ a range of strategies to reduce pesticide use and residues in the foods they produce and put on the market:

- Monitoring pesticide usage and residues
- Consulting with growers, including advice on integrated pest management and on alternative pest control systems
- Designing and providing decision tools, including crop-specific advisory sheets and frameworks for pesticide selection
- Prohibiting or restricting the use of certain pesticides
- Publishing monitoring results, for example, on corporate websites
- Promoting organically grown foods

NGO Initiatives—Ranking Residue Contents

In the United States, the Environmental Working Group has designed a report card to score pesticide residues in food products.[13] The report card, which is based on government agency monitoring and published monitoring data, shows scores for each analyzed commodity based on a number of residue characteristics, and a combined (or total) score in these categories:

- Percentage of samples with detectable residues
- Percentage of samples with two or more pesticides
- Average number of pesticides found on each sample
- Average total concentration of pesticides found
- Maximum number of pesticides found on a single sample
- Total number of pesticides found on a single commodity

In the Netherlands, Natuur and Milieu (a Dutch NGO), has designed and used a ranking system based on the following components, and a calculated total score[14]:

- For any residue not exceeding an MRL: 1 penalty point
- For each residue of a pesticide with neurotoxic effects: 2 penalty points
- For each residue exceeding an MRL or resulting from the use of a pesticide not authorized for use in the Netherlands (applicable also to imported foods): 4 penalty points
- For each residue exceeding an MRL and resulting from the use of a pesticide not authorized for use in the Netherlands (applicable also to imported foods): 8 penalty points

Conclusions

Polls have shown repeatedly that consumers are concerned about pesticide residues in food. Maximum residue limits are trading standards, which prescribe the maximum amount of particular pesticides legally permitted. These limits (MRLs) are generally based on the level of pesticides expected if good agricultural practice is followed.[10] Other MRLs may reflect only that a pesticide is no longer authorized for use, leading to zero tolerance, and to potential problems for food exporters overseas. Since many consumers feel that current good agricultural practice is not good enough for their own or their childrens' safety, governments, retailers and NGOs have initiated actions to reduce residues. Strategies focus on production methods (integrated, organic) as well as product quality (residue-reduced or residue-free foods).

Acknowledgments

Contributions from Stephanie Williamson, Pesticide Action Network U.K., are gratefully acknowledged.

References

1. European Food Safety Authority. Risk perception and food safety: Where do European consumers stand today? February 2006. http://www.efsa.eu.int/press_room/press_release/1340_en.html (accessed May 2006).
2. Food Standards Agency. Consumer concern over the use of pesticides to grow food, March 2004. http://www.food.gov.uk/multimedia/pdfs/pestresconsumeresearch.pdf (accessed January 2005).
3. Croft, D. Removing pesticides from the food chain, December 2002. http://www.pan-uk.org/pestnews/pn58/pn58p9.htm (accessed January 2006).
4. Committee on Toxicity of Chemicals in Food, Consumer Products and the Environment. Risk assessment of mixtures of pesticides and similar substances, September 2002. http://www.food.gov.uk/multimedia/pdfs/reportindexed.pdf (accessed January 2006).
5. Beaumont, P.; Buffin, D. A cocktail of problems, March 2002. http://www.pan-uk.org/pestnews/pn55/pn55p10.htm (accessed March 2006).
6. Andersson, A.; Jansson, A. The Swedish monitoring of pesticide residues in food of plant origin 2004. National (Swedish) Food Administration. http://www.slv.se/upload/dokument/Rapporter/Bekampningsmedel/2005Z2005_17eng_Livsmedelsverket.pdf (accessed January 2006).

7. Maximum residue limits—Chemical groups. Australian Food Standards No 1.4.2. http://www.foodstandards.gov.au/foodstandardscode/ (accessed January 2006).
8. World Health Organization. Public health impact of pesticides used in agriculture. 1990, ISBN 92-4-1561-39-4.
9. Ekstrom, G.; Hemming, H.; Palmborg, M. Swedish pesticide risk reduction 1981–1995: food residues, health hazard, and reported poisonings. *Rev. Environ. Contam. Toxicol.* **1996**, *147*, 119–147.
10. Food Safety Authority. Progress on an agency action plan to minimize pesticide residues in food, May 2004. http://www.food.gov.uk/multimedia/pdfs/fsa040502.pdf (accessed January 2006).
11. The Co-operative Group. Co-op and the responsible use of pesticides. Consumer issues website. http://www.co-op.co.uk (accessed February 2006).
12. Barker, K. Co-operative retail's pesticide reduction program, July 2005. http://www.pan-europe.info/conferences/PURE%20workshop%202005/Kevin%20Barker%20Co-op.pdf (accessed March 2006).
13. Environmental Working Group. Report card: Pesticides in produce 2005. http://www.foodnews.org/reportcard.php (accessed December 2005).
14. Muilerman, H. Dutch supermarket residue campaign. *Pestic. News* **2006**, *71*, 6–7.

3

Human Health: Endocrine Disruption

Evamarie Straube
and Sebastian
Straube

Introduction

In general, endocrine disruption in humans can result from genetic disorders,[1] diseases,[2] medical treatments, mental and physical[3] stress, and chemical exposures. Endocrine disruption through chemical agents forms the focus of the present chapter.

Substances relevant for such endocrine-disrupting chemical exposures include pesticides (organochlorines such as DDT,[4] other organohalogens such as dibromochloropropane,[5] some organophosphates, carbamates, dithiocarbamates, phthalates), polychlorinated biphenyls,[6] some solvents,[7] metals such as cadmium, lead, and manganese,[8] phytoestrogens, and isoflavonoids.[9] Furthermore, endocrine disruption can be caused by smoking and alcohol use,[3] and by certain drugs, for example, glucocorticoids, hypnotics, antihypertensives, neuroleptics, and H_2-antihistaminies.[2]

Endocrine disruption can affect various endocrine systems. For example, thyroid hormone inhibition has been reported in humans after occupational exposure to amitrol and mancozeb.[10] Insulin levels can be affected by streptozotocin, which is toxic to pancreatic beta cells.[11]

Arguably the most extensive knowledge exists for chemical exposure affecting the reproductive system, which is discussed in further detail below.

Of note, for some endocrine-disrupting chemicals, non-monotonic dose–response relationships have been described with hormonal disruption occurring at relatively low levels of exposure.[12]

Mechanisms of Endocrine Disruption

Diverse mechanisms of endocrine disruption by chemical agents have been described. Endocrine disruption can be caused by xenohormones. Xenoestrogens, such as endosulfan, toxaphene, dieldrin, DDT, bisphenol A, nonylphenols, and dibutylphthalates,[7] mimic the physiological effects of estrogens. Xenoantiestrogens have effects opposite to those of xenoestrogens. For example, dioxin exerts its inhibitory effect by enhancing the expression of enzymes that degrade the estrogen receptors.[13] Antiandrogenic effects may result from competitive antagonism at androgen receptors. This was demonstrated for vinclozolin and DDE, the stable metabolite of the DDT.[7] Sometimes a xenobiotic and its metabolite (such as DDT and DDE) can exert their effects at different targets in the organism.

Pesticide-induced enzymes such as UDP-glucuronyl transferase and monooxygenases can degrade hormones (e.g., testosterone). Furthermore, the pesticides endosulfan, mirex, and DDT can increase the elimination of androgens by stimulating cytochrome P450.[14,15] Pesticide exposure can also disrupt hormonal status by inhibiting enzymes. For example, inhibition of the aromatase system can lead to an increase in testosterone levels and a decrease in the formation of estradiol from testosterone.[16] Inhibitors of the aromatase system include prochloraz, imazalil, propiconazole, fenarimol, triadimenol, triadimefon, and dicofol.[17]

Endocrine disruption can furthermore involve hormone transport proteins: for example, polychlorinated biphenyl can induce thyroid disruption, and this may involve its sulfated metabolites which bind to the thyroid hormone transport protein transthyretin with high affinity.[18]

Moreover, recent research demonstrates that endocrine disruption can involve epigenetic alterations, which may be preserved into the third generation, as has been shown for DNA methylation changes that are associated with lead exposure.[19]

Endocrine Disruption with Occupational Exposure in Women

Pesticide exposure has been linked with reproductive difficulties and menstrual abnormalities. For example, prolonged time to pregnancy, reduced fecundability, reduced fertility as well as infertility have been described for pesticide-exposed women and for women working in agriculture and greenhouses, presumably related to pesticide exposure.[20–24] Furthermore, pesticide exposure has been associated with long cycles, missed periods, and intermenstrual bleeding.[25]

Endocrine Disruption with Occupational Exposure in Men

In men, pesticides may adversely affect sperm count and quality as well as impacting the levels of testosterone, follicle-stimulating hormone, and luteinizing hormone.[26]

We have found changes in sex hormone concentrations after low-dose occupational exposure to pesticides (mainly pyrethroids, carbamates, and organophosphates). With chronic occupational pesticide exposure, we found a higher level of testosterone in comparison to control persons. There also was a reduction in estradiol levels during and after the application season in pesticide applicators.[16] Another study[27], however, found an increase in estradiol concentration in pesticide-exposed men; it may be that variations in the nature and timing of the exposure can account for these differences.

Endocrine Disruption with *In Utero* Exposure

Historically, diethylstilbestrol (DES) was the first recognized example of a xenobiotic eliciting a hormonal effect. This now very well-described paradigm serves to illustrate the potential consequences of *in utero* exposure to endocrine-disrupting chemicals. The treatment of pregnant women with DES leads to an increase in the incidence of adenocarcinoma of the vagina in their daughters[28] and malformations of the external genitals in their sons[29] and grandsons.[30] The risk of breast cancer after age 40 was also increased in women with prenatal DES exposure, as were the risks for spontaneous abortions, premature births, and ectopic pregnancies in later pregnancies of women exposed *in utero*.[31–33] Treatment with DES has, furthermore, been reported to have an effect on sexual orientation[34] and handedness.[35]

Acknowledgment

We acknowledge the contribution of our previous co-author, Wolfgang Straube, to whose memory this chapter is dedicated.

References

1. Achermann, J.C.; Ozisik, G.; Meeks, J.J.; Jameson, J.L. Genetic causes of human reproductive disease. *J. Clin. Endocrinol. Metab.* **2002**, *87* (6), 2447–54.
2. Turner, H.E.; Wass, J.A.H. Gonadal function in men with chronic illness. *Clin. Endocrinol.* **1997**, *47* (4), 379–403.
3. Vermeulen, A.; Kaufman, J.M. The age associated decline in testicular function: Partial androgen deficiency of the ageing male (PADAM). *Menopause Rev.* 1999, *IV* (2), 23–35.
4. Welch, R.M.; Levin, W.; Conney, A.H. Estrogenic action of DDT and its analogs. *Toxicol. Appl. Pharmacol.* **1969**, *14* (2), 358–67.
5. Potashnik, G.; Porath, A. Dibromochloropropane (DCBP): A 17-year reassessment of testicular function and reproductive performance. *J. Occup. Environ. Med.* **1995**, *37* (11), 1287–92.
6. Bitman, J.; Cecil, H.J. Estrogenic activity of DDT analogs and polychlorinated biphenyls. *J. Agric. Food Chem.* **1970**, *18* (6), 1108–12.
7. Massaad, C.; Entezami, F.; Massade, L.; Benahmed, M.; Olivennes, F.; Barouki, R.; Hamamah, S. How can chemical compounds alter human fertility? *Eur. J. Obstet. Gynecol. Reprod. Biol.* **2002**, *100* (2), 127–37.
8. Gennart, J.P.; Buchet, J.P.; Roeis, H.; Ghyselen, P.; Ceulemans, E.; Lauwerys, R. Fertility of male workers exposed to cadmium, lead or manganese. *Am. J. Epidemiol.* **1992**, *135* (11), 1208–19.
9. Jacobs, M.N.; Lewis, D.F. Steroid hormone receptors and dietary ligands: A selected review. *Proc. Nutr. Soc.* **2002**, *61* (1), 105–22.
10. Cocco, P. On the rumors about the silent spring. Review of the scientific evidence linking occupational and environmental pesticide exposure to endocrine disrupting health effects. *Cad. Saude Publica* **2002**, *18* (2), 379–402.
11. Bolzan, A.D.; Bianchi, M.S. Genotoxicity of streptozotocin. *Mutat. Res.* **2002**, *512* (2–3), 121–34.
12. Schmidt, C.W. The lowdown on low-dose endocrine disruptors. *Environ. Health Perspect.* **2001**, *109* (9), 420–21.
13. Yonemoto, J. The effects of dioxin on reproduction and development. *Ind. Health* **2000**, *38* (3), 259–68.
14. Savas, U.; Griffin, K.J.; Johnson, E.F. Molecular mechanisms of cytochrome P-450 induction by xenobiotics: An expended role for nuclear hormone receptors. *Mol. Pharmacol.* **1999**, *56* (5), 851–57.
15. Waxman, D.J. P-450 gene induction by structurally diverse xenochemicals: Central role of nuclear receptors CAR, PXR, and PPAR. *Arch. Biochem. Biophys.* **1999**, *369* (1), 11–23.
16. Straube, E.; Straube, W.; Krüger, E.; Bradatsch, M.; Jacob-Meisel, M.; Rose, H.-J. Disruption of male sex hormones with regard to pesticides: Pathophysiological and regulatory aspects. *Toxicol. Lett.* **1999**, *107* (1–3), 225–31.
17. Vinggaard, A.M.; Hnida, C.; Breinholt, V.; Larsen, J.C. Screening of selected pesticides for inhibition of CYP19 aromatase activity *in vitro*. *Toxicol. In Vitro* **2000**, *14* (3), 227–34.
18. Grimm, F.A.; Lehmler, H.J.; He, X.; Robertson, L.W.; Duffel, M.W. Sulfated metabolites of polychlorinated biphenyls are high-affinity ligands for the thyroid hormone transport protein transthyretin. *Environ. Health Persp.* **2013**, *121*, 657–62.
19. Sen, A.; Heredia, N.; Senut, M-C.; Land, S.; Hollocher, K,; Lu, X.; Dereski, M.O.; Ruden, D.M. Multigenerational epigenetic inheritance in humans: DNA methylation changes associated with maternal exposure to lead can be transmitted to the grandchildren. *Sci. Rep.* **2015**, *5*, 14466–82.
20. Fuortes, L.; Clark, M.K.; Kirchner, H.L.; Smith, E.M. Association between female infertility and agricultural work history. *Am. J. Ind. Med.* **1997**, *31* (4), 445–51.
21. Smith, E.M.; Hammonds-Ehlers, M.; Clark, M.K.; Kirchner, H.L.; Fuortes, L. Occupational exposures and risk of female fertility. *J. Occup. Environ. Med.* **1997**, *39* (2), 138–47.
22. Abell, A.; Juul, S.; Bonde, J.P. Time to pregnancy among female greenhouse workers. *Scand. J. Work. Environ. Health* **2000**, *26* (2), 131–6.

23. Idrovo, A.J.; Sanin, L.H.; Cole, D.; Chavarro, J.; Caceres, H.; Narvaez, J.; Restrepo, M. Time to first pregnancy among women working in agricultural production. *Int. Arch. Occup. Environ. Health* **2005**, *78* (6), 493–500.
24. Bretveld, R.W.; Thomas, Ch.M.G.; Scheeepers, P.T.J.; Zielhuis, G.A.; Roeleveld, N. Pesticide exposure: The hormonal function of the female reproductive system disrupted? Reprod. *Biol. Endocrinol.* **2006**, *4*, 30–54.
25. Farr, S.L.; Cooper, G.S.; Cai, J.; Savitz, D.A.; Sandler, D.P. Pesticide use and menstrual cycle characteristics among premenopausal women in the agricultural health study. *Am. J. Epidemiol.* **2004**, *160* (12), 1194–204.
26. Mehrpour, O.; Karrari, P.; Zamani, N.; Tsatsakis, A.M.; Abdollahi, M. Occupational exposure to pesticides and consequences on male semen and fertility: A review. *Toxicol. Lett.* **2014**, *230* (2), 146–56.
27. Oliva, A.; Spira, A.; Multigner, L. Contribution of environmental factors to the risk of male infertility. *Hum. Reprod.* **2001**, *16* (8), 1768–76.
28. Herbst, A.L.; Ulfelder, H.; Poskanzer, D.C. Adenocarcinoma of the vagina. Association of maternal diethylstilbestrol therapy with tumor appearance in young women. *N. Engl. J. Med.* **1971**, *284* (15), 878–81.
29. Bibbo, M.; Gill, W.B.; Azizi, F.; Blough, R.; Fang, V.S.; Rosenfield, R.L.; Schuhmacher, G.F.P.; Sleeper, K.; Sonek, M.G.; Wied, F.; Wied, G.L. Follow-up study of male and female offspring of DES-exposed mothers. *Obstet. Gynecol.* **1977**, *49* (1), 1–8.
30. Klip, H.; Verloop, J.; van Gool, J.D.; Koster, M.E.; Burger, C.W.; van Leeuwen, F.E. Hypospadias in sons of women exposed to diethylstilbestrol in utero: A cohort study. *Lancet* **2002**, *359* (9312), 1102–7.
31. Kaufman, R.H.; Adam, E.; Hatch, E.E.; Noller, K.; Herbst, A.L.; Palmer, J.R.; Hoover, R.N. Continued follow-up of pregnancy outcomes in diethylstilbestrol-exposed offspring. *Obstet. Gynecol.* **2000**, *96* (4), 483–9.
32. Palmer, J.R.; Wise, L.A.; Hatch, E.E.; Troisi, R.; Titus-Emshoff, L.; Strohsnitter, W.; Kaufmann, R.; Herbst, A.L.; Noller, K.L.; Hyer, M.; Hoover, R.N. Prenatal diethylstilbestrol exposure and risk of breast cancer. *Cancer Epidemiol. Biomarkers Prev.* **2006**, *15* (8), 1509–14.
33. Gaskins, A.J.; Missmer, S.A.; Rich-Edwards, J.W.; Williams, P.L.; Souter, I.; Chavarro, J.E. Demographic, lifestyle, and reproductive risk factors for ectopic pregnancy. *Fertil. Steril.* **2018**, *110* (7), 1328–37.
34. Ehrhardt, A.; Meyer-Bahlburg, H.; Rosen, L.; Feldman, J.; Veridiano, N.; Zimmermann, I.; McEwen, B.S. Sexual orientation after prenatal exposure to exogenous estrogen. *Arch. Sex. Behav.* **1985**, *14* (1), 57–77.
35. Scheirs, J.G.M.; Vingerhoets, A.J.J.M. Handedness and other laterality indices in women prenatally exposed to DES. *J. Clin. Exp. Neuropsychol.* **1995**, *17* (5), 725–30.

4

Human Health: Pesticides

Kelsey Hart and
David Pimentel

Introduction

Since the first use of DDT for crop protection in 1945, the total amount of pesticides used in agriculture worldwide has been staggering. In 1945, about 50 million kg of pesticides were applied worldwide. Today, global usage is currently at about 2.5 billion kg/yr, an approximate 50-fold increase since 1945.[1] Unfortunately, the toxicity of most modern pesticides is more than 10-fold greater than those used in the early 1950s,[1] so the potential hazards have increased as well. In fact, studies have linked exposure to pesticides with a variety of human health problems, from asthma to cancer.[2] High levels of exposure to toxic pesticides can even result in fatal poisonings. In addition, we are discovering that we can be unknowingly exposed to pesticides and pesticide residues through the food we eat, water we drink, and air we breathe,[3] and that both short- and long-term exposure to pesticides can lead to chronic health effects. Based on the available data, estimates are that human pesticide poisonings and related illnesses in the United States cost about $933 million each year.[3]

Exposure to Pesticides

While farmers and pesticide applicators typically are exposed to higher levels and more kinds of pesticides, the general public is also exposed—often unwittingly—to pesticides and pesticides residues in their daily lives. For example, about 35% of the foods purchased by American consumers have detectable levels of pesticide residues.[4] This estimate is, in fact, conservative because we currently test for only about one-third of the pesticides in use.

In addition to food contamination, the public can also be exposed to pesticides in other ways. The principle exposure of the general public in the United States occurs in the home; about 90% of all U.S. households use pesticides on their lawn, garden, and/or the inside of the home.[5] Drinking water can also contain significant chemical and pesticide residues; at present, more than 10% of U.S. rivers and 5% of U.S. lakes are measurably polluted with pesticides.[5] Groundwater supplies can also be polluted when pesticides seep into aquifers or wells.[6] Finally, public exposure to pesticides can occur through accidents or spills, or even through the air during application, when pesticides drift from the target area into more populated towns and cities. In fact, only 25%–50% of pesticides applied by aircraft under ideal weather conditions actually reach the target area.[1]

Acute Effects: Pesticide Poisonings

In 1945, when synthetic pesticides were first used, few pesticide poisonings were reported. But by the late 1960s, both pesticide usage and toxicity had increased so dramatically that the number of human pesticide poisonings was substantial.[1] Unfortunately, this trend continued into the year 2000 (Figure 1). Just in the last decade, the total number of pesticide poisonings in the United States has increased from 67,000 in 1989, to the current level of 110,000 per year.[2] Worldwide, the increased use of pesticides results in approximately 26.5 million cases of occupational pesticide poisonings each year, and an unknown number of non-occupational pesticide poisonings.[7] Of all these estimated poisoning episodes, about 3 million cases are hospitalized, resulting in approximately 220,000 fatalities and about 750,000 cases of chronic illness every year.[8]

Poisonings can occur when pesticides contact the skin or eyes, are inhaled, or are ingested. Typically, pesticides can have acute local effects on the area they directly contact—skin, eye, or respiratory tract irritation—in addition to acute systemic effects. Different pesticides act on different systems in the body in a variety of ways, but most common insecticides—organophosphates and organo-chlorines (i.e., Parathion)—have acute neurotoxic effects. This means that they impair normal functioning of the brain and/or the spinal cord, which can result in tremors, paralysis, seizures, and other systemic effects. Large degrees of exposure or exposure to highly toxic pesticides can be fatal.

Chronic Effects: Cancer and Other Health Concerns

Exposure to pesticides, though, does not always occur at a level sufficient to produce these acute symptoms. Many people are exposed to low levels of pesticides over a long period of time through their occupations or in the food they eat or water they drink. If no overt symptoms occur close to the time of exposure, we often assume that no damage is being done. However, pesticides have been associated with numerous chronic illnesses and health problems, especially in some highly sensitive individuals.

Chronic effects of pesticides are diverse and can affect most systems of the human body. The major types of chronic health effects that pesticides can have are neurological effects, respiratory and

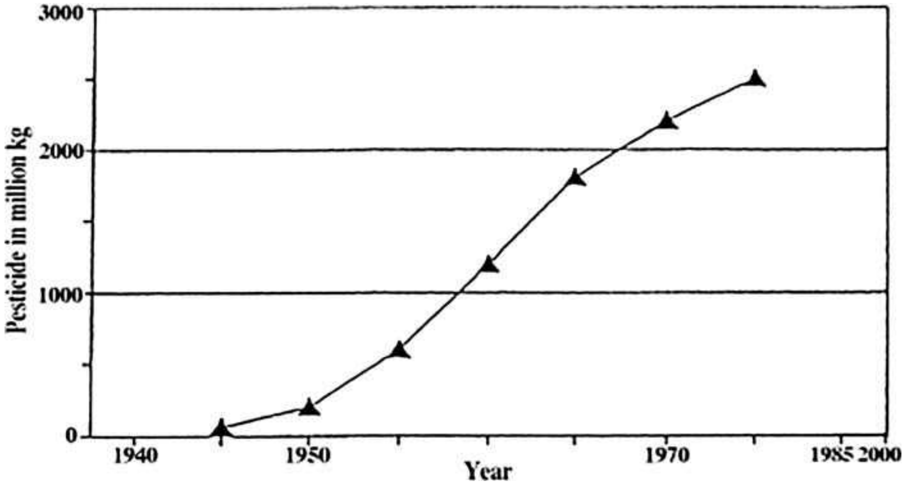


FIGURE 1 Trend in annual world pesticide use.[2]

reproductive effects, and carcinogenic effects. The chronic neurotoxic effects of pesticide exposure are not well understood, but there is some evidence linking pesticide exposure with symptoms like fatigue, muscle weakness, and sensory disturbances as well as cognitive effects such as memory loss, language problems, and learning impairment. The malady organophosphate-induced delayed poly-neuropathy (OPIDP) is well documented and includes irreversible neurological effects.

In addition to their neurotoxic effects, pesticides can have adverse effects on the respiratory and reproductive systems. For example, 15% of a group of professional pesticide applicators suffered asthma, chronic sinusitis, or chronic bronchitis as compared with 2% of people who used pesticides infrequently.[9] Studies have also linked pesticides with reproductive effects such as infertility and fetal deformities, but the data are still inconclusive.[10] Some pesticides have been shown to cause testicular dysfunction or sterility in animals, and similar effects are suspected in humans, but are less well understood at present.[10] Sperm counts in males in Europe and the United States, for example, declined by about 50% between 1938 and 1990[11]—a time period during which the use of synthetic pesticides increased about 30-fold and toxicity per pound increased about 10-fold.[1] At present, there is evidence that human sperm counts continue to decrease by about 2% per year.[3]

U.S. data indicate that 18% of all insecticides, and about 90% of all fungicides, are carcinogenic,[12] and many studies have shown that risks for certain types of cancers are higher in people—such as farmworkers and pesticide applicators—who are more frequently exposed to certain pesticides. Certain pesticides have been shown to induce tumors in lab animals; there is some evidence to suggest that they may have similar effects in humans.[10]

Many pesticides are also estrogenic—they mimic or interact with the hormone estrogen—linking them to the increased breast cancer rate among some groups of women in the United States. The breast cancer rate rose from 1 in 20 in 1960 to 1 in 8 in 1995.[13] There was a concurrent increase in pesticide use during that time period, and, although it has not been concretely linked to the increase in breast cancer rates, some studies suggest that exposure to pesticides is related to breast cancer incidence. Pesticides that interfere with the body's endocrine—hormonal—system can also have reproductive, immulogic, or developmental effects; these effects are well documented in animals and just starting to be understood in humans.[10] While endocrine disrupting chemicals may appear less dangerous at first glance—hormonal effects rarely result in acute poisonings or sudden death—their effects on reproductive and developmental processes may prove to have far-reaching and even more sinister consequences.[10]

Effects on Children

The negative health effects that pesticides can have—both acute and chronic—can be more significant in children than in adults, for several important reasons. First, children have much higher metabolic rates than adults, and their ability to activate, detoxify, and excrete toxic compounds is different from that of adults. Also, because of their smaller physical size, children are exposed to higher levels of pesticides per unit of body weight. In addition, certain types of pesticides are inherently more dangerous for children than for adults.[2,10] For example, the organophosphate and carbamate classes of pesticides adversely affect the nervous system by inhibiting cholinesterase, a critical enzyme, and can produce difficulty breathing, partial or total paralysis, convulsions, coma, or death. This problem is particularly significant for children since their brains are more than five times larger in proportion to their body weight than an adult's brain, making cholinesterase even more vital. In a California study, 40% of the children working in agricultural fields had blood cholinesterase levels below normal, a strong indication of organophosphate and carbamate pesticide poisoning.[14] In addition, a study in England and Wales has shown that 50% of all pesticide poisoning incidents in those countries involved children less than 10 years of age.[15] Use of pesticides in the home is also linked to childhood cancer.[16]

Because these studies and others have demonstrated that children's increased sensitivities to toxicants is a significant concern, some pesticide regulations have recently been reevaluated to provide special protections for children and infants. The Food Quality Protection Act of 1996 requires the EPA to specifically address risks that pesticides pose to infants and children and provides for an additional safety factors to take into account the potentially greater exposure and/ or sensitivity to pesticide effects on infants and children.

Economic Costs and Conclusions

Although no one can place a precise monetary value on a human life, the economic "costs" of human pesticide poisonings have been estimated. For our assessment, we use the conservative estimate of $2.2 million per human life—the average value that the surviving spouse of a slain New York City policeman receives.[17] Available estimates suggest that human pesticide poisonings and related illnesses in the United States cost about $933 million per year.[3] Pesticide use, though, provides a substantial net agricultural return of $12 billion per year.[2] Are the public health risks associated with pesticide use a great enough concern to warrant a reduction in pesticide use?

Given the known—and suspected—adverse effects of pesticide use on public health discussed in previous sections, it seems fairly clear that our present levels and methods of pesticide use present significant public health dangers and concerns. However, the abrupt and complete cessation of synthetic pesticide use is not without substantial costs of its own. Termination of pesticide use would result in huge crop losses—the food supply would be severely reduced and a huge number of people would starve in a short time,[18] an especially dire consequence given the 3 billion people—half the world's population—currently malnourished worldwide.[19]

Clearly, it is essential that all the costs and benefits—economic, environmental, and social/health—of pesticide use be considered when current and future pest control programs are being developed and evaluated. A recent study estimated that the environmental and social costs related to U.S. pesticide use—crop losses, public health effects, pesticide resistance, water pollution, and other environmental effects—total $8.3 billion each year.[3] Furthermore, another study has shown that U.S. pesticide use can be reduced up to one-half without any reduction in crop yields and cosmetic standards, and only a minimal 0.6% increase in food costs.[20] Therefore, it is clear that our current methods of chemical pesticide use need to be reconsidered and evaluated with the above-outlined public health effects and the related economic costs in mind, toward the goal of the development of sound, sustainable pest management practices that maximize the benefits of pesticide use while at the same time minimizing the adverse effects that pesticides can have on human health.

References

1. Pimentel, D. Protecting crops. In *The Literature of Crop Science*; Olsen, W.C., Ed.; Cornell University Press: Ithaca, NY, 1995; 49–66.
2. Pimentel, D.; Hart, K.A. Ethical, environmental, and public health implications of pesticide use. In *Perspectives in Bioethics*; Galston, A., Ed.; Johns Hopkins University Press: Baltimore, MD, 2000, in press.
3. Pimentel, D.; Greiner, A. Environmental and socio economic costs of pesticide use. In *Techniques for Reducing Pesticide Use*; Pimentel, D., Ed.; John Wiley & Sons: Chichester, 1997; 51–78.
4. FDA. Food and drug administration pesticide program residues in foods—1989. *J. Assoc. Anal. Chem.* **1990**, *73*, 127A–146A.
5. Pimentel, D.; Wilson, C.; McCullum, C.; Huang, R.; Dwen, P.; Flack, J.; Tran, Q.; Saltman, T.; Cliff, B. Economic and environmental benefits of biodiversity. *BioSci.* **1997**, *47* (11), 747–757.
6. Pimentel, D.; Acquay, H.; Biltonen, M.; Rice, P.; Silva, M.; Nelson, J.; Lipner, V.; Giordano, S.; Horowitz, A.; D'Amore, M. Assessment of environmental and economic impacts of pesticide use. In *The Pesticide Question: Environment, Economics and Ethics*; Pimentel, D., Lehman, H., Eds.; Chapman & Hall: New York, 1993; 47–84.
7. UNEP. *Global Environmental Outlook*; United Nations Environment Program: Nairobi, 1997.

8. WHO. *Our Planet, Our Health: Report of the WHO Commission on Health and Environment;* World Health Organization: Geneva, 1992.
9. Weiner, B.P.; Worth, R.M. Insecticides: Household use and respiratory impairment. In *Adverse Effects of Common Environmental Pollutants*; MSS Information Corporation: New York, 1972; 149–151.
10. Colborn, T.; Dumanoski, D.; Myers, J.P. *Our Stolen Future: Are We Threatening Our Fertility, Intelligence, and Survival? A Scientific Detective Story*; Dutton: New York, 1996.
11. Carlsen, E.; Giwercman, A.; Keiding, N.; Skakkebaek, N.E. Evidence for decreasing quality of semen during past 50 years. *Br. Med. J.* **1992**, *305*, 609–613.
12. NAS. *Regulating Pesticides in Food*; National Academy of Sciences: Washington, DC, 1987.
13. McCarthy, S. Congress takes a look at estrogenic pesticides and breast cancer. *J. Pestic. Reform* **1993**, *13* (4), 25.
14. Repetto, R.; Baliga, S.S. *Pesticides and the Immune System: The Public Health Risks*; World Resources Institute: Washington, DC, 1996.
15. Casey, P.; Vale, J.A. Deaths from pesticide poisoning in England and Wales: 1945–1989. *Human Exp. Toxicol.* **1994**, *13*, 95–101.
16. Leiss, J.K.; Savitz, D.A. Home pesticide use and childhood cancer: A case-control study. *Am. J. Public Health* **1995**, *85*, 249–252.
17. Nash, E.P. What's a Life Worth; New York Times: New York, 1994.
18. Lehman, H. Values, ethics, and the use of synthetic pesticides in agriculture. In *The Pesticide Question: Environment, Economics and Ethics*; Pimentel, D., Lehman, H., Eds.; Chapman & Hall: New York, 1993; 347–379.
19. WHO. *Micronutrient Malnutrition—Half the World's Population Affected*; World Health Organization: Geneva, 1996; 78, 1–4.
20. Pimentel, D.; McLaughlin, L.; Zepp, A.; Lakitan, B.; Kraus, T.; Kleinman, P.; Vancini, T.; Roach, W.J.; Graap, E.; Keeton, W.S.; Selig, G. Environmental and economic impacts of reducing U.S. agricultural pesticide use. In *Handbook on Pest Management in Agriculture*; Pimentel, D., Ed.; CRC Press: Boca Raton, FL, 1991; 679–718.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

5

Nanoparticles

Alexandra
Navrotsky

Introduction

Nanoparticles have no exact definition, but they are aggregates of atoms bridging the continuum between small molecular clusters of a few atoms and dimensions of 0.2–1 nm and chunks of solid containing millions of atoms and having the properties of macroscopic bulk material. In water, nanoparticles include colloids; in air, they include aerosols. Nanoparticles are ubiquitous. We pay to have them. We pay more to not have them. They occur as dust in the air, as suspended particles that make river water slightly murky, in soil, in volcanic ash, in our bodies, and in technological applications ranging from ultratough ceramics to microelectronics. They both pollute our environment and help keep it clean. Microbes feast on, manufacture, and excrete nanoparticles.

Understanding nanoparticle formation and properties requires sophisticated physics, chemistry, and materials science. Tailoring nanomaterials to specific applications requires both science and Edisonian inventiveness. Applying them to technology is state-of-the-art engineering. Tracing their transport and fate in the environment invokes geology, hydrology, and atmospheric science. Applying them to improving soil fertility and water retention links soil science and agriculture to surface chemistry. Understanding their biological interactions brings in fields ranging from microbiology to medicine. Probing the impact of nanoparticles on humans and of human behavior on the production and control of nanoparticles requires the behavioral and social sciences, e.g., in dealing with issues of automotive pollution. The purpose of this review is to describe some of the unique features of nanoparticles and to discuss their occurrence and importance in the natural environment.

Although we often think of the natural environment as that part of the planet which we can see, a somewhat broader definition includes the "critical zone": the atmosphere, hydrosphere, and shallow portion of the solid earth that exchange matter on a geologically short time scale, on the order of tens to thousands of years. This critical zone affects us directly, and our activities influence it. Because of the active chemical reactions continuously taking place in the critical zone, and because its temperatures and pressures are relatively low and it is dominated by water, solids are constantly being formed and decomposed. Many of these solids start out as nanoparticles; many remain so. In a yet broader sense, our entire planet from crust to core, the solar system, and the galaxy are part of our environment.

Physical Chemistry of Nanoparticles

A major feature of nanoparticles is their high surface-to-volume ratio. Figure 1 shows the volume fraction within 0.5 nm of the surface for a spherical particle of radius *r*. One can think of this fraction either as the fraction of atoms likely to be influenced by processes at the surface, or as the fraction of the volume of a material that could be taken up by a 0.5-nm coating of another material. In the first case, because the surface dominates chemical reactivity, the increased surface to volume ratio means that nanoparticles dominate chemical reactions. In the second case, the ability to carry a substantial coating offers a mechanism for the transport of nutrients or pollutants.

Many oxides are *polymorphic,* exhibiting several crystal structures as a function of pressure and temperature. Often, nanosized oxide particles crystallize in structures different from that of large crystals of the same composition.[1] Examples are γ-Al_2O_3, a defect spinel rather than α-Al_2O_3, corundum, γ-Fe_2O_3, the defect spinel maghemite rather than α-Fe_2O_3, hematite, and the anatase and brookite forms of TiO_2 rather than rutile. From arguments based on transformation sequences and the occurrence of phases, it was long argued that there may be a crossover in phase stability at the nanoscale if the structure which is metastable for large particles has a significantly lower surface energy.[2] This has been proven for alumina and titania in recent calorimetric studies (Figure 2).[3,4] The resulting transformation enthalpies and surface energies, and those of other related systems are shown in Table 1. Another interesting feature is that the hydrous phases AlOOH boehmite and FeOOH goethite have significantly lower surface energies than their anhydrous counterparts, Al_2O_3 and Fe_2O_3.[5,6] Whether this is a general feature of hydrous minerals with hydroxylated surfaces is not yet known.

As particles become less than about 10 nm in size, their x-ray diffraction patterns are broadened sufficiently that they begin to appear "x-ray amorphous" (Figure 3). This term lacks exact definition. High-resolution electron microscopy may still detect periodicity, and short-range order is certainly present.[7] The identification of structure in 1–10 nm particles is very difficult, and phases are empirically described as, for example, "two line ferrihydrite," based on x-ray diffraction patterns.[8]

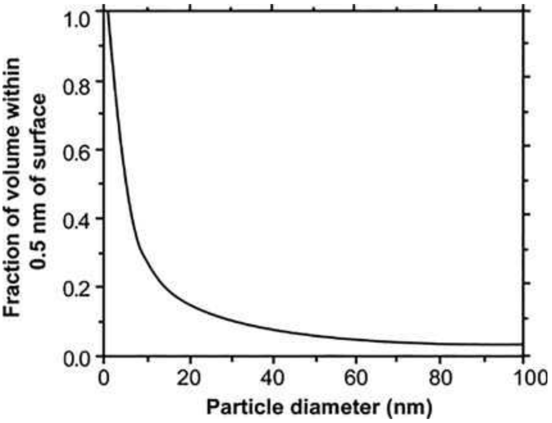


FIGURE 1 Volume fraction of a nanoparticle within 0.5 nm of the surface as a function of particle radius.
Source: Navrotsky,[41] Kluwer Academic Publishers.

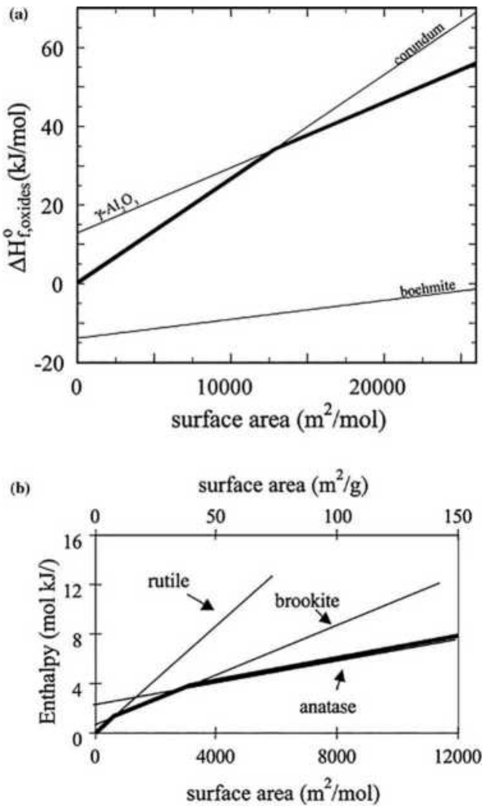


FIGURE 2 (a) Enthalpies of alumina polymorphs as a function of surface area. (**Source:** From McHale et al.[3]) (b) Enthalpies of titania polymorphs as a function of surface area. The heavy lines show the stable polymorphs in each size range. (**Source:** Ranade,[4] PNAS.)

TABLE 1 Energetic Parameters for Oxide and Oxyhydroxide Polymorphs

Formula	Polymorph	Metastability (kJ/mol)	Surface Energy (J/m^2)
Al_2O_3[a]	Corundum (α)	0	2.6
	Spinel (γ)	13.4	1.7
Fe_2O_3[b]	Hematite	0	0.8
	Maghemite	20	0.8
TiO_2[c]	Rutile	0	2.2
	Brookite	0.7	1.0
	Anatase	2.6	0.4
AlOOH[d]	Diaspore	0	?
	Boehmite	4.9	0.5
FeOOH[b]	Goethite	0	0.3
	Lepidocrocite		0.3

[a] From McHale et al.[3]

[b] From Majzlan.[6]

[c] From Ranade.[4]

[d] From Majzlan.[5]

Nanoparticles in Soil and Water

Soil is a complex aggregate of inorganic, organic, and biological material.[9] Its constituents of largest size are rocks and gravel, small animals, plant roots, and other debris. Smaller mineral grains, clumps of organic matter, and microorganisms make up an intermediate size fraction. The smallest particles, ranging into the nanoscale, are clays, iron oxides, and other minerals. These are often heterogeneous and coated by other minerals and organic matter. The entire composite is porous and hydrated. The percolation of water in soil transports both nanoparticles and dissolved organic and inorganic species. The texture and porosity, as well as the chemical composition and pH, are crucial to biological productivity. The surfaces of nanoparticles provide much of the chemical reactivity for both biological and abiotic processes.

Major aluminosilicate minerals in soils include clays, zeolites, and poorly crystalline phases (Table 2). These can change their water content in response to ambient conditions, often swelling in wet seasons, and shrinking in dry seasons. These nanophase materials are major controllers of soil moisture and permeability. Iron and manganese oxides are another class of major sol minerals. Their extensive polymorphism at the nanoscale makes them highly variable. They sequester and/or transport and make available the essential plant nutrient iron, as well as other essential transition metals (cobalt, copper, zinc, etc.). They frequently carry coatings of other metal oxides and oxyhydroxides, including toxic metals such as lead and chromium. They also frequently have organic coatings. Sulfates, including the jarosite–alunite family of hydrated [(K, Na), (Al, Fe)] sulfates, are another important constituent. In alkaline and arid environments, other sulfates and halides form, and their formation, dissolution, and transport is a major issue in heavily irrigated regions. How much these processes are controlled by nanoscale phenomena is not known.

Groundwater is constantly in touch with soil and rock, and minerals are dissolving and precipitating as it flows. The load of fine sediments in streams and groundwater can be substantial, especially during

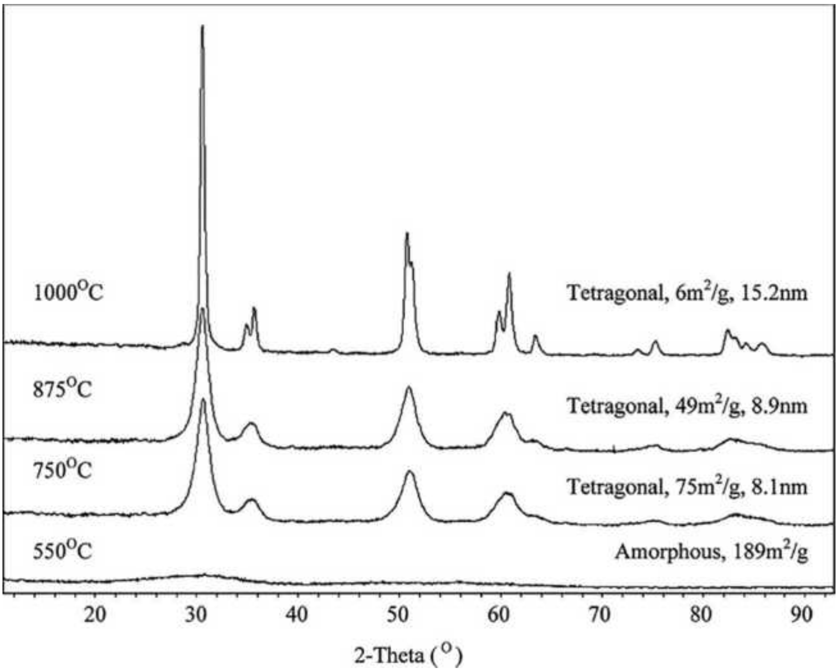


FIGURE 3 Powder x-ray diffraction patterns of sol–gel zirconia heated at various temperatures. The structure and average particle diameters are indicated.
Source: Pitcher and Navrotsky.[42]

TABLE 2 Major Soil Minerals and Constituents

Type	Composition	Structures
Clay	Hydrated aluminosilicate	Layered
Zeolite	Hydrated aluminosilicate	Three-dimensional porous
Salts	NaCl, Na_2SO_4, $CaSO_4$	Ionic crystals
Carbonates	$CaCO_3$–$MgCO_3$–$FeCO_3$	Calcite, dolomite, others
Allophane	Hydrous aluminosilicate gel	Amorphous
Iron oxides	$Fe2O_3$, FeOOH	Various polymorphs
Aluminum oxides	AlOOH, $Al(OH)_3$	Various polymorphs
Quartz	SiO_2	Quartz
Manganese oxides	Mn_2O_3, MnOOH, MnO_2	Various polymorphs
H_2O	H_2O	Water, ice, vapor
Organics	C–H–N–O	Large surface area amorphous colloids
Jarosite-alunite	Alkali (Fe, Al) sulfates	Ionic double salts

spring floods. The Missouri River is called “the Big Muddy” because of its load of particulate matter, a large fraction of which is of nanoscale dimensions. The yearly flooding of the Nile, depositing fertile soil with its large nanoparticle content, made ancient Egyptian civilization flourish. Today, one of the major concerns of our system of dams, especially in the arid western United States, is interference with the normal cycle of sediment transport and “silting up” of the lakes behind the dams. Silt is partly nanoparticles.

Contaminants and pollutants in water can be transported as aqueous ions (dimensions < 0.5 nm), as molecular clusters (0.5–2 nm), as nanoparticles (2–100 nm), as larger colloids (100–1000 nm), and as macroscopic particles (>1 μm). These size range distinctions are rather arbitrary and serve to illustrate the continuity between the dissolved and the solid state. Several examples illustrate this complexity. Aluminum oxyhydroxide particles can transport transition metals such as nickel, cobalt, and zinc, seemingly as adsorbed coatings. Initially thought to be loosely bound metal complexes at the surface of the aluminum oxyhydroxide mineral grain, these are now realized to be precipitates, only a few atomic layers thick, of mixed double hydroxides of the hydrotalcite family, in which anions such as carbonate play an essential role.[10] The transport of plutonium through groundwater is a concern in old plutonium processing facilities such as the Hanford, WA atomic energy reservation, in the Nevada nuclear test site, and in the planned nuclear waste repository at Yucca Mountain, Nevada. There remain questions of permeability and the adhesion of particles to the rock and engineered barrier walls, of colloid transport, of biological transport, and of mineral precipitation which can change the rate of progress of a contamination plume. Linking laboratory scale, field scale, and simulation studies of nanoparticle transport is an essential area of research for understanding radioactive and chemical contamination and geologic processes involving uranium and other actinides.[11]

When particles are below 5 nm in size, several other effects must be considered. Whereas for larger particles, most of the atoms are in specific planes or faces, for smaller ones, an increasing number of surface atoms must sit at the intersection of facets, in presumably even higher energy sites. An alternate, more macroscopic way of describing this is to consider the surface as curved, rather than as a series of planes. Then the surface energy per unit area is no longer a constant, but potentially increases quite rapidly with decreasing particle size. This unfavorable energy may be relaxed by the adsorption of various molecules on the surface, and there is evidence that the adsorption coefficient of organics rises steeply at very small particle size.[12]

The flocculation of colloids depends on the surface charge; the pH of which the surface is neutral is the “point of zero charge.”[13] Does this depend on particle size? This is an area of active research.

How do nanocrystals form from solution? The classical picture of nucleation and growth by addition of single atoms or ions is probably inadequate.[14] There is increasing evidence for clusters of atoms or

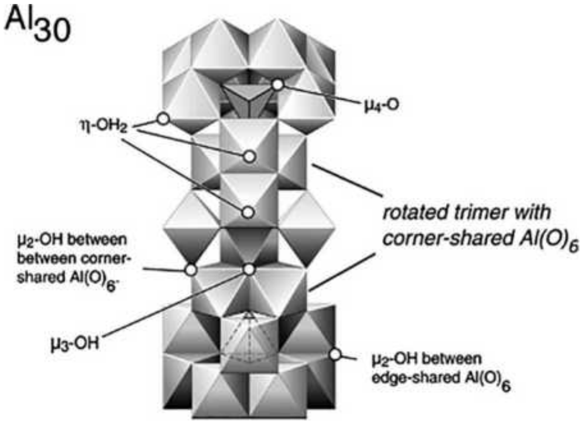


FIGURE 4 $Al_2O_8Al_{28}(OH)_{56}(H_2O)_{24}{}^{18+}(aq)$ (often called Al_{30}) cluster of 2 nm dimensions, which is intermediate in structure and properties between isolated ions and solid aluminum.
Source: W.H. Casey, personal communications. From Rowsell and Nazar.[15]

ions in solution which contain 5–50 atoms and clearly show some of the structural features of the solid. An example is the Keggin-molecular cluster containing 13 aluminum atoms shown in Figure 4.[16] It appears stable over a wide range of neutral to basic pH, and is probably a major precursor to and a dissolution product of aluminum oxyhydroxides.[17] The growth of TiO_2 anatase may occur by the oriented attachment of ~3-nm particles.[18] The growth of zeolites templated by organics may involve 3-nm cuboctahedral clusters.[19] Nanoclusters have been invoked in the growth of sulfides in ore-forming solutions.[20] Characterization of such nanoscale precursors in aqueous solutions remains a major challenge.

The coarsening and phase transformation of nanoscale precipitates upon heating is equally important to the sol–gel synthesis of ceramics and the geologic compaction and diagenesis of buried sediments. Loss of water, loss of surface area, and phase transformations to the stable bulk polymorph are closely interlinked.[14] A nanoparticle with dimensions below 5 nm probably cannot maintain defects or dislocations; they can migrate to the surface and be annihilated.[21] An aggregate of such single domain nanocrystals, with disorder and impurities at their surfaces, may be a common morphology in nature. Such aggregates give smaller average particle size from x-ray peak broadening than from gas adsorption [Brunauer–Emmett–Teller (BET)] measurements.

Magnetic nanoparticles cannot hold a direction of magnetization for a long time because of thermal fluctuations.[22] The magnetic iron oxides found in magnetotactic bacteria, which are single domain particles, neither too large nor too small, provide orientation in the Earth's magnetic field (see below). On a geologic time scale (millions of years), magnetization of largely nanophase iron oxides provides a record of the variation of the Earth's magnetic field through time, including periodic reversals of north and south poles. Thus the ability or inability of nanoparticle oxides to retain magnetization is of critical importance.

Nanoparticles in the Atmosphere

Atmospheric particles include dust (rock and soil), sea salt, acids including sulfuric, organics (especially carbon), inorganics, and, of course, water and ice (Table 3). The atmosphere can carry particles of spherical equivalent diameters from 1 to 10^5 nm. Often, a trimodal distribution of particle sizes, with peaks in number density near 5, 50, and 300 nm, is seen.[23] The smaller particles account for most of the reactive surface area but little of the total mass.

Atmospheric particles affect the environment in many ways.[23,24] They reduce visibility (smog, haze) primarily through light scattering. They act as nuclei for water vapor condensation and cloud formation.

TABLE 3 Atmospheric Nanoparticles

Liquid droplets
Water
Sulfuric acid
Nitric acid
Sea water and other salt solutions
Organics
Solid particles
Ice (H_2O)
NaCl
Na_2SO_4
$CaSO_4 \cdot 2H_2O$
$NaNO_3$
$H_2SO_4 \cdot 4H_2O$
$HNO_3 \cdot 3H_2O$
C (graphite, amorphous, fullerenes, nanotubes)
SiO2
Iron oxides
Clays
Organics
Many particles have core–shell structures and coatings

They are involved in radiative forcing, changing the ratio of absorbed sunlight to reflected sunlight. Thus they are implicated in global climate change. Their effect on radiative forcing can be positive (more energy adsorbed) or negative (more energy reflected), leading to warming or cooling, respectively.[23,24] Their net effect is a subject of vigorous research and controversy.

Anthropogenic particles contribute disproportionately to the fine particle fractions.[23] These may have very significant effects on climate and (see below) health. Soot (carbon) from burning coal and oil and from automobile (especially diesel) emissions contributes greatly to the atmospheric load of nanoparticles.

Particles in the atmosphere travel a long way. Dust from Africa is seen in Florida; industrial emissions from China are detected in North America. Particles are removed from the atmosphere by diffusion and gravitational settling (aided by small particles coalescing into larger ones) and by rain. The residence time of nanoparticles in the atmosphere ranges from minutes to days.[23]

Atmospheric nanoparticles are more involved in gas phase reactions than particles in soil and water.[24] Their formation may involve combustion synthesis, as in industrial or automobile emission. Mineral nanoparticle surfaces may catalyze the oxidation of SO_2 and NO_2, leading to sulfuric and nitric acid. These acids can exist as gaseous species, liquids, or solid hydrates at low temperature. Nanoparticles are invoked in the depletion of atmospheric ozone by catalytic production of reactive chlorine compounds. Changes of phase (liquid to solid) are critical to the chemistry of sodium chloride and sodium nitrate particles, with their water content being controlled by available humidity.

Mineral dust particles may provide critical nutrients (e.g., iron) to the surface of the ocean far from land. The ocean's biological productivity is often limited by the availability of these nutrients; thus such inorganic nanoparticles may significantly influence the global cycling of carbon through ocean biomass.[25]

Nanoparticles in Sediments, Rocks, and the Deep Earth

The debris of rock weathering is brought down river to the ocean in sediments consisting of nanoscale particles of clay, small quartz grains, and other minerals. Indeed, the terms "clay" and "silt" have a classic connotation of size fraction, although the former also implies a structural group of minerals, the layered aluminosilicates. In the ocean, carbonates precipitate, dissolve, and reprecipitate as a complex function of depth.[26] Both silica and various polymorphs of calcium carbonate (calcite, aragonite, vaterite) are produced by organisms such as diatoms, foraminifera, and corals. Their debris rains down on the ocean bottom, forming sediments which often show annual cycles in composition and texture and which bear records of climate change, shifts in ecosystems, and catastrophic events such as meteor impacts.[27] These sediments start off largely nanoscale. They coarsen and dehydrate with time and depth of burial. The evolution of their organic matter leads to petroleum. The evolution of their minerals, involving coarsening and compaction, called diagenesis, leads to rocks such as limestones and shales. The nanoscale processes that take place (dehydration and organic loss, phase transformation, coarsening and densification) are natural analogs of ceramic processing which starts with nanoscale precipitates or gels.

Natural processes involving changes in temperature, pressure, acidity, and oxygen fugacity cause the concentration of trace metals into ore deposits. These often occur in hydrothermal systems, spatially contained circulations of hot, pressurized, metal-rich aqueous solutions. Our ability to mine low-grade deposits by chemical leaching techniques brings us into the world of nanoparticles and reactions at mineral surfaces. There is increasing evidence that microorganisms play an active role in ore deposition.[28,29] Hot springs at the surface produce deposits of nanoscale amorphous silica and other minerals, which may also be closely linked to microbial activity.[30]

At temperatures above a few hundred degrees Centigrade and pressures above a few kilobars, coarse-grained metamorphic and igneous rocks predominate. The interior of the Earth is layered, with seismic discontinuities delineating the crust, upper mantle, transition zone, lower mantle, and core. These discontinuities represent regions of rapidly changing density, mineralogy, and chemistry.[31] Ongoing phase transitions and chemical reactions can decrease the grain size of a material and render it easier to deform.[32] Thus nanoscale phenomena, occurring at specific locations, may play a disproportionate role in processes such as subduction, plate tectonics, earthquake generation, and volcanism. Shock processes, (e.g., meteor impact, nuclear detonation) also produce nanoparticles.

When a volcano erupts explosively, a plume of dust particle is sent into the atmosphere, sometimes reaching the stratosphere. These particles make beautiful sunsets but they also exert a significant cooling effect on climate for several years and pose a significant aviation hazard. Combining sedimentation, coarsening, subduction, volcanism, and weathering, there is an ongoing global geochemical cycle of nanoparticles, analogous in some ways to global geochemical cycles of elements such as carbon. However, the mass balances, or imbalances, in global nanoparticle production and consumption through time have not been characterized.

Nanoparticles beyond the Earth

In the early stages of planet formation, dilute and more or less uniform gas condensed to form a series of mineral particles, with the order of condensation described by thermodynamic calculations based on the volatility and stability of these phases.[37] The more refractory oxides condensed earlier than those with higher volatility. These particles accreted, under the influence of gravity, to form our solar system. What was the nature of these initial particles? What was their size distribution? Were they crystalline or amorphous? Were metastable polymorphs formed? While the initial high temperatures might argue against such metastability, the low pressures and condensation from a vapor argue for it. In technological processes, chemical vapor deposition produces nanoscale amorphous silica "snow," and combustion produces soot and inorganic nanoparticles. The role of nanoparticles in planetary accretion has not yet been explored. The change in stability at the nanoscale, which will be different for various compositions and polymorphs, may alter the sequence of condensation of phases. Are the particles now present in space as interplanetary dust partly or mostly nanoparticles?

The surfaces of the Moon and Mars, subject to "space weathering" by bombardment with meteorites of all sizes, contain an extensive fine grained dust or soil layer.[32] Samples of lunar soil, brought back by the Apollo missions, contain a distribution of particle sizes of spherules and irregular shards. Their particle size distribution appears not to have been a subject of active interest, but clearly a significant number are in the nanoregime. The red surface of Mars appears to be dominated by various fine-grained or nanophase iron oxides. Until Martian sample return missions, planned to occur in the next decade or two, bring some of this material to Earth, we must rely on remote sensing technology (spectroscopic techniques) and instrumentation on Martian landers (possibly Mossbauer spectroscopy, x-ray fluorescence, and x-ray diffraction) to obtain information on the composition and structure of Martian soil. Considering the difficulty of characterizing iron oxide nanoparticles in the best laboratories on Earth, definitive conclusions about the nature of Martian soil are unlikely until we have some samples in hand. Meteorites believed to be from Mars contain micron-sized spherules, which were proposed to be biological in origin. This sparked much recent controversy and it is by no means settled whether these structures are fossil microorganisms or the product of inorganic nanoscale crystal growth processes.[33,34]

Nanoparticles and Life

Microbial communities are rich in the production and utilization of nanoparticles.[35] Table 4 lists some examples. In addition to aerobic respiration (the enzymatic oxidation of carbohydrates and other organics with molecular oxygen to produce water, carbon dioxide, and energy stored as high-energy phosphate linkages) organisms use many other strategies to extract energy from the environment. The following biological reactions produce or consume nanoparticles. Dissolved Mn(II) or Fe(II) can be oxidized by oxygen, producing Mn(III), Mn(IV), or Fe(III) oxide nanoparticles, while organics can be oxidized by manganese or iron oxides, producing soluble Mn(II) and Fe(II) species. Some bacteria can also utilize the U(IV)–U(VI) couple as an energy source. Because hexavalent uranium is much more soluble than tetravalent, biological processes that accelerate its production are of concern in modeling nuclear waste leaching. The sum of these two groups of redox processes is the oxidation of organics by oxygen, akin to respiration. The important difference is that the organic food source and the oxygen source can be spatially separated in the sharp gradients in oxygen and organic contents that frequently occur in sediments, and different communities of organisms participate in the two processes. In marine sediments, sulfate is the dominant biological electron acceptor and is more important than oxygen. Bacterial sulfate reduction produces sulfide which often precipitates as nanophase metal sulfide minerals. Sulfide and sulfur oxidizing bacteria typically live in specialized environments where there is enough oxygen to oxidize sulfur but not so much that chemical oxidation swamps biological oxidation. This oxidation consumes solid sulfur and sulfides, and produces soluble sulfate.

TABLE 4 Example of Interaction of Microorganisms and Nanoparticles

Class of Organisms	Example	Nanoparticle Interaction
Iron and manganese oxidizing bacteria	*Thiobacillus*	Oxidize soluble Mn^{2+} and Fe^{2+} to insoluble higher oxides
Iron and manganese reducing bacteria	*Shewenella*	Reduce insoluble Mn and Fe oxides to soluble forms oxidize sulfide to sulfate to sulfide precipitate
Sulfur reducing bacteria	*Thiobacillus*	Reduce sulfate to sulfur in sulfide
Magnetotactic bacteria	*Aquaspirillum magnetotacticum*	Nanoparticles of Fe_2O_3, Fe_3O_4, and/or iron sulfides
Uranium reducing bacteria	*Geobacter, shewenella*	Soluble $U^{6+} \rightarrow$ insoluble U^{4+}
Fungi	Specific strains unknown	Oxidize Mn^{2+}, precipitate MnO_2
Diatoms	Various	Precipitate silica
Foraminifera	Various	Precipitate $CaCO_3$ calcite and aragonite

Organisms utilize nanoparticles in processes other than respiration. Bacterial precipitation of sulfide minerals, e.g., ZnS and UO_2, may also be a mechanism of detoxification.[36] Similar detox processes may occur in plants. Magnetotactic bacteria synthesize and align single domain magnetic iron oxide and iron sulfide particles in structures called magnetosomes.[37] Such bacteria align themselves both north–south and vertically in the Earth's magnetic field. The navigational (homing) capabilities of bees, pigeons, and probably other higher organisms utilize magnetic field orientation sensed by magnetic iron oxide particles in their brains. Similar particles, although at lower abundance, occur in many mammals, including *Homo sapiens*.[38] There has been a debate in the public sector whether the magnetic fields produced by high-voltage power lines are potentially dangerous to human health. In contrast, the use of magnets in alternative medicine, and the market for magnetic pillows, back supports, etc., suggests, or at least hopes for, a beneficial effect of the interaction of magnetic fields with animals. Key to either harmful or helpful biological effects is a mechanism for the magnetic field to interact with living cells. Interaction with biological magnetic nanoparticles may provide such a mechanism, but very little is known at present.

Nanoparticles have other documented health effects.[39] When inhaled into the lungs, particles cause an inflammatory response, which contributes to allergies, asthma, and cancer.[48] The detailed mechanism of this response, and how it depends on surface area, particle size, or specific particle chemistry, is not clear. The harmful effects of inhaled particles may be enhanced by other pollutants, particularly ozone, typically present in smog. Nanoparticles penetrate deep into the lungs. Many are returned with exhaled air, some stick to the surfaces of the alveoli, and some may even penetrate into general blood circulation and be transported to other organs. Studies linking detailed nanoparticle characterization, biochemical and physiological processes, and health effects are just beginning to be carried out. It is likely that not all particles have comparable effects, and understanding which are the most dangerous could lead to rational, rather than arbitrary, emission standards for automotive and industrial particulates.

In the early Earth, prebiotic processes culminated in the origin of life.[40] Because the synthesis of complex organic molecules competes with their destruction by hydrolysis and other degradation, it is possible that the most successful synthesis could have occurred in sheltered and catalytic environments, such as those provided by mineral surfaces, nanoparticle surfaces, and pores within mineral grains. Present-day organisms utilize a wide variety of elements (e.g., Fe, Co, Ni, Cr, Zn, Se) in specific enzymes. Although large amounts of such elements are toxic, trace amounts are essential. The active centers in enzymes utilizing these trace elements often consist of clusters of metal atoms, sometimes associated with sulfide. Are these fine-tuned by evolution from earlier simpler metal clusters and nanoparticles existing in the environment? Thus nanoparticles may play a role not just in the sustenance of life but in its origin.

Conclusions

Nanoparticles play diverse roles in the environment and are involved in both abiotic and biologically mediated chemical and physical processes. Their high surface area, chemical reactivity, polymorphism, and unique properties involve nanoparticles in a disproportionately large fraction of the chemical reactions occurring on and in the Earth and other planets. Understanding this involvement is itself evolving into a new field of study in the environmental and Earth sciences, which is beginning to be called "nanogeoscience." Nanogeoscience will take its place alongside other new areas such as astrobiology and biogeochemitry, fields that link physical, chemical, and biological processes viewed in the context of the long time and distance scales natural to the Geosciences.

References

1. Navrotsky, A. Thermochemistry of Nanomaterials. In *Nanoparticles and the Environment: Reviews in Mineralogy and Geochemistry*; Banfield, J.F., Navrotsky, A., Eds.; Mineralogical Society of America: Washington, DC, 2001; Vol. 44, 73–103.

2. Garvie, R.C. The occurrence of metastable tetragonal zirconia as a crystallite size effect. *J. Phys. Chem.* **1965**, *69*, 1238–1243.
3. McHale, J.M.; Auroux, A.; Perrotta, A.J.; Navrotsky, A. Surface energies and thermodynamic phase stability in nanocrystalline alumina. *Science* **1997**, *277*, 788–791.
4. Ranade, M.R.; Navrotsky, A.; Zhang, H.Z.; Banfield, J.F.; Elder, S.H.; Zaban, A.; Borse, P.H.; Kulkarni, S.K.; Doran, G.S.; Whitfield, H.J. Energetics of nanocrystalline TiO_2. *Proc. Natl. Acad. Sci.* **2002**, *99* (Suppl. 2), 6476–6481.
5. Majzlan, J.; Navrotsky, A.; Casey, W.H. Surface enthalpy of boehmite. *Clays Clay Miner.* **2000**, *48*, 699–707.
6. Majzlan, J. Ph.D. Thesis; University of California at Davis, 2002.
7. Janney, D.E.; Cowley, J.M.; Buseck, P.R. Structure of synthetic 2-line ferrihydrite by electron nanodiffraction. *Am. Mineral.* **2002**, *85* (9), 1180–1187.
8. Schwertmann, U.; Cornell, R.M. *Iron Oxides in the Laboratory*, 2nd Ed.; Wiley-VCH: Weinheim, 2000; 188.
9. Singer, M.J.; Munns, D.N. *Soils: An Introduction*, 3rd Ed.; Simon and Schuster Company: Upper Saddle River, NJ, 1991.
10. Thompson, H.A.; Parks, G.A.; Brown, G.E., Jr. Ambienttemperature synthesis, evolution, and characterization of cobalt–aluminum hydrotalcite-like solids. *Clays Clay Mater.* **1999**, *47*, 425–438.
11. Ragnarsdottir, K.V.; Charlet, L. Uranium behavior in natural environments. *Environ. Mineral.* **2000**, *9*, 245–289.
12. Zhang, H.; Penn, R.L.; Hamers, R.J.; Banfield, J.F. Enhanced adsorption of molecules on surfaces of nanocrystalline particles. *J. Phys. Chem. B.* **1999**, *103*, 4656–4662.
13. Hunter, R.J. *Foundations of Colloid Science*; Oxford University Press: New York, 1993; Vol. 1.
14. Banfield, J.F.; Zhang, H. Nanoparticles in the Environment. In *Nanoparticles and the Environment: Reviews in Mineralogy and Geochemistry*; Banfield, J.F., Navrotsky, A., Eds.; Mineralogical Society of America: Washington, DC, 2001; Vol. 44, 2–58.
15. Rowsell, J.; Nazar, L.F. Speciation and thermal transformation in alumina sols: Structures of the polyhydroxyoxoaluminum cluster $[Al_{30}O_8(OH)_{56}(H_2O)_{26}]^{18+}$ and its Keggin moeté. *J. Am. Chem. Soc.* 2000, *122*, 3777–3778.
16. Casey, W.H.; Phillips, B.L.; Furrer, F. Aqueous Aluminum Polynuclear Complexes and Nanoclusters: A Review. In *Nanoparticles and the Environment: Reviews in Mineralogy and Geochemistry*; Banfield, J.F., Navrotsky, A., Eds.; Mineralogical Society of America: Washington, DC, 2001; Vol. 44, 167–190.
17. Furrer, G.F.; Phillips, B.L.; Ulrich, K.U.; Poethig, R.; Casey, W.H. The origin of aluminum flocs in polluted streams. *Science* 2002, *297* (5590), 2245–2247.
18. Penn, R.L.; Banfield, J.F. Oriented attachment and growth, twinning, polytypism, and formation of metastable phases: Insights from nanocrystalline TiO2. *Am. Mineral.* **1998**, *83*, 1077–1082.
19. de Moor, P.P.E.A.; Beelen, T.P.M.; Komanschek, B.U.; Beck, L.W.; Wagner, P.; Davis, M.E.; Van Santen, R.A. Imaging the assembly process of the organic-mediated synthesis of a zeolite. *Chem. Eur. J.* **1995**, *5*, 2083–2088.
20. Luther, G.W.; Theberge, S.M.; Richard, D.T. Evidence for aqueous clusters as intermediates during zinc sulfide formation. *Geochim. Cosmochim. Acta* **1999**, *64*, 579.
21. Jacobs, K.; Alivisatos, A.P. Nanocrystals as model systems for pressure-induced structural phase transitions. In *Nanoparticles and the Environment: Reviews in Mineralogy and Geochemistry*; Banfield, J.F., Navrotsky, A., Eds.; Mineralogical Society of America: Washington, DC, 2001; Vol. 44, 59–104.
22. Rancourt, D.G. Magnetism of earth, planetary, and environmental nanomaterials. In *Nanoparticles and the Environment: Reviews in Mineralogy and Geochemistry*; Banfield, J.F., Navrotsky, A., Eds.; Mineralogical Society of America: Washington, DC, 2001; Vol. 44, 217–292.

23. Anastasio, C.; Martin, S.T. Atmospheric Nanoparticles. In Nanoparticles and the Environment: *Reviews in Mineralogy and Geochemistry*; Banfield, J.F., Navrotsky, A., Eds.; Mineralogical Society of America: Washington, DC, 2001; Vol. 44, 293–349.
24. Ramanathan, V.; Crutzen, P.J.; Kiehl, J.T.; Rosenfeld, D. Aerosols, climate, and the hydrological cycle. *Science* **2001**, *294*, 2119–2124.
25. Martin, J.H.; Coale, K.H.; Johnson, K.S.; Fitzwater, S.E.; Gordon, R.M.; Tanner, S.J.; Hunter, C.N.; Elrod, V.A.; Nowicki, J.L.; Coley, T.L.; Barber, R.T.; Lindley, S.; Watson, A.J.; Vanscoy, K.; Law, C.S.; Liddicoat, M.I.; Ling, R.; Stanton, T.; Stockel, J.; Collins, C.; Anderson, A.; Bidigare, R.; Ondrusek, M.; Latasa, M.; Millero, F.J.; Lee, K. Testing the iron hypothesis in ecosystems of the equatorial Pacific Ocean. *Nature* **1994**, *371*, 123–129.
26. Millero, F.J. *Chemical Oceanography*, 2nd Ed.; CRC: Boca Raton, FL, 1996.
27. Broecker, W.S. *The Great Ocean Conveyor*; AIP Conference Proceedings; Columbia University Palisades: New York, 1992; Vol. 347, 129–161.
28. Labrenz, M.; Druschel, G.K.; Thomsen-Ebert, T.; Gilbert, B.; Welch, S.A.; Kemner, K.M.; Logan, G.A.; Summons, R.E.; de Stasio, G.; Bond, P.L.; Lai, B.; Kelly, S.D.; Banfield, J.F. Formation of sphalerite (ZnS) deposits in natural biofilms of sulfate-reducing bacteria. *Science* **2000**, *290*, 1744–1745.
29. Ehrlich, H.L. Microbes as geologic agents: Their role in mineral formation. *Geomicrobiol. J.* **1999**, *16*, 135–153.
30. Konhauser, K.O.; Phoenix, V.R.; Bottrell, S.H.; Adams, D.G.; Head, I.M. Microbial–silica interactions in Icelandic hot spring sinter: Possible analogues for some Precambrian siliceous stromatolites. *Sedimentology* **2001**, *48*, 415–433.
31. Anderson, D.L. *Theory of the Earth*; Blackwell Scientific Publications: Brookline Village, MA, 1989.
32. Karato, S.; Li, P. Diffusion creep in perovskite: Implications for the rheology of the lower mantle. *Science* **1992**, *255*, 1238–1240.
33. Sasaki, S.; Nakamura, K.; Hamabe, Y.; Kurahashi, E.; Hiroi, T. Production of iron nanoparticles by laser irradiation in a simulation of lunar-like space weathering. *Nature* **2001**, *410*, 555–557.
34. Buseck, P.R.; Dunin-Borkowski, R.E.; Devouard, B.; Frankel, R.B.; McCartney, M.R.; Midgley, P.A.; Posfai, M.; Weyland, M. Magnetite morphology and life on Mars. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98* (24), 13490–13495.
35. Gibson, E.K.; McKay, D.S.; Thomas-Keprta, K.L.; Wentworth, S.J.; Westall, F.; Steele, A.; Romanek, C.S.; Bell, M.S.; Toporski, J. Life on mars: Evaluation of the evidence within Martian meteorites ALH84001, Nakhla, and Shergotty. *Precambrian Res.* **2001**, *106* (1–2), 15–34.
36. Nealson, K.H.; Stahl, D.A. Microorgansims and Biogeochemical Cycles: What Can We Learn from Layered Microbial Communities? In *Geomicrobiology: Interaction Between Microbes and Minerals; Review in Mineralogy*; Banfield, J.F., Navrotsky, A., Eds.; Mineralogical Society of America: Washington, DC, 1997; Vol. 35, 5–34.
37. Suzuki, Y.; Banfield, J.F. Geomicrobiology of Uranium. In. *Uranium: Mineralogy, Geochemistry and the Environment; Review in Mineralogy*; Burns, P.C., Finch, R., Eds.; Mineralogical Society of America: Washington, DC, 1999; Vol. 38, 388–432.
38. Stolz, J.F. Magnetotactic Bacteria: Biomineralization, Ecology, Sediment Magnetism, Environmental Indicator. In *Biomineralization Process of Iron and Manganese: Modern and Ancient Environments*; Catena Supplement 21; Skinner, H.C.W., Fitzpatrick, R.W., Eds.; Catena: Destedt. Germany, 1992; 133–146.
39. Kirschvink, J.L.; Walker, M.M.; Diebel, C.E. Magnetitebased magnetoreception. *Curr. Opin. Neurobiol.* **2001**, *11* (4), 462–467.
40. Guthrie, G.D.; Mossman, B.T. *Health Effects of Mineral Dusts: Reviews in Mineralogy*; Mineralogical Society of America: Chelsca, MI, 1993; Vol. 28.

41. Navrotsky, A. Nanomaterials in the environment, agriculture, and technology (NEAT). *J. Nanopart. Res.* **2000**, *2*, 321–323.
42. Pitcher, M.; Navrotsky, A. unpublished data.

Bibliography

Nakashima, S.; Ikoma, M.; Shiota, D.; Nakazawa, K.; Maruyama, S. Geochemistry and the origin and evolution of life: A tentative summary and future perspectives. *Precursors Chall. Investig. Ser.* **2001**, *2*, 329–344.

6

Pharmaceuticals: Treatment

Diana Aga and
Seungyun Baik

Introduction

The presence of pharmaceutical chemicals and their byproducts in soil, wastewater effluents, surface water, and drinking water sources has become a growing concern over the past two decades. Improvements in analytical methods coupled with large-scale surveys have revealed the broad range of persistent pharmaceuticals that are cycling through our wastewater-to-drinking water cycle. Non-metabolized active ingredients and transformation products of veterinary and human pharmaceuticals are introduced into the environment through the effluents of municipal wastewater treatment plants (WWTPs), pharmaceutical formulation facilities, and through the land application of animal waste and sewage sludge. Approximately 10 million tons of sewage sludge and manure containing residues of pharmaceuticals are used to fertilize croplands each year. Public awareness of the potential problems related to pharmaceutical pollution, widely known as "emerging contaminants," has brought this issue to the forefront in the water and wastewater treatment industries. Hence, advanced water treatment systems are being evaluated to potentially eliminate these emerging contaminants from effluents of WWTPs and from drinking water sources. This entry aims to provide an overview on the occurrence of pharmaceuticals in the environment and recent studies that investigate promising treatment technologies to eliminate emerging contaminants from wastewater and drinking water systems.

Although the human risk associated with chronic exposure to pharmaceuticals in the environment remains unclear, evidence of detrimental ecological impacts is growing. The environmental contamination by pharmaceutical residues, especially antibiotics, may have profound environmental effects at several levels. While the promotion of antibiotic resistance in pathogenic microorganisms has been the major concern associated with the presence of antibiotics in the environment, other issues such as endocrine disruption in fish and wildlife, plant uptake, and phytotoxicity are also significant and warrant discussion. Therefore, the second goal of this entry is to summarize current knowledge on the ecological impacts of pharmaceutical pollution.

Occurrence and Impacts of Pharmaceuticals in the Environment

The presence of pharmaceutical residues in terrestrial and aquatic systems, resulting largely from discharges of municipal WWTPs and the land application of animal wastes, is now well documented in the literature.[1–3] As depicted in Figure 1, residues of human pharmaceuticals and their metabolites may eventually enter surface water, groundwater, and drinking water systems after passing through WWTPs. While most active ingredients of drugs are metabolized in the body, or removed during wastewater treatment, others remain intact and persist in the environment. Low levels of persistent pharmaceuticals can eventually end up in finished drinking water and distribution system (tap) water, when using source waters that have been affected by effluents from WWTPs.[4]

Veterinary pharmaceuticals, particularly antibiotics used for therapeutic purposes and for growth promotion, are also finding their ways into the environment. While some antibiotics used in animal production are decomposed quickly after being excreted, others remain stable during manure storage and end up in agricultural fields upon manure application. Additionally, antibiotics are widely used in fish farms and may enter the aquatic environment via direct discharge. For example, the antibiotic oxytetracycline has been detected in groundwater that has been affected by fish farming at sub-parts-per-billion concentrations.[5]Highly polar pharmaceuticals are susceptible to leaching and may therefore reach the groundwater aquifer. For example, the high frequency of detection of sulfonamide antibiotics in groundwater from various sites in the United States can be attributed to the relatively high water solubility and poor biodegradability of these drugs.[6,7] The detection of sulfonamides in groundwater from wells deeper than 50 ft, located downgradient from an animal feeding operation in the United States, indicates the persistence and mobility of this class of antibiotics.[8] Similar results were observed in Germany where sulfonamides were detected in groundwater located downgradient from an agricultural field where sewage sludge was used for irrigation.[9] The concentrations of pharmaceuticals in groundwater are typically lower than in surface water because infiltration through the soil profile removes some

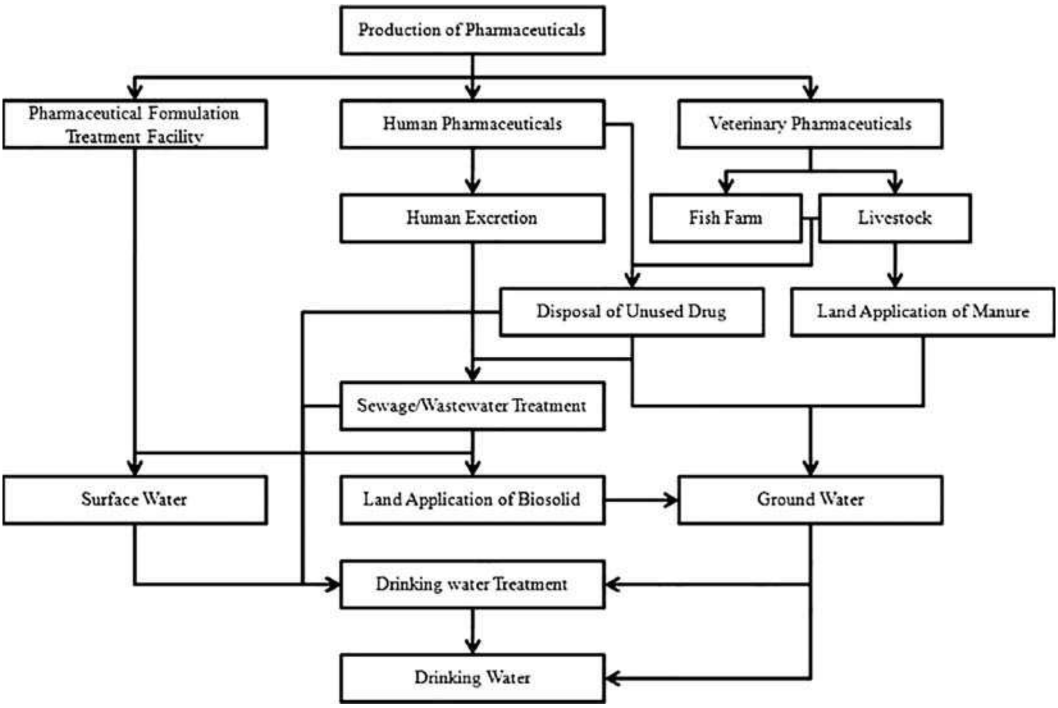


FIGURE 1 Exposure routes of pharmaceuticals in the environment.

fraction of the organic pollutants before entering groundwater systems.[10] Nevertheless, contamination of groundwater by pharmaceuticals is a significant issue because these compounds are not readily biodegradable under anoxic conditions.[9] The persistence of pharmaceuticals in the groundwater aquifer is a concern because groundwater is a major source of drinking water in many areas in the United States and around the world.[8] If the contaminated groundwater is used as drinking water source, pharmaceuticals may eventually reach finished drinking water systems[4] and may not be degraded during conventional drinking water treatment processes.

In the first comprehensive study on the occurrence of pharmaceuticals in the United States surface waters, Kol-pin et al.[1] reported detection of 95 organic contaminants, which included 30 antibiotics, 12 prescribed drugs, 4 nonprescribed drugs, and 6 drug metabolites. A later report summarized the occurrence of 80 pharmaceuticals and drug metabolites in the aquatic systems for eight different countries.[11] The concentrations of pharmaceuticals in surface water and groundwater depend on several factors, such as removal processes employed in the WWTPs, source variability, dilution, retardation, and weather events.[12]

Table 1 shows examples of the types of pharmaceuticals frequently detected in the environment and their typical removal rates (high, medium, low) in conventional activated sludge (CAS) systems during wastewater treatment.[13,14] Consequently, these pharmaceuticals are often detected in receiving surface waters at a wide range of concentrations as depicted in Figure 2. The variability in pharmaceutical concentrations in

TABLE 1 Classification of Pharmaceuticals Based on Their Typical Removal Rates in CAS Systems

Group	Pharmaceutical	Usage
High removal (>65%)	Acetaminophen (ACE)	Analgesic (non-NSAID)
	Ibuprofen (IBP)	Analgesic (NSAID)
	Naproxen (NAP)	Analgesic (NSAID)
	Paroxetine (PRX)	Antidepressant
	Iopamidol (IOM)	Iodinated contrast agent
	Caffeine (CAF)	Psychoactive stimulant
Medium removal (30%–65%)	Sulfamethoxazole (SMX)	Antibiotic
	Gemfibrozil (GFB)	Lipid regulator
	Atenolol (ATN)	β-Blocker
	Propranolol (PRN)	β-Blocker
	Ranitidine (RTD)	Antihistamine (Zantac)
	Fluoxetene (FXT)	Antidepressant
	Iopromide (IOP)	Iodinated contrast agent
Low removal (<30%)	Diclofenac (DCF)	Analgesic (NSAID)
	Mefenamic acid (MFN)	Analgesic (NSAID)
	Ciprofloxacin (CIP)	Antibiotic
	Erythromycin (ERY)	Antibiotic—macrolide
	Roxythromycin (ROX)	Antibiotic—macrolide
	Trimethoprim (TMP)	Antibiotic
	Clofibric acid (CLO)	Lipid regulator
	Carbamazepine (CBZ)	Anticonvulsant
	Dilantin (DLT)	Anticonvulsant
	Meprobamate (MPB)	Anti-anxiety drug

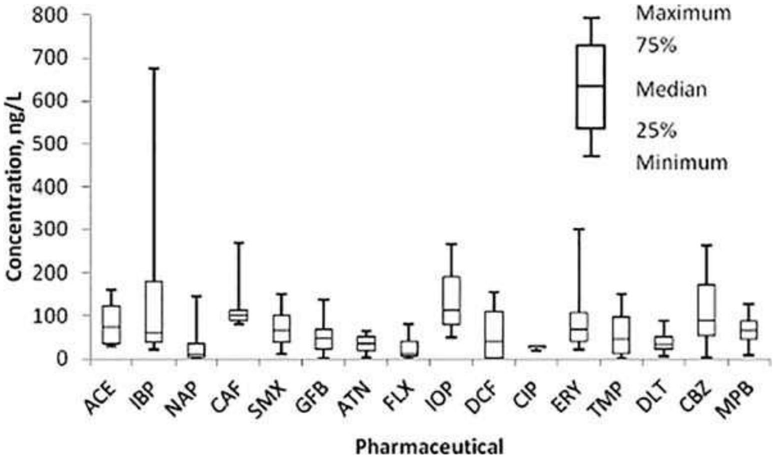


FIGURE 2 Typical concentrations of selected pharmaceuticals in surface waters in the U.S. The middle line in the box plot indicates median value of the data; The upper edge (hinge) of the box indicates 75th percentile of the data set; the lower hinge indicates 25th percentile. The whiskers indicate minimum and maximum data values.
Source: Kolpin, D.W. et al.[1]; Benotti, M.J. et al.[4]; Snyder, S.A. et al.[15]; Westerhoff, P. et al.[16]; Focazio, M.J. et al.[17]; Nodler, K. et al.[18]; Boyd, G.R. et al.[19]; Ferrer, I. et al.[20].

the environment can be attributed to the types of WWTPs, the frequency and time of sampling, and the biodegradability of the pharmaceuticals in a given environmental conditions, among others.

Despite the relatively low concentrations of pharmaceuticals typically found in the environment, the ecological effects of pharmaceutical pollution cannot be ignored. Several standard toxicity tests reveal that while acute toxicity is not a concern because of the high effective concentrations needed to elicit observable acute effects, chronic toxicity may be important at environmentally relevant concentrations.

Acute and chronic toxicity of several pharmaceuticals, including non-steroidal anti-inflammatory drugs (NSAIDs) and β-blockers, against phytoplankton, zooplankton, and other aquatic organisms[21–23] have been reported. Similarly, chronic toxicity studies toward aquatic organisms of lipid-lowering agents,[24,25] neuroactive pharmaceuticals (antidepressants),[22,26] and anti-epileptic drugs[22] have been conducted at concentrations typically found in the environment. Some of the documented ecological effects of pharmaceuticalsin the environment are summarized in Table 2. Both the U.S. Environmental Protection Agency (USEPA) and European Union (EU) have recognized residues of pharmaceuticals and personal care products as emerging contaminants of concern that may require future environmental regulation if their persistence in the environment proves to be ecologically significant. In fact, diclofenac, ibuprofen, triclosan, and clofibric acid have been identified as future emerging priority candidates by the EU water framework directive, which is a priority substance list that is updated every 4 years.[27]

TABLE 2 Reported Ecotoxicity for Selected Pharmaceuticals Using Various Test Organisms

Pharmaceutical	Effect Observed	Effective Concentration	Environmentally Relevant[a]	References
Atenolol	Fathead minnow	1.0 mg/L (21 days)—NOEC	No	[28]
		3.2 mg/L (21 days)—LOEC	No	
Carbamazepine	*Daphnia magna*	0.492 μg/L—multigeneration (up to 6 generations) effect	Yes	[29]

(*Continued*)

TABLE 2 (*Continued*) Reported Ecotoxicity for Selected Pharmaceuticals Using Various Test Organisms

Pharmaceutical	Effect Observed	Effective Concentration	Environmentally Relevant[a]	References
	Japanese medaka	6.15 mg/L—swim speed effect (9 days)	No	[30]
	Oryzias latipes	45.87 mg/L—LC50 (96 hours)	No	[31]
	Thamnocephalus platyurus	>100 mg/L—LC50 (24 hours)	No	[31]
Clarithromycin	*Oryzias latipes*	>100 mg/L—LC50 (96 hours)	No	[31]
	Thamnocephalus platyurus	94.23 mg/L—LC50 (24 hours)	No	[31]
Diclofenac	*Daphnia magna*	0.355 µg/L—multigeneration (up to 6 generations) effect	Yes	[29]
	Hyalella azteca	0.467 mg/kg—LC50 (72 hours) with sediment	No	[32]
	Rainbow trout	25 µg/L—accumulation test on gall (21 days)	Yes	[33]
Fluoxetene	Sheepshead minnow	>2.0 mg/L—LC50 2.0 mg/L—LOEC 1.87 mg/L—NOEC (All for 96 hours)	No	[34]
	Western mosquitofish	0.546 mg/L—LC50 (7 days)	No	[35]
	Western mosquitofish	0.5 µg/L—increasing lethargy in 59–159 days	Yes	[35]
Ibuprofen	*Oryzias latipes*	>100 mg/L—LC50 (96 hours)	No	[31]
	Planorbis carinatus	17.1 mg/L—LC50 (48 and 72 hours)	No	[36]
	Planorbis carinatus	>5.36 mg/L—LOEC 5.36 mg/L—NOEC (All for 21 days)	No	[36]
	Thamnocephalus platyurus	19.59 mg/L—LC50 (24 hours)	No	[31]
Metoprolol	*Daphnia magna*	1.170 µg/L—multigeneration (up to 6 generations) effect	Yes	[29]
Paracetamol (Acetaminophen)	Wheat	22.4 mg/L—damage in 21 days exposure	No	[37]
Propranolol	Fathead minnow	1.0 mg/L—NOEC (female) 3.4 mg/L—LOEC (female) 0.1 mg/L—NOEC (male) 1.0 mg/L—LOEC (male) (All for 21 days)	No (Yes for NOEC male)	[38]
	Rainbow trout	1.0 mg/L—NOEC 10 mg/L—LOEC (All for 10 days)	No	[39]
	Oryzias latipes	11.40 mg/L—LC50 (96 hours)	No	[31]
	Thamnocephalus platyurus	10.31 mg/L—LC50 (24 hours)	No	[31]
Verapamil	Juvenile rainbow trout	2.72 mg/L—LC50 (96 hours)	No	[40]

Note: NOEC, no observed effect concentration; LOEC, lowest observed effect concentration; LC50, 50% lethal concentration.

[a] Environmentally relevant—"Yes" for concentrations of compounds are in the ranges in surface water and WWTP effluent reported through 2011.

Treatment of Pharmaceuticals in the Aquatic Systems

Alternative Systems for Water Treatment

Activated sludge treatment, which relies on microbes to biodegrade contaminants in wastewater, is the most widely used waste water treatment system in the United States and around the world. While many pharmaceuticals are removed partially or completely during CAS treatment, there are a significant number of pharmaceuticals that have very little to no removal. Therefore, alternative treatment systems are being explored to improve removal efficiencies for these trace organic contaminants. Examples of treatment systems that are suspected to be more efficient are membrane bioreactors (MBRs)[41–44]; membrane treatment with nano-, micro-, and ultrafiltration (NF/MF/UF); and reverse osmosis (RO).[45] Granular activated carbon (GAC) is also used as another alternative system in the Unites States and Canada for removal of trace organic matters through filtration and adsorption.[45] Recent studies have also shown that advanced oxidation processes, such as ozonation and UV–H_2O_2 disinfection, which are employed during tertiary treatment before discharging the treated wastewater, could have additional benefits in removing pharmaceuticals from water.[15,46,47]

For drinking water treatment systems, the removal of many pharmaceuticals during the first three steps (coagulation, flocculation, and sand filtration) before disinfection is typically incomplete.[16,48] On the other hand, similar to wastewater treatment, pharmaceuticals may be effectively removed by activated carbon adsorption, ozone oxidation, and membrane filtration.[49] Chlorination, the most common disinfection step, may also remove pharmaceuticals but produce unwanted by-products.[16]

Alternative water treatment systems are summarized and compared in Table 3.[50,51] These systems are advantageous because they offer increased removal of pharmaceuticals and other organic contaminants that cause undesirable odor and taste in water. To enhance the efficiency of treatment, combining systems such as MBRs followed by activated carbon adsorption system maybe applied.

TABLE 3 Summary and Comparison of Alternative Systems: Membrane Process, Activated Carbon Adsorption Process, and Advanced Oxidation Process for Contaminant Removal

Process		Operation and Application	Important Notes
Membrane process	Reverse osmosis (RO)	• Pressure driven (150–400 psi)	• Best removal performance • Possible for desalination
		• Removes organics with molecular weight >100	• Needs pretreatment such as UF or MF to prevent plugging and fouling; hence, high cost
		• Demineralization	• Low product recovery (30%–85%)
	Nanofiltration (NF)	• Pressure driven (80–200 psi)	• Good removal performance
		• Removes organics with molecular weight between 100 and 500	• Needs pretreatment such as UF or MF to prevent plugging and fouling; hence, high cost
		• Removes NOM	• Medium product recovery (70%–90%)
	Ultrafiltration (UF)	• Pressure driven (15–60 psi)	• Replacement for CT • High product recovery (80%–95%) • Higher cost than CT
		• Pretreatment for NF or RO	• Sensitive with water temperature and viscosity

(*Continued*)

TABLE 3 (*Continued*) Summary and Comparison of Alternative Systems: Membrane Process, Activated Carbon Adsorption Process, and Advanced Oxidation Process for Contaminant Removal

Process		Operation and Application	Important Notes
	Microfiltration (MF)	• Macromolecule removal includes viruses	• Needs other process to increase treatment performance
		• Pressure driven (5–40 psi)	• Replacement for CT
			• High product recovery (95%–98%)
			• Higher cost than CT
		• Pretreatment for NF or RO	• Sensitive with water temperature and viscosity
		• Bacteria removal	• Needs other process to increase treatment performance
Activated carbon adsorption	Powder activated carbon (PAC)	• Used in early steps of treatment process	• Good performance in combination with MF or UF membrane systems
			• Usually used for drinking water treatment
			• Good for emergency situations, with high organic pollutants
			• Handling problem—dry PAC causes dust problem
		• Controls taste or odor	• May pass through filters and enter the final treated water
		• Applied for small or mid-sized plants with moderate or severe taste, odor, or organic contaminants problem	• Cannot be mixed with chlorine
	Granular activated carbon (GAC)	• Used in adsorption beds or tanks	• Replacement of conventional gravity filter
			• Possible for WWTPs
			• Fouling by chemicals—need to consider breakthrough for design
			• Backwash requirement
		• Gravity or pressure driven	• Need to monitor carbon bed depth
		• Applied for plants with moderate or severe taste, odor, or organic contaminants problem	• Needs carbon regeneration
Advanced oxidation process	Ultraviolet (UV)	• Chemical treatment	• Disinfection benefit
		• Low pressure ($\lambda = 253.7$ nm) or medium pressure ($\lambda = 180–1370$ nm)	• Controls taste and odor
		• UV radiation with H_2O_2 addition (hydroxyl radical oxidation)	• Relatively high cost and complexity
		• Combined with ozonation	• Treatment effectiveness is weakened by turbidity or color in water
		• NOM in water may inhibit or promote oxidation	• Unwanted by-products

Note: NOM, natural organic matter; CT, conventional treatment.

Removal Efficiencies in Municipal WWTP Systems

Several reports have shown that MBRs generally have higher removal efficiencies for NSAIDs than CAS systems. For example, it was demonstrated that up to 98%–100% ketiprofen and 86%–89% naproxen were removed in MBRs.[41,44] In addition, MBRs also have better removal efficiencies for the antibiotics roxythromycin, sulfamethoxazole, trimethoprim, and diclofenac relative to the CAS systems.[43,44] However, certain compounds, such as carbamazepine, remain mostly undegraded in both conventional WWTP and MBRs.[43] A study compared different filter systems such as MBR, MF/UF, and RO in removing pharmaceuticals using WWTP primary effluent as the feed water.[45] In contrast to other studies, most pharmaceuticals were removed well (except for phenytoin and meprobamate) in the MBR, including a 90% removal of carbamazepine. These differences in results suggest that design and operation of MBRs can be optimized to increase removal efficiencies of pharmaceuticals in wastewater. Additionally, it was demonstrated that more than 90% of all pharmaceuticals tested were removed in the RO system alone; however, the use of combined systems such as UF/RO, MF/RO, and MBR/RO resulted in more than 99% removal for all pharmaceuticals. However, membrane systems are expensive and may not be affordable to many municipalities. In addition, membrane systems require that the brine, in which the rejected pharmaceuticals are concentrated after removing pharmaceuticals from mother water, must be appropriately treated.

Granular activated carbon (GAC) adsorption systems were also investigated in two facilities, one with regular regeneration of GAC and the other without regular replacement/regeneration of GAC.[45] As summarized in Table 4, the latter facility had poor removal efficiencies for most of the pharmaceuticals tested. Hence, it is important to determine the breakthrough for individual pharmaceuticals so that regeneration or replacement GAC can be scheduled on a timely fashion to be most effective in removing pharmaceutical residues in the water.

Ozone oxidation has been investigated for sulfonamides, macrolides, and iodinated contrast media, as well as other acidic pharmaceuticals spiked in MBR-treated wastewater effluent.[46] As may be expected, increased ozone concentrations resulted in increased degradation of pharmaceuticals. However, while the sulfonamides and macrolides were very sensitive toward ozone degradation and were completely removed with high ozone concentration, iodinated contrast media were only 40% degraded. Ozone oxidation experiments have been performed to investigate the removal of 17 pharmaceuticals, as well as other personal care products and endocrine-disrupting chemicals, in bench-top pilot plants and in one full-scale WWTP system. The removal efficiencies of pharmaceuticals by ozonation can be classified into four groups: >80% removal, 80%–50% removal, 50%–10% removal, and <20% removal.[15] Most of the pharmaceuticals exhibited >80% removal; diazepam, phenytoin, and ibuprofen had 80%–50% removal; and iopromide and meprobamate were in the group of 50%–10% removal. In a separate

TABLE 4 Comparing Removal Efficiencies of Pharmaceuticals for GAC Facilities with and without Regular Regeneration or Replacement of Activated Carbon

Pharmaceuticals	Removal (%) in Facility 1 (with Regular GAC Regeneration)	Removal (%) in Facility 2 (without Regular GAC Regeneration)
Caffeine	>41.2	16.3
Carbamazepine	>54.5	15.6
Phenytoin (Dilantin)	>44.4	22.7
Erythromycin–H_2O	>44.4	7.9
Gemfibrozil	>16.7	8.2
Ibuprofen	>9.1	16.4
Iopromide	>69.7	72.0
Meprobamate	>16.7	13.3
Sulfamethoxazole	>83.3	83.8

study using two different wastewater matrices, one with effluent water from a conventional WWTP and the other with effluent from an MBR system, it was shown that the removal efficiencies by ozonation, ozone–UV, and H_2O_2–UV may be slightly lower in the presence of higher dissolved organic carbon.[47]

The removal of 13 selected pharmaceuticals in full-scale WWTP with CAS system, followed by advanced treatment by ozonation, was evaluated.[52] Interestingly, most of the pharmaceuticals that are typically poorly removed (<20% removal) in the secondary clarifier, such as crotamiton, sulfapyridine, and roxithromycin, were almost completely-degraded upon ozonation. An exception to the pharmaceuticals effectively removed by ozonation was carbamazepine (which was only <5% degraded by ozonation).

In a separate study, the efficiency of UV oxidation was examined with and without H_2O_2 for atenolol, carbamazepine, phenytoin, meprobamate, primidone, and trimethoprim, using three different wastewater matrices.[53] Results of this study indicated that the nature of the aqueous matrix, most likely defined by the amount and composition of the natural organic matter, is an important factor in optimizing the removal efficiencies of pharmaceuticals by advanced oxidation. For example, removal efficiencies between 16% and 95% were observed for the pharmaceuticals tested in one WWTP using a high power of UV (fluence of 700 mJ/cm^2) with H_2O_2. However, removal efficiencies of the same pharmaceuticals in water from other sampling locations with high organic matter content were relatively lower (10%–85%), although the same UV treatment conditions were used.

Removal Efficiencies in Drinking Water Treatment Systems

Polar and persistent pharmaceuticals may eventually enter the drinking water systems.[49] Since conventional processes used in drinking water treatment may not completely remove pharmaceuticals, other advanced oxidation processes have been considered for treatment of pharmaceuticals in drinking water. Ozonation appears to be the most effective way to disinfect the water and at the same time oxidize organic chemicals via direct reaction with ozone, or through reactions with hydroxyl radicals (OH) formed during ozonation.[15] For disinfection purposes, monochromatic UV at 254 nm is used for drinking water treatment. In UV advanced oxidation process, the oxidation of pharmaceuticals is enhanced by addition of H_2O_2 to facilitate formation of hydroxyl radicals and promote indirect photolysis.[54] The oxidation of diclofenac, a frequently detected NSAID in the environment, was evaluated using ozone, UV, and UV–H_2O_2. Ozonation showed 100% removal, while UV–H_2O_2 oxidation showed only 52% removal of diclofenac at conditions that corresponded to a 35% decrease in total organic carbon, after 90 minutes of oxidation time.[55] The antidepressant pharmaceutical fluoxetene (trade name of Prozac), which is typically detected at parts per billion levels in U.S. streams,[1] was used as test compound to evaluate the efficiency of direct and indirect photolysis in removing pharmaceuticals in water. The degradation rate for fluoxetene reached up to 9.60×10^9/M.s for indirect photolysis.[561] It has been shown that the use of TiO2 catalyst can significantly increase the photolysis rate constants for NSAIDs, including diclofenac, naproxen, and ibuprofen in water.[57] Experimental oxidation rates of the hydroxyl radical for β-blockers, atenolol, metoprolol, and propranolol, were found to be 7.05×10^9, 8.39×10^9, and 1.07×10^9/M/s, respectively.[58] Ozonation of tetracycline showed the complete removal within 4–6 minutes of ozonation despite the maximum degradation of 40% total organic carbon after 2 hours of ozonation.[59] Finally, photo-Fenton (Fe(II)) reaction is another oxidation method that can be applicable to drinking water treatment systems owing to its cost effectiveness and ease of operation.[60] It was demonstrated that diclofenac can be 100% degraded within 60 minutes of photo-Fenton oxidation.[60]

Organic pollutants in drinking water sources, including pharmaceuticals, can be removed by activated carbon adsorption.[16] Laboratory-scale tests mimicking full-scale drinking water treatment system using activated carbon-showed good removal efficiencies of trace organic compounds, except for the more hydrophilic pharmaceuticals (e.g., clofibric acid and ibuprofen were only 40% removed) after 3 hours of contact time.[61]

Water filters can provide an additional step, at a point-of-use consumption treatment, to remove trace levels of pharmaceuticals and other drinking water contaminants from tap water. The removal efficiencies of chlorination by-products, including trihalomethanes, bromodichloromethane, dibromochloromethane, and bromoform, by "Envirofilter" (made of nutshell carbon and commercial water filters) have been reported.[62] Water filters are made with activated carbon to remove dissolved organic matter,[62,63] together with polar and ion-exchange resins to remove charged species such as metal ions.[64,65] Since activated carbon has been shown to have high removal efficiencies for many pharmaceuticals in wastewater,[16,66–71] it is reasonable to expect that water filters will be effective in removing trace levels of pharmaceuticals from drinking water.

Conclusion

The decreasing amount of clean water resources for drinking water and for food production has become one of the most challenging problems in the world.[72] To alleviate the shortage in water supply, reuse of treated wastewater for irrigation[72–74] and as drinking water source[75] is becoming more and more common. Therefore, pharmaceuticals can enter the groundwater and drinking water systems, as shown in Figure 1 depicting the routes of entry of pharmaceuticals into the environment. It is encouraging to see a recent survey reporting that only very few pharmaceuticals from biosolid application and wastewater irrigation can be transported to the groundwater aquifer.[72] Because only lipophilic pharmaceuticals tend to sorb onto the biosolids, and these pharmaceuticals in turn are not mobile in soil, contamination of groundwater aquifer from biosolids application may not pose a significant source of pharmaceutical pollution. With the exception of four pharmaceuticals (two iodinated contrast media of diatrizoate and iopamidol, carbamazepine and sulfamethoxazole) most of the 52 pharmaceuticals being targeted for analysis were not detected in the said study. Nevertheless, optimization of the design and operating conditions of wastewater treatment systems is key to eliminating pharmaceuticals and their deleterious ecological effects in the environment. Advanced oxidation processes are very promising treatment technologies that are waiting to be tested and implemented under full-scale treatment plants.

References

1. Kolpin, D.W.; Furlong, E.T.; Meyer, M.T.; Thurman, E.M.; Zaugg, S.D.; Barber, L.B.; Buxton, H.T. Pharmaceuticals, hormones, and other organic wastewater contaminants in US streams, 1999–2000: A national reconnaissance. *Environmental Science & Technology* **2002**, *36* (6), 1202–1211.
2. Aga, D.S.; O'Connor, S.; Ensley, S.; Payero, J.O.; Snow, D.; Tarkalson, D. Determination of the persistence of tetracycline antibiotics and their degradates in manure-amended soil using enzyme-linked immunosorbent assay and liquid chromatography-mass spectrometry. *Journal of Agricultural and Food Chemistry* **2005**, *53* (18), 7165–7171.
3. Batt, A.L.; Bruce, I.B.; Aga, D.S. Evaluating the vulnerability of surface waters to antibiotic contamination from varying wastewater treatment plant discharges. *Environmental Pollution* **2006**, *142* (2), 295–302.
4. Benotti, M.J.; Trenholm, R.A.; Vanderford, B.J.; Holady, J.C.; Stanford, B.D.; Snyder, S.A. Pharmaceuticals and endocrine disrupting compounds in US drinking water. *Environmental Science & Technology* **2009**, *43* (3), 597–603.
5. Avisar, D.; Levin, G.; Gozlan, I. The processes affecting oxytetracycline contamination of groundwater in a phreatic aquifer underlying industrial fish ponds in Israel. *Environmental Earth Sciences* **2009**, *59* (4), 939–945.
6. Lindsey, M.E.; Meyer, M.; Thurman, E.M. Analysis of trace levels of sulfonamide and tetracycline antimicrobials, in groundwater and surface water using solid-phase extraction and liquid chromatography/mass spectrometry. *Analytical Chemistry* **2001**, *73* (19), 4640–4646.
7. Karthikeyan, K.G.; Meyer, M.T. Occurrence of antibiotics in wastewater treatment facilities in Wisconsin, USA. *Science of the Total Environment* 2006, *361* (1–3), 196–207.

8. Batt, A.L.; Snow, D.D.; Aga, D.S. Occurrence of sulfonamide antimicrobials in private water wells in Washington County, Idaho, USA. *Chemosphere* **2006**, *64* (11), 1963–1971.
9. Richter, D.; Massmann, G.; Taute, T.; Duennbier, U. Investigation of the fate of sulfonamides downgradient of a decommissioned sewage farm near Berlin, Germany. *Journal of Contaminant Hydrology* **2009**, *106* (3–4), 183–194.
10. Katz, B.G.; Griffin, D.W.; Davis, J.H. Groundwater quality impacts from the land application of treated municipal wastewater in a large karstic spring basin: Chemical and microbiological indicators. *Science of the Total Environment* **2009**, *407* (8), 2872–2886.
11. Heberer, T. Occurrence, fate, and removal of pharmaceutical residues in the aquatic environment: A review of recent research data. *Toxicology Letters* **2002**, *131* (1–2), 5–17.
12. Musolff, A.; Leschik, S.; Moder, M.; Strauch, G.; Reinstorf, F.; Schirmer, M. Temporal and spatial patterns of micropollutants in urban receiving waters. *Environmental Pollution* **2009**, *157* (11), 3069–3077.
13. Jones, O.A.H.; Voulvoulis, N.; Lester, J.N. Human pharmaceuticals in wastewater treatment processes. *Critical Reviews in Environmental Science and Technology* **2005**, *35* (4), 401–427.
14. Sipma, J.; Osuna, B.; Collado, N.; Monclús, H.; Ferrero, G.; Comas, J.; Rodriguez-Roda, I. Comparison of removal of pharmaceuticals in MBR and activated sludge systems. *Desalination* **2010**, *250* (2), 653–659.
15. Snyder, S.A.; Wert, E.C.; Rexing, D.J.; Zegers, R.E.; Drury, D.D. Ozone oxidation of endocrine disruptors and pharmaceuticals in surface water and wastewater. *Ozone-Science & Engineering* **2006**, *28* (6), 445–460.
16. Westerhoff, P.; Yoon, Y.; Snyder, S.; Wert, E. Fate of en-docrine-disruptor, pharmaceutical, and personal care product chemicals during simulated drinking water treatment processes. *Environmental Science & Technology* **2005**, *39* (17), 6649–6663.
17. Focazio, M.J.; Kolpin, D.W.; Barnes, K.K.; Furlong, E.T.; Meyer, M.T.; Zaugg, S.D.; Barber, L.B.; Thurman, M.E. A national reconnaissance for pharmaceuticals and other organic wastewater contaminants in the United States-II) Untreated drinking water sources. *Science of the Total Environment* **2008**, *402* (2–3), 201–216.
18. Nodier, K.; Licha, T.; Bester, K.; Sauter, M. Development of a multi-residue analytical method, based on liquid chromatography-tandem mass spectrometry, for the simultaneous determination of 46 micro-contaminants in aqueous samples. *Journal of Chromatography A* **2010**, *1217* (42), 6511–6521.
19. Boyd, G.R.; Palmeri, J.M.; Zhang, S.Y.; Grimm, D.A. Pharmaceuticals and personal care products (PPCPs) and endocrine disrupting chemicals (EDCs) in stormwater canals and Bayou St. John in New Orleans, Louisiana, USA. *Science of the Total Environment* 2004, *333* (1–3), 137–148.
20. Ferrer, I.; Zweigenbaum, J.A.; Thurman, E.M. Analysis of 70 Environmental Protection Agency priority pharmaceuticals in water by EPA Method 1694. *Journal of Chromatography A* **2010**, *1217* (36), 5674–5686.
21. Huggett, D.B.; Brooks, B.W.; Peterson, B.; Foran, C.M.; Schlenk, D. Toxicity of select beta adrenergic receptor-blocking pharmaceuticals (β-blockers) on aquatic organisms. *Archives of Environmental Contamination and Toxicology* **2002**, *43* (2), 229–235.
22. Ferrari, B.; Paxeus, N.; Lo Giudice, R.; Pollio, A.; Garric, J. Ecotoxicological impact of pharmaceuticals found in treated wastewaters: Study of carbamazepine, clofibric acid, and diclofenac. *Ecotoxicology and Environmental Safety* **2003**, *55* (3), 359–370.
23. Ferrari, B.; Mons, R.; Vollat, B.; Fraysse, B.; Paxeus, N.; Lo Giudice, R.; Pollio, A.; Garric, J. Environmental risk assessment of six human pharmaceuticals: Are the current environmental risk assessment procedures sufficient for the protection of the aquatic environment? *Environmental Toxicology and Chemistry* **2004**, *23* (5), 1344–1354.
24. Donohue, M.; Baldwin, L.A.; Leonard, D.A.; Kostecki, P.T.; Calabrese, E.J. Effect of hypolipidemic drugs gemfibrozil, ciprofibrozil, ciprofibrate, and clofibric acid of peroxisomal beta-oxidation in primary cultures of rainbow-trout hepatocytes. *Ecotoxicology and Environmental Safety* **1993**, *26* (2), 127–132.

25. Nunes, B.; Carvalho, F.; Guilhermino, L. Acute and chronic effects of clofibrate and clofibric acid on the enzymes acetylcholinesterase, lactate dehydrogenase and catalase of the mosquitofish, *Gambusia holbrooki*. *Chemosphere* **2004**, *57* (11), 1581–1589.
26. Brooks, B.W.; Foran, C.M.; Richards, S.M.; Weston, J.; Turner, P.K.; Stanley, J.K.; Solomon, K.R.; Slattery, M.; La Point, T.W. Aquatic ecotoxicology of fluoxetine. *Toxicology Letters* **2003**, *142* (3), 169–183.
27. Ellis, J.B. Pharmaceutical and personal care products (PPCPs) in urban receiving waters. *Environmental Pollution* **2006**, *144* (1), 184–189.
28. Winter, M.J.; Lillicrap, A.D.; Caunter, J.E.; Schaffner, C.; Alder, A.C.; Ramil, M.; Ternes, T.A.; Giltrow, E.; Sumpter, J.P.; Hutchinson, T.H. Defining the chronic impacts of atenolol on embryo-larval development and reproduction in the fathead minnow (*Pimephales promelas*). *Aquatic Toxicology* **2008**, *86* (3), 361–369.
29. Dietrich, S.; Ploessl, F.; Bracher, F.; Laforsch, C. Single and combined toxicity of pharmaceuticals at environmentally relevant concentrations in Daphnia magna-A multigenerational study. *Chemosphere* **2010**, *79* (1), 60–66.
30. Nassef, M.; Matsumoto, S.; Seki, M.; Khalil, F.; Kang, I.J.; Shimasaki, Y.; Oshima, Y.; Honjo, T. Acute effects of triclosan, diclofenac and carbamazepine on feeding performance of Japanese medaka fish (*Oryzias latipes*). *Chemosphere* **2010**, *80* (9), 1095–1100.
31. Kim, J.W.; Ishibashi, H.; Yamauchi, R.; Ichikawa, N.; Takao, Y.; Hirano, M.; Koga, M.; Arizono, K. Acute toxicity of pharmaceutical and personal care products on freshwater crustacean (*Thamnocephalus platyurus*) and fish (*Oryzias latipes*). *Journal of Toxicological Sciences* **2009**, *34* (2), 227–232.
32. Oviedo-Gomez, D.G.C.; Galar-Martinez, M.; Garcia-Medina, S.; Razo-Estrada, C.; L.Gomez-Olivan, M. Diclofenac-enriched artificial sediment induces oxidative stress in Hyalella azteca. *Environmental Toxicology and Pharmacology* **2010**, *29* (1), 39–43.
33. Mehinto, A.C.; Hill, E.M.; Tyler, C.R. Uptake and biological effects of environmentally relevant concentrations of the nonsteroidal anti-inflammatory pharmaceutical Diclofenac in Rainbow Trout (*Oncorhynchus mykiss*). *Environmental Science & Technology* **2010**, *44* (6), 2176–2182.
34. Winder, V.L.; Sapozhnikova, Y.; Pennington, P.L.; Wirth, E.F. Effects of fluoxetine exposure on serotonin-related activity in the sheepshead minnow (*Cyprinodon variegatus*) using LC/MS/MS detection and quantitation. *Comparative Biochemistry and Physiology C-Toxicology & Pharmacology* **2009**, *149* (4), 559–565.
35. Henry, T.B.; Black, M.C. Acute and chronic toxicity of fluoxetine (selective serotonin reuptake inhibitor) in western mosquitofish. *Archives of Environmental Contamination and Toxicology* **2008**, *54* (2), 325–330.
36. Pounds, N.; Maclean, S.; Webley, M.; Pascoe, D.; Hutchinson, T. Acute and chronic effects of ibuprofen in the mollusc *Planorbis carinatus* (*Gastropoda: Planorbidae*). *Ecotoxicology and Environmental Safety* **2008**, *70* (1), 47–52.
37. An, J.; Zhou, Q.X.; Sun, F.H.; Zhang, L. Ecotoxicological effects of paracetamol on seed germination and seedling development of wheat (*Triticum aestivum* L.). *Journal of Hazardous Materials* **2009**, *169* (1–3), 751–757.
38. Giltrow, E.; Eccles, P.D.; Winter, M.J.; McCormack, P.J.; Rand-Weaver, M.; Hutchinson, T.H.; Sumpter, J.P. Chronic effects assessment and plasma concentrations of the betablocker propranolol in fathead minnows (*Pimephales promelas*). *Aquatic Toxicology* **2009**, *95* (3), 195–202.
39. Owen, S.F.; Huggett, D.B.; Hutchinson, T.H.; Hetheridge, M.J.; Kinter, L.B.; Ericson, J.F.; Sumpter, J.P. Uptake of propranolol, a cardiovascular pharmaceutical, from water into fish plasma and its effects on growth and organ biometry. *Aquatic Toxicology* **2009**, *93* (4), 217–224.
40. Li, Z.H.; Li, P.; Randak, T. Ecotoxocological effects of short-term exposure to a human pharmaceutical Verapamil in juvenile rainbow trout (*Oncorhynchus mykiss*). *Comparative Biochemistry and Physiology C-Toxicology & Pharmacology* **2010**, *152* (3), 385–391.

41. Kimura, K.; Hara, H.; Watanabe, Y. Removal of pharmaceutical compounds by submerged membrane bioreactors (MBRs). *Desalination* **2005**, *178* (1–3), 135–140.
42. Quintana, J.B.; Weiss, S.; Reemtsma, T. Pathway's and metabolites of microbial degradation of selected acidic pharmaceutical and their occurrence in municipal wastewater treated by a membrane bioreactor. *Water Research* **2005**, *39* (12), 2654–2664.
43. Celiz, M.D.; Perez, S.; Barcelo, D.; Aga, D.S. Trace Analysis of Polar Pharmaceuticals in Wastewater by LC-MS-MS: Comparison of Membrane Bioreactor and Activated Sludge Systems. *Journal of Chromatographic Science* **2009**, *47* (1), 19–25.
44. Tambosi, J.L.; de Sena, R.F.; Favier, M.; Gebhardt, W.; Jose, H.J.; Schroder, H.F.; Moreira, R. Removal of pharmaceutical compounds in membrane bioreactors (MBR) applying submerged membranes. *Desalination* **2010**, *261* (1–2), 148–156.
45. Snyder, S.A.; Adham, S.; Redding, A.M.; Cannon, F.S.; DeCarolis, J.; Oppenheimer, J.; Wert, E.C.; Yoon, Y. Role of membranes and activated carbon in the removal of endocrine disruptors and pharmaceuticals. *Desalination* **2007**, *202* (1–3), 156–181.
46. Huber, M.M.; Gobel, A.; Joss, A.; Hermann, N.; Loffler, D.; McArdell, C.S.; Ried, A.; Siegrist, H.; Ternes, T.A.; von Gunten, U. Oxidation of pharmaceuticals during ozonation of municipal wastewater effluents: A pilot study. *Environmental Science & Technology* **2005**, *39* (11), 4290–4299.
47. Gebhardt, W.; Schroder, H.F. Liquid chromatography-(tan-dem) mass spectrometry for the follow-up of the elimination of persistent pharmaceuticals during wastewater treatment applying biological wastewater treatment and advanced oxidation. *Journal of Chromatography A* 2007, 1160 (1–2), 34–43.
48. Vieno, N.; Tuhkanen, T.; Kronberg, L. Removal of pharmaceuticals in drinking water treatment: Effect of chemical coagulation. *Environmental Technology* **2006**, *27* (2), 183–192.
49. Vieno, N.M.; Harkki, H.; Tuhkanen, T.; Kronberg, L. Occurrence of pharmaceuticals in river water and their elimination a pilot-scale drinking water treatment plant. *Environmental Science & Technology* **2007**, *41* (14), 5077–5084.
50. HDR Engineering, ed. *Handbook of Public Water Systems*, 2nd Ed.; John Wiley and Sons: New York, 2001.
51. Baruth, E.E., ed. *Water Treatment Plant Design*, 4th Ed.; McGraw-Hill: New York, 2005.
52. Nakada, N.; Shinohara, H.; Murata, A.; Kiri, K.; Managaki, S.; Sato, N.; Takada, H. Removal of selected pharmaceuticals and personal care products (PPCPs) and endocrine-disrupting chemicals (EDCs) during sand filtration and ozonation at a municipal sewage treatment plant. *Water Research* **2007**, *41* (19), 4373–4382.
53. Rosario-Ortiz, F.L.; Wert, E.C.; Snyder, S.A. Evaluation of UV/H_2O_2 treatment for the oxidation of pharmaceuticals in wastewater. *Water Research* **2010**, *44* (5), 1440–1448.
54. Pereira, V.J.; Weinberg, H.S.; Linden, K.G.; Singer, P.C. UV degradation kinetics and modeling of pharmaceutical compounds in laboratory grade and surface water via direct and indirect photolysis at 254 nm. *Environmental Science & Technology* **2007**, *41* (5), 1682–1688.
55. Vogna, D.; Marotta, R.; Napolitano, A.; Andreozzi, R.; d'Ischia, M. Advanced oxidation of the pharmaceutical drug diclofenac with UV/H_2O_2 and ozone. *Water Research* **2004**, *38* (2), 414–422.
56. Lam, M.W.; Young, C.J.; Mabury, S.A. Aqueous photochemical reaction kinetics and transformations of fluoxetine. *Environmental Science & Technology* **2005**, *39* (2), 513–522.
57. Mendez-Arriaga, F.; Esplugas, S.; Gimenez, J. Photocatalytic degradation of non-steroidal anti-inflammatory drugs with TiO2 and simulated solar irradiation. *Water Research* **2008**, *42* (3), 585–594.
58. Song, W.H.; Cooper, W.J.; Mezyk, S.P.; Greaves, J.; Peake, B.M. Free radical destruction of beta-blockers in aqueous solution. *Environmental Science & Technology* **2008**, *42* (4), 1256–1261.
59. Khan, M.H.; Bae, H.; Jung, J.Y. Tetracycline degradation by ozonation in the aqueous phase: Proposed degradation intermediates and pathway. *Journal of Hazardous Materials* **2010**, *181* (1–3), 659–665.

60. Perez-Estrada, L.A.; Malato, S.; Gernjak, W.; Aguera, A.; Thurman, E.M.; Ferrer, I.; Fernandez-Alba, A.R. Photo-fenton degradation of diclofenac: Identification of main intermediates and degradation pathway. *Environmental Science & Technology* **2005**, *39* (21), 8300–8306.
61. Simazaki, D.; Fujiwara, J.; Manabe, S.; Matsuda, M.; Asami, M.; Kunikane, S. Removal of selected pharmaceuticals by chlorination, coagulation-sedimentation and powdered activated carbon treatment. *Water Science and Technology* **2008**, *58* (5), 1129–1135.
62. Ahmedna, M.; Marshall, W.E.; Husseiny, A.A.; Goktepe, L.; Rao, R.M. The use of nutshell carbons in drinking water filters for removal of chlorination by-products. *Journal of Chemical Technology and Biotechnology* **2004**, *79* (10), 1092–1097.
63. Humbert, H.; Gallard, H.; Suty, H.; Croue, J.P. Natural organic matter (NOM) and pesticides removal using a combination of ion exchange resin and powdered activated carbon (PAC). *Water Research* **2008**, *42* (6–7), 1635–1643.
64. An, H.K.; Park, B.Y.; Kim, D.S. Crab shell for the removal of heavy metals from aqueous solution. *Water Research* **2001**, *35* (15), 3551–3556.
65. Rengaraj, S.; Moon, S.H. Kinetics of adsorption of Co(II) removal from water and wastewater by ion exchange resins. *Water Research* **2002**, *36* (7), 1783–1793.
66. Ternes, T.A.; Meisenheimer, M.; McDowell, D.; Sacher, F.; Brauch, H.J.; Gulde, B.H.; Preuss, G.; Wilme, U.; Seibert, N. Z. Removal of pharmaceuticals during drinking water treatment. *Environmental Science & Technology* **2002**, *36* (17), 3855–3863.
67. Snyder, S.A.; Westerhoff, P.; Yoon, Y.; Sedlak, D.L. Pharmaceuticals, personal care products, and endocrine disruptors in water: Implications for the water industry. *Environmental Engineering Science* **2003**, *20* (5), 449–469.
68. Kim, S.D.; Cho, J.; Kim, I.S.; Vanderford, B.J.; Snyder, S.A. Occurrence and removal of pharmaceuticals and endocrine disruptors in South Korean surface, drinking, and waste waters. *Water Research* **2007**, *41* (5), 1013–1021.
69. Yu, Z.R.; Peldszus, S.; Huck, P.M. Adsorption characteristics of selected pharmaceuticals and an endocrine disrupting compound-Naproxen, carbamazepine and nonylphenol-on activated carbon. *Water Research* **2008**, *42* (12), 2873–2882.
70. Yu, Z.; Peldszus, S.; Huck, P.M. Adsorption of selected pharmaceuticals and an endocrine disrupting compound by granular activated carbon. 1. Adsorption Capacity and Kinetics. *Environmental Science & Technology* **2009**, *43* (5), 1467–1473.
71. Yu, Z.; Peldszus, S.; Huck, P.M. Adsorption of Selected Pharmaceuticals and an Endocrine Disrupting Compound by Granular Activated Carbon. *2. Model Prediction. Environmental Science & Technology* **2009**, *43* (5), 1474–1479.
72. Ternes, T.A.; Bonerz, M.; Herrmann, N.; Teiser, B.; Andersen, H.R. Irrigation of treated wastewater in Braunschweig, Germany: An option to remove pharmaceuticals and musk fragrances. *Chemosphere* **2007**, *66* (5), 894–904.
73. Chefetz, B.; Mualem, T.; Ben-Ari, J. Sorption and mobility of pharmaceutical compounds in soil irrigated with reclaimed wastewater. *Chemosphere* **2008**, *73* (8), 1335–1343.
74. Xu, J.; Chen, W.P.; Wu, L.S.; Green, R.; Chang, A.C. Leach-ability of some emerging contaminants in reclaimed municipal wastewater-irrigated turf grass fields. *Environmental Toxicology and Chemistry* **2009**, *28* (9), 1842–1850.
75. Reungoat, J.; Macova, M.; Escher, B.I.; Carswell, S.; Mueller, J.F.; Keller, J. Removal of micropollutants and reduction of biological activity in a full scale reclamation plant using ozonation and activated carbon filtration. *Water Research* **2010**, *44* (2), 625–637.

II

COV: Comparative Overviews of Important Topics for Environmental Management

7

Buildings: Climate Change

Lisa Guan and
Guangnan Chen

Introduction

The greenhouse effect is a natural warming process of the earth. It is caused by the greenhouse gases, such as water vapor, carbon dioxide (CO_2), and methane (CH_4), which trap long-wave radiation and then radiate the energy in all directions, warming the earth's surface and atmosphere (Figure 1).[1] Without the heat-trapping greenhouse gases, scientists estimate that the average earth surface temperature would likely to be some 30°C colder than it is today, or –18°C instead of the present mild average 15°C.[1]

The enhanced greenhouse effect is additional to the natural process of greenhouse effect and is mainly due to human activities (human induced), including burning fossil fuels, land clearing, and agriculture, which change the makeup of the atmosphere and lead to an increased concentration of greenhouse gases.[1] This has the potential to cause significant changes in the global climate system (referred to as climate change), including increased temperature, changed patterns of rainfall, tropical cyclone activities, and other extreme climatic events.

Climate change has now become one of the most important global environmental issues facing the world today. It is now widely recognized as having significant potential to seriously affect the integrity of our ecosystems and human welfare.[2] The effects, or impacts, of climate change may be in physical, ecological, social, and/or economic areas.

In this entry, the likely future climate change is first presented. The cycling interaction between climate change and buildings is then discussed, which includes both aspects of the implication of climate change on building performance and the contribution of buildings to the process of human-induced climate change. The potential strategies for building design and operation are then highlighted, in order to reduce the greenhouse gas emissions from buildings and to prepare the buildings to withstand a range of possible climate change scenarios.

Likely Future Climate Change

Due to uncertainties in future emissions and concentrations of greenhouse gases, as well as the climate system's response to the changing conditions and the natural influences (e.g., changes in the sun and volcanic activity), it may be difficult to accurately predict the extent of climate changes.[2] However, the advancements in climate model simulations, combined with more and more observed data on climate changes, led the Intergovernmental Panel on Climate Change (IPCC) to predict the following likely scenarios of climate changes[2]:

- Average global surface temperature will likely rise a further 1.1–6.4°C (2.0–11.5°F) during the 21st century. It is expected that the average rate of warming is very likely to be at least twice as large as that experienced during the 20th century.
- Warming will not be evenly distributed around the globe and will vary with different seasons. For example, land areas will warm more than oceans, high latitudes will warm more than low latitudes, and winters will be warming more than summers in most areas.
- There will be significant changes to the amount and pattern of precipitation, including an increase in droughts, tropical cyclones, and extreme high tides. The changes in precipitation, either an increase or a decrease, will vary from region to region.
- The global average sea level is estimated to rise by 18–59 cm (7.2 to 23.6 in.) by 2100 relative to 1980–1999. Current model projections also indicate substantial variability in future sea level rise between different locations.
- Increases in the intensity of extreme weather events, such as storms, heat waves, and drought, are also predicted.
- In regard to the projection of future climate change, it is noted that there are different levels of confidence (e.g., in terms of accuracies and reliability) for different climate parameters. For

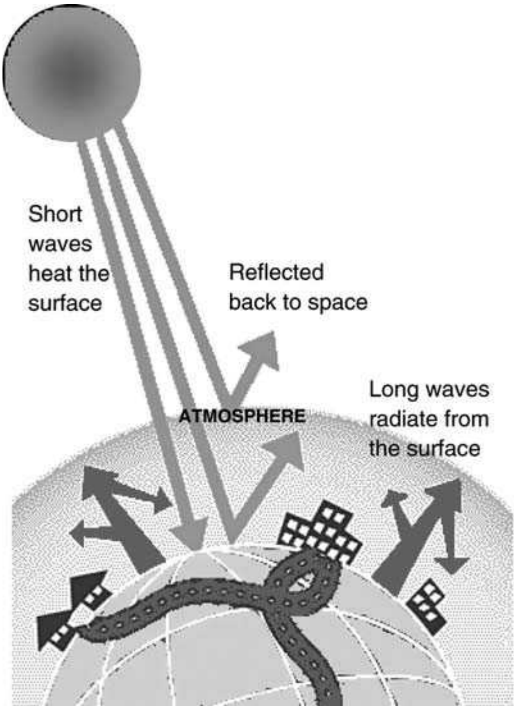
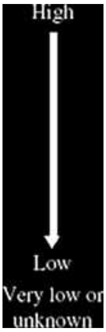


FIGURE 1 Natural/enhanced greenhouse effect.

TABLE 1 List of Climate and Associated Scenario Variables, Ranked Subjectively in Decreasing Order of Confidence

Climate Variable	Confidence
Atmospheric CO_2 concentration	
Global—mean sea level	
Global—mean temperature	
Regional seasonal temperature	
Regional temperature extremes	
Regional seasonal precipitation and cloud cover	
Regional potential evapotranspiration	
Changes in climatic variability (e.g., El Niño, daily precipitation regimes)	
Climate surprises (e.g., disintegration of the West Antarctic Ice Sheet)	

Source: Hulmer and Sheard.[3]

example, it is believed that there is higher confidence in the projection of increases in carbon dioxide concentrations and rises in sea level than in storminess or intense precipitation events. A list of climate and associated scenario variables, ranked subjectively in decreasing order of confidence, had been recommended by Hulme and Sheard[3] and is shown in Table 1.

Implication of Climate Change on Buildings

Climate change is likely to affect both the performance of existing building stock and the design of new buildings. For example, under climate change, buildings will have more overheating hours in summer and less underheating hours in winter, thus use more cooling energy and less heating energy. Where heating and cooling are provided by different fuels, this could have a significant influence on the design and operations of energy delivery systems.

Because climate change entails new climatic conditions for the building industries, it is expected that climate change will have a significant impact on the design, construction, and performance of buildings, as well as the health and productivity of people living and working inside them.[4] These impacts may include the following:

- *Higher building energy consumption*: Climate change may require higher capacity and more uses of air-conditioning equipment to provide comfort indoor environment. For example, it has been predicted that for air-conditioned office buildings in Australia, the cooling load may increase by 2%–47%, depending on the assumed future climate scenarios, as well as different locations.[5] The increases of total building energy use would range from 0.4% to 15.1%. However, due to the potential decrease of heating energy in winter, skin-dominated buildings located in cold regions could receive some benefits from the climate change. Moreover, the expected increased stringencies in building energy codes around the world would also offset some increases of building energy use due to climate change.[6,7]
- *Deteriorating internal thermal environment*, such as more overheating in summer. It has been found that for air-conditioned office buildings in Australia, when the annual average temperature increase exceeds 2°C "threshold," the risk of current office buildings subjected to overheating will be significantly increased.[5] When the increase of external air temperature is more than 5°C, all the Australian office buildings would suffer from the overheating problem regardless where they are located. This could have significant implications on people's health and capacity to work and productivity. It has been estimated that global warming is currently contributing to the death of about 160,000 people every year.[8]

- *Structural integrity*, such as more severe wind and snow loading, and foundation movement. For example, shrinkage or expansion in clay soils can lead to foundation movement and cracking of walls, which may cause damage in building structure. In Northwest Norway, severe damage to buildings was caused by the hurricane on January 1, 1992, and several buildings collapsed due to heavy snow loads on roofs during the 1999/2000 winter.[9]
- *External fabric*: Durability of external fabric becomes shorter due to increased storm, rain, flood, and other weather conditions. For example, the strength and durability properties of concrete may be influenced by the changes in its environment (e.g., temperature, humidity).
- *Construction process* may be disrupted due to adverse weather condition. This may have implication in the project planning and the associated challenges to complete the project on time and within the budget.
- *Service infrastructure* may become inadequate. For example, changing weather patterns and more frequent and intense storm may lead to the drainage problem. The existing drainage system in many parts of the world may not be able to cope with increased storm loads.

Contribution of Buildings to Human-Induced Climate Change

The climate system is a dynamic system in transient balance.[10] A change of external and/or internal climate forcing imposed on the planetary energy balance would cause a corresponding change in global temperature.[10] Overall, scientists have now been able to reach a broad consensus and provide overwhelming evidence to suggest that human activities are having a discernible influence on the global climate.[2] In particular, it has been found that the recent rapid increase in global temperature is closely aligning with the strong growth in use of fossil fuel over the past 50 years.

The construction and operation of modern buildings consume a considerable amount of energy and materials and therefore contribute significantly to the process of human-induced climate change. Figure 2 shows the world greenhouse gas emissions by sector, end use, activity, and gas types.[11] It can be seen that buildings, including both commercial buildings and residential buildings, account for 15.3% of world greenhouse gas emission, which is greater than the sectors of transportation, agriculture, and waste. Indeed, buildings are one of the most significant infrastructures in modern societies.

Worldwide, the Worldwatch Institute estimates that the construction and operation of buildings is responsible for 40% of the world's total energy use, 30% of raw materials consumption, 55% of timber harvests, 16% of freshwater withdrawal, 35% of global carbon dioxide (CO_2) emissions, and 40% of municipal solid waste sent to landfill.[12] In 2004, it was estimated that worldwide, the total emissions from the building sector, including the electricity consumed, were 8.6 Gt CO_2, 0.1 Gt CO_2–eq N_2O, 0.4 Gt CO_2–eq CH_4, and 1.5 Gt CO_2–eq halocarbons (including CFCs and HCFCs).[13]

Basically, the impact of buildings on the human-induced climate change is through three routes, including building operational energy, building embodied energy, and building-related refrigerants. Operational energy includes all energy used for mechanical services [e.g., heating, ventilation, and air conditioning (HVAC) systems], electrical services (e.g., lighting and other office appliances), and hydraulic services (e.g., pumping system). Embodied energy includes all energy used for the production of building materials, their transportation and handling, and building construction processes. Building-related refrigerants include the refrigerants used in air-conditioning systems (e.g., Freon used in compressors and chillers) and other building appliances (e.g., refrigerators and freezers), which may have ozone depletion potential (e.g., depleting the ozone in the upper atmosphere) and/or global warming potential that persists in the upper atmosphere, trapping the radiation emitted by the earth.

Previous studies of Life Cycle Assessment (LCA) have shown that the CO_2 emission from the sources of building embodied energy and building-related refrigerants is often fairly small for commercial office buildings. For instance, based on the assumption of a building having a 40 years life span, it has been

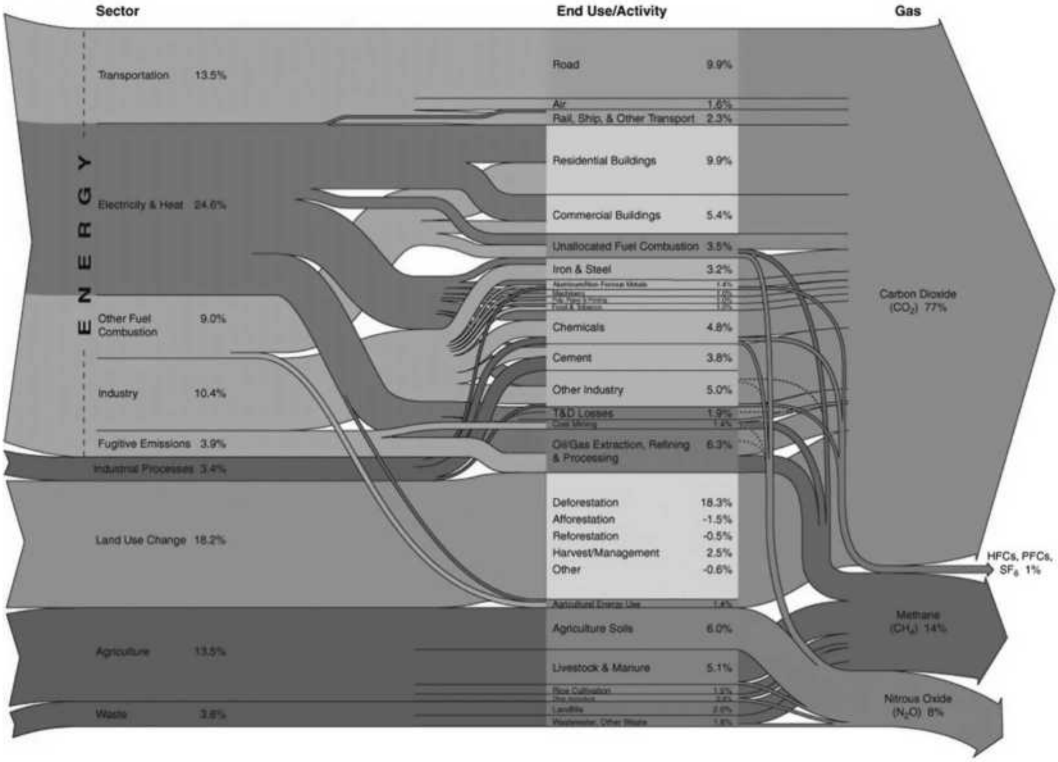


FIGURE 2 World greenhouse gas emissions by sector, end use, activity, and gas types.
Notes: All data are for 2000. All calculations are based on CO_2 equivalents, using 100 years global warming potentials from the IPCC (1996), based on a total global estimate of 41,755 $MtCO_2$ equivalent. Land use change includes both emissions and absorptions. Dotted lines represent flows of less than 0.1% of total greenhouse gas emissions.
Source: World Resources Institute.[11]

shown that the embodied energy emissions contributed only approximately 8%–10% of building total emission.[14] More than 90% of building energy consumption occurs during the use/operational phase, so the energy performance of the buildings is particularly important, especially for buildings with 24 hours occupancy.

Relationship between Climate Change and Buildings

As discussed above and also illustrated in Figure 3, the cycling interaction between climate change and buildings is of dynamic nature. Climate change, for instance, would generally lead to more use of air conditioning, which leads to more greenhouse gas emission and then climate change. Therefore, both climate change and buildings are essentially the cause and the effect of each other.[15] On one hand, climate change is expected to impact on many aspects of building design, construction, and operation. On the other hand, buildings have also contributed significantly to the human-induced climate change. They have produced more greenhouse gas emission than the sectors of transportation, agriculture, and waste.

Because greenhouse gas concentrations are still continuing to increase, the process of climate change will continue and may accelerate. This requires the building sector to develop suitable strategies, including the enhancement in building energy codes, to mitigate the greenhouse gas emissions, which should

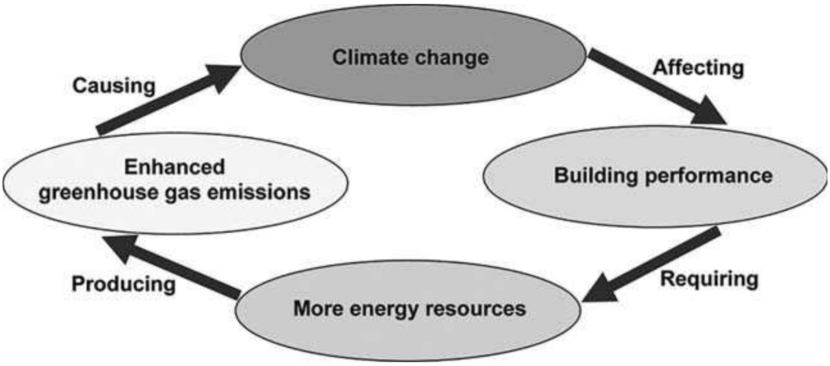


FIGURE 3 The cycling interaction between global warming and buildings.

be not only environmentally effective but also economically efficient. In parallel, appropriate adaptation strategies are also needed to prepare buildings to withstand the future inevitable climate change. Adaptation to climate change has now become one of the key requirements for buildings.

It is noted that there may be both potential synergy and conflict between adaptation and mitigation measures and strategies. This will require the development of integrated rather than separate responses.[16–18] Moreover, it seems that there is a tight intertwining between the issues of adaptation and mitigation.[19,20] The energy-efficient and renewable energy technologies, for instance, often can play dual roles in mitigating greenhouse gas emissions while increasing adaptive capacity by making buildings more disaster resilient. An effective response to climate change may not always be costly and could even be both profitable and socially beneficial.

Potential Strategies for Building Design and Operation

Buildings are one of the most significant infrastructures in modern societies, having significant impact on health and productivity of people living and working inside them. In particular, as buildings have a very long life span, typically 50–100 years, it is very important that all the current and future building stocks be designed and maintained to perform satisfactorily in future climates.

For both new and existing buildings, potential mitigation may include the utilization of renewable energy and low emission energy, energy efficiency, and energy conservation. Appropriate adaptation may also be achieved by focusing on major building design parameters for new buildings or retrofit of building envelope and/or internal heat sources such as lighting systems for existing buildings. In particular, those improved building energy codes currently adopted around the world, including the requirement of thermal conductance values for building envelope, lighting power densities for internal heat sources, and HVAC system efficiency for building mechanical services,[6,7] could play key roles in both the potential mitigation and adaptation of buildings. Overall, the aim is to ensure that all the existing and new buildings not only perform and operate satisfactorily in the new environment but also satisfy the environmental performance criteria of sustainability.

Building Envelope

The building envelope is a significant determinant of energy required to heat and cool a building. Building designers can use the building envelope as a “filter” to accept or reject solar radiation and outside air, depending on the need for heating, cooling, ventilation, and lighting.[13] The heat storage capacity of the building envelope can also be used to reduce peak thermal loads. As a result, a well-designed building shell can substantially reduce the building cooling and heating demands and lighting energy through the provision of daylighting, particularly at the perimeter zones. This can result in not only substantial capital cost savings by enabling smaller sizing of the air-conditioning plant to be used but also significant operating cost saving by reducing the quantity of air circulation and fan power, and partial load situation for the HVAC systems.

It is particularly important that building designers employ energy-efficient principles, as well as integrated building design approaches at design and construction of the building envelope. Overall, for the building envelope, the potential mitigation and adaptation strategies in the face of climate change may include the following:

- Installation of insulation in foundation, walls, and/or roofs to minimize conductive heat loss/gain.
- Optimum design of windows, including careful determination of window-to-wall ratio (WWR) and selection of types of glass and shading to optimize use of solar heat and light.
- Sealing gaps to prevent draughts and heat infiltration to minimize uncontrolled movement of air into the building.
- Utilizing thermal mass, reflective roofs, and trees for shade, etc., if it is appropriate.

Internal Heat Sources

The internal heat sources include the occupants, lights, electric appliances, and machines. Their influence may be not only directly on the total energy consumption of the building but also on the building thermal loads, which indirectly influence the energy consumption of HVAC systems. The selection of lights, for example, can have significant impact on the building energy and thermal performance. Advances in higher-efficiency light sources, lower-loss ballasts and control gear, and more efficient luminaries will therefore enable lighting energy usage to be significantly reduced. Developments in the technology of daylighting will further increase the potential lighting energy savings. The potential mitigation and adaptation strategies to manage internal heat sources may include the following:

- Optimum integration of lamp (lighting sources), ballasts (for electric-discharge lamps), and fixtures (e.g., reflectors, diffusers, and/or polarizing panels). It was found that to maintain the same lighting level of 320 lux, lighting power densities can vary from 2 to 5 W/m^2 (averaged over time) for having state-of-the-art lighting technology to 25–35 W/m^2 for having only mediocre level of efficiency.[21]
- Purchasing energy-efficient office equipment and running it wisely. Energy-efficient office equipment can use less than half the energy as the standard models—at no or little additional capital cost.[21] Recommendations may include the use of “energy star” office equipment and “energy labeling” appliance. It was found that every dollar invested in energy efficiency, at a 20%–30% saving rate, is equivalent to increasing net operating income by 3%–4% and net asset value by $2.50–$3.75.[22]

Using the building simulation technique, it has been shown that for Australian office buildings, if the building total internal load density, which includes all internal heat sources of lighting, plug load, and occupants, is reduced from 43 to 21.5 W/m^2, the building total energy use under the future 2070 high scenario (the worst case) can be reduced by up to 89–120 kWh/m^2 per annum and the overheating problem could be completely avoided and the office building will perform as good as at the current climate scenario.[23]

Building Mechanical Services

Building mechanical service or HVAC systems are used to control factors affecting thermal comfort, with the ultimate purpose being to provide a clean, noise-free, and efficient working environment.[24] Since air-conditioning systems account for around one-third of the total energy use in typical office buildings, the proper design and selection of HVAC systems are critical for the performance of these buildings.

Generally, according to the purpose of the building, as well as local site information and climate condition, a building may be designed as naturally ventilated, mechanically ventilated, air conditioned, or a combination of these three options (hybrid mixed-mode system). Potential mitigation and adaptation strategies for design and selection of building service systems include the following:

- Using passive design for heating, cooling, and ventilation. A building relying entirely on air conditioning should be avoided as much as possible.
- Utilization of renewable energy and/or low emission energy for the operation of the building. It was found that on clear days throughout the year, reductions of conventional power use of at least 60% can be achieved with optimum photovoltaic cladding densities targeted to lighting and small power load demands.[25]
- Selecting energy-efficient technologies. Such technologies may include evaporative cooling, desiccant cooling, chilled ceilings, displacement ventilation, cogeneration and district heating and cooling, active solar and heat pump systems, and underground earth pipes.[26]
- Good control of HVAC systems. It was found that the performance of the HVAC system is subject to the significant influence of its control strategy and maintenance. Building energy management systems (BEMSs) can also be a useful tool for energy saving in the range of 5%–40%.[26]
- Good commissioning, operation, and maintenance of HVAC systems. It was found that of the deficiencies found in the retro-commissioning of buildings, 85% were related to HVAC systems.[22] In the same study, it was also reported that energy savings between 7% and 29% may be achieved with paybacks ranging from 0.2 to 2.1 years.

Recommendations on Implementation of Strategies

Overall, it is recommended that the potential mitigation and adaptation strategies should be taken at the early planning and design stage. This is because the thermal loads gained/lost from the building envelope and internal heat sources are the “base load” for the selection of HVAC equipment. Appropriate selection and design of these systems are therefore vital for the energy efficiency of the building. Without a suitable choice, all later work will be built on an unsatisfactory foundation. It is also noted that the mitigation and adaptation strategies taken at the early design stage will future-proof the building and be more effective in the longer term. Although the initial capital cost of taking the measures may be higher, the overall cost in many cases would actually be much lower than having to do retrofit strengthening at a later stage.[16]

Building regulation and design codes can also play a crucial role in the anticipation and avoidance of risk.[27] Because climate change would have a major impact on the frequency of extreme weather events, building codes need to be reviewed regularly in order to maintain a proper level of reliability and be adhered to in practice.[9] For instance, changes in the frequency of storms will have building code implications.[28] However, it is argued that simply raising performance standards as a response to climate change, without dealing with the issue of noncompliance with existing standards, may not be desirable. It may run the risk of undermining the legitimacy of regulation generally.[27]

In the face of inevitable climate change, the shifting from a reliance on historical data to a reliance on predicted future data may be also necessary, as it is increasingly important to precisely predict the conditions under which buildings and other infrastructure will need to withstand in the future. It is understood that using different sets of design conditions will have significant implication on the cost and performance of buildings. For example, using the building simulation technique, it has been shown that for typical Australian office buildings, the required cooling capacity may increase from 28% to 59% if the new buildings were designed using the future project climate (e.g., 2070 high scenario) rather than the current climate condition.[5] In order to maintain comfort indoor condition, a further increase of 4%–10% may be required in addition to the possible increase of 27%–47% cooling load if the buildings were designed at current climate conditions.

Conclusion

The climate change induced by the emissions of greenhouse gases is one of the most important global environmental issues facing the world today. Buildings are one of the most significant infrastructures in modern societies, and they need to be prepared to withstand climate change. On one hand, climate change is going to impact on many aspects of buildings, including both building design and building operation. On the other hand, buildings have also contributed significantly to the process of human-induced climate change. In this entry, both aspects of knowledge have been discussed.

It has been suggested that the potential mitigation and adaptation strategies should focus on major energy-related factors, such as design and construction of building envelope, design and careful selection of the air-conditioning system, and selection of management of internal heat sources including both lighting systems and electrical equipment. In many cases, it has been demonstrated that the energy-efficient and renewable energy technologies can play dual roles in mitigating greenhouse gas emissions while increasing adaptive capacity by making buildings more disaster resilient.

It has also been recommended that energy efficiency principles, such as adopting passive design principles for heating, cooling, ventilation, and lighting, and integrated design approaches should be adopted at the early design stage and be maintained in the whole life of buildings. These strategies have the potential of not only being environmentally effective but also economically efficient and socially beneficial.

References

1. Healey, J., Ed. *Climate Change—Issues in Society*; The Spinney Press: Sydney, 2003; Vol. 184, ISBN 1-876811-93-5.
2. IPCC. Climate change 2007: The physical science basis. *Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*; Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K.B., Tignor, M., Miller, H.L., Eds.; Cambridge University Press: Cambridge and New York, 2007, 996.
3. Hulme, M.; Sheard, N. *Climate Change Scenarios for Australia*. Climatic Research Unit; Norwich, 1999, 6 pp.
4. Guan, L. Global warming: Impact on building design and performance. *Encycl. Energy Eng. Technol.* **2009**, *1* (1), 1–6.
5. Guan, L. Implication of global warming on air-conditioned office buildings in Australia. *Build. Res. Inf.* **2009**, *37* (1), 1–12.
6. ANSI/ASHRAE/IESNA Standard 90.1-2007. Energy standard for buildings except low-rise residential buildings.
7. Building Code of Australia (BCA)—Australian Building Codes Board (ABCB), 2010.
8. Bhattacharya, S. Global warming kills 160,000 a year. *New-Scientist.com News Service*, 17, 01 October 2003, available at http://www.newscientist.com/article.ns?id=dn4223.
9. Lisø, K.R.; Aandahl, G.; Eriksen, S.; Alfsen, K. Preparing for climate change impacts in Norway‘s built environment. Build. Res. Inf. **2003**, *31* (3), 200–209.
10. McGuffie K.; Henderson-Sellers, A. A Climate Modeling Primer, 3rd Ed.; John Wiley and Sons, Ltd.: Chichester, 2005.
11. *World GHG Emissions Flow Chart*; World Resources Institute, Retrieved on 09 February 2011, available at http://cait.wri.org/figures.php?page=/World-FlowChart.
12. Fenner, R.A.; Ryce, T. *A comparative analysis of two building rating systems. Part 1: Evaluation. Proc. Inst. Civil Eng., Eng. Sustain*, **2008**, *161* (ES1), 55–63. doi: 10.1680/ensu.2008.161.1.55.
13. Levermore, G.J. A review of the IPCC assessment report four, Part 1: The IPCC process and greenhouse gas emission trends from buildings worldwide. *Build. Serv. Eng. Res. Technol.* **2008**, *29* (4), 349–361.

14. Australian Greenhouse Office. *Australian Commercial Building Sector Greenhouse Gas Emissions 1990–2010*; Canberra, 1999; ISBN 18-76536-195.
15. Degelman, L.O. Which came first—Building cooling loads or global warming? A cause and effect examination. *Build. Serv. Eng. Res. Technol*. **2002**, *23* (4), 259–267.
16. Lowe, R. Really rethinking construction. *Build. Res. Inf.* **2001**, *29* (5), 409–412.
17. Steemers, K. Towards a research agenda for adapting to climate change. *Build Res. Inf.* **2003**, *31*, 291–301.
18. Lowe, R. Lessons from climate change: A response to the commentaries. *Build. Res. Inf.* **2004**, *32* (1), 75–78.
19. Camilleri, M.; Jaques, R.; Isaacs, N. Impacts of climate change performance on building in New Zealand. *Build. Res. Inf.* **2001**, *29* (6), 440–450.
20. Mills, E. Climate change, insurance and the buildings sector: Technological synergisms between adaptation and mitigation. *Build Res. Inf.* **2003**, *31* (3–4), 257–277.
21. Sustainable Energy Development Authority (SEDA). Tenant energy management handbook—Your guide to saving energy and money in the workplace, 2000, ABN 80-526-465-581.
22. Cull, S.L. Commissioning: Retrocommissioning. *Encycl. Energy Eng. Technol.* **2007**, *1* (1), 200–206.
23. Guan, L. Adaptation to global warming by changing internal loads of buildings. *Proceedings of International Conference on Building Energy and Environment (COBEE)*, Dalian, China, July 13–16, 2008.
24. ASHRAE. *ASHRAE Handbook: Fundamentals*; American Society of Heating, Ventilating and Air-Conditioning Engineers (ASHRAE): Atlanta, 2001.
25. Jones, A.D.; Underwood, C.P. Cladding strategies for building-integrated photovoltaics. *Build. Serv. Eng. Res. Technol*. **2002**, *23* (4), 243–250.
26. Levermore, G.J. A review of the IPCC assessment report four, Part 2: Mitigation options for residential and commercial buildings. *Build. Serv. Eng. Res. Technol.* **2008**, *29* (4), 363–374.
27. Lowe, L. Preparing the built environment for climate change. *Build. Res. Inf.* **2003**, *31* (3), 195–199.
28. Larsson, N. Adapting to climate change in Canada. *Build. Res. Inf.* **2003**, *31* (3), 231–239.

8

Economic Growth: Slower by Design, Not Disaster

Peter A. Victor
and Tim Jackson

Introduction

Economic growth is a very recent phenomenon, dating back two or three centuries and limited in its extent to only some parts of the world. While economic growth has dramatically improved the lives of billions of people, billions more remain in abject poverty. The mainstream view is that all countries should strive for economic growth, and through a process of convergence where poorer countries grow the fastest, all the people of the world will eventually enjoy a high and secure material standard of living.

There are many problems with this vision of growth for all. Foremost is the mounting evidence that the world's economy is already bumping up against and even surpassing the biophysical limits of the planet. This is showing up in terms of global and local environmental degradation and resource scarcity. Consequently, a different vision is emerging, one in which rich countries no longer pursue economic growth as a primary objective, leaving room for the economies of poorer countries and regions to expand, at least temporally. An additional motivation for low and no growth is to reduce competition with other species with which humans share the planet.

This chapter examines the possibility of rich countries managing without growth, by which is meant achieving improvements in well-being without economic growth. That such an outcome is possible is illustrated with the use of LowGrow SFC, an interactive macroeconomic model of the Canadian economy. The chapter concludes with a discussion of policy directions implied by such a scenario with special emphasis on employment policies and funding public services in a low-/no-growth economy.

A Brief History of Economic Growth

In terms of human history, economic growth is a very recent phenomenon. World gross domestic product (GDP) was only 14% higher at the end of the first millennium than when it began. GDP is the "total unduplicated value of the goods and services produced in the economic territory of a country or region during a given period" (Statistics Canada). GDP per capita was no higher at all, slightly less even, owing to population growth. From the year 1000 to 1400, world GDP increased by 140%, but GDP per capita grew only 60% and at a barely perceptible rate, as population expanded. Over the next 400 years, world GDP increased a further 140%, and although GDP per capita increased by 30% by 1700, it fell back again in 1800 to the same level as in 1400. Average world GDP per capita declined further in the early 1800s, after which economic growth accelerated. By 1900, world GDP was over five times greater than in the previous 100 years, growing at an average annual rate of 1.7%, and GDP per capita was more than three times greater. The rate of economic growth increased in the 20th century such that annual world GDP grew almost ninefold at an average annual rate of 2.2% and GDP per capita rose 2 1/2 times.

These figures,[1] approximate as they are especially early on, show that economic growth has been exceptional in human history. The vast majority of people have lived in circumstances where they had no reason to believe that their children's lives would be materially any different from their own. To think otherwise is a very modern idea dating back perhaps a dozen or so generations, and even then, only in some parts of the world.

Of course, global averages conceal much of the huge variation in experiences from country to country and region to region. In the 18th and 19th centuries, most of the world's economic growth was confined to Europe and ex-colonies of European powers. Some regions were impoverished in the process, 19th-century India being a prime example through the deliberate destruction of cotton manufacturing at the behest of British industrialists. The rest of the world continued to experience lives in which economic growth had little or no impact and was not part of the lived experience of most people. Even in places where growth rates were highest, there were substantial areas of poverty. In the 20th century and into the 21st century, economic growth spread to more parts of the globe but unevenly with 735 million people in 2015 still living in extreme poverty (less than US$1.90 per day) and 3.4 billion living below US$5.50 per day, about the price of a single cup of coffee in the United States. While the number living in extreme poverty declined by 40% since 1990, the number living below $5.50 only declined by 5%, and the rate of reduction in extreme poverty is slowing down.[2] Contrary to earlier expectations that global economic growth would close the gap between rich and poor, the spread has increased, just as it has done within even the richest countries.[3]

All this economic growth has required a massive increase in the use of natural resources. In the 20th century, the global use of construction materials, ores and industrial materials, fossil fuels, and biomass increased eightfold, almost as much as GDP.[4] Since 2000, the global use of these resources has risen even faster than GDP. The much-vaunted "green growth" which depends fundamentally on increasing GDP with declining use resources is nowhere to be seen.[5] The increasing extraction, use, and disposal of materials have resulted in large-scale, adverse environmental impacts. One prominent list includes the following: climate change, ocean acidification, stratospheric ozone depletion, overloading of the nitrogen and phosphorus cycles, global freshwater uses, land system change, biodiversity loss, atmospheric aerosol loading, and chemical pollution.[6] Then, there is the looming threat of peak oil from conventional sources[7] and the more general concern that the age of cheap energy is coming to an end,[8] as well as an emerging scarcity of "critical" minerals for new technologies.[9] All this plus serious doubts about future prospects for increases in productivity[10,11] and concerns over secular stagnation,[12] make for a bleak outlook for real, long-term, comprehensive, global economic growth. With the world human population approaching 8 billion, projected to rise to nearly 10 billion by mid-century, the question arises as to whether economic growth for all is a viable option in the 21st century and if not, what are the alternatives?

One alternative is to shift the over-riding economic priority of the richest economies away from the single-minded pursuit of growth and to reduce “growth dependency.” After all, economic growth has only held this position as the single most important economic policy objective since about 1960. “There is in fact hardly a trace of interest in economic growth as a policy objective in the official or professional literature of western countries before 1950.”[13] And even when it was introduced, it was as a means to fulfill other policy objectives such as full employment, rather than as an end in itself. Now the pursuit of economic growth is deemed so important that it is customary for policy proposals across many domains including environment, education, and the arts to be judged in terms of their implications for growth, or one of its surrogates such as competitiveness or productivity.

If the poor countries of the world are to benefit from economic growth, and the world economy is to function within the “safe operating space” of the planet,[6] then rich countries, those that have benefited the most from economic growth in the past, must be prepared to make room for them. Otherwise, disaster threatens, brought about by the excessive pressure on Earth systems.

One approach to this predicament based on principles of distributive justice would be to determine fair shares of access to global resources, accounting also for the interests of other species. This is the kind of dialogue that is underway in the slow progression of international climate negotiations, illustrating most profoundly the difficulties in reaching and then implementing agreement in a global world divided into national and regional power blocs. Nonetheless, a country or group of countries could adopt a view of its fair share independently of an international agreement and set these as boundaries within which their economy must function. Depending on the specification of the boundaries and the capability of economies to adjust, economic growth, measured in the conventional way as an increase in real (inflation) adjusted GDP, could conceivably continue for a time, but incidentally rather than as a primary objective.

What would such development entail? Total environmental (including resource) impact is the product of GDP and impact per unit of GDP (e.g., GDP × greenhouse gas (GHG) emissions per unit of GDP or GDP × energy used per unit of GDP). Therefore, if environmental impact is to decline as GDP grows, environmental impact per unit GDP must decline faster than the rate of economic growth. This is one meaning of “green” growth, and it underlies the downward-sloping portion of the Environmental Kuznets Curve.[14] (The Environmental Kuznets Curve is an inverted “U” with a measure of environmental impact plotted on the y-axis and GDP plotted on the x-axis. The hypothesis is that in the early stages of economic growth, environmental impact rises to a maximum after which it declines as growth continues.) There are examples of obvious, local problems such as urban air quality whose history can be described by an Environmental Kuznets Curve, but it is a poor description of global materials and energy use or global environmental impacts over the past century.

Whether green growth defined in this way (i.e., impact/GDP declining faster than GDP increases) is possible at the global level remains an open question. It is complicated by the fact that many environmental and resource depletion problems relate to stocks such as accumulating GHGs in the atmosphere, diminishing rain forests, and declining fish stocks. Reducing the flows that determine these stocks may be insufficient to bring the stocks to the required levels fast enough or at all. If it is not possible to sufficiently decouple economic growth from its environmental and resource impacts, growth will have to cease and even turn negative for a time as proponents of degrowth argue.[15] Otherwise, it will not be possible to bring the global economy back within the planetary boundaries that are already being exceeded and others which will be exceeded if present trends are not reversed. Given what we know about the state of the environment, and concern over supplies of low-cost energy and critical materials, there is a very strong case on ethical and practical grounds for rich countries to take the lead in managing without growth.

Economists Question Growth

At the same time as economic growth was reaching the pinnacle of policy objectives, dissenting voices were beginning to be heard. One of the most widely prominent was John Kenneth Galbraith. In *The Affluent Society*[16] published in 1958 and revised through multiple editions, Galbraith compared private affluence in the United States with public squalor. He also questioned the efficacy of dealing with poverty through a general rise in incomes. Many academic economists regarded Galbraith as more of a political commentator than a serious economist because of his disdain for theoretical economics, and on these tenuous grounds, they resisted his arguments. The same could not be said of British economist Ezra Mishan who published *The Costs of Economic Growth* in 1967.[17] Mishan was a highly regarded and well-published expert in "welfare economics," the field within mainstream economics that is concerned with the relationship between economic activity and well-being. Thus, although Mishan's analysis of the costs of economic growth was aimed at a broad audience, no one could dismiss the author as not really understanding modern economic theory.

Perhaps this is one reason why Mishan's critique of economic growth, unlike Galbraith's, ignited a heated debate that went on for several years between him and Wilfred Beckerman, another well-established British economist. Beckerman wrote "Why We Need Economic Growth"[18] and *In Defence of Economic Growth*.[19] Later, Beckerman wrote *Small Is Stupid*[20] in response to Schumacher's widely read *Small Is Beautiful*,[21] Schumacher's critique of modern industrialized economies. Many of Schumacher's arguments about the optimal scale of an economy were anticipated, echoed, and augmented by other economists such as Kenneth Boulding in his seminal essay "The Economics of the Coming Spaceship Earth,"[22] N. Georgescu-Roegen,[23] who explored the implications of the second law of thermodynamics for economics and economic growth, and Herman Daly who has promoted a steady-state economy for more than three decades.[24] The publication of *The Limits to Growth*[25] in 1972 addressed similar themes using systems dynamics but was roundly and largely unfairly criticized especially by economists (e.g. Maddison,[1] pp. 90–94). It remains influential to this day. A useful summary of the state of the growth and no-growth debate, largely where it was left in the 1970s, can be found in Olson and Landsberg's collection of essays.[26] (For a contemporary assessment of the Limits to Growth debate see the review by Jackson and Webster.[27])

After this flurry of publications in the 1960s and 1970s, the growth debate subsided. In the late 1990s, the criticisms of growth resurfaced stronger than ever with economists such as Douthwaite, *The Growth Illusion*,[28] Daly, *Beyond Growth: The Economics of Sustainable Development*,[29] and Booth, *The Environmental Consequences of Growth: Steady-State Economics*[30] leading the charge. By this time, the transdisciplinary, ecological economics was almost 20 years old, with a dedicated peer-reviewed journal, *Ecological Economics*, publishing 12 times a year including many papers dealing with problematic aspects of economic growth. In the first part of the 21st century, it is impossible to keep up with the many papers, reports, blogs, conferences, media entries, and YouTube videos questioning economic growth. And there are many books such as Hamilton, *Growth Fetish*,[31] Booth, *Hooked on Growth*,[32] Victor, *Managing without Growth*,[33] Speth, *The Bridge at the End of the World*,[34] Brown, *Right Relationship*,[35] Jackson, *Prosperity without Growth*,[36] Schor, *Plenitude*,[37] Dietz and O'Neill,[38] *Enough Is Enough*, Czech, *Supply Shock*,[39] and von Weizsäcker and Wijkman, *Come On!*[40] just to name a few in the English language alone.

The remainder of this chapter describes an investigation into what might be possible in an economy in which economic growth ceases. In particular, the following question is addressed: is it possible to have low unemployment, reduced GHG emissions and other environmental pressures, reduced income inequality, and reduced hours of paid work while maintaining reasonable debt levels in the public and household sectors, all in the absence of economic growth? And, if so, what policy frameworks or initiatives would be required? LowGrow SFC, an interactive macroeconomic systems dynamics model, was developed for the Canadian economy specifically to help answer these questions. In the next section, LowGrow SFC is described in fairly general terms and three very different scenarios for Canada to the year 2067 are presented. Brief comments on policy directions suggested by the simulation of low/no growth are followed by a more detailed consideration of employment in a no-/low-growth economy and the implications for government finance and the provision of public services.

Exploring Low and No Growth in Canada with LowGrow SFC

LowGrow SFC is a quantitative model of the Canadian economy designed to explore future scenarios. LowGrow SFC has been calibrated using Canadian data with simulation results reported from 2017 to 2067. The scenarios are not predictions. Rather, they are offered as consistent and plausible future possibilities intended to feed discussions about current choices and alternative futures. The model integrates three primary spheres of interest within a system dynamics framework: (1) the environmental and resource constraints on economic activity; (2) a full account of production, consumption, employment, and public finances in the "real economy" at the level of the nation state; and (3) a comprehensive account of the financial economy, including the main interactions between financial agents. LowGrow SFC conforms quite closely to standard economic frameworks. Data are drawn directly from the Canadian national accounts and some of the behavioral relationships in the model are estimated econometrically on the basis of time-series data from the Canadian economy. At the same time, LowGrow SFC departs from more conventional economic modeling approaches by incorporating time-lags, feedbacks and expectations in the model, and also by allowing for some potentially radical variations on "typical" macroeconomic policy.

The theoretical basis for LowGrow SFC draws heavily on the post-Keynesian macroeconomic approach of Godley and Lavoie (2012), which places a particular emphasis on a full and consistent account of the relationships between monetary stocks and flows within and between different financial sectors. In the aftermath of the 2008 financial crisis, so-called stock-flow consistent (SFC) modeling has gained a particular traction because of its ability to provide a comprehensive account of financial transactions in the economy and to map the impact of these on financial balance sheets—something that was conspicuously missing in the run-up to the crisis. LowGrow SFC is articulated in terms of six interrelated sectors: households, firms, banks, government, a central bank and the "rest of the world" (or "foreign" sector). It models a range of financial assets and liabilities including deposits, loans, mortgages, government bonds, and firms' equities.

In LowGrow SFC, as in the economy that it represents, economic growth is driven primarily by net investment which increases the capital stock and hence labor productivity, combined with growth in the labor force. The extent to which the full productive capacity of the economy is employed depends on aggregate demand comprising expenditures by the private and public sectors on consumption and investment, and international trade. LowGrow SFC incorporates numerous features related to the overall environmental performance of the economy. For instance, a key focus of the model is on climate change and GHG emissions. LowGrow SFC includes a sub-model of the electricity sector. This is useful for assessing the economic and environmental effects of a transition to renewable sources of electricity and the widespread electrification of the economy.

One of the most important elements in the model relates to investments undertaken to reduce environmental impact. Key to a future in which economies reduce the burden they place on the biosphere is a shift from "brown" investment in activities that increases environmental impacts to "green" investment in activities that reduces them. Some green investments such as cost-competitive, energy-efficient equipment can add to the productive capacity of the economy. Other green investments such as seawalls built to protect coastlines from rising sea levels, protect productive capital but do not add to it, reducing the rate of economic growth. A further consideration that is important for determining macroeconomic outcomes is whether green investment is "additional" or "non-additional." Additional green investment increases total investment expenditures whereas "non-additional" green investment simply displaces brown investment without adding to total investment. Only additional green investment increases GDP.

LowGrow SFC generates the values of many variables relevant to an assessment of the performance of the economy. Seven of these are incorporated in a Sustainable Prosperity Index (SPI) for evaluating the scenarios:

- GDP per capita (more relevant to well-being than GDP).
- The rate of unemployment.
- The ratio of government debt to GDP.
- The ratio of household debt to household net worth.
- The Gini coefficient on household incomes (a measure of inequality).

- The average hours of paid work.
- The Environmental Burden Index (EBI).

The EBI is designed to capture the environmental impacts of economic activity notably absent from GDP. While comprehensive in scope, the EBI lacks specificity other than with respect to GHG emissions. Both the SPI and EBI are preliminary and could be improved with better data. However, they share with GDP the redeeming feature that they emerge from a model of the system in whose performance we are interested and so can be used to measure the effect of measures designed to make the system work better.

Scenarios for the Canadian Economy

Former British Prime Minister, Margaret Thatcher, insisted that "there is no alternative" to the market economy and economic growth. She was mistaken. There are in fact many alternatives. We describe three scenarios for the Canadian economy. None of them is a prediction of the future. Rather they are intended to illustrate some of the possibilities facing Canada, to inform discussion and debate, and to suggest the kinds of choices available, not just in Canada but in other advanced economies. The three scenarios are summarized in Table 1.

The Base Case scenario is a benchmark against which other scenarios can be compared. It is a description of what would happen, broadly speaking, at the national level, if current trends continue through and beyond mid-century. The GHG Reduction scenario includes several measures specifically to reduce GHG emissions: a price on GHG emissions from the electric power sector rising from $0 in 2017 to $300 per ton over 15 years, gradual electrification of road and rail, and substantial expenditures to reduce GHG emissions from other sources. In addition to these GHG reducing measures, the Sustainable Prosperity scenario includes more green investment aimed at reducing a wider set of environmental impacts, transfer payments rising to $20 billion per year to reduce inequality and poverty, a shorter work year and a slower rate of population growth.

Figure 1 shows GDP per capita for the three scenarios. Under the Base Case, per capita GDP increases from $50,000 in 2017 to $97,000 in 2067, with an average growth rate of 1.3%. (Unless otherwise stated, values are in 2007$.) This is essentially a conventional, growth-based view of the future, in which the economy as a whole (taking into account population growth of around 44%) increases 2.8 times by the year 2067. The GHG Reduction scenario has a somewhat lower average growth rate in GDP per capita of 1.1%, with incomes in 2067 achieving a level of $87,000 per year. In this scenario, some of the green investment is "unproductive" in the sense that it does not add to the capital stock with which labor produces output. Hence, labor productivity is not as high as in the Base Case and so neither is GDP. The

TABLE 1 Three Scenarios for Canada 2017–2067

Scenario	Main Features
1. Base case	Scenario 1 • Continuation of current trends and relationships
2. GHG reduction	Scenario 1 plus: • Carbon price on GHG emissions from electricity generation • GHG abatement by non-electric industrial sources • Electrification of road and rail transport
3. Sustainable prosperity	Scenario 2 plus: • Switch from brown to green investment • Increased transfer payments to reduce income inequality and reduce poverty • Lower rate of population growth • Reduced average hours worked

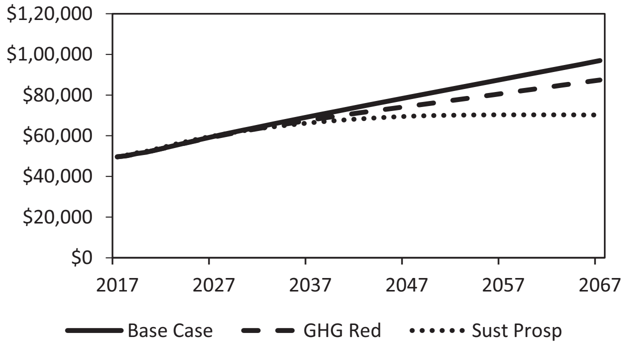


FIGURE 1 GDP per capita in the LowGrow SFC scenarios.

projected reduction in GDP growth is at the high end of the range of published estimates.[41] It also runs counter to the view that a green economy grows faster than a brown one.[42]

The most marked difference, however, is between the Base Case and the Sustainable Prosperity scenario. The latter illustrates a stabilization of per capita income at a level above current income levels. Specifically, the GDP per capita in 2067 is $70,000, an average annual increase of 0.7% over the period. More significantly, GDP and GDP per capita are essentially stable over the final 20 years of the scenario because of the stabilization of population. This scenario thus illustrates a transition from a growth-based economy to an economy managing without growth. The lower rate of economic growth and ultimately its cessation altogether result from the reduced investment in brown capital and the consequential lower increase in labor productivity combined with a reduction in average hours of paid work. Conventional wisdom suggests that such a transition toward what is effectively a steady-state economy is impossible without causing irreparable damage to prosperity and well-being in society. But Figure 2 suggests that this undesirable outcome can be avoided. In fact, the composite SPI described above rises significantly in the Sustainable Prosperity scenario despite falling in both the other two scenarios. Starting from a base of 100 in 2017, the SPI falls precipitously by more than 50% in the Base Case. Even in the GHG Reduction scenario, the SPI declines 10%. In the Sustainable Prosperity scenario, by contrast, the SPI increases 35% from 2017 to 2067.

To understand these differences, all of the components of the SPI must be considered. One of these is GDP per capita, which tends to push the SPI upwards, the higher the level of GDP. Principal among the factors that favor the Sustainable Prosperity scenario over the Base Case is the EBI, which includes the negative impact of GHG emissions and other environmental pressures which are significantly reduced in the Sustainable Prosperity scenario. The EBI for the Base Case more than triples over

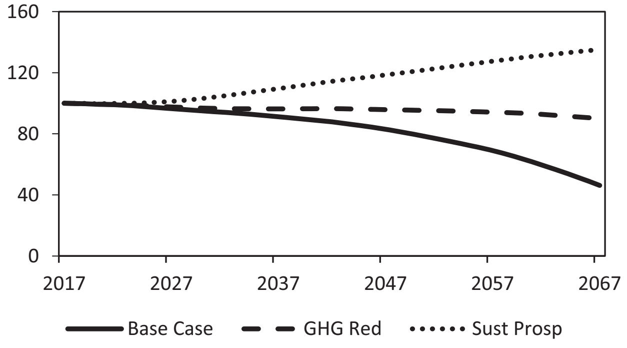


FIGURE 2 SPI in the LowGrow SFC scenarios.

the period of the scenario, as GHGs continue to rise and little is done to offset other environmental impacts from the economy. The EBI for the GHG Reduction scenario performs significantly better. The main reason for this is a significant decline in GHG emissions, which suppresses the rise in the EBI and in doing so has a notably positive effect on the SPI as compared with the Base Case. However, because of the continued expansion of economic output in the GHG Reduction scenario, the reduction in GHG emissions while substantial falls far short of the Canadian government's target of 80% reduction by 2050 from 2005 levels. By comparison, the Sustainable Prosperity scenario more than meets this target. Some of the decline in the EBI, in particular for the Sustainable Prosperity scenario, is also due to the deliberate policy of shifting the electricity sector toward renewable energy sources induced by the carbon price on GHG emissions. The rest of the decline comes from the lower level of economic activity in this scenario.

Two specific social measures adopted in the Sustainable Prosperity scenario also contribute to the improved performance of this latter case over the other two scenarios. The first of these is the redistributive fiscal policy in which transfer payments are progressively increased from 2020 and distributed preferentially to the lower-income categories. Also in the Sustainable Prosperity scenario, annual average hours of paid work decline from more than 1,700 h/yr in 2017 to less than 1,400 h/yr in 2067. Fewer hours working offers more opportunities for people to enjoy time with their families and friends, perhaps volunteering in the community or taking advantage of increased leisure, all of which contribute to people's well-being and quality of life and help increase the SPI.

LowGrow SFC keeps track of the financial flows among sectors as well as sector balance sheets. From this information, two measures of financial prudence included in the SPI are generated by the model: government debt and household loans to value (i.e., net worth). In the Base Case and GHG Reduction scenarios, government debt to GDP increases slightly but more so in the Sustainable Prosperity scenario where it rises from 55% to 70% by the end of the run. This has a negative impact on the SPI. Nonetheless, even at the end of the run, the debt-to-GDP ratio remains at a level that has been far surpassed by many countries without the collapse of their economies reaching 250% in Japan in 2016.

With government running a deficit, it follows from the SFC of the model and the financial behaviors of the other sectors that the overall net lending position of the household sector is positive in all three scenarios, leading to a healthy position in terms of household net worth. There are minor increases in the ratio of household loans to net worth in all three scenarios, with the smallest being in the SP scenario.

In summary, the results discussed in this section suggest that the Sustainable Prosperity scenario remains a realistic alternative to the conventional wisdom of continual exponential growth, outperforming the Base Case in several important ways over the next half a century. Even the financial indicators of a low-growth economy can, under the right conditions, remain relatively sustainable. Investment portfolios change, productivity growth declines, consumption expenditures stabilize, but the economy is nonetheless still financially resilient, its social outcomes improve, and its environmental burden on the planet is dramatically reduced.

Recognizing that the results described here are exploratory, it is fair to conclude that they support the following three conclusions:

1. The pursuit of economic growth at the expense of a deepening environmental crisis has a very high probability of catastrophe.
2. Substantial reductions in GHG emissions can clearly be achieved without massive changes to the structure of society. But the impact on the rate of growth, while modest, is larger than others have suggested and at odds with those who see a "green" economy as growing faster than a "brown" one. Furthermore, this "green growth" scenario falls well short of the Government of Canada's GHG target reduction of 80% by 2050.
3. Sustainable prosperity is attainable, but it will require a major reorientation of society's priorities toward improvements in social equity, economic security, and environmental quality. These changes may well lead to low- or no-growth economy, but they will also deliver a better quality

of life. The result may not be entirely incompatible with capitalism, but it will look very different from the over-financialized consumer capitalism of the early 21st century and may well be worthy of a different name altogether.

Policy Directions for a Low-/No-Growth Economy

The achievement of sustainable prosperity will require some very significant changes in the economy and in society at large. Policy and policy-related measures that drive the Sustainable Prosperity scenario include the following:

- *Investment*: a shift from brown to green investment and from private to public goods through changes in taxation and expenditures.
- *Labor force*: stabilization through changing age structure of the population and population stabilization.
- *Population*: stabilization through a declining fertility rate and changes to immigration policy.
- *Poverty and inequality*: trickle down replaced with focused antipoverty programs that address the social determinants of illness and provide more direct income support.
- *Technological change*: slower, more discriminating, preventative rather than end-of pipe, through technology assessment and changes in the education of scientists and engineers.
- *Government expenditures*: a declining rate of increase until stabilization is reached.
- *Trade*: a neutral net trade balance and diversification of markets.
- *Workweek*: shorter, more leisure through changes in compensation, work organization and standard working hours, and active market labor policies.
- *Environment and energy*: a comprehensive environmental program of pricing, regulation, green investment, land use planning, and education to achieve a rapid transition from fossil fuels to renewable sources of energy, significant reductions in material and energy throughput, and a reversal of habitat destruction conversion.
- *Consumption and lifestyles*: more public goods and fewer positional (status) goods through changes in taxation and marketing.
- *Localization*: fiscal and trade policies to strengthen local economies.

The next two sections look more closely at two specific policy areas in relation to the low/no-growth scenario: strategies for full employment and funding government programs.

Economic Growth and Employment

In 1960, the UN World Economic Survey stated that “the reinterpretation of the objective of full employment under the United Nations Charter to embrace the goal of economic growth marks a second fundamental change in public policy thinking.”[43] This statement from the UN is based on the insight derived from the early work on economic growth by Harrod, Domar, and others that if aggregate expenditure required for full employment in the short run expands the productive capacity of the economy, further increases in aggregate expenditure will be required in the future if full employment is to be maintained. This relationship between growth and employment is accentuated if the size of the labor force is increasing as well.

Equation 1 expresses the relationship between GDP, productivity, the labor force, and unemployment:

$$\text{GDP} = P(1-u)L \tag{1}$$

where GDP is the real gross domestic product, P is the productivity (real GDP per employed person), L is the labor force (employed plus unemployed persons), and u is the unemployment rate (unemployed/labor force).

Between 1971 and 2016, Canadian real GDP grew 234%, productivity increased by 50%, the labor force rose by 125%, and the unemployment rate increased from 6.2% to 7.1%. While the increase in GDP had a positive impact on employment (and vice versa), it was more than offset by the increase in productivity (P). The net effect was an increase in the rate of unemployment (u).

This is a classic dilemma. A growing economy stimulates employment, an increase in productivity reduces it. How can the advantages of increased productivity be realized in an economy that is not growing without causing high unemployment? One way is to reap the benefits of increased productivity as more leisure rather than more goods and services. This can be accomplished by reducing the average number of hours worked by an employed person so that unemployment for a few becomes more leisure for the many. If more people worked fewer hours, it should be possible to have high levels of employment without relying on economic growth.

Some counterfactuals from the past show how this could work. From 1971 to 2016, the average hours worked per year by a Canadian employee decreased by 10.9%. If the decrease in average hours worked had been 13.8%, the rate of unemployment would have been 4% not 7.1% in 2016 given the same increases in GDP and the labor force. At an average of 1647 hours of work per employed Canadian, employees in Canada would still have been working about the same number of hours per year as the average employee in Sweden and the United Kingdom and more than the average employee in seven other OECD countries, in some cases substantially more. Had there been no decrease in the average hours worked between 1971 and 2016, the rate of unemployment would have been 17.3% not 7.1% for the same increases in GDP, productivity, and the labor force.

These calculations show that the average length of the work year, which includes vacation days, can have a marked impact on the rate of unemployment. By spreading the same amount of work among a larger number of employees, the unemployment rate can be lowered and the relationship, as shown by the above examples, is strong. For this reason, researchers have examined the potential for reductions in the average number of hours worked per employee to contribute to full employment. From the standpoint of managing without growth, the benefits of increased productivity would be experienced as increases in leisure and reduced impacts on the environment rather than as increases in output, consumption, and environmental impacts.

A shorter average work year is one of the factors included in the Sustainable Prosperity scenario illustrated in Figures 1 and 2. Over the 50 years of the scenario, the work year declines by 28% to 1400 in 2067. This compares with levels already approached or surpassed in 2015 in Belgium (1427), the Netherlands (1420), Norway (1408), and Germany (1370). In general, European countries have been more proactive than Canada and the United States in reducing the working time as an instrument of employment policy. During the 2008/2009 recession, some countries, Germany for example, mitigated the impact on employment by relying more heavily on reductions in work time.

The arithmetic of reducing the rate of unemployment by reducing the average hours each employed person works is compelling. Achieving such gains in employment in the real world is another matter, but in a review of studies of the employment effects of working time reductions, Bosch finds that most show a gain of “25–70 percent of the arithmetically possible effect.”[44] Bosch has examined the European experience, and the six conditions he identified as particularly important for the success or failure of this policy are summarized in Table 1. He points out that the general political conditions must be suitable for a policy of reducing work time to reduce unemployment. There must be acceptance from employees, trade unions, and employers and support of the State.

Looking at working time policy in the future, Bosch concludes that “shorter working hours are an indicator of prosperity.”[44] They have been in the past, though more recently we have seen the emergence of a sector of the labor force that is “overemployed,” working long hours and “failing to achieve a desired balance in their lives between paid work, family life, personal, and civic time.”[45] These are usually men with higher levels of education in management positions. Simultaneously, there are people who are underemployed and poorly paid, more often than not women. These circumstances contribute to and accentuate rising income inequality.

TABLE 2 Policies for Reducing the Workweek

1. Wage compensation—"If working time reductions and pay increases are negotiated as a total package, then the compensatory increase for the working time reduction can be offset by lower pay rises." This could become more difficult with no or low growth.
2. Changes in work organization—"Larger reductions in working time generally have to be accompanied by changes in work organization"; otherwise, firms will rely on overtime and the employment effects will not materialize.
3. Shortages of skilled labor—"An active training policy is an indispensable supplement to working-time policy" to ensure that there are people with the necessary skills to pick up the slack when skilled workers reduce their hours.
4. Fixed cost per employee—Such as benefits paid on a per-employee basis rather than an hourly basis are an obstacle to reducing working hours because it is costly to employers. Canada shares with most Western European countries the practice of financing statutory social programs through contributions that are usually a proportion of earnings or through taxation, minimizing this fixed cost problem.
5. The evolution of earnings—"The decreasing rate of real wage rises in most industrialized countries has reduced the scope for implementing cuts in working time and wage increases simultaneously." This would be a serious obstacle unless there is widespread support for seeking prosperity without growth though it can be mitigated by a more equal distribution of income. "One fundamental precondition for the working time policy pursued in Germany and Denmark, for example, was a stable and relatively equal earning distribution".
6. The standardization of working hours—Any reduction in standard working hours must strongly influence actual hours worked. If it merely generates more overtime for those already with jobs, it will fail to increase employment. Work reorganization will be required to allow more flexibility in hours worked.

Source: Summarized from Bosch.[44]

Layard in his work on economics and happiness concludes "that people over-estimate the extra happiness they will get from extra possessions" because of habituation. "The required correction is towards lower work effort and thus lower consumption."[46] This means that a shorter work year would not only contribute to reducing unemployment but may also increase the general level of happiness for employees who find themselves better off working fewer hours, for less income and consuming at lower levels.

Funding Public Services in a Low-/No-Growth Economy

Economic growth provides government with increasing resources without increasing tax rates. In times of rapid growth, receipts from corporation profits taxes, personal income taxes, and value-added taxes tend to increase faster than the economy as a whole, allowing governments to provide more services, invest more in infrastructure, redeem outstanding debt, reduce tax rates, or some combination of all these. Governments welcome these circumstances. They have as much to gain from economic growth as anyone. How might this be different in an economy that eschewed economic growth as a policy objective?

Insight into this matter can be gained by considering in some detail the Sustainable Prosperity scenario in which economic growth slows and eventually ceases. In particular, we look at the projected ratio of government debt to GDP, the government's net lending position, and government expenditure in total and per capita. We have already seen that in this scenario, the ratio of government debt to GDP rises from 55% to 75% over the 50-year simulation, which though not necessarily desirable, is well within the bounds of viability based on the experience of other countries and in Canada as well. It is interesting to note here that modern money theorists advise against the use of the debt-to-GDP measure as an indicator of long-run resilience, on the grounds that in countries with sovereign monetary systems such as Canada, the United Kingdom, and the United States, the state does not have a budget constraint comparable to that of a household.[45] The argument is that government can always pay debts denominated in their own currency. In the Sustainable Prosperity scenario, government borrowing continues while GDP stabilizes leading to a rise in the debt-to-GDP ratio. Another reason for thinking that this is unlikely to be a significant problem is that government borrowing is projected to decline, approaching zero by the end of the simulation. This evolution of the government's net lending position coincides, as it must under SFC, with a decline in lending by the other sectors of the economy.

Government expenditures on consumption and investment are assumed in the model to be constant proportions of GDP, modified by countercyclical expenditures to moderate increases in unemployment above a target rate. These expenditures represent the utilization by government of the output of the economy. In the Sustainable Prosperity scenario, the sum of these government expenditures increases gradually from 25% in 2017 to 29% by 2040, remaining at that level as GDP stabilizes. As well as these expenditures, government supplements those with low incomes and supports selected business activities. These "transfer" payments affect the incomes of non-government sectors to pay for expenditures, which are included in GDP in household consumption and private sector investment. It would be double-counting to also include them in government expenditure in the calculation of GDP. However, they are included in the calculation of government net lending.

In the Sustainable Prosperity scenario, government expenditure per capita stabilizes about 70% higher than in 2017 in real, inflation-adjusted terms. This substantial increase should allay any concern of insufficient government funding for services in the absence of economic growth.

Conclusion

There are many reasons for considering how rich economies might manage without growth: growth rates are slowing down, biophysical constraints to continued growth are becoming more apparent, mounting evidence indicates that higher incomes do not make people happier beyond a level of per capita incomes far surpassed in rich countries, economic inequality continues to rise, and despite decades of substantial economic growth many social and environmental problems remain. It is past time to set aside contributions to economic growth as a criterion for assessing measures to improve well-being for humans and other species over the long term and to replace it with the pursuit of sustainable prosperity in all its various dimensions.

Acknowledgments

This chapter is based on Victor, P.A. *Managing without Growth: Slower by Design, Not Disaster*, 2nd edition 2019 and in particular, Chapter 11 which is co-authored with Tim Jackson.

References

1. Maddison, A. *The World Economy: A Millennial Perspective*; OECD: Paris, 2001.
2. World Bank. *Poverty and Shared Prosperity, Piecing Together the Poverty Puzzle*; World Bank Group: Washington, DC, 2018.
3. Alvaredo, F. et al., *World Inequality Report 2018*; World Inequality Lab: Paris, 2018.
4. Krausmann, F.; Gingrich, S.; Eisenmenger, N.; Erb, K.-H.; Haberl, H.; Fisher-Kowalski, M. Growth in global materials use, GDP and population during the 20th century. *Ecol. Econ.* **2009**, *68* (10), 2696–2705.
5. Ekins, P.; Hughes, P. et al. *Resource Efficiency: Potential and Economic Implications. A Report of the International Resource Panel*; UNEP: Paris, 2016.
6. Rockstrom, J. et al. A safe operating space for humanity. *Nature* **2009**, *24* (461), 472–475.
7. Sorrell, S. *Global Oil Depletion*; Technology and Policy Research Centre: London, 2009.
8. Ayres, R.U.; Warr, B. *The Engine of Economic Growth*; Edward Elgar: Cheltenham, 2009.
9. Coulomb, R.; Dietz, S.; Godunova, M.;Nielsen, T.B. '*Critical Minerals Today and in 2030: An Analysis for OECD Countries', Environment Working Paper No. 91*; ESRC Centre for Climate Change Economics and Policy, Grantham Research Institute on Climate Change and the Environment: London, 2015.
10. Gordon, R.J. *The Rise and Fall of American Growth: The US Standard of Living since the Civil War*; Princeton University Press: Princeton, NJ, 2016.

11. Jackson, T.; Victor, P. Productivity and work in the ‘green economy. Some theoretical reflections and empirical tests’. *Environ. Innov. Soc. Trans.* **2011**, *1*, 101–108
12. Jackson, T. The post-growth challenge: Secular stagnation, inequality and the limits to growth. *Ecol. Econ.* **2019**, *156*, 236–246.
13. Arndt, H.W. *The Rise and Fall of Economic Growth: A Study in Contemporary Thought*; Longman Cheshire: Melbourne, 1978. 13.
14. Dinda, S. Environmental Kuznets curve hypothesis: A survey. *Ecol. Econ.* **2004**, *49* (4), 431–455.
15. Schnieder, F.; Kallis, G.; Martinez-Alier, M. Crisis or opportunity? Economic degrowth for social equity and ecological sustainability. Introduction to this special issue. *J. Cleaner Prod.* **2010**, *18* (6), 511–518.
16. Galbraith, J.K. *The Affluent Society*; Houghton Mifflin: Boston, MA, 1958.
17. Mishan, E.J. *The Costs of Economic Growth*; F.A. Praeger: New York, 1968.
18. Beckerman, W. *Why we need economic growth. Lloyds Bank Rev.* **1971**, *102*, 1–15.
19. Beckerman, W. *In Defence of Economic Growth*; J. Cape: London, 1974.
20. Beckerman, W. *Small Is Stupid*; Duckworth: London, 1995.
21. Schumacher, E.F. *Small Is Beautiful: Economics As If People Mattered*; Harper and Row: London, 1973.
22. Boulding, K.E. The economics of the coming spaceship earth. In *Environmental Quality in a Growing Economy*; Jarrett, H., Ed.; Johns Hopkins Press: Baltimore, MD, 1966, 3–14.
23. Georgescu-Roegen, N. *The Entropy Law and the Economic Process*; Harvard University Press: Cambridge, MA, 1971.
24. Daly, H. *Steady-State Economics: The Economics of Biophysical Equilibrium*; W.H. Freeman: San Francisco, CA, 1977.
25. Meadows, D.H.; Meadows, D.L.; Randers, J.; Behrens III, W.W. *The Limits to Growth*; Earth Island Limited: London, 1972.
26. Olson, M.; Landsberg, H.H., Eds. *The No-Growth Society*; W.W. Norton and Company: New York, 1973.
27. Jackson, T.; Webster, R., *Limits revisited. A review of the limits to growth debate.* A report to the APPG on limits to growth published under Creative Commons, CC BY-NC-ND 4.0., 2016.
28. Douthwaite, R. *The Growth Illusion*; New Society Publishers: Gabriola Island, 1994 (also, a second edition in 1999).
29. Daly, H. *Beyond Growth: The Economics of Sustainable Development*; Beacon Press: Boston, MA, 1996.
30. Booth, D.E. *The Environmental Consequences of Growth: Steady-State Economics*; Routledge: London, 1998.
31. Hamilton, C. *Growth Fetish*; Allen and Unwin: Sydney, 2003.
32. Booth, D.E. *Hooked on Growth: Economic Addictions and the Environment*; Rowman and Littlefield: Lanham, MD, 2004.
33. Victor, P.A. *Managing without Growth: Slower by Design, Not Disaster*; Edward Elgar Publishing: Camberley, 2008 (2nd edition, 2019).
34. Speth, J.G. *The Bridge at the End of the World: Capitalism, the Environment, and Crossing from Crisis to Sustainability*; Yale University Press: New Haven, CT, 2008.
35. Brown, P.; Carver, G. *Right Relationship*; Berrett-Koehler Publishers Inc.: San Francisco, CA, 2009.
36. Jackson, T. *Prosperity without Growth*; Earthscan: London, 2009 (2nd edition, 2017).
37. Schor, J. *Plenitude: The New Economics of True Wealth*; The Penguin Press: New York, 2010.
38. Dietz, R.; O’Neill, D. *Enough is Enough*; Berrett-Koehler Publishers: San Francisco, CA, 2013.
39. Czech, B. *Supply Shock: Economic Growth at the Crossroads and the Steady State Solution*; New Society Publishers: Gabriola Island, 2013.
40. Von Weizsäcker, E.; Wijkman, A. *Come On! Capitalism, Short-Termism, Population and the Destruction of the Planet—A Report to the Club of Rome*; Springer: New York, 2018.

41. Ekins, P. Ecological modernisation and green growth: Prospects and potential. In *Handbook on Growth and Sustainability*; Victor, P.A.; Dolter, B. Eds.; Edward Elgar Publishing: Cheltenham and Northampton, MA, 2017, 107–137.
42. Bassi, A., *UNEP GER Modeling Work. Technical Background Material, V.4*; Millennium Institute: Arlington, VA, 2011.
43. United Nations Development Programme. *Human Development Report 2006*; UNDP: New York, 2006.
44. Bosch, G. Working time reductions, employment consequences and lessons from Europe. In *Working Time: International Trends, Theory and Policy Perspectives*; Golden, L., Figart, D.M., Eds.; Routledge: London, 2000, 177–211, p. 192.
45. Figart, D.M.; Golden, L. Introduction and overview, understanding working time around the world. In *Working Time: International Trends, Theory and Policy Perspectives*; Golden, L., Figart, D.M., Eds.; Routledge: London, 2000, 1–17.
46. Layard, R. Happiness and public policy: A challenge to the profession. *Econ. J.* **2006**, *116*, C24–C33.

9

Food–Energy–Water Nexus

Nemi Vora

Introduction

The United Nations established 17 distinct goals (referred to as the sustainable development goals or SDGs) for the 2030 agenda for global sustainable development ranging from climate action to economic growth. It has been noted that the agenda can only be achieved successfully if their interconnections are recognized and incorporated into planning (Nilsson, Griggs, and Visbeck 2016). For instance, SDG 2 outlines ending hunger, providing nutrition, achieving food security, and promoting sustainable agriculture. It is closely related to SDG 14 of life below water and SDG 15 of life on land, which in turn requires planning for quality and quantity supply of water (SDG 6), renewable, affordable energy (SDG 7), and sustainable consumption and production (SDG 12). The concept of interconnections between Food–Energy–Water (referred to as the FEW nexus) is one such piece to a much larger puzzle of managing for the SDGs. Recently, international and national funding calls on the FEW nexus (National Science Foundation, National Science Foundation (U.S.–China), European Commission 2015) have mobilized research around the topic. These organizations recognize the challenges and benefits of managing the FEW nexus, as individual management has often resulted in unintended consequences.

The connections between food, energy, and water resources are apparent in daily lives, although numerous examples exist of policies focusing on managing one resource but resulting in straining the others deeply connected with them. One such an example is the well-intended, but damaging irrigation policies in Gujarat two decades ago. Gujarat, a relatively arid state in western India, saw a rise in agriculture productivity through investments in irrigation infrastructure, land reforms, and advances in seed technology since the 1980s (Mathur and Kashyap 2000). Due to erratic precipitation patterns, groundwater irrigation became central to the success of Gujarat's farmers. The government supported groundwater use through a series of initiatives including access to cheap electricity for pumping water out (Narula et al. 2011). However, decades of excessive pumping coupled with low recharge levels resulted in severe groundwater depletion, posing a threat for the future of agriculture in the region. As the water levels declined, farmers pumped deeper, putting undue pressure on the power grid and the government for continuing to subsidize expensive electricity (Narula et al. 2011). Thus, managing electricity for promotion of agriculture without considering impacts on water use resulted in severe and long-lasting consequences. In 2012, a drought in eastern India resulted in farmers turning to groundwater for irrigation. Sudden excessive pumping and lack of water for power plants caused a massive grid failure. Comparable situations were reported during California's recent 7-year drought (Webber 2015). From an energy policy perspective, the promotion of biofuels from food and oilseed crops, and their effects on water, land, and biodiversity have been a center of debate and contention for a while (Searchinger et al. 2008). The context of these issues may be different, but they demonstrate the potential tradeoffs that may occur if resources are managed without regarding their connections.

The concept of the FEW nexus is not new; similar calls for integrated water resource management (IWRM; an approach to promote simultaneous development of water and associated resources for social and economic benefits) date back to 1962 (Lloyd 1963). Similarly, the integrated natural resource management concept was put forward to couple agrarian objectives with ecosystem services (Twomlow, Love, and Walker 2008). In 2011, Hoff et al. highlighted the concept of "nexus approach" to simultaneously target food, energy, and water security for the background paper at the Bonn Conference (Hoff 2011). With the FEW nexus, a clearly defined scope of focusing on three essential resources was put forth. In the scientific literature, many variations of the term FEW nexus exist with studies coining the terms Energy–Water–Food (EWF) nexus, Water–Energy–Food (WEF) nexus, Climate–Land–Energy–Water (CLEW nexus), etc.; however, the underlying message of the interlinkages remains the same.

Defining the Food–Energy–Water Nexus

There is no single agreed-upon definition of FEW nexus in the literature, apart from recognizing the FEW nexus as linkages. It often makes the scope of the work more complex as numerous connections exist between these three resources. For example, growing food requires water for agriculture and energy to employ machinery on farm and transport food. Global agriculture is the largest consumer of water amounting to 69% of total withdrawals, while food supply chain accounts for 30% of global energy consumption (Dubois 2011). Water is integral in generating energy from renewable and fossil resources alike: water is required in thermoelectric power plants, and for hydraulic fracturing to obtain natural gas. Renewable resources, such as hydropower, and biofuels production also heavily depend on water availability. Energy is

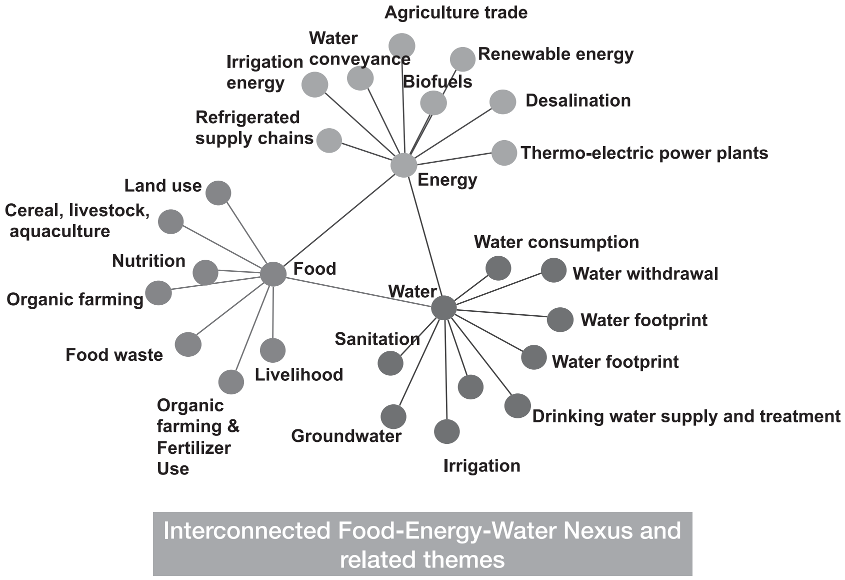


FIGURE 1 Common research and policy themes under the umbrella of Food–Energy–Water nexus.

required to withdraw, treat, and convey water for municipal, agriculture, and industrial use. Desalination, an important alternative for fresh water in water-stressed countries, is also highly energy intensive.

While still relevant to the nexus, some studies are often not included in the FEW nexus if they focus on less than three resources (i.e. food–water, energy–water), or if they go beyond the three to include climate, livelihoods, and biodiversity in their analysis. Additionally, the focus of FEW nexus studies is often on physical interactions and environmental concerns alone, while the crucial social component is overlooked. It has also been argued that calling a study "a nexus approach" does not increase its scientific merit, and therefore FEW nexus does not add anything new to the literature (Wichelns 2017). Here, I adopt the perspective that establishing universal criteria for inclusion in the FEW nexus is not as important as recognizing the existence of interconnections, both hidden and visible, between resources, and resources and the rest of the environment. Thus, managing one resource requires effort to identify and evaluate the influence on/of other resources.

Admittedly, merely recognizing the interconnections may not simplify their management, it does initiate a discourse. In the aforementioned case of Gujarat, the issue of expensive subsidies and groundwater depletion was tackled by separating and tightly rationing farm electricity supply from the rest of the rural areas, while providing superior quality supply on a pre-announced schedule. The "Jyotigram (electrified village)" scheme was instrumental in overhauling Gujarat's rural electrification while reducing groundwater overdraft and boosting non-farm rural economy (Shah et al. 2008). Using the irrigation–energy nexus to tackle not only energy-water tradeoffs but also to benefit rural economy has been touted a win–win scheme and is now being implemented across other Indian states.

The scientific community has largely agreed upon managing resources by recognizing their interconnections, and it has been widely acknowledged to adopt a systems perspective in tackling the complex issue (Bazilian et al. 2011, Liu et al. 2018). Systems analysis recognizes that real-world issues do not strictly arise from a linear chain of cause and effect but are rather a result of complex feedback loops between individual components. As such, it recognizes that a system is often greater than the sum of its parts. Thus, rather than focusing on individual components, systems analysis examines how each component interacts with other components and the external environment. The result of these multiple interactions, often unobservable by studying just one component, can be understood through a holistic perspective. In a more practical sense, this is done by employing a collection of models and assumptions from different disciplines to represent a simplified version of reality. This approach helps identify potential side effects and address the issue of considering biophysical and socio-economic interactions together. The following section briefly talks about already existing models and methods used in the FEW nexus context.

Quantitative Analysis of the FEW Nexus

Much has been written regarding the FEW nexus and models to quantify linkages (Bazilian et al. 2011, Newell, Goldstein, and Foster 2019). As the research under the wide umbrella of FEW nexus studies is highly interdisciplinary, a series of methods and their combinations have been used to model one or more FEW interactions. Studies have used methods such as life cycle assessment, network analysis, footprint approach, optimization, input–output analysis, computable general equilibrium models, and hydrological models to model FEW interactions. These methods are adopted from disparate disciplines including ecology, social science, probability and statistics, thermodynamics, and hydrology. In their review articles, Albrecht et al. and Newell et al. provide a detailed list enumerating the studies of various methods that have been used to tackle FEW challenges (Albrecht, Crootof, and Scott 2018, Newell, Goldstein, and Foster 2019). Therefore, instead of repeating such a list, a few examples are provided here that show the use of interdisciplinary methods to help answer specific FEW issues at different geographical scales.

Vora et al. used social network analysis combined with environmental life cycle assessment to assess energy and greenhouse gas emissions associated with irrigated food trade in the United States (Vora et al. 2017). Davis et al. developed an optimization framework using spatially explicit data and crop water use models to assess redistribution of crops at an international scale with the goal to reduce global water use, without compromising nutritional benefits or loss of crop diversity (Davis et al. 2017). Ramaswami et al. developed a framework around cities as emerging demand centers and quantified FEW nexus impacts associated with meeting current demand from cities (Ramaswami et al. 2017). Using regression models, Herrera-Estrada et al. assessed the effect of drought on air emissions from electricity sector with a focus on regions heavily relying on hydropower and thermoelectric power plants (Herrera-Estrada et al. 2018). Keskinen et al. utilized a combination of mathematical models pertaining to hydrology and climate to understand tradeoffs between new hydropower development and its effect on soil fertility and local agriculture in Mekong Basin (Keskinen et al. 2015). Not all of the studies cited here framed their work explicitly within the FEW nexus contexts, nevertheless their focus on inter-linkages makes them nexus studies. The key takeaway is that there is no singular method and discipline that can provide a comprehensive quantification of FEW linkages, but requires interdisciplinary approach spanning both natural and social systems.

While these studies piece together different methods to develop a FEW nexus frameworks, there are also already existing models that help quantify some of these linkages. Pollitt et al. provide a detailed assessment of large-scale models along with their drawbacks and advantages (Pollitt et al. 2010), and Bazilian et al. provide a brief list of sectoral and integrated models and methods specifically pertaining to the FEW nexus (Bazilian et al. 2011). Some of the notable and widely used models include Climate, Land-Use, Energy and Water Strategies (CLEW) by International Atomic Energy Agency (Howells et al. 2013), Global Agro Ecological Zones (GAEZ) by Food and Agriculture Organization (Fischer et al. 2012), and the Integrated Solutions for Water, Energy, and Land (ISWEL) project that links multiple open-source models such as MESSAGE for energy modeling and GLOBIOM for land-use modeling (Willaarts et al. 2018).

Method and Model Selection

As there is no dearth of different methods and models that can be used for FEW nexus modeling, selecting an appropriate one depends on several factors: (1) scope and objective of the work, (2) availability and quality of data, (3) appropriate spatial scale, and (4) usefulness of analyzing long-, medium- and short-term temporal trends. Table 1 lists each factor and relevant questions for choosing appropriate models and methods.

TABLE 1 Decision Factors Influencing Model and Method Selection for the FEW Nexus

Scope	• Goal of the study (e.g. new policy formulation, infrastructure decisions, disaster preparedness) • Already known information about resource linkages (e.g. groundwater levels affecting irrigation, fossil heavy grid) • Defining system boundary of the work (e.g. production-based impacts vs. consumption impacts, inclusion/exclusion of other resource connections)
Data	• Availability of open data and time/money budget if none available (i.e. budget for purchasing data or gathering yourself) • Sensitivity of the chosen method to the size of data (e.g. certain methods may only work for larger datasets) • Quality of the datasets and quantification of uncertainty (precision of data collection methods, survey methodologies used, representativeness of data) • Appropriate assumptions and relying on substitutable data to fill data gaps (e.g. using national averages when local not available) • Age of the data (are 10-year old datasets appropriate to represent the current situation)?
Spatial and temporal scales	• Spatial scale of the study (local, metropolitan, state/province, basin level) • Integrating multiple datasets and models available for different spatial scale (e.g. hydrology models developed at basin scale vs. energy models for specific electricity markets and grids) • Available data reconciliation methods and their complexity (e.g. downscaling aggregate global data to local through cross-entropy, spatial estimation) • Relevance of long-term impacts on study results (e.g. climate-induced change on water availability, increase in crop yields) • Usefulness of incorporating historical trends

Scope

Although FEW nexus is circular, there is generally a particular resource of interest that is established (e.g. reservoir water management for agriculture, municipal, and energy supply) and subsequent linkages are then assessed. If there are previous assessments and known information about historical connections between resources, it should be incorporated in the study. For example, decision on upgrading irrigation technologies should consider impact on groundwater levels. Additionally, depending on the work, it should be determined whether the study should be limited to FEW nexus linkages alone or FEW nexus-everything studies that consider economy-wide impacts as well (Lant et al. 2018). This relates to setting a system boundary to contain the complexity of the system. A system boundary delineates impacts and resources to be considered, and those that will be out of the scope for the work. However, system boundary should be selected carefully as to not leave out relevant and important impacts. This can be done based on expert judgment, literature review of similar case studies, and use of screening tools such as economic input–output life cycle assessment (Hendrickson et al. 1998).

In the FEW nexus discourses, choosing a systems boundary is particularly important due to the connectedness of the products and services through global supply chains. Global trade and goods exchange are driven by competitive advantage and endowment of resources between regions, domestic and international policies and agreement, and politics. Depending on the production practices and environmental regulations in the production region, trade can alleviate or increase environmental impacts associated with consumption (Dalin and Rodríguez-Iturbe 2016). Studies have shown that domestic policies promoting human and environmental health have driven production out of the country causing environmental damages elsewhere (Nesme et al. 2016, Plevin et al. 2010). Therefore, careful consideration needs to be made to assess whether distant impacts should be incorporated as to not under/overestimate the resource use and impacts. FEW nexus studies have used the concept of embodied (virtual) resource/impacts trade to incorporate the role of global supply chains and connect distant locations (Kastner, Erb, and Haberl 2014, Konar et al. 2016). Here, the embodied trade refers to not the physical trade but indirect use of resources and resulting impacts from the production of specific goods and services in distant locations. Finally, the scope of the analysis may need to be adjusted depending on whether there are existing models that can be used to establish and quantify linkages or new models need to be built from scratch and required investment of time, effort, and interdisciplinary expertise.

Data

Freely available quality data are a necessity for carrying out insightful quantitative analysis and modeling. Numerous international agencies (UN Statistics division, International Energy Agency, Food and Agriculture Organization, World Resource Institute, etc.) and country-specific agencies (US Department of Agriculture, US Geological Survey, Government of India Planning Commission, Statistics Canada, etc.) collect and host data (see Data.Gov, data.gov, data.europa.eu, data.un.org for a single point of access to disparate databases) pertaining to food supply, agriculture, energy use, trade, water withdrawal statistics. Data collection is a time-consuming and money-intensive process, and therefore necessary data may not be available for a specific year, a time period, or a specific location. To this end, appropriate assumptions and substitutions may need to be made to fill in data gaps. These assumptions can be made with the help of scientific principles such as mass–energy balance, engineering equations, or based on peer-reviewed literature. A common method for data substitutions is based on similarity, be it geographical or technological: for example, substituting lack of data for one developed country with another developed country with similar gross domestic product (GDP).

However, many of the datasets are usually limited at country level and/or by time scale. Therefore, success of any integrated management approach hinges on data being available at an appropriate spatial scale and for a longer time frame. Therefore, new data collection and compilation techniques are needed to analyze systems at specific scales with recent/current data and for impoverished countries that may not be able to spend money on census surveys. Currently, the research work is driven toward bypassing traditional government surveys and using novel techniques such as satellite imagery, remote sensing, and cellphone data to gain important insights on living conditions (Azzari, Jain, and Lobell 2017, Blumenstock, Cadamuro, and On 2015). Although, more work is needed before such methods can replace the need for traditional surveys. Finally, methods such as sensitivity analysis and uncertainty quantification should be used to help convey quality of the data. Sensitivity analysis is useful in understanding which parameters affect the results the most and indicates where data collection improvements can be made. Uncertainty quantification helps in conveying variability and reliability of the results. Due to limited scope, these concepts are mentioned here in brief. However, the open-source life cycle assessment textbook from Matthews et al. is a good source for introductory material on conducting both types of analysis (Matthews, Hendrickson, and Matthews 2015).

Spatial and Temporal Scales

As shown with aforementioned examples of quantitative studies, the spatial scale for FEW analyses has varied from local to international. However, the appropriate scale in which these systems should be analyzed for effective policymaking and decisions is still ambiguous. While local challenges of drought and crop failure remain, the global supply chains have extended the system boundary and made it more fragile as distant events can cause cascading failures. Marston et al. noted in their analysis that depletion of major ground water aquifers in the United States would not only impact domestic food consumption but also affect distant countries such as Japan relying on food exports from these regions (Marston et al. 2015). Boundaries for water management can also be unclear as interconnectivity between water basins can go beyond political boundaries. There are numerous examples of water conflicts within a country such as water wars in the Western United States between different states (Tory 2018) and international issues such as the Brahmaputra river basin spanning China, India, Bangladesh, and Bhutan (Yang et al. 2016). The interconnectedness of economic and political tensions and extreme weather events to energy supply has been evident in numerous examples of energy crisis across the world (Hamilton 2011). Thus, the connections between places are not only based on political or physical boundaries but also through common pool of resources, infrastructure, and social systems. To integrate this complexity in analysis will require redefining how distant FEW systems are connected as regulations may stop at political boundaries, but the actual environmental impact may be larger beyond the traditional perceived boundaries. Recently, Lant et al. (2019) discussed the need of "mesoscale" analysis to capture differences that cannot be captured by analysis at two extremes of local scale vs. international scale (Lant et al. 2019). Based on the chosen temporal aspect (long term vs. short term), climate narratives provided by the shared socio-economic pathway projections and national climate assessments can be incorporated.

There is also ambiguity regarding the best approach to analyze such complex interactions: specifically pertaining to a bottom-up approach that builds up from a smaller system or a top-down approach that looks at interactions between different sectors and economy. We need integration of both approaches (top-down, bottom-up) and multiple spatial and temporal scales as decisions are made at various levels from household and local community level actions to national and international policies and agreement. As one size does not fit all, analyzing such complex systems will require both multi-sector and multilevel actions.

The Criticism and Cautions for the Food–Energy–Water Nexus

A few have criticized the overuse of the word FEW nexus indicating that while it is undisputable that these resources are interconnected, the phrase is ambiguous in terms of prescribing a framework in tackling these issues (Cairns and Krzywoszynska 2016, Wichelns 2017). The critics point out that by setting such arbitrary bounds on only looking at three resources, the FEW nexus misses the opportunity to consider other factors such as human health, livelihood, and farm chemicals. However, as mentioned earlier, the study of FEW nexus should be looked at in the much larger context of assessing connections between SDGs. While FEW nexus has indeed become a buzzword (Cairns and Krzywoszynska 2016), the interest garnered can be successful in raising awareness at both policy and stakeholder levels. Additionally, having a general umbrella under which such case studies, policies, and research can be compiled provides a huge opportunity to mobilize large-scale funding, exchange knowledge, and support interdisciplinary and interagency collaboration.

An important consideration that nexus proponents should keep in mind is that when managing for tradeoffs, one interest lobby may be stronger than the others and argue for their interest over an optimal solution (Jensen 2013). Furthermore, the economic cost, time, and expertise required for an integrated approach will be much higher than traditional silos approach and would involve substantial work toward capacity building, making the decision process lengthy (Liu et al. 2018). Additionally, an integrated approach may not be the panacea as considering too many factors at once may make a system unsolvable or at best provide a solution that may not be agreeable to all.

Conclusion

It is important to note that beyond the three resources, the usefulness of the concept lies in the term "nexus thinking." A nexus approach involves addressing the interlinkages and feedback between different systems, so synergies are promoted and tradeoffs minimized. However, in some instances, even a FEW nexus approach may not provide optimal outcomes and overlook unintended consequences on interdependent systems such as human livelihood or biodiversity. Therefore, a context-specific analysis that prioritizes local stakeholders' benefits should be adopted in policymaking with an overarching goal of achieving the sustainable development agenda. Going forward, the FEW nexus literature should explicitly recognize that there is no one suitable method, model, or spatial/temporal scales to conduct such complex analysis. But interdisciplinary multi-scale effort is needed to understand the issues in entirety. Good quality open data is at heart of such analysis and future work and studies should focus on novel data collection and data treatment approaches that take care of bridging the spatial and temporal resolution gaps, conducting hybrid analysis that incorporate both top-down and bottom-up approaches, and adopt/develop methods that provide a robust statistical framework for empirical data analysis.

Despite its drawbacks, FEW nexus has been successful in mobilizing the scientific and policy community to go beyond the "silos thinking" approach and recognize the need for interdisciplinary work. By incorporating three large policy and scientific communities of agriculture, hydrology, and energy analysis, FEW nexus has also merged expertise within these communities, with studies formulating novel frameworks and techniques to overcome issues of data availability, spatial and temporal scope, and model integration. FEW nexus has also drawn the scientific communities' attention back to SDG interlinkages and driven the policy community to engage in quantitative prospective analysis before making large-scale decisions. The penetration of FEW nexus in humanities has also incorporated more dialogue through stakeholder engagement. Finally, the primary takeaway lesson from the FEW nexus for environmental management is to adopt holistic approach and recognize that there may be hidden or visible links connected to the system being managed, and therefore, these links need to be identified and incorporated in the management plan.

References

Albrecht, Tamee R., Arica Crootof, and Christopher A. Scott. 2018. "The water-energy-food nexus: A systematic review of methods for nexus assessment." *Environmental Research Letters* 13 (4):043002.

Azzari, George, Meha Jain, and David B. Lobell. 2017. "Towards fine resolution global maps of crop yields: Testing multiple methods and satellites in three countries." *Remote Sensing of Environment* 202:129–141.

Bazilian, Morgan, Holger Rogner, Mark Howells, Sebastian Hermann, Douglas Arent, Dolf Gielen, Pasquale Steduto, Alexander Mueller, Paul Komor, and Richard S.J. Tol. 2011. "Considering the energy, water and food nexus: Towards an integrated modelling approach." *Energy Policy* 39 (12):7896–7906.

Blumenstock, Joshua, Gabriel Cadamuro, and Robert On. 2015. "Predicting poverty and wealth from mobile phone metadata." *Science* 350 (6264):1073–1076.

Cairns, Rose, and Anna Krzywoszynska. 2016. "Anatomy of a buzzword: The emergence of 'the water-energy-food nexus' in UK natural resource debates." *Environmental Science & Policy* 64:164–170.

Dalin, Carole, and Ignacio Rodríguez-Iturbe. 2016. "Environmental impacts of food trade via resource use and greenhouse gas emissions." *Environmental Research Letters* 11 (3):035012.

Davis, Kyle Frankel, Maria Cristina Rulli, Antonio Seveso, and Paolo D'Odorico. 2017. "Increased food production and reduced water use through optimized crop distribution." *Nature Geoscience* 10 (12):919–924. doi: 10.1038/s41561-017-0004-5.

Dubois, Olivier. 2011. *The State of the World's Land and Water Resources for Food and Agriculture: Managing Systems at Risk*. Earthscan: London.

European Commission. 2015. "Integrated approaches to food security, low-carbon energy, sustainable water management and climate change mitigation Horizon 2020 funding call." https://cordis.europa.eu/programme/rcn/664566/en.

Fischer, Günther, Freddy O. Nachtergaele, Sylvia Prieler, Edmar Teixeira, Géza Tóth, Harrij Van Velthuizen, Luc Verelst, and David Wiberg. 2012. *Global Agro-Ecological Zones (GAEZ v3.0) – Model Documentation*. IIASA; FAO: Laxenburg; Rome.

Hamilton, James D. 2011. *Historical Oil Shocks*. National Bureau of Economic Research: Washington, DC.

Hendrickson, Chris, Arpad Horvath, Satish Joshi, Octavio Juarez, Lester Lave, H. Scott Matthews, Francis C. McMichael, and Elisa Cobas-Flores. 1998. "Economic input-output-based life-cycle assessment (EIO-LCA)." *Mental*. https://www.researchgate.net/profile/H_Matthews/publication/242142910_Economic_Input-Output-Based_Life-Cycle_Assessment_EIO-LCA/links/0c96053b2c2bd76615000000.pdf

Herrera-Estrada, Julio E, Noah S. Diffenbaugh, Fabian Wagner, Amy Craft, and Justin Sheffield. 2018. "Response of electricity sector air pollution emissions to drought conditions in the western United States." *Environmental Research Letters* 13 (12):124032.

Hoff, Holger. 2011. *Understanding the Nexus. Background Paper for the Bonn 2011 Conference: The Water, Energy and Food Security Nexus*. Stockholm Environment Institute: Stockholm.

Howells, Mark, Sebastian Hermann, Manuel Welsch, Morgan Bazilian, Rebecka Segerström, Thomas Alfstad, Dolf Gielen, Holger Rogner, Guenther Fischer, and Harrij Van Velthuizen. 2013. "Integrated analysis of climate change, land-use, energy and water strategies." *Nature Climate Change* 3 (7):621.

Jensen, Kurt Mørck. 2013. "Viewpoint–swimming against the current: Questioning development policy and practice." *Water Alternatives* 6 (2):276–283.

Kastner, Thomas, Karl-Heinz Erb, and Helmut Haberl. 2014. "Rapid growth in agricultural trade: Effects on global area efficiency and the role of management." *Environmental Research Letters* 9 (3):034015.

Keskinen, Marko, Paradis Someth, Aura Salmivaara, and Matti Kummu. 2015. "Water-energy-food nexus in a transboundary River Basin: The case of Tonle Sap Lake, Mekong River Basin." *Water* 7 (10):5416–5436.

Konar, Megan, Jeffrey J. Reimer, Zekarias Hussein, and Naota Hanasaki. 2016. "The water footprint of staple crop trade under climate and policy scenarios." *Environmental Research Letters* 11 (3):035006.

Lant, Christopher, Jacopo Baggio, Megan Konar, Alfonso Mejia, Benjamin Ruddell, Richard Rushforth, John L. Sabo, and Tara J. Troy. 2018. "The US food–energy–water system: A blueprint to fill the mesoscale gap for science and decision-making." *Ambio*:1–13.

Lant, Christopher, Jacopo Baggio, Megan Konar, Alfonso Mejia, Benjamin Ruddell, Richard Rushforth, John L. Sabo, and Tara J. Troy. 2019. "The US food–energy–water system: A blueprint to fill the mesoscale gap for science and decision-making." *Ambio* 48 (3):251–263.

Liu, Jianguo, Vanessa Hull, Hugh Charles, Jonathan Godfray, David Tilman, Peter Gleick, Holger Hoff, Claudia Pahl-Wostl, Zhenci Xu, Min Gon Chung, and Jing Sun. 2018. "Nexus approaches to global sustainable development." *Nature Sustainability* 1 (9):466.

Lloyd, Emlyn Howard. 1963. "Design of water-resource systems; new techniques for relating economic objectives, engineering analysis, and government planning." *Journal of the Royal Statistical Society: Series A (General)* 126 (3):480–481.

Marston, Landon, Megan Konar, Ximing Cai, and Tara J. Troy. 2015. "Virtual groundwater transfers from overexploited aquifers in the United States." *Proceedings of the National Academy of Sciences* 112 (28):8561–8566.

Mathur, Niti, and Surendra P. Kashyap. 2000. "Agriculture in Gujarat: Problems and prospects." *Economic and Political Weekly* 35:3137–3146.

Matthews, H. Scott, Chris T. Hendrickson, and Deanna H. Matthews. 2015. "Life cycle assessment: Quantitative approaches for decisions that matter." Retrieved June 1, 2016. https://www.lcatextbook.com/

Narula, Kapil Kumar, Ram Fishman, Vijay Modi, and Lakis Polycarpou. 2011. "Addressing the water crisis in Gujarat, India." Columbia Water Center White Paper, March 2011.

National Science Foundation. 2018. "Innovations at the nexus of food, energy and water systems (INFEWS)." https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505241.

National Science Foundation (U.S.-China). "Innovations at the nexus of food, energy, and water systems (INFEWS: U.S.-China)." https://www.nsf.gov/pubs/2018/nsf18096/nsf18096.pdf.

Nesme, Thomas, Solène Roques, Geneviève S Metson, and Elena M. Bennett. 2016. "The surprisingly small but increasing role of international agricultural trade on the European Union's dependence on mineral phosphorus fertiliser." *Environmental Research Letters* 11 (2):025003.

Newell, Joshua Peter, Benjamin Paul Goldstein, and Alec Foster. 2019. "A 40-year review of food-energy-water nexus literature with a focus on the urban." *Environmental Research Letters* 14: 073003.

Nilsson, Måns, Dave Griggs, and Martin Visbeck. 2016. "Policy: Map the interactions between sustainable development goals." *Nature News* 534 (7607):320.

Plevin, Richard J., Andrew D. Jones, Margaret S Torn, and Holly K. Gibbs. 2010. "Greenhouse gas emissions from biofuels' indirect land use change are uncertain but may be much greater than previously estimated." *Environmental Science & Technology* 44 (21):8015–8021.

Pollitt, Hector, Anthony Barker, Jennifer Barton, Elke Pirgmaier, Christine Polzin, Stephan Lutter, Friedrich Hinterberger, and Andrea Stocker. 2010. *A Scoping Study on the Macroeconomic View of Sustainability. Final Report for the European Commission, DG Environment, SERI (Sustainable Europe Research Institute)*. Cambridge Econometrics. http://seri.at/wp-content/uploads/2009/12/SERICE-2010_Macroeconomic-view-of-sustainability.pdf.

Ramaswami, Anu, Dana Boyer, Ajay Singh Nagpure, Andrew Fang, Shelly Bogra, Bhavik Bakshi, Elliot Cohen, and Ashish Rao-Ghorpade. 2017. "An urban systems framework to assess the transboundary food-energy-water nexus: Implementation in Delhi, India." *Environmental Research Letters* 12 (2):025008.

Rothberg, Daniel. 2018. "States accuse Arizona water agency of gaming Lake Mead, undermining Colorado River drought plans." *The Nevada Independent*, April 17, 2018.

Ruddell, Benjamin L., Elizabeth A. Adams, Richard. Rushforth, and Vincent C. Tidwell. 2014. "Embedded resource accounting for coupled natural-human systems: An application to water resource impacts of the western US electrical energy trade." *Water Resources Research* 50 (10):7957–7972.

Searchinger, Timothy, Ralph Heimlich, Richard A. Houghton, Fengxia Dong, Amani Elobeid, Jacinto Fabiosa, Simla Tokgoz, Dermot Hayes, and Tun-Hsiang Yu. 2008. "Use of US croplands for biofuels increases greenhouse gases through emissions from land-use change." *Science* 319 (5867):1238–1240.

Shah, Tushaar, Sonal Bhatt, Ramesh K. Shah, and Jayesh Talati. 2008. "Groundwater governance through electricity supply management: Assessing an innovative intervention in Gujarat, western India." *Agricultural Water Management* 95 (11):1233–1242.

Tory, Sarah. 2018. "A Southwest water dispute reaches the Supreme Court." *High Country News*, January 23, 2018.

Twomlow, Stephen, David Love, and Sue Walker. 2008. "The nexus between integrated natural resources management and integrated water resources management in southern Africa." *Physics and Chemistry of the Earth, Parts A/B/C* 33 (8–13):889–898.

Vora, Nemi, Apurva Shah, Melissa M. Bilec, and Vikas Khanna. 2017. "Food-energy-water nexus: Quantifying embodied energy and GHG emissions from irrigation through virtual water transfers in food trade." *ACS Sustainable Chemistry & Engineering* 5: 2119–2128.

Webber, Michael E. 2015. "A puzzle for the planet." *Scientific American* 312 (2):62–67.

Wichelns, Dennis. 2017. "The water-energy-food nexus: Is the increasing attention warranted, from either a research or policy perspective?" *Environmental Science & Policy* 69:113–123.

Willaarts, Barbara, Simon Langan, Juraj Balkovic, Peter Burek, Edward Byers, Arnulf Deppermann, Stefan Frank, Matthew Gidden, Peter Greve, and Petr Havlik. 2018. *Integrated Solutions for Water, Energy and Land Progress Report 3*. United Nations Industrial Development Organization (UNIDO) and International Institute for Applied Systems Analysis (IIASA): Laxenburg.

Yang, Y. C. Ethan, Sungwook Wi, Patrick A. Ray, Casey M. Brown, and Abedalrazq F. Khalil. 2016. "The future nexus of the Brahmaputra River Basin: Climate, water, energy and food trajectories." *Global Environmental Change* 37:16–30.

10

Geographic Information System (GIS): Land Use Planning

Egide Nizeyimana
and Jacob Opadeyi

Introduction

Agricultural scientists are required to provide information needed to address land degradation and land use conflicts confronting the world today. As the population increases and land becomes a commodity in many parts of the world, careful planning of the use of land must be undertaken to accommodate conflicting people's needs and preserve and/or protect the environment. The decision about the use of land must be made based on analyses of each potential use in terms of its economic and biophysical suitability to the specific tract of land and possible impact to environmental degradation.

Geographic information system (GIS) and related technologies (e.g., remote sensing, global position system) have proven to be a valuable tool in land use planning activities. The GIS approach is important in this area because it provides functions to capture, store, organize, and analyze spatially referenced data. Moreover, GIS has been coupled to a variety of models (e.g., crop productivity, hydrology, and water quality simulations) and may be an important component of spatial decision support systems (SDSS) and land resource information systems (LRIS). GIS enhances model flexibility and efficiency in these systems where it is often regarded as a centralized data analysis, management, and planning system. As a result, decision makers and land use planners in private companies, universities, and at various levels of the government are using GIS to develop spatial environmental databases, perform land evaluations, and analyze and manage resources. There is no doubt that the use of GIS and GIS-based systems in land use planning activities will continue to increase in the future, as more detailed digital environmental datasets become available and the capability of computers to handle large volumes of data increases.

GIS Use in Land Use Planning Activities

The goal of land use planning is to make decisions about the use of land and resources.[1] Its implementation is often driven by future people's needs in terms of productivity and environmental sustainability. Land use planning is important in highly populated communities primarily due to conflicts between competing uses and interests of users.[2] In this case, planning activities are tailored to make the optimal use of the limited land resources. In general, the land use planning involves sequentially an organization of thoughts and establishment of long-term goals, a land evaluation that includes appraising alternatives, and finally designing and implementing the plan. Land evaluation is the most important aspect of this process.

The information within a GIS consists of a spatial component represented by points (e.g., well locations), lines (e.g., streams, road networks), or polygons (e.g., soil delineations) and attribute data or information that describes characteristics of the spatial features. The spatial entity is referenced to a geographic coordinate system and is stored either in a vector or raster format.[3] GIS is primarily used in the development of spatial databases and land evaluation, a procedure in the land use planning process that is aimed at determining the suitability of land units to current and alternative uses and the potential impact of each on the environment.

Development of Spatial Databases

GIS stores, retrieves, and allows efficient manipulation of database information. It provides powerful analysis and relational database facilities to modify and/or integrate spatial data from different sources and resolutions as well as advanced visualization functions to display output data in the form of interpretative maps. In this case, land attributes and qualities are derived from geographic databases and used to determine land suitability, limitations, or ratings for various land use types. The analysis results may be presented in tabular or graphical form and are intended to provide key information necessary for land users or planners to make meaningful decisions about land management and conservation and/or land use planning.

Site Suitability Analyses

A site suitability analysis typically involves the assessment of the level of affinity a specific land has for a particular use. Soil information available in soil databases is rarely enough for site evaluations. In addition to soils, the analysis often integrates local information on landforms, current land uses, the relative location of the land, and associated social and political restrictions. The proposed use may have additional limitations that should be taken into account. For example, an analysis for suitable sites for land application of sewage sludge should consider the physical, chemical, and biological properties of the waste in soil and water.[4,5]

GIS is efficient in identifying and ranking sites for various land use planning activities. First, site-specific analyses often require many and detailed data sources. Second, GIS overlay features, logical operations, and display functions are tailored to speed up data processing and therefore allow efficiently suitability class assignment and graphical display of results. A good example is the use of GIS for locating appropriate sites for forestland application of sewage waste.[6] The authors derived physical site suitability ratings for an area in Vermont based on EPA guidelines[7] and merged them with social and political restrictions of the state and counties to derive a land applicability classification. Similar GIS-based approaches have been used to locate sites for solid waste disposals.[8]

Linking GIS and Models

The spatial databases and associated attribute data described above may be part of a GIS/model graphical user interface (GUI). These models may be those that determine land suitabilities or those that simulate water quality for environmental impact assessments. As an entity of GIS/model interface, GIS allows easy access to database attributes by various algorithms, statistical software, and environmental models for a variety of land use analyses. In addition to model parameterization, a full integration of GIS and models also allows the user to interact with various modules, select data input and module utilities, and display graphical and/or tabular representations of modeling results.[9] The use of this type of interface reduces significantly the processing time and resources required to develop input data and run the model. GIS/model linkage has been accomplished for field- and watershed-scale hydrology/water quality models such as the leaching and chemistry estimation (LEACHM)[10] and agricultural non-point source pollution (AGNPS)[11] models. Results of these analyses are used to support management and land use planning activities in the farm or watershed.

The GIS/model interface may also be part of an SDSS. In addition to modeling parameterization, SDSS offer the users with functions to evaluate different land use scenarios necessary for making management recommendations and/or land use planning decisions. Results may be used by farmers to adjust management practices of distinct fields in a given farm or may be used by field officers to help farmers set priorities while providing technical assistance for nutrient management planning. Figure 1 shows a flow diagram indicating data integration in a GIS with models for water quality assessment in watersheds.

GIS has been incorporated in many land evaluation systems so that planners and public officials can take advantage of its spatial modeling and visualization capabilities. An example of such a system is the USDA–NRCS land evaluation and site assessment (LESA).[12] The land evaluation portion of LESA determines soil productivity levels, farm size and agricultural sales volume; the site assessment portion deals with factors such as location, amount of nonagricultural land, zoning restrictions, etc.[13] GIS, in GIS-based LESA systems, is used in both modules and at all levels of the analysis.

Development of Land Resource Information Systems

A number of land information delivery systems have been developed in recent years to make spatial data available to users for application in various aspects of land use planning. These systems, also called LRIS, are multipurpose systems that integrate geographic databases and GIS tools to analyze, record, report, and display spatial data relationships. Some of these systems have been embedded into the World Wide Web (WWW) for quick and easy access and analysis via Internet browsers.

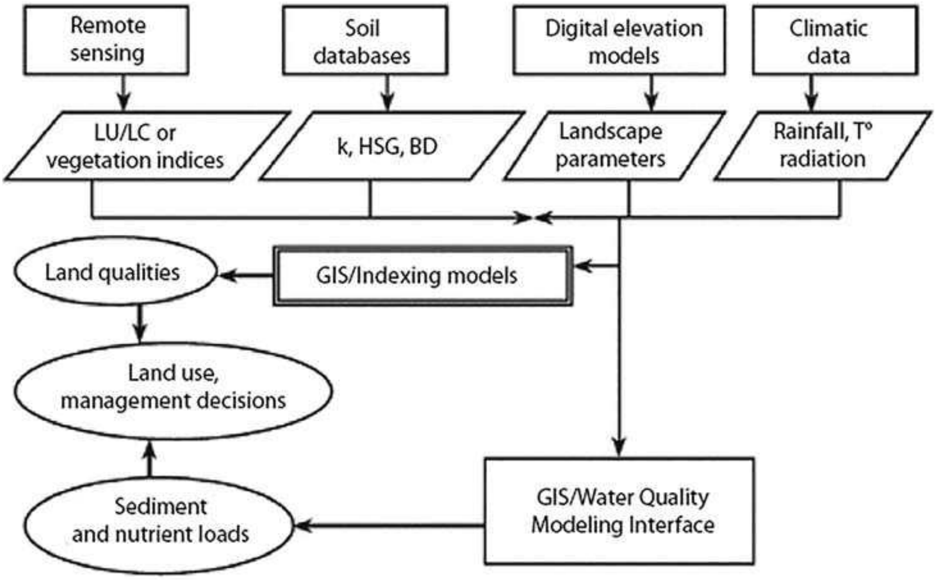


FIGURE 1 Integration of GIS, spatial database, and models.

Reliability of GIS-Based Analysis Results

GIS and GIS-based systems used in spatial data analyses for land use planning have grown in recent years. GIS is particularly attractive in these areas because it allows overlay of spatial data sets and the merging and analysis of attribute data from different sources. The resulting data are obtained using data from digital and paper maps of different scales or acquired at different resolutions such as in the case of digital elevation models (DEM) and remote sensing imagery. The combination of such data layers may produce unrealistic data and consequently lead to erroneous predictions. The question is how reliable are these results when used for developing land use and management plans. A discussion on the propagation of error and uncertainty in GIS-based systems is provided elsewhere.[14]

The accuracy of GIS-based land evaluations is a function of the quality of attribute data and mapping delineations of databases, and the type of model or assessment scheme used in the analyses. Various algorithms for assessing the quality of GIS analysis results as affected by error and uncertainty in GIS layers, and their propagation are part of popular GIS and image analysis software (e.g., Arc/Info, ERDAS Imagine). The results, which are often in the form of reliability diagrams are not used by average GIS practitioners. This is probably due to the fact that they are not easy to understand or interpret. Similarly, models vary depending on how each represents various processes of the system. Lumped-parameter models and indexing/ ranking schemes are mostly used in land evaluations because they are easy to parameterize. However, these models ignore spatial variations of parameters throughout the field, watershed, or region of study. Furthermore, models originally designed for fields and watersheds are often applied to regional analyses, thus adding some level of uncertainty in modeling predictions. For example, most land use planning programs use conventional methods of land evaluation. Each land parameter is given a range of values with corresponding ratings showing its suitability to crop production. These indexes are added or multiplied to create a single index that is to rank land units. The method is simple, but carries a high uncertainty because breaks between two ranges or ranks are subjective.

The effect of map scale and resolution on environmental assessment and modeling output data has been subject to many studies. Raster-based GIS systems require that a grid cell size be defined prior to the analysis. However, as pixel size increases above the resolution of the original data, the spatial variability decreases. This causes a decrease of the predictive power of generated input parameters particularly for small land areas.

Conclusions

GIS has been used primarily in land evaluation, a procedure in the land use planning process that deals with determining land suitability to current and alternative uses and the potential impact of each use on the environment. GIS and GIS-based systems are often considered as integrated spatial information systems (ISIS) for data analysis needed in land use management and planning activities. Despite the advantages of GIS outlined above, the land use planner should be aware of the error and uncertainty in GIS and GIS-based analyses resulting from digitizing and scaling inaccuracies, data conversion between vector and raster formats, and others. Finally, GIS implementation in land use planning activities can be expensive. In addition to costs associated with hardware and software purchase and maintenance, a high level of technical expertise is required to perform complex modeling tasks in land evaluations for alternative uses and to sustain databases. Nonetheless, the demand for GIS and GIS-based analysis systems in land use planning is to increase in the future as more detailed spatial datasets become available.

References

1. FAO. *Guidelines for Land Use Planning*; FAO Development Series 1; Soil Resources, Management and Conservation Service; FAO: Rome, 1993; 96 pp.

2. Brinkman, R. Recent developments in land use planning, with special reference to FAO. In *The Future of the Land: Mobilizing and Integrating Knowledge for Land Use Options*; Fresco, L.O., Stroosnijder, L., Bouma, J., van Keulen, H., Eds.; Wiley: Chichester, 1994; 13–21.
3. Burrough, P.A. Matching databases and quantitative models in land resource assessment. *Soil Use Manage.* **1989**, *5*, 3–8.
4. Nizeyimana, E.; Petersen, G.W.; Looijen, J.C. Land use planning and environmental impact assessment using GIS. In *Environmental Modeling with GIS and Remote Sensing*; Skidmore, A., Prins, H., Eds.; Taylor and Francis Ltd.: London, 2002; in press.
5. Petersen, G.W.; Nizeyimana, E.; Miller, D.A.; Evans, B.M. The use of soil databases in resource assessments and land use planning. In *Handbook of Soil Science*; Summes, M.E., Ed.; CRC Press: Boca Raton, FL, 2000.
6. Hendrix, W.G.; Buckley, J.A. Use of a geographic information system for selection of a sites for land application of sewage waste. *J. Soil Water Conserv.* **1992**, *92*, 271–275.
7. U.S. Environmental Protection Agency. *Process Design Manual for Land Treatment of Municipal Waste*; Environmental Protection Agency: Cincinnati, OH, 1981, PB 299655/1.
8. Weber, R.S.; Jenkins, J.; Leszkiewicz, J.J. Application of geographic information system technology to landfill site selection. *Proceedings of the Application of Geographic Information Systems, and Knowledge-Based Systems for Land Use Management*, November 12–14; VPI and State University: Blacksburg, VA, 1990.
9. Petersen, G.W.; Nizeyimana, E.; Evans, B.M. Application of GIS to land degradation assessments. In Methods for Assessment of Soil Degradation; Lal, R., Blum, W.H., Valentine, C., Stewart, B.A., Eds.; CRC Press: Boca Raton, FL, 1997; 377–391.
10. Inskeep, P.P.; Wraith, J.M.; Wilson, J.P.; Snyder, R.D.; Ma-cur, R.E.; Gaber, H.M. Input parameter and model resolution effects on predictions of solute transport. *J. Environ. Qual.* **1996**, *25*, 453–462.
11. Tim, U.S.; Jolly, R. Evaluating agricultural nonpoint source pollution using integrated geographic information systems and hydrology/water quality models. *J. Environ. Qual.* **1994**, *23*, 25–35.
12. Wright, L.E.; Zitzmann, W.; Young, K.; Googins, R. LESA—Agricultural land evaluation and site assessment. *J. Soil Water Conserv.* **1983**, *38*, 82–86.
13. Daniels, T. Using LESA in a purchase of development rights program. *J. Soil Water Conserv.* **1990**, *45*, 617–621.
14. Heuvelink, G.B.M. *Error Propagation in Environmental Modeling with GIS*; Taylor & Francis: London, 1998; 127 pp.

11

Industrial Networks

Sven E. Jørgensen

The Importance of System Networks

An important property known from ecosystems, operation in networks, can be implemented in industries by building a network of several industries. This would facilitate the possibilities to find a better matter and energy (or rather, work energy or exergy) efficiency, understood as the entire efficiency, equal to the sum of all outputs (products) divided by the sum of all inputs. Inputs and outputs may be either matter or work energy (exergy; see the entries about exergy and eco-exergy). The matter can eventually be multiplied by the price or costs of products and of inputs to express matter efficiency in economic terms.

Industrial symbiosis in the form of a network is similar to the ecological network of an ecosystem, i.e., the network consists of flows of energy, matter, and information.

An Example of an Industrial Network

The use of industrial networks has been tested in the Danish town Kalundborg. The network (for further details, see Jørgensen[1]) is shown in Figure 1.

The heart of the system is the Asnæs Power Station (APS), the largest power plant in Denmark. Half of the plant is fueled by coal and half by a new fuel called orimulsion, a bituminous product made from Venezuelan tar sand. The use of orimulsion was introduced in 1998 and gave an 18% reduction in carbon dioxide emission compared with the use of 100% coal but, due to the sulfur content of Venezuelan tar, gave a higher production of gypsum by the sulfur dioxide scrubber. By exporting half of the former waste energy, APS has reduced the fraction of available energy directly discharged by about 80%. Since 1981, the municipality of Kalundborg has eliminated the use of 3,500 oil-fired residential furnaces by distributing heat from the power plant through a network of well-insulated underground pipes. The homeowners have paid for the pipeline but receive cheap reliable heat in return.

The power plant also supplies cooling water that is warmed 7–8°C to an on-site fish farm producing about 200 tons of trout per year. The APS also delivers process steam to its neighbors, Novo Nordisk (a pharmaceutical and biotechnological industry) and Statoil (an oil refinery). The APS produces 70,000 tons of ash that is sold for road building and cement production.

In 1993, APS installed a sulfur dioxide scrubber that produces calcium sulfate or industrial gypsum. Gypsum is the primary ingredient of wallboard, and ASP is the primary supplier of gypsum to the

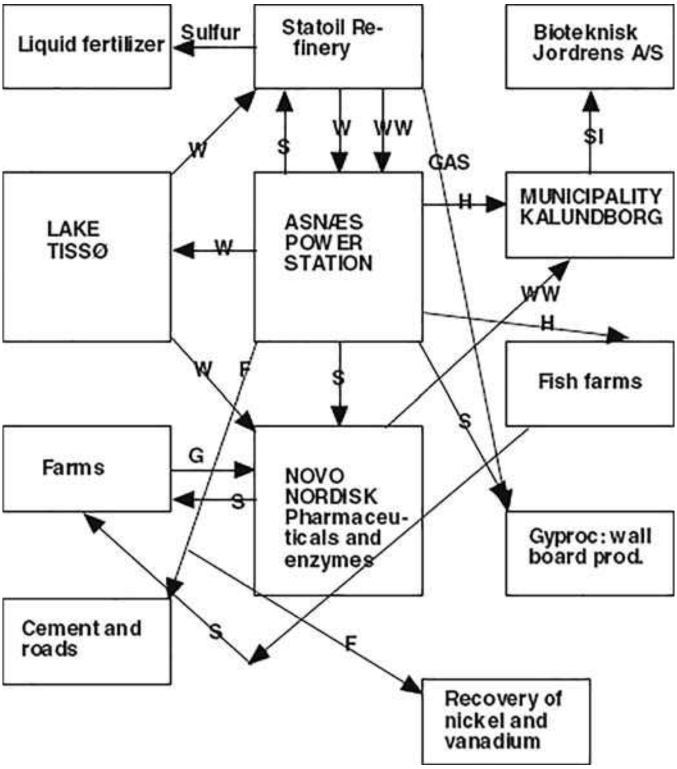


FIGURE 1 The industrial symbiosis at Kalundborg, Denmark. The flows of energy and matter make up an "ecological" network that implies that the overall efficiency of the utilization of the exergy input to the network is high. W, water; F, fly ash; WW, wastewater; S, steam; Sl, sludge; H, heat; G, grains.

wallboard factory Gyproc. The refinery has refinery gas as a waste product that is sold after desulfurization as natural gas to be used in the town or elsewhere. The desulfurization process produces liquid sulfur that is used to produce a liquid fertilizer, ammonium thiosulfate.

Novo Nordisk and its sister firm Novozymes are world leaders in the production of insulin and enzymes. The plant employs 1,000 people, and the products are produced by fermentation processes. Based on agricultural crops, valuable products are produced by microorganisms. The waste is a nutrient-rich sludge that can be used by the nearby farmers as animal feed (the yeast sludge) or fertilizers (the microorganisms sludge). More than 3,000 tons of the two types of sludge are produced daily. The yeast sludge is sold, and the microorganism sludge is given away, due to the firm's concern for disposal security. Distribution of the sludge was the least cost-effective way to comply with environmental regulations.

A new partner, A/S Bioteknisk Jordrens (meaning Biotechnical Soil Purification), joined the industrial symbiosis in 1999. The company uses municipal sludge as a nutrient in a bioremediation process to decompose pollution by toxic substances in contaminated soil.

The Benefits of Industrial Networks

It is clear from a review of the industrial ecological network in Kalundborg that there is an enormous potential to save energy or, rather, work energy or exergy in the use of industrial symbiosis. In this context, the Nordiske Affaldsbørs (Nordic Waste Exchange) could also be mentioned, an Internet page that gives information about wastes produced in the Nordic countries of Finland, Iceland, Norway, Sweden, and Denmark. The idea is that what may be waste for one industry could maybe be used as a raw material for another. For instance, used solvent from the pharmaceutical industry may be applicable for the production of paint or dyestuff. It is very important to see such opportunities if we want to reduce the loss of exergy in the form of fuel or raw-material waste to the environment. The idea behind industrial symbiosis—an ecological-industrial network—has been used in agriculture, but it could be applied much more widely, which would make agriculture much more sustainable.[1]

The industries have already adopted several of the characteristic properties of ecosystems, because it has been profitable for the companies. The discipline behind this application of ecological principles in industries is called industrial ecology.[2] The most significant losses of work energy in the industries are waste and the ineffective use of energy. These losses can be reduced by recycling and reuse, which the industries are using increasingly, but only to the extent that it is profitable. If several companies, industries, and the local municipalities are forming an ecological symbiotic network, it is, however, possible to reduce the overall losses of work energy or exergy and at the same time increase the profit. The possibilities of finding an optimum solution are generally bigger for a large system than for a small system. The experience gained so far from an industrial symbiotic network is so positive that it is surprising that it has not found a wider application.

Changes to the rules, taxes, and legislation for discharge of wastes may also increase the motivation for companies to recycle and reuse.[1] If the industries have to pay for the discharge of wastes or pay more for raw materials, it is a question of economic optimization how much the company should recycle and reuse the waste. The conclusion is clear: A higher direct utilization efficiency of work energy or exergy coupled with increased use of recycling and reuse are the key methods to decrease the eco-exergy losses by industries. Moreover, the use of ecological symbiotic networks increases the possibilities for recycling and reuse.

References

1. Jørgensen, S.E. *Eco-Exergy as Sustainability.* WIT: Southampton, 2006; 220 pp.
2. Ayres, R.U.; Ayres, L.W. *Handbook of Industrial Ecology.* Edward Elgar: Cheltenham, 2002; 680 pp.

12

Land Restoration

Richard W. Bell

Introduction

Land degradation is widespread globally[1–3] and is devaluing the natural capital of land as well as compromising the provision of soil ecosystem services. Arresting the degradation of land should be a high-priority goal. However, restoration of land that has previously been degraded represents an untapped opportunity to increase the natural capital of land and the provision of key soil and plant ecosystem services, particularly those related to plant productivity and water supply. Many forms of disturbance such as overgrazing, excessive tillage, overirrigation lead to land degradation including water erosion, wind erosion, decline of soil structure, acidification, salinization, waterlogging, and decline in soil fertility. Land disturbance after mining is comparatively extreme, but research and adaptive management on these sites have fostered the development of sound principles and practices of land restoration. In the present entry, most examples of land restoration are drawn from mine land restoration to illustrate the principles of the emerging discipline of land restoration. While land restoration can be quite site specific in its practice, there are general principles that can be applied and these have a degree of relevance to most forms of land degradation.

Land as a Finite Resource

For sustainable environmental management, land should be regarded as a finite global resource. The critical surface layer of the finite land resource is the soil profile that varies from a few millimeters' thickness to several meters. While soils are continually forming, the rate is slow, equivalent to 1 mm every 100 years[4] or even less.[5] Land may be created by reclamation of submerged areas along shorelines or in wetlands, and river deltas create new land surfaces as they deposit sediments when they reach the sea or lakes. Locally, these processes that create new land surfaces may contribute significantly to available land resources, but at a global scale, the amounts created are small relative to the total area and to the areas being degraded. Moreover, there is already evidence that in low-lying coastal zones, land is being inundated by rising sea levels or eroded by storm surges associated with rising sea levels. These processes are offsetting gains from natural or anthropogenic land creation. There seems little alternative but to regard land as a finite resource, to slow and prevent degrading processes, and to implement restoration programs on land that has suffered degradation.

Natural Capital and Soil Services

While most emphasis in the past was on food and fiber production from land, the concept of its value is broadening. Land is a form of natural capital that underpins several critical ecosystem services.[6] The soil ecosystem services outlined in Table 1 comprise supporting, regulating, provisioning, and cultural services.[6] The emergence of the concept of natural capital is beginning to redefine the notion of land and its value. Production of food is still a critical function of land, indeed one whose importance is predicted to rise over the next several decades as the challenge of providing sufficient quantity and nutritional quality of food for a rising population becomes more evident.[7,8] In addition, the role of land and soil in regulating major elemental cycles, the hydrological cycle, and the disposal of wastes and dead organic matter is gaining increasing importance.

Land Degradation

Land as a finite resource is susceptible to a range of degrading processes that limit its productivity, its land use potential, and its value in providing ecosystem services. Conacher and Conacher[9] define land degradation broadly as "alteration of all aspects of the natural (biophysical) environment by human actions, to the detriment of vegetation, soils, landforms and water (surface and subsurface, terrestrial and marine) and ecosystems." More broadly, land degradation can be defined as those processes that lower the natural capital of land and compromise the provision of soil ecosystem services.

The processes of land degradation, their measurement, impact, and management have been dealt with else- where.[10–12] Most commonly, the impact of land degradation is described in terms of areas of land degraded.[13] However, this approach has several limitations. Apart from the difficulty of acquiring good quality data that are consistent across large areas to make such estimates, the different impacts of degrading processes makes it impossible to compare processes based only on area affected. For example, the degradation of soil structure by tillage may have an incremental effect on crop yields but cause a quantum increase in water runoff. The degradation of land by salinity, which raises soil salinity levels above a critical threshold for growth of most plants, prevents continuation of current land use options, while opening up others (such as salt-tolerant vegetation). By contrast, soil acidification may have no effect on plants until a critical threshold is crossed, leading to the release of toxic Al^{3+} concentrations that impair plant growth. Hence, the key to predicting the effects of land disturbance is to understand the threshold values of soil properties for plant growth and for

TABLE 1 Societal Soil Ecosystem Services

Supporting	Physical stability and support for plants Renewal, retention, and delivery of nutrients for plants Habitat and gene pool
Regulating	Regulation of major elemental cycles Buffering, filtering, and moderation of the hydrological cycle Disposal of wastes and dead organic matter
Provisioning	Building material
Cultural	Heritage sites, archaeological preserver of artifacts Spiritual value, religious sites, and burial grounds

Source: Adapted from Robinson et al.[6]

ecosystem services. The differences among degrading processes are also critically important in planning the restoration of degraded land.

Alternatives to assessing and comparing only the land areas suffering degradation are to estimate the value of lost production from the degradation or to estimate the cost of restoration. The latter is generally considered to be critical since restoration costs will be a major impediment to reversing the effects of degradation. The cost per hectare for land restoration on mine sites is much larger than that on production lands not only because of the more extreme degradation but also because of the greater capacity to pay for restoration. Costing the restoration works for different forms of degradation and different areas of degraded land provides a basis for prioritization of restoration works. However, costs of restoration commonly only consider land productivity, rather than biodiversity and the full range of ecosystem services. If the natural capital of land in total was considered, there may be less variation in costs of land restoration among different forms of disturbance.

Land Restoration Definition

The term *land restoration* has a generic meaning, first proposed by Hobbs and Norton,[14] to indicate "that restoration occurs along a continuum and that different activities are simply different forms of restoration." The term *land rehabilitation,* as defined by Aronson et al.,[15] is also appropriate when applied to land degraded by mining. In the United States, the term *reclamation* is commonly used to describe activities that replace ecological functions by planting different vegetation to what previously grew.

Land restoration will usually focus on re-establishing ecological functions such as nutrient cycling, hydrological balance, and ecosystem resilience.[16] In some situations, restoring the original flora is in addition a realistic and appropriate goal. A case study based on the application of this goal is outlined below for restoration of eucalyptus forest after bauxite mining. Further expansion of the concept of land restoration should consider its role in the recovery of natural capital and soil ecosystem services.

Land Restoration Principles

The process of land restoration comprises the following: determination of the end land use, definition of the main limiting factors for restoration and means of alleviating them, and, finally, planning and implementation of a restoration program including monitoring and evaluation against success or completion criteria. This entry focuses on the first two components, while the latter component is described in more detail by Hobbs.[16]

End Land Use

Development of a land restoration strategy is generally considered to need a well-defined end land use. The key constraints that need to be alleviated depend on the land use envisaged. Hence, definition of the end land use is often considered to be a prerequisite to the restoration of degraded land. From the defined end land use, it is possible to identify the stakeholders whose interests need to be considered, the scope of the restoration challenge, and the prime constraints that will have to be alleviated.[16,17] It is also a prerequisite for the establishment of the measurable goals and targets for restoration that are used in setting success or completion criteria. End land use will largely define the complexity and difficulty of the restoration task and the costs associated with achieving a successful outcome. At one end of the spectrum, the complete restoration of the pre-existing ecosystem is a challenging goal. It may require decades of systematic research and continuous improvement of the restoration procedures before that goal is achieved and before the success criteria can be defined and validated. On the other hand, simply achieving a stable land surface or creating a pasture suited to low-intensity grazing would be more easily achieved.

Existing land use on surrounding areas, or prior land use at the site, often determine the end land use. However, degradation such as salinization may alter the substrate for plant growth so radically as to make the pre-existing land use impossible or undesirable. Furthermore, on mine sites, the alteration in landforms may require a change in land use in order to achieve land stability. In the case where land use change is necessary, stakeholders will need to be engaged to arrive at an acceptable alternative end land use.

A great diversity of end land uses has been applied during land restoration. Apart from natural ecosystems, agriculture, forestry, nature conservation, grazing, housing, wetlands, amenity and recreational facilities, and waste disposal are all possible options. Sociopolitical and economic factors will generally determine the selection of end land use particularly if the land is in a protected area, on production land, or in a densely settled zone.

While a well-defined end land use is necessary to set realistic goals and targets for restoration, there are risks associated with setting a highly prescriptive end land use if the restoration technology has not been well developed and based on solid research. The development of best practice for restoration of a particular type of degraded land may require decades of research and adaptive management. Hence, the premature setting of the end land use and targets for achievement may lock in an inferior set of outcomes. Using an adaptive management approach will allow continuous improvements in restoration practice to be made and tested. As improved practices and outcomes become possible, new benchmarks for completion can be set and new possibilities for end land use emerge. The bauxite-mining case study below illustrates this process.

Diagnosis of Limiting Factors

Degradation of land takes many forms and is triggered by many agents.[10] The types of constraints and their severity and the consequences for the restoration plan will clearly vary from site to site. Correct diagnosis of the key constraints and identification of likely feasible solutions are a prerequisite for successful restoration of degraded land. Identification of the key agents causing degradation is also essential. At a mine site, the active cause of degradation is mining and its associated disturbance. In other cases, the degrading agent may be tillage, overgrazing, wind erosion, saline groundwater discharge, etc. Apart from the biophysical constraints related to substrate properties, landform, climate, and hydrology, limiting factors may be associated with socioeconomic factors that prevent the degrading processes from being arrested. In the present entry, the focus is on the biophysical constraints to restoration (Table 2).

Biophysical Factors Limiting Land Restoration

Climate

The growing conditions for plants when restoring degraded land are determined primarily by climate. Climate is the main limitation on potential land uses for each site. Species that are indigenous to the site will usually be well adapted to the rainfall, temperature, and extreme conditions that occasionally occur in a given climatic regime such as drought, heat stress, and frost. However, if the substrate is unsuitable for the indigenous species, or the end land use requires a different selection of species such as agricultural species, then the chosen species need to be well adapted to the site climate.

When considering climate-related constraints, it is not just the average conditions that should be examined, but also the frequency and severity of extreme events, such as drought, heavy rain storms (e.g., cyclones), frost, snow, hail, etc. The coincidence of extreme events with commencement of a land restoration project may prevent seed germination or cause the loss of seed and topsoil due to erosion.

TABLE 2 Biophysical Factors Limiting Land Restoration, Their Consequences for Site Stability and Plant Growth, and Common Treatment Methods

Factor	Constraint	Consequence	Treatment
Climate	Drought	Failure of germination, poor emergence or establishment, plant death	Irrigation, drought-tolerant or drought-avoiding species, adjust time of sowing
	High temperature	Poor germination and emergence, plant death	Mulching, retention of crop/plant residue
	Low temperature/ frost	Delayed emergence, plant death, poor seed set	Frost- or cold-tolerant species, adjust time of sowing
	Extreme rainfall/ wind events	Water or wind erosion episodes, loss of seeds	Contour banks, soil cover by foliage, mulch or stubble, windbreaks
Landform	Slope	Land slippage, soil creep, unsafe conditions for machinery operation	Deep-rooted plants, reshape to lower slope angle, contour banks, engineering design, prevent water run-on from upslope
	Runoff	Water loss, sediment loss, downslope deposition	Contour banks, reshape to lower slope angle, improve infiltration and soil water storage
	Exposure	Drought, high winds, extreme temperatures	Tolerant plant species especially as windbreaks
	Aspect	High winds, extreme temperatures	Tolerant plant species especially as shade plants
Hydrology	Runoff	Reduced soil water storage, waterlogging, flooding downslope	Contour banks, increased drainage intensity
	Limited profile water storage	Drought, runoff, increased groundwater recharge	Tolerant plant species, deep ripping, treatment of subsoil chemical constraints
	Groundwater discharge	Water-filled voids, acid mine drainage, waterlogging, salinity	Containment of water, wetland treatment ponds, drainage
Substrate properties	Acidity	Poor plant growth especially roots, nutrient deficiencies, plant death	Lime, acid-tolerant species, P fertilizer
	Alkalinity	Poor plant growth, nutrient deficiencies, plant death	Gypsum, leaching, alkaline tolerant species, acidifying materials
	Salinity	Poor plant growth, plant death	Leaching, salt-tolerant species, drainage
	Sodicity	Soil dispersion, crusting, poor seedling emergence, runoff, water erosion	Gypsum, leaching, organic matter addition
	Nutrient deficiency	Poor plant growth	Fertilizer (mineral and organic)
	Metal toxicity	Poor plant growth, plant death	Lime, tolerant species, burial or capping of substrate, removal of substrate, phytoremediation
	Low water availability	Poor plant growth, plant death	Irrigation, mulching, organic matter, deep ripping, adjust time of sowing, drought tolerant species
	Waterlogging	Poor plant growth, plant death	Drainage, tolerant plant species
	Poorly structured soils	Crusting, poor water holding capacity, poor root growth	Mulching, organic matter, gypsum
	Mycorrhiza	Poor plant growth, nutrient deficiency	Topsoil management, inoculation of nursery plants
	Rhizobium	Nitrogen deficiency	Topsoil management, inoculation, liming acid soils
	Soil microbes	Slow mineralization of soil organic matter	Topsoil management

Source: Modified from Bell.[18]

Landform

Mining creates voids below ground level and waste material that is stacked above ground level. Hence, the overall slope angle of land on a degraded mine site will be increased relative to the pre-existing landform. Hence, the creation of stable, non-eroding surfaces is generally the first major goal of restoration after mining. The key landform factors that limit the achievement of stable surfaces are slope angle, elevation, aspect, and surface drainage. Maximum slope angles are generally set by regulation and vary among jurisdictions. In West Australia, a maximum of 20° is the guideline for restored mine slopes such as the outer surface of waste rock dumps.[19] The change in landform is less of a consideration with other forms of disturbance apart from mining.

Hydrology

Any significant removal or disturbance of vegetation, change in surface soil properties, increase in average slope angle, or change in drainage density on a degraded site will alter the water balance. Reduced vegetation cover will increase runoff and there may also be increased deep drainage. The increase in slope angle on disturbed sites will generally increase the proportion of rainfall that becomes runoff. Erosion and downstream flooding and/or sedimentation are the likely consequences of altered hydrology unless precautions are taken to avoid these effects. Where sulfidic substrates are excavated by mining and stored in contact with moisture and oxygen, acid mine drainage may be discharged into groundwater or surface water. The discharge of acidic water alters downstream ecology and may damage the infrastructure it contacts, such as concrete structures in bridges.

Substrate Properties

Physical *properties:* Degrading processes such as erosion, overgrazing, or excessive tillage may alter the physical properties of the surface substrate in ways that decrease its suitability for plant establishment and growth, particularly by changing water storage and availability. Mining substrates and mineral processing residues are commonly poorly sorted, which alters physical properties such as available water capacity, porosity, soil strength, crusting, and susceptibility to wind or water erosion.

Water erosion strips away topsoil, decreases soil depth, and exposes subsoil material with different texture, lower organic matter levels, and degraded soil structure. The eroded soil has less favorable physical properties for water infiltration, water holding capacity, seed germination, and root growth. Wind erosion selectively removes clay and humus from the soil surface, increasing the prevalence of coarse materials. Fine sand deposits from wind erosion may bury topsoil and vegetation. Dust from bare surfaces can be a health hazard especially if it contains alkali salts or other toxic elements.

The passage of heavy machines on agricultural, forest, and mine sites causes compaction of the substrate, and this may be a major constraint to plant growth by restricting root depth. In drought-prone environments, the failure of roots to penetrate to depth may cause stunting or death of plants. Deep ripping to break the compacted layer is generally necessary in order to achieve deep root growth in mine pits.

Chemical *properties:* Plant growth may be hampered by low nutrient levels, acidity, alkalinity, salinity, sodicity, low organic matter levels, and excessive levels of toxic elements or compounds in the substrate. Therefore, effective land restoration depends on thorough chemical characterization of the substrate for these likely chemical constraints. Where feasible, any substrate that is likely to hinder plant growth should be isolated or buried under more benign materials so that root contact with it is avoided or minimized. This approach, which is practiced on mine sites, is clearly not possible on agricultural land. Soil and plant analysis are the most common methods for predicting likely nutrient deficiency or toxicity in the substrate.[20] Fertilizer applications can usually be effective in correcting nutrient deficiencies although determination of appropriate rates of application requires expert judgement or decision support systems and depends on the end land use. For acidity, alkalinity, salinity, or ion toxicities in the substrate, plant tolerance is the most cost-effective strategy for achieving successful plant growth. This may mean selecting a different suite of species or ecotypes to those that existed before mining unless the local species are already adapted to those constraints.

Biological *properties:* The disturbance of topsoil during mining, through stripping, transport, replacement, and/or storage, all have negative effects on soil biological activity.[21] Maximum soil microbial activity is retained when topsoil is used immediately after stripping without a period of storage. Storage for extended periods should be avoided, but in those mining operations where long-term storage is unavoidable, the stockpile should be uncompacted, should be <2 m deep, and should support a vegetative cover that maintains soil biological activity. Restoration of microbial biomass in replaced topsoil on revegetated mine sites may take 7–10 years.[21,22] In southwest Australia, the presence or absence of the pathogen *Phytophthora cinnamomi* in topsoil determines how it should be stored and reused.

Plant Establishment

Except in rare circumstances, vegetation cover is a requirement for land restoration. The alleviation of substrate constraints that limit plant establishment or growth is a prerequisite for successful vegetation establishment. Where a choice of substrate materials is available, the most benign of these should be placed at the surface to create a favorable root zone. This will minimize the need for expensive amelioration procedures for any adverse soil conditions. Topsoil, if available, is generally the most favorable substrate.

In land restoration, plant establishment is pursued through direct application of seed, the application of topsoil containing viable seed banks, transplanting of seedlings, or combinations of two or more approaches. Topsoil typically contains a substantial seed bank. Use of topsoil for revegetation can achieve large cost savings by avoiding seed collection and spreading. Topsoil is most commonly used for revegetation when the indigenous species of the area are to be replaced. However, where the topsoil contains seed of exotic species or has a large proportion of weeds in the seed bank, it may be preferable to avoid its use as the seedbed. In addition to the seed bank, topsoil contains soil microorganisms that are required for the restoration of soil functions such as organic matter decomposition and nutrient cycling. The nutrient stores in topsoil may be critically important for the production of adequate biomass in the restored ecosystem. The physical conditions in the topsoil are generally more favorable for seed germination and emergence than from other substrates. However, well-characterized subsoils or regolith materials may be suitable for revegetation when topsoil is absent or where topsoil is unsuitable due to chemical constraints or heavy weed infestation.

In some plant communities, seed is stored in the canopy of key species rather than in the topsoil. Placement of cut branches on the soil surface may be used in these cases to maximize recruitment of indigenous species.

Transplanting nursery-raised seedlings is a reliable approach for ensuring rapid plant establishment and canopy cover on a degraded site. Transplanted seedlings compete effectively against weeds and have greater survival under grazing by herbivores than plants recruited from direct seeding. For species that are difficult to propagate from seed, nursery-raised seedlings using tissue culture or other vegetative means of propagation may be the only reliable method for introducing those species during land restoration. However, raising nursery seedlings is relatively expensive, especially if only few plants of a large number of species need to be treated in this way.

Ecosystem Restoration

According to Hobbs and Norton,[14] the following are the ecosystem characteristics that should be considered when setting goals for land restoration: vegetation composition, structure, pattern, and heterogeneity; species interactions; and ecosystem function, dynamics, and resilience. A set of measures is needed to determine the success of restoration. They need to be not only low-cost and reliable indicators of present condition and function but also predictors of future trajectories for the restored ecosystem. Composition and structure of vegetation are the most commonly used indicators.[23] For pattern, heterogeneity, dynamics, and resilience, the indicators are less advanced in their development, in part because these indicators can only be identified and validated from long-term data. As a nutrient cycling indicator, microbial biomass has been proposed,[22] while Koch and Hobbs[23] concluded that soil organic matter was the best indicator of the restoration of nutrient cycling processes for bauxite mine restoration.

Increasingly, the land restoration activities in the mining industry are being assessed against completion criteria or success criteria.[17] These are legal instruments established following negotiation between regulators, mining companies, government advisers, and the community. Their aim is to provide certainty for all stakeholders about the restoration process to be followed and the expected outcome. They are designed to avoid future liability to government agencies or private landowners once mining has ceased and mine ownership has passed to new owners.

Case Study: Bauxite Mine Restoration

The restoration of land after bauxite mining in southwest Australia is recognized globally for its excellent practice.[24] Current practice is based on more than 40 years of research and development and adaptive management. There are important lessons to be learned about land restoration from this case study. Bauxite mining in the dry sclerophyll eucalyptus forest of southwest Australia is a shallow surface mining operation. The current restoration goal is to restore the forest values.[24] One hundred percent of the species are now routinely returned. Nutrient cycling appears to be on a trajectory towards restoring the nutrient stores and the fluxes of nutrients present in the premining forest.[22,25] Hydrological balance is disturbed for up to 12 years after land clearing for bauxite mining and subsequent revegetation.[26] Thereafter, water levels return to premining levels, but may drop below premining levels due to the increase in leaf area index relative to premining levels.[27] The restored forest is resilient to fire.[25] Completion criteria have been developed for bauxite mine restoration and several areas have been determined to meet the designated targets.[24]

The present restoration practice as described by Koch[28] is the result of four major revisions in the goals over 40 years and several other significant improvements in practice. The first rehabilitation simply planted *Pinus radiata* as a single species plantation. This was followed by a *Eucalyptus saligna* plantation and then by a goal to restore a diverse forest rather than a plantation. At this stage, it was thought that planting of local eucalyptus would not succeed because of the existence of *P. cinnamomi* in the soils and its threat to the survival of a wide range of native species. Further research demonstrated that the reconstructed profile would produce low risk of *P. cinnamomi* infection in susceptible species, and hence, it was decided to change the end land use goal to that of a forest compatible with the jarrah forest, using jarrah and marri as the overstory species.[24] Finally, it was decided to revise the end land use goal to achieve the restoration of the jarrah forest.[17,25] The selection of this goal was based on research breakthroughs that demonstrated that it was possible to stimulate the germination and emergence of recalcitrant species and hence reach close to 100% species return.

The main learning from this case study is that reaching the point when practices enable full ecosystem restoration takes several cycles of research and adaptive management. It will be based on a systematic program of research into biotic and abiotic constraints. A flexible approach from regulators enabled end land use goals to be revised over time as new research demonstrated the potential to achieve more challenging goals and targets. The end land use goals set in 1963 would have only resulted in exotic pine plantation on the mined bauxite pits. The present end land use goal is a fully functioning jarrah forest that can be integrated into existing forest management programs and achieve the multiple land use goals for the forest estate.[24]

A comprehensive report of the background research on bauxite mine restoration is found in *Restoration Ecology Special Issue* of 2007. It is a worthy model for study and emulation in land restoration on mine sites.

Case Study: Restoration of Land Affected By Dryland Salinity

Unlike the first case study of restoration on mined land, where disturbance is localized, dryland salinity has a more widespread impact and requires a landscape-scale response for land restoration. This case study is illustrated using examples of dryland salinity in southwest Australia and Northeast Thailand. By 2003, about 1 million ha in the wheatbelt of southwest Australia was affected by dryland salinity.[29] In Northeast Thailand, up to 30% of land could potentially become salt affected.[30]

Dryland salinity demonstrates how perturbation in water balance can have devastating consequences for the natural capital of landscapes. Whereas the native plants were predominantly deep-rooted perennial species, those agricultural species that replaced them were predominantly shallow rooted and annual. Annual plants use water only during their growing season, and their usage is limited by the fact that roots are generally confined to the surface 50–100 cm. Thus, the additional water under agricultural species is distributed to increased runoff, causing erosion and waterlogging, and to increased recharge to groundwater (Table 3). Williamson et al.[31] similarly concluded that dryland salinity in Northeast Thailand was triggered by deforestation of the uplands to produce crops like kenaf and cassava.

Two attributes of the landscape that gave rise to dryland salinity in southwest Australia were the deeply weathered regolith and the accumulation of salts from rainfall accretion.[33] Salt contents in extreme cases of up to 20,000 tons/ha have been reported,[34] most of it stored below 5 m depth.[32] Prior to clearing the native vegetation, plant roots in the upper 5 m of the regolith were largely separated from the salt bulge below, and the semipermeable aquifer at the base of the regolith was often dry.[32] However, with increased recharge, the aquifers have filled, causing water levels to rise at the rate of 0.2–1 m/yr. After a 20–30 years period of groundwater rise, saline groundwater discharge is observed commencing generally in valley floor land- forms. In Northeast Thailand, the origin of salt discharge is not rainfall accretion but salt mobilized from halite sequences in the Mesozoic sediments that underlie the Korat Plateau.[31]

Reversing Dryland Salinity

Ultimately, restoring the preclearing water balance is the only complete solution to the dryland salinity problem. This requires treatments in recharge zones of landscapes to decrease recharge rates. The species that can mimic recharge rates that existed before clearing will therefore probably need to be deep rooted and perennial. They will also have to be adapted to a variety of soil conditions and climatic regimes across the affected environment. Finally, it is imperative that the species chosen to fulfill the above functions are economically viable within the farm enterprise to accelerate their adoption by land managers.

In order to manage dryland salinity, it is necessary to understand the groundwater systems as well as water balance components. With intense winter rainfall, in landscapes extensively covered by shallow-rooted annual species, recharge can occur virtually anywhere in the landscape that is not

TABLE 3 Changes in Water Balances for Cleared Catchments before and after Clearing

Catchment	Year	Rainfall (mm)	Interception (mm)	Evapotranspiration (mm)	Change in Water Storage (mm)	Change in Groundwater Storage (mm)	Stream Flow (mm)
Wights forested	1975	1027	130	855	−28	−11	81
Wights cleared	1985	1147	0	565	–	−21	115
Lemon forested	1975	739	74	656	4	−1	5
Lemon cleared	1983	821	38	708	–	−19	56

Source: Williamson.[32]

actively discharging.[33] Computer modeling adds weight to the conclusion that the only fully effective revegetation solutions for salinity control in southwest Australia are with deep-rooted perennial vegetation over most of the whole catchment.[33] Even systems like agro-forestry that place a high density of woody shrubs and small trees in rows 30 m apart were insufficient in the modeling scenarios to restore water balance and achieve complete control of salinity. Continued reliance in the farming system on annual shallow-rooted crops such as cereals is problematic because these crops will allow continued recharge. Loss of these species is also problematic because they are the main source of income for farmers.

Until recharge control treatments start to decrease saline groundwater discharge, treatments are also needed in the discharge areas. These may include both engineering treatments to alleviate waterlogging[35] as well as vegetation options that cope with saline waterlogged conditions.[36] In Northeast Thailand, there has been considerable investigation of salt-tolerant species that can be grown on various classes of salt-affected soil.[30]

The case study on dryland salinity may therefore serve as a useful model for landscape-scale restoration. As with the case of land restoration after bauxite mining, the present approach has involved background research to understand the underlying physical processes (landscape water balance, water fluxes, hydrogeology) and to develop effective solutions (the effect of land use and vegetation type on water balance). The present set of strategies to control dryland salinity has evolved out of several phases of research and adaptive management leading to current understanding and solutions.

Conclusions

Land is a form of natural capital that is essentially finite and non-renewable. Every effort must be exerted to avoid degradation of land because degradation diminishes its natural capital value and compromises the ecosystem services provided by soil. The restoration of degraded land has the potential to increase its natural capital value and enhance ecosystem services. Since large areas of land globally are degraded, there is substantial scope for increasing ecosystem services by restoring degraded land. Land restoration is a relatively new discipline, which, along with restoration ecology, consists of successful practices for land restoration at a site-specific scale. Examples of best practice can be found in restored mine sites. The challenge remaining is to scale up land restoration from site-specific cases on mine sites to regional or landscape scales for a variety of degrading land disturbances. Most success at both scales has, to date, been concerned with restoring key ecosystems functions such as organic matter accumulation, nutrient cycling, and water balance.

References

1. Lal, R.; Stewart, B.A., Eds. Soil Restoration. In *Advances in Soil Science;* CRC Press: Boca Raton, Florida, 1992; Vol. 17.
2. Kaiser, J. Wounding Earth's fragile skin. Science **2004**, *304*, 1616–1618.
3. Reich, P.; Eswaran, H. Soil and trouble. Science **2004**, *304;* 1614–1615.
4. McKenzie, N.; Jacquier, D; Isbell, R.F.; Brown, K. *Australian Soils and Landscapes. An Illustrated Compendium;* CSIRO Publishing: Collingwood, 2004; 416.
5. Pillans, B. Soil development at a snail's pace: Evidence from a 6 Ma soil chronosequence on basalt in north Queensland. Geoderma **1997**, *80,* 117–128.
6. Robinson, D.A.; Lebron, I.; Vereecken, H. On the definition of the natural capital of soils: A framework for description, evaluation, and monitoring. Soil Sci. Soc. Am. J. **2009**, *73*, 1904–1911.
7. Lal, R. Soil degradation as a reason for inadequate human nutrition. Food Secur. **2009**, *1*, 45–57.
8. *Schaffnit-Chatterjee, C. The Global Food Equation. Food Security in an Environment of Increasing Scarcity; Deutsche Bank Research: Frankfurt, Germany 2009.*

9. Conacher, A.J.; Conacher, J. *Rural Land Degradation in Australia;* Oxford Univ. Press: Melbourne, Australia, 1995.
10. Lal, R.; Blum, W.H.; Valentine, C.; Stewart, B.A., Eds. *Methods for Assessment of Soil Degradation*; CRC Press: Boca Raton, Florida, 1998.
11. Stocking, M. A. Land degradation. In *International Encyclopedia of the Social & Behavioral Sciences*; Smelser, N.J., Baltes, P.B., Eds.; 2001; 8242–8247.
12. Bossio, D.; Geheb, K.; Critchley, W. Managing water by managing land: Addressing land degradation to improve water productivity and rural livelihoods. Agric. Water Manage. **2010**, *97*, 536–542.
13. FAO; UNDP; UNEP. Land degradation in south Asia: Its severity, causes and effects upon the people; World Soil Resources Reports. Food and Agriculture Organization of the United Nations: Rome, 1994.
14. Hobbs, R.J.; Norton, D.A. Towards a conceptual framework for restoration ecology. Restor. Ecol. **1996**, *4*, 93–110.
15. Aronson, J.; LeFoc'h, E.; Floret, C.; Ovalle, C.; Pontanier, R. Restoration and rehabilitation of degraded ecosystems in arid and semiarid regions. II. Case studies in Chile, Tunisia and Cameroon. Restor. Ecol. **1993**, *1*, 168–187.
16. Hobbs, R.J. Restoration ecology. In *Encyclopedia of Soil Science;* Lal, R., Ed.; Marcel Dekker: New York, 2002; 1153–1155.
17. Ward, S. Success criteria. In *Encyclopedia of Soil Science*; Lal, R., Ed.; Marcel Dekker: New York, 2002; 1156–1160.
18. Bell, R.W. Principles of land restoration. In: *Encyclopedia of Soil Science*; Lal, R., Ed.; Marcel Dekker: New York, 2002; 766–769.
19. Department of Minerals and Petroleum. *Draft Guidelines for Preparing Mine Closure Plans*; Department of Minerals and Petroleum: Perth, 2010.
20. Bell, R.W. Diagnosis and prognosis of soil fertility constraints for land restoration (Ch.16). In *Remediation and Management of Degraded Lands*; Wong, M.H., Wong, J.W.C., Baker, A.J.M., Eds.; Lewis Publishers: Boca Raton, Florida, 1999, 163–173.
21. Jasper, D.A.; Sawada, Y.; Gaunt, E.; Ward, S.C. Indicators of reclamation success—Recovery patterns of soil biological activity compared to remote sending of vegetation. In *Land Reclamation: Achieving Sustainable Benefits;* Fox, H.R., Moore, H. M., McIntosh, A.D., Eds.; A.A. Balkema: Rotterdam, 1998; 21–24.
22. Jasper, D.A. Beneficial soil microorganisms of the Jarrah forest and their recovery in bauxite mine restoration in southwestern Australia. Restor. Ecol. **2007**, *15,* S74–S84.
23. Koch, J.M.; Hobbs, R.J. Synthesis: Is Alcoa successfully restoring a Jarrah forest ecosystem after bauxite mining in Western Australia? Restor. Ecol. **2007**, *15,* S137–S144.
24. Gardner, J.H.; Bell, D.T. Bauxite mining restoration by Alcoa World Alumina Australia in Western Australia: Social, political, historical, and environmental contexts. Restor. Ecol. **2007**, *15*, S3–S10.
25. Grant, C.D.; Ward, S.C.; Morley, S.C. Return of ecosystem function to restored bauxite mines in Western Australia. Restor. Ecol. **2007**, *15*, S94–S103.
26. Croton, J.T.; Reed, A.J. Hydrology and bauxite mining on the Darling Plateau. Restor. Ecol. **2007**, *15*, S40–S47.
27. Bari, M.A.; Ruprecht, J.K. *Water yield response to land use change in south-west Western Australia,* Salinity and Land Use Impacts Series Report No. SLUI 31; Department of Environment: Perth, Australia, 2003.
28. Koch, J.M. Alcoa's mining and restoration process in south western Australia. Restor. Ecol. **2007**, *15*, S11–S16.
29. McFarlane, D.; George, R.J.; Cacetta, P.A. The extent and potential area of salt-affected land in Western Australia estimated using remote sensing and digital terrain models. In *1st National Salinity Engineering Conference, 9–12 November 2004, Perth Australia,* Conference Proceedings; Dogramaci, S., Waterhouse, A., Eds.; Institution of Engineers: Australia, 2004; 55–60.

30. Yuvaniyama, A. Managing problem soils in northeast Thailand. In *Natural Resource Management Issues in the Korat Basin of Northeast Thailand: An Overview*; Kam, S.P., Ho- anh, C.T., Trebuil G., Hardy, B., Eds.; IRRI Limited Proceedings No. 7; 2001; 147–156.
31. Williamson, D.R.; Peck, A.J.; Turner, J.V.; Arunin, S. Groundwater hydrology and salinity in a valley in Northeast Thailand. Groundwater Contam. Int. Assoc. Hydrogeologists Publ. **1989**, *185,* 147–154.
32. Williamson, D.R. Land degradation processes and water quality effects: Waterlogging and salinisation (Ch. 17). In *Farming Action Catchment Reaction;* Williams, J., Hook, R.A., Gascoigne, H., Eds.; CSIRO: Melbourne, 1998; 162–190.
33. Clarke, C.J.; George, R.J.; Bell, R.W.; Hatton, T.J. Dryland salinity in southwestern Australia: Its origins, remedies, and future research directions. Aust. J. Soil Res. **2002**, *40,* 93–113.
34. Moore, G.W. Salinity. In *Soil Guide. A Handbook for Understanding and Managing Agricultural Soils;* Moore, G.W., Ed.; Bulletin 4343; Department of Agriculture: Western Australia, 2004; 146–158.
35. Bell, R.W.; Mann, S. Amelioration of salt and waterlogging-affected soils: Implications for deep drainage. In *1st National Salinity Engineering Conference, 9–12 November 2004, Perth Australia,* Conference Proceedings; Dogramaci, S., Waterhouse, A., Eds.; Institution of Engineers: Australia, 2004; 95–100.
36. Barrett-Lennard, E.G. *Saltland Pastures in Austalia. A Practical Guide,* 2nd Ed.; Land, Water and Wool Sustainable Grazing on Saline Lands Sub-program: Canberra, 2003.

13

Limits to Growth

Brian D. Fath

Introduction

In thermodynamic-based systems science terminology, the Earth is considered a closed system. This means that it is open to energy exchanges, but not matter exchanges. Regarding energy, the Earth receives on average 1,300 W/m^2 of solar radiation at the top of the atmosphere and radiates or reflects a near equal amount back to space (the current increase in greenhouse gases causes a small difference in the radiation balance enough to be responsible for the current observed warming). The amount of matter on the Earth is more or less fixed, with only negligible exchanges from receiving meteorite impacts or outgassing. The Earth's matter is distributed in the following way: The Earth has a polar radius of 6.37×10^3 km, giving it 501.1 million km^2 of surface area, of which 70.8% is covered with water. The overall volume of the Earth is 1.08×10^{12} km^3, with a mass of 5.98×10^{24} kg. Those physical dimensions are hard evidence of the limits of the available space and material resources. The *Blue Marble* (Figure 1) shows this celestial body in all its complexities, opportunities, and beauty; and it also shows starkly the scale and boundaries of the one world that humans and all known species have. This photo, taken during the Apollo 17 Lunar Mission in 1972, coincided with the emergence of the modern environmental movement (more on that below). It was not the first environmental movement, but those ideas from the 1970s still cast a strong shadow on the discussions today that deal closely and explicitly about living in balance on the Earth as represented in among other things, the United Nations Sustainable Development Goals.

The ideas of bio-physical limits impacting human growth and resource consumption are not new. At the turn of the 19th century, Thomas Malthus was already writing about the convergence of population growth and food supply. He foresaw an inevitable turning point in which the finite resources of the planet could not support an exponentially growing population. His dour predictions later became the face of resource-limited, doom-and-gloom environmentalism, referred to as Malthusian. This discouraging perspective is a label that most modern environmentalists try to avoid, while still holding the reality of his concerns. At the time Malthus was writing, the global human population was around 1 billion, and today is approaching 8 billion, moving from abundant nature and scarce humanity to abundant humanity and scarce nature. Yet, there is not an absolute food shortage in terms of total calories produced, only hunger due to regional and distributional dilemmas. In fact, obesity is a more serious and growing problem in many countries than food deficiencies. Nonetheless, the effort that humans have taken to supply this food has meant the conversion of most arable land to agriculture leading to extensive direct impacts

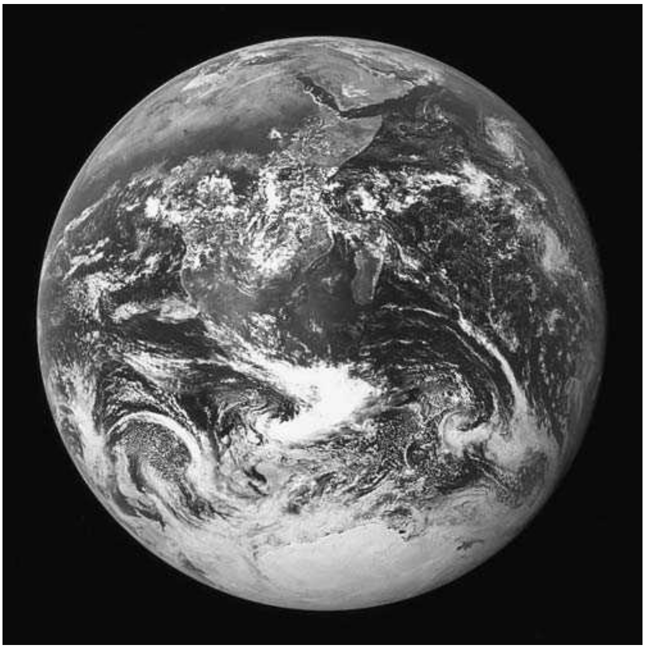


FIGURE 1 The *Blue Marble*, view of the Earth from Apollo 17 (1972).

such as habitat loss, soil degradation, and biodiversity loss. It has also led to the doubling of nitrogen flows through industrial fixation (as a result of the Haber–Bosch process) and the tripling of phosphorus flows (benefitting from massive industrial, fossil fuel-based mining efforts). The proximate result of increasing the amount of these typically limiting nutrients expands food supply, but the production of both nitrogen and phosphorus is highly fossil fuel-intensive creating energy supply dependencies. Furthermore, excess nitrogen and phosphorus that runs off agricultural fields creates massive eutrophication problems that are evident in almost every heavily populated estuary and water body in the world. Agricultural production to meet the world's needs is also responsible for rapid groundwater withdrawals, pesticide applications, and abundant greenhouse gas emissions at numerous stages from the fossil fuel use, to soil alterations, to methane from rice and livestock production. So, while the verdict is not in regarding absolute constraint of food production on human population growth, it does appear clear that we are in fact not feeding the people of the planet in a sustainable manner, nor whether it is even possible to do so at this scale. Malthus' concerns should not be dismissed lightly.

Environmental Limits

Recognition of limits can also be found in the seminal writing of George Perkins Marsh. Marsh wrote *Man and Nature* in 1864, which was one of the first scientific treatises on the impacts humans have on nature and the consequential conditions that coincide. Marsh warned that we should take notice of the scale and extent to which humans can continue to modify nature for our own benefit. Presciently, he wrote:

> A certain measure of transformation of terrestrial surface, of suppression of natural, and stimulation of artificially modified productivity becomes necessary. This measure man has unfortunately exceeded.
>
> **Marsh (1964)**

It is interesting that he recognized that humans will and must modify their environments similar to any species. There is a feedback in which the modifications, the other ecological interactions, and the environmental conditions reach a dynamic balance if the system is to sustain over time. Of course, this does not mean that the systems become rigid without change or variation because innovation, adaptation, evolution, information gain, and learning all allow for continual resilience and flexibility to meet the self-enhanced, recursive dynamic conditions. The path dependency between the interaction of the ecosystem constituents is one of the key features of living systems. Nonetheless, when humans erode too much land, deforest too many hectares, divert too much water, degrade too much habitat, etc., there are limits to which the ecosystem can recover and continue to provide the services we have come to expect. The diminishment first comes in the form of provisioning services when farms and fisheries fail, but then impacts cascade to supporting and regulating services, undermining the capacity of the land to regenerate and function sustainably. Continuation of an approach that overuses and abuses the natural resources will lead to eventual ruin, as Marsh noted in many earlier civilizations, which has been the topic of a recent plethora of research and books describing the collapse of complex societies (e.g., Tainter 1988, Diamond 2005, Kriwaczek 2010, Cline 2014).

The United States of America was settled on the notion of boundless space and opportunity, a land of cornucopia, which influenced not only the profligate physical resource use but also the psychological engagement that one had with resources. However, around the time that Marsh was writing, the transcontinental railroad was being completed. In 1869, the Union Pacific and Central Pacific railroads were joined in Promontory Point, Utah Territory, enabling rapid transportation access across the United States. Travel that previously took 4–6 months could be completed in 6 days. A few decades following both Marsh and transcontinental rail, another measure of limits was recognized in the closing of the Western frontier. The 1890 census showed the first time the disappearance of a contiguous frontier line of a migrating population. The westward wave of expansion was not endless as it had been (naively and myopically) perceived. Historian Frederick Jackson Turner used this moment to refer to the "closing of the American Frontier" (1893). At this bifurcation point, it should have been obvious that the solution to resource shortages, for example, in the form of eroded agricultural land, could not be found by simply relocating to the next plentiful area. While there was still plenty of "under-utilized land" (from a human economic perspective), the new reality meant filling in the middle states rather than experiencing an expansionary boundary. It should have been a time to rethink and reformulate our relationship with land and nature, thus confronting limits and working within constraints. However, the notion of boundless resources remains deeply held by many and institutionalized in many core economic practices (e.g., debt-based money supply, Ponzi-style retirement benefits, high future discount factors). Therefore, recognizing and accepting bio-physical limits was not the first impulse when faced with this new reality.

Ethical Limits

In the 1940s, limits were given another dimension by Aldo Leopold and his *Land Ethic* referring now not only to physical limits but also moral ones. He questioned, why would we pursue certain things (namely economic growth through resource consumption and technologies with unintended consequences) if the result is to destroy and degrade the life and ecosystems around us, loss of species and wild places. He eloquently wrote:

> Our grandfathers were less well-housed, well-fed, well-clothed than we are. The strivings by which they bettered their lot are also those which deprived us of [passenger] pigeons. Perhaps we now grieve because we are not sure, in our hearts, that we have gained by the exchange. The gadgets of industry bring us more comforts than the pigeons did, but do they add as much to the glory of the spring?
>
> **Leopold (1949)**

He elaborated this new perspective in what he called a Land Ethic:

> A thing is right when it tends to preserve the integrity, stability, and beauty of the biotic community. It is wrong when it tends otherwise.
>
> **Leopold (1949)**

His Land Ethic puts an environmental ethic clearly on the table as to what is right and wrong, expanding the boundaries of moral concern to the living environment and to the land itself, thus providing new limits of our actions. This ethic is informed by an ecological as well as a systems perspective. Leopold called on humans to constrain certain activities if those actions were detrimental to natural systems. He recognized, for example, the intricate balance cascading between the predator wolves and soils passing through the grass and deer. Too few wolves meant consequences for the soils. There are limits to the control which humans can exert over the ecosystem with a desire for favored species in certain times and places (e.g., the domestic chicken is the most abundant bird in the world, and *Zea mays* (corn) the most abundant plant). In the ensuing decades, we have better knowledge of complex systems, and with this we can have better management, but that does not obliviate the presence of limits imposed from outside and fostered from within. The Land Ethic allows one to see that beyond our own immediate sphere we interact with other bounded spheres, thus implying that the limit we sense is in fact an indication of another system beyond that we are pushing up against. In a zero-sum interaction, expanding our boundary, for example, into nature, diminishes nature. However, there are many non-zero-sum interactions and relations prevalent in ecosystems and complex adaptive systems. The challenge is not to completely disengage when we encounter another system, but to acknowledge it, respect it, work to understand it, and then try to find a way to engage it for the benefit of both parties (see, e.g., Fiscus and Fath 2019).

Socio-Economic Limits

These ideas of limits coalesced in a seminal work by Meadows et al. sponsored by the Club of Rome in the *Limits to Growth* report released in 1972. Their conclusion was based on the results of one of the first global systems dynamics models developed that included state variables representing human population, agricultural productivity, industrial production, resources, and pollution. The dire results were that under no scenario would unlimited growth be able to continue, and the only steady state was found by strict conditions of stable population, 100% use of renewable resources, and investment in business that equaled depreciation. The report was an important contribution to the debate about limits from a bio-physical perspective but was largely ignored by economists and politicians since the approach did not fit their existing growth-oriented mental model. The study was dismissed out-of-hand as having no relevance since their models—absent of any environmental resources or feedbacks—could grow forever. The Meadows et al. model was viewed as Malthusian and overly doom-and-gloom, a stigma that dogged the environmental community for decades. Nonetheless, the Club of Rome work continued as decadal update reports were published showing the projections were in line with reality.

Coincident with the release of the original *Limits to Growth*, the modern environmental movement was gaining traction (as stated above, spurred on by images such as the *Blue Marble* and the Spaceship Earth concept). In the United States, it was a period of aggressive federal policy protections for the environment (Table 1), most passing with very wide margins in both the House and Senate and signed by both Republican (Nixon and Ford) and Democratic (Carter) Presidents. Shy of 50 years later, the pendulum has swung far the other way. Concerted efforts within the United States Congress to dismantle or weaken these Acts, in particular the Endangered Species Act (ESA), is rampant; including news just today (August 13, 2019) of efforts to sign a Presidential Executive order to weaken ESA protections. It would be humorous if it were not sad that one prime argument against the ESA is the stated negative impact that it may have on the economy. However, measuring the success of this

TABLE 1 Major United States Legislation Passed in the 1970s during the Resurgent Environmental Movement

Coastal Management Zone (1972)
Clean Air Act (1972)
Clean Air Act Amendments (1977)
Comprehensive Environmental Response, Compensation and Liability Act (1980)
Endangered Species Act (1973)
Federal Land Policy and Management Act of 1976
Fishery Conservation and Management Act of 1976
Marine Mammal Protection Act (1972)
National Environmental Policy Act (1970)
National Forest Management Act (NFMA) of 1976
Forest and Rangeland Renewable Resources Planning Act of 1974
Noise Pollution and Abatement Act of 1972
Marine Protection, Research and Sanctuaries Act of 1972
Resource Conservation and Recovery Act (RCRA), enacted in 1976
Safe Drinking Water Act (SDWA) (1974)
Surface Mining Control and Reclamation Act of 1977 (SMCRA)
Toxic Substances Control Act of 1976

legislation by its economic contribution is fundamentally misunderstanding its purpose. Framing the ESA as a failure due to its hindrance of capital gain is historical revisionism. The language of the ESA was clear that protections were being put in place precisely because of a lack of concern for the environment and an alarming over-prioritization of economic growth at the expense of ecological limits. The first sentence of the ESA reads:

> (a) Findings—The Congress finds and declares that—(1) various species of fish, wildlife, and plants in the United States **have been rendered extinct as a consequence of economic growth and development untempered by adequate concern and conservation** (emphasis added).
>
> **Endangered Species Act (1973)**

The whole point of the legislation was to put a brake on growth and remember the limits of nature. It was also encouraging that the drafters of the legislation appreciated and recognized the interconnectedness of nature and that species are part of larger ecosystems.

> (b) Purposes—The purposes of this Act are to provide a means whereby **the ecosystems upon which endangered species and threatened species depend may be conserved** … (emphasis added).
>
> **Endangered Species Act (1973)**

The tension between pressures for economic growth and environmental constraints continued and elevated in the ensuing decades.

Flourishing within Limits

The idea of limits has been renewed but with a more positive outlook. In particular is the perspective that the presence of a limit does not need to invoke immediately negative connotations. For example, Jane Jacobs (2000) has focused on the opportunities that constraints and limits bring about. She insightfully remarked that

> Natural principles of chemistry, mechanics and biology are not merely limits. They're invitations to work along with them.

In other words, our understanding of thermodynamics, the periodic table, and biological principles should help to design smarter and better performing systems than without this knowledge. Ecosystems and socio-economic systems are self-organizing and thrive even in the presence of constraints. Think back to the energy flows mentioned in the opening, around 1,300 W/m^2, yet that is enough energy to drive atmospheric circulation, global ocean currents, and complex ecosystems in almost every niche on the planet. Nature has evolved to utilize this energy in a very efficient and also robust manner, resulting in highly complex, diverse, well-functioning systems that arose and are maintained all within the real-time bio-physical constraints.

The message to learn from ecosystems is elaborated in Jørgensen et al. (2015), *Flourishing within Limits: Following Nature's Way*. The authors, all systems ecologists, have drawn upon ecological theory of growth and development to identify a number of attributes that are evident in ecosystems. Using these properties and approaches—within the known bio-physical constraints—ecosystems are able to thrive and flourish, as the most diverse, complex, integrated, and sustainable systems on the Earth.

In a nutshell, the idea is that ecosystems are constrained by both the available resources as inputs and outputs as they reside as gradient-enhancing conduits between these two flows (see Figure 2, Fath 2017). In addition to solar energy, the rate of material availability is controlled by the bio-geochemical cycles at local and global scales. The rates of water, carbon, nitrogen, phosphorus, etc. all have local and global

TABLE 2 Fourteen Properties Observed in Ecosystems

1. Ecosystems conserve matter and energy
2. There are no trashcans in nature
3. All processes (in nature and society) are irreversible
4. All life uses largely the same biochemical processes
5. Ecosystems are open systems and require an input of work energy to maintain their function
6. An ecosystem uses surplus energy to move further away from thermodynamic equilibrium
7. Ecosystems use three growth and development forms: (1) biomass, (2) network, (3) information
8. Ecosystems select the pathways that move it most away from thermodynamic equilibrium
9. Ecosystems are organized hierarchically
10. Ecosystems have a high diversity in all levels of the hierarchy
11. Ecosystems resist (destructive) changes
12. Ecosystems work together in networks that improve the resource use efficiency
13. Ecosystems contain an enormous amount of information
14. Ecosystems have emergent system properties

Source: After Jørgensen et al. (2015).

Ecosystem Input Constraints
- Solar radiation
- Global carbon cycle
- Rate of nutrient cycling
- Rate of hydrological cycle

Ecosystem Output Constraints
- Rate of decomposition
- Rate of accumulation of unwanted byproducts
- Finding others to take your waste

FIGURE 2 Some basic constraints that ecosystems have evolved to thrive within.

constituents that influence what is available and these constraints then lead to the type of ecosystem observed whether it is a tropical rain forest, deciduous forest, steppe grassland, or Arctic tundra, for example. On the output side, the main constraint is that the products and flows out cannot exceed the capacity of the environment to assimilate them. Often this occurs by linking processes such that the output of one process is the input to another in an integrated, coupled fashion. The most obvious example is photosynthesis and respiration in that the former takes in CO_2 and gives off O_2, while the latter does the opposite. Other wastes, such as organic ones, need to be decomposed and returned back to elemental constituents. Therefore, the rate of decomposition is a main factor for the uptake of outflows, which are biologically mediated and influenced by temperature and water (warm, wet environments have faster rates of decomposition). The biomass foundation for fossil fuels occurred during a time period when for various reasons, the material was unable to be decomposed and therefore was buried under conditions that promoted the conversion to fossil fuels. Today's society is generating plastic pollution faster than it can be decomposed, and will likely also leave a recognizable and discernable (and perhaps useful) layer for future generations, and also likely that some organisms will eventually evolve better processes to decompose plastics thus reaching a new balance between production and consumption.

There are of course key differences between nature and society, namely a temporal aspect that ecosystems happen in the present, in real time—although the consequences, building biomass, building soils, establishing drainage patterns, and biogeochemical cycles, etc. have lasting effects. Socio-economic systems operate with a more future-oriented perspective, in that while the actions are short term, there is thought for a future pay-off. For example, a lion takes only what is needed to satisfy the immediate hunger. There is no utility in killing more prey at that moment as they would be absconded by competitors or would spoil; there is no profit motive or line of credit. A hunter today has in mind to take as many prey (or fish or trees or minerals, etc.) as possible since the surplus can be exchanged for other goods or services or turned into a storable currency commodity. It is not clear how to reconcile this basic difference. Humans cannot and should not give up the intellectual foresight that has evolved, but rather it could be used to see the systemic consequences of actions that exceed the limits, that exceed the regenerative rates provided by nature's processes. The remaining question is both philosophical and one of management, which is: How to align our activities within these available flows and constraints?

Conclusion

If humanity desires to stay within limits, then a first consideration is to know where those limits are. How much resources can humans use from the environment and still be within the limits? A standard ecological concept is one of carry capacity, the maximum number of individuals that can be maintained in an area without degrading that area. Carry capacity is a dynamic concept—it can move higher as we innovate new efficient methods and technologies as observed in agriculture since the time of Malthus; yet, it can also move lower as we despoil forests and fields with logging, erosion, and desertification as observed by Marsh and other witnesses to the collapse of civilizations. Metrics such as Ecological Footprint and Ecological Biocapacity that try to make estimates of both the consumption (footprint) and the resource (Biocapacity) have shown that we have outstripped our resource base. Ecological Footprint exceeds Biocapacity, and at a global level, attention is drawn to a notion called overshoot day, which for 2019 occurred on July 29. Meaning on that date, humanity had already used all the resources "sustainably available" to it for that year. After this date, the remainder of the year represents overshoot or dipping into future reserves and leaving future generations with reduced resources. Given this temporal aspect, we see that limits are not simply bio-physical but also ethical as Leopold instructed. Yet, a rigorous debate continues to this day as to whether there are limits of any kind or not. This creates tension between perspectives on policies and approaches, as some see limits as absent or challenges to overcome, while others want to recast the human scale to accommodate them. The planetary boundaries highlighted through the text give a clear indication of one scale of limits. If perhaps we break beyond the Blue Marble and our progeny lives elsewhere than within the Earth's gravitational pull, then this will shift and delay the ultimate confrontation with those concrete limits. Can this line always be receding into the future? I prefer to take care of the home that we do have on which we have co-evolved rather than rely on one that is only in our imagination.

References

Cline, E.H. 2014. *1177 B.C.: The Year Civilization Collapsed*. Princeton, NJ: Princeton University Press, 237 pp.

Diamond, J.M. 2005. *Collapse: How Societies Choose to Fail or Succeed*. London: Penguin Books.

Endangered Species Act, 16 U.S.C. §§1531-1544 (1973).

Fath, B.D. 2017. Systems ecology, energy networks, and a path to sustainability. Prigogine lecture. *International Journal of Ecodynamics* 12(1), 1–15.

Fiscus, D.A., Fath, B.D. 2018. *Foundations for Sustainability: A Coherent Framework of Life–Environment Relations*. London: Academic Press.

Jacobs, J., 2000. *The Nature of Economies*. New York: Vintage Books, 208 pp.

Jørgensen, S.E. Fath, B.D., Nielsen, S.N., Pulselli, F., Fiscus, D., Bastianoni, S. 2015. *Flourishing within Limits to Growth: Following Nature's Way*. New York: Earthscan.

Kriwaczek, P. 2010. *Babylon: Mesopotamia and the Birth of Civilization*. London: Atlantic Books, 310 pp.

Leopold, A. 1949. *A Sand County Almanac, and Sketches Here and There*. New York: Oxford University Press.

Marsh, G.P. 1864. *Man and Nature; or, Physical Geography as Modified by Human Action*. Cambridge, MA: Belknap Press.

Meadows, D.H., Meadows, D.L., Randers, J., Behrens III, W.W. 1972. The Limits to Growth; A Report for the Club of Rome's Project on the Predicament of Mankind. New York: Universe Books. 205 pp.

Tainter, J.A. 1988. *The Collapse of Complex Societies*. Cambridge: Cambridge University Press.

Turner, F.J. 1893. The Significance of the Frontier in American History. Annual Report of the American Historical Association, *pp. 197–227.*

14

Nuclear Energy: Economics

James G. Hewlett

Introduction

In an entry published in the *Encyclopedia of Energy Engineering and Technology,* I attempted to answer the following question: Is nuclear power economic? Unfortunately, I was unable to give a definitive answer to this question. Most of the work on that entry was done in the 2004–2005 time frame, and at that time, there was very little information on the realized costs of building nuclear power plants in the United States. There was some information on the realized cost of building nuclear power plants abroad (mainly in the Far East). However, that information was at best "sketchy" and difficult to use to estimate costs in the United States. Additionally, in the United States, a number of cost estimates of building hypothetical ("paper") plants at hypothetical ("Middletown USA") sites were published. History has shown that such estimates are always too low.[1] In short, at that time, there was a large amount of uncertainty about nuclear power plant construction costs. Thus, the best that could be done was to conclude that nuclear capital costs would have to be less than $1500 per kW before nuclear power would be economic. (This was substantially less than the cost of building nuclear power plants in the Far East.)

Over the last few years, the economic and political environment has changed, and thus, the question of whether nuclear power is economic needs to be reevaluated. First, a number of estimates of building actual nuclear power plants at actual sites have recently become available. This, by itself, would reduce some of the uncertainty about nuclear power plant construction costs. However, these estimates are two to three times higher than the ones made in the mid-2000s. Additionally, power plant construction costs in general have increased substantially. For example, the realized costs of building coal-fired power plants and wind farms have increased by about 80%–100%.[2] Such increases in construction costs would have major effects on the economics of nuclear power.

Second, unlike when the original entry was written, for a variety of reasons, utilities are now far more interested in nuclear power. In 2006, licensing activity at the U.S. Nuclear Regulatory Commission (NRC) was limited to four early site permit approvals. By issuing an early site permit (ESP), the U.S. NRC approves one or more sites for a nuclear power plant, independent of an application for a combined license to build and operate a power plant. This ESP can be valid for up to 40 years. Since then, 16 utilities have filed applications with the NRC to build and operate a total of 28 nuclear units. Given this increased licensing activity, a reexamination of the underlying economics of nuclear power is certainly in order.

Third, in the 2004–2005 time frame, most forecasts of coal and natural gas prices in 2015–2020 time frame were about $1.45 and $6 ($2009) per mmBtu, respectively.[3] Since then, fossil fuel prices increased substantially and then fell. As of 2010, most forecasts of fossil fuel prices are now greater than the ones made 5–6 years ago. Increased fossil fuel prices will clearly affect the economics of nuclear power.

The purpose of this entry is to reexamine the economies of nuclear power in light of the changed economic environment. Before proceeding with the analysis, two comments about the scope of the analysis will be made. First, in this entry, it is assumed that the investment decision—e.g., the decision to build a nuclear unit or coal-fired power plant—will be made by the owner of a traditional utility. This utility owns and operates other power plants, and the operation of all of them is interrelated. That is, if the utility chooses to operate power plant X less, it would have to operate power plant Y more to meet demand. Otherwise, "the lights would go out." In such cases, the decision maker's objective is to build the unit that minimizes the total cost of building the plant in question and operating all power plants. Suppose, for example, that the decision maker has the choice of building a nuclear or coal-fired power plant. The decision maker will calculate the total system costs if the nuclear plant is built and compare that estimate with the total system costs if the coal plant is built. The decision maker would choose to build the plant that yields the lower total system costs.

The available software to estimate total system costs is complex and expensive and requires many assumptions, and thus, using this approach is beyond the scope of this entry. However, if the alternative to the nuclear power plant is another baseload plant type operating in the same portion of the merit order (baseload demand), total system costs would be minimized by choosing the plant type that has the lower "stand-alone" or levelized cost. (Levelized costs will be defined below. Additionally, because electricity is costly to store, demand will vary over the day, month, and/or year. The portion of total demand that does not vary is called baseload, and the units that are used to meet this demand will run at close to full capacity over the entire year.) This is because the operation of both units under consideration will have the same effect on the operation of the other units. For example, suppose that the levelized cost of building and operating a nuclear power plant for 40 years is 6 cents per kWh, and the levelized cost of building and operating a combined-cycle natural gas–fired power plant is 8 cents per kWh. If both units are assumed to operate in the baseload mode, then the operation of both units will have the same effect on the operation of the other units. In such a case, total system costs would be minimized by building the nuclear plant.

Thus, in this entry, the alternatives to the nuclear power plant are two other baseload plant types—namely, coalfired and combined-cycle natural gas–fired units. By limiting the comparisons to other baseload plant types, the analysis becomes much more tractable and transparent. Unfortunately, by just computing levelized stand-alone costs, many of the renewable technologies must be excluded from the analysis. The stand-alone cost of building and operating a wind farm, for example, can be computed. However, total system costs may not be minimized by building the wind farm even though that plant type has lower stand-alone costs. This is because the effects of the operation of the wind farm and the nuclear unit on the operation of the other units will not be the same.

Second, since fossil fuel prices will probably increase over time, there is a time dimension to the question of whether nuclear power is economic. Because of the recession and utility conservation programs, additional baseload capacity will probably not be needed until around 2020. Thus, in this analysis, the first year of a unit's operation is assumed to be 2020. By focusing on the mid-term, the carbon capture and storage (CCS) technologies will not be considered. Recently, a CCS task force was formed with a goal of bringing 5 to 10 commercial-size CCS units online by 2016.[4] Even if this goal, which is very ambitious, is met, to demonstrate that the technology works, the units would have to operate for probably 5–10 years. In all probably, it would take 15–20 years before the CCS technologies would be commercially available on a widespread basis.

Capital Costs for Nuclear and Coal-Fired Powerplants

One of the major uncertainties in any analysis of the economics of nuclear power deals with the construction cost estimates. This section begins with a discussion about nuclear and coal-fired powerplants' overnight capital cost estimates. Overnight cost is defined as the cost of building a power plant instantaneously at some point in time. It is also a direct measure of the value of the land, labor, and materials needed to build a nuclear power plant. Thus, differences in overnight costs reflect differences in the values of the land, labor, and material needed to build the same unit.

It is obviously impossible to build a plant overnight, so the second part of this section describes how the total project costs are derived from the overnight costs. To do this, a number of important assumptions are needed, and in many studies, they are not articulated. This section will also show why comparisons of total project costs must be made with great care. The fuel costs for coal and natural gas–fired power plants are discussed in Section 4. The other assumptions will be briefly discussed in Section 6. These include nonfuel operating costs and nuclear fuel costs.

Nuclear and Coal-Fired Power Plant Overnight Costs

Prior to about 2007, most analyses/discussions about nuclear power plant overnight capital costs tended to focus on either realized costs of units built in the Far East (mainly Japan) or on the estimated costs of building generic units at generic sites in the United States.[5] Each of these sources had their own set of problems. There are always problems with transferring the experience of reactors built in foreign countries to the United States. Also, publicly available foreign overnight capital cost data are not well documented, so all the costs may not be included in the reported figures. Additionally, research has indicated that cost estimates of generic units built on generic sites were always too low, so there were problems with using the resulting estimates. Nonetheless, the analyses/discussions that based their cost estimates on foreign reactors tended to use overnight nuclear capital costs of about \$2700 per kW (2009 dollars).[6] The ones that used the cost estimates of generic units tended to use overnight nuclear construction costs of about \$1500–\$2000 per kW (2009 dollars).

Over the last few years, as part of the process of getting approval from the state public utility commissions (PUCs) to proceed with their NRC licensing activities, utilities filed cost estimates of building actual powerplants at actual sites. In some cases, it was possible to determine what cost items were and were not included in the estimates. More important, if the overnight construction costs were not directly reported on the filings, they could be directly estimated. On average, these overnight nuclear construction cost estimates, shown in Table 1, were about \$4000 per kW.

These estimates are clearly much better than the ones based on generic units built at generic sites. Because problems with the transfer of foreign cost information to the United States are avoided, they are much better than the estimates based on realized overnight costs of reactors built in Japan. However, they are also much higher than the ones based on generic designs at generic sites and were also much greater than the realized overnight costs of units built in Japan.

There are probably at least four reasons why the recent U.S. overnight cost estimates are greater than the realized costs of reactors built the Far East (mainly Japan). First, there are clearly cultural factors at play. Second, over the last 15 years, there has been a slow but relatively constant expansion of nuclear power in Japan, so Japanese builders are further down their learning curves than their counterparts in the United States. (The South Texas project is being built by Toshiba, a large Japanese firm that has built

TABLE 1 Estimated Overnight Costs and Lead Times of Selected Proposed Nuclear Plants

Owner	Plant Type	Plant	Capacity (mWe)	Costs (2009 Dollars per kW of Capacity)	Lead Times (Years)
Tennessee Valley Authority	ABWR	Bellefonte	1371	$3164	NA
Florida Power and Light	ESBWR	Turkey Point	3040	$3811	5–6
Progress Energy (Florida)	AP1000	Levy County	2212	$4541	5
South Carolina Electric and Gas	AP1000	Summer	2234	$4089	5 (unit 2) and 8 (unit 3)
Southern	AP1000	Vogtle	2200	$4535	6
NRG	ABWR	South Texas	2700	$3758	NA
Average				$3983	

Source: Adapted from Du and Parsons.[7]

Note: A correction was made to the Vogtle Estimate reported in Du and Parsons.[7]

ABWR, advanced boiling water reactor; ESBWR, economically simplified boiling water reactor; AP1000, advanced pressurized water reactor.

some nuclear power plants in Japan to another. The cost estimates for that project are not that much different from the others. Thus, the learning may not be transferred from one country to another). Third, it is not clear whether all the costs are being reported in the Japanese figures. This is especially true for the so-called owners' costs. Costs which are the ones incurred by the utility over and above the ones paid to the firms building the unit. In fact, it was never clear if the reported Japanese costs even included any owners' costs. Lastly, many of the major components for the proposed U.S. reactors, such as the reactor vessel, are being manufactured in Asia, so some exchange rate issues could exist.

There are also a number of reasons why the overnight nuclear construction cost estimates obtained from the PUC filings are higher than the ones for generic units. As was just noted, research has shown that generic cost estimates at generic sites are always too low. Additionally, a number of analysts have argued that the growth in overnight costs was due to increases in commodity prices—the prices of iron, steel, cement, and so on—that occurred from around 2005 to 2008.[8] However, about 40%–50% of the overnight cost of building a nuclear power plant is labor related, and therefore, the effects of increases in commodity prices on total construction costs are probably modest. Moreover, the cost estimates based on generic units might have assumed that all of the components were manufactured in the United States. Over the last 10 years, the value of the dollar relative to the yen has fallen. Since many of the major nuclear components for proposed U.S. plants will be imported from Japan, some of the increases in overnight costs could be due to the fall of the dollar. This, however, would depend upon how the contracts with the Japanese firms are structured, and there is very little public information about this. Lastly, with the rapid expansion of nuclear power in China, some bottlenecks in the production of the major components that would increase costs appeared.

The increases in nuclear power plant overnight construction cost estimates raise the question of whether the current cost levels will be permanent. The effects of commodity price increases and/or decreases in the value of the dollar on the cost estimates are very unclear. Thus, the effects of decreases in commodity prices (or increases in the value of the dollar relative to the yen) on costs would also be unclear. Additionally, even with the rapid expansion of nuclear power in the Far East, in the long run, markets would adjust, and the bottlenecks would no longer exist. Because of the complexity of the technology, it would take a number of years for existing firms to build new production facilities and for new firms to enter the market. Thus, the bottlenecks may not be removed for a number of years.

Lastly, currently, only a few firms produce the large nuclear components and build the power plants. In such cases, the behavior of one firm (e.g., Westinghouse) could possibly affect the behavior of its competitor (e.g., General Electric). In economist's jargon, such a market structure is called an oligopoly.

TABLE 2 Estimated Overnight Construction Costs of Selected Proposed Coal-Fired Power Plants

Owner	Plant	Capacity (mWe)	Costs (2009 Dollars per kW of Capacity)
Florida Power and Light	Glades	1960	$2130
Duke Power	Cliffside	800	$2124
AMP Ohio	Megis Co.	960	$3277
AEP Swepco	John W Turk Jr	600	$2508
Average			$2510

Source: Adapted from Du ans Parsons.[7]

One major characteristic of an oligopolistic market is price "stickiness"—i.e., when the firm's costs fall, prices will fall, but with a lag. Thus, even if the costs of the firms that produce the components and build the plants fall, it would take some time before reductions in costs would be reflected in reductions in the prices charged to the utilities.[9,10]

The availability of some actual overnight cost estimates from the PUC filing reduces some of the uncertainty in nuclear power plant construction costs. Unfortunately, the large increases in the estimates introduce another source of uncertainty—namely, whether the current cost levels will be permanent. Given current publicly available information, one can only speculate why overnight capital costs increased and whether they will fall in the future. Moreover, no nuclear power plants in the United States have been built on time and on budget, and cost overruns in the two units under construction in Europe have occurred. Thus, even with a number of detailed cost estimates of building actual units at actual sites, there is still a considerable amount of uncertainty about nuclear overnight construction costs. Indeed, until a few units are actually built, nuclear construction costs are essentially unknown.

Table 1 also shows estimated date of commercial operation leadtimes for a number of proposed nuclear units. As can be seen from Table 1, with the exception of one unit, the estimated leadtimes ranged from 6 to 7 years. It should be noted that these lead time estimates were made just before or at the beginning of the recession. Recently, the date of commercial operation for a number of these units was moved back simply because the capacity was not needed. In all probability, the utilities will also move back the construction start date and keep leadtimes the same. If, however, they begin construction and then revise their estimated date of commercial operation, lead times could actually increase. In the 1970s and 1980s, many utilities building nuclear power plants also increased leadtimes because of the lack of need for capacity. I have shown elsewhere that these actions also affected overnight costs.[11] Thus, this issue is important.

Lastly, Table 2 shows that the average estimated overnight cost of building a number of recent coal-fired power plants was about $2500 per kW. A 2007 Massachusetts Institute of Technology (MIT) study estimated that the costs of building a wide range of coal-fired powerplants were about $1280 to $1360 per kW (2005 dollars).[12] Thus, the cost of building coal-fired power plants also increased substantially, and therefore, the cost growth was not limited to nuclear power.

Derivation of Total Project Costs

The focus of the preceding section was on overnight costs, because they are a direct measure of the values of the land, labor, and materials needed to build a power plant. However, the analysis will use total project (capital) costs. The best way to explain how total capital costs are computed is to outline the steps taken to derive an estimate of the total cost of building any power plant. The first step is to prepare a detailed "bottom-up" estimate using current commodity prices, labor wage rates, and so on. This bottom-up estimate is the product of the estimated quantities of the land, labor, and materials needed to build the unit times the prices of these inputs. Suppose, for example, that the firm is making an estimate in 2010 of the cost of building a nuclear unit that will become operational in 2020. Given a 6-yr lead time, construction must begin in 2015. Thus, the firm would first make a bottom-up estimate using 2010 prices and wage rates. This is the estimate of the overnight costs using 2010 prices. In the example shown in Table 3, the 2010 estimated overnight cost is $4000 per kW of capacity.

Next, the cost of building the unit overnight in 2015—the year that construction of the unit begins—must be estimated. To do this, the firm would make assumptions about how commodity prices, labor wage rates, and so on would change from 2010 to 2015. In the example shown in Table 3, commodity prices, labor wage rates, etc., are assumed to increase at a rate equal to the general inflation rate of 3% per year. Given this assumption, the 2015 overnight cost would be about $4600 per kW ($\$4000 \times 1.03^5$). Then, assumptions about how the funds are expended over each year of the construction period and how prices change over the construction period would be used to compute total costs in the dollars of the year the funds are expended. In the example shown in column 3 in Table 3, total construction costs excluding financing charges would be about $5000 per kW of capacity. Note that simply because of increases in prices and wages over the 2010–2020 period, costs increased by more than $1000 per kW.

Lastly, there is the issue of how to include the financing costs in total capital costs and analyses of the economics of nuclear power. Because of the complexity of this issue, there are wide variations in both the financing rates and the costs that are included in the total capital cost estimates. In some analyses, including the present one, for a variety of reasons, the total capital cost estimates that are computed do not include any financing costs. In such cases, financing issues are accounted for elsewhere.[7] In other analyses, the total capital cost estimates do include financing costs. Sometimes, the total capital cost estimates include just interest charges on the monies that are actually borrowed.[6] If the analysis is used by a utility that is subject to state-level rate-of-return regulation, the financing charges included in the total capital cost estimate often consist of interest on the funds that are borrowed (debt financing) and an implied charge for the funds that are internally generated (equity financing).

It is always tempting to compare total capital cost estimates that are reported in the media. Columns 4–7 in Table 3 illustrate that this must be done with great care. In particular, these columns show total capital cost estimates using four different assumptions about the financing costs that are included

TABLE 3 Derivation of Total Capital Costs

	Total Capital Costs						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Year	Percent of Total Overnight Costs Spent in Each Year of Construction Period	Expenditures in 2015 Dollars (per kW)[a]	Expenditures in Dollars of Year Funds Expended (per kW)[a]	Financial Costs Not Included (per kW)	Column 4 Plus Financing Charges—Lower Financing Rate (per kW)[b]	Column 4 Plus Financing Charges—Higher Financing Rate (per kW)[c]	Column 4 Plus Just Debt Component (per kW)[d]
2015	10.00%	463.7	$463.7	$463.7	$695.9	$925.1	$560.2
2016	15.00%	$695.6	$716.4	$716.4	$1004.8	$1273.9	$838.6
2017	20.00%	$927.4	$983.9	$983.9	$1289.7	$1559.3	$1116.0
2018	30.00%	$1391.1	$1520.1	$1520.1	$1862.2	$2147.1	$1670.8
2019	15.00%	$695.6	$782.9	$782.9	$896.3	$985.5	$833.8
2020	10.00%	$463.7	$537.6	$537.6	$575.2	$603.1	$554.8
Total Capital Costs per kW		$4637.1	$5004.6	$5004.6	$6324.1	$7494.1	$5574.1
Total Capital Costs Billions of Dollars		$ 10.2	$ 11.0	$ 11.0	$ 13.9	$ 16.5	$ 12.3

[a] Overnight capital costs in $2010 of $4000 per kW of capacity and a 3% annual escalation rate in costs were used to derive the data in columns 2 and 3

[b] A 7% financing rate was used.

[c] A 12 % financing rate was used.

[d] An 8 % debt rate was used.

in them. As can be seen from this table, the total capital costs range from $5000 to $7500 per kW, depending upon how the financing costs are reported. Again, comparisons of cost estimates without detailed knowledge about what is included in them should not be done.

Environmental Costs and Regulations

In the long run, another important factor is the environmental costs of generating electricity from fossil fuel–fired and nuclear power plants. The fossil fuel–fired power plant's capital cost estimates shown in Table 2 include the expenses needed to meet sulfur dioxide and nitrogen oxide limits imposed by the Environmental Protection Agency. The estimates also include the costs of meeting current federal and state water discharge regulations. Additionally, in the analysis reported in Section 6, an explicit fee on carbon dioxide emissions will be included. The costs of all existing laws and regulations affecting fossil fuel–fired power plants, and an important proposed one are, therefore, included in the analysis.

Nuclear power has its own set of environmental costs—namely, the possibility of exposing the public to radiation, decommissioning, and radioactive waste disposal. The cost estimates of the units shown in Table 1 reflect designs that met NRC requirements as of about 2007 and 2008. These designs may, however, have to be changed because of additional NRC requirements, which could have cost implications. The design of the AP1000, a 1100-MW pressurized water reactor, was approved by the NRC in 2006. However, a number of changes to that design have been approved, and one additional change is currently under review by the NRC. Similarly, the design of the ABWR, a 1600-MW boiling water reactor developed by GE, was approved by the NRC in 1997. Currently, GE is in the process of renewing the NRC approval of that design. Additionally, in 2009, a number of design changes were submitted by GE to the NRC. The NRC is currently reviewing these changes.

Nuclear power decommissioning deals with the dismantlement of the plant and the decontamination of the site so it can be used for other purposes. The NRC must approve the utilities' plans, set residual radiation standards, and oversee the actual dismantlement of the plant. To insure that funds will be available when units are decommissioned, the state PUCs require that monies be placed in trusts. In the analysis presented in Section 6, it was assumed that decommissioning would cost $600 million and would occur 40 years after the plant's date of commercial operation.

The spent fuel from a light water reactor will be radioactive for millions of years, and the disposal of that waste is a major economic and political issue. There are two basic methods of disposing of the spent fuel from nuclear reactors: geological disposal and reprocessing/recycling. Reprocessing/recycling consists of extracting usable fuel from the waste and using it in other reactors. Currently, this is done in a few countries, most notably France and Japan. Needless to say, reprocessing/recycling is very controversial. It is very expensive and, at least in its current form, the reprocessed spent fuel could be used for military purposes. Thus, there are major proliferation concerns with reprocessing. Lastly, reprocessing has its own set of waste disposal problems.

With the passage of the Nuclear Waste Policy Act (NWPA) of 1982, the United States formally chose direct geological disposal of the waste. The initial act directed that the Department of Energy (DOE) study the feasibility of burying the waste at a number of sites. In 1988, however, Congress directed DOE to focus on just one site—Yucca Mountain, Nevada—and in 2003, the President formally chose Yucca Mountain as the country's high-level waste repository site. Under the NWPA, the state of Nevada could veto the President's decision, which they did. This veto was then overturned by Congress, and in 2008, the DOE submitted an application to the NRC for their approval to build the repository.

In 2010, the Obama Administration decided to stop work on Yucca Mountain, and DOE formally requested the NRC to permit them to withdraw the application. As of early 2011, the NRC has yet to publicly announce their decision. Additionally, the Administration abolished the Office of Civilian Radioactive Waste Management—the office within DOE that was managing the Yucca Mountain project. In response to the Obama Administration's actions, a number of lawsuits were filed by various states and localities in Federal Court. These lawsuits claim that DOE does not have the authority to unilaterally stop work on Yucca Mountain. As of early 2011, the ultimate outcomes of these lawsuits are unknown, and it is unclear what would happen if the courts ruled against DOE.

The NWPA requires that the government collect a fee of one mill (one-tenth of a cent) per kilowatt hour of electricity generated from nuclear power, and this charge is included in the analysis. Given the state of the U.S. spent fuel disposal policy, it is impossible to say anything about ultimate waste disposal costs. Regardless of what the ultimate cost is, at least for geological disposal, many of the costs will be incurred by future generations, and because of the "magic of compound interest," the total costs "today" will be small. To illustrate the effects of discounting, the yearly estimated costs of building and operating Yucca Mountain will be used. These costs were derived from a 2008 estimate prepared by DOE and the total undiscounted costs were about $100 billion. (Since Yucca Mountain has been abandoned, these estimates should be viewed as the cost of some hypothetical repository.) The cumulative costs are shown in Figure 1. As can be seen from this figure, roughly 50% of the expenditures will be made from year 50 to year 150, and the bulk of the expenditures will be made from year 30 to year 80. The point here is that the bulk of the costs of a geological repository will probably be incurred many years in the future.

Because of the complexities involved in geological disposal of the waste, it is quite possible that the ultimate cost of such a repository will be much greater than $100 billion. For argument's sake, suppose that the yearly expenditures are five times those shown in Figure 1. Figure 2 shows the present value "today" of the yearly expenditures using various discount rates. Using a 2% discount rate, which is very low, the present value "today" of the yearly expenditures is about $200 billion. Many private sector discount rates range from 7.5% to 10%, and if these rates were used, the present value "today" of $500 billion s incurred over a 150 years period would be less than $50 billion—roughly one tenth of the undiscounted costs.

Because of the discounting process, very large costs imposed on future generations will appear to be very small "today." Consequently, some intergenerational equity issues dealing with evaluating the back-end costs exist. Some economists have attempted to include intergenerational fairness considerations in discounting, but unfortunately, there is no consensus about how to incorporate equity issues in discounting.[14,15] Thus, there are some equity issues dealing with nuclear power that cannot be resolved with economic analysis.

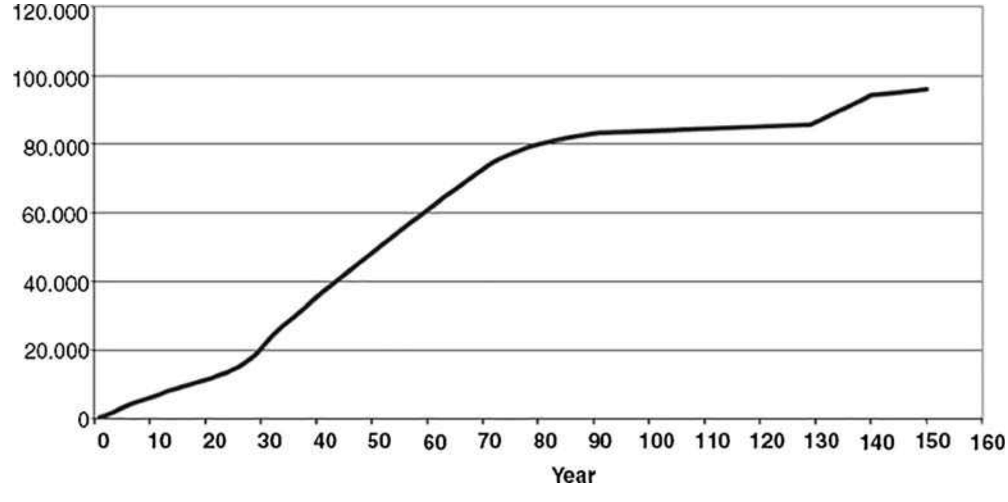


FIGURE 1 Cumulative cost of building and operating a hypothetical geological nuclear waste repository (millions of 2007 dollars).

Source: Adapted from Analysis of the Total System Life Cycle Cost of the Civilian Radioactive Waste Management Program, Fiscal Year 2007.[13]

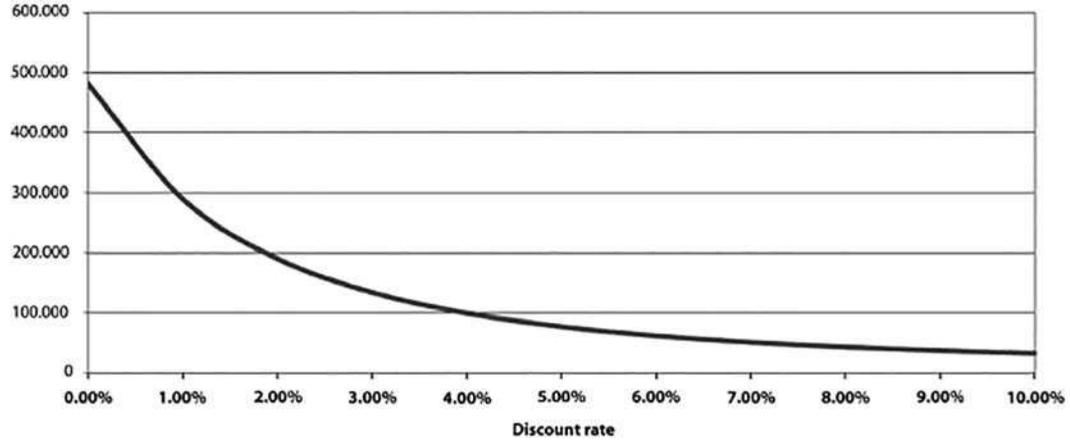


FIGURE 2 The effects of discounting on the cost of building and operating a hypothetical nuclear waste repository (present value of costs in millions).
Note: The costs are assumed to be 5 times the ones shown in Figure 1.
Source: Adapted from Analysis of the Total System Life Cycle Cost of the Civilian Radioactive Waste Management Program, Fiscal Year 2007.[13]

Discount Rates

The choice of the appropriate discount rate is another very important assumption in any analysis of the economics of nuclear power. To compare the economics of different plant types, costs that are incurred in the future must be discounted back of the present (“today"). This discounting process will account for the fact that “a dollar today is worth more than a dollar tomorrow.” Nuclear capital costs are much greater and the operating costs are much less when compared with fossil fuel–fired power plants, especially natural gas–fired units. Thus, a larger percentage of the total cost of building and operating a nuclear plant is incurred “upfront” when compared with a fossil–fired power plant. The higher the discount rate, the greater would be the weight placed on the “up-front” costs, and thus, higher discount rates would tend to favor natural gas and, to a lesser extent, coal-fired power plants relative to a nuclear unit.

Since most spreadsheets have routines that calculate present values, the mechanics of discounting are trivial. However, the same cannot be said for the choice of the discount rate. The discount rate should reflect the risk of the project—i.e., building and operating a power plant. There is a long literature on estimating discount rates, but unfortunately, in most cases, the necessary financial data (stock prices and bond yields) are at the utility and not at the project level.[16] This presents some problems when utilities own assets that have different risks. In particular, most but not all utilities own transmission and distribution facilities in addition to generating plants. The risks associated with building and operating the transmission/distribution (TD) system are very different from the risks of building and operating powerplants. Thus, the risks reflected in observed utility level data are some type of weighed average of the risks of the TD and generating plant assets. Additionally, the risks reflected in recent utility level data largely deal with operating factors because very few baseload power plants are currently under construction. Thus, project specific risks (i.e., discount rates) for large-baseload power plants cannot be estimated with actual data. Indeed, this is one of the classic problems in finance—the choice of a discount rate when a low-risk firm undertakes a high-risk investment.

Since the choice of a discount rate is largely judgmental, the analysis described in Section 6 will use two of them. (See Table 4.) Financial theory states that investors will require higher returns for bearing risks that are nonrandom.[16] (Financial economists often refer to nonrandom risk as nondiversifiable or systematic risk. Even when observable data can be used, the estimation of such risks is very difficult).

TABLE 4 Assumption Used in the Analysis

	Plant Type		
Assumption	Nuclear	Coal	Combined-Cycle Natural Gas
Unit size (mWe)	1,000	1,000	1,000
Capacity factor	85.00%	85.00%	85.00%
Heat rate	10,400	8,870	6,800
Overnight costs (dollars per kW of capacity)	NA	2300	850
Lead times (years)	6	4	3
Fixed O&M costs (dollars per kW of capacity)	96	51	23
Variable O&M (dollars per kwh)	0.0004	0.00357	
Fuel costs (mills per kwh)	7	NA	NA
Waste fee (mills per kwh)	1	NA	NA
Decommissioning (million $s)	600	NA	NA
CO2 fee for each dollar per metric ton CO2 (2009 $/mmBtu)[a]	0	0.095	0.053
Escalation/Inflation Rates:			
Inflation rate	2.00%	2.00%	2.00%
Annual O&M real escalation rates	1.00%	1.00%	1.00%
Annual fuel cost real escalation rate	0.50%	0.20%	1.70%
Annual real capital cost escalation rate	0.00%	0.00%	0.00%
Financial:			
Tax rate	37.00%	37.00%	37.00%
Debt fraction—higher rate	40.00%	40.00%	40.00%
Debt fraction—lower rate	60.00%	60.00%	60.00%
Cost of debt capital—higher rate	8.00%	8.00%	8.00%
Cost of debt capital—lower rate	6.50%	6.50%	6.50%
Cost of equity capital—higher rate	15.00%	15.00%	15.00%
Cost of equity capital—lower rate	11.00%	11.00%	11.00%
Weighted after-tax cost of capital—higherrate[b]	11.02%	11.02%	11.02%
Weighted after-tax cost of capital—lower rate[b]	6.04%	6.04%	6.04%

Source: Adapted from Du and Parsons.[7]

Note: All costs are in 2009 dollars. Except for the financial assumptions, the others were generally obtained from Du and Parsons.[7]

[a] These values were used to compute the increase in fuel costs per kilowatt hour because of a carbon fee. For example, the increase in fuel costs for a coal plant caused by a fee of \$20 per metric ton carbon was computed as follows: $20^{*}.095^{*}(8870/1000)=16.86$ mills per kwh.

[b] These values were used as the discount rates. These rates cannot be compared with the ones used in any study that excluded corporate income taxes.

This is because random risks can be eliminated by constructing portfolios of diversified assets. Thus, the discount rate will be a direct function of the level of nonrandom risk. The higher discount rate assumes that the nonrandom risks of building and operating any power plant are 50% greater than the risk for the average investment. The lower discount rate assumes that the nonrandom risks are about 20% less than the risk for the average investment. The higher and lower discount rates are consistent with the nonrandom risks observed in the airlines/telecommunications and manufacturing industries, respectively. The higher discount rate is also slightly greater than the one used for nuclear power in the MIT study, and the lower rate is slightly less than the one used for fossil fuel–fired power plants in the MIT study.[6]

To properly interpret the differences in the two rates, the distinction between risk reduction and risk shifting becomes important. One example of risk shifting, as opposed to risk reduction, is the cost-based rate-of-return regulation of utilities. Under this form of regulation, utilities can recover all the costs for projects that are ex ante prudently expended but ex post uneconomic. In such cases, some of the risks are being shifted from utility shareholders to electricity consumers. State-level rate-of-return regulation, therefore, does not reduce risks but rather shifts some of the risks to consumers. Another example is long-term fixed-price purchase power contracts between deregulated generating companies and the regulated transmission and distribution firms. Again, such contracts do not reduce risks but instead shift the risks associated with volatile wholesale electricity prices from utility shareholders to consumers. In short, the difference in the two discount rates shown in Table 4 reflects differences in the underlying risks of building and operating power plants and not who bears that risk.

As Table 3 shows, the total cost of building a 2200-MW-nuclear power plant could be in the range of $10 to $15 billion. Given the size of many U.S. utilities, the failure of a $10-billion project may result in the firm's bankruptcy or insolvency. Thus, many utilities who are planning to build nuclear power plants are attempting to dilute bankruptcy possibilities by forming joint ventures or, in one case, a merger. The discussion above abstracts from bankruptcy risk per se because it a function of the size of the project relative to the size of the firm along with the underlying risk of the project. In other words, the risk of a project is not a function of the firm's size but instead is determined by the variability in the underlying costs and revenues.

Cost of the Alternative

A fourth factor affecting the decision to build a nuclear power plant is the cost of the alternative. As was noted in Section 2, coal-fired power plant construction costs have escalated over the last few years. This cost growth has introduced some uncertainties related to the permanency of the increases. However, environmental considerations aside, most of the uncertainty related to the cost of generating electricity from fossil fuel–fired power plants deals with fuel prices. Historical and projected coal prices are shown in Figure 3. As this figure shows, after adjusting for inflation, coal prices fell until about the year 2000 and then increased by about 30% from 2005 to 2010. Two projections suggest that coal prices will remain relatively constant at about $2.00 per mmBtu over the 2010–2020 period, and the other one shows coal prices increasing to their 2009 levels. (In this section, all the prices are in 2009 dollar. Additionally, the first year of the unit's commercial operation is assumed to be 2020.)

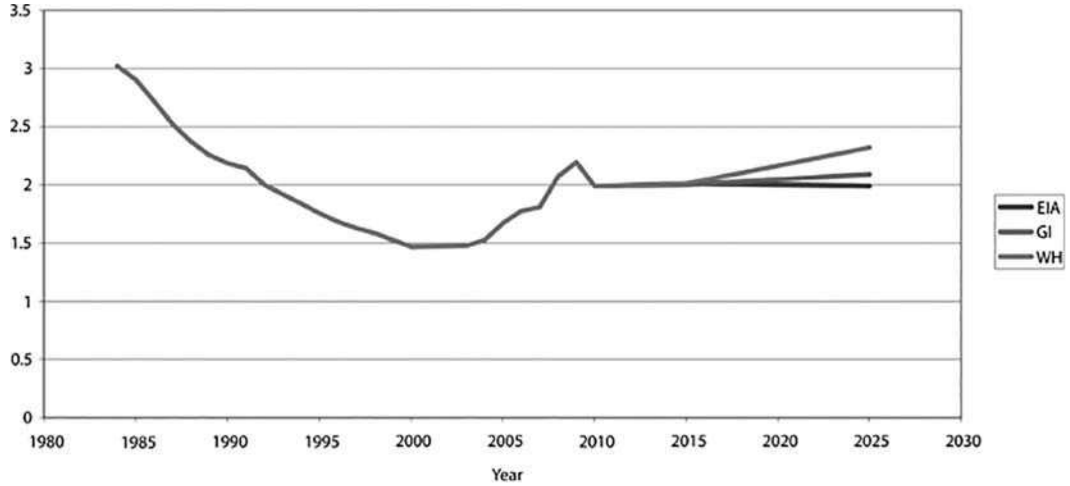


FIGURE 3 Delivered price of coal to electric utilities, 1984–2025 (2009 dollars per mmBtu).
Note: EIA, Energy Information Administration; GI, IHS Global Insights; WH, Wood MacKenzie Company. For the GI and WH projections, only the years 2015 and 2025 were reported.
Source: Annual Energy Outlook, 2010.[17]

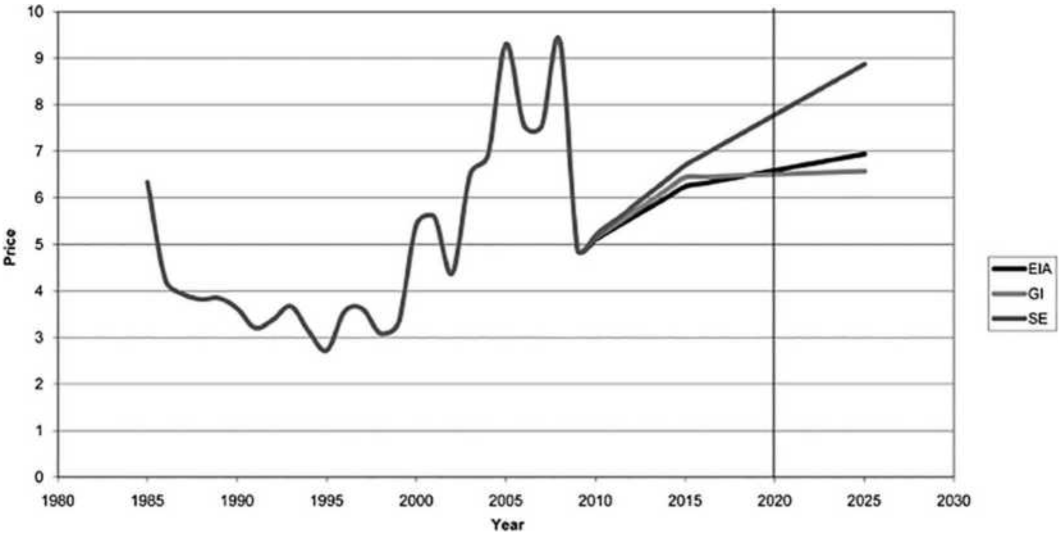


FIGURE 4 Delivered price of natural gas to electric utilities, 1985–2025 (2009 dollars per mmBtu).
Note: EIA, Energy Information Administration; GI, IHS Global Insights; SE, Strategic Energy and Economic Research. For GI and SE projections, only the years 2015 and 2025 were used.
Source: Annual Energy Outlook, 2010.[17]

Historically, natural gas prices have been much more volatile than coal prices (see Figure 4). After natural gas prices were deregulated in the early 1980s, they fell from more than $6 to about $3–4 mmBtu. In about 1998, natural gas prices began to increase, and by 2008, they increased to more than $9.00 mmBtu. Then, partly because of the recession, they fell by more than 100% to about $5 mmBtu. All three projections shown in Figure 4 have natural gas prices increasing to more than $6 mmBtu by about 2014. Over the 2015–2020 period, two of the three projections have natural gas prices increasing by relatively small amounts, whereas the third one has prices continuing to increase to more than $8.00 mmBtu by 2020.

As Table 5 shows, there are also wide regional variations in prices of coal and, to a lesser extent, natural gas. These variations are largely due to the cost of getting the fuel from the source of supply to the end users. Regions of the country that are far away from the source of supply tend to have higher costs, and the ones located close to the coal mines or gas fields have lower costs. Thus, the economics of nuclear power have a regional dimension.

TABLE 5 2010 Delivered Price of Coal and Natural Gas to Electricutilities (2009 Dollars per mmBtu)

Census Region	Coal	Natural Gas
New England	3.14	5.70
Mid Atlantic	2.23	5.03
South Atlantic	2.86	5.87
East North Central	1.88	4.30
East South Central	2.24	4.66
West North Central	1.25	4.62
West South Central	1.55	4.40
Mountain	1.55	5.00
Pacific	2.24	4.89
US average	1.99	4.85

Source: Annual Energy Outlook, 2010.[17]

Results

As was noted in Section 2, the cost of building nuclear power plants is highly uncertain. There are issues dealing with the permanency of current construction cost levels and also their accuracy. Wide regional variations in coal prices and uncertainty about future natural gas prices also exist. Because of these factors, point estimates of the levelized cost of generating electricity from coal, natural gas, and/or nuclear power plants have little value. The general approach used here is, therefore, to derive various combinations of nuclear overnight capital costs and "current" coal (natural gas) prices that would result in nuclear power and coal (natural gas) being equally economic—i.e., the levelized costs of coal (natural gas) and nuclear power are the same. The levelized cost is defined as the constant real price of electricity that would result in the net present value of the project (discounted revenues less discounted costs) equal to zero. The method used to compute the levelized costs can be found in the work of Du and Parsons.[7] Then, the combinations of fossil fuel prices and overnight nuclear capital costs that would result in nuclear power being economic (or uneconomic) could be determined. While this is not ideal, it is the best that can be done, given all the uncertainty.

To implement this approach, it was necessary to "fix" all the other variables affecting the economics of nuclear power, and unfortunately, there is some uncertainty in all of them. Some unreported sensitivity analyses suggested that variations in leadtimes, heatrates, and nonfuel Operations and Maintenance (O&M) costs had relatively minor effects on the basic conclusions of this analysis. Additionally, as was discussed in Section 2, there is uncertainty about the permanency of the increases in the coalfired power plant overnight construction costs. As will be seen shortly, fixing the coal capital costs will not have any major impact on the basic conclusions of this analysis.

The results of a comparison of the economics of nuclear power relative to coal-fired powerplants using the higher discount rate are shown in Figure 5a. (In this section, except where noted, all costs and prices are in 2009 dollars.) The solid line shows the combinations of overnight nuclear capital expenses and current coal prices that would result in both plant types having the same levelized cost. Thus, any combination of overnight nuclear capital costs and coal prices that fall in the region denoted as E would result in nuclear power being economic. Similarly, any combination of these two factors that fall in the region U would result in nuclear power being uneconomic relative to coalfired power plants.

As was noted above, overnight nuclear capital costs derived from a small number of PUC filings averaged about $4000 per kW. Given overnight capital costs of $4000 per kW, projected coal prices would have to be about $6 per mmBtu in 2020 before nuclear power would be competitive with coal-fired power plants (point A in Figure 5a). These coal prices are about two and a half to three times their 2010 levels and are also much greater than the projected values. Coal prices of $6 per mmBtu are also much greater than their 2010 levels in regions of the United States that are not close to the coal reserves.

Similarly, given projected coal prices in the midterm of about $2 mmBtu, nuclear overnight capital costs would have to fall to levels roughly comparable with the ones for the coal-fired power plants, before nuclear power would be economic. Given $2 per mmBtu coal prices and the nuclear operating cost assumptions shown in Table 4, the operating costs of both plant types are about the same. Thus, for nuclear power to be competitive with coal-fired powerplants, the overnight capital costs of the two technologies would have to be similar (point B in Figure 5a).

The results of the comparison of nuclear power to coalfired power plants using the lower discount rate are shown in Figure 5b. These results are qualitatively similar to the ones using the higher discount rate. That is, if the overnight nuclear capital costs would be about $4000 per kW, coal prices in 2020 would have to be much greater than the ones reported in a number of recent studies. Additionally, using projected coal prices in 2020 of about $2 per mmBtu, overnight nuclear capital costs would have to fall to about $2500 per kW before nuclear power would be competitive with coal-fired power plants.

To summarize, if nuclear overnight capital cost estimates found in a number of recent PUC filing are at all indicative of what it would actually cost to build a nuclear power plant, and if the external costs of carbon dioxide (CO_2) emissions are ignored, nuclear power is not competitive with efficient coal-fired

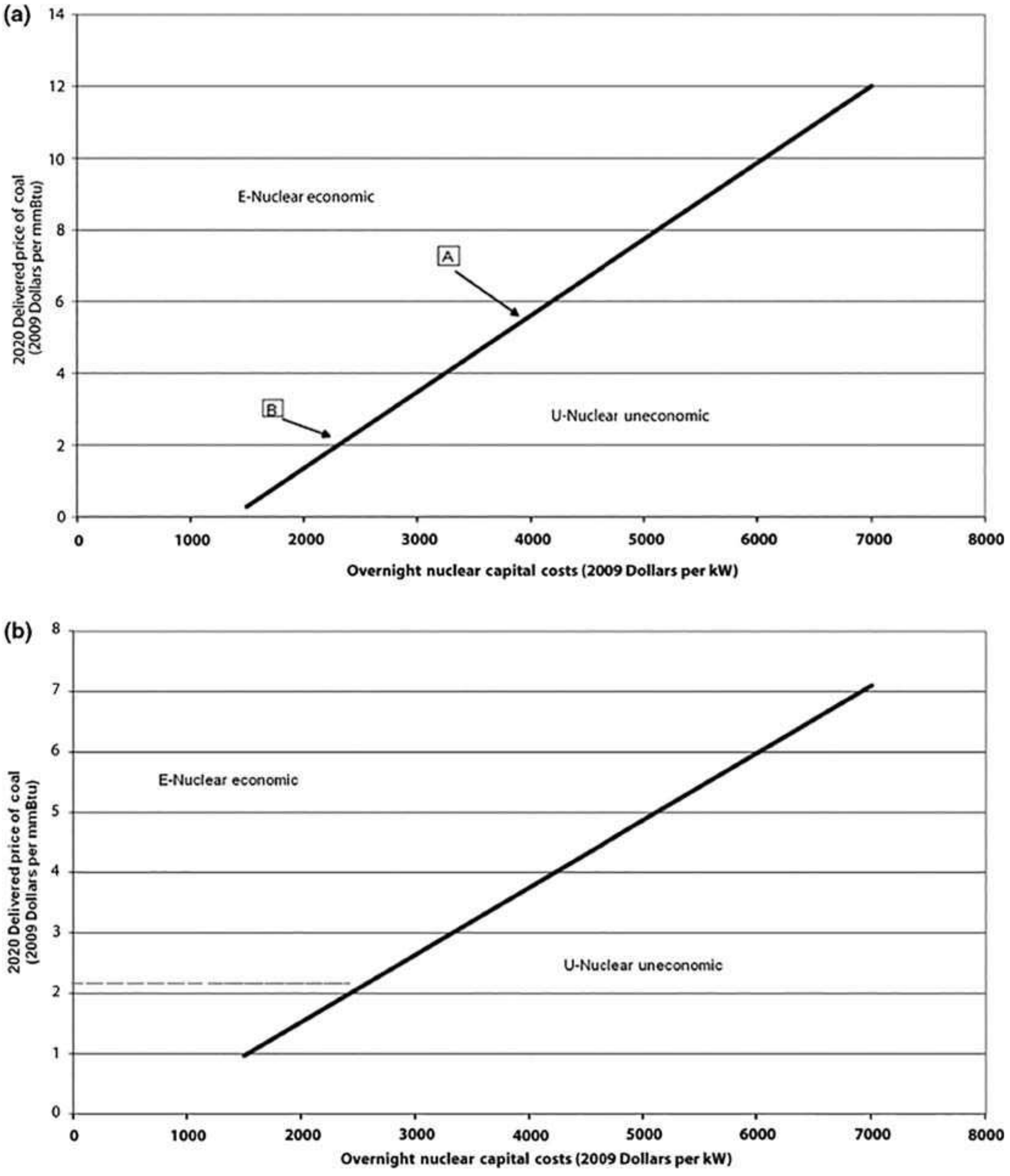


FIGURE 5 The economics of a nuclear power plant relative to a coal-fired power plant: (a) higher discount rate (b) lower discount rate.

Note: The solid line shows the combinations of 2020 coal prices and nuclear capital costs that would produce the same levelized cost for both plants. The combinations of coal prices and nuclear capital costs that would fall in the region E (U) would produce levelized costs that are lower (higher) for the nuclear plant than for the coal plant.

power plants. Given nuclear capital costs of about $4000 per kW, coal prices would have to increase substantially before nuclear power would be competitive with coal-fired power plants. Also, given the assumptions shown in Table 4, these conclusions do not depend upon the assumed discount rate. As will be seen shortly, the same is not true for comparisons of nuclear power with natural gas–fired power plants.

The economics of nuclear power relative to an efficient combined-cycle natural gas–fired power plant using the higher discount rate is shown in Figure 6a. Given overnight nuclear capital costs of $4000 per kW, natural gas prices would have to increase to about $9.50 per mmBtu before nuclear power would be economic relative to gas–fired power plants (point A in Figure 6a). Natural gas prices of over $9 per mmBtu in 2020 are much higher than their 2010 levels and are also higher than their projected levels shown in Figure 4. Similarly, using the higher discount rate and 2020 natural gas prices

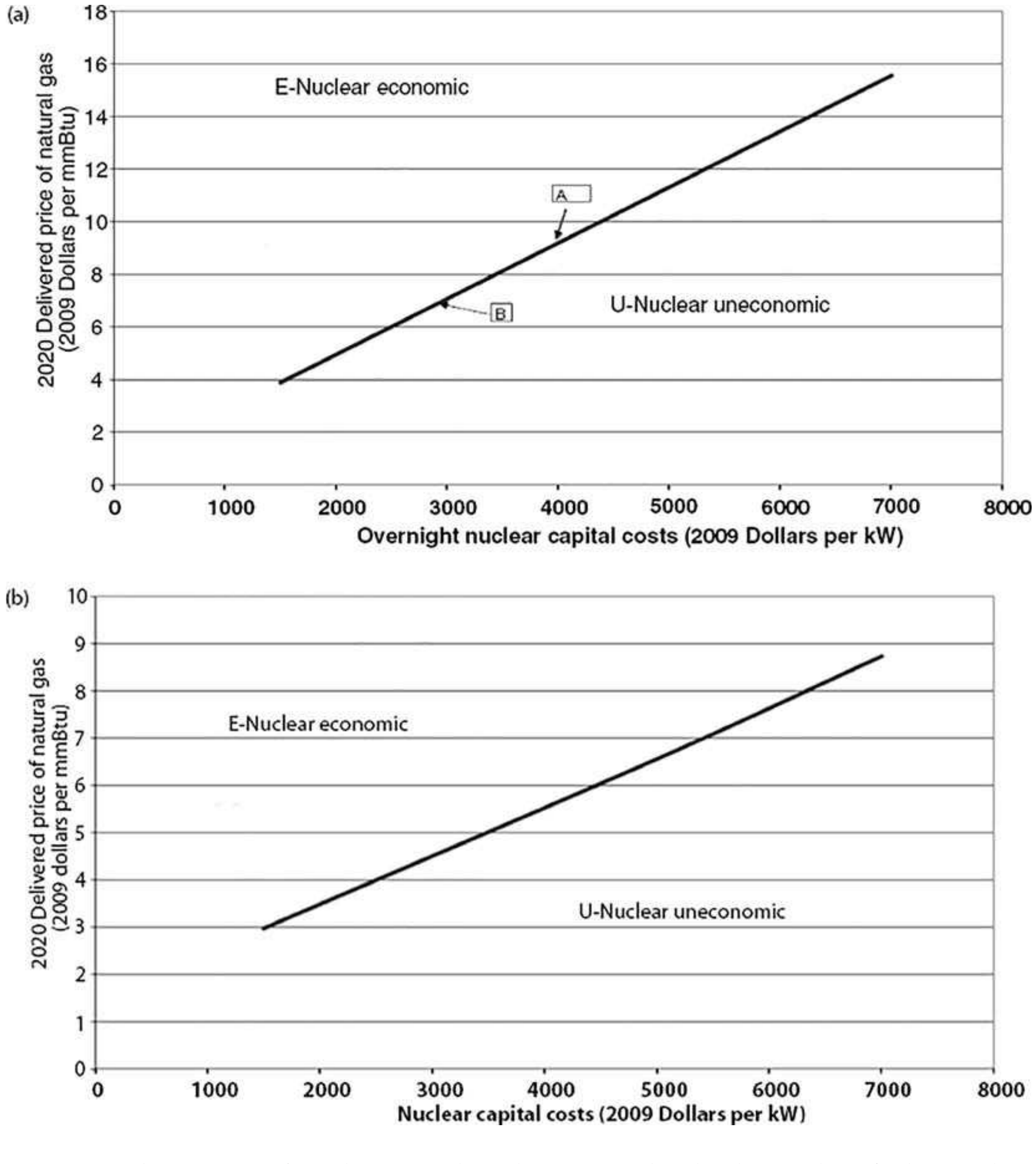


FIGURE 6 The economics of a nuclear power plant relative to a combined cycle natural gas–fired power plant: (a) higher discount rate (b) lower discount rate.

Note: The solid line shows the combinations of 2020 natural–gas prices and nuclear capital costs that would produce the same levelized cost for both plants. The combinations of gas prices and nuclear capital costs that would fall in the region E (U) would produce levelized costs that are lower (higher) for the nuclear plant than for the gas–fired plant.

of about \$7 per mmBtu, overnight nuclear capital costs would have to be about \$3000 per kW before nuclear power would be competitive with natural gas-fired power plants.

Given overnight nuclear capital costs of \$4000 per kW, roughly 75% of the total cost of generating electricity from the nuclear unit is incurred upfront in terms of construction costs. However, about 75% of the total cost of generating electricity from a combined-cycle natural gas–fired plant is incurred in future years. Again, the lower the discount rate, the lesser will be the weight placed on the up-front costs, and thus, lower discount rates would favor nuclear power. The comparison of the economics of nuclear power plants relative to gas–fired power plants using the lower discount rate, shown in Figure 6b, suggests that this is the case. Using the lower discount rate and nuclear capital costs of about \$4000 per kW, natural gas prices would have to increase from their 2010 levels of about \$5.00 to about \$6.00 per mmBtu by 2020 before the nuclear power plant would be economic. Natural gas prices of about \$6 per mmBtu in 2020 are slightly lower than two of the three projections shown in Figure 4. Thus, if building and operating any power plant is perceived to be a relatively low-risk endeavor, even if the external costs related to CO_2 emissions are ignored, nuclear power is marginally economic relative to combined-cycle natural gas–fired power plants.

This result was based on the assumption that real natural gas prices would increase at an annual rate of .75% from 2021 to 2061. At least through 2035, this escalation rate is relatively high. If a lower escalation rate in annual real natural gas prices of 0.3% were assumed, and given nuclear overnight capital costs of \$4000 per kW, 2020 real natural gas prices would have to be about \$7 per mmBtu before a nuclear power plant would be competitive with a combined-cycle natural gas–fired power plant. Natural gas prices of \$7 per mmBtu in 2020 are greater than two of the three projection shown in Figure 4. Thus, using lower discount rate, nuclear power would be extremely marginally competitive with combined-cycle natural gas–fired power plants.

As was noted above, the fossil fuel capital costs include the expenses needed to meet current sulfur dioxide and nitrogen oxide emission levels. However, the external costs related to CO_2 emissions have not been included. The size of these costs is, however, highly uncertain. Indeed, the extent to which CO_2 emissions have resulted in global warming and the associated costs of the warming of the atmosphere are still being debated. However, one recent study recommended using external costs in 2020 of about \$7 to \$42 or perhaps about \$80 per metric ton of CO_2 emissions if the Earth warms faster than expected.[18] Additionally, some recent analyses of the effects of recently proposed U.S. "cap-and-trade" programs estimated that in 2020, the "price" of CO_2 emissions would range from about \$20 to \$100 per metric ton in 2020.[19] Lastly, in Europe's CO_2 cap-and-trade program, over the last few years, CO_2 emission permits have been trading in the range of 15€–20€ per metric ton.

Because of these uncertainties, the best that could be done is to compute the combinations of nuclear capital costs and CO_2 prices that would result in nuclear power being economic relative to coal and combined-cycle natural gas–fired power plants. This, of course, requires that coal and natural gas prices in 2020 be fixed at \$2.00 and \$7.00 per mmBtu, respectively. The results of this exercise for coal and natural gas–fired power plants are shown in Figures 7 and 8, respectively. As before, both the higher and lower discount rates were used.

These results suggest that if nuclear power construction costs were about \$4000 per kW, CO_2 prices in 2020 would have to exceed \$20–\$40 per metric ton before nuclear power would be economic relative to coal-fired power plants (see Figure 7). These CO_2 prices are in the same range as ones used in a number of discussions. Thus, once the external costs related to CO_2 emissions from very efficient coal-fired power plants are considered, nuclear power could very well be economic relative to coal-fired power plants. However, using the higher discount rate, the price of CO_2 emissions would have to be somewhat higher (about \$50–\$60 per ton) before nuclear power would be competitive with combined-cycle natural gas–fired power plants. As was just noted, using the lower discount rate, nuclear power would be very marginally economic relative to gas–fired power plants even if the price of carbon were zero.

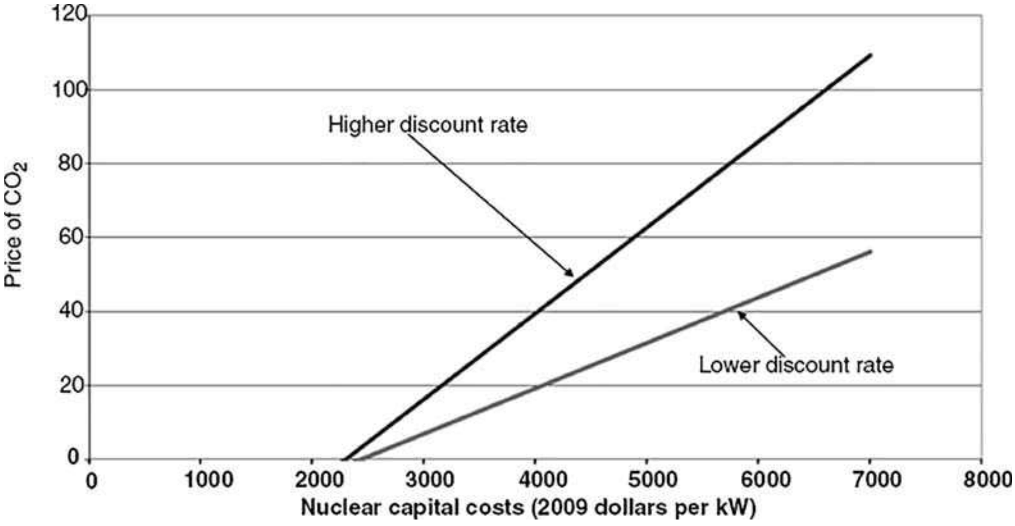


FIGURE 7 The economics of a nuclear power plant relative to a coal-fired power plant: costs of CO_2 emissions included.

Note: The solid line shows the combinations of CO_2 prices and nuclear capital costs that would produce the same levelized cost for both plants. The combinations of coal prices and nuclear capital costs that would fall in the region above (below) or to the left (right) would produce levelized costs that are lower (higher) for the nuclear plant than for the coal plant. Real coal prices in 2020 were assumed to be about $2.10 per mmBtu. The carbon prices are in 2009 dollars per metric ton.

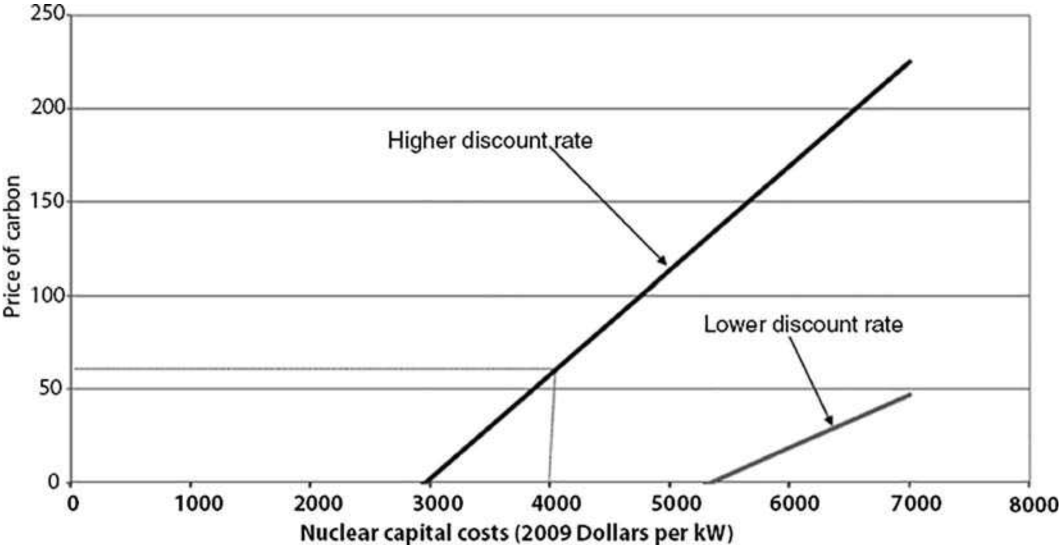


FIGURE 8 The economics of a nuclear power plant relative to a combined-cycle gas–fired power plant: cost of CO_2 emissions included.

Note: The solid line shows the combinations of CO_2 prices and nuclear capital costs that would produce the same levelized cost for both plants. The combinations of natural gas prices and nuclear capital costs that would fall in the region above (below) or to the left (right) would produce levelized costs that are lower (higher) for the nuclear plant than for the gas plant. Real natural prices in 2020 were assumed to be about $7.00 per mmBtu. The carbon price is in 2009 dollars per metric ton.

Conclusions

This entry attempted to answer the following question: Is nuclear power economic? Two major factors will influence the economics of nuclear power, and since there are major uncertainties with both of them, it is impossible to give an unqualified answer. The most obvious one is the cost of building a nuclear power plant. In a number of recent PUC rate cases, overnight nuclear power plant capital costs of around $4000 per kW were used. If these estimates are indicative of realized overnight costs, environmental considerations aside, then nuclear power would generally not be economic. The analysis in this entry suggests that overnight nuclear power plant capital costs would have to fall to between $2500 and $3000 per kW before nuclear power would be economic. The one exception to this conclusion is if a very low discount rate is used.

The second factor is the cost of the alternative, which would be either coal-fired power plants or combined-cycle natural gas–fired units. The major uncertainty here is the enactment of policies limiting CO_2 emissions. The economics of nuclear power would be greatly improved if all of the external costs related to global warming were included in the cost of generating electricity from fossil fuel–fired power plants. It must be noted that nuclear power has its own set of environmental costs in the form of nuclear waste disposal and decommissioning. Since these costs will be incurred over hundreds of years, because of discounting, the back-end costs are very small "today." Some have objected to discounting expenses incurred over very long time periods because this procedure represents a strong incentive to impose large costs on future generations. However, such equity considerations are outside the realm of economic analysis.

Lastly, there are probably at least two reasons why some utilities are interested in nuclear power even though the economics appears to be unfavorable. First, recent research using data from the United Kingdom found that nuclear power can be used as a hedge against volatile natural gas prices if the utility was subject to price regulation.[20] It is possible that utilities are using nuclear power as a hedge. This might partially explain why most of the utilities interested in nuclear power are located in states that have not been deregulated. Second, a number of utilities that are interested in nuclear power and their state-level regulators are assuming that eventually some type of explicit carbon price will be enacted. The analysis in the present entry does show that the economics of nuclear power is improved considerably when the cost of CO_2 emissions from fossil fuel–fired power plants is included.

Addendum—The Impact of the Fukushima Nuclear Disaster on the Cost of Nuclear Power

On March 11, 2011, as a result of a major earthquake and tsunami in Japan, the Fukushima nuclear power plant lost all of its on-site and off-site power, which is needed to operate the plant's safety systems. At least as of May 2011, it appears that the earthquake damaged the transmission facilities needed to supply off-site power, and the tsunami damaged all of the on-site power sources, including the emergency on-site backup diesel generators. As a result, without any source of power, the safety systems could not cool the reactor, and the fuel literally melted. The spent fuel is submerged in water in large storage pools, which cool the high-level waste. When the plant lost all power, and water could not cool the spent fuel, this resulted in a number of explosions. Because of these explosions and attempts to cool the reactor and spent fuel from off-site water sources, significant amounts of radiation were discharged into the air and water surrounding the plant.

As of May 2011, the owner of Fukushima and the Japanese government estimates that the plant will not be stabilized (cooled) until January 2012. Thus, it could take months or even years before authorities can determine what exactly happened to the plant. Indeed, it took 5 years before workers could enter the Three Mile Island plant to determine the exact nature of that accident. Consequently, at this point, it is impossible to determine with any degree of confidence what the impact of the Fukushima disaster will be on the cost of nuclear power. Nevertheless, it is possible that the disaster could affect the economics of nuclear power in a number of ways. First, there could be design changes to existing and new power plants that will increase capital costs, especially in the areas of backup power and spent fuel storage. Second, the disaster will clearly affect public acceptance of nuclear power, and the licensing hearings at the NRC could become more controversial, which would increase leadtimes. Third, in other research, I found evidence of a very small (1 to 2 percentage points) risk premium on the common stock of nuclear U.S. utilities resulting from the accident at Three Mile Island.[21] Over time, capital markets could have similar reactions to the Fukushima disaster. This would increase the cost of financing the construction and operation of a nuclear plant.

Acknowledgments

I would like to thank my past and present colleagues at the Energy Information Administration for their many useful discussions with me about the economics of nuclear power and their assistance in my research in studying this subject. However, the views and opinions stated in this entry are the author's alone and do not represent the official position of the Energy Information Administration or the United States DOE. All errors are the sole responsibility of the author.

References

1. Merrow, E.; Phillips, K.; Myers, C. *Understanding Cost Growth and Performance Shortfalls in Pioneer Process Plants;* The Rand Corporation: Santa Monica, CA, 1981.
2. Wiser, R.; Bolinger, M. *2008 Wind Technologies Market Report;* Lawrence Berkley National Laboratory: Berkley CA, 2009, available at http://www.nrel.gov/analysis/pdfs/46026.pdf (accessed April 2011).
3. *Annual Energy Outlook: 2004,* DOE/EIA-0383 (2004); U.S. Department of Energy, Energy Information Administration: Washington, DC, 2004.
4. *Report of the Interagency Task Force on Carbon Capture and Storage,* DOE/FE-0001 (2010); U.S. Department of Energy, Assistant Secretary for Fossil Energy: Washington, DC, 2010, available at http://fossil.energy.gov/programs/sequestration/ccstf/CCSTaskForceReport2010.pdf. (accessed April 2011).
5. *Projected Costs of Generating Electricity: 2010 Update;* Nuclear Energy Agency-International Energy Agency, Organization of Economic Co-operation and Development: Paris, France, 2010.
6. Deutch, J.; Moniz, E.; Joskow, P. *The Future of Nuclear Power*; Massachusetts Institute of Technology: Cambridge, MA, 2003, available at http://web.mit.edu/nuclearpower/pdf/nuclearpowerfull.pdf (accessed April 2011).
7. *World Energy Outlook: 2008;* International Energy Agency: Paris, France, 2008.
8. Du, Y.; Parsons, J. *Update of the Cost of Nuclear Power,* Working Paper 09-004; Center for Energy and Environmental Research, Massachusetts Institute of Technology: Cambridge, MA, 2009.
9. McCabe, M. Principals, agents, and the learning curve: The case of steam electric power plant construction. *J.Ind. Econ.* **1996**, *20* (4), 240–270.
10. Cantor, R.; Hewlett, J. The economics of nuclear power: Further evidence of learning, economies of scale and regulatory effects. *Resour. Energy* **1988**, *10* (4), 315–335.
11. Hewlett, J. Why were the nuclear power plant cost and leadtime estimates so wrong?. In *Nuclear Power at the Crossroads;* Lowinger, T., Hinman, G., Eds.; University of Colorado Press: Bolder CO, 1994; 121–148.
12. Deutch, J.; Moniz, E. *The Future of Coal;* Massachusetts Institute of Technology: Cambridge, MA, 2007, available at http://web.mit.edu/coal/ (accessed April 2011).
13. *Analysis of the Total System Life Cycle Cost of the Civilian Radioactive Waste Management Program, Fiscal Year 2007,* DOE/RW-0591; U.S. Department of Energy, Office of Civilian Radioactive Waste Management: Washington, DC, 2008.

14. Portney. P.; Weyant, J. *Discounting and Intergenerational Equity;* Resources for the Future: Washington, DC, 1999.
15. Sumaila, U.; Walters, C. Intergenerational discounting: A new intuitive approach. *Ecol. Econ.* **2005,** *10* (2), 135–142.
16. Copeland, T.; Weston, F. *Financial Theory and Corporation Policy,* 3rd Ed.; Addison-Wesley: New York, 1988.
17. *Annual Energy Outlook, 2010,* DOE/EIA-0383 (2010); Energy Information Administration: Washington, DC, 2010.
18. Parry, I.; Williams, R. Is a carbon tax the only good climate policy. In *Resources*; Resources for the Future: Washington DC, Fall 2010, 176; 38–41, available at http://www.rff.org/resourcesno176/parry (accessed April 2011).
19. *Energy Market and Economic Impacts of the American Power Act of 2010,* SR/OIAF/2010-01; U.S. Department of Energy, Energy Information Administration: Washington, DC, 2010.
20. Roques, F.; Newbery, D. Nuclear power: A hedge against uncertain gas prices and carbon prices. *Energy J.* **2006**, *27* (4), 72–95.
21. Hewlett, J. *Investor Perceptions of Nuclear Power*, DOE/ EIA-0446; Energy Information Administration, U.S. Department of Energy: Washington, DC, 1984.

15

Remote Sensing and GIS Integration

Egide Nizeyimana

Compatibility Issues between Remote Sensing and GIS

A GIS can be defined as a set of computer tools for capturing, storing, analyzing, and displaying spatially referenced data. The data within a GIS consists of two elements: spatial entities represented by points (e.g., well locations), lines (e.g., streams, road networks), and polygons (e.g., soil delineations) and attribute data or information that describes characteristics of the spatial features. The spatial entity is referenced to a geographic coordinate system and is stored in either a vector or raster model. GIS is primarily a platform that integrates spatial information from variable data sources and provides tools to overlay and analyze it.

RS is compatible with GIS because the information acquired by imaging sensors carried aboard satellites and airplanes is geographic in nature, referenced to a known coordinate system, and in a grid layout, a raster data model commonly found in most popular GIS software packages. In a raster model environment, the analysis is performed pixel by pixel. Current commercial GIS-based software packages have vector and raster capability analyses and allow data conversion from one data model to the other. Nonimaging systems are also important to RS/GIS integration because they provide point data, which at georeferenced locations, are often combined with other environmental data for site-specific assessments.

Linking Remote Sensing and GIS

Remote Sensing as a Source of Spatial Data

RS is one of the most important sources of land use/cover information used in GIS analyses. It provides information on the location and spatial and temporal distribution of land cover on the Earth's surface. Land use/cover distributions may be derived from aerial photographs after they are corrected for relief displacement and distortions caused by camera angle, and registered to a coordinate system. Delineations of different land use/cover types on photographs are made from visual interpretations of characteristics (e.g., tone, texture and color) aided by optical devices. Digital maps of these classes are created by digitizing and processing boundaries between land uses or by scanning photographs covering the area of interest and screen-digitizing their boundaries.

Land use distributions on the Earth's surface may also be obtained using RS imaging systems. In this case, RS provides opportunities in the area of land use planning which would not be otherwise available. While aerial photographs are effective and appropriate for analyses of small areas, RS offers tremendous advantages when planning for large areas such as river basins and regions. The fact that the land surface is observed from reflected/emitted energy for large areas and over a wide range of wavelengths allows for easy differentiation of existing land uses (e.g., wetlands, degraded landscapes).[1] Earth observation satellites also offer a repetitive coverage of the land, thus providing the possibility of monitoring land use pattern changes over time. Gathering information on land uses at several time intervals is particularly important in the monitoring and land evaluation stages of the land use planning process. Factors such as plant stress, crop growth and yields that serve as measures of agricultural productivity can be rapidly estimated following digital processing and analysis of RS imagery. RS has been used in many instances to monitor land surface conditions such as soil degradation and soil salinity.[2] It is true that RS cannot replace field mapping of land use/cover. However, RS supplements provide information that would not otherwise be available to land use planners and managers. There is no doubt that the development of sensors of high spatial resolution (1 m and higher) presently in orbits would increase the use of RS in land use planning and management.

As indicated before, RS and GIS technologies are highly compatible primarily because of the nature of RS as a source of spatial land use/land cover used in various environmental applications. Remote sensed imagery is a grid cell layer, a data format easily handled by GISs. In the RS/GIS integration, GIS appears as a platform that stores and integrates spatial data from different sources including remote sensing and output information needed for environmental analyses depending on the type of analysis sought. Land use/cover distributions derived from RS are input to GIS, GIS-based systems (SDSS), and models (e.g., hydrologic/water quality, crop yields, primary productivity) commonly used in land evaluation for land use planning and management. GIS has been proven to be a valuable tool in integrating RS-derived data with climate and land surface parameters (soils, terrain, etc.) to generate digital maps of ecosystem productivity or vulnerability to environmental factors.[3] The parameterization of the Terrestrial Ecology Model (TEM) and a forest ecosystem model using GIS- and RS-derived data at regional scales has also been accomplished.[4,5] The RS/GIS approach has been adopted by many state and federal agencies involved in environmental assessments. For example, in the GAP Analysis Program, the U.S. Geological Survey (USGS) and collaborative institutions map and/or model potential natural habitats of native vertebrate species from remotely sensed data across the country, and use GIS map overlay tools to determine the degree of richness of habitats in these species.[6]

Finally, GIS and RS applications have promoted the development of other high-resolution spatial technologies that enhance the land use assessment and planning. Some of these are the global positioning system (GPS) and digital orthophoto quadrangles (DOQ). GPS allows the user to record accurately and rapidly geographic coordinates of any location in the field with precisions ranging from several meters to a centimeter. Soil, terrain and land use attributes observed or measured can then be input to a GIS along with their precise locations and extent. GPS coupled with a GIS can improve the accuracy of land quality mapping by increasing the spatial variability of soil and landscape attributes. DOQs, on the other hand, are digital images of aerial photographs that were corrected to remove relief displacement and distortion caused by the camera angle. The USGS distributes single-band, 256-scale, gray-scaled DOQs at 1 m grid resolution. Although DOQs have not been used extensively in land use planning in the past, their high resolution and photograph-like characteristics make them a potential source of data in this area. Furthermore, repetitive DOQs acquired at different times can be used to monitor the magnitude of changes in land use over time. The relationship between GIS, RS, and other data sources is shown in Figure 1.

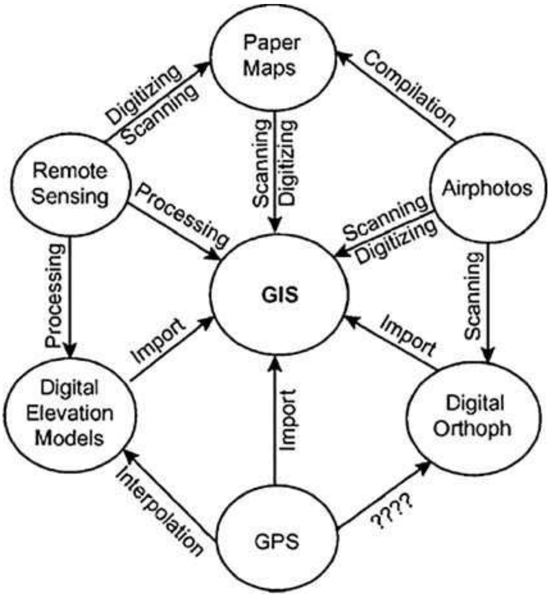


FIGURE 1 Relationships between remote sensing and GIS.
Source: Adapted from Nizeyimana and Petersen.[7]

Applications of Remote Sensing/GIS in Soil Science

The need for geospatial data for use in various spatial database development and analyses in industry, government, and universities have increased the demand for remotely sensed data. In the area of soil science, the remote sensing technology is used to acquire data, and digitally process and analyze it. Subsequent analyses may involve error assessment, and data conversion from raster to vector format and vice versa before the data is merged with other datasets in GIS for specific analyses. Table 1 summarizes remote sensing techniques used in various aspects of soil science and potential limitations for use in GIS and advantages of each. These are laboratory approaches, field methods, and aircraft-/satellite-based methods. The latter involves interpretation and analyses of digital images, color composites, or radiances.

Remote Sensing/GIS integration for Site-Specific Farming

The goal of site-specific farming (SSF) or precision farming is to optimize the profitability of a farm practicing variable management according to soil conditions found at each site. The concept is based primarily on the fact that soil chemical and physical properties that affect crop production (e.g., pH, nutrients, available water, impeding layers) vary spatially across agricultural landscapes Management practices such as fertilizer and pesticide applications, irrigation water, and crop varieties should, therefore, be prescribed according to this soil variability. The profitability associated with variable rate applications of SSTs should avoid or reduce waste and the risk for environmental pollution because these agrochemicals are applied to the field only in amounts needed for optimal crop growth.

In an SST, the real-time detection of continuous soil variable is made possible by mobile devices including sensors and differential GPS units while agrochemicals are applied using variable rate field applicators mounted on farm equipments. These sensors are based on the same principles as RS and have been developed to detect, directly or indirectly, soil properties such as soil moisture, soil texture, nitrates, etc. The data acquired, along with their respective locations, are often integrated with spatial information from other sources such as soil databases, climate, landscape, and satellite/airborne remote sensing to delineate meaningful management zones within the farm.

TABLE 1 Documented Uses of Remote Sensing in Various Areas of Soil Science

Application	Advantages	Potential Limitations	References
Laboratory methods	Provides accurate measurements of reflectance values	Provide point data rather than areal extent of soil properties	[8]
Field methods	Easily related to on-site conditions	Provide limited data coverage; measurements are affected by soil conditions (soil roughness, moisture), sun angle, etc.	[9]
Airborne -/ satellite-based methods	Provides areal extent and temporal coverage of soil properties	Radiance measurements affected by atmospheric and soil-surface conditions	(–)
a) Interpretation of digital images	Good results when data are well analyzed and interpreted	Image analysis can be costly; requires experienced technician for good results; field verification or prior knowledge of area required for best results	[10]
b) Interpretation of color composites	Easy to use and rapid interpretations	Intensive field verification is needed for good results due to the fact that interpretation is based on differences in tone and physical characteristics of objects; requires experienced interpreter for good results	[11]
c) Interpretation of radiances	Relatively easy, cheap, quantitative; data can be normalized to remove environmental effects	Results often unreliable	[12]

Source: Adapted from Nizeyimana and Petersen.[13]

References

1. Petersen, G.W.; Nizeyimana, E.; Evans, B.M. Applications of geographic information systems in soil degradation assessments. In *Methods for Assessment of Soil Degradation*; Lal, R., Blum, W.H., Valentine, C., Rose, B.A., Eds.; Advances in Soil Science; CRC Press: Boca Raton, FL, 1997; 377–391.
2. Raina, P.; Joshi, D.C.; Kolarkar, A.S. Mapping of soil degradation by remote sensing on Alluvial plan, Rajasthan, India. Arid Soil Res. Rehab. **1993**, *7*, 145–161.
3. Parrish, D.A.; Townsend, L.; Saunders, J.; Carney, G.; Langston, C. USEPA Region 6 Comparative Risk Project: Evaluating Ecological Risk. EPA Unpublished Report, 1993.
4. Pan, Y.; McGuire, A.D.; Kicklighter, D.W.; Melillo, J.M. The importance of climate and soils for estimates of net primary production: a sensitivity analysis with the terrestrial ecosystem model. Global Change Biol. **1996**, *2*, 5–23.
5. Lathrop, R.G., Jr.; Abler, J.D., Jr.; Bognar, J.A., Jr. Spatial variability of digital soil maps and its impact on regional ecosystem modeling. Ecol. Model. **1995**, *82,* 1–10.
6. Scott, J.M.; Davis, F.; Csuti, B.; Noss, R.; Butterfield, B.; Groves, C.; Anderson, H.; Caicco, S.; D'Erchia, F.; Edwards, T.C., Jr.; Ulliman, J.; Wright, R.G. GAP Analysis: a geographic approach to protection of biological diversity. Wildlife Monogr. **1993**, *123,* 1–41.
7. Nizeyimana, E.; Petersen, G.W. Land use planning and environmental impact assessment using GIS. Environmental Modeling Using GIS and Remote Sensing. 2001.
8. Stoner, E.R.; Baumgardner, M.F. Characteristic variations in reflectance surface of soils. Soil Sci. Soc. Am. J. **1981**, *45,* 1161–1165.

9. Gausman, H.W.; Leamer, R.W.; Noriega, J.R.; Rodriguez, R.R.; Wiegand, C.L. Field-measured spectrometric reflectance of disked and non-disked soil with and without wheat straw. Soil Sci. Soc. Am. J. **1977**, *41,* 493–496.
10. Connors, K.F.; Gardner, T.W.; Petersen, G.W. Digital analysis of the hydrologic components of watersheds using simulated SPOT imagery. Proceedings of Workshop on Hydrologic Applications of Space Technology; IAHS: Cocoa Beach, FL, 1985; Vol. 160, 355–365.
11. Bocco, G.; Palacio, J.; Valenzuela, C.R. Gully erosion modeling using GIS and geomorphologic knowledge. ITC J. **1990**, *3,* 253–261.
12. Pickup, G.; Nelson, D.J. Use of landsat radiance parameters to distinguish soil erosion, stability, and deposition in Arid Central Australia. Remote Sens. Environ. **1984**, *16,* 195–209.
13. Nizeyimana, E.; Petersen, G.W. Remote sensing applications to soil degradation assessments. In *Methods for Assessment of Soil Degradation;* Lal, R., Blum, W.H., Valentine, C., Rose, B.A., Eds.; CRC Press: Boca Raton, FL, 1997; 393–405.

16

Solid Waste: Municipal

Angelique
Chettiparamb

Introduction

Waste is normally understood as something that has no value. The *Oxford Dictionary* defines waste as "unwanted or unusable material, substance, or by-product" (http://oxforddictionaries.com/definition/waste). The United Nations Statistical Division defines it as

> Wastes are materials that are not primary products (produced for the market) and for which the generator has no further use in terms of production, transformation or consumption and therefore wants to dispose of [cited in Basel Convention, 2006][1] (p. 140).

The centrality of "value" and "use" in defining waste draws attention to ways and means by which these come to be attributed to waste. Since this attribution can vary greatly with users, with time, and across societies, it becomes difficult to objectively define waste. As Davoudi[2] (p. 131) says

> … it is difficult to determine the boundaries between what *is* and what is *not* waste. In other words it is difficult to determine exactly when a material ceases to be a resource (with social, economic and environmental values) and *becomes* a waste [original italics].

Understanding waste and its value or lack of it as an attribute bestowed upon it by society allows for an examination of societal values and practices that lead to the formation of waste. Waste can thus be understood in two distinct but complementary ways: first, in terms of its intrinsic qualities, which then suggest particular use and nuisance values, and second, its socially constructed qualities that draw attention to the political economy and the socio-cultural practices that reproduce the worth or worthlessness of waste.

This entry is concerned with "municipal solid waste" (MSW). It includes waste produced by households, small businesses, and industry, which are of a similar nature and therefore collected by the municipal authorities. Wastes such as construction and demolition wastes, heavy industrial wastes, etc., are excluded from this definition.[3] However, it is not uncommon to find contamination of municipal waste by small quantities of other types of waste, and often these can be made up of harmful substances such as pathological hospital waste, abattoir waste, harmful industrial waste, or even feces.[4]

The aim of this entry is to provide a broad-ranging overview of the nature of MSW and managerial practices associated with it. Particularly stressed are the historical growth of MSW as an environmental problem and the geographical variations—the global commonalities and differences—in the nature of MSW. The historical trajectories and geographical constancies in the nature of solid waste in turn have given rise to particular normative principles for the management of MSW. These are reviewed. These principles emanate from an understanding of the intrinsic nature of solid waste, but are, however, mediated by place- and time-specific ways of organizing governance. The patterning of these ways of organizing governance is finally reviewed.

Waste: A Historical Overview

Waste has always been an issue since human beings started living in societies. The severity of waste management has, however, varied, culminating in it becoming a serious environmental issue today. Compared with present times, in pre-industrial times, material possessions and artifacts were limited and consequently valued more highly. Everything that could be repaired and reused would be repaired and reused. Also, with limited population numbers and densities, more space for disposal and limited waste generation, waste was, relatively speaking, manageable and less polluting. In terms of waste composition, in most societies during these times, domestic waste was made up of ash from fires (for cooking and heating), wood, bones, and vegetable waste.[5] These were allowed to either compost or decay aerobically (i.e., decomposition that takes place in the presence of abundant oxygen, which then do not result in foul odors). This is a viable form of managing organic waste if the quantity of waste produced is less and if the waste is spread out to decay.

The above situation changed with industrialization and consequent urbanization, first, in the Western, now developed world. Rising population numbers in cities led to high densities and less space for aerobic waste disposal. Besides, waste generation itself increased with the availability of cheaper goods through mass production and the rise of a consumer culture. The nature of waste generated also saw a diversification with new waste materials such as plastics entering into the waste stream in significant quantities and materials such as ash declining due to the emergence of water-, gas-, oil-, and electricity-based central heating systems and cooking appliances. Roberts[6] reports that during this period, waste was being dumped in streets, alleys, open land, or common water sources, where pigs and other animals could devour them. Thus, waste management during the early periods of industrialization was almost non-existent and very similar in both England and the United States.

The history of institutionalization of waste management practices in the United Kingdom begins with the Public Health Act of 1848. This is when management of waste was directly linked to public health outcomes. The subsequent Public Health Act of 1875 required all local authorities to arrange for the removal and disposal of domestic solid waste.[5] In the United States, the first public health code was enacted in 1866 in New York City. With the establishment of a correlation between sanitation and public health and the occurrences of a series of epidemics in different states, the federal government of the United States came to establish the National Board of Health in 1879. Waste handling and disposal systems then came to be gradually set up in the United States. However, dumping in landfill sites (mostly near water sources) was the most common disposal method adopted.[6] In Britain, additionally, incineration plants (destructors) were constructed in around 250 locations in an attempt to recover energy.[5]

The environmental impacts of unsanitary disposal methods were soon felt. Thus, in 1929, the federal government in the United States first issued restrictions on the location of waste dumps by requiring these to be away from river banks. Waste disposal at sea was prohibited in 1934 and restrictions on disposing waste 20 miles within the shoreline came into existence in the 1930s.[6] During this time, the concept of a sanitary landfill (waste alternated with layers of sand or other inert material) was developed in Britain. The sanitary landfills were thought to be less polluting and less dangerous than the previous waste dumps. Further, explicit planning powers over waste management sites were granted to local authorities by the Town and Country Planning Act of 1947. The concept of sanitary landfills subsequently spread to the United States, and the 1930s and 1940s saw the establishment of a number of sanitary landfills in the country. However, in overall terms, these sanitary landfills were still just around 37% of the total landfills.[5,6]

The post-war years saw a rise in consumer culture. Easy access to mass-produced material goods soon saw an increase in the quantity of waste generation. Also, new nonbiodegradable products started entering the waste stream. Although the practice of sanitary landfills grew both in the United States and the United Kingdom, the unseen environmental effects of landfill sites started to surface. This, together with increasing environmental awareness, led to more proactive intervention in waste management. For instance, the Resource Conservation and Recovery Act, which aimed at conserving resources and reducing waste, thereby protecting the environment, was enacted in the United States in 1976. Similarly, in the United Kingdom, the Control of Pollutions Act was enacted in 1974, which then later evolved into the Environmental Protection Act 1990 wherein waste minimization and recycling are priorities in waste management.[5,6]

The above broad overview of the history of waste management in the West, particularly the United States and United Kingdom, show that differing socio-economic conditions through time have led to differences in waste characteristics and notions of how it should be handled and treated. Overall economic growth and development have, in the past, resulted in a growth in the quantity of solid domestic waste generated, and the very composition of waste. Historic differences in solid waste management (SWM) practices can today be seen in geographical differences in SWM practices. Thus, in developing countries, the quantity of waste generated tends to be less and the composition of waste tends to be distinctly different from the composition of waste in developed countries. However, waste characteristics can also differ with climate, seasons, the location, and the developmental status of specific locations. The next section further explores the nature and quantity of domestic solid waste and geographical differences in the same.

Waste Characteristics and Waste Quantities

There are many ways of classifying MSW and different degrees to which component streams can be differentiated. The forms of classification used in a particular instance will normally depend on the purpose of classification. In this entry, to understand the environmental pollution associated with domestic solid waste in broad terms, and the geographical variations in the type and severity of environmental pollution, a broad classification adopted by the United Nations Environmental Programme (UNEP)[4] is used. This classification is useful to understand some of the major differences in the nature of waste produced in developed and developing regions.

The UNEP[4] classifies MSW into two: organic wastes and inorganic wastes. Organic wastes can be further classified into putrescible, fermentable, and non-fermentable organic wastes. The putrescible components within domestic waste can decompose very rapidly with the production of unpleasant odors (e.g., food). The fermentable components also decompose rapidly, but do not give rise to the unpleasant odors associated with putrescible waste (e.g., crop). Non-fermentable wastes decompose very slowly (e.g., wood).

One of the major differences in waste characteristics between the developed and the developing world is in the higher proportionate amount of organic material present in the municipal waste generated in developing countries. This can be seen in Table 1 below.

The proportionate weight of putrescibles in waste in India is around 4.5 times more than that of Paris (75.2% vs. 16.3%). Significant difference can also be seen in the quantity of paper present in waste. Generally, the quantity of paper in the waste varies with the level of economic development of an area. Therefore, from Table 1, it can be seen that the quantity of waste in Paris and California is 40.9% and 40.8%, respectively, while the quantity of waste present in India is just 1.5%. This difference is mainly due to low levels of packaging combined with high levels of formal and informal reuse, recycling, and recovery practices widely prevalent in developing countries. Further, waste from warm and humid tropical

TABLE 1 Comparison of Solid Waste Characteristics Worldwide (% wet wt)

Location	Putrescibles	Paper	Metals	Glass	Plastics, Rubber, Leather	Textiles	Ceramics, Dust, Stones	Wt (g)/ cap/ day
Bangalore, India[1]	75.2	1.5	0.1	0.2	0.9	3.1	19.0	400
Manila, Philippines[2]	45.5	14.5	4.9	2.7	8.6	1.3	27.5	400
Asunción, Paraguay[2]	60.S	12.2	2.3	4.6	4.4	2.5	13.2	460
Seoul, Korea[3]	22.3	16.2	4.1	10.6	9.6	3.8	33.4”	2000[a]
Vienna, Austria[4]	23.3	33.6	3.7	10.4	7.0	3.1	18.9[b]	1180
Mexico City, Mexico[5]	59.8[c]	11.9	1.1	3.3	3.5	0.4	20.0	680
Paris, France[4]	16.3	40.9	3.2	9.4	8.4	4.4	17.4	1.430
Australia[7]	23.6	39.1	6.6	10.2	9.9		9.0	1.870
Sunnyvale, California, USA[6]	39.4[d]	40.8	3.5	4.4	9.6	1.(1)	1.3	2000
Bexar Comity, Texas, USA[6]	43.8”	34.0	4.3	5.5	7.5	2.0	2.9	1816

[a] Includes briquette ash (average).
[b] Includes “all others.”
[c] Includes small amounts of wood, hay, and straw.
[d] Includes garden waste.
Source: Data from UNEP (2005, p. 2).

and subtropical countries (where most of the developing world in located) also tend to be high in plant content, while waste from colder climates may have higher levels of ash (if coal or wood is used).

Besides differences in waste characteristics, a significant observation from Table 1 is the differences in overall quantities of waste production. Lower levels of consumption, lower levels of packaging in consumption practices, as well as higher levels of formal and informal reuse and recycling mentioned earlier, result in significantly low levels of waste production and accumulation. Thus, Table 1 shows that California produces 5 times more waste per capita per day than does Bangalore in India. However, it must be remembered that the management of solid waste, especially collection and disposal, continues to remain a significant municipal challenge in most developing countries and can thus result in less overall quantities of waste collected and disposed.[7]

The above account draws a sharp distinction between waste characteristics found in the developed world and the developing worlds. It is, however, useful to remember that the boundaries of this distinction may be blurred in practice as particular locations in developed countries may reveal characteristic very similar to those found in developing countries and vice versa. Therefore in many ways, the above distinction provides a caricature. It nevertheless performs a pragmatic function in that it alerts an observer to potential differences based on level of development that may be found in municipal waste characteristics and quantities. The actual use of this distinction as applied to particular locations must, in the end, be necessarily informed by concrete realities.

While the quantity and composition of waste may vary with time and location, two significant attributes remain intrinsic to municipal waste regardless of the context: first, the site of production of waste is diffused and variable, and, second, municipal waste has potential environmental nuisance values. Solid waste management practices instituted by municipalities worldwide reflect the challenges posed by these intrinsic characteristics and therefore certain managerial practices have come to be identified as constitutive of municipal waste management worldwide. The next section provides an overview of such managerial practices.

Municipal Solid Waste Management Practices

Municipal solid waste management practices can be generally said to comprise of storage and collection practices (including practices of collection for reuse, recycle, and recovery), transportation practices (sometimes including practices of transferring waste from one form of transport to the other), and disposal practices. All, or some of the above, may occur through formal practices designed and implemented by the municipal authority or through informal voluntary initiatives organized individually or in community groups.

Storage and collection practices include various types of storage receptacles designed for waste production units such as households/commercial establishments and community storage receptacles for storage of collected waste. Organization of proper storage of waste is essential as in its absence waste can be dumped in public spaces/drains. Storage units may also be custom made to suit particular waste management technologies such as the use of receptacles of particular design to suit waste transportation vehicles. Further, storage systems may include practices to encourage reuse/recycle/recovery by using specific color-coded waste receptacles that enable the streamlining of various streams.

Transportation practices consist of the movement of collected waste to disposal sites or transfer sites (where waste is collected and then transferred to vehicles suitable for transporting to more distant final disposal sites). Collection and transportation practices may also include managerial contracts forged with the private sector or voluntary sector for the collection and transportation of waste produced in dispersed locations.

Disposal practices include the many options through which waste is disposed, such as incineration plants, compost plants, landfills sites, and anaerobic biogasification plants, for the end disposal of wastes that are not reused and recycled. Disposal options can include material recovery options (such as composting) or energy recovery options (such as anaerobic biogasification plants or energy recovery from landfill sites). Disposal options may also aim to just get the waste out of sight by either dumping it in waterways, quarries, and other similar areas, or by transporting it afar (often to less developed areas). Further, disposal practices need not always be organized by municipal authorities, but can also be found at individual, household/establishment, or community level.

The above SWM practices are typically organized in accordance with generally acknowledged principles. Some of the principles that are widely adopted in SWM practices today are reviewed below.

Normative Principles in SWM

The "waste hierarchy" is a normative principle that has had a significant impact on practices of SWM in recent times (for instance, the principle has been very influential in the formulation of European waste management practices).[8] A diagrammatic representation of this is given in Figure 1.

The waste hierarchy advocates a waste management system oriented toward a hierarchy of managerial options that target waste reduction, reuse, recycling, and recovery, in that order of preference, thus reducing the amount of waste that needs to be finally disposed in a landfill. Strategic aims such as "zero waste" initiatives' aimed at minimizing waste toward zero by reducing consumption, arranging for reuse by repair or modification, and maximizing recycling, thereby bringing materials back into nature or the marketplace, are based on the normative principle promoted by the "waste hierarchy".

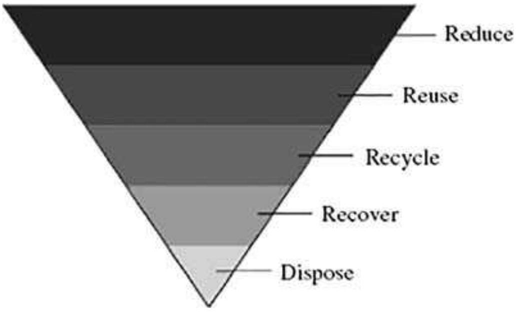


FIGURE 1 Waste hierarchy.
Source: Chettiparamb et al. (p. 327).[7]

Such initiatives are typically aimed toward instilling a re-valuation of waste as a resource rather than as something to be disposed off.[9]

Life cycle assessment (LCA) in waste has influenced waste reduction practices since around the 1990s.[10] It focuses on the environmental burdens associated with a product, a process, or an activity. LCA encompasses the entire life cycle of the system studied, including material and energy flows in raw materials, their acquisition, manufacturing processes, usage, and waste disposal/treatment, leading to what is known as a "cradle to grave approach." Thus, LCA helps expand the framework of analysis to beyond the relatively limited confines of the actual SWM system, thereby bringing in wider environmental impacts, some of which may have a greater impact on the environment than just the analysis of the SWM system itself. The LCA can thus lead to an improvement analysis that suggests ways in which the impacts on the environment can be reduced.[9] In practice, however, the LCA, although normatively influential, has been difficult to implement mainly because of the enormity of data demanded, problems in the reliability of the data, and other pragmatic difficulties in employing the technique.[11]

Integrated solid waste management (ISWM) is yet another principle that has been influential in organizing whole SWM systems. ISWM is based on the idea that any one method of SWM may not be adequate in addressing the challenges that the various streams of waste present. Instead of viewing the many streams of solid waste as distinct and separate, ISWM proposes that the most appropriate method must be adopted for each stream in a way that the sum total of the environmental impact is kept to a minimum. Thus, paper and other recyclable material must be recycled, while compostable material must be composted or digested anaerobically, and so on. The overall key aim is to minimize landfill while maximizing recycling and recovery.

The above account suggests that there is relative stability surrounding understandings of ideal ways of managing MSW. Studies of the governance of MSW, however, show wide variations, including shifting relative emphasis on these normative principles occurring both historically and within and across particular contexts. These variations are reviewed in the next section.

Policy Drivers, Policy Regimes, and Modes of Governance in SWM

The nature of the "political economy" of a region has been identified as a key determinant of MSW. The term "political economy" is defined by O'Brien[12] (p. 270) as "a regulated social framework for transacting values, comprising an arrangement of practices, relationships and institutions." He argues that industrialized societies are primarily "rubbish societies" in that these societies are infused with a relationship to waste and wasting that allows their social, political, and cultural systems and their own self-understanding to develop and change. If the centrality of wasting processes in industrialized societies is acknowledged, a political economy around wasting needs to develop, what O'Brien terms as "rubbish relationships'[12] (p. 287). Further, it then becomes necessary to valorize and promote particular institutional relations that organize and stabilize rubbish flows in ways that allow the reproduction and survival of society. This valorization and promotion reconfigures SWM practices, thus lending more credibility to particular normative principles resulting in patterns in what has been termed as "drivers,"[13] "policy regimes,"[14] or "modes of governance"[15] in SWM. These then give rise to new terminologies such as "waste citizenship" (how waste is accessed),[16] waste networks and waste flows (how waste is distributed),[17] and "waste commons" (hard rubbish meant for scavenging),[18] to name a few.

The study by Wilson[13] sought to identify key drivers that have led to the adoption of particular mechanisms or strategies in SWM both historically and around the world. A survey conducted by Wilson[13] among "colleagues" identified six key drivers for the organization of SWM practices: 1) public health; 2) environmental protection; 3) resource value of waste; 4) closing the loop; 5) institutional and responsibility issues; and 6) public awareness. The relative emphasis among these varies over time and differs with local circumstances. Wilson[13] argues that public health emerged as a key driver in the 19th century and has been more or less taken for granted in developed countries. It, however, still persists as a key driver in emerging and developing countries with hot climates. The environmental protection driver came to the fore in the 1970s and still remains a key driver in the developed world. Energy/climate change has now become central to this agenda. In developing countries, on the other hand, the focus remains on initial steps such as the phasing out of uncontrolled disposal and the promotion of clean development mechanisms. The resource value of waste was a driver until the early 20th century underpinning major industrial economies in 19th century London, and 20th century China, Soviet Union, and Eastern Europe. In developed countries, however, this driver was soon replaced by a more holistic approach to waste management. The resource value of waste remains a key driver in developing countries as large numbers of the poor survive on reuse, recycling, and recovery practices. Further, the import of waste as an industrial raw material continues in some developing countries, thus making waste integral to those economies. The waste hierarchy came to be advocated in the late 1970s and triggered a more holistic approach to resource management. Wilson[13] argues that this now means that waste prevention and recycling are key priorities in the developed world with sustainable production and consumption, integrated production policies, and zero waste policies emerging as key drivers. Notions of institutional responsibility emerged as a key driver when municipalities were charged with waste collection responsibilities in the late 19th century. While in the developed world, this responsibility is more or less taken for granted, with even an extended notion involving the private sector emerging, in developing countries, the ability to discharge this function is still limited with notions of capacity building and good governance emerging as key drivers. Finally, Wilson[13] argues that public awareness has become a key driver with waste management moving up in people's priorities as living standards increase. The environment, climate change, and resource management is informing this increase in public awareness in the developed world. In the developing world, however, public awareness remains focused on food, shelter, security, and livelihoods, with waste drawing attention only when these are threatened. The study concludes that

> "There is no one, single driver that can be seen as 'dominant'; rather, all of the six groups of drivers are important, and the balance between them will vary between countries depending on local circumstances, and indeed between stakeholder groups depending on their particular perspective"—Wilson (pp. 205–206).[13]

The "development drivers" described above therefore have a key role in organizing practices and selecting particular normative orientations and stabilizing these within institutional relations and a variety of governing instruments.

Gille[14] extends Young's[19] concept of "resource regimes" to waste. He defines it as follows:

> At their [resource regime's] core is a structure of rights and rules, which implies a certain distribution of advantages and disadvantages. Social institutions determine what wastes, and not just what resources are considered valuable by society, and these institutions regulate the production and distribution of waste in empirically tangible ways (p. 1056).

Gille explores the history of waste management in Hungary and identifies three distinct regimes in the history. He calls these "the metallic regime (1948–1974), the efficiency regime (1975–1984), and the chemical regime (1985–present)" (p. 1056). In the metallic regime, Gille argues that waste was seen as a particular kind of material—something that was like metal scrap and thus discrete, non-toxic, and infinitely recyclable and reusable. Under this benevolent perception of waste, the key actors were class-conscious workers and citizens who would collect and find new uses for all sorts of wastes. The efficiency regime was marked by a "monetized concept of waste: [where] waste was seen as a cost of production, and waste reduction and reuse were seen as steps to increase efficiency" (p. 1057). Policy tools thus included financial incentives and professionals with economic and technical expertise came to be valued. In the chemical regime, waste came to be seen as useless and even harmful. Waste liquidation came to be very important, and scientists, engineers, and the chemical industry became key players. The waste regime concept thus traces the broad patterns that structure certain normative values in particular periods and locations. **Table 2** Modes of governing MSW.

TABLE 2 Modes of Governing MSW

Mode	Components				
Disposal	Governmental rationality (policies and programmes)	Governing agencies	Institutional relations	Governmental technologies (examples)	Governed entities
	Economic efficiency	Local authorities	Devolved hierarchy	Dustbins	Municipal waste
	Public health Environmental efficiency	Regulator		Weekly collections Landfill sites Contracts Best Practicable Environmental Option assessment	Ratepayers
Diversion	Reducing the (global) environmental impact of landfill (EU Landfill Directive 1999 *Waste Strategy WOO*)	European Union	Multilevel	Performance targets and auditing	Successively lower government tiers
		DEFRA	Strongly hierarchical	New policy instruments	Individuals as passive citizens
		Local authorities		Funding mechanisms and criteria Education campaigns	Differentiated waste streams
Eco-efficiency	Reducing the environmental impacts of waste; recovering value (waste hierarchy; meeting targets)	Local authorities, waste contractors, community-waste-sector organizations	Heterarchy	Kerbside collections	Individuals as active citizens
			Networks	New technologies	Differentiated waste streams
				Reuse and reduction practices (e.g., nappies, compost)	
Waste as resource	Reducing the environmental impacts of waste; social and economic benefits	Nongovernmental organizations and networks	Solidarity	Provision of alternative infrastructures and collections	Individuals as community members
			Community		Waste as a resource

Source: Data from Bulkeley and Watson (p. 2740).

Bulkeley and Watson[15] develop the notion of "modes of governance" in waste in order to capture both structures and processes of governing while also recognizing the plurality of modes through which governance is established. They define a mode of governing as

> A set of governmental technologies deployed through particular institutional relations through which agents seek to act on the world/other people in order to attain distinctive objectives in line with particular kinds of governmental rationality (p. 2739).

The authors analyze SWM in the United Kingdom and identify four distinctive, but coexisting, modes of governance. These are the disposal mode, the diversion mode, the ecoefficiency mode, and the waste as resource mode. The various components of waste management associated with these different modes are reproduced in Table 2 above.

The three examples of development drivers, regimes, and modes of governance reviewed above illustrate how the management of solid waste can be very different across geographical regions even if the fundamental building blocks of managing waste—those that emanate from its intrinsic qualities—remain the same.

Conclusions

This entry has provided a broad overview of the nature of MSW and the management of it. It has highlighted that the definition of what waste is, will depend on particular contexts as the attribution of value and non-value takes place subjectively and contingently. Further, the history of waste management practices shows that the nature of municipal waste has varied with the level of industrialization and climatic conditions of a location. This variation can also be seen today in the difference in waste characteristics of developed and developing countries.

The dispersed nature of production sites of MSW and the potential environmental nuisance value of MSW can, however, be said to be intrinsic to waste. Fairly consistent management practices have arisen from these intrinsic attributes. Further, certain normative principles have also arisen to guide the organization of these managerial practices. The selection and mobilization of the normative principles, however, take place within broader governance frameworks that have been variously termed as development drivers, waste regimes, or modes of governance in waste. These have also been reviewed. In conclusion, this entry has shown that while the intrinsic nature of SWM has led to selective widely stabilized managerial practices, the organization of these practices and the governance of the same may be embedded in very different governance frameworks, thereby giving rise to a myriad of permutations in the actual practices of SWM.

References

1. Basel Convention, 2006. *Vital Waste Graphics 2.* Basel Convention: Geneva. Available at http://www.grida.no/files/publications/vital-waste2/VWG2_p3to46.pdf (accessed January 2012).
2. Davoudi, S. Governing waste: Introduction to special issue. J. Environ. Plan. Manage. **2009**, *52* (2), 131–136.
3. DEFRA. *Local Authority Collected Waste—Definition of Terms,* 2011. Available at http://www.defra.gov.uk/statistics/environment/waste/la-definition/ (accessed February 2012).
4. UNEP. *Solid Waste Management,* 2005; Vol. 1. Available at http://www.unep.org/publications/contents/pub_details_search.asp?ID=3799 (accessed February 2012).
5. Waste Online. *History of Waste and Recycling Information Sheet;* 2004. Available at http://dl.dropbox.com/u/21130258/resources/InformationSheets/HistoryofWaste.htm (accessed January 2012).
6. Roberts, J. *GARBAGE: The Black Sheep of the Family: A Brief History of Waste Regulation in the United States and Oklahoma;* Office of the Secretary of Environment: State of Oklahoma, 2011. Available at http://www.deq.state.ok.us/lpdnew/wastehistory/wastehistory.htm (accessed January 2012).
7. Chettiparamb, A.; Chakkalakkal, M.; Chedambath, R. In my backyard!: An alternate model of solid waste management. Int. Plan. Stud. **2011**, *16* (4), 313–331.
8. Schmidt, J.H.; Holm, P.; Merrild, A.; Christense, P. Life cycle assessment of the waste hierarchy—A Danish case study on waste paper. Waste Manage. **2007**, *27* (11), 1519–1630.
9. Glavic, P.; Lukman, R. Review of sustainability terms and their definitions. J. Cleaner Prod. **2007**, *15* (18), 1875–1885.
10. Sundqvist, J.O. System analysis of organic waste management schemes—Experiences of the ORWARE model. In *Resource Recovery and Reuse in Organic Waste Management;* Lens, P., Hamelers, B., Hoitink, H., Bidlingmaier, W., Eds.; IWA Publishing: London, 2004.
11. Ekvall, T.; Aseefa, G.; Bjorklund, A.; Eriksson, O.; Finnveden, G. What life-cycle assessment does and does not do in assessments of waste management. Waste Manage. **2007**, *27*, 989–996.
12. O'Brien, M. Rubbish values: Reflections on the political economy of waste. Sci. Cult. **1999**, *8* (3), 269–295.
13. Wilson, D.C. Development drivers for waste management. Waste Manage. Res. **2007**, *25* (3), 198–207.

14. Gille, Z. Actor networks, modes of production, and waste regimes: Reassembling the macro-social. Environ. Plan. A **2010**, *42,* 1049–1064.
15. Bulkeley, H.; Watson, M. Modes of governing municipal waste. Environ. Plan. A **2007**, *39,* 2733–2753.
16. *Gutberlet, J. Recovering Resources—Recycling Citizenship: Urban Poverty Reduction in Latin America; Ashgate: Aldershot, U.K., 2008.*
17. Davies, A.R. *The Geographies of Garbage Governance;* Ashgate Publishing: Aldershot, U.K., 2008.
18. Lane, R. The waste commons in an emerging resource recovery regime: Contesting property and value in Melbourne's hard rubbish collections. Geogr. Res. **2011**, *49* (4), 395–407.
19. Young, O. *Resource Regimes: Natural Resources and Social Institutions;* University of California Press: Berkeley, CA, 1982.

Bibliography

1. *Baud, I.; Post, J.; Furedy, C. Solid Waste Management and Recycling: Actors, Partnerships and Policies in Hyderabad, India and Nairobi, Kenya;Kluwer Academic Publishers: Dordrecht, the Netherlands, 2010.*
2. Bulkeley, H.; Watson, M.; Hudson, R.; Weaver, P. Governing municipal waste: Towards a new analytical framework. J.Environ.Policy Plan. **2005**, *7* (1), 1–23.
3. Davoudi, S. Planning for waste management: Changing discourses and institutional relationships. Progr. Plan. **2000**, *53*, 165–216.
4. Davoudi, S. Scalar tensions in the governance of waste: The resilience of state spatial Keynesianism. J. Environ. Plan. Manage. **2009**, *52* (2), 137–156.
5. Tchobanoglous, G.; Kreith, F. *Handbook of Solid Waste Management*, 2nd Ed.; McGraw Hill: New York, 2002.
6. UN-HABITAT. *Solid Waste Management in the World's Cities: Water and Sanitation in the World's Cities 2010.* UN-HABITAT, 2010. Available at http://www.unhabitat.org/pmss/listItemDetails.aspx?publicationID=2918 (accessed October 2012).
7. Zhu, D.; Asnani, P.U.; Zurbrugg, C.; Anapolsky, S.; Mani, S. *Improving Municipal Solid Waste Management in India: A Sourcebook for Policy Makers and Practitioners;* The World Bank: Washington, D.C., 2008.

17

Sustainability and Planning

Richard Cowell

Introduction

This chapter offers a critical assessment of the role of planning in promoting sustainability. "Planning" can be defined broadly as an activity by which humans seek consciously to shape their collective future,[1] but it embraces activities with enormous variation in scope and scale. Planning can be undertaken from the local through to the national scale, and it entails the detailed regulation of land use and management as well as the setting of strategic direction for an array of public and private sector actions.

The first half of this entry charts the emergence of a worldwide interest in "planning for sustainability" and outlines key features of the dominant, conventional approach. This typically involves the following components: identifying sustainability goals, allocating actions, seeking integration between governmental sectors, engagement of different stakeholder groups, and appraisal and monitoring.

The second half of the entry assesses the extent to which planning for sustainability has achieved results and offers explanations for these outcomes. It notes that there have been significant shortfalls in implementation, partly attributable to the deficiencies of planning as a force for change but also because of the fundamentally contested nature of sustainable development as an objective. In response, two alternative perspectives to the conventional approach to planning for sustainability are discussed: (1) adaptive planning and transition management, which emphasizes the complexity of the task and the need to respond to uncertainty, and (2) a more critical perspective, which highlights how planning systems have actually exerted leverage over business as usual.

Some preliminary points of definition are required. A distinction is often made between "sustainability," a concept with an environmental focus that describes the long-term continuation of a system, and "sustainable development," a concept that is more overtly concerned with social progress and distributive equity as well as environmental quality. However, there is much debate about the meaning of either term and little consistency in the way that they are used.[2] This handbook entry adopts the term

TABLE 1 The Two Main Categories of Planning for Sustainability

Type of Approach	Scale	Focus	Power
National sustainability planning	National, with focused attention on individual sectors	Achieving environmental targets consistent with sustainability	Overarching coordination, target-setting, backed by monitoring, oversight arrangements and sector-specific measures
Land use or spatial planning	Mainly local, but subject to national policy guidance	Land use implications arising from array of environmental, social, and economic demands	Regulation of land use change through zoning procedures and spatial steering; wider coordinating role

"sustainability" and argues that rather than focusing on *ex ante* definition, practitioners must be alert to the particular interpretations of the concept that get used in any given planning process, and to what is actually being sustained.

"Planning," too, requires some further explanation. Although planning can be undertaken by any kind of organization—public, private, or voluntary sector—here, we are primarily concerned with planning as a collective, public activity, undertaken by governments at all levels (local, regional, national). Planning can thus be seen as one of a number of tools available to government to promote sustainability, working alongside, or sometimes in conjunction with, regulations, market-based instruments, voluntary measures, and so on.

There are a number of categories of approach to "planning for sustainability" that are covered within this entry. The first category is the creation of "national sustainable development plans" designed specifically to promote the achievement of sustainable development (as encouraged by Agenda 21), which tends to provide high-level strategic direction across an array of activities.[3] An exemplar is the National Environmental Policy Plans (NEPPs) that were drawn up for the Netherlands from the 1980s to the early years of the 21st century. The second category of approaches is the incorporation of sustainability objectives into existing planning processes for coordinating the use of land ("land use planning" or "spatial planning"). Many countries have planning systems that seek to steer the use and development of land, and provide a framework for making decisions about development projects. Key features of the two categories are set out in Table 1 (above).

History and Context

Ever since "sustainable development" first arrived on global policy agendas, planning has been identified as a means of achieving it. A commitment to producing national sustainability plans was a key outcome of the United Nations Conference on Environment and Development, held in Rio in 1992, with its main statement of principles—Agenda 21—regarding "environmentally sound physical planning" as essential to sustainable development.[4] The European Union, too, has given planning a significant role in shifting from reactive responses to environmental problems toward anticipatory and strategic solutions.[5]

Over the same period, sustainable development took hold as a powerful ethical and intellectual framework for existing planning activities, notably land use planning or spatial planning, but also planning across a range of sectors—transport, economic development, tourism, and so on. By way of definition, land use planning usually concerns the management of competing uses of land, typically through regulatory means. This differs somewhat from "spatial planning," which—in the form adopted in the Netherlands or France, for example—has a wider role in coordinating the actions of all sectors that have spatial consequences (transport, economic development, etc.). For convenience, "land use planning" is the term used here, but the analysis also applies to forms of spatial planning. Many countries around the world have instituted requirements that their land use or spatial planning procedures promote sustainability.[6–11]

For many commentators, the connections between sustainability and planning are self-evident. The planning profession has long identified itself with shaping the future[12] and sustainability is a long-term goal, bound up with moral responsibilities to future generations. Thus, if sustainability can be regarded as a preferred future state of affairs—whether that is applied to an ecosystem, a nation, or a city—planning provides a vehicle for steering society toward it. Planning offers the prospect that the transition to sustainability can be orderly, rather than precipitated by ecological, social, or economic crises.

Planning also provides important coordinating and integrating functions. Planning offers mechanisms for identifying and fostering "win-win-win" solutions between economic development, environmental protection, and social equity, the prospect of which has always been central to the political appeal of sustainable development. Planning has also been identified as having procedural virtues for promoting sustainability, in that the processes of formulating plans can allow for the engagement of civil society, diverse stakeholders, and foster wider learning.[13]

However, if it is widely asserted that we should "plan for sustainability," the reality is more complex. At a fundamental level, some would contest the possibility of steering complex societies and unpredictable ecosystems toward pre-given goals; for Lindblom, the best societies can hope for is "muddling through."[14] Governments and societies vary in the faith that they place in (state) planning compared with markets as means of coordination. States also vary in the wider effectiveness of their governance institutions, with developing countries especially often lacking capacity. Consequently (as discussed further below), many efforts to plan for sustainability fail to fully achieve their goals, for all sorts of reasons—some economic, some political, and some practical.

Thus, to understand the prospects of planning for sustainability, one needs to consider both ideals and what happens in practice, as this entry seeks to do.

Planning for Sustainability—A Conventional Perspective

This account of planning for sustainability is organized around what can be regarded as the conventional perspective, as depicted in Figure 1 below. It echoes core features of long-standing rational planning models: exhibiting the procedural stages of "survey-analyze-plan-monitor," and with a clear relationship between means and ends. Although this traditional, instrumental view of planning has come under criticism from various directions,[15] it still provides a useful structuring device for this review. Moreover, this perspective remains deeply entrenched in planning practice in many arenas, and concern for sustainability has arguably reinforced its position.[16] Problems with this perspective are discussed later in this entry.

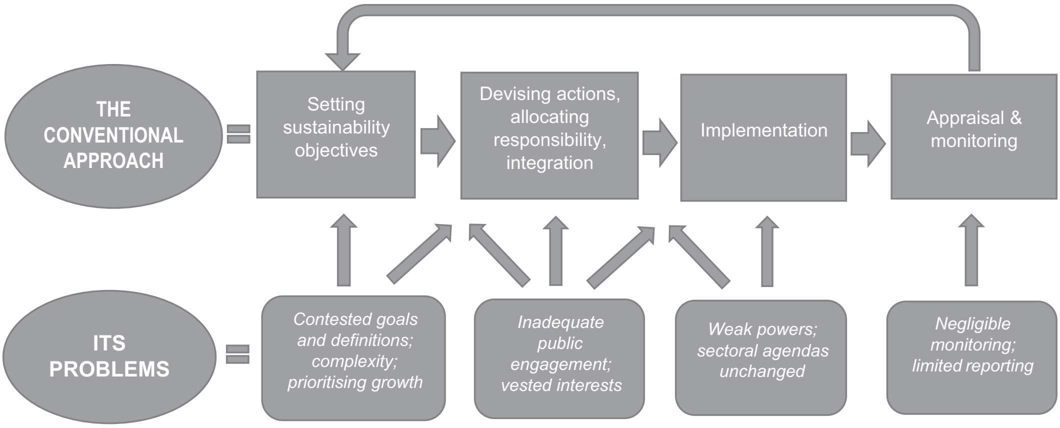


FIGURE 1 The conventional approach to planning for sustainability and its problems.

Setting Objectives and Interpreting Sustainability

At the core of the conventional view of planning for sustainability is the assumption that one can specify what makes a society more sustainable and express this in objectives toward which development can be steered. These objectives might be environmental constraints, trends or targets, or some vision of a more sustainable urban form. Objectives might be designed to avoid losses of biodiversity; to reduce the depletion of primary resources and/or replace them with renewable or less polluting substitutes; to limit the release of certain wastes or pollutants; or to express social goals for literacy, health, poverty, access to services, and so on.

In effect, this part of the process is an exercise in defining outcomes, which seek to capture "the overall quality or sustainability of human well-being and the eco-systems on which it ultimately depends."[17] Sometimes the process of determining objectives takes place within the planning process, but very often objectives are "imported" from other policy arenas. A pre-eminent example is climate change, where targets for greenhouse gas emission reductions cascade down from international agreements into national action plans and then into planning. Other examples include air quality standards, which may emanate from debates in public health. Where the objectives relate closely to land use, then they may be expressed in vision statements, diagrams, or maps.

Of course, there is no singular, neutral definition of sustainability, but a range of interpretations with quite different developmental consequences. Planning for sustainability is often therefore more than an exercise in applying pre-given goals; it forms part of the process of working out how sustainability should be interpreted. Planning processes may exhibit "weak" forms of sustainability,[18] so-called either because they explicitly prioritize economic growth or because they give no clear priority between economic, social, and environmental goals. The goal of planning for sustainability may thus be expressed as finding a "balance" between these goals, allowing for environmental quality to be traded off for economic development.

Alternatively, planning approaches may pursue "stronger" conceptions of sustainability.[18] Here, the guiding assumption is that society should be precautionary in the face of significant, irreversible losses of valued environmental assets—be they planetary life support systems or treasured landscapes—and give priority to maintaining their value over time. The implications for planning are that environmental constraints are pre-eminent, goals should be set to reflect this, and that development trajectories must be adjusted to observe these constraints and avoid degrading the ability of the environment to support human welfare. Stronger conceptions of sustainability draw support from justice to future generations, and the idea that they should inherit an equivalent stock of "environmental capital"[19]; or from biological notions of observing the "carrying capacity" of an environment to sustain a particular activity without deterioration.[20]

One example of a stronger approach to planning for sustainability is the NEPPs of the Netherlands. From the 1980s, the Dutch Environment Ministry began developing plans that sought to respond to the worsening and interconnected nature of environmental problems through what was regarded as a "holistic" planning process. A key starting point was analysis of the level of pollution reduction and resource efficiency that was required "to achieve sustainable development in one generation."[21] From this, the NEPPs identified targets and brought together steps for reducing pressures on the environment and "closing substance cycles" (e.g., reducing and recycling wastes).

Land use planning offers examples of weaker and stronger interpretations of sustainability. For example, since the early 1990s, U.K. governments have required all land use plans to promote sustainable development; however, national guidance has often interpreted this in weak terms—as achieving simultaneously economic growth, inclusive social progress, protecting the environment, and making prudent use of natural resources.[22] Nevertheless, one can point to land use plans at local and regional levels that have sought to determine the environmental capacity of their area to accommodate development, and introduce policies to ensure that development does not exceed it.[23,24] Observing environmental limits remains an important idea in planning for sustainability.[25]

Devising Actions and Allocating Responsibility

The creation of a plan can itself exercise persuasive effects. There is a long history of planning processes pursuing some form of "visionary idealism" to stimulate public, stakeholder and political support and promote change, many of them connected to aspects of sustainability.[26] However, the second stage of the conventional approach to planning for sustainability usually entails specifying the particular actions required to achieve the objectives of the plan.

Various approaches can be used to identify those activities most responsible for placing unsustainable pressures on the environment—be they industry, government or the public, or particular sectors of activity (transport, housing, agriculture, etc.)—and thus requiring action. One such approach is "backcasting," a form of futures study. The backcasting process begins by envisioning a desirable end state—in this context, one deemed to be sustainable—then works backward from that to identify actions that will move society toward it.[27]

Depending on the political culture of the country concerned, the process of preparing a sustainability plan may involve efforts to secure the participation of different sectors of society. This participation can take place at the objectives-forming stage as well as in the selection of actions. The rationale for participation is at least fourfold:

- To try to achieve some "buy-in" from those stakeholders with the capacity to affect outcomes, and reduce the chance of conflict at the implementation stage
- To draw in knowledge and ideas beyond that possessed by government alone
- To create potentially more just outcomes by conferring recognition and opportunities for participation on a range of social groups[28]
- To create greater democratic legitimacy and societal support for the transitions that sustainability requires.[13]

Some theories of planning see the participatory processes of planning as key to its role in achieving change, notably collaborative planning theory.[15] The interactions and dialogue that takes place in planning processes—where they require parties to make and defend arguments in relatively open arenas—can help shift the attitudes and beliefs of those involved. Planning exercises thus help foster wider learning, develop new knowledge and relations between groups of people, and assist in finding common ground between contending parties. This, in turn, builds up the capacity of society for resolving future problems.

Collaboration between stakeholders was a key element of producing the NEPPs in the Netherlands. Government and business constructed sectoral plans for reducing environmental impacts and technological change bound together by voluntary agreements called "covenants" (albeit often with the threat of tighter regulation should voluntary compliance fail). The Dutch NEPPs also incorporated environmental non-governmental organizations (NGOs) into the planning process. In many countries, the existing land use or spatial planning system has long-provided opportunities for the public and environmental groups to contribute to the decision-making process,[29] and planning for sustainability builds on these traditions.

If planning is to have some strategic role in promoting sustainability across society, then the process of allocating responsibility and devising actions may entail "environmental policy integration."[30] This is a process by which key policy sectors coordinate their actions toward shared sustainability objectives, cooperate on implementation, and reduce conflicting objectives. Examples would be agricultural policy and transport policy being attuned to the delivery of biodiversity or air quality objectives, rather than just their own developmental goals. In procedural terms, integration may be achieved by various innovative governance arrangements, including altering the objectives of sectors, or through various forms of deliberation (i.e., getting key departments round the table) and policy appraisal, as discussed below.

Closely connected with the assignment of responsibility is the identification of actions and policy instruments—the means by which the plan will actually be implemented. Where sustainability goals are being integrated into existing land use planning processes, then the powers of those processes will dictate how the goals can be achieved. In many countries, this entails the incorporation of policies into a plan, with some spatial expression of future goals, which then informs zoning controls or some other regulatory process to achieve compliance. Where sustainability is concerned, these regulations can embrace[31]:

- Promoting sustainable urban form (by controlling building density, or directing development to locations accessible by walking, cycling or public transport)
- Protecting sensitive, valued spaces from damaging development (e.g. for biodiversity)
- Requiring that new developments are designed and built in ways that reduce their adverse impacts.

With national sustainability plans, the plan may provide a long-term framework for shorter- and medium-term actions to be undertaken by a range of actors—government departments, industrial sectors, and the community and voluntary sector. Goals may therefore ultimately be implemented by an array of measures: standards and regulations, market-based instruments like green taxes, channeling investment or research funding, and various forms of voluntary mechanisms (like information campaigns, promoting best practice, etc.).

Appraisal and Monitoring

Forms of appraisal and monitoring are integral to conventional approaches to planning for sustainability and may be used either ex post or ex ante to guide the planning process. Ex post entails monitoring the effects of the plan after implementation (either progress toward goals or of wider unintended effects), and using the information to make adjustments or revisions in cyclical, learning processes. Ex ante appraisal entails assessing the likely effects of the plan while it is being formulated, perhaps to ensure that it is contributing sufficiently to sustainability.

The use of appraisal techniques has become a significant part of planning for sustainability and links to wider suites of tools used in environmental management, many drawing on international principles and methodologies.[10] Strategic Environmental Assessment (SEA) is required in many countries for plans, programs, and policies likely to affect the environment (e.g., under European Union Directive 2001/42/EC). Through SEA, the draft policies of plans are assessed for their environmental impacts and, in many formulations, this entails assessment against sustainability criteria and objectives.[32] SEA can also enable the comparison of different options and, potentially, facilitate debate over alternatives, as well as fostering environmental policy integration by identifying where sectoral policies are poorly aligned with sustainability goals.

Environmental indicators may be used to assess changes in the economy, society, and environment, and signal whether those changes lead us toward or away from sustainability. Some indicators reflect a particular environmental parameter. For example, in the United Kingdom, the percentage of new development located on brownfield land (i.e., land that has been built on before) is an important sustainability indicator for the land use planning system, with the goal being that increasing development on brownfield sites will reduce sprawl, slow down the consumption of "greenfield" land, and support compact, less car-dependent urban forms.

The search for better appraisal and measurement systems has been a constant feature of planning for sustainability debates: where "better" can be taken as more comprehensive, more integrated and quantifiable (i.e., combining different effects into single units), more comprehensible to diverse audiences or decision-makers, or better linked to core principles of sustainability like environmental limits. One example is the "ecological footprint," which seeks to provide a proxy measure of human demands on the environment by assessing how much biologically productive land and sea is appropriated to maintain a given consumption pattern and assimilate the waste produced.[33] There are other approaches also seeking to capture the totality of material and ecological flows associated with "urban metabolisms,"[34] including "exergy," a concept that derives from thermodynamics and ecosystem health analysis.[35]

Rising in profile from 2010 onwards has been interest in assessing the effects on ecosystem services; an approach that reflects a long-standing belief in environmental appraisal circles that for decision-makers to take environmental effects seriously they need to be expressed in economic terms.[36]

Has Planning for Sustainability Had Any Effect?

Across the world, numerous plans have been produced that would claim to be promoting sustainability. In many countries, sustainable development has become “mainstreamed” as a core goal of land use or spatial planning—with almost all such plans, and much national planning policy too, now making reference to the importance of sustainability.[7,8,10,11] What is less clear is how far these planning *outputs* have promoted more sustainable *outcomes*. This section reviews the effects of planning for sustainability and then explains why outcomes rarely match aspirations.

One can point to beneficial outcomes in a number of spheres. Sustainability-oriented land use plans have contributed to an incremental greening of urban development. By providing locational guidance, plans have helped steer development to sites where its impacts have been lower than they might otherwise have been, reducing inappropriate development in areas of high existing environmental value, and focused development on advantageous locations—such as sites that encourage access by public transport or foot, or face reduced risks from flooding. The planning system also confers greater protection to environmental assets, such as valued wildlife sites, and requires more exacting mitigation and compensation measures when development has damaging effects. Over the same period, plans have mobilized policies to enhance the environmental performance of built development, for example, by promoting better design, or as one of a suite of tools for ensuring development is adapted to the risks of climate change. Plans have developed innovative regulations,[31,37] such as the “One Planet Development” policy in Wales, allowing development in the countryside where it demonstrably attains a low Ecological Footprint.[38]

Nevertheless, the overall effects of planning on the sustainability of development have been modest, and uneven between issues and places. Unsurprisingly perhaps, land use planning has made the greatest progress where sustainability agendas link to traditional planning concerns, such as land allocation and siting policies, or issues like transportation where technical fixes are available. Where particular environmental issues are of recognized importance, then planning has addressed them, such as water scarcity in parts of the United States and Australia.[10,39] The contribution of planning is often less clear in relation to global, systemic issues, like reducing greenhouse gas emissions, poverty or resource depletion, where wider societal transformation is required.

With national sustainability planning, too, most strategies have gravitated “toward the cosmetic rather than the ideal.”[40] The NEPPs from the Netherlands were most successful in reducing wastes, pollution, and the use of toxic materials; however, progress has been greatest in areas where industry saw the potential for economic gains from improving environmental performance, and where technological solutions were readily available.[41] Moreover, the NEPP strategies have struggled to address issues caused by mass consumption and growing personal mobility, such as carbon emissions from transport. If there has been any decoupling of economic growth from environmental degradation, it is mostly relative decoupling (falling environmental impact per unit of growth) rather than absolute decoupling (environmental impacts falling in absolute terms). Many countries have still not produced national sustainability plans, and those that have done so too rarely institute cycles of reporting and updating.[3]

Thus, despite 30 years of “planning for sustainability” around the globe, progress in fostering more sustainable patterns of development has been patchy at best, and often disappointing. As noted above, nations differ greatly in their capacity for organized and effective planning.[42] That said, success is partial in many developed countries, including those with strong traditions of strategic plan-making and reputations for environmental concern, for example, the Netherlands[43] and Sweden.[44] Despite some modest gains, the overwhelming perception is one of a significant “implementation deficit,” in which the outcomes achieved through planning fail to match the rhetoric of sustainability or meet the more challenging targets. How might this “deficit” be explained?

Some see it as a simple time lag between aspiration and realization. Plans take time to prepare and implement; consequently, they can omit newer, emerging issues, like sustainable urban food.[45]

Ambitious planning for sustainability also confronts the limited power of planning mechanisms to effect change. The regulatory powers of land use planning do not apply to every sphere of social or technical change. It can be used to regulate new built development—indeed, new "eco-cities" or "sustainable communities" with higher sustainability credentials than the norm are frequently held up as exemplars of what can be achieved.[46] However, planning has limited scope to reshape existing buildings or economic systems, especially those that unfold at wider spatial scales. Thus, land use planning might be successful in steering the location of supermarkets, even influence the environmental performance of their built form (perhaps by requiring certain energy efficiency standards), but exercises little control over the systems of food production and consumption channeled through them. Research suggests that planning professionals have sometimes molded planning for sustainability toward orthodox planning tools and approaches rather than looking at problems afresh.[11,45]

However, there are fundamental problems in conceiving of this underachievement as an "implementation gap," with the expectation that this will be closed over time with "more knowledge," "better practices," or "improved participation."

One problem with the conventional, linear view of planning for sustainability is the assumption that sustainability has an agreed meaning that can be translated into clear consensual goals, which can inform singular blueprints for progress. In reality, although there may be broad agreement on the principles of sustainability, translating this into more specific interpretations in the planning process can cause dispute, as fundamental questions are raised about human wants and needs, the relationship between state and market, and environmental values. Planning for sustainable development cannot therefore be a value-free, technical exercise.[2] Even within the environmental sphere, there is often significant scientific uncertainty about the causes of change, and the position (or existence) of ecological thresholds or "capacities," which affects the definition of targets or indicators.

Planning for sustainable development has also exposed the difficulties of achieving outcomes that are win-win for social, economic, and environmental goals. Synergies between goals can be found: for example, planning for more walking-friendly cities has benefits for human health, environmental protection, and local economies where congestion is reduced. In many other circumstances, goals cannot readily be reconciled without winners and losers. Protecting valued landscapes or habitats can displace local livelihoods, including those which are highly sustainable.[9,45] Displacement can also be caused by the expansion of renewable energy like solar power.[47] Newly enhanced or created sustainable communities can make property too expensive for poorer existing residents.[39] Environmental constraints may challenge particular directions of economic growth. Planning for sustainability may thus struggle to align competing conceptions of what it means for development to be sustainable.[7]

One also needs to understand the political and economic context in which planning for sustainability takes place, particularly macro-economic shifts that have made communities and nations dependent on the private sector to deliver development, and public sector austerity programs that reduce the resources for long-term environmental activity.[48] The vulnerabilities this creates are illustrated by particular projects, for example, Masdar Eco-City in Abu Dhabi. This new settlement was planned with very high sustainability ambitions, but the 2008 financial crash undermined the economics of the project, leading to the most innovative environmental dimensions being abandoned.[49] More widely, economic concerns may explain why planning authorities may be reluctant to exercise stringently the regulatory powers they do possess, where that could threaten jobs and growth. Plans that have sought to promote "strong" interpretations of sustainability, by seeking to keep development within environmental limits, have faced resistance when confronting what "seem to be inexorable upward trends in production, consumption, and mobility."[7]

This also explains why many exercises in planning for sustainability have adopted "weaker" interpretations of the concept, seeking only to achieve some "balance" between economic, environmental, and social goals. Sometimes the interpretation of sustainability has been weaker still, with planning goals framed in ways that accommodate rather than question unsustainable patterns of production and demand, such as for road space, mineral resources, or speculative property assets, causing contradictions. For example, although Dutch spatial plans sought to encourage greener forms of travel, they still promoted the Netherlands' international role as a transport hub, incorporating goals for expanding seaports and airports, with deleterious environmental consequences.[50] Here, one can see that the problem is not a straightforward "implementation deficit," but that sustainable development has been defined in ways that accommodate the status quo.

These difficulties have raised further questions about the procedural dimensions of planning for sustainability.

- *Participation and expertise*: There is evidence to support the argument that wider societal participation helps support plans that emphasize sustainability.[8] However, securing "adequate" public engagement in planning processes encounters a number of problems. The range of complex ecological issues that sustainability has brought into planning processes raises questions about how different forms of knowledge are handled and tested. A common risk is that technical elites and industry representatives dominate the process, to the exclusion of other groups. These dilemmas, in turn, highlight the risks of collaborative planning processes that give privileged access to key economic players, which act to protect their immediate interests; Van der Straaten[51] suggests that the NEPPs underperformed because powerful actors were able to weaken targets.
- *Futurity and uncertainty*: For all that planning and sustainability both seek to respond to long time horizons, the realities on which present actions will be based are constantly shifting in the light of new knowledge and social changes. This creates tensions between instituting long-term plans and remaining responsive to change[52] without caving in to expedient short-termism.[53]
- *Environmental policy integration*: The idea that planning for sustainability requires "integration" has also been easier to say than to achieve in practice.[30] The kind of voluntarist, collaborative approaches to integration that planning processes tend to foster has proved insufficient in shifting dominant sectoral objectives toward sustainability goals.[54] National sustainability strategies are still often led by environment ministries, which then struggle to exercise much influence over other sections of government. Ironically, land use planning is often left managing the side effects of sectoral policies that national sustainability strategies have failed to integrate with sustainability goals.

The overall evidence of the years since Agenda 21 is that planning is a desirable but insufficient ingredient in progressing sustainable development. For some, the problem is the concept of sustainability itself, which has been accused of being vacuous, vague, too contradictory and too readily captured by powerful interests. Some have shifted their attention to goals that seem to offer firmer anchors for progress, like resilience or decarbonization,[26] but such concepts may be no less prone to interpretive and political challenges when used in actual planning.

Perhaps the problems lie partly in the planning approach? The implementation problems discussed above expose how conventional approaches to planning for sustainability, depicted in Figure 1, tend to adopt an "ineffective engineering ideal,"[21] seeing society as a technical system that can be steered from the center to deliver targets. This strategy can work for narrow issues like pollution control but performs less well for complex social problems. However, if the above analysis shows planning for sustainability to be a more complex and challenging task than it might at first appear, it also points the way to two alternative perspectives.

Rethinking Planning for Sustainability

Adaptive Planning and Transition Management

If initial interest in using planning to promote sustainability was often guilty of naive optimism in what could be achieved, one by-product of these implementation difficulties has been a better understanding of the problem.

> Sustainable development is pursued in a world of multidimensional, intersecting and dynamic complex systems. We cannot expect to describe them fully, much less predict future effects. We may lack even suggestive evidence about many emerging problems, whose influences will ripple unpredictably through complex socio-ecological systems.[55]

One response has been to advocate approaches to promoting sustainability that are better suited to this reality, by showing "explicit appreciation of complexity and uncertainty, likelihood of surprise, and need for flexibility and adaptive capacity."[55] Some would describe these approaches as "adaptive planning"[56]; others have developed concepts of "transition management."

In practice, adaptive planning and transition management embrace elements of the conventional approach to planning for sustainability, but with different emphases. They incorporate an interest in developing shared long-term objectives, sustainability visions, and being anticipatory, as well as the use of indicators and monitoring to assess progress. However, there is less emphasis on "the old plan-and-implement model aimed at achieving particular outcomes"[55] in favor of "a more open-ended, process oriented philosophy." Rather than pursuing a single vision of sustainability, there is a concern to explore a diversity of different routes in a more iterative way, using a portfolio of ideas including "transition experiments," to develop circles of learning and adaptation. Adaptive planning takes an iterative approach with a greater emphasis on "monitoring, researching, and adjusting," allowing the redirection of previous management goals and activities "in light of new information and surprise,"[57] which could be an environmental or social change. There is less accent on achieving specific goals and more on institutional designs that prove resilient in the event of failure.

Some transition management approaches also take a much wider view of what needs to be changed if sustainability is to become a reality; from more targeted, incremental improvements in existing development forms to "promoting the sustainable reconfiguration of entire socio-technical systems."[58] This reflects the fact that "the power to shape structural change in society and technology is distributed across a multitude of actors and societal subsystems."[59] Transition management also emphasizes fostering innovation.

There is limited evidence for the actual effects of transition management approaches,[60] but the ideas have been applied in the Netherlands. Dissatisfied by the poor performance of the NEPPs, and keen to "revitalize" the process,[58] policy makers incorporated transition management in the fourth NEPP, released in 2001, and committed the government to sustainability transitions in energy, transport, water management, natural resource use, and agriculture.[55] The process involved creating a series of "platforms" in which stakeholders came together, to develop "transition themes" and pathways to those goals, including "transition experiments" to explore in practical terms how to progress along those pathways.[58]

Whatever its theoretical merits, the Netherlands' initial experiences raise questions about how far, in practice, elements of transition management have actually exerted pressure on business-as-usual. Smith and Kern[58] show how, when applied to the energy sector, the desire to incorporate relevant, powerful actors from industry and government departments into the process left it open to being captured by them; a wider risk of stakeholder engagement, as noted above. The result was that the transitions management approach failed to displace a narrow focus on technological costs and benefits or to exert much influence on wider energy policy. Absent also was any wider public debate about the transition goals or pathways. Other commentators have expressed a concern that transition management does not avoid the problems of conventional, linear approaches to planning for sustainability in assuming that there could be a consensus among diverse stakeholders about what constitutes a sustainable development path.[61]

Arguably, transitions management approaches also display another problem of much planning for sustainability: a tendency to focus excessively on new, more innovative forms of development, and give insufficient attention to how societies challenge and displace existing, unsustainable development forms.[62] However, planning can have a role in this, as explained below.

Planning and the Politics of Sustainability Transitions

The second alternative perspective considered here requires us to look beyond the intended goals of planning, to understand how planning processes have actually served to promote more sustainable forms of development. This requires us to view planning less as an activity on its own and more in relation to wider systems of public policy, and the politics of decision making. Viewed in this way, the impact of planning on sustainability may be messier, and less obvious in the short term, but more profound over the long term.[7,63]

One important feature of planning processes is that they provide "opportunity structures"[64]—including consultation procedures over plans and projects, legal review mechanisms, and various forms of environmental assessment—that allow the involvement of the public and interest groups, and through which argument about social purpose can take place. A distinctive feature of the opportunity structures provided by planning is the scope they offer for cross-scalar discussions, that is, through engagement about future development projects or land uses participants may also come to question the wider policies and values that lie behind them. Participants may raise fundamental concerns about the environmental sustainability of certain forms of development, and press for lower-impact alternatives.[45,65] The arguments generated can also exert pressure on environmentally damaging forms of development through extra costs, delays, and the ever-present threat of veto. Over time, the cumulative effect of these discursive and material impacts has sometimes undermined previously dominant policy approaches, loosening the grip of powerful interests, fostering acceptance of tightening environmental conditions, and opening up long-held policy norms.[63]

One can see these dynamics in a number of contexts. Challenges made through (and beyond) planning processes to new road schemes and their environmental impacts have helped reinforce arguments for a change in transport policy, away from predicting travel demand and building road space to accommodate it, toward greater emphasis on public transport and managing travel demand. Airport expansion planning has also become an important venue for debates about the environmental risks and societal worth of unfettered aviation growth.[7,66] Hudalah et al.[67] chart the way in which environmental NGOs in Indonesia exploited moments of opportunity in project and plan decision making to form new coalitions, and increase the weight given to wider development agendas based on conserving ecological functions. Planning is important because it is where broad policies and otherwise abstract sustainability concepts meet real, concrete environmental changes, stimulating public engagement and demanding solutions.

For all that it might be messy, conflictual, and unintended, one can see how existing planning systems have generated the kind of debate and policy learning that transitions to sustainability are often deemed to demand.[55] It is often conflict as much as cooperation that forces new perspectives onto the agenda. Moreover, this function of planning has reinforced sustainability transitions by subjecting unsustainable development to intensive questioning and, in some cases, presaging its demise—a dimension of change that the technology/innovation-centered approaches of transitions management tend to neglect.[62]

The challenge is to maintain openness of planning processes to wide-ranging debate, often in face of efforts by governments to keep fundamental questions of need and risk off the agenda.[7]

Conclusion

A central problem with the proposition that one can "plan for sustainability" is that sustainability is not a concept that can be defined precisely and then implemented.[68] Rather, planning is one of a number of arenas in which the meaning of sustainability is debated and constructed, often in the face of significant challenges. Indeed, the experiences of planning have been important for showing that formulating policies for sustainability can open up quite fundamental questions about needs, demands, and environmental value. Such questions are implicit in asking—"what is it that we should sustain?" Because of the disagreements surrounding such fundamental questions, progressing toward sustainable development is inevitably a process of "muddling through,"[14] even when plans look clear on paper, and likely to be contested—especially perhaps if plans are seeking to make a real difference to unsustainable trends.

However, although the conventional, linear view of planning for sustainability displays a number of weaknesses, planning is a vital component of social progress toward sustainable development. Through the intended goals and actions of plans, more sustainable forms of development can be delivered; at the same time, the spaces for deliberation and challenge that planning creates form a key part of the "ongoing territorial struggle"[69] to understand and institutionalize more sustainable forms of growth.

This experience suggests that when examining processes of planning for sustainability, one needs to do more than look at whether the core features of the conventional approach are in place: clear goals and targets, well-chosen actions, mechanisms for public and interest group engagement, and systems for monitoring and review. One should also ask the following questions: what are the opportunity structures for influencing the plan and who is able to access and exploit them? What scope is there for raising alternative development strategies and challenging dominant but unsustainable development priorities?[70]

References

1. Rose, E. Philosophy and purpose in planning. In *The Spirit and Purpose of Planning*, 2nd Ed.; Bruton, M.J., Ed.; Hutchinson: London, 1984; 31–65.
2. Springett, D.; Redclift, M. Sustainable development: History and evolution of the concept. In *Routledge International Handbook of Sustainable Development*; Redclift, M., Springett, D., Eds.; Routledge: London, 2015; 3–38.
3. Steurer, R.; Martinuzzi, A. (2007) 'Editorial: From environmental plans to sustainable development strategies', *Eur. Environ.* **2007**, *17*, 147–151.
4. United National Conference on Environment and Development (UNCED). *Agenda 21: The United Nations Programme of Action from Rio*; United Nations Department of Public Information: New York, 1992; para 7.28.
5. CEC. *Towards Balanced and Sustainable Development of the European Union*; Office for Official Publications of the European Commission: Luxembourg, 1999.
6. Bishop, K. Planning to save the planet? Planning's green paradigm. In *British Planning Policy in Transition*; Tewdwr-Jones, M., Ed.; UCL Press: London, 1996; 205–219.
7. Owens, S.; Cowell, R. *Land and Limits: Interpreting Sustainability in the Planning Process*; Routledge: London, 2002.
8. Conroy, M.; Berke, P. What makes a good sustainable development plan? An analysis of factors that influence principles of sustainable development. *Environ. Plan. A* **2004**, *36*, 1381–1396.
9. Lestrelin, G.; Castella, J.; Bourgoin, J. (2012) Territorialising sustainable development: The politics of land-use planning in Laos. *J. Contemp. Asia 42* (4), 581–602.
10. Davidson, K.; Arman, M. (2014) Planning for sustainability: An assessment of recent metropolitan strategies and urban policy in Australia. *Aust. Plan.* **2014**, *51* (4), 296–306.
11. Persson, C. Deliberation and doctrine? Land use and spatial planning for sustainable development in Sweden. *Land Use Policy* **2013**, *34*, 301–313.
12. Davies, J.G. *The Evangelistic Bureaucrat: A Study of a Planning Exercise in Newcastle upon Tyne*; Tavistock: London, 1972.
13. Blowers, A. *Planning for a Sustainable Environment*; Town and Country Planning Association: London, 1993.
14. Lindblom, C. The science of 'muddling through'. *Public Adm. Rev.* **1959**, *19*, 79–88.
15. Healey, P. *Collaborative Planning: Shaping Places in Fragmented Societies*; Macmillan: Basingstoke, 1997.
16. Rydin, Y. Sustainable development and the role of land use planning. *Area* **1995**, *27* (4), 369–377.

17. Adger, W.N.; Jordan, A. Sustainability: Exploring the processes and outcomes of governance. In *Governing Sustainability*; Adger, W.N., Jordan, A., Eds.; Cambridge University Press: Cambridge, 2009; 5.
18. Jacobs, M. *The Green Economy*; Pluto Press: London, 1991.
19. Pearce, D.W.; Turner, R.K. *Economics of Natural Resources and the Environment*; Harvester Wheatsheaf: Hemel Hempstead, 1988.
20. Meadows, D.H.; Randers, J.; Meadows, D.L. *Limits to Growth: The Thirty Year Update*; Earthscan: London, 2005.
21. Van der Straaten, J.; Ugelow, J. Environmental policy in the Netherlands: Change and effectiveness. In *Rhetoric and Reality in Environmental Policy. The Case of the Netherlands in Comparison with Britain*; Wintle, M.; Reeve, R., Eds.; Ashgate: Aldershot, 1994; 127.
22. Department of Transport, Environment and the Regions. *A Better Quality of Life: A Strategy for Sustainable Development*; Cmnd 4345; The Stationery Office: London, 1999.
23. Cowell, R.; Owens, S. Sustainability: The new challenge. In *Town Planning into the 21st Century*; Blowers, A., Evans, B., Eds.; Routledge: London, 1997; 15–32.
24. Counsell, D. Attitudes to sustainable development in the housing capacity debate: A case study of the West Sussex structure plan. *Town Plan. Rev.* **1999**, *70* (2), 213–229.
25. Haines-Young, R.; Potschin, M.; Cheshire, D. *Defining and Identifying Environmental Limits for Sustainable Development. A Scoping Study*; Final Overview Report to Defra, Project Code NR0102; Defra: London, 2006.
26. Gunder, M. Sustainability: Planning's saving grace or road to perdition? *J. Plan. Educ. Res.* **2006**, *26* (2), 208–221.
27. Miola, A., Ed. *Backcasting Approach for Sustainable Mobility*; European Commission Joint Research Centre: Luxembourg, 2008.
28. Schlosberg, D. *Defining Environmental Justice. Theories, Movements, and Nature*; Oxford University Press: Oxford, 2007.
29. Lowe, P.; Goyder, J. *Environmental Groups in Politics*; Allen and Unwin: London, 1983.
30. Lafferty, W.M. From environmental protection to sustainable development: The challenge of decoupling through sectoral integration. In *Governance for Sustainable Development. The Challenging of Adapting Form to Function*; Lafferty, W.M., Ed.; Edward Elgar: Cheltenham, 2004; 191–220.
31. Gurran, N.; Gilbert, C.; Phibbs, P. (2015) 'Sustainable development control? Zoning and land use regulations for urban form, biodiversity and green design in Australia', *J. Environ. Plan. Manage.* **2015**, *58*(11), 1877–1902.
32. Therivel, R.; Partidario, M. *The Practice of Strategic Environmental Assessment*; Earthscan: London, 1996.
33. Wackernagel, M.; Rees, W. *Our Ecological Footprint: Reducing Human Impact on the Earth*; New Society Publishers: Gabriola Island, 1996.
34. Curry, R.; Ellis, G. Material flow analysis. In *Routledge Companion to Environmental Planning*; Davoudi, S., Blanco, H., Cowell, R., White, I., Eds., 2019, 346-357, Routledge, Abingdon UK and New York.
35. Jørgensen, S.E. *Eco-Exergy as Sustainability*; WIT Press: Ashurst, 2006.
36. Cowell, R.; Lennon, M. The utilization of environmental knowledge in land use planning: Drawing lessons for an ecosystem services approach. *Environ. Plan. 'C': Gov. Policy* **2014**, *32*, 263–282.
37. Harris, N. Exceptional spaces for sustainable living: The regulation of one planet developments in the open countryside. *Plan. Theory Pract.* 2019 *20*(1), 11–36.
38. Jepson, E.J.; Haines, A.L. Zoning for sustainability: A review and analysis of the zoning ordinances of 32 cities in the United States. *J. Am. Plan. Assoc.* **2014**, *80* (3), 239–252.
39. Long, J. Constructing the narrative of the sustainability fix; sustainability, social justice and representation in Austin TX. *Urban Stud.* **2016**, *53* (1), 149–172.

40. Meadowcroft, J. National sustainable development strategies: Features, challenges and reflexivity. *Eur. Environ.* **2007**, *17*, 152–163.
41. Van Muijen, M. The Netherlands: Ambitious on goals—Ambivalent on action. In *Implementing Sustainable Development: Strategies and Initiatives in High Consumption Societies*; Lafferty, W., Meadowcroft, J., Eds.; Oxford University Press: Oxford, 2000; 142–173.
42. Jänicke, M.; Weidner, H., Eds. *National Environmental Politics: A Comparative Study of Capacity-Building*; Springer: Berlin, 1997.
43. Weale, A. *The New Politics of Pollution*; Manchester University Press: Manchester, 1992.
44. Fudge, C.; Rowe, J. Ecological modernisation as a framework for sustainable development: A case study in Sweden. *Environ. Plan. A* **2001**, *33*, 1527–1546.
45. Adams, D., Scott, A.J.; Hardman, M. Guerrilla warfare in the planning system: Revolutionary progress towards sustainability? *Geogr. Ann. B* **2013**, *95* (4), 375–387.
46. Pandis Iverot, S.; Brandt, N. The development of a sustainable urban district in Hammarby Sjöstad, Stockholm, Sweden? *Environ. Dev. Sustain.* **2011**, *13*, 1043–1064.
47. Yenneti, K.; Day, R.; Golubchikov, O. Spatial justice and the land politics of renewables: Dispossessing vulnerable communities through solar energy mega-projects. *Geoforum* **2016**, *76*, 90–99.
48. Flint, J.; Raco, M., Eds. *The Future of Sustainable Cities. Critical Reflections*; Policy Press: Bristol, 2012.
49. Jensen, B. Masdar City: A critical retrospective. In *Under Construction: Logics of Urbanism in the Gulf Region*; Wippel, S., Bromber, K., Steiner, C., Krawietz, B., Eds.; Routledge: London, 2016; 45–54.
50. Woltjer, J. *Consensus Planning: The Relevance of Communicative Planning Theory in Dutch Infrastructure Development*; Ashgate: Aldershot, 2000.
51. Van der Straaten, J. The Dutch NEPP: To choose or to lose. *Environ. Polit.* **1992**, *1* (1), 45–71.
52. Smith, A.; Stirling, A. Moving outside or inside? Objectification and reflexivity in the governance of sociotechnical systems. *J. Environ. Policy Plan.* **2007**, *9* (3/4), 351–373.
53. Boston, J. *Governing for the Future. Designing Democratic Institutions for a Better Tomorrow*; Emerald Group Publishing: Bingley, 2016.
54. Degeling, P. The significance of ‘sectors’ in calls for urban publish health intersectoralism: An Australian perspective. *Policy Polit.* **1995**, *23* (4), 289–301.
55. Kemp, R.; Parto, S.; Gibson, R. Governance for sustainable development: Moving from theory to practice. *Int. J. Sustain. Dev.* **2005**, *8* (1/2), 12–30.
56. Holling, C.S. *Adaptive Environmental Assessment and Management*; Wiley: Chichester, 1978.
57. Lessard, G. An adaptive approach to planning and decision making. *Landsc. Urban Plan.* **1998**, *40*, 81–87.
58. Smith, A.; Kern, F. The transitions storyline in Dutch environmental policy. *Environ. Polit.* **2009**, *18* (1), 79–98.
59. Newig, J.; Voss, J.; Monstadt, J. Editorial. Governance for sustainable development in the face of ambivalence, uncertainty and distributed power. *J. Environ. Policy Plan.* **2007**, *9* (3/4), 185–192.
60. Jordan, A. The governance of sustainable development: Taking stock and looking forwards. *Environ. Plan. C Gov. Policy* **2008**, *26*, 17–33.
61. Walker, G.; Shove, E. Ambivalence, sustainability and the governance of socio-technical transitions. *J. Environ. Policy Plan.* **2007**, *9* (3/4), 213–225.
62. Shove, E.; Walker, G. Transitions ahead: Politics, practice, and sustainable transition management. *Environ. Plan. A* **2007**, *39* (4), 763–770.
63. Cowell, R.; Owens, S. Governing space: Planning reform and the politics of sustainability. *Environ. Plan. C Gov. Policy* **2006**, *24* (3), 403–421.
64. Kitschelt, H. Political opportunity structures and political protest: Anti-nuclear movements in four democracies. *Br. J. Polit. Sci.* **1986**, *16*, 58–95.

65. Freudenberg, N.; Steinsapir, C. Not in our backyards: The grassroots environmental movement. *Soc. Nat. Res.* **1991**, *4* (3), 235–245.
66. Hayden, A. Stopping Heathrow Airport expansion (for now) lessons from a victory for the politics of sufficiency. *J. Environ. Policy Plan.* **2014**, *16* (4), 539–558.
67. Hudalah, D.; Winarso, H.; Woltjer, J. Planning by opportunity: An analysis of peri-urban environmental conflicts in Indonesia. *Environ. Plan. A* **2010**, *42* (9), 2254–2269.
68. Owens, S. Land, limits and sustainability: A conceptual framework and some dilemmas for the planning system. *Trans. Inst. Br. Geogr.* **1994**, *19*, 439–456.
69. Dierwechter, Y. *Urban Growth Management and Its Discontents. Promises, Practices and Geopolitics in U.S. City Regions*; Palgrave Macmillan: New York, 2008.
70. Owens, S.; Cowell, R. *Land and Limits: Interpreting Sustainability in the Planning Process*, 2nd Ed.; Routledge: London, 2010.

18

Sustainable Development

Mark A. Peterson

Introduction

This entry defines sustainable development and its three basic aspects. Because sustainable development is a relatively new concept, a short history and description of the drivers that lead to sustainable development are described. Then, a sustainable energy future is presented. To be sustainable as a society requires cooperation and collaboration rather than command and control management. A very key aspect of sustainable development, social synergy, is covered. Also, because sustainable development is relatively new but essential to build a better and viable future, children from the earliest age through college need to learn to understand and apply the concepts of sustainable development. A paragraph is included that describes current efforts in the United States to incorporate education in sustainable development (ESD) in K–12 and college curricula.

Sustainable development encompasses stewardship of many areas of human and planetary life. In business, one of the key motivators is to implement sustainable development measures to be profitably successful indefinitely. To do this requires that businesses show their due diligence to both society and the environment while maximizing profits. Sustainability reporting assists businesses in assessing their efforts. Sustainability reports are both management and public relations tools. An innovative, very effective, time-tested form of sustainable development, “renting a service” rather than “selling a product,” is covered. Then, sustainable development for community vitality is described. As the sizes, types, and socioeconomics of communities vary considerably, so do all of the related aspects of evolving them to be sustainable. Reference is made to a web site that thoroughly describes all aspects. Some of the subjects covered on the web site are briefly summarized in this entry. The final section covers the vast, opening market of sustainable development in developing countries.

What Is Sustainable Development?

Sustainable development is development that meets the needs of the present without compromising the ability of future generations to meet their needs. Sustainable development has three aspects:

1. Social (people)
2. Environmental (planet)
3. Economic (profits)/prosperity

All development affects all three aspects. All three aspects are interdependent. Thus, being mindful of these interdependencies in management and leadership decisions will result in the best overall solution—a win-win-win solution that maximizes success and minimizes any negative social, environmental, and economic costs. This is called managing the triple bottom line of people, planet, and profits. This is also called whole systems thinking[1,2] because all relevant factors are considered as a whole. The role of engineers is to help their clients be successful. This requires integrated whole systems thinking that covers all related liabilities that a company or community (their client) may have and provides the most efficient and profitable solution to the challenge. Often, the best whole system solution is also the most efficient and most sustainably profitable.

The environmental (planet) aspect is significantly affected by energy consumption and management, including: the entire national power infrastructure and distribution, transportation, plus the construction and renovation of all residential, commercial, and industrial facilities.

History, Environmental Degradation, and National Security

During the last century, while fossil fuels were abundant and cheap, those fuels fulfilled a majority of our energy conversion needs. The mounting problem is that combustion emissions have fouled the environment in a number of ways, resulting in increases in respiratory illnesses, mercury pollution, and a rise in global temperatures. The quantity of easily retrieved fossil fuels is significantly depleted. Coal is still relatively abundant, but it does not burn cleanly. Technologies need to be developed to both mine the coal safely and to burn it cleanly. Regarding petroleum, many countries that are not friendly to the United States control most of the remaining easily extractable sources. Many national security advisors have indicated the urgency of severing our dependence on foreign oil as a part of an overall strategy for the security of the United States and as a means to prevent oil-related conflicts.[3,4] It is also apparent that there is a need to protect and allow the environment to regenerate. The effect of using fossil fuels extensively and inefficiently is that we are simultaneously poisoning the environment and ourselves. Nuclear power emissions are clean, but the nuclear power industry has significant obstacles such as storage of radioactive wastes for many thousands of years. In addition, there are security concerns to safeguard radioactive material from being stolen for production of atomic weapons.

Governors and mayors are taking action to implement clean energy technologies. In response to clear signs of increased cost from continuing to use fossil fuels and scientific evidence showing that by burning fossil fuels we are initiating a possibly devastating global warming trend[5] that could flood coastal cities, disrupt the food chain, and change climate patterns significantly, many states have taken the initiative and enacted renewable energy portfolios to fund the transition to renewable energy resources. Many remaining states are in the process of developing their own renewable energy portfolios. These renewable energy portfolios provide significant state- and utility-sponsored financial incentives for the commercial, industrial, and residential use of renewable energy systems and fuels. On the city level, many mayors from major cities around the world have made commitments to cut greenhouse gas emissions to slow the rate of global warming.[6] Many of these cities are coastal and could be severely impaired or destroyed from rising sea levels from global warming. So, civic action to switch to cleaner energy options is beginning in earnest.

Energy Future

What this means to energy engineers is that petroleum-derived fuels are on their way out over the next half century. Hydrogen (where the hydrogen is derived from renewable energy sources), ethanol, biodiesel, and other forms of renewable fuels are on their way in. Direct and indirect conversion of solar energy, including wind, biomass, wave/tidal power, and small-scale hydroelectric power will increasingly be part of the energy infrastructure that energy engineers will design and build. The bottom line with energy is that it needs to be relatively nonpolluting and indefinitely available. It is a very dynamic time for energy engineers as the entire, worldwide energy picture transitions to clean renewable technologies. This will eventually add a lot of stability to the world economy, the world political environment and to everyone's lives. The stability will come from the fact that renewable energy technologies can be used to tap the natural energy resources that are available everywhere. Stability will also come, as the environment regenerates, the climate stabilizes and resources remain available for our sustenance.

How We Socially and Professionally Interact with Each Other Determines Our Degree of Success

Another aspect of sustainability has to do with how well we interact and collaborate. In the past, management of most activities was by a top-down hierarchy. Now humanity is evolving and it is driven by high levels of sophistication in technology and communications, which has resulted in individual knowledge and skill level increases. Plus, many families are now structured such that individual adults and children are taking more independent responsibilities for the many aspects of their lives. This has all created a desire by many people to be more intimately involved in solutions rather than just letting someone else do the thinking. In this new economy, teamwork and open communication are important to bring all stakeholders together, whether in a business or community, to manage by consensus and cooperation. Everyone affected should have the opportunity to be involved in the solution, even if just by being informed as the planning and decision-making are in process. The outcome then is one that promotes efficiency, for the simple reason that all persons affected are involved, which promotes enthusiasm and "buy-in." Typically with this process, more work and time are invested up-front such that all aspects are considered and thus everything proceeds more efficiently down-stream.[7]

Creative, Cooperative, Design and Planning Teamwork[2]

The ASHRAE GreenGuide recommends "integrated design teams" that have all of the design, economic, planning, and other related disciplines involved up-front to create better designs. If this process is not used, design typically proceeds in a series of "handoffs" that tend to compound problems, as each succeeding team designs "around" any incompatibilities that the previous designers have already finished. This adds unnecessary complexities and inefficiencies, which increase construction and life-cycle costs. Through coordination and the collaboration of designers, architects, engineers, and key players, an integral design can be created that functions as an efficient system and not as a collection of parts that are force-fit together. This approach has enabled design teams to design very energy efficient, comfortable, aesthetically and environmentally friendly buildings at or less than the conventional price per square foot of traditionally designed buildings. So, working as a team from the beginning is the most efficient way of designing because potential conflicts are resolved up-front rather than later at a higher cost. Effectually, when people are creating and they know that their contributions are respected, superior planning and designs are achieved. The upfront work of coordinating planning and brainstorming sessions with many people of diverse backgrounds can be a challenge. However, the results and life-cycle costs are almost assuredly optimal.

Education

The United Nations has declared the decade from 2005 to 2014 as The Decade of Education for Sustainable Development.[8]

There are many national teams around the world that have taken the lead to work across the educational spectrum—including public and private education, primary and higher education, independent, charter, and home-schooling—to incorporate ESD in their curricula. The U.S. Partnership for The Decade of Education for Sustainable Development[9] was formed to facilitate implementation of ESD in the United States.

Effective education can demonstrate the inter-relationship and interdependence of people, planet, and profits in all life activities. We have the technology to transition to a clean energy future and to manage materials in a cyclic manner. We know that pollution is causing global environmental change. We also know that teamwork and cooperation get better results than working in a hierarchical or isolated manner. Education in sustainable development will show that the current "consume and throw away" economy no longer works for the benefit of humanity and life. Rather, there is a need for cyclic, whole-systems thinking that integrates all relevant factors into the best, longest lasting results.

Sustainability Reporting

Realizing the importance of being "sustainable" and understanding that they are good long-term investments, many companies are developing sustainability reports. These reports are strong management tools that show how well a company is progressing with their sustainability, their corporate social responsibility (CSR), and their goals to continually improve. They provide an openness that stakeholders (employees, stockholders, regulatory agencies, customers, and community leaders) expect so they know if a company is a good place to work, a good investment, or a good neighbor. These reports are available to the public (free download).[10] For engineers, preparation of sustainability reports involves the collection and analysis of energy and environmental data. Demonstration of energy savings plans and associated pollutant emission reductions can have significant public relations and market share value. This can promote higher sales and flow of investor capital as companies prove their social responsibility and long-term viability.

"High 5!—Communicating your Business Success through Sustainability Reporting—A Guide for Small and Not-So-Small Businesses," from the Global Reporting Initiative, describes the benefits of sustainability reporting:

Sustainability reporting has many advantages that benefit different areas of your business. Some benefits are purely financial while others deal with customer or employee satisfaction. Sustainability reporting helps organizations identify and address their current and potential risks, saving time and money in the short and long term. As the public becomes more aware of your efforts, customer loyalty and credibility of your business will greatly increase. When you take a deeper look into your daily business operations through sustainability reporting you will be able to discover new opportunities.

Businesses continuously seek to generate income and acquire a competitive edge by identifying new market opportunities and determining current and potential risks. When your organization embraces a sustainability perspective, i.e., simultaneously addressing social, environmental and economic issues, you can benefit from cost-savings and improvements in product quality and employee performance.

When considering how to improve financial performance, many organizations only look at financial aspects, such as the cost of purchasing goods, personnel costs, or tax payments. However, working on environmental and social issues can also positively affect your financial bottom line.

Sustainability reporting is the way to identify these potential benefits and realize the economic gains. It helps you achieve your business goals by setting up a continuous improvement process based on target setting and progress measurement. All in all, sustainability reporting helps you to acquire that competitive advantage.

Renting vs. Buying (A Sustainability Innovation)[11,12]

Here is an example of systems thinking to ensure that a product has minimal environmental impact plus high social and economic value:

In today's consumer/throw-away economy, typically a product is manufactured and sold. There is producer incentive to minimize the use of labor and material resources put into a product, thus saving on cost. The product is sold as cheaply as possible to maximize sales. The product eventually wears out and is disposed of. A product that breaks and is disposed of soon after its warranty period is best for sales, so that a consumer will go out and buy a replacement. This is obviously a wasteful scenario, which is prevalent in commerce today. However, that is changing.

A more efficient scenario that many companies have successfully deployed for years is to manufacture and rent a product. Then, there is an incentive to maximize the utility and life of the product, thus ensuring an income to the owner/renter for as long as possible into the future. (This is what sustainable development is all about—maintaining economic flow for as long as possible into the future.) When this product is manufactured, due care in manufacturing processes and sufficient material are used to maximize durability, reliability, and longevity. Customer satisfaction is high from having a reliable product, which also builds name recognition and increases desirability. The product is designed to be easily maintainable, again, to maximize longevity, and also to facilitate dismantlement of the product when it has come to the end of its useful life. Thus, the parts can be easily remanufactured and reused or segregated for efficient recycling of the raw materials. So, maximum utility and income is achieved from the product, and most of the resources that went into manufacturing and maintaining the product throughout its life are recycled with minimal impact to the environment.

So, engineers with sustainable development in mind might be thinking of the "rented" product scenario, which is inherently efficient, rather than the "sold" product scenario, which is inherently wasteful. For example, Interface, Inc., the largest carpet manufacturer in the world and a company that has committed itself to be as sustainable as possible, is using this renting concept in its products. They rent carpet tiles. As the carpet eventually wears, tiles are replaced and recycled to make new tiles.

Community Development to Decrease Energy Costs[13]

Over the last three-quarters of a century, there was an assumption that gasoline and diesel fuel would be cheap and plentiful, indefinitely. Thus, urban sprawl developed because of the low cost of owning and operating one or more cars. This is not the case anymore! Now, with the realization that cars are expensive to both purchase and operate, there are efforts by many cities to re-establish neighborhoods that have all of the amenities needed for occupants all within a short distance that can be covered on foot, with a bicycle, or with convenient public transportation. These cities are excluding automobiles from certain areas and some are charging admission fees for cars to enter semirestricted areas. This has helped to revitalize many city commercial districts because of the park-like feeling of being in these areas without the noise and exhaust from cars. With the economic benefits that have been realized by these arrangements, more and more urban and suburban cities are pursuing these types of commercial district renovations in their cities.

Obviously, society has made a major investment that established urban/suburban sprawl. That investment now needs to be made sustainable. Consequently, there will be major investments in clean fuels such as biodiesel, ethanol, and renewably derived hydrogen to power the huge fleet of vehicles in the United States. A remaining economic burden to maintain suburban living is the maintenance of roads. However, that may be relieved as convenient, modern, and cleaner public transit busses, trains, and other guided vehicles are developed into transportation networks, thus decreasing the number of cars on the road. Energy engineers will be integral to this transition in transportation.

So, the paradigm for energy engineers will evolve as community structures around urban and suburban cities change and improve for energy efficiency. There will probably be a tendency to create nodes, where residents have a majority of the amenities nearby. Then, clean, comfortable, energy-efficient public transportation and express lanes for cars will connect each node.

Overview of the Many Aspects of Sustainable Communities Infrastructure and Nature

The web site http://www.conservationeconomy.net/IN-DEX.CFM (courtesy of the Ecotrust) summarizes the many and various aspects of sustainable communities and preserving our natural assets. Many social factors that play a role in the culture and systems changes associated with sustainable development are thoroughly explained on the web site. Engineers should know these social interactions, which are cohesive and essential to sustainable communities. Readers are encouraged to visit the web site to get a flavor of all of the aspects of sustainability or to narrow in on aspects that are of particular interest to them. The following is a narrative summary of some of the entries on the web site:

A conservation economy describes how social capital, natural capital, and economic capital can be synergetic and sustainable.

Social capital covers fundamental needs, which include: the strong need for local sources of food; accessible, healthy shelter; healthy environment and access to healthcare; plus access to knowledge about the interconnectedness of us to our environment and to each other. The section on community discusses collaborative processes that honor: social equity, which promotes prosperity for all; security from fear and violence; recognition of the wealth and strength in our cultural diversity and establishment of a will to preserve it; plus, local celebrations to honor a sense of place with the community and environment. It proceeds to describe the importance of enjoying beauty and play, to relieve stress from our busyness; learning to welcome transitions that improve communities as a whole; and establishment of civic society where all residents can manage their communities collaboratively.

Natural capital includes the atmosphere, biosphere, and earth. To sustain it requires ecological land use, which includes connected wild lands, in which indigenous animals, plant and people can coexist together. Protected core reserves can be set aside for native plants and animals to thrive, without interference. Wildlife corridors can connect reserves such that animals may migrate freely and parks can be established to act as buffer zones between developed and undeveloped areas. Productive rural areas can be re-established through: sustainable agriculture, which does not cause runoff of pollutants into streams and is more in harmony with natural processes; sustainable forestry, which thins rather than clear-cuts stands of trees; sustainable fisheries, which establish quotas, such that species are not depleted; and eco-tourism, to provide natural getaways and education for people that want to learn more about sustainability and to be close to nature. Compact towns and cities, with human scale neighborhoods, green buildings, convenient transit access, ecological infrastructure, and urban growth boundaries will create healthy, vibrant communities.

Economic capital in healthier communities and commerce will tend to expand through synergies with social and environmental capital, thus building prosperity.

Energy Engineering for Developing Countries[13]

Approximately four billion of the six and one half billion people on the planet live in extreme poverty, where a poor sanitation infrastructure results in disease, there is minimal economic productivity, and there is significant economic burden on governments and aid agencies. However, recently, it has been found that given some of the modern necessities such as water wells, electronic communication, and dependable energy, people that are impoverished can quickly and enthusiastically become productive to their communities and not be an economic burden.

The use of small photovoltaic power systems in villages has literally energized villages into minieconomic zones of relative wealth and flow of capital, thus creating self-sufficiency rather than dependence.

Just providing electricity for lighting, water pumps, and small power tools can tremendously boost the productive capabilities of a village.

The establishment of cell phone repeaters and wireless infrastructure is much less expensive than running miles of telephone cable. A limited number of cell phones in villages have also promoted prosperity because farmers and merchants can effectively communicate and market their products.

As developing countries continue to develop, water, energy, and communication will be vital to their success. So, energy engineers will be a key part of this process. Many multinational and small companies and financial institutions are tapping into bringing these four billion people into the world of commercial success and out of the world of poverty. So, engineers will play an important part in this next major step toward a sustainable world.

Conclusion

Though we are just now seeing the start of a transition to sustainable development, it is clearly economically, ecologically, and socially advantageous to choose this means to success, prosperity, well-being, stability, and peace. We have an abundance of all of the material and energy resources we need. We have the ability, knowledge, and conscience to do the best for humanity, all of life, and the planet. We have the inspiration and insight to transition to a sustainable world. Now all it takes is the willingness to accept change and make the transition to sustainability. The outcome will be a world that is more prosperous and stable than it is now. Sustainable development is a goal to embrace and make part of our daily decisions.

References

1. *Waage, S. Ants, Galileo, and Gandhi—Designing the Future of Business Through Nature, Genius, and Compassion; Greenleaf Publishing: Sheffield S3 8GG, U.K., 2003; [www.greenleaf-publishing.com].*
2. Stecky, N. *Introduction to the ASHRAE Greenguide for Leed, Globalcon 2005 Proceedings CD.* Available at https://www.aeecenter.org/store/detail.cfm?idZ895&category_idZ6.
3. Woolsey, R.J.; Hunter, L.; Amory, L. *Energy Security: It Takes More Than Drilling.* Available at http://www.csmonitor.com/2002/0329/pns02-coop.html.
4. Woolsey, R.J.; McFarlane, B. Conference Report: *Renewable Energy in America: The Call for Phase II, Session 3: Using Renewable Energy to Meet Our National Security Needs;* December 2004. Available at http://www.acore.org/pdfs/04policy_report.pdf.
5. Available at http://www.insnet.org/ins_headlines.rxm?custZ2&idZ1234.
6. Available at http://www.insnet.org/ins_headlines.rxm?custZ2&idZ1248.
7. Anderson, C.; Katharine, R. *The Co-Creator's Handbook,* http://www.globalfamily.net.
8. Available at http://portal.unesco.org/education/en/ev.php-URL_IDZ27234&URL_DOZDO_TOPIC&URL_SECTIONZ201.html.
9. Available at http://www.uspartnership.org.
10. Available at http://www.CorporateRegister.com.
11. Williard, B. *The Sustainability Advantage—Seven Business Case Benefits of a Triple Bottom Line*; New Society Publishers: British Columbia, V0R 1X0, Canada, 2002; [http://www.newsociety.com].
12. *Doppelt, B. Leading Change Toward Sustainability—A Change-Management Guide for Business, Government and Civil Society; Greenleaf Publishing: Sheffield S3 8GG, U.K., 2003; [http://www.greenleaf-publishing.com].*
13. Available at http://magma.nationalgeographic.com/ngm/data/2001/07/01/html/ft_20010701.3.html.
14. Available at http://www.csrwire.com/print.cgi?sfArticleIdZ3709.

19

Urban Agriculture

Natalia Fath

> How we eat determines, to a considerable extent, how the world is used.
>
> **Wendell Berry (2010)**

Introduction

The practice of growing food in cities, or urban environments, is known as urban agriculture. Integration of food production into the urban ecosystem is not a new phenomenon; however, the reasons behind this growing practice have changed over the past few centuries. Urban agriculture arose during difficult times, such as during war or depression, when the growth of vegetables on any available plots was encouraged as a response to dwindling food supply and food shortages. In addition, this was considered as an activity that provided jobs and income and boosted spirits and community cohesion during harsh times. The concept of “allotment gardens” was developed in Germany as a response to rapid industrialization in Europe throughout the 19th century (Holmer and Drescher 2005). Growth of cities during that time was accompanied by growing food insecurity and by allowing citizens to use open spaces to grow food, originally called “the gardens of the poor”, the administrations of the cities, thus, contributed to the idea of urban farming. The late 19th century witnessed the initiative by Detroit’s Mayor, Hazen S. Pingree, when the citizens were asked to utilize any available land to grow vegetables as a response to economic slowdown (Austin 2019) – coincidentally, now that Detroit is again a leader in urban farming due to recent economic collapse with the loss of the automobile industry. These original initiatives were further developed during World War I and World War II (WWII) and received a new meaning – they were called “victory gardens”. Misallocation and devastation of agricultural land in Europe and loss of farmers to military recruitment led to the growth in popularity of these fruit and vegetable gardens, specifically in the United States, Canada, the United Kingdom, Germany, and Australia (Drescher 2001). In the United States alone, these gardens produced about 1.45 million quarts of canned fruits and vegetables in 1918 (Schumm 2018). The United States Department of Agriculture (USDA) further continued to encourage expansion of the victory gardens during the hardships of WWII. Thus, the existence of agricultural production in urban environments has evolved to be more systematic and expected.

The purpose of establishing those gardens in the cities has shifted since the end of WWII. During times of relative food abundance, these practices may be seen from a different perspective. While still providing food, which is especially crucial for food-insecure districts, and part-time or full-time jobs, urban farms also now looked at as a place for social gathering or as a green area that helps fighting air pollution in cities. Advantages of urban farming include but are not limited to social, economic, and health benefits. However, there are some risks involved as well, which include but are not limited to required high inputs of energy and time, and high probability of polluted urban soil and air. It has been attempted to estimate the proportion of food that is provided by urban agriculture globally. According to van Veenhuizen and Danso (2007), 15%–20% of the world's food supply comes from urban agriculture. A recent publication by Clinton et al. (2018) suggests that the above numbers are highly overestimated, perhaps because of inclusion of peri-urban areas and livestock production. The authors claim that if the space is utilized at its maximum, and the practices in place are intense, then urban agriculture could produce ~5% of the global crop production.

Types of Urban Farming

Food production in cities can take various forms. There are different projects and activities that have been developing and have now firmly established themselves in urban environments. Several major categories should be mentioned in respect to urban farming: (1) community-based, (2) institutional, (3) public, and (4) commercial (Steele 2017). The borderlines between them are not rigid and, in particular cases, may carry the features of two or three categories. Urban agriculture is a flexible concept and adjusts to the local conditions, thus acquiring various forms. There is not a widespread agreement on the classification of types of urban farming, however, there are types that are commonly recognized. Specifically, community and backyard gardens; rooftop and balcony gardening; growing in vacant lots, right-of-ways, and parks; aquaponics; hydroponics; fruit trees and orchards; market farms; raising livestock and beekeeping (Steele 2017). In addition, it can be stated that post-harvest activities, for example, "creating value-added products in community kitchens, farmers' markets and road-side farm stands, marketing crops and products, and addressing food waste" (Steele 2017), should also be considered as overall urban agricultural activities.

Community and backyard gardens involve growing food for the family or the neighborhood; rooftop gardening utilizes available space on roofs while growing in vacant lots puts a productive use on otherwise abandoned land; hydroponics practices soilless food growth; and aquaponics combines the latter with raising fish. Some cities allow raising chickens or keeping beehives in urban environments.

Additionally, the classification system of urban farming types can explore the following categories: location, people involved, products grown, economic activities, product destination, scale, and technology used (RUAF Foundation 2019). However, currently, urban agriculture researchers have been using their own approach to typify this activity, which has led to a significant variety of divisions and subdivisions of local urban farming systems. Thus, it appears to be beneficial to continue developing a consistent viewpoint of types of urban farms (van Veenhuizen and Danso 2007).

Regardless of the classification system used, it is imperative to realize that food production in cities is fully integrated into the urban economic and ecological systems, and thus should be an important part of urban planning policies (RUAF Foundation 2019).

Benefits of Urban Farming

Social and Cultural Benefits

The practice of growing food in cities can be viewed as an excellent opportunity for social integration, cultural exchanges, and education for both adults and youth. Through collective action and advocacy, city dwellers build and strengthen their communities, as urban farms in their various forms mostly appear to be based on a sense of community (Holland 2004). Working as a team on underutilized land plots and transforming them into productive environments that can feed them or can be sold, residents improve the quality of their immediate surroundings and develop bonding (see Figure 1). Of particular importance, relevant to urban food production is the opportunity for people of different ages to interact which promotes intergenerational bond, and the chance for disadvantaged or segregated groups such as people from different socioeconomic classes or different backgrounds, orphans, women, recent immigrants without jobs, elderly and disabled people, to get involved in community building, to become engaged citizens, and to maintain decent livelihoods (Santo et al. 2016).

Urban farming, thus, can be an "agent of change" as it encourages social bonding and activities that have the potential to promote social, economic, and environmental changes at the local level (Holland 2004).

Urban farms can also play an important role in educational functions – participants, and especially youth, learning about ecology and the process of food growth, which in turn, can encourage growth in agricultural literacy and reconnection to nature (Santo et al. 2016). Advancing urban dwellers' knowledge on how their food is grown and distributed, may lead to the appreciation of the urgent need for cities to adopt "circular" food production system rather than current "linear" to ensure their own sustainability. Specifically, urban "outputs", like organic waste production, can and should become an "input" into urban food production system (Deelstra and Girardet 2000). Recycling and composting of organic matter and nutrient recycling on local urban farms reduce dependence on delivered "inputs" and keeps this waste from landfills. By participating in the food-producing process, citizens foster an understanding and appreciation of systems thinking and the cycles of life; and therefore, the importance of urban sustainability as the world is predominantly urban and the pattern is predicted to continue into the future.

Economic Benefits

As some types of urban agriculture utilize vacant lots, it leads to increased property values, which in turn generates increased tax revenues (Steele 2017). This is particularly true in economically disadvantaged neighborhoods (see figure 1). Among other economic benefits, it can be mentioned that this activity

FIGURE 1 Baltimore, MD, USA – low-income neighborhood, underutilized land. Real Food Farm, Perlman Place. Planted fruit trees (fig). (*Photo credit: N. Fath.*)

adds jobs to the local economy and stimulates capital investment and redevelopment (Santo et al. 2016). Guitart et al. (2012) analyzed published research on community gardens. Their review found that all 13 of the studies which examined property values reported increased property values associated with the existence of the community garden. Similar research findings can be found in Voicu and Been (2008).

It can be argued that urban agriculture makes fresh produce more affordable. This is particularly important for low-income neighborhoods and for poorer countries where between 50% and 70% of income on being spent on food. However, based on the currently available studies, it can be concluded that the criteria and methods used to assess specific profits and other economic impacts differ markedly and/or are lacking (van Veenhuizen and Danso 2007). In addition to the direct economic benefits to the participants, urban agriculture also encourages the growth of small enterprises; farming activities require necessary agricultural inputs that can vary from compost and organic pesticides production to tool production and managing of the outputs that can include packaging, marketing, and transportation of products (Homem de Carvalho 2001).

Besides the above discussed economic advantages, urban farms surplus can be donated to food banks which, in turn, can help reduce money spent on food for low-income neighborhoods. In addition, when vacant land is utilized by urban agriculture activities, it can save maintenance costs for the city as illegal waste dumping and/or vandalism will decrease (Steele 2017).

It is important to note that there is a need for more research on economic outputs of urban farming, profitability of commercial food production in cities, long-term studies on employment opportunities and on neighborhoods' indicators, discussed above. Economic outcomes are the "least documented aspect of urban agriculture" (Hodgson et al. 2011).

Health Benefits

There are several aspects relevant to this topic that should be discussed, specifically, improved food security and physical and emotional well-being. Poorer neighborhoods often find themselves living in so-called "food deserts". USDA defines food deserts as "parts of the country vapid of fresh fruit, vegetables, and other healthful whole foods, usually found in impoverished areas" (*Nutrition Digest* 2019). To qualify as a food desert, "at least 500 people and/or at least 33 percent of the census tract's population must reside more than one mile from a supermarket or large grocery store (for rural census tracts, the distance is more than 10 miles)" (*Nutrition Digest* 2019). When analyzing the ability of urban agriculture to improve food security, one should look at the individual/household and municipal levels. Neither, though, should be understood as a substitute to traditional food retail (Santo et al. 2016). While not the whole answer to urban food deserts, farming in these locations provides participants with access to some fresh produce that otherwise is unavailable; thus, improving their food security and providing a better diet that includes more fresh fruits and vegetables (Zezza and Tasciotti 2008, Smith and Harrington 2014). Numerous studies report participants, their households, as well as their neighbors and friends, gaining access to a variety of seasonally appropriate foods while saving money for other needs (Corlett et al. 2003, Wakefield et al. 2007, Kortright and Wakefield 2011). However, more empirical evidence and longitudinal studies are necessary to support the claim that urban farming can significantly amend food security over time in poorer urban neighborhoods (Siegner et al. 2018). While there is accumulated evidence that supports increased produce consumption among the gardeners (Alaimo et al. 2008, Litt et al. 2011), experts suggest it is not an overall significant indicator of improved food security and/or quality of the diets in low-income households (Hallsworth and Wong 2013). Based on physical land availability in cities, along with other barriers, urban agriculture's capacity to ameliorate food security is not strongly supported; however, it can be a supplemental solution for the urban poor (Badami and Ramankutty 2015).

Improved access to fresh produce, fruits and vegetables, may lead to better nutrition, and thus can be considered a viable strategy to mitigate certain diet-related diseases, such as obesity and diabetes (Steele 2017). However, through many studies and research done, it appears that via urban farming activities most notable and well-documented gains are in improved mental health and physical health in participants. This is accomplished through acquiring a purpose, seeing the results of one's effort, and being physically active (Santo et al. 2016). In addition, simply being close to nature has been observed to be a stress-relieving activity (Armstrong 2000, Brown and Jameton 2000, Wolf and Robbins 2015). Increased green cover in cities associated with urban agriculture activities contributes to such ecosystem services as air filtration and temperature moderation, thus also benefiting the overall well-being of urban residents (Wolf and Robbins 2015).

Environmental Benefits

Besides social and human health benefits of urban agriculture, arguably, environmental advantages associated with this activity are the dominant ones. Among particular benefits to the environment are improved air quality and mitigation of the urban heat island effect. Extensive research studies support that replacing impervious surfaces with vegetation and trees, in particular, leads to cleaner air, as vegetation filters particulate matter, and to cooler temperatures during summer. For example, in the Phoenix metropolitan area, increased tree canopy cover by 25% leads to a total cooling of 7.9°F as compared to a bare neighborhood, and addition of greenery to residential backyards reduces average neighborhood temperatures by 0.4°F–0.5°F (Middel and Chhetri 2014). This phenomenon appears to be of particular importance for low-income neighborhoods as they tend to have fewer trees and grass cover and, thus, higher summertime temperatures (Jenerette et al. 2011).

Among other environmental advantages of urban agriculture are its ability to maintain and support biodiversity via creation of habitats; its contribution to recycling of nutrients when organic waste is turned into a resource – soil (see Figures 2 and 3); its ability to increase water infiltration, thus reducing stormwater runoff and decreasing the flooding potential (van Veenhuizen 2006, Santo et al. 2016).

FIGURE 2 Composting bin at aquaponics farm, Baltimore, MD, USA. Center for a Livable Future, JHU. (*Photo credit: N. Fath.*)

FIGURE 3 Composting bins at urban farm, Towson University, MD, USA. (*Photo credit: N. Fath.*)

Urban farming should also be looked at from the perspective of one of the most pressing challenges present day – scarcity of freshwater and agriculture being the largest user of freshwater. Some types of urban farming contribute to lessening this issue by collecting rainwater on site and using it for irrigation (see Figure 4). Others are capable of re-using wastewater, thus lessening the demand for freshwater supply (Haysom 2009). A particular type of urban agriculture, aquaponics, attempts to recycle water used for fish and produce production. While initially the system requires a significant input of water, the continued operation is based on this water being re-used and accompanied only by some expected losses due to evaporation, spillage, etc., which could be replenished by rainwater collection (Love et al. 2015). United Nations' Food and Agriculture Organization, FAO, is particularly invested in exploring the possibilities of integrating aquaponics in drier climate cities, such as in the Near East and North Africa (FAO 2019).

The overall reduced ecological footprint of the cities practicing urban agriculture leads to the discussion of its positive influence on one of the dominant topics discussed today, climate change. There are many challenges faced by growing cities and continued urbanization throughout the world; however, climate change is recognized as one of the most pressing challenges (IPCC 2018). The most obvious way how urban farming can lessen the contribution of food production to climate change is by reducing carbon footprint. Increase in urban green spaces, including fruit trees and crops grown, will lead to increased carbon sequestration (Thornbush 2015).

By practicing food production close to the consumers, the greenhouse gas emissions associated with the food transportation, storage, and cooling will be reduced (Kulak et al. 2013). Numerous case studies exist that support the above, and while this has not been quantified on a global scale, urban agriculture can and should be promoted as both climate change mitigation and adaptation measures.

As a concluding remark on the benefits of urban agriculture, it is important to note that the extent of these benefits directly depends on the policies in practice, the type of urban farming, and the local environment. Summary of the benefits is presented in Table 1.

FIGURE 4 Rain barrels at urban farm, Towson University, MD, USA. (*Photo credit: N. Fath.*)

TABLE 1 Benefits of Urban Agriculture

Categories	Benefits
Social	• Builds and strengthens communities • Reconnection with food production • Renders an educational venue • Promotes social, economic, and environmental activism • Promotes social inclusion
Environmental	• Increases urban green spaces • Improves air quality • Reduces Urban Heat Island effect • Promotes biodiversity • Recycles organic waste • Increases stormwater infiltration • Reduces runoff • Reduces water usage and reuses wastewater • Reduces carbon footprint
Health	• Improves food security in urban food deserts • Improves access to fresh food and, thus, overall nutrition • Encourages physical activity • Cleaner air → improves overall health • Increases urban green spaces → improves psychological well-being and mental health

Challenges of Urban Farming

Securing Land for Food Production in Cities

One of the major obvious challenges and constraints is the lack of space in world cities to be used for growing food. On the other hand, there are many cities where plenty of underutilized or abandoned land is available. Particularly, in the cities that have experienced post-industrial decline and out-migration, such as in the city of Detroit, USA. While thousands of acres of land have been given over to unemployed workers for food growing in American cities of Detroit and New York (van Veenhuizen 2006), it is imperative to enhance access to this available vacant land by creating appropriate policies. These policies will need to enable sustainable urban agriculture and remove unnecessary obstacles that exist currently in securing land for food production in cities as agriculture traditionally refers to the rural areas; thus, urban agriculture currently has no institutional home (RUAF Foundation 2019).

Additional problem is that worldwide urban land value typically rises, and therefore urban farmers have to look for free land to use. Municipal land-use regulations do not guarantee long-term use of these plots, and frequently only temporary farms or gardens can be established (Steele 2017). Given the notable time and initial money and resources necessary to be invested in developing a productive urban farm, the temporary status may not appear to be an attractive venue to pursue as they are vulnerable to redevelopment.

A relatively new phenomenon called "guerrilla gardening" has been growing around the world over the last two or three decades, and it arose in part as a response to the difficulty of securing land for food production in the cities. The name suggests that the gardening is done without the securing legal rights to the land that is being used, thus making them vulnerable.

Health Risks

While food production in urban areas promotes some substantial benefits to human health (they were discussed above), it also may present some health risks to the participants and the nearby community. Specifically, if the gardens are not organic, farmers and consumers may be exposed to doses of fertilizers and pesticides that are used. This risk, however, is not pertinent only to urban agriculture, but to rural agriculture as well.

One of the most common concerns associated with urban farming is the high probability that urban air and soil may be contaminated with particulate matter, heavy metals, and others. These may result from heavy traffic congestion that is typical for cities, from industrial activities that tend to be concentrated in urban areas, and possibly, from waste dumping areas in or around the cities. Accumulation of these pollutants in air or soil may negatively impact health of the farmers and consumers; however, those can be reduced if adequate measures and regulations are put in place (van Veenhuizen and Danso 2007).

Remediation Costs

As mentioned in the section above, it is plausible that some land plots in urban areas have higher levels of contamination, and thus it is imperative that soil samples are analyzed prior to farming activities. If soil contamination is confirmed, then land requires remediation. Current remediation methods include soil removal, washing, or capping. They are costly and therefore are limiting factors in growth of outdoor urban agricultural systems (Wortman and Lovell 2013).

Eliminating contact between food grown and the ground via construction of raised beds is a cost-effective way to establish food production on sites where soil has been contaminated (see Figure 5).

Energy Usage

Certain types of urban agriculture, particularly, indoor activities such as vertical farming, greenhouses, hydroponics, and aquaponics, do require substantial energy input for their proper operation, especially

FIGURE 5 Raised beds at urban farm, Towson University, MD, USA. (*Photo credit: N. Fath.*)

if these facilities are located in cold or water-scarce regions. In the world where 64.5% of electricity production still falls on fossil fuels, coal being the largest source, this is a concern (IEA 2019).

An aquaponics farm, which integrates hydroponics (soilless plant cultivation) with aquaculture (fish farming), where both components complement each other to generate two products at once, crops and fish, is a good practice to reduce agricultural water footprint as it recirculates water initially added to the system, but it has been proved that it requires a significant energy usage. While the food produced at the aquaponics farm in Baltimore, Maryland, USA did not contain antibiotics, synthetic pesticides, or chemical fertilizers, the energy use was remarkably high, particularly during the winter months. When input costs were compared to market prices on tilapia, it demonstrated a net economic loss (Love et al. 2015). High energy costs forced the shutdown of a hydroponics farm in Buffalo, New York and its relocation to southwest Texas (Santo et al. 2016).

Alternative types of energy should be considered where applicable, in particular, in respect to the construction of new greenhouse systems or retrofitting conventional greenhouses. Solar-powered greenhouse systems are more popular in Europe and can achieve up to 70% of energy saving (Taki et al. 2018).

Summary of the challenges and risks are presented in Table 2.

TABLE 2 Challenges of Urban Agriculture

Categories	Problems/Limitations
Social	• If air and soil contamination → high remediation costs • Unavailable land due to other uses or unsuitable • Expensive or inaccessible land • Lack of zoning regulations and policies, which include urban farming as part of future urban development • Increased fresh produce consumption does not notably lessen food insecurity • Increased fresh produce consumption does not notably improve diets overall
Environmental	• Some instances require high energy input • Likelihood of air and soil being contaminated • If in drier climates, then increased water demand (drinking water used for irrigation)
Health	• Potential health risks to farmers and consumers from soil contaminants and air pollutants if adequate preventative remediation measures not taken

Future Projections

Urban agriculture appears to be growing in popularity around the world, although the goals behind their establishment range from aesthetics to communal activity, to actual crop production in low-income "food deserts". Assessment of feasibility and sustainability or these food-producing practices should be further continued. It is implied that the land available for urban agriculture is limited given all the traditional land-uses in cities. Nevertheless, the productivity of that land can be high and it can achieve high yields, but currently material and labor resources are used inefficiently (McDougall et al. 2019).

It seems unlikely that urban agriculture can or will replace rural agriculture, as the majority of food calories in the form of cereal grains will continue to be produced outside the cities. However, urban agriculture offers an important supplement, particularly to food impoverished areas, it promotes a strong sense of community in terms of caring and investing in local place and educates people on the cycles of nature and where food comes from. With the growth of the industrial food production system in the mid-20th century in the United States, a huge disconnect arose between farming and most people. This gap can be breached by promoting farming practices in cities.

The scale of food production in cities can vary from small community gardens to large urban farms, such as the one that is currently under construction in Paris and is scheduled to open in 2020. It is projected to be the world's largest urban rooftop farm, which will also incorporate closed-water hydroponic vertical system spanning overall ~14,000 m^2 (150,695 ft^2). They are expecting to cultivate more than 30 different plant species using entirely organic methods, with fruit and vegetable daily production amounting to ~1,000 kg in the high season (*The Guardian* 2019). This could become a globally acknowledged model for sustainable urban food production.

References

Alaimo, K., Packnett, E., Miles, R.A., & Kruger, D.J. 2008. Fruit and vegetable intake among urban community gardeners. *Journal of Nutrition Education and Behavior*, 40(2), 94–101.

Armstrong, D. 2000. A survey of community gardens in upstate New York: Implications for health promotion and community development. *Health and Place*, 6, 319–327.

Austin, D. 2019. Hazen S. Pingree Monument. HistoricDetroit.org. https://historicdetroit.org/buildings/hazen-s-pingree-monument

Badami, M.G. & Ramankutty, N. 2015. Urban agriculture and food security: A critique based on an assessment of urban land constraints. *Global Food Security*, 4, 8–15. doi:10.1016/j.gfs.2014.10.003

Berry, W. 2010. "The pleasures of eating." In *Food* edited by Brooke Rollins & Lee Bauknight, 21–28. Southlake: Fountainhead Press.

Brown, K.H. & Jameton, A.L. 2000. Public health implications of urban agriculture. *Journal of Public Health Policy*, 21(1), 20–39. doi:10.2307/3343472

Clinton, N. et al. 2018. A global geospatial ecosystem services estimate of urban agriculture. *Earth's Future*, 6, 40–60. doi:10.1002/2017EF000536

Corlett, J.L., Dean, E.A., & Grivetti, L.E. 2003. Hmong gardens: Botanical diversity in an urban setting. *Economic Botany*, 57(3), 365–379.

Deelstra, T. & Girardet, H. 2000. Urban Agriculture and Sustainable Cities. *Growing Cities, Growing Food: Urban Agriculture on the Policy Agenda*.

Drescher, A.W. 2001. The German Allotment Gardens – A Model for Poverty Alleviation and Food Security in Southern African Cities? Published in the *Proceedings of the Sub-Regional Expert Meeting on Urban Horticulture*, Stellenbosch, South Africa, Jan. 15–19. FAO/University of Stellenbosch. http://www.cityfarmer.org/germanAllot.html

FAO. 2019. Every Drop Counts. http://www.fao.org/fao-stories/article/en/c/1111580/

Guitart, D., Pickering, C., & Byrne, J. 2012. Past results and future directions in urban community gardens research. *Urban Forestry & Urban Greening*, 11(4), 364–373. doi:10.1016/j.ufug.2012.06.007

Hallsworth, A. & Wong, A. 2013. Urban gardening: A valuable activity, but…. *Journal of Agriculture, Food Systems, and Community Development*, 3(2), 11–14.

Haysom, G. 2009. Urban agriculture and food security. *The Sustainability Institute. Sustainable Neighbourhood Design Manual.* Chapter 8. https://pdfs.semanticscholar.org/9b05/a3ad15a9d1e-b57227e01e3387d0966888fec.pdf

Hodgson, K., Caton Campbell, M., & Bailkey, M. 2011. *Urban Agriculture: Growing Healthy, Sustainable Places*. Chicago, IL: American Planning Association Planning Advisory Service.

Holland, L. 2004. Diversity and connections in community gardens: A contribution to local sustainability. *Local Environment,* 9(3), 285–305. doi:10.1080/1354983042000219388

Holmer, R.J. & Drescher, A.W. 2005. In Christine Knie (ed.): Urban and Peri-Urban Developments – Structures, Processes and Solutions. Southeast Asian-German Summer School Program. 2005 in Cologne/Germany, Oct. 16–29: 149–155 Allotment Gardens of Cagayan de Oro: Their Contribution to Food Security and Urban Environmental Management.

Homem de Carvalho, J.L. 2001. Prove: Small Agricultural Production Virtualization Programme. *Urban Agriculture Magazine*, No 5. Appropriate Methods for Urban Agriculture. Leusden, RUAF.

IEA. 2019. Electricity Statistics. https://www.iea.org/statistics/electricity/

IPCC. 2018. Global Warming of 1.5°C. *Special Report.* https://www.ipcc.ch/sr15/

Jenerette, G.D., Harlan, S.L., Stefanov, W.L., & Martin, C.A. 2011. Ecosystem services and urban heat riskscape moderation: Water, green spaces, and social inequality in Phoenix, USA. *Ecological Applications*, 21(7):2637–2651.

Kortright, R. & Wakefield, S. 2011. Edible backyards: A qualitative study of household food growing and its contributions to food security. *Agriculture and Human Values*, 28(1), 39–53.

Kulak, M., Graves, A., & Chatterton, J. 2013. Reducing greenhouse gas emissions with urban agriculture: A life cycle assessment perspective. *Landscape Urban Plan*, 111, 68–78.

Litt, J.S., Soobader, M.-J., Turbin, M.S., Hale, J.W., Buchenau, M., & Marshall, J.A. 2011. The influence of social involvement, neighborhood aesthetics, and community garden participation on fruit and vegetable consumption. *American Journal of Public Health*, 101(8), 1466–1473.

Love, D.C., Uhl, M.S., & Genello, L. 2015. Energy and water use of a small-scale raft aquaponics system in Baltimore, Maryland, United States. *Aquacultural Engineering*, 68, 19–27. doi:10.1016/j.aquaeng.2015.07.003

McDougall, R., Kristiansen, P., & Rader, R. 2019. Small-scale urban agriculture results in high yields but requires judicious management of inputs to achieve sustainability. *PNAS*, 116(1), 129–134. doi:10.1073/pnas.1809707115

Middel, A. & Chhetri, N. 2014. City of Phoenix Cool Urban Spaces Project: Urban Forestry and Cool Roofs. Center for Integrated Solutions to Climate Challenges, Arizona State University. https://d3dqsm2futmewz.cloudfront.net/docs/dcdc/website/documents/NOAA_PHX_UrbanSpaces_Rep.pdf

Nutrition Digest. 2019. USDA Defines Food Deserts. American Nutrition Association. Vol. 28, No 2. http://americannutritionassociation.org/newsletter/usda-defines-food-deserts

RUAF Foundation. 2019. Urban Agriculture: What and Why? https://www.ruaf.org/urban-agriculture-what-and-why

Santo, R., Palmer, A., & Kim, B. 2016. Vacant Lots to Vibrant Plots: A Review of the Benefits and Limitations of Urban Agriculture. *CLF Report*. Johns Hopkins University. https://www.jhsph.edu/research/centers-and-institutes/johns-hopkins-center-for-a-livable-future/_pdf/research/clf_reports/urban-ag-literature-review.pdf

Schumm, L. 2018. America's Patriotic Victory Gardens: During Both World Wars, America's Agricultural Production Became a Powerful Military Tool. https://www.history.com/news/americas-patriotic-victory-gardens

Siegner, A., Sowerwine, J., & Acey, C. 2018. Does urban agriculture improve food security? Examining the nexus of food access and distribution of urban produced foods in the United States: A systematic review. *Sustainability*, 10(9), 2988. doi:10.3390/su10092988

Smith, V.M. & Harrington, J.A. 2014. Community food production as food security: Resource and market valuation in Madison, Wisconsin (USA). *Journal of Agriculture, Food Systems, and Community Development*, 4(2), 61–80.

Steele, K. 2017. Urban Farming: An Introduction to Urban Farming, from Types and Benefits to Strategies and Regulations. Vitalyst Health Foundation. www.vitalysthealth.org

Taki, M., Rohani, A., & Rahmati-Joneidabad, M. 2018. Solar thermal simulation and applications in greenhouse. *Information Processing in Agriculture*, 5(1), 83–113. doi:10.1016/j.inpa.2017.10.003

The Guardian. 2019. World's Largest Urban Farm to Open – on a Paris Rooftop. Aug. 13. https://www.theguardian.com/cities/2019/aug/13/worlds-largest-urban-farm-to-open-on-a-paris-rooftop

Thornbush, M. 2015. Urban agriculture in the transition to low carbon cities through urban greening. *AIMS Environmental Science*, 2(3), 852–867.

van Veenhuizen, R. & Danso, G. 2007. Profitability and Sustainability of Urban and Peri-Urban Agriculture. *Agricultural Management, Marketing and Finance Occasional Paper*. Food and Agriculture Organization of the United Nations. http://www.fao.org/3/a-a1471e.pdf

van Veenhuizen, R. 2006. Cities Farming for the Future, Urban Agriculture for Green and Productive Cities. International Institute of Rural Reconstruction and ETC Urban Agriculture. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.124.4555&rep=rep1&type=pdf

Voicu, I. & Been, V. 2008. The effect of community gardens on neighboring property values. *Real Estate Economics*, 36(2), 241–283.

Wakefield, S., Yeudall, F., Taron, C., Reynolds, J., & Skinner, A.L. 2007. Growing urban health: Community gardening in South-East Toronto. *Health Promotion International*, 22(2), 92–100.

Wolf, K.L. & Robbins, A.S. 2015. Metro nature, environmental health, and economic value. *Environmental Health Perspectives*, 123(5), 390–398.

Wortman, S.E. & Lovell, S.T. 2013. Environmental challenges threatening the growth of urban agriculture in the United States. *Journal of Environmental Quality*, 42(5), 1283–1294.

Zezza, A. & Tasciotti, L. 2008. *Does Urban Agriculture Enhance Dietary Diversity? Empirical Evidence from a Sample of Developing Countries.* Rome: FAO.

III

CSS: Case Studies of Environmental Management

20

Cell Tower Procurement: Public School Placement

Joshua Steinfeld

Introduction

Personal wireless service providers, independently and through use of brokers, have installed cell towers at or close to public schools in the United States. The open space surrounding schools, typically baseball and football fields, allows for optimal transmission of radio frequency (RF) waves between towers. Furthermore, public school districts, in comparison with private enterprises, have been more easily won over in the cell tower proposal process. Surges in experimental research regarding health hazards to RF emission, especially the apparent increased susceptibility of children to RF radiation, have sparked controversy over the exact locations of cell towers.

The purpose of this manuscript is to present an argument opposing the placement of cell towers at public schools. First, scientific studies related to animals and humans are provided to show that RF waves may be harmful to humans, and there is concern over electromagnetism. Second, voluntary initiatives and cell tower proposal processes are discussed, highlighting the Rockville, Maryland, community's precedential victory in opposition to the placement of a cell tower at Wootton High School. Third, ethics and public policy considerations address the need for streamlining nationwide community efforts by amending the Telecommunications Act of 1996 to disallow the placement of cell towers near public schools. Fourth, areas for further research are presented, which include the compilation of data sets, tracking of exposed students at school, and new theory to address causality issues related to competing risks. Finally, the chapter concludes by providing commentary regarding invisible risks.

Impact and Study of RF Hazards

Recent Trends

Concerns regarding the safety of cell towers at school have risen across the country. In 2010, Vista del Monte Elementary School in Palm Springs, CA, was pegged as having a reputation for being a cancer school. In 2005, a cell tower was erected on campus. Since then, 12 people have been diagnosed with cancer, affecting those who worked closest to where the cell tower was installed, where the field-strength readings were highest.[1]

From 1975 to 2000, childhood cancer rates had increased dramatically by a rate of 32%. Some of the most severe and deadly cancers such as acute lymphocytic leukemia, brain, kidney, and bone cancer also increased considerably.[2] In 2004, there were 36 million prescriptions of sleeping medications. As of 2009, 56 million prescriptions were outstanding, a whopping 56% increase. In 2004, the number of residents using cell phones was 109 million. By 2009, the number was up to 271 million.[3] Public health officials and environmental experts alike have been searching for environmental stimuli that may be contributing to the increased childhood cancer rates and sleep deprivation. The placement of cell towers at schools has been an area of focus (Figures 1 and 2). This Florida community elementary school also has a narrow-band transmission device installed on the tower (Figure 3). Installation of narrow-band cell towers is generally more restrictive than wide-band cell towers because of the narrow band's compact, piercing wavelength. Electromagnetic stimuli emitted from cell devices have become an area of intense research interest.

Background Research

Just to get an idea of the strength of an electromagnetic force emitted from a cell tower in comparison with a more well-known object, cell phone technology operates on frequencies up to 3 gigahertz (GHz), and a microwave oven cooks food at 2,450 megahertz (MHz).[4] Three GHz is equal to 300,000 MHz! Coulomb's law states that the force of an electromagnetic field is proportional to the magnitude of the charge and inversely proportional to the square of the separation. From Coulomb's law, we can derive the Inverse law, which suggests that holding the magnitude of the charge constant, the electromagnetic force emitted on a subject increases exponentially as the subject moves closer to the source.[5] Using

FIGURE 1 Public elementary school.

FIGURE 2 Wide-band cell tower at elementary school.

FIGURE 3 Narrow-band cell tower at elementary school.

Ampere's law, which combines the magnetostatic equation for determining the magnitude of a magnetic field with Stokes' theorem dealing with surface area, it is possible to determine the strength of an electromagnetic force through a closed path that may be tangent or indirectly exposed to the source.[5,6] Coulomb's law and the resulting derivation of the Inverse law demonstrate key findings in the discussion of cell towers. The farther away the cell tower site is from the subject, the lower the field strength absorbed. Furthermore, due to the exponential nature of the Inverse law, being close means being really close. Ampere's law is especially helpful in determining indirect exposure to force strength in cities where waves regularly bounce off other buildings.

A given material is composed of atoms. Each atom consists of electrons orbiting a central positive nucleus. Electrons also spin around their own axis. The orbital array of activity occurring between magnetic forces of protons and electrons in atoms results in an organized disarray of unpredictability. This unpredictable orbital disarray is the normal, unaltered state of the atom. When an external RF field is applied, the bombardment of electrons (from the source) stimulates host atom movement changes (in the subject), and the atom has a magnetic moment.[7] The orbital path of the electrons is brought into a slight sense of organization, throwing normal behavior out of whack. It is the electron stimulus emitted from cell towers that has been of much focus in experimental studies.

Radio Frequency Research on Animals and Humans

Scientists and researchers have proceeded cautiously regarding the use of human subjects in testing the effects of RF exposure. As a result, numerous testing on animals has been done to learn about the effects of electromagnetic radiation on living organisms. A 1997 study on mice demonstrated the effects of radiation on prenatal development and resulted in a progressive decrease in the number of newborns per dam, ending in irreversible infertility. In a subsequent 1999 study, mice exposed for just 24–72 hours to weak electromagnetic waves increased the activity of natural killer cells by 130%. Meanwhile, exposure to microwave irradiation had no effect on the activity of natural killer cells.[8] Nonetheless, microwave stimulus interfered with cell immunity of mice, increasing T-cell proliferation in response to stimulus.[9]

Studies on other living organisms have also been conducted. In Germany, behavioral abnormalities were observed in a herd of dairy cows that grazed near a cell tower for over 2 years, leading to reduction in milk yield and increased health problems.[10] In Russia, the effects of electromagnetic radiation on sea urchin embryos were tested. Only sea urchins with preexisting weakened viability were impacted by the electromagnetic radiation, in which case the electromagnetic radiation stimulated the onset of early development of embryos.[11]

While much more rare than electromagnetic testing on animals, some testing on humans has taken place. It was discovered that electromagnetic fields affect the central nervous system in humans because visual reaction time was prolonged and scores on short-term memory tests were lower in high-intensity exposure test groups.[12] Also, in a controlled study aimed to investigate the impact of low-force electromagnetic fields on healthy humans, human subjects were exposed to a 900-MHz electromagnetic field and intermittently pulsed with 217 MHz. It was determined that low-force fields have no effect on nocturnal hormone secretion under polysomnographic controls. However, cortisol production increased, which is transient by classification, indicating the organism (human subject in this case) adapted to the stimulus.[13] It was unclear if any mental impediments or genetic responses may have taken place in addition to the increase in cortisol production, but it was certain that the human subjects endured a cellular response. Nonetheless, current research indicates that genes are at risk, even at low-force electromagnetic fields.

Genes that ward off cancer and other illness may be inhibited when a cell receives stimuli from the environment. The National Institute of Environmental Health Sciences has been using genomic techniques to determine the behavior of promoter-proximal paused polymerase (Pol II) with and without environmental stimuli.[14] Pol II is known to have a role in fighting disease. Transcriptional responses to environmental stimuli can cause alterations in Pol II distribution, gene expression, and epigenetic chromatin signatures, leading to transcription dysregulation that can cause etiology of cancers.[15] Additionally, recent work has revealed that signal-response pathways are loaded with Pol II prior to final gene activation, further enhancing the opportunity for cellular changes to take place as a result of harmful environmental stimuli.[16,17] Molecules transported from environmental stimuli can inhibit the signal-response pathways' ability to pause release of Pol II.[18] Pol II pausing is necessary in providing an accessible chromatin architecture for gene promoters that inevitably fight disease.[19,20] It has been determined that Pol II pausing facilities' precise control and coordination of genes is a crucial regulatory step in rate-limiting the expression of DNA damage responsive genes.[21]

The extent to which scientists and doctors alike understand the impact of RF waves on humans is unquestionable in its fineness. For example, RF waves are routinely manipulated in clinical medicine to achieve exact thermal dosimetry and thermal pattern poisoning of tumors. The important nuance to remember is that the beneficial uses of RF waves in medicine are based on the destructive qualities of the high-energy RF waves.[22]

Voluntary Initiative

Students Against Cell Towers

In 1963, a group of citizens became activists in opposition to the proposal of a Con Edison power plant on Storm King Mountain in New York.[23] If the Storm King success story gave life to an entire environmental movement, then the success of Students Against Cell Towers (SACT) in opposing a cell tower at Wootton High School in Rockville, Maryland, solidified the environmental movement's existence by the hundreds of cell tower opposition advocacy groups that spawned across the nation since the SACT community voluntary initiative spanned from 2003 to 2005. Although SACT is no longer an active organization, its over 100 former members carry with them knowledge of the vital considerations regarding placement of cell towers at schools.

In 2000, Cingular Wireless began an aggressive campaign to install cell towers at public schools in Montgomery County, Maryland. The high schools were targeted first, perhaps because younger children are known to be especially susceptible to the radiation emitted from cell towers, although we do not know for sure what Cingular's strategy was. Additionally, Cingular Wireless first targeted public high schools located in communities with relative economic disadvantages and therefore less likelihood of organizing against a cell tower proposal. Cingular installed cell towers at Wheaton High School, Sherwood High School, and Kennedy High School. The students in these school districts come from families with median household incomes of $55,562, $57,260, and $60,296, respectively. It was not until 2003 that Cingular approached the wealthier Wootton and Walter Johnson High Schools, with median household incomes of $74,655 and $77,568, respectively.[24] Cingular never approached Whitman High School and Churchill High School with cell tower proposals; families of students in these schools have median household incomes of $113,788 and $140,222, respectively. Cingular Wireless' schematic timeline for public school cell tower proposals aimed to first test the will of communities who were economically and educationally disadvantaged before targeting wealthier, resourceful communities. Eventually, the residents of the Wootton and Walter Johnson school districts successfully rejected the cell towers due to the organization of a community voluntary initiative.

Cingular Wireless hired an experienced attorney to implement objectives related to the Wootton cell tower proposal, the same attorney who had previously handled the installment of cell towers at ten other public schools in Montgomery County. With little to no resistance coming from the Montgomery County public schools where cell tower installation was already in place, there was no need for Cingular to hire more than one person, an attorney rightfully so, to execute public school cell tower proposals in Montgomery County. An emphasis is placed on Cingular's need for just a one-man proposal team to demonstrate two things. First, it may be surprising how little publicity and resistance cell tower proposals typically receive, considering how big and visible they are. Second, Cingular's one-man show enabled Cingular to circumvent the proposal process. According to the Telecommunications Act of 1996, personal wireless service providers are required to notify in writing students, employees, and local residents of any proposal to erect a cell tower on public school property. Cingular's attorney had the ability to make sole judgment in his decision to refuse to notify the community regarding the plan to install a cell tower at Wootton High School.

In attempting to reject Cingular's proposal to erect a cell tower at Wootton High School, local residents formed a coalition called SACT. Despite any protections that may be offered by the Telecommunications Act of 1996, it clearly states that cell towers may not be rejected because of health concerns of nearby subjects. However, since Wootton High School is situated in a valley with Frost

Middle School perched up on the adjacent valley ridge, the RF waves emitted from the cell tower at Wootton would be passing through Frost Middle School, creating the allowance for a new interpretation of the act and a platform to remain steadfast. Next, it was discovered by SACT that the principal of Wootton High School had the exclusive authority to decide on whether to allow a 150-ft cell tower to be erected next to the football field at Wootton. SACT contacted local neighborhood associations, cluster school principals and administrators of Frost Middle School and Fallsmead Elementary School, Rockville City management, Montgomery County executive offices, parent–teacher association groups, and other perceived interest groups to spread word of the issue and provide scientific research related to health risks of RF exposure. Through discussion with local government and administration, SACT realized that the community had the right to a town hall meeting prior to the consideration of a cell tower proposal. Over 100 advocates holding greater than 1,400 petitions standing against the placement of a cell tower at Wootton High School arrived at the town hall meeting held in the Wootton library to greet Cingular's attorney and his science expert, with state representatives and media in attendance anxiously awaiting the confrontation. When Cingular's scientific expert utilized research no more recent than the 1960s, it became clear even to Cingular's attorney that SACT's research was plausible and that cell tower radiation may indeed be harmful to humans. For the benefit of Cingular's attorney, it was likely that he did not believe in his own company's stance; instead, his involvement was probably based on solidaristic group loyalties.[25]

An insight deserving mention was a key leadership tactic used by SACT to retrieve petition signatures. One of the exemplary practices of leadership is to inspire a shared vision.[26] Considering the perceived lack of concrete scientific data regarding human exposure to cell towers, one of the most effective ways to gain support for the petitions was to focus on the negative aesthetics of a 150-ft tower that would be visible from a Rockville resident's nearby home. Realizing the eyesore created by cell towers and the fact that opposition due to aesthetic concerns frequently arise, personal wireless service providers have begun camouflaging the cell towers (Figure 4). Figure 4 shows a cell tower with elaborate camouflage meant to make the cell tower look like a tree. After further survey of the surrounding area, it was discovered the camouflage was an attempt to help the cell tower blend in with a tree line overlooked by a nearby middle school (Figure 5). The canvas makes the cell tower more difficult to notice and provides personal wireless service providers with a prompted solution to aesthetic concerns.

A key ingredient behind this chapter's explicit focus on the placement of cell towers at public schools as opposed to all schools, in general, was illustrated by the Wootton principal's authority to decide for or against installation of the cell tower. If Wootton were a private school, the principal may not have been dictated the authority of decision maker or, in the event that authority was dictated, may not have had the wherewithal to acknowledge responsibility to local residents not affiliated with the school. SACT was a community effort predominately fueled by advocates not directly associated with Wootton High School. The idea that public school principals and all public school administrators, in general, are public servants lends additional support for the case against cell towers at public schools in particular. However, the stance against cell towers at public schools applies to all schools because of the RF-related health hazards.

Discussion of Cell Tower Placement Process

The Telecommunications Act of 1934 established the Federal Communications Commission (FCC) as the regulatory authority over communications activities in the United States. Because digital cell phones were not available to the public until 1988, controversy over the placement of cell towers is a new phenomenon. It was not until the Energy Policy Act of 1992 and the Telecommunications Act of 1996 when the federal government realized the issues of RF signal strength and tower placement, respectively. In 2003, SACT became one of the first nongovernmental organizations to address hazards related to cell

FIGURE 4 RF tower with elaborate camouflage.

FIGURE 5 The tallest tree in this image is the cell tower.

tower placement near public institutions. However, a number of transformations in the telecommunications industry have taken place since the Telecommunications Act of 1934, which have shaped the current regulatory environment. The years 1945 and 1952 marked the first major oversight by the FCC on over-the-air television, regulating the spectrum allocations and color standards. In 1968, telecommunications service providers were authorized by the FCC to attach equipment to preexisting above-the-ground electrical lines. And, in 1992, the FCC ruled to let the market decide the appropriate standards for digital cell phones and related equipment.[27]

The political actors typically involved in the process of determining the placement of cell towers complicate the ability of community voluntary initiatives to succeed in opposing the placement of cell towers. Achieved by a 1999 amendment to the Telecommunications Act of 1934, the local government has authority over state and federal governments on the issue of tower placement.[28] However, the amendment to the Telecommunications Act of 1934 is pursuant to the Telecommunications Act of 1996 specified requirement of the federal government to assist licensees' pursuit of preferred sites.[29] Federal involvement in the tower placement process results in streamlining of policy action. Streamlining results in the shrinking of the policy window for community advocates, which reduces community advocates' opportunity to introduce their own policies.[30] Additionally, personal wireless service providers sometimes hire independent facilities siting companies who offer comprehensive tower placement services, from lobbying of local government and communities to addressing zoning regulatory concerns. Leasing of sites and fulfillment of regulatory and registration requirements may also be taken care of by facilities siting companies (Figures 6–8). "The local zoning authorities should therefore be aware that a facilities siting company may not be seeking the sites that are of most interest to particular Commission licensees [personal wireless service providers], but rather seek general sites on highly elevated locations in the hopes of leasing the sites, in turn, to Commission licensees."[31] It would be intuitive to reason that the existence of broker special interests in the placement of cell towers would be an additional obstacle to voluntary initiatives striving to oppose cell towers. However, the contrary is sometimes true. Policy brokers, such as independent facilities siting companies, are interested in maintaining a sustainable level of conflict in order for services being offered to remain in high demand.[32] Lingering, yet not overpowering, community opposition is welcomed by facilities siting companies. To combat the influence of the federal government and policy brokers in the tower placement process, community organizations need to adopt a policy of political efficacy, arising from political participation as a means to exert influence.[33] Political efficacy is especially important when it comes to the upholding of FCC guidelines on tower height and field strength. According to the Energy Policy Act of 1992, electric and magnetic field strength must be made public.[34] Yet, according to the Code of Federal Regulations pertaining to personal communications services, height-above average-terrain (HAAT) and field-strength guidelines may be waived if all parties involved agree.[35] If the community does not involve local government, the community will have no voice on the issue, and the already-flaccid federal guidelines will leave school children and staff insurmountably exposed.

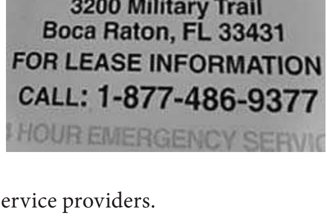


FIGURE 6 Brokers lease to wireless service providers.

FIGURE 7 Warning of RF emission.

FIGURE 8 Record keeping by the FCC.

Procurement and Contracting

The cell tower placement process as it involves local, state, and federal governance can be a separate, yet parallel set of activities to procurement and contracting processes. While law and statutes specific to cell tower placement may be directed at issues of allowability, the procurement and contracting functions deal specifically with the actual purchase in terms of cell tower specifications, features, pricing, contract terms and conditions between the public and private entity, solicitation for prospective vendors, competition for contract award, and contract administration, so that the private entities fulfill its obligations for the duration of the contract, among other procurement activities. In order to execute these functions, local government procurement units, which are typically housed in a department of procurement, finance, or budgeting, engage a contracting process involving pre-solicitation, solicitation, proposal evaluation, contract award, contract administration, and contract close-out.[36]

The pre-solicitation process involves procurement planning and communication with potential vendors who aim to install cell towers and provide cellular service to the area. Here, the government engages a process of discovery in which it learns about products and services and may hold formal or informal discussions with vendors. These discussions may involve policy matters, issues of feasibility for placement or service provision, the technology, and initial concerns regarding contract terms and conditions that may eventually follow. Generalities regarding the process for government's selection of the vendor for actual cell tower placement may also be discussed. These discussions may take place between a procurement agent and company representative over the phone, an in-person meeting, in writing, or at a trade show or conference. The pre-solicitation phase also enables government to prepare its formal solicitation document according to vendor concerns and considerations in conjunction with its own needs and priorities.

Oftentimes, wireless service providers may initially contact government to place cell towers on public properties. However, once government decides that it will entertain the possibility of placing a cell tower on public property, then it typically must engage a competitive process that allows the various companies an opportunity to compete for the contract. At the next phase, solicitation, the government may issue a request for proposals (RFP) or an invitation to bid (ITB). The RFP seeks proposals from vendors and provides them an opportunity to win the contract by setting forth a product and service to government, including various factors such as price, specifications, and commitments. Meanwhile, an ITB is a more rigid form of solicitation, in which the government sets forth its required criteria and the vendors simply respond regarding how they would perform to contract.

The justification behind having vendors compete for the contract award is so that government can be in a position to choose from the best product and service, i.e. the vendor who will commit the greatest amount of revenues at the most favorable contract terms and conditions for government and its community stakeholders. When vendors are required to compete for the contract award, then they will aim to increase the rent that they will commit to pay to government, or the public entity, for use of the property, as well as bolster their services or offerings in order to be the more attractive vendor. Furthermore, the vendors may make special accommodations in attempts to ameliorate public concerns that may have been vocalized during the pre-solicitation process such as issues of tower unsightliness, proximity, equipment dimensions, transmission power, servicing, and maintenance.

Once the RFPs or responses to the ITB have been submitted by the vendors to the procuring entity, then the procuring entity must evaluate these proposals to decide which vendor will be awarded with the contract. In this proposal evaluation phase, a team is formed to evaluate the documents submitted by the vendors. It is important to note that the proposal evaluation team members, per public procurement mandate, are usually required to consist of public personnel that were not involved in the pre-solicitation or solicitation phases in order to avoid conflict of interest. It may be the case that various personalized relationships were formed between the procurement agent and company representative during these early discussions, and it is important that the proposals are evaluated objectively according to the actual product and service to be delivered without consideration of prior personal relationships. Once the proposal evaluation team is in place, the evaluations are made based on factors of responsibility and responsiveness.

Responsibility refers to matters of vendor track record including their past performance on contracts with other public entities and the vendor's financial capitalization. The notions are that a vendor should be well capitalized in order to ensure that it has the resources to fulfill contractual obligations for the duration of the contract term and that past performance on contracts is an indicator of a company's ability to deliver its products and services as promised. Responsiveness deals with the extent to which a vendor's proposal reflects the needs as set forth by the public entity in the language of the RFP or ITB. A scoring methodology is then typically developed by the proposal evaluation team such as a weighted or other numerical scoring schemata, which assigns scores to various facets of the vendor proposals. The proposal that receives the highest score by the proposal evaluation team will be awarded with the contract.

Next, the contract award phase involves a debriefing process so that the winning and losing bidders are notified of the proposal evaluation team's selection. In this phase, any deviations between the vendor's proposal submitted and the RFP or ITB are addressed and a negotiation process may ensue between the selected vendor and procuring entity regarding actual contract terms and conditions. For example, some of the requirements set forth by the public entity in the RFP or ITB for the cell tower may be questioned by the vendor. Since the vendor has expertise regarding its own products and services, including the technology and equipment, it may be that the public entity is requiring features or specifications of the cell tower that may not be reasonable or feasible according to the vendor. Oppositely, the vendor's proposal as evaluated may lack some aspects of responsibility or responsiveness that need to be addressed, which may include community concerns or interests regarding cell towers, and it may be that the public entity has the greater knowledge and awareness in this particular political area. Of course, any deviations between the vendor's proposal and the RFP or ITB may reflect modifications in price, i.e., the rent that the public entity will be charging. If there are aspects of the RFP or ITB that the vendor cannot guarantee or conform to, then the public entity may offset those deficiencies by charging higher rent or setting forth other collateral commitments.

The first three phases of contracting involve pre-award processes, while contract administration and close-out are two phases that deal with post-award processes. Contract administration involves the actual delivery of products and services according to the contractual agreement known as performance to contract. This involves planning, preparation, and installation of the cell tower, and the administrative functions and procedures that will be involved such as oversight of construction, facilities monitoring, inspection, payment of rent, and other facets of interaction, communication, and coordination between the vendor and procuring entity. If the products and services delivered by the vendor are not consistent with the contract terms and conditions, which may include the manner that tower construction is taking place, the actual tower, or its servicing, then various modifications to the contract may be stipulated by the procuring entity. Meanwhile, if the procuring entity does not deliver its property site in the condition as promised, then the procuring entity may need to institute various change orders that could force it to reduce or avoid its rent payment obligations from the vendor for a specified time period or allow the vendor to make changes to its products or services. Thus, various change clauses will be written into the contract during negotiations in the event that unintended scenarios or problems arise during contract administration. Of course, it is difficult if not impossible to imagine and surmise all the potential scenarios and obstacles that may surface for a given procurement. If the vendor fails to adhere to the contract terms and conditions or does not adequately respond to the procuring entity's inquiry into contract violations, then the contract may be terminated by the procuring entity.

The final procurement phase is contract close-out. This phase involves quality assurance that the vendor has delivered as promised by installing a cell tower according to the requirements set forth by the procuring entity in the contract, and that servicing and maintenance are being executed accordingly. Any accounts receivable or rent payments owed to the public entity are accounted for and contract evaluation takes place to ensure that all contractual obligations are satisfied. Contract close-out is also important because of the lessons learned. The procuring entity, as the procurement unit within the government's department of procurement, finance, or budgeting, likely has other potential property sites in the municipality that could be attractive to wireless service providers. Hence, the lessons learned from the procurement and contracting process on the current contract may be applied and incorporated by the procuring entity into future cell tower procurements.

Meanwhile, the contract close-out phase also entails evaluation of the vendor on their performance to the contract, which could subsequently serve as performance evaluation data for the same or other procuring entities as they evaluate any future proposals by the vendor for aspects of responsibility. Whereas the early years of cell tower placement on public school and other public properties involved much more devolved processes that were more inconsistently applied on a case by case basis, even within the same municipality, the prevalence and debate surrounding cell tower placement has partially shifted the public authority on cell tower placement from public schools and school boards, and those public units or divisions having operations on the actual property, to procurement departments acting as agents on behalf of public schools or other public units. As a result, lessons learned from cell tower procurements can be more aptly applied to future contracts in order to address previous contract failures that may involve public concern and outcry. Essentially, the cell tower procurement is managed by the same procuring entity within a given municipality. Therefore, it has the opportunity to develop policy, politics, and process expertise from experience similar to the private entities who repeatedly write cell tower proposals to government.

The procuring entity does face a myriad of challenges though. The procuring entity must comply with local, state, and federal procurement law and policy, in addition to any local, state, and federal law and policy governing cell tower placement specifically. Additionally, there must be due consideration given to voluntary initiatives and advocacy groups. At times, these laws and policies may be conflicting or difficult to streamline into procurement and contracting methods.

Ethics and Public Administration

The three elements of corporate social responsibility are market actions, externally mandated actions, and voluntary actions.[36] Personal wireless service providers fail to address all three elements of social responsibility when dealing with the placement of cell towers. First, personal wireless service providers have poorly responded to market actions in their use of policy brokers to ensnare the tower placement process. Second, the mandated actions of the FCC related to HAAT and field-strength guidelines allow regulatory thresholds to be exceeded if no opposing voluntary organization is present at scheduled hearings. Third, voluntary actions that aim to avoid students' exposure to cell towers are not taken. In fact, the current trend is just the opposite, in which schools are targeted because of the surrounding open space that allows for enhanced RF wave transmission (Figures 9 and 10).

The ability of voluntary initiatives to oppose the placement of cell towers at public schools relies upon the formation of a nucleus of zealous participants; charismatic leadership alone does not result in the

FIGURE 9 RF tower overlooking playing fields.

FIGURE 10 RF tower installed adjacent to football field.

type of rapid expansion of the coalition that is necessary to fight seasoned corporate interests such as personal wireless service providers.[37] In addition, the media may be needed to intervene in facilitating changes in public perceptions. The media can help provide a transition and a way for the public to digest new policy initiatives.[38] In some cases, even with high visibility, the strongest coalitions are unable to defeat polyarchal interest groups on a particular issue, be it the placement of a cell tower at yet another school. Success depends on forging relationships with government officials as much as administrative competence.[39] Because of the variations in local government across the nation, including inconsistencies in the law and process governing zoning and other enforcement departments, there is a need to adopt a customizable approach when attempting to oppose a cell tower.

The SACT's ability to reject a cell tower proposal at Wootton High School but failure to impact policy change nationally brings into question the federalist debate. Hamilton, Madison, and Jay desired a strong central authority in their staunch support of federal government and ratification of the U.S. Constitution.[40] Opposing the federalists, Patrick Henry led the antifederalist approach arguing for decentralization and states' rights.[41] The federal government's determination to simultaneously support personal wireless service providers' cell tower installation campaign at public schools while allowing local communities to decide for themselves is rooted in both federalist and antifederalist modes. Nonetheless, the federal government's dual role is authoritarian.[42]

The cry for federal regulation disallowing placement of cell towers at or near public schools involves regulatory policy making, which inevitably will indulge or deprive one specific interest or another.[43] The regulatory approach offered does not necessarily favor some sort of Weberian chain-of-command hierarchy originating from the top down[44] but rather prefers reactionary cultural movement that reflects Thelen and Mahoney's,[45] Hacker's,[46] and Sheingate's[47] ideas on institutional change according to evolving assumptions of administrative and technological environments. If the dominant approach were favored, the solution would be to charge personal wireless service providers higher rental fees to compensate for any associated health care costs that may result.[48,49] Other rational socioeconomic approaches would seek to place value on the cost of the loss of human life and monetary damages due to the terminal pain and suffering induced by cell towers. Unfortunately, the most commonly observed failure in public management and the one associated with the egalitarian approach is a lack of ability to exert authority.[50]

Areas for Further Research

A logical area for further research is to track and survey humans. Tracking students who attend schools with cell towers on premises is the best way to determine the danger of cell towers to students. Tracking students while in school and in years beyond may help determine whether cell towers are indeed harmful to children or humans in general. An effective data set for a future study would need to take into account differentiation between student exposure at elementary, middle, and high schools, considering that young children are known to be more susceptible to RF exposure. Additionally, the data would need to differentiate between students who had many instances of intense exposure, such as student-athletes on a sports field containing a cell tower, and students who had fewer instances of intense exposure. Previously, this type of study was not possible. It was only since 2002 that the personal wireless service providers started targeting schools. Today, cell towers are erected on school properties across the country. However, obtaining access to student records for data collection of this type of study invokes the support of government. The resources and sheer number of people who would need to be mobilized by such a study requires congressional backing. Furthermore, the scientific community has questioned the methods that would underwrite such a study. Areas of inquiry that have been considered to be obstacles to a human study of this type are the competing risks when experimenting for causality. Oftentimes, it is difficult to determine which risk is the source of the illness. For example, did cancer clusters in the area form as a result of contaminated drinking water or exposure to a cell tower? New research indicates that margins of error in causality can be reduced. Building on time-dependent predictive accuracy measures,[51,52] the coefficient of the distribution of false positives among event-free subjects can be adjusted to reflect nearest neighbor (in our case, the highly exposed student-athletes on the sports fields containing cell towers) estimation of the distribution of input variables representing true positive incidents and length of exposure of competing risks in order to help determine causality.[53] The false-positive value is manipulated by a coefficient that is calculated from the estimate of true positive incidents in order to offset causality miscalculations stemming from overlapping incident rates of competing risks. Unconcerned citizens and proponents of cell tower placement on school property argue that cancer clusters related to cell towers have not emerged. However, it is impossible to predict what the health impacts may be over time. An evaluation technique is needed for cumulating, comparing, and contrasting varied results in order to establish an applied theory and framework.[54]

Conclusion

Just 2 years after the victory at Storm King Mountain paved the way for environmentalism, Olson's *The Logic of Collective Action* (1965) opened our eyes to the intuitiveness of equality and why the events of the 1960s were unlikely to ever recur.[55] Much different from the equal rights movement, the successful voluntary initiative at Wootton High School has not led to nationwide legislation and is embedded in an issue that is invisible. The inability to see the RF waves somehow precludes from our psyche the notion of harm. Invisible risks skew the indifference curve that guides our behavior in responding to risk. An indifference map is our collection of indifference curves and helps to shape our order of preferences.[56] The typical reaction to an invisible risk is a delay in response. After a 5-year pause, personal wireless service providers are once again proposing cell towers at public schools in Montgomery County, Maryland, this time at Whitman High School in Bethesda, Maryland.[57] Bethesda community advocates are citing Wootton's 2005 rejection of a cell tower as precedent in Montgomery County.[58] Currently, only the ten cell towers that were installed prior to the Wootton campaign exist in Montgomery County. Also, in the sense of a health-wise decision at Wootton, why have the ten previously erected towers not been taken down?

Cell towers should not be placed at schools. Scientific experimental research on animals and humans is conclusive that RF waves increase animal and human risk to cancer and other illnesses. Also, in the case of public schools, students have little to no choice in deciding whether or not to attend a particular school. To address personal wireless service providers' desire to bolster cell phone connectivity in communities, state and local governmental zoning boards should work together to designate specific areas where the placement of cell towers will be permitted. Assuming no cell tower is placed at or near a public school, the designated zones will enable the public to make their own decision regarding living or spending time near cell tower sites.

Community voluntary initiatives are the answer if the goal is to reject a cell tower proposal. Federal governments' involvement, not from the standpoint of assisting personal wireless service providers meet their objective of cell tower placement in a given community but from the standpoint of bringing together the nationwide advocacy groups and coalitions that have both successfully and unsuccessfully defended their schools from RF wave penetration, is the long-term solution to keeping schools safe from cell towers. Also, although the prospect of fair gamesmanship in the process of appeal against cell tower placement at public schools was not presented to be optimistic, consideration of bureaucratic red tape, or in this case, purposeful lack thereof, is essential to successfully implementing a strategy that leads to the rejection of a personal wireless service provider's cell tower proposal. In any case, policies that engage citizens in their own communities and ask them to do their own policy analysis are generally more preferable to those policies that do not.[59]

References

1. Milham, S. *Dirty Electricity, Electrification and the Diseases of Civilization*; Universe: New York, 2010.
2. The Stop Cancer before It Starts Campaign, Press Report, May 8, 2003, Cancer Prevention Coalition, available at http://preventcancer.com/publications/pdf/M803.htm (accessed September 2011).
3. LeBeau, C. *Insomnia, Fatigue, and Cell-Phone Towers*; Vital Health Publications: West Allis, WI, 2010.
4. Lai, H. Biological effects of radiofrequency radiation from wireless transmission towers. In *Cell Towers, Wireless Convenience? Or Environmental Hazard?* Levitt, B., Ed.; New Century Publishing: Markham, Canada, 2000; 65.
5. Schwartz, M. *Principles of Electrodynamics*; McGraw Hill: New York, 1972; 2–0, 14–49.
6. Edminister, J.; Nahri-Dekhordi, M. *Electromagnetics*, 3rd Ed.; McGraw Hill: New York, 2011; 174.
7. Sadiku, M. *Elements of Electromagnetics*; Oxford University Press: New York, 2010; 350.
8. Fesenko, E.; Novoselova, E.; Semiletova, N.; Aganova, T.; Sadovnikov, V. Stimulation of murine natural killer cells by weak electromagnetic waves in the centimeter range. *Biofizika* 1999, 44 (4), 73–41.
9. Fesenko, E.; Makar, V.; Novoselova, E.; Sadovnikov, V. Microwaves and cellular immunity. Effect of whole body microwave irradiation on tumor necrosis factor production in mouse cells. *Bioelectrochem. Bioenerg.* 1999, 49 (1), 29–35.
10. Loscher, W. Extraordinary behavior disorders in cows in proximity to transmission stations. *Der Praktische Tierarz* 1998, 79 (1), 437–444.
11. Galat, V.; Mezhevikina, L.; Zubin, M.; Lepikhov, K.; Khramov, R.; Chailakhian, L. Effect of millimeter waves on the early development of the mouse and sea urchin embryo. *Biofizika* 1999, 44 (1), 137–140.
12. Chiang, H.; Yao, G.; Fang, Q.; Wang, K.; Lu, D.; Zhov, Y. Health effects of environmental electromagnetic fields. *J. Bioelectr.* 1989, 8 (1), 127–131.
13. Mann, K.; Wagner, P.; Brunn, G.; Hassan, F.; Hiemke, C.; Roschke, J. Effects of pulsed high-frequency electromagnetic fields on the neuroendocrine system. *Neuroendocrinology* 1998, 67 (2), 139–144.
14. Nechaev, S.; Fargo, D.; dos Santos, G.; Liv, L.; Gao, Y.; Adelman, K. Global analysis of short RNA's reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila. *Science* 2010, 327 (5963), 335–338.
15. Adelman, K. Transcriptional Responses to the Environment Group, Chromatic Signatures and Gene Expression. Principal Investigator. National Institute of Environmental Health, available at http://niehs.nih.gov/research/atniehs/labs/lmc/tre/index.cfm (accessed September 2011).
16. Muse, G.; Gilchrist, D.; Nechaev, S.; Shah, R.; Parker, J.; Grissom, S.; Zeitlinger, J.; Adelman, K. RNA polymerase is poised for activation across genome. *Nat. Genet.* 2007, 39 (12), 1507–1511.

17. Zeitlinger, J.; Stark, A.; Kellis, M.; Hong, J.; Nechaev, S.; Adelman, K.; Levin, M.; Young, R. RNA polymerases stalling at developmental control genes in the Drosophila embryo. *Nat. Genet.* 2007, 39 (12), 1512–1516.
18. Boettiger, A.; Ralph, P.; Evans, S. Transcriptional regulation: Effects of promoter proximal pausing on speed, synchrony, and reliability. *Comput. Biol.* 2011, 7 (5), 1–14.
19. Gilchrist, D.; Nechaev, S.; Lee, C.; Gosh, S.; Collins, J.; Li, L.; Gilmour, D.; Adelman, K. NELF-mediated stalling of Pol II can enhance gene expression by blocking promoter-proximal nucleosome assembly. *Genes Dev.* 2008, 22 (14), 1921–1933.
20. Gilchrist, D.; dos Santos, G.; Fargo, D.; Xie, B.; Gao, Y.; Li, L.; Adelman, K. Pausing of RNA polymerase II disrupts DNA-specified nucleosome organization to enable precise gene regulation. *Cell* 2010, 143 (4), 540–541.
21. Adelman, K.; Kennedy, M.; Nechaev, S.; Gilchrist, D.; Muse, G.; Chineov, Y.; Rogatsky, I. Immediate mediators of the inflammatory response are poised for gene activation through RNA polymerase stalling. *Proc. Nat. Acad. Sci.* 2009, 106 (43), 18207–18212.
22. Kasevich, R. Brief overview of the effects of electromagnetic fields on the environment. In *Cell Towers, Wireless Convenience? Or Environmental Hazard?* Levitt, B., Ed.; New Century Publishing: Markham, Canada, 2000; 170.
23. Anzevino, J. Preserving scenic and historic sites: The dilemma of siting cell towers and antennas in sensitive areas. In *Cell Towers, Wireless Convenience? Or Environmental Hazard?* Levitt, B., Ed.; New Century Publishing: Markham, Canada, 2000; 169.
24. Public School Review. Public Elementary, Middle, and High Schools, Data, 2011, available at http://www.publicschoolreview.com (accessed September 2011).
25. Dunleavy, P. *Democracy, Bureaucracy, and Public Choice*; Prentice Hall: London, 1991; 28.
26. Kouzes, J.; Posner, B. *The Leadership Challenge*, 4th Ed.; Jossey-Bass: San Francisco, 2007; 16–18, 142–143.
27. Ismail, S. Transformative Choices: A Review of 70 Years of FCC Decisions, FCC Staff Working Paper 1; Federal Communications Commission: Washington, DC, 2010; 1, 18, 19.
28. A Bill to Amend the Communications Act of 1934, 106th Congress, 1st Session; U.S. Senate: Washington, DC, 1999; S.1538.
29. New Wireless Tower Siting Policies, Fact Sheet #1; Wireless Telecommunications Bureau, Federal Communications Commission: Washington, DC, 1996; 1–2.
30. Kingdon, J. *Agendas, Alternatives, and Public Policies*, 2nd Ed., Longman: New York, 2003; 165.
31. National Wireless Tower Siting Policies, Fact Sheet #2; Wireless Telecommunications Bureau, Federal Communications Commission: Washington, DC, 1996; 8.
32. Sabatier, P. An advocacy coalition framework of policy change and the role of policy-oriented learning. *Policy Sci.* 1988, 21 (1), 141.
33. Sabatier, P. Political science and public policy. *Political Sci. Polit.* 1991, 24 (2), 145.
34. Energy Policy Act of 1992, Section 2118, Electric and Magnetic Fields Research and Public Information Dissemination (RAPID), Public Law 102–486; 42 U.S.C. 13478.
35. Code of Federal Regulations, Title 47, Calculation of Height above Average Terrain and Field Strength Limits; Personal Communications Services, 2010; 24.53, 24.236.
36. Thai, K.V. *Introduction to Public Procurement*. Herndon: Virginia, 2013.
37. Steiner, J.; Steiner, G. *Business, Government, and Society*; McGraw Hill: New York, 2012; 131.
38. Downs, A. *Inside Bureaucracy*; Little, Brown, and Company: Boston, 1967; 7, 9.
39. Sabatier, P.; Mazmanian, D. The implementation of public policy: A framework of analysis. *Policy Stud.* 1980, 8 (4), Special no. 2, 550.
40. Nalbandian, J. Reflections of a "pracademic" on the logic of politics and administrations. *Pub. Admin. Rev.* 1994, 54 (6), 531.
41. Kesler, C.; Rossiter, C. *The Federalist Papers*; New America Library: New York, 1999, vii–xii.

42. Henry, P. Legitimate government. In *The Anti-Federalist Papers*. Borden, M., Ed.; Michigan State University Press: East Lansing, MI; Par. 5.
43. Steinfeld, J. American authoritarian democracy: Vietnam War, Kosovo War, and Overseas Contingency Operation. In *Proceedings of the 32nd Annual Southeastern Conference for Public Administration*, New Orleans, LA, Sept. 21–24, 2011.
44. Miller, H. Weber's action theory and Lowi's policy types in formulation, enactment, and implementation. *Policy Stud*. 1990, 18 (4), 895.
45. Weber, M. *The Theory of Social and Economic Organization*; The Free Press: New York, 1947.
46. Thelen, K.; Mahoney, J. *Explaining Institutional Change: Ambiguity, Agency, and Power*; Cambridge University Press: Cambridge, England, 2010.
47. Hacker, J. Policy drift: The hidden politics of welfare state retrenchment. In *Beyond Continuity: Institutional Change in Advanced Political Economies*; Streeck, W., Thelen, K., Eds.; Oxford University Press: Oxford, 2005; 17.
48. Sheingate, A. Rethinking rules: Creativity in the House of Representatives. In *Explaining Institutional Change: Ambiguity, Agency, and Power*; Thelen, K., Mahoney, K., Eds.; Cambridge University Press: Cambridge, England, 2010.
49. Weber, M. *Economy and Society, an Outline of Interpretive Sociology*; Roth, G., Wittich, C., Eds.; University of California Press: Berkeley, CA, 1978; Vol. 1.
50. Weber, M. *Economy and Society, an Outline of Interpretive Sociology*; Roth, G., Wittich, C., Eds.; University of California Press: Berkeley, CA, 1978; Vol. 2.
51. Hood, C. *The Art of the State: Culture, Rhetoric, and Public Management*; Oxford University Press: Oxford, 1998; 40.
52. Heagerty, P.; Lumley, T.; Pepe, M. Time-dependent ROC curves for censored survival data and diagnostic marker. *Biometrics* 2000, 56 (1), 337–344.
53. Heagerty P.; Zheng, Y. Survival mode predictive accuracy and ROC curves. *Biometrics* 2005, 61 (1), 921.
54. Saha, P.; Heagerty, P. Time-dependent predictive accuracy in the presence of competing risks. *Biometrics* 2010, 66 (4), 999–1011.
55. Lowi, T. American business, public policy, case-studies, and political theory. *World Polit*. 1964, 16 (4), 688.
56. Hirschman, A. *Shifting Involvements: Private Interests and Public Affairs*; Blackwell: Oxford, 1985; 79.
57. Anderton, C.; Carter, J. *Principles of Conflict Economics*; Cambridge University Press: New York, 2009; 29.
58. Barnes, A. Cell Tower Proposed for Walt Whitman High School; USA 9 News: Bethesda, MD, 2010, available at http://wusa9.com/news/local/stay.aspx?stayid=95703&catid=158 (accessed September 2011).
59. Cropper, M. No cell tower at Whitman. Rockville Gazette, 2011, available at http://ww2.gazette.net/stories/03102010/montlet175220_32598.php (accessed September 2011).
60. Clarke, J.; Ingram, H. A founder: Aaron Wildavsky and the study of public policy. *Policy Stud*. 2010, 38 (3), 574.

21

Community-Based Monitoring: Ngarenanyuki, Tanzania

Aiwerasia V.F. Ngowi, Larama M.B. Rongo, and Thomas J. Mbise

Introduction

Community-based monitoring was initiated in Ngarenanyuki, Tanzania, to study the impacts of pesticides on health and the environment. This entry is organized and divided into the following main sections: "Introduction," "Methodology," "Results and Discussion," and "Conclusion." Illustrations are included in the "Introduction," "Methodology," and "Results and Discussion" sections. The main goal of the study was to reduce exposure to pesticides among the farmers in Ngarenanyuki by training the farmers on health impacts related to exposure to pesticides, how to monitor such impacts, and how to reduce the risks.

Incidences of poisoning from pesticides are estimated to be highest in developing countries, despite the higher use of pesticides in developed countries.[1,2] The monitoring of pesticides and their health impacts on farmers and the public in general, which is normally performed by qualified researchers, is not sufficiently practiced in many developing countries, owing to financial constraints and to competing research interests. It therefore makes more sense to empower communities themselves to monitor the impact of pesticides and to take decisions that might reduce the risks to themselves and to their environment. Community-based monitoring of the impacts from pesticides enable those communities to determine whether or not the chemicals they are already exposed to, or might be exposed to, present any sort of hazard to their health and a potential threat to their environment.

Community pesticide-surveillance methods have been successfully used in the Asia Pacific[3] and could therefore be considered appropriate in Tanzania and other Southern African countries, both for establishing better data on the extent of pesticide poisoning and to raise awareness among farmers themselves. Using their own system of observation and evaluation of risks, for example, Malaysian plantation workers have developed the Community Pesticides Action Monitoring (CPAM) approach, in which they succeeded in documenting the health effects of airborne pesticides, identifying paraquat in particular as a major problem.[4]

They then proceeded to take action to prevent further exposure of plantation workers to paraquat. Communities in Kasargod District, Kerala (India), after investigation, monitoring, and documentation, using a CPAM approach, identified endosulfan as the major pesticide causing health and environmental problems and subsequently called for the ban on endosulfan to prevent further exposure and damage to the communities.[5]

Incorrect pesticide handling and management is thus known to be unsafe to both human and environmental health and jeopardizes biodiversity.[6] It is also further evident that most rural communities in Tanzania depend on farming and agribusinesses to earn their living. However, traditional agricultural production in the country has come under continuous pressure from globalization and other market forces, with the result that high-input agriculture has come to play an increasingly major role in the economies of rural communities. Although the use of pesticides in combating pests and diseases is widely encouraged among the farmers in these communities to promote production, less emphasis has been placed on safety practices and the proper handling and management of materials.

Ngarenanyuki Ward gives an example of a community in Tanzania where the majority of vegetable farmers believed that, without pesticide use, crop production would have been impossible. In a previous study carried out by the Work and Health in Southern Africa (WAHSA) team (unpublished), it was found that mixing three to five different types of pesticides in a single spray mix was a common practice in Ngarenanyuki, and that farmers did not understand what was written on the label or the meaning of the colors on the containers. They simply applied pesticides because a neighbor had applied them, and not because they had identified a particular pest problem. The farmers were also found to have mixed pesticides without following the doses recommended on the label, sometimes doubling or trebling the dosage regardless.

Retail outlets for pesticides in Ngarenanyuki were also found not to have been registered with the regulatory authority and therefore appeared to be selling pesticides illegally. It was further noted that, because the shop owners tended to repackage or dispense pesticides in other containers, they were observed on occasion to be left with empty pesticide containers, which they apparently destroyed by burning them at the marketplace. Farmers were also observed in some instances to have stocked substantial amounts of pesticides to cater for the whole year, owing to perceived shortages in the local village and the reported distances they had to travel in order to purchase pesticides in towns, frequently mentioned as Arusha or Moshi. Some of these stored pesticides were also observed to have become obsolete, which were then likely to create fresh problems of disposal as shown in Figure 1.

The unintended outcomes of pesticide exposures are difficult to reverse once they have been established and are in themselves expensive. Although advances in acute pesticide-poisoning surveillance and treatment in developed countries have led to some achievements in control, pesticide poisoning remains a public health problem globally, particularly in developing countries.[7] Those applying pesticides need to understand the effects of these chemicals to the environment and to their own health and the resulting costs. Alternative pest management strategies that are cheaper and friendlier to end users and the environment need to be promoted.

A workshop was organized for Southern Africa Development Community (SADC) registrars of pesticides in Arusha on October 13–14, 2006. Participants suggested that WAHSA-TPRI (Table 1) should pilot the tool used in the Asia Pacific to establish a systematic mechanism for pesticide monitoring and data collection, with a view to determining the extent of pesticide exposures, injuries, and diseases at the community level. WAHSA-TPRI then selected Ngarenanyuki as the study area, based on their working experience in Northern Tanzania in health hazards posed by pesticides and on the knowledge that farmers in the area were especially at risk with regard to pesticide poisoning.

FIGURE 1 Hazardous practices observed in Ngarenanyuki, 2006–2007. (left) Haphazard disposal of empty containers. (right) Dispensing/repackaging of pesticides in retail shops.

TABLE 1 The WAHSA Program Was Established in October 2004 as a Regional Initiative in Southern Africa to Build Capacity in the Region in Occupational Health

One of its key programs was its project ***Action on Health Impacts of Pesticides,*** which aimed to:

- Improve pesticide-safety materials for the SADC region
- Intervene to reduce pesticide usage
- Improve on agricultural policies and pesticide registration
- Enhance knowledge and improve surveillance about pesticide exposures and health impacts in the region
- Foster a strong regional network for information exchange and consultation

This entry reports on the process involved in the establishment of a community-monitoring team in Ngarenanyuki and on the preliminary results of the monitoring exercise. The authors hope that the findings will facilitate a process to identify those resources required to reduce pesticide use, the development of an action plan to access and mobilize these resources, and the further establishment of an effective system of communication among members of the community on pesticide use and access to any other information with regard to pesticide poisoning.

Methodology

This initiative was intended to pilot the CPAM approach that has already been used successfully in Asian countries. The WAHSA-TPRI Team was trained on the subject through their link with a non-governmental organization, AGENDA for Environment and Responsible Development, and was then employed to mobilize the community in Ngarenanyuki.

The initiative adopted a participatory research methodology by involving farmers in the collection and analysis of pesticide-related data. Data collection tools were developed by making use of Community Pesticide Action Kits (CPAKs), and training materials were developed in the regional language of Swahili. As a result of these activities, the community was sensitized and a subsequent rise in awareness was noted. CPAKs were produced by an ASEAN team of citizens' groups and farmer schools as a tool for action and advocacy, encouraging community education/empowerment. It contains modules that address various aspects of concern such as Warning! Pesticides are a Danger to Your Health; Breaking the Silence: Pesticides in Plantations; Profiting from Poisons: The Pesticides Industry; Drop Pesticides! Build a Sustainable World; Pesticides Destroy our World; Women and Pesticides; Keeping

Watch: Pesticides Laws; How to say NO! to Pesticides: Community Organizing; and Seeking out the Poisons: A Guide to Community Monitoring. The modules are not complete in themselves but need additional materials in local languages.

After securing the community's consent and the involvement of farmers' representatives and communities in capacity building, the program of community-based data collection and analysis was started and the monitoring exercise was implemented. Selected farmers worked in collaboration with the WAHSA experts to monitor and record issues related to pesticide use and exposures in the Ngarenanyuki villages.

Through a series of village meetings, the ward government in collaboration with the ward extension officer invited the farmers to participate in the training. Thirty farmers were selected by the farmers themselves from two villages (Uwiro and Olkung'wado) to represent each subvillage. No farmers were selected from Ngabobo, Kisimiri Juu, and Kisimiri Chini as communication became difficult. Ngarenanyuki as a whole is situated between Mt. Kilimanjaro and Mt. Meru, the first and third tallest mountains in Africa, respectively. The terrain thus consists of rocks, hills, valleys, rivers, and streams, which make some areas impassable during the wet season. One of the villages left out did not cooperate well with the others as they are believed to grow cannabis, a plant that is illegal in Tanzania.

A 6-day training of 25 representative farmers from the Ngarenanyuki was conducted by WAHSA-TPRI scientists, who had expertise in agricultural extension, agronomy, toxicology, entomology, plant pathology, and environmental science, and covered the following topics:

1. Pesticide use and their impacts on human health and environment, where farmers learned about pesticide use around the world, including examples of the negative impacts of pesticides on human health and the environment.
2. Pesticide identification and classification according to their acute toxicity, where farmers learned to identify the types of pesticides used and how they are classified according to their acute toxicity by the World Health Organization (WHO). One such classification the farmers learned was according to the different chemical families, such as organochlorines, organo-phosphates, carbamates, pyrethroids, and so on.
3. Pesticide label identification and interpretation, where farmers learned to read and understand pesticide labels, the various pictograms, and colored warning signs on containers, including the interpretation of the various toxicity symbols.
4. Pesticide handling and management, where farmers learned how to handle pesticides properly to safeguard themselves, their families, their neighbors, and their surroundings. They learned to observe which pesticides are used in their area, how they are used, and to observe the protective measures that are taken during mixing and application. They also learned how they might get contaminated during the handling of pesticides.
5. Pesticide storage and disposal of empty pesticide containers, where farmers learned about proper storage and disposal of surplus pesticides and their empty containers. They also learned how the improper storage of pesticides and the careless disposal of empty containers could form a risk for children, foodstuff, freshwater supplies, farm animals, and so on.
6. Recognizing the signs and symptoms of pesticide poisoning, where farmers learned the different signs and symptoms of pesticide poisoning and how to recognize them. They also learned how to distinguish these from other signs and symptoms that are simply due to poor health.
7. Pest identification and management, where farmers learned how to identify different insects, distinguishing genuine pests from more beneficial insects and symptoms of common vegetable diseases. Farmers were introduced to the basic principles of pest control methods. They were thus equipped with a practical knowledge of insects as an important component in pest management and on how to protect their crops from insect attack with a view to reducing the insecticide load on the environment.

8. Introduction to Integrated Pest Management (IPM), where farmers learned the principles of IPM as a sustainable approach to managing pests by combining biological, cultural, physical, and chemical tools in a way that minimized financial, health, and environmental risks. Farmers were further informed that one of the primary missions of IPM was to assist them in producing profitable crops, using environmentally and economically sound approaches.
9. Spraying equipment and techniques, where farmers learned about spray equipment [such as the knapsack sprayer, the motorized ultra low volume (ULV) sprayer, and so on]. They also learned about the handling, maintenance, and spraying techniques with regard to this equipment.
10. Reducing pesticide costs, where farmers learned to assess actual costs of pesticide use to include direct and indirect costs.
11. Participatory data collection methodology, where farmers learned about methods of data collection and analysis. They were introduced to the kind of data needed and data collection procedures, using different techniques with different data collection tools. Demonstration and practical sessions on how to handle and record data were also held with the farmers.

Establishment of Community Pesticide Monitoring Team

The 25 trained representative farmers were divided into teams of at least three people each, who then became the focal point for monitoring and recording of all pesticide incidences in Ngarenanyuki, and who also worked closely with the WAHSA-TPRI Team, including those who had been working with communities in the Arumeru district in research and training in their respective fields.

Data Collection

Consent forms were developed to be completed by those individuals who agreed to participate in the Community Monitoring Project. Three data collection tools were developed: a questionnaire, a checklist, and a self-surveillance form. They were designed to cover all areas of interest in community pesticide monitoring through interviews, observation, and self-examination of pesticide exposure.

The tools were pretested for validity and consistence in Mlangarini ward with a sample of 30 farmers. Adjustments and other improvements were made to the tools prior to final data collection in Ngarenanyuki. A self-surveillance form without pictograms was preferred, owing to some confusion arising from the meaning of the pictograms.

Farmers were organized into teams to conduct crop surveys and recognize damages, assess losses, and collect insect pests for identification. The farmers went out into their respective villages to collect information on pesticides used, perceptions on pesticide hazards, poisoning, and symptoms using the questionnaire. They also used the checklist to observe and record pesticides available in the area, means of storage, and use of protective equipment. Each farmer contacted by the team members was asked to do self-surveillance and record pesticide use conditions and practices, as well as poisoning signs and symptoms experienced.

The Community Pesticide Monitoring Teams needed technical support and close follow-up to ensure consistency in data collection and in transferring the knowledge gained to the entire Ngarenanyuki community. However, the teams were fully prepared in getting the message across to the community and to involve them in providing relevant information regarding pesticide issues. Moreover, the village leaders were made responsible for making a close follow-up of the team and the villagers involved.

Data Analysis

Analysis of information collected was performed using two different approaches. Structured interviews were conducted by the farmers using questionnaires, and the information was tallied, before it was tidied up and analyzed with the aid of the SPSS (Statistical Package for the Social Sciences) computer software to obtain frequencies. The data that were collected through observation on the basis of checklists were manually analyzed by the farmers themselves using flip charts and colored pens.

Results and Discussion

The training of the farmers was meant to prepare the community in taking responsibility themselves for monitoring the negative impacts of pesticides in their area. Subject matter specialists conducted the training with the aim of building the capacity of participants to make the right judgement and decisions when dealing with pest and pesticide issues. During the training sessions, farmers expressed keen interest in learning how to recognize insects (beneficial and harmful), to recognize the signs and symptoms of pesticide exposure, to practice safer pesticide handling and management, to understand proper spraying techniques and the maintenance of knapsacks, and to understand the benefits of participatory data collection techniques.

General Information

The data from the farmers in Ngarenanyuki Ward were collected between February and April 2007 by the farmers trained on pesticide monitoring and analyzed using the SPSS computer software. While the majority of the 120 farmers were males (90%), the average highest education level recorded was that of primary education (76%); hence, functional literacy was not a problem in this community. Agriculture (98.3%) was the major income-generating activity, although some farmers also kept livestock.

Pesticides Used in Ngarenanyuki

Thirty different types of pesticides commonly used in Ngarenanyuki were identified by the farmers. The major groups of pesticides used included insecticides, fungicides, and, to a lesser extent, herbicides. The most widely used insecticide and fungicide were chlorpyrifos (72.5%) and mancozeb (69.2%), respectively. Only 36.6% of the pesticides used in Ngarenanyuki had full registration, while some had provisional or experimental registration. It is mandatory for pesticides intended for use in Tanzania to go through a registration process, which involves efficacy and quality tests before they are approved for general, restricted, or experimental use. Pesticides under experimental use are not expected to be sold in retail shops. There was also the presence of one class 1b pesticide (chlorfenvinphos) and banned/restricted pesticides such as DDT.

Pesticide Availability, Affordability, and Application

The majority of pesticides (86.7%) were locally available in Ngarenanyuki, and a considerable proportion of farmers (65.8%) could afford to buy pesticides. Those unable to buy mostly obtained their pesticides on credit, and paid after harvesting. Most farmers (68.3%) claimed to have a pesticide application timetable, the most prominent approach being that of applying pesticides whenever insects or disease symptoms appeared.

Pesticide Mixing

The majority of farmers (90%) mixed more than one pesticide in a single application. The main reason given for mixing was to kill all pests and diseases at a go and to improve the quality of leaves and fruits (54%) in the field. A few (25%) said they preferred mixing to ease the workload and in order to cover larger areas with one treatment, while some (55%) said they had simply followed the pesticide retailer's advice. The mixing exercise was widely done in respective farms (89%). The common mixtures normally contained more than one fungicide and one insecticide, although some mixtures were found with around three fungicides and two insecticides.

Frequency of Pesticide Application, Number of Risk Days per Year, Spraying Equipment, and Pesticide Storage

The scale of environmental pollution was fairly evident as pesticides could be smelled all over the farms and in nearby residential areas, causing health problems (such as cough, sneezing, excessive difficulty breathing, and chest pains) to both sprayers and those who found themselves in the path of the sprays. The farmers worked out 52 risk days per year, as the majority (73%) of 120 farmers applied pesticides once a week and fewer (18%) applied the pesticides twice a week. The most common spraying equipment was the knapsack (76%), while in some cases (21%), buckets were also used.

Most respondents (57%) stored pesticides in a pesticide store, and in some cases, storage took place in sitting or living rooms (12%), in general stores (13%), and in bedrooms (7%). Pesticides were also found to be stored in toilets (1%). The choice of storage areas was often determined by their offering protection against thieves.

Adherence to Pesticide Label Instruction

The study by the farmers revealed that many of them did read the instructions on the pesticide containers, but only few actually followed the instructions as shown in Table 2. An example is the mixing of ULV formulations in water sprays while instructions given on the label are for direct application without dilution. The following were the arguments put forward for not following instructions: that some labels were only written in English, that the farmers were not familiar with conventional signs and symbols, and that some containers had no labels at all, having been dispensed from another container. The repackaging and dispensing of smaller quantities was found to be a common, albeit illegal, practice and it was felt that this needed greater attention, since this practice has negative implications for efforts to reduce the worst effects of pesticide poisoning by the implementation of proper labeling and instructions.

Disposal of Pesticide Containers

The major mode of disposal of empty pesticide containers by most of the farmers was by simply throwing the containers away in the farm surroundings and by burning. It was observed that some empty containers from the pesticide retail shops were also thrown or burnt at the marketplace. It was also revealed that some farmers did reuse empty pesticide containers for domestic purposes such as buying cooking oil and kerosene and for local brewing (Table 3).

TABLE 2 Adherence to Pesticide Label Instruction by Farmers in Ngarenanyuki during a Pevious Farming Season (December 2006 to March 2007)

Response toward Pesticide Label Instructions	Number of Farmers (N = 120)	% of Farmers
Always read instructions	72	60
Follow instructions	45	38
Sometimes follow instructions	34	28
Sometimes read instructions	28	23
Trained on pesticide issues	16	13
Get information on pesticides	7	6

TABLE 3 Modes of Disposal of Empty Pesticide Containers in Ngarenanyuki (December 2006 to March 2007)

Mode of Disposal	Number of Farmers (N = 120)	% of Farmers
Burn	41	34
Throw away on the farm	35	29
Bury in the farm surroundings	18	15
Sell back to pesticide vendors	8	7
Throw in the toilet	7	6
Use for other domestic uses	2	2

Pesticide Poisoning

The self-surveillance form was used without pictograms, to record the signs and symptoms of pesticide poisoning, owing to the confusion arising from the use of pictograms. During the pretest, farmers did not understand what the pictograms meant, and since the majority were able to read and write, it was agreed that there would be no need to include picto-grams in the surveillance form and that the list of signs and symptoms provided in the form was sufficient until proper research had been undertaken to determine what visual aid would be considered appropriate for the target audience.

The majority of the farmers (69.2%) had experienced pesticide poisoning in the previous farming season, owing to exposure, much of which had occurred more than 3 times to a single farmer. Pesticide poisoning was characterized by signs and symptoms known from previous studies to be related to pesticide exposures. Cypermethrin–profenofos mixture and profenofos were mostly associated with poisoning, and the action taken by many of those exposed (43.3%) was to drink milk, while a few respondents had attended hospitals for a proper medical examination. A considerable high proportion (57.5%) had been admitted more than 3 times, owing to pesticide poisoning (Table 4). Validation of poisoning through biological monitoring was not possible during this pilot stage but has been planned in future surveillances.

The action taken by the 25 farmers who fully participated in the pilot study was to intensify the training by initiating capacity-building sessions in all villages in Ngarenanyuki. They held community pesticide monitoring training in every village meeting, gave feedback to the WAHSA-TPRI Team on the farmers' reaction, and suggested what further input they needed from the team.

Conclusion

This pilot enabled the building of Ngarenanyuki farmers' capacities to assess their own health and environment as far as pesticides were concerned, analyze the situation, develop a plan of action, and work toward improving their condition. It facilitated the farmers' capacity so that they could take control and work with pesticides more safely and so become healthier. This program therefore works to benefit not only the farming community in the long run but also those consumers who would otherwise be forced to eat contaminated crops, and it contributes to the health of the environment as a whole.

Relevant data relating to pesticides, such as their availability, their usage, the farmers' handling practices, risk perception, and behavior, all gathered during the pilot project, enabled farmers in Ngarenanyuki and the WAHSA-TPRI Team to properly document the incidents and adverse events resulting from pesticide use. The initial evaluation of the association of the observed adverse event and pesticide exposure revealed that different pesticide-related tasks gave rise to signs and symptoms of pesticide poisoning and that skin and eye problems, for example, needed more attention during interventions.

The impact of the sensitization and awareness-raising seminars has been dramatic. The disposal site at the marketplaces vanished and the mistake of mixing ULV formulations with water has also been abandoned by the trained farmers and their associates. The farmers realized that the formulation was suspended in water and they were spraying water in some areas instead of pesticide. The formation

TABLE 4 Pesticide Poisoning, Circumstances, and Action Taken in Ngarenanyuki (December 2006 to March 2007)

Event		Number of Farmers (N = 120)	% of Farmers
Pesticide poisoning	Affected by pesticides in the last farming season	83	69
	Not sure	18	15
Occurrence of effects in the last farming season	Once	11	9
	Twice	12	10
	Thrice	9	8
	More than three times	26	22
Pesticides used	Profenofos + cypermethrin	32	27
	Profenofos	25	21
	Mancozeb (Dithane)	14	12
	Endosulfan	14	12
	Triadimenol	12	10
	Chlorothalonil	12	10
	Lambda-cyhalothrin	8	7
	Mancozeb (Ivory)	8	7
	Deltamethrin	1	1
	Copper sulfate	1	1
Action taken after pesticide Poisoning	Drank milk	52	43
	Went to the hospital	34	28
	No action taken	4	3
	Washed with water	2	2
Number of times admitted due to pesticide poisoning	Once	20	17
	Twice	23	19
	Thrice	8	7
	More than three times	69	58

of the community monitoring teams enhanced the whole process of data collection and action being taken. This has also provided a base for the sustainability of the project as the team continues to be in close contact and collaboration with the WAHSA-TPRI Team through the training of other farmers and in responding to their queries on pests, pesticides, monitoring pesticide use, their application, and the disposal of obsolete pesticides and empty pesticide containers.

The project was well received by the Ngarenanyuki community and has shown that if it is applied elsewhere, it will help in changing risk behaviors and in reducing the negative impact of pesticide exposures in communities. It is therefore recommended that the program be implemented systematically in Ngarenanyuki and be extended to other communities in Tanzania such as Mang'ola, in Karatu District, where the current use of pesticides appears to be indiscriminate.

Acknowledgments

We are indebted to the small-scale vegetable farmers in Ngarenanyuki for their cooperation. This project could not have been implemented without support from AGENDA for Environment and Responsible Development for which we are grateful. We appreciate the work done by the WAHSA-TPRI Team and the contribution of experts from Selian Agricultural Research Institute (Dr. Hussein Man-soor) and Mikocheni Research Institute (Dr. Ruth Minja) during the farmers' training.

References

1. WHO. *Public Health Impact of Pesticide Used in Agriculture*; Geneva, 1990.
2. ILO Chemicals in the Working Environment. In *World Labour Report;* International Labour Office: Geneva, Switzerland, 1994.
3. Murphy, H.H.; Hoan, N.P.; Matteson, P.; Abubakar, A.L. Farmers' self-surveillance of pesticide poisoning: A 12-month pilot in northern Vietnam. Int. J. Occup. Environ. Health **2002**, *8* (3), 201–211.
4. Rengam, S. Breaking the silence: Women struggle for pesticide elimination. In *Silent Invaders*; Jacobs, M., Dinham, B., Eds.; Zed Books Ltd.: London, 2003.
5. Quijano, R.F. *Endosulfan Poisoning in Kasargod, Kerala, India*: *Report on a Fact-Finding Mission*; Pesticide Action Network Asia and the Pacific: Penang, 2002.
6. Ngowi, A.V.F.; Mbise, T.J.; Ijani, A.S.M.; London, L.; Ajayi, O.C. Smallholder vegetable farmers in Northern Tanzania: Pesticides use practices, perceptions, cost and health effects. Crop Prot. **2007**, *26*, 1617–1624.
7. Ngowi, A.V.F. *Health Impact of Exposure to Pesticides in Agriculture in Tanzania*; PhD Thesis, Acta Universitatis Tamperensis 890; University of Tampere, 2002.

22

Developing Countries: Pesticide Health Impacts

Aiwerasia V.F. Ngowi, Catharina Wesseling, and Leslie London

Introduction

The use of potentially hazardous chemicals is increasing in developing countries whose populations have the least capacity to protect themselves. Hundreds of thousands of people die annually from the effects of use, misuse, or accidental exposures to pesticides.[1,2] Developing nations in Africa, Asia, and Latin America comprise more than 75% of the total world population, use 25% of the world's pesticides, yet account for 99% of deaths caused by these toxic agents.[3]

Health Impacts

Acute Pesticide Poisoning

Two decades ago, the World Health Organization estimated that three million cases of acute pesticide poisoning resulting in 220,000 deaths occur worldwide each year, the majority in developing countries.[3] However, it is well recognized that these figures are an underestimate because of underdiagnosis and/or underreporting. Diagnostic difficulties are prominent in developing countries,[4–6] owing to insufficient medical training and high background levels of ill health.

Organophosphorus insecticides are the most common agents involved in acute pesticide poisonings, accounting for between 50% and 80% of all poisonings in Asia[7] and are a major public health concern in most African countries, where approximately 80% of the workforce is involved in agricultural work. In Central America, pesticides identified as causing most poisonings between 1992 and 2000 were paraquat; aluminum phosphide; the organophosphates methyl parathion, methamidophos, monocrotophos, chlorpyrifos, terbufos, and ethoprophos; the carbamates carbofuran, methomyl and aldicarb; and endosulfan.[8]

Part of the reason for this picture is the continued use in developing countries of pesticides no longer registered for use in the developed world, because of their high toxicity, and the substitution of persistent organochlorines with organophosphate insecticides.

Fatality rates and lifelong disability resulting from pesticide poisoning in developing countries are exacerbated by poor diagnosis and delayed treatment, resulting in both human suffering and economic losses.

High rates of unintentional poisoning, mostly occupational, have been reported in rural agricultural working and urban populations worldwide.[9,10] Nearly 66,000 cases of acute pesticide poisonings occur annually in Nicaragua.[11] Mass poisonings by pesticides in developing countries have typically resulted in high numbers of fatalities.

In the remote Andean village of Tauccamarca in October 1999, 42 children were poisoned after eating a school breakfast contaminated with the organophosphate pesticide methyl parathion, resulting in 24 deaths before the children could reach medical treatment.[12]

However, it is only a limited number of the most extreme cases in developing countries, which appear to be documented. Less high-profile cases are common but unrecorded. For example, a methomyl-poisoning incident involving 11 female flower farm workers in Arusha, Tanzania, in March 2004 was reported in the press, but absence of adequate local investigation mechanisms prevented its documentation in the peer-reviewed literature.

Deliberate self-harm is a major problem in the developing world. Pesticides are commonly used as suicidal agents throughout developing nations and are associated with high mortality rates, causing an estimated 300,000 deaths annually in the Asia Pacific.[1,2,8,9,13] In India, suicide using aluminum phosphide was reported as so common that postmortem examinations on deceased bodies were said to be routinely conducted by staff wearing respirators for personal protection from released fumes.[14]

Underlying factors that make individuals at risk for self-harm are both social (including domestic problems, poverty, social isolation, and financial hardship) and medical.[1] Farmer indebtedness, widespread in many developing countries characterized by unequal economic systems, is an important factor driving high rates of suicide. More recent findings suggest that pesticides, particularly organophosphates, may be more than agents in suicidal attempts; they are also part of the causal pathway because of their neurotoxicity and the possible links between organophosphate exposure, depression, and impulsivity, mediated through effects on neurotransmitters such as serotonin.[9] In a context where the above social risk factors for depression are common in developing countries, further exposure to neurotoxic pesticides may substantially increase the risks of suicide.

Chronic Health Impacts Unknown

Although long-term consequences of pesticide poisoning are well recognized in the literature, relatively few studies of long-term health effects of pesticide exposure have been conducted among working populations in developing countries. Underdiagnosis is accentuated for long-term health consequences that require greater diagnostic capacity. Dermal exposure routes for developing country workers are also common but are an underdocumented yet critical pathway for systemic poisonings, both acute and chronic. Consequently, the extent of chronic health impacts of pesticides in developing country workers is poorly characterized.

However, there is little reason to believe that their impact would be any less than that in developed countries. Indeed, high levels of background morbidity and poor social conditions are likely to aggravate pesticide toxicity. For example, research among South African farm workers highlighted the link between chronic lifetime undernutrition, organophosphate exposure, and impaired neurological performance on tests of vibration threshold.[15]

Azoospermia (absent sperm), oligospermia (low sperm count in semen), and low fertility have been documented in more than 26,000 workers previously exposed to 1,2-dibromo-3-chloropropane (DBCP) on banana and pineapple plantations in more than 12 countries.[16]

Weak Surveillance for Hazards and Impact

Although a critical public health tool for the control of pesticide poisoning, surveillance in developing countries is bedeviled by multiple problems such as lack of access to health care for poisoning survivors, lack of human resources, diagnostic skills and equipment to identify cases, and weak information systems. Acute poisoning rates are consequently underestimated and may selectively undercount certain types of poisoning (occupational circumstances) and certain risk groups (women and migrant workers). Lack of professional competence and conflict of interest arising from compensation system levies may also lead occupational poisonings to be misreported as suicide.[10] As a result, inferences from review of flawed data may lead to mistaken policy decisions.[17]

To improve information on the extent of pesticide poisoning in developing countries, surveillance systems for acute health effects from pesticides are being established in developing nations. In 1998, almost 6000 pesticide poisonings were reported in five of the seven Central American countries generating an estimate (corrected for underreporting) of 30,000 pesticide poisonings annually in the region.[11] Poisoning rates reported in an intensified surveillance intervention in South Africa increased 10-fold in the study area compared to a control area.[17] Recently, WHO in collaboration with partners initiated a community intervention, the Global Public Health Initiative, to prevent self-harm by pesticide poisoning.[18]

Weak Regulation and Enforcement

Vulnerable economies and weak infrastructure in developing nations hinder their ability to regulate the use of pesticides, particularly when macroeconomic pressures promote deregulation and restrict public spending required to implement regulatory controls. As a result, marketing and advertising of pesticides are often uncontrolled. Incorrectly labeled or unlabeled formulations, including readymade solutions in soft drink bottles and other containers, are commonly sold at open stands. In South Africa, the repackaging of aldicarb granules into small-volume packets sold by street vendors[19] for domestic pest control has been linked to increasing numbers of poisonings in urban areas. Low retail prices, sometimes associated with subsidy policies, promote risky pesticide use. Weaknesses in sustainable international and national agricultural and chemical management policies manifest in a reliance on "safe-use" strategies. Yet, evidence has shown that the so-called "good agricultural practices" and "safe use" are ineffective in controlling risks in developing countries, principally because many measures assumed to enable safe use are not feasible in developing countries, particularly under tropical or adverse climatic conditions.[20]

Low Levels of Worker and Community Awareness

Farmers and farm workers rarely have access to adequate training in pesticide safety or advice on the complicated management of pesticides. Hot climates are a disincentive to use of protective clothing, and many workers and farmers lack access to water for washing hands or exposed skin, increasing the risks of contamination. Recognition of pests and their predators is generally low, leading to overreliance on routine pesticide applications to control pests; knowledge of product selection, application rates, and timing is poor; different products are often combined in the belief that the effect will be greater; re-entry periods after spraying are not known; and without knowledge of alternatives, farmers often assume that the only solution to pest problems is to spray more frequently.[21]

Pesticides are often stored improperly in or around farmers' homes, increasing family member's access.[21,22] In some instances, empty pesticide containers are reused to store water and food, resulting in serious poisonings.

Import/Export of Banned and Restricted Compounds

Pesticides banned or restricted in developed countries are often easily available in developing countries. These include pesticides causing significant acute and chronic morbidity (such as class I and II organophosphates and paraquat) and organochlorines earmarked for eradication under the Stockholm Convention on persistent organic pollutants (POPs) (particularly dieldrin, lindane, and chlordane). Endosulfan, a candidate pesticide for inclusion under the POPs treaty, has been responsible for a series of poisonings in Benin[21] and developmental impacts on children in Kerala, a state in India.[21,23]

The use of *p*-*p′*-dichlorodiphenyltrichloroethane (DDT) continues to be permitted for malaria control in developing countries, where malaria remains endemic, despite its known hazards for wildlife and controversial adverse effects on human health.[24,25] Ironically, DDT use in Africa has increased since the Stockholm Convention came into effect.[26] As a result, it is still produced for export in at least three countries. Because of its ongoing usage for public health vector control, unauthorized use for agricultural purposes remains a concern in developing countries, particularly where regulatory controls are weak. The presence and persistence of DDT and its metabolites worldwide are still problems of great global relevance to public health.

Although the Prior Informed Consent (PIC) procedure, upgraded in status from a voluntary agreement to an international convention known as the Rotterdam Convention, seeks to protect developing countries from harms arising from import of chemicals banned or restricted in exporting countries, the effectiveness of the Convention has been questioned. For example, the data requirement for a Severely Hazardous Pesticide Incident report, used under the Convention to add a pesticide onto the controlled list, lacks mechanisms suited to developing country conditions. This is usually because developing countries lack the infrastructure to collect the required data. Similarly, the process of adding a pesticide onto the POPs list (e.g., endosulfan) is often met in practice with strong resistance. Even when pesticides are restricted by the POPs and Rotterdam conventions, compliance with the obligations contained in the conventions may often be poor, despite developed countries ratifying the conventions. The PIC and POP secretariat could assist local contacts such as the Designated National Authorities (DNAs) in developing simple tools that will be used to collect relevant data and help in the establishment of suitable mechanisms for the flow of data and information in member countries.

Over the past few years, pressure from non-governmental organizations and discussions within the Food and Agriculture Organization and other intergovernmental bodies has recognized that greater effort must be put into restricting the availability of pesticides based not only on their acute toxicity, traditionally measured through the WHO classification system, but also on the capacity of particular pesticides to cause long-term toxic effects with chronic exposure. These initiatives, particularly the call for a progressive ban on highly hazardous pesticides, offer some hope for better protections for developing country populations, but are still in development.

Lack of Technical and Laboratory Capacity

Many developing countries suffer from a lack of human and technical resources, aggravated by the global brain drain and weak economies. As a result, few developing countries are able to monitor pesticide residues. Most developing countries do not have laboratories capable of conducting analyses for pesticides and their residues, particularly at standards that meet good laboratory practice. Where laboratory capacity is available, it is usually to service residue testing of agricultural exports destined for consumers in developed countries. Produce grown for domestic consumption is rarely monitored.

Environmental media such as water and soil are rarely tested, and, even then, usually only on a research basis. Isolated studies of lactating women in Southern Africa have confirmed the presence of high levels of DDT metabolites in breast milk in populations living in malaria-endemic areas subject to DDT applications. Yet, despite provisions arising from the POPs treaty to undertake routine testing to monitor the impact of DDT use, there is no system for biological monitoring for DDT metabolites in place in Southern Africa. As a result, many infants in the region are substantially exposed through cross-placental transfer and breastfeeding, with potential adverse impacts on childhood neurodevelopment.

Research capacity to identify problems and develop prevention strategies is also constrained by limited investments in capacity building in relevant scientific fields in developing countries. As a result, there is neither proactive monitoring nor information systems usage to effect adequate responses to pesticide problems identified. It is critical to foster South–South learning to promote best practice because what applies in the north may be different to what happens in the south. There is a need to build southern capacity because reliance on the north at times perpetuates many of the problems. Indeed, there is much expertise in the south from which both north and south can learn.

Pest Control Policies

Unlike many developed countries, agricultural policies in many developing countries have emphasized short-term economic gains at the expense of environmental sustainability or human health. Few developing countries have adopted integrated pest management or pest reduction strategies. The dominant "pesticide culture" assumes that the use of pesticides to control pest as the first option is the norm, is reinforced by advertising and marketing practices, and is often encouraged by agricultural credit policies and development aid. Much needs to be done to enhance research and development to support pesticide reduction for agriculture and public health, and to strengthen the capacity in developing countries to develop monitoring systems and research capacity to deal with the problems of pesticides in developing nations. Reducing deaths from pesticide poisoning through restrictions on the availability of pesticides can be accomplished based on a prior evaluation of national agricultural needs and the development of a plan to encourage substitution with less toxic pesticides without loss of agricultural output,[27] bearing in mind that policies aiming towards sustainable chemical-free agriculture would be the ideal long-term solution.

Conclusions

Underestimations of acute and long-term effects of pesticide in developing countries occur due to underdiagnosis and/or underreporting. The impact of pesticide poisoning is also unknown because of weak surveillance for hazards and impact, import/export of banned or restricted compounds, lack of technical and laboratory capacity, weak regulations and enforcement, low level of worker and community awareness, and inappropriate pest control policies. Enhancing research and development to support pesticide reduction for agriculture and public health and strengthening capacity to develop monitoring systems are critically important for developing countries to deal with problems concerning pesticides.

References

1. Konradsen, F.; van der Hoek, W.; Cole, D.C.; Hutchinson, G.; Daisley, H.; Singh, S.; Eddleston, M. Reducing acute poisoning in developing countries—Options for restricting the availability of pesticides. Toxicology **2003**, *192* (2–3), 249–261.
2. Konradsen, F.; Dawson, A.H.; Eddleston, M.; Gunnell, D. Pesticide self-poisoning: Thinking outside the box. Lancet **2007** *369* (9557), 169–170.
3. WHO. *Public Health Impact of Pesticide Used in Agriculture;* World Health Organization and United Nations Environment Programme: Geneva, 1990.
4. Corriols, M.; Marin, J.; Berroteran, J.; Lozano, L.M.; Lundberg, I.; Thörn, A. The Nicaraguan Pesticide Poisoning Register: Constant underreporting. Int. J. Health Serv.**2008** *38* (4), 773–87.
5. Mbakaya, C.F.L.; Ohayo-Mitoko, G.J.A.; Ngowi, A.V.F.; Mbabazi, R.; Simwa, J.M.; Maeda, D.N.; Stephens, J.; Hakuza, H. The status of pesticide usage in East Africa. Afr. J. Health Sci. **1994**, *1*, 37–41.

6. London, L.; Myers, J.E. Critical issues for agrichemical safety in South Africa. Am. J. Ind. Med. **1995**, *27,* 1–14.
7. He, F.; Xu, H.; Quin, F. Intermediate myasthenia syndrome following acute organophosphates poisoning—An analysis of 21 cases. Hum. Exp. Toxicol. **1998**, *17,* 40–45.
8. Henao, S.; Arbelaez, M.P. Epidemiologic situation of acute pesticide poisoning in Central America, 1992–2000. Epidemiol. Bull. **2002**, *23* (3), 5–9.
9. London, L.; Flisher, A.J.; Wesseling, C.; Mergler, D.; Kromhout, H. Suicide and exposure to organophosphate insecticides: Cause or effect. Am. J. Ind. Med. **2005**, *47,* 308–321.
10. Wesseling, C. Multiple health problems in Latin America. In *Silent Invaders: Pesticides, Livelihoods and Women's Health*; Jacobs, M., Dinham, B., Eds.; ZED Books: London, 2003.
11. Corriols, M.; Marin, J.; Berroteran, J.; Lozano, L.M.; Lundberg, I. Incidence of acute pesticide poisonings in Nicaragua: A public health concern. Occup. Environ. Med. **2009**, *66*, 205–210.
12. Rosenthal, E. The tragedy of Tauccamarca: A human rights perspective on the pesticide poisoning deaths of 24 children in the Peruvian Andes. Int. J. Occup. Environ. Health **2003**, *9*, 53–58.
13. Chowdhury, A.N.; Banerjee, S.; Brahma, A.; Biswas, M.K. Pesticide poisoning in non-fatal deliberate self-harm: A public health issue. Study from Sundarban delta, India. J. Psychiatry **2007**, *49,* 262–266.
14. Levine, R.S.; Doull, J. Global estimates of acute pesticide morbidity and mortality. Rev. Environ. Contam. Toxicol. **1992**, *129,* 29–50.
15. London, L. Occupational epidemiology in agriculture: A case study in the Southern African context. Int. J. Environ. Occup. Health **1998**, *4*, 245–256.
16. Slutsky, M.; Levin, J.L.; Levy, B.S. Azoospermia and oligospermia among a large cohort of DBCP applicators in 12 countries. Int. J. Occup. Environ. Health **1999**, *5* (2), 116–122.
17. London, L.; Bailie, R. Challenges for improving surveillance for pesticide poisoning: Policy implications for developing countries. Int. J. Epidemiol. **2001**, *30* (3), 564–570.
18. WHO. *Safer Access to Pesticides: Community Interventions*; World Health Organization and International Association for Suicide Prevention: Geneva, 2006.
19. Rother, A. Falling through the regulatory cracks: Street selling of pesticides and urban youth in South Africa. Int. J. Occup. Environ. Health **2010**, *16,* 202–213.
20. Wesseling, C.; Ruepert, C.; Chavarri, F. Safe use of pesticides: A developing country's point of view. In *Encyclopedia of Pest Management;* Marcel Dekker, Inc.: New York, 2003.
21. Ngowi, A.V.F.; Maeda, D.N.; Wesseling, C.; Partanen, T.J.; Sanga, M.P.; Mbise, G. Pesticide handling practices in agriculture in Tanzania: Observational data on 27 coffee and cotton farms. Int. J. Occup. Environ. Health **2001**, *7,* 326–332.
22. Dinham, B.; Malik, S. Pesticides and human rights. Int. J. Occup. Environ. Health **2003**, *9*, 40–52.
23. Saiyed, H.; Dewan, A.; Bhatnagar, V.; Shenoy, U.; Shenoy, R.; Rajmohan, H.; Patel, K.; Kashyap, R.; KuLkarni, P.; Rajan, B.; Lakkad, B. Effect of endosulfan on male reproductive development. Environ. Health Perspect. **2003**, *111*, 1958–1962.
24. Bouwman, H.; Becker, P.J.; Schutte, C.H.J. Malaria control and longitudinal changes in levels of DDT and its metabolites in human serum from KwaZulu. Bull. W. H. O. **1994**, *72* (6), 921–930.
25. Eskenazi, B.; Chevrier, J.; Rosas, L.G.; Anderson, H.A.; Bornman, M.S.; Bouwman, H.; Chen, A.; Cohn, B.A.; de Jager, C.; Henshel, D.S.; Leipzig, F.; Leipzig, J.S.; Lorenz, E.C.; Snedeker, S.M.; Stapleton, D. The Pine River statement: Human health consequences of DDT use. Environ. Health Perspect. **2009**, *117,* 1359–1367.
26. van den Berg, H. Global status of DDT and its alternatives for use in vector control to prevent disease. Environ. Health Perspect. **2009**, *117,* 1656–1663.
27. Manuweera, G.; Eddleston, M.; Egodage, S.; Buckley, N.A. Do targeted bans of insecticides to prevent deaths from selfpoisoning result in reduced agricultural output? Environ. Health Perspect. **2008**, *116* (4), 492–495.

23

Insulation: Facilities

Wendell A. Porter

Introduction

Insulation is rated in terms of thermal resistance, called the *R*-value, which indicates the resistance to heat flow. Although insulation can slow all types of heat flow—conduction, convection, and radiation—its greatest impact is on conduction.

The higher the *R*-value is, the greater the insulation effectiveness is.[1,2] The *R*-value of thermal insulation depends on the type of material, the thickness, and the density. When calculating the *R*-value of a multilayered installation, the *R*-values of the individual layers are added.

The effectiveness of an insulated wall or ceiling also depends on how and where the insulation is installed. For example, compressed insulation will not give its full rated *R*-value. The overall *R*-value of a wall or ceiling will also be somewhat different from the *R*-value of the insulation itself because some heat flows around the insulation through the studs and joists thermal bridging. With careful design, this short-circuiting can be reduced.

The key to an effective insulation system is proper installation of quality insulation products. A building should have a continuous layer of insulation around the entire building envelope (Figure 1). Studies show that improper installation can cut performance by 30% or more.

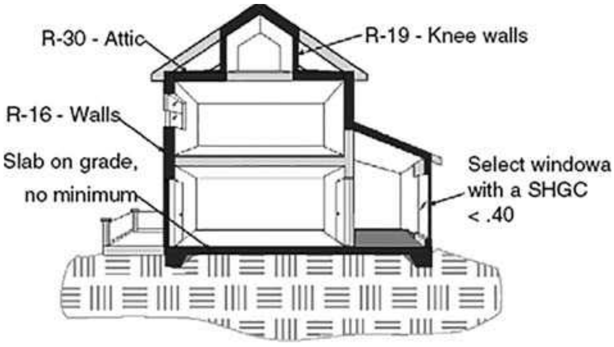


FIGURE 1 Building envelope insulation.

Insulation Materials

The wide variety of insulation materials makes it difficult to determine which products and techniques are the most cost effective (Table 1). Whatever product is chosen, install it per the manufacturer's specifications.

Here are short descriptions of a few of the insulation products available today:

- Fiberglass insulation products come in batt, roll, and loose-fill form, as well as a semirigid board material. Many manufacturers use recycled glass in the production process of fiberglass building insulation, with most using between 20% and 30% recycled glass in their product. Fiberglass is used for insulating virtually every building component—from walls to attics to ductwork.
- The term mineral wool refers to both slag wool and rock wool. Slag wool is manufactured from industrial waste product. It is primarily (~75%) produced from iron ore blast furnace slag, a by-product of smelting. Rock wool is fireproof and produced from natural rocks—basalt primarily—under high heat. Mineral wool insulation is available as a loose-fill product, batts, semirigid, or

TABLE 1 Comparison of Insulating Materials

Material	Typical R-Value (per inch)
Batts, blankets, and loose-fill insulation	
Mineral wool and fiberglass	2.2–4.0
Cellulose (loose-fill)	3.0–3.7
Cotton (batts)	3.0–3.7
Perlite (loose-fill)	2.5–3.3
Foam insulation and sheathing	
Polyisocyanurate	6.0–6.5
Closed-cell, spray polyurethane	5.8–6.8
Open-cell, low-density polyurethane	3.6–3.8
Extruded polystyrene	5.0
Molded expanded polystyrene (beadboard)	4.0
Fiberboard sheathing (blackboard)	1.3
Air-krete	3.9
OSB sheathing (3/8 in.)	0.5
Foil-faced OSB	Depends on installation
Polyicynene	3.6

rigid board. Usage of this product has decreased as more and more building codes require active sprinklering of buildings.

- Cellulose insulation, primarily made from post-consumer recycled newsprint with up to 20% ammonium sulfate and/or borate flame retardants, is installed in loose-fill, wall-spray (damp), dense-pack, and stabilized forms. Because of its high density, cellulose can help reduce air leaks in wall cavities, but air sealing other areas of air infiltration, such as under wall plates and band joists, must be performed to obtain an effective air barrier. However, given certain conditions and applications, cellulose may hold moisture.
- Molded expanded polystyrene (MEPS), often known as beadboard, is a foam product made from molded beads of plastic. MEPS is used in several alternative building products discussed in this entry, including insulated concrete forms and structural insulated panels (SIPs).
- Extruded polystyrene (XPS), also a foam product in rigid board form, is a homogenous polystyrene produced primarily by three manufacturers with characteristic colors of blue, pink, and green.
- Polyisocyanurate, foil-faced rigid board, is insulating foam with one of the highest available *R*-values per inch.
- Closed-cell, high-density spray polyurethane is used both for cavity insulation and as insulating roofing materials [often referred to as spray polyurethane foam (SPF)]. It has structural properties, good adhesive properties, and good compressive strength.
- Open-cell, low-density polyurethane foam is used primarily to seal air leaks and provide an insulating layer. Produced primarily from petrochemicals, some of these products are now manufactured in part from soybeans.
- Aerated concrete, including lightweight, autoclaved (processed at high temperature) concrete, can provide a combination of moderate R-values and thermal mass for floors, walls, and ceilings, in addition to structural framing.
- Reflective insulation is often used between furring strips on concrete block walls to reflect the heat. Note that reflective insulation products differ from radiant barriers in that they include a trapped air space as part of the product. These trapped air spaces may be a result of the way the reflective insulation is manufactured or installed.

Determine actual R-values and costs from manufacturers or local suppliers.

Note that many new types of insulation are rapidly becoming incorporated into conventional construction. However, always research a material's characteristics and suitability to a particular situation before buying any new product. For instance, many new insulation products require covering for fire rating.

Insulation and the Environment

There has been considerable study and debate about the potential negative environmental and health impacts of insulation products.[3] These concerns range from detrimental health effects for the installer to depletion of the earth's ozone layer.

- Fiberglass and mineral wool—questions about effects on health from breathing in fibers. In 2001, the International Agency for Research on Cancer changed its classification for fiberglass and mineral wool from "possible human carcinogen" to "not a known human carcinogen."
- Cellulose—concerns about dust inhalation during installation to VOC emissions from printing inks (these are now almost entirely vegetable-based) and limited evidence of toxicity of boric acid flame retardants. Long-term fire retardancy is unknown. Limited health and safety research has been performed on these products.
- Foam products and chlorofluorocarbons—for years, many foam products contained chlorofluorocarbons (CFCs), which are quite detrimental to the earth's ozone layer. The CFCs were the blowing agent that helped create the lightweight foams. Current blowing agents are

- Expanded polystyrene—pentane, which has no impact on ozone layer, but may increase the potential for smog formation.
- Extruded polystyrene, polyisocyanurate, and poly-urethane—use primarily hydrochlorofluorocarbons (HCFCs), which are 90% less harmful to the ozone layer than CFCs. Some companies are moving to non-HCFC blowing agents.
- Open-cell polyurethane, including the products made by Icynene, Inc. and Demilec, Inc., as well as the newer soy-based foams—use water, which is much less detrimental than other blowing agents (Table 2).

Insulation Strategies

In general, commonly used insulation products are the most economical. Prices can vary according to in staller and location. Review all of the choices, as they offer different *R*-values, suggested uses, and environmental and health considerations.

TABLE 2 Comparison of Insulation Materials (Environmental Characteristics and Other Information)

Type of Insulation	Installation Method(s)	R-Value per Inch (RSI/m)[a]	Raw Materials	Pollution from Manufacture	Indoor Air Quality Impacts	Comments
Fiber insulation *Cellulose*	Loose fill; wall-spray (damp); dense-pack; stabilized	3.0–3.7 (21–26)	Old newspaper, borates, ammonium sulfate	Vehicle energy use and pollution from newspaper recycling	Fibers and chemicals can be irritants. Should be isolated from interior space	High recycled content; very low embodied energy
Fiberglass	Batts; loose fill; semi-rigid board	2.2–1.0 (15–28)	Silica sand; limestone; boron; recycled glass, phenol formaldehyde resin or acrylic resin	Formaldehyde emissions and energy use during manufacture; some manufactured without formaldehyde	Fibers can be irritants, and should be isolated from interior spaces. Formaldehyde is a carcinogen. Less concern about cancer from respirable fibers	
Mineral wool	Loose fill; batts; semi-rigid or rigid board	2.8–3.7 (19–26)	Iron-ore blast furnace slag; natural rock; phenol formaldehyde binder	Formaldehyde emissions and energy use during manufacture	Fibers can be irritants, and should be isolated from interior spaces. Formaldehyde is a carcinogen. Less concern about cancer from respirable fibers	Rigid board (e.g., Roxul) can be an excellent foundation drainage and insulation material
Cotton	Batts	3.0–3.7 (21–26)	Cotton and polyester mill scraps (especially denim)	Negligible	Considered very safe	Two producers; also used for flexible duct insulation
Perlite	Loose fill	2.5–3.3 (17–23)	Volcanic rock	Negligible	Some nuisance dust	

(*Continued*)

TABLE 2 (*Continued*) Comparison of Insulation Materials (Environmental Characteristics and Other Information)

Conversion Factors

R-Value Conversions	To Get	Multiply	By
Thermal resistance (R)	RSI (m2 C/w)	R (ft2 hF/Btu)	0.1761
Insulation R/unit thickness	RSI/mm	R/in.	0.00693

In the chart the heading is R-value per inch (RSI/m); to obtain this number, the RSI/mm is divided by 1000.

Foam insulation

Polyisocyanurate	Foil-faced rigid boards; nail-base with OSB sheathing	6.0–6.5 (42–45)	Fossil fuels; some recycled PET; pentane blowing agent; TCPP flame retardant; aluminum facing	Energy use during manufacture	Potential health concerns during manufacture. Negligible emissions after installation	Phaseout of HCFC ozone-depleting blowing agents completed
Extruded polystyrene (XPS)	Rigid board	5.0 (35)	Fossil fuels; HCFC-142b blowing agent; HBCD flame retardant	Energy use during manufacture. Ozone depletion	Potential release of residual styrene monomer (a carcinogen) and HBCD flame retardant	Last remaining insulation material with ozone-depleting blowing agents
Expanded polystyrene (EPS)	Rigid board	3.6–1.4 (25–31)	Fossil fuels; pentane blowing agent; HBCD flame retardant	Energy use during manufacture	Potential release of residual styrene monomer (a carcinogen) and HBCD flame retardant	
Closed-cell spray polyurethane	Spray-in cavity-fill or spray-on roofing	5.8–6.8 (40–4.7)	Fossil fuels, HCFC-141b (through early 2005) or HFC-245fa blowing agent; nonbrominated flame retardant	Energy use during manufacture, global-warming potential from HFC blowing agent	Quite toxic during installation (respirators or supplied air required). Allow several days of ailing out prior to occupancy	
Open-cell, low-density polyurethane	Spray-in cavity-fill	3.6–3.8 (25–27)	Fossil fuels and soybeans; water as a blowing agent; nonbrominated flame retardant	Energy use during manufacture	Quite toxic during installation (respirators or supplied air required). Allow several days of ailing out prior to occupancy	
Air-Krete	Spray-in cavity-fill	3.9 (27)	Magnesium oxide from seawater; ceramic talc	Negligible	Considered very safe	Highly fire-resistant; inert; remains friable

(*Continued*)

TABLE 2 (*Continued*) Comparison of Insulation Materials (Environmental Characteristics and Other Information)

.	Conversion Factors					
R-Value Conversions		To Get		Multiply		By
Radiant barrier						
Bubble back	Stapled to framing	Depends on installation	Aluminum; fossil fuels	Energy use during manufacture	Minimal offgassing from plastic	Exaggerated R-value claims have been common
Foil-faced	Stapled to framing;	Depends on installation	Aluminum; fossil	Energy use during	Minimal offgassing from	Exaggerated Æ-value claims
polyethylene foam	requires air' space for radiant benefit		fuels; recycled polyethylene	manufacture	polyethylene	have been common. Recycled content in some
Foil-faced paperboard sheathing	Stapled to framing; requires air' space for radiant benefit	Depends on installation	Aluminum; fossil fuels; recycled paper	Energy use during manufacture	Considered very safe	High recycled content. Structural sheathing available (e.g., Thermo-Ply®))
Foil-faced OSB	Most common as attic sheathing	Depends on installation	Wood fiber; formaldehyde binder in OSB; aluminum	Energy use and VOC emissions during manufacture	Formaldehyde emissions	Primary benefit is reduced heat gain

[a] RSI/m: The standard unit of measurement in the United States has been the Imperial unit. The country is converting to the International System (SI) unit—or metric standard—which predominates internationally. To differentiate like terms, you may find "SI" added to the term symbol. For example, RSI refers to the Æ-value in International System (metric) units.

Critical Guidelines

When installing any insulating material, the following guidelines are critical for optimum performance[4]:

- Seal all air leaks between conditioned and unconditioned areas
- Obtain complete coverage of the insulation, especially around doors and windows
- Minimize air leakage through the material with air sealing measures
- Avoid compressing insulation
- Avoid lofting (installing too much air) in loose-fill products

Foam Insulation Strategies

Foam products are primarily economical when they can be applied in thin layers as part of a structural system or to help seal air leaks. Examples include:

- Exterior sheathing over wall framing
- Forms in which concrete can be poured
- As part of a structural insulated panel for building walls
- Spray-applied foam insulation

Floor Insulation

Slab-on-Grade

Slab-on-grade floors consist of a concrete slab poured over at least four inches of compacted gravel or sand and a layer of 10-mil polyethylene used as a vapor barrier. In hot, humid climates, most buildings are built with concrete slab-on-grade.

For colder climates, slabs lose energy as a result of heat conducted outward toward the perimeter of the slab. Insulating the exterior edge of the slab with R-10 rigid insulation can reduce winter heating bills by 10%–20%.

Raised Floor

Raised floor systems (wood and concrete) have specific requirements depending on climate zones. Consult the local building code for specific details.

Wall Insulation

Walls are the most complex component of the building envelope to provide adequate thermal insulation, air sealing, and moisture control.

Concrete Wall Insulation

Foundation walls and other masonry walls are usually built of concrete block or poured concrete. Insulating concrete block walls is more difficult than insulating framed walls.

Insulating Concrete Block Cores

Builders can insulate the interior cores of concrete block walls with insulation such as:

- Vermiculite R-2.1 per inch (See Figure 2)
- Polystyrene inserts or beads R-4.0–5.0 per inch
- Polyurethane foam R-5.8-6.8 per inch

Unfortunately, as shown in Figure 2, the substantial thermal bridging in the concrete connections between cores continues to depreciate the overall *R*-value. This approach is only a partial solution to providing a quality, well-insulating wall. Other techniques, as explained in the next few pages, provide more cost-effective solutions.

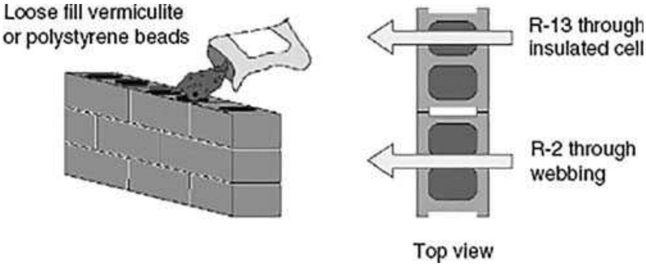


FIGURE 2 Insulating concrete block cores (R-4–R-6 overall).

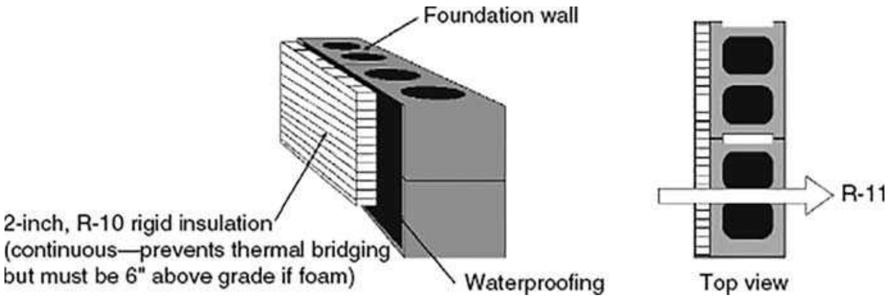


FIGURE 3 Exterior foam insulation (R-11–R-12 overall).

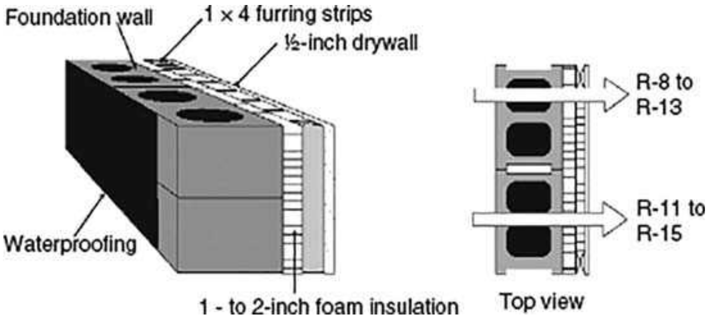


FIGURE 4 Interior foam insulation (R-10–R-14 overall).

Exterior Rigid Fiber Glass or Foam Insulation

Rigid insulation is generally more expensive per *R*-value than mineral wool or cellulose, but its rigidity is a major advantage (Figure 3). However, it is difficult and expensive to obtain R-values as high as in framed walls.

Interior Foam Wall Insulation

Foam insulation can be installed on the interior of concrete block walls (Figure 4); however, it must be covered with a material that resists damage and meets local fire code requirements. Half-inch drywall will typically comply, but furring strips will need to be installed as nailing surfaces. Furring strips are usually installed between sheets of foam insulation; however, to avoid the direct, uninsulated thermal bridge between the concrete wall and the furring strips, a continuous layer of foam should be installed underneath or on top of the nailing strips.

Interior Framed Wall

In some cases, designers will specify a framed wall on the interior of a masonry wall (Figure 5). Standard framed wall insulation and air-sealing practice can then be applied.

Insulated Concrete Form Systems

Insulated concrete forms (ICFs) are permanent rigid plastic foam forms that are filled with reinforced concrete to create structural walls with significant thermal insulation (Figure 6). The foam is typically

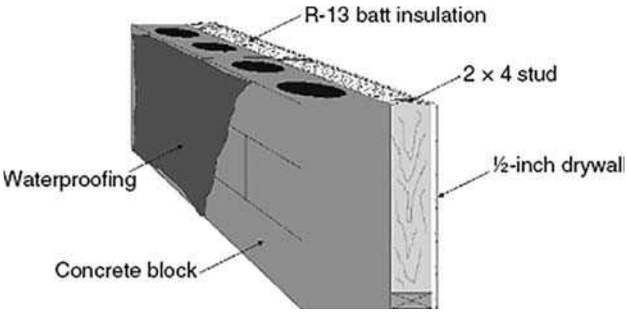


FIGURE 5 Interior framed wall insulation (R-11–R-13 overall).

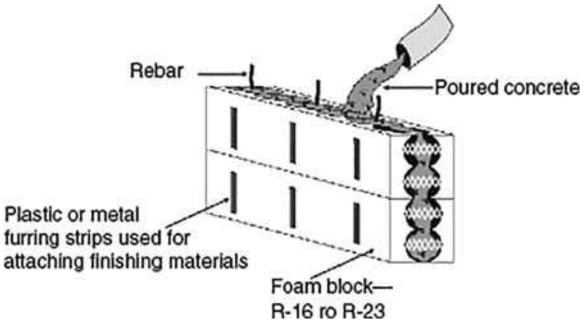


FIGURE 6 Insulated concrete foam system (R-17–R-24 overall).

either expanded polystyrene (EPS) or extruded polystyrene (XPS) and occasionally polyurethane, but it may also be made from a composite of cement and foam insulation or a composite of cement and processed wood.[5]

The concrete will be one of several shapes: flat, waffle-or screen-grid, or post-and-beam, depending on the specific form design. The Portland Cement Association (PCA) reports that in 1994, 0.1% of all new homes used ICFs in above-grade walls (about 1100 new homes). That number rose to 1.2% in 1999, 2.7% in 2001, and increased to 3.8% in 2002, which would be 50,639 homes, according to U.S. Census Data.

Above-grade ICF walls cost more to build than typical wood-framed walls. As wood-framed walls approach the thermal insulation value of ICFs, the cost differential will decrease. In most cases, materials' costs (concrete and forms) are primarily responsible for increased costs, while labor costs are often similar to wood framing. Cost premium depends on relative material prices, labor efficiency for each system, engineering necessity, and its effect on other practices or trades, among other factors.

The cost premium for ICF houses is smaller in areas such as high-wind regions, which require additional labor, time, and materials for special construction of woodframed houses. According to an NAHB Research Center study, costs are estimated to increase by 1%–8% of total house cost over a wood-framed house.

Lightweight Concrete Products

Lightweight, air-entrained concrete is an alternative wall system (Figure 7). Autoclaved aerated concrete (AAC), sometimes referred to as precast autoclaved aerated concrete (PAAC), which can be shipped either as blocks or panels, combines elevated *R*-values (compared to standard concrete) with thermal mass.

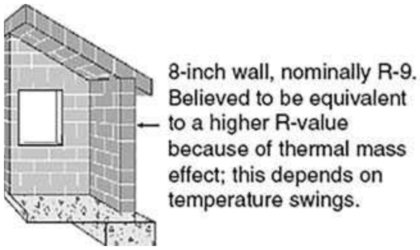


FIGURE 7 Lightweight concrete products (R-1.1 per inch plus).

TABLE 3 2 × 4 Framed Wall Problems and Solutions

Problem	Solution
Small space available for insulation	Install continuous exterior foam sheathing and medium (R-13) to high (R-15) density cavity insulation
Enclosed cavities are more prone to cause condensation, particularly when sheathing materials with low R-values are used	Install a continuous air barrier system. Use continuous foam sheathing
Presence of wiring, plumbing, ductwork, and framing members lessens potential R-value and provides pathways for air leakage	Locate mechanical systems in interior walls; avoid horizontal wiring runs through exterior walls; use air sealing insulation system

2 × 4 Wall Insulation

Throughout the United States, debates on optimal wall construction continue.[6] Table 3 summarizes typical problems and solutions in walls framed with 2 × 4 studs. In addition to standard framing lumber and fasteners, the following materials will also be required during construction:

- Foam sheathing for insulating headers.
- 1 × 4 or metal T-bracing for corner bracing (Figure 8).

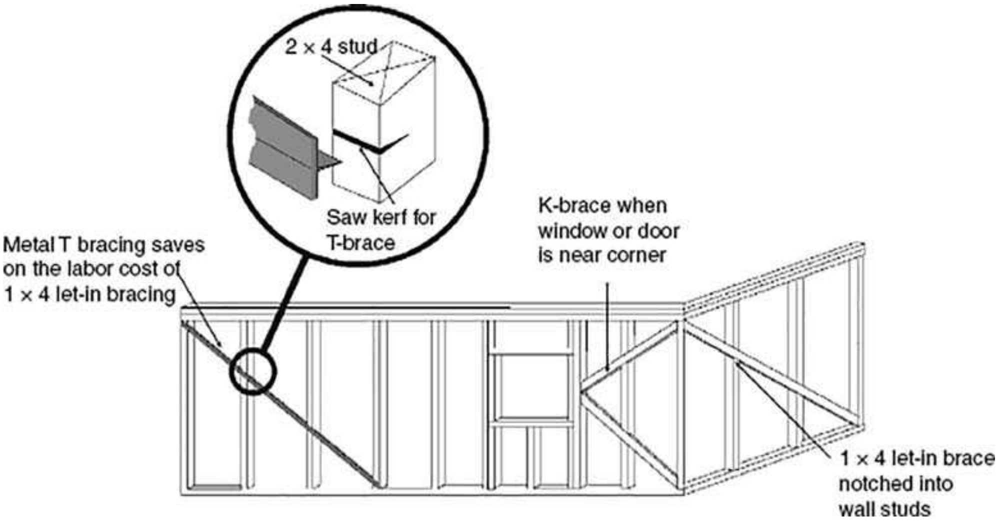


FIGURE 8 Let-in bracing.

- R-13 batts for insulating areas during framing behind shower/tub enclosures and other hidden areas.
- From the Florida Building Code, Residential (FBC-R): "R307.2 Bathtub and shower spaces. Bathtub and shower floors and walls above bathtubs with installed shower heads and in shower compartments shall be finished with a nonabsorbent surface. Such wall surfaces shall extend to a height of not less than 6 ft above the floor."
- Caulking or foam sealant for sealing areas that may be more difficult to seal later.

Avoid Side Stapling

Walls are usually insulated with batts that have attached vapor retarder facing. Many builders question whether it is best to side staple or face staple batt insulation. The common arguments are that face stapling results in less compression, while side stapling interferes less with drywall installation.

The ideal solution should focus on where the kraft paper (vapor retarder) is rather than on how it is installed.[7]

The face stapling question is an appropriate question in northern or "heating-dominated" climates. In northern areas, vapor retarders should be installed on the "warm" side of the wall cavity. In southern or "cooling-dominated" climates, the vapor retarder should be on the outside surface of the wall cavity. Because of this, the use of unfaced batts is recommended in hot-humid climates (Figure 9).

Unfaced batts are slightly larger than the standard 16-or 24-inch stud spacing and rely on a friction-fit for support. Because unfaced batts are not stapled, they can often be installed in less time. In addition, it is easier to cut unfaced batts to fit around wiring, plumbing, and other obstructions in the walls.

Blown Loose-Fill Insulation

Loose-fill cellulose, fiberglass, and rock wool insulation can also be used to insulate walls. This insulation is installed with a blowing machine and held in place with a glue binder or netting (Figure 10). This technique can provide good insulation coverage in the stud cavities; however, it is very important that excess moisture in the binder be allowed to evaporate before the wall cavities are enclosed by an interior finish. Keep in mind that insulation products are not replacements for proper air sealing technique.

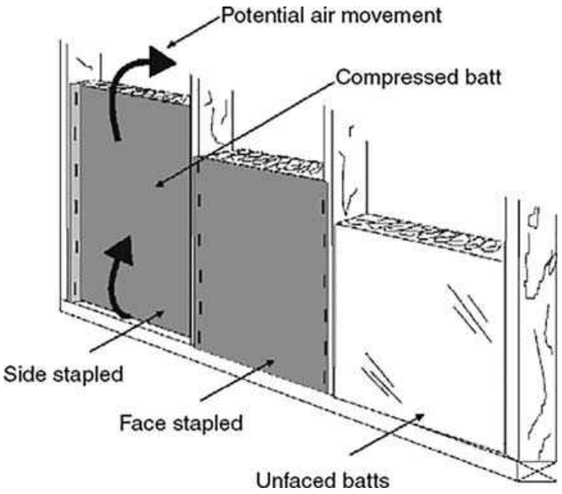


FIGURE 9 Insulating walls with batts.

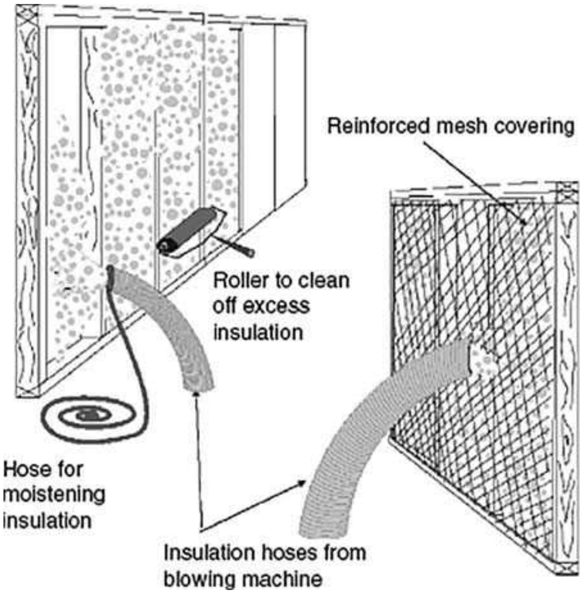


FIGURE 10 Blown sidewall insulation options.

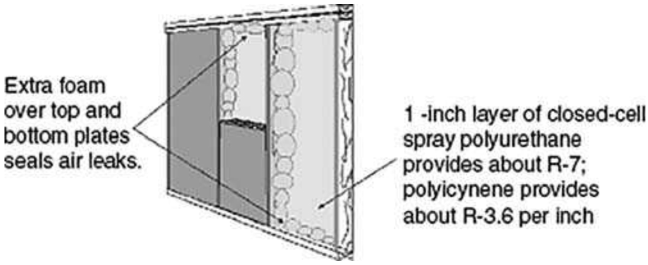


FIGURE 11 Blown foam insulation.

Blown Foam Insulation

Some insulation contractors are now blowing polyurethane or polyicynene insulation into the walls and ceilings of new buildings (Figure 11). This technique provides high *R*-values in relatively thin spaces and seals air leaks effectively. The economics of foam insulation should be examined carefully prior to their application.

Structural Insulated Panels

Another approach to wall construction is the use of structural insulated panels (SIP), also known as stress-skin panels (Figure 12).[8] They consist of 4-inch or 6-inch thick foam panels onto which structural sheathing, such as oriented strand board (OSB), cement fiber board, or various types of metal have been attached. They reduce labor costs, and because of the reduced framing in the wall, they have higher *R*-values and less air leakage than standard walls.

SIPs are generally 4 feet wide and 8 to 12 feet long. There are a wide variety of manufacturers, each with its own method of attaching panels together. Procedures for installing windows, doors, wiring, and plumbing have been worked out by each manufacturer. Some SIPs come from the factory with

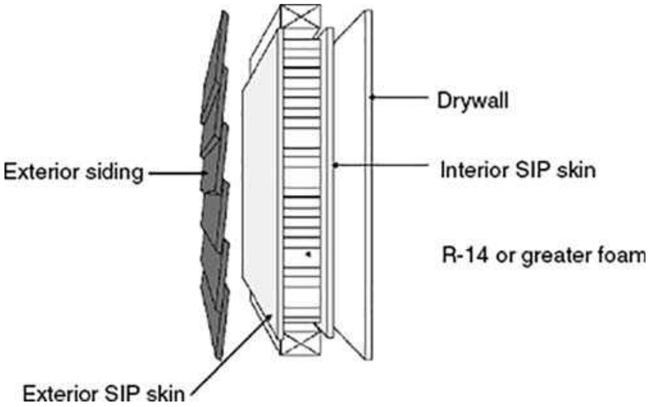


FIGURE 12 Structural insulated panels (SIP)

preinstalled windows. In addition to their use as wall framing, SIPs can also be used in ceilings, floors, and roofs.

While buildings constructed with SIPs may be more expensive than those with standard framed and insulated walls, research studies have shown that SIP-built buildings have higher average insulating values per inch than most commonly used insulation materials. Due to their typical modular style of construction, infiltration losses are also reduced. Thus, they can provide substantial energy savings. Be sure to follow local building codes with regard to termites, including leaving a 6" inspection zone above finished grade.

The performance of any SIP wall depends on its component materials and installation processes. There are a few important variables to take into consideration when building with SIP systems:

1. Panel fabrication (proper panel gluing, pressing, and curing) is critical to prevent delamination.
2. Panels must be flat, plumb, and have well-designed connections to ensure tightness of construction.
3. Though SIPs offer ease of construction, installers may need training in installing the system being used.
4. Fire rating of SIP materials and air-tightness of SIP installation affect the system's fire safety.
5. There may be potential insect and rodent mitigation issues, depending on SIP materials and construction.
6. Proper HVAC design and installation must take the SIP system being used into account.

Metal Framing

Builders and designers are well aware of the increasing cost and decreasing quality of framing lumber. As a consequence, interest in alternative framing materials, such as metal framing, has grown. While metal framing offers advantages over wood, such as consistency of dimensions, lack of warping, and resistance to moisture and insect problems, it has distinct disadvantages from an energy perspective.

Metal framing is an excellent conductor of heat. Buildings framed with metal studs and plates usually have metal ceiling joists and rafters, as well. Thus, the entire structure serves as a highly conductive thermal grid. Insulation placed between metal studs and joists is much less effective due to the extreme thermal bridging that occurs across the framing members.

The American Iron and Steel Institute is well aware of the challenges involved in building an energy efficient steel structure. In their publication *Thermal Design Guide for Exterior Walls* (Publication RG-9405), the Institute provides information on the thermal performance of steel-framed buildings. Table 4 summarizes some of their findings.

TABLE 4 Effective Steel Wall *R*-Values

Cavity Insulation	Sheathing	Effective Overall R-Value
11	2.5	9.5
11	5	13
11	10	18
13	2.5	10
13	5	14

Moisture-related problems have been reported in metal frame buildings that do not use sufficient insulated sheathing on exterior walls. Metal studs cooled by the air conditioning system can cause outdoor air to condense, leading to mildew streaks (or ghosting), where one can see the framing members on the inside and outside of a home. In winter, studs covered by cold outside air can also cause streaking. Attention to proper insulation techniques can alleviate this problem.

2 × 6 Wall Construction

There has been interest in hot, humid climates in the use of 2 × 6s for construction. The advantages of using wider wall framing are:

- More space provides room for R-19 or R-21 wall insulation.
- Thermal bridging across the studs is less of a penalty due to the higher *R*-value of 2 × 6s.
- Less framing reduces labor and material costs.
- There is more space for insulating around piping, wiring, and ductwork.

Disadvantages of 2 × 6 framing include:

- Wider spacing may cause the interior finish or exterior siding to bow slightly between studs.
- Window and door jambs must be deeper, resulting in additional costs.
- Walls with substantial window and door area may require almost as much framing as 2 × 4 walls and leave relatively little area for actual insulation.

The economics of 2 × 6 wall insulation are affected by the number of windows in the wall because each window opening adds extra studs and may require the purchase of a jamb extender. Walls built with 2 × 6s having few windows provide positive economic payback. However, for walls in which windows make up over 10% of the total area, the economics become questionable.

Ceilings and Roofs

Attics over flat ceilings are usually the easiest part of a building's exterior envelope to insulate. They are accessible and have ample room for insulation. However, many homes use cathedral ceilings that provide little space for insulation.[9] It is important to insulate both types of ceilings properly.

Attic Ventilation

In the summer, properly designed ventilation reduces roof and ceiling temperatures, thus potentially saving on cooling costs and lengthening the life of the roof. In winter, roof vents expel moisture which could otherwise accumulate and deteriorate insulation or other building materials.

At present, several research studies are investigating whether attic ventilation is beneficial. For years, researchers have believed the cooling benefits of ventilating a well-insulated attic to be negligible. However, some experts are now questioning whether ventilation is even effective at moisture

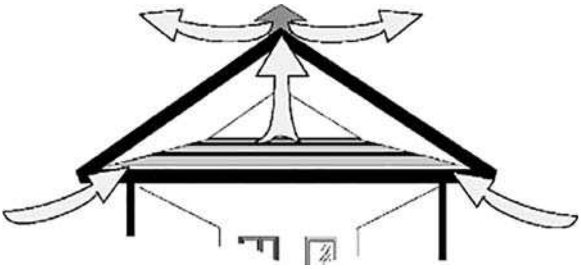


FIGURE 13 Attic ventilation through soffit and ridge vents.

removal. The Florida Building Code, Residential now provides provisions for "conditioned attic assemblies" (Section R806.4) as long as certain conditions are met. When attic ventilation is provided, ventilation openings shall be provided with corrosion-resistant wire mesh with 1/8 in. minimum to 1/4 in. maximum openings. Total net free ventilation area shall not be less than 1–150 of the area of the space ventilated. An exception for 1–300 is provided in the code.

Vent Selection

The amount of attic ventilation needed is based on state building code requirements. If ventilating the roof, locate vents high along the roof ridge and low along the eave or soffit. Vents should provide air movement across the entire roof area (Figure 13). There are a wide variety of products available, including ridge, gable, soffit, and mushroom vents.

To allow for proper airflow in attic spaces, it is common practice to install a rafter baffle at the soffit. This will prevent insulation from sealing off the airflow from the soffit vent to the attic space.

The combination of continuous ridge vents along the peak of the roof and continuous soffit vents at the eave provides the most effective ventilation. Ridge vents come in a variety of colors to match any roof. Some brands are made of plastic and covered by cap shingles to hide the vent from view.

Manufacturer or product testing is being performed by a variety of organizations to verify leak-free operation of continuous ridge vents in high wind situations. Care should be taken to ensure that the vents chosen are appropriate for hurricane-prone areas.

Powered Attic Ventilator

Electrically powered roof ventilators can consume more electricity to operate than they save on air conditioning costs and are not recommended for most designs. Power vents can create negative pressures in the home, which may have detrimental effects, such as (Figure 14):

- Drawing outside air into the home
- Removing conditioned air from the home through ceiling leaks and bypasses
- Pulling pollutants, such as radon and sewer gases, into the home
- Backdrafting fireplaces and fuel-burning appliances

Attic Floor Insulation Techniques

Either loose-fill or batt insulation can be installed on an attic floor. Generally, blowing loose-fill attic insulation is usually less expensive than installing batts or rolls. Most attics have either blown fiberglass, rock wool, or cellulose. Ceilings with a rise greater than 5 and a run of 12 (5 over 12) should not be insulated with blown-in insulation.

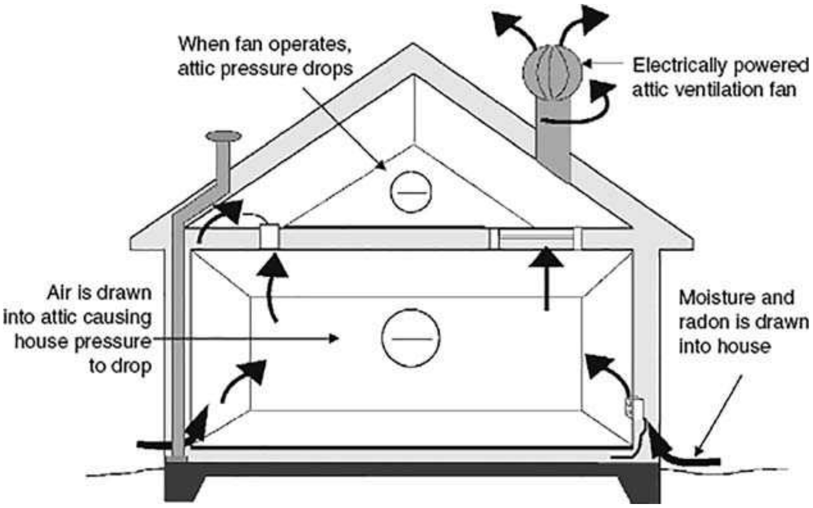


FIGURE 14 Attic ventilation through powered ventilation.

Loose-Fill Attic Insulation

Steps for installing loose-fill attic insulation[10]:

7. Seal attic air leaks, as prescribed by fire and energy codes.
8. Follow manufacturer's and state building code clearance requirements for heat-producing equipment found in an attic, such as flues or exhaust fans. One example of attic blocking is shown later in this entry.
9. Use baffles to preserve ventilation space at eave of roof for soffit vents.
10. Insulate the attic hatch or attic stair. Foam boxes are available for providing a degree of insulation over a pull-down attic stairway.
11. Determine the attic insulation area; based on the spacing and size of the joists, use the chart on the insulation bag to determine the number of bags to install. Table 5 shows a sample chart for cellulose insulation. Cellulose is heavier than fiberglass for the same *R*-value. Closer spacing of roof joists and thicker dry-wall is required for larger *R*-values. Check this detail with the insulation contractor. Weight limits and other factors at R-38 insulation levels are shown in Table 5 for the three primary types of loose fills.
12. Avoid fluffing the insulation (blowing with too much air) by using the proper air-to-insulation mixture in the blowing machine. A few insulation contractors have "fluffed" (added extra air to) loose-fill insulation to give the impression of a high *R*-value. The insulation may be the proper depth, but if too few bags are installed, the *R*-values will be less than claimed.

TABLE 5 Blown Cellulose in Attics

			2 × 6 Joists Spaced 24 in. on Center		2 × 6 Joists Spaced 16 in. on Center	
R-Value at 75°F	Minimum Thickness (in.)	Minimum Weight (lb/ft2)	Coverage per 25-lb Bag (ft2)	Bags per 1000 ft2	Coverage per 25-lb Bag (ft2)	Bags per 1000 ft2
R-40	10.8	2.10	12	83	13	77
R-32	8.6	1.60	16	63	18	56
R-24	6.5	0.98	21	48	23	43
R-19	5.1	0.67	37	27	41	24

13. Obtain complete coverage of the blown insulation at relatively even insulation depths. Use attic rulers (obtainable from insulation contractors) to ensure uniform depth of insulation.

Batt Attic Insulation

Steps for installing batt insulation:

14. Seal attic air leaks, as prescribed by fire and energy codes.
15. Block around heat-producing devices, as described in Step 2 for loose-fill insulation.
16. Insulate the attic hatch or attic stair as described in Step 4 for loose-fill insulation.
17. Determine the attic insulation area based on the spacing and size of the joists, order sufficient R-30 insulation for the flat attic floor. Choose batts that are tapered—cut wider on top—so that they cover the top of the ceiling joists. (See Figure 15).
18. When installing the batts, make certain they completely fill the joist cavities. Shake batts to ensure proper loft. If the joist spacing is uneven, patch gaps in the insulation with scrap pieces. Try not to compress the insulation with wiring, plumbing, or ductwork. In general, obtain complete coverage of full-thickness, noncompressed insulation.
19. Attic storage areas can pose a problem. If the ceiling joists are shallower than the depth of the insulation (generally less than 2 ×10s), raise the finished floor using 2 × 4s or other spacing lumber. Install the batts before nailing the storage floor in place (see Figure 16).

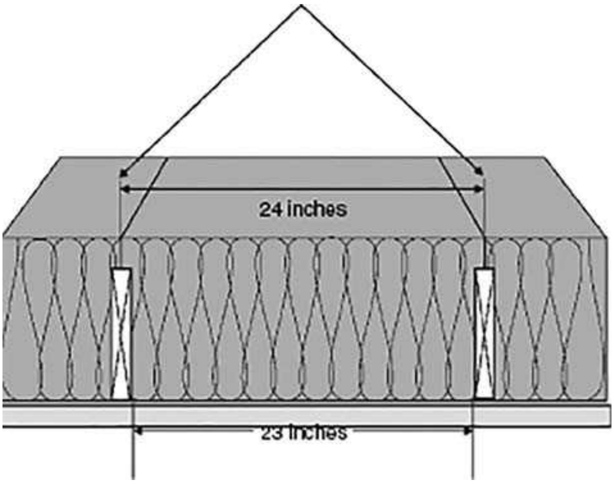


FIGURE 15 Full-width ceiling batt insulation.

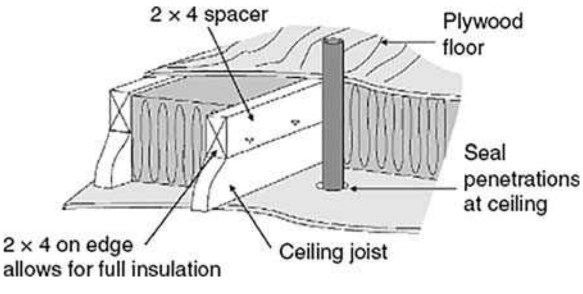


FIGURE 16 Ceiling insulation under attic storage floor.

Note that often attic framing is not designed for storage. Check engineered loads of framing before increasing loads and piggy-backing ceiling joists.

Preventing Air Flow Restrictions at the Eave

One problem area in many standard roof designs is at the eave, where there is not enough room for full R-30 insulation without preventing air flow from the soffit vent or compressing the insulation, which reduces its *R*-value. Figures. 17 and 18 show several solutions to this problem. If using a truss roof, purchase raised heel trusses that form horizontal overhangs. They should provide adequate clearance for both ventilation and insulation.

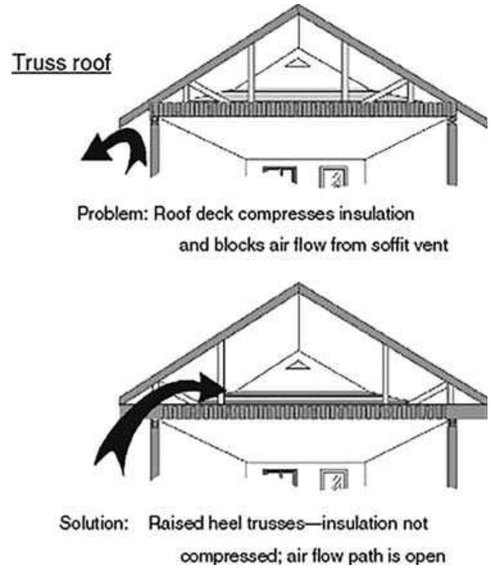


FIGURE 17 Soffit air ventilation—raised heel trusses.

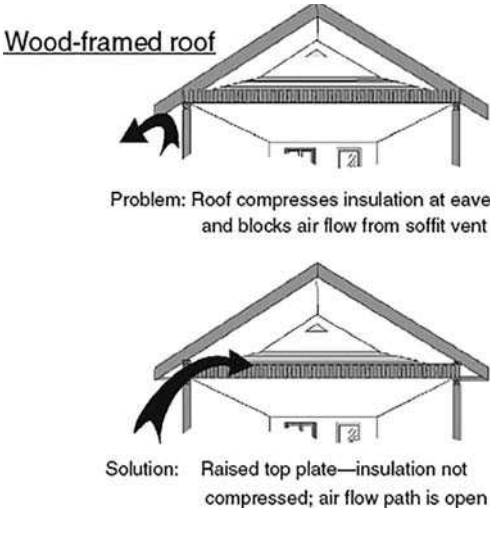


FIGURE 18 Soffit air ventilation—raised top plate.

In stick-built roofs, where rafters and ceiling joists are cut and installed on the construction site, an additional top plate that lies across the top of the ceiling joists at the eave will prevent compression of the attic insulation. Note: This needs to be a double plate for bearing unless rafters sit directly above joists. The rafters sitting on this raised top plate allow for both insulation and ventilation.

Cathedral Ceiling Insulation Techniques

Cathedral ceilings are a special case because of the limited space for insulation and ventilation within the depth of the rafter. Fitting in a 10-in. batt (R-30) and still providing ventilation is impossible with a 2 × 6 or 2 × 8 rafter (R-19 or R-25, respectively).

Building R-30 Cathedral Ceilings

Cathedral ceilings built with 2 × 12 rafters can be insulated with standard R-30 batts and still have plenty of space for ventilation. Some builders use a vent baffle between the insulation and roof decking to ensure that the ventilation channel is maintained.

If 2 × 12s are not required structurally, most builders find it cheaper to construct cathedral ceilings with 2 × 10 rafters and high-density R-30 batts, which are $8^{1/4}$ in. thick (Table 6).

Some contractors wish to avoid the higher cost of 2 × 10 lumber and use 2 × 8 rafters. These roofs are usually insulated with R-19 batts.

In framing with 2 × 6 and 2 × 8 rafters, higher insulating values can be obtained by installing rigid foam insulation under the rafters. Note that the rigid foam insulation must be covered with a fire-rated material when used on the interior of the building. Drywall usually meets this requirement.

Scissor Trusses

Scissor trusses are another cathedral ceiling framing option. Make certain they provide adequate room for both R-30 insulation and ventilation, especially at their ends, which form the eave section of the roof.

Any sized rafter; blown-in cellulose, fiberglass, or rock wool held in place; provide 1 in. ventilation space above.

Difficulties with Exposed Rafters

A cathedral ceiling with exposed rafters or roof decking is difficult and expensive to insulate well. Often, foam insulation panels are used over the attic deck, as shown in Figure 19. However, to achieve R-30, four to seven inches of foam insulation are needed. Ventilation is also a problem.

In homes where exposed rafters are desired, it may be more economical to build a standard, energy efficient cathedral ceiling, and then add exposed decorative beams underneath. Note that homes having tongue-and-groove ceilings can experience substantially more air leakage than solid, drywall ceilings. Install a continuous air barrier, sealed to the walls above the tongue-and-groove roof deck and held in place; provide 1 in. ventilation space above.

TABLE 6 Cathedral Ceiling Insulation Options

Rafter	Batt
2 × 8	R-19
2 × 10	R-25
2 ×10	Moderate density R-30
2 × 12	Standard density R-30

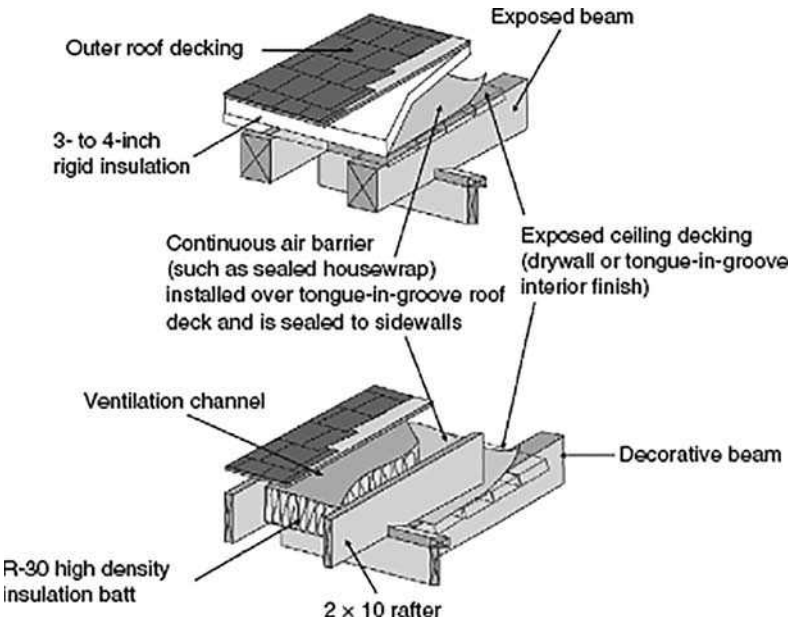


FIGURE 19 Insulating exposed rafters.

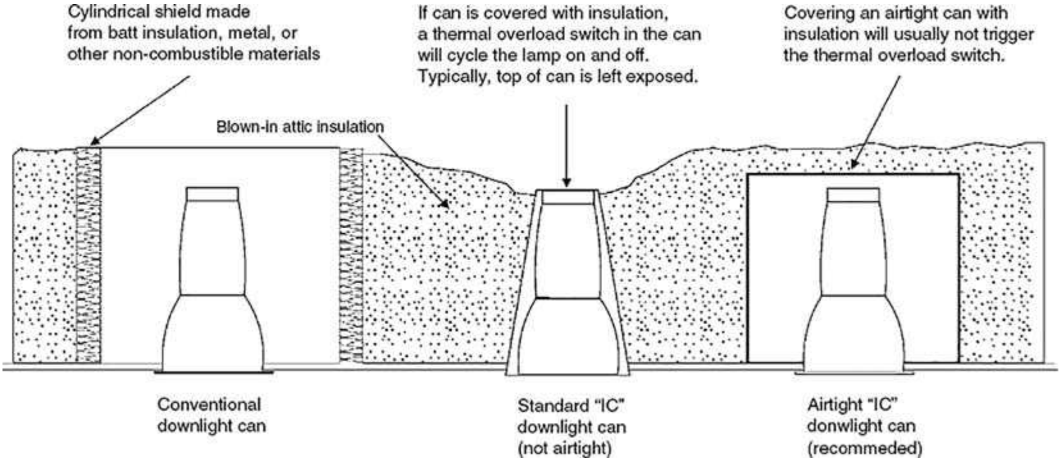


FIGURE 20 Recessed lighting insulation.

Recessed Lights

Standard recessed fixtures require a clearance of several inches between the sides of the lamp's housing and the attic insulation. In addition, insulation cannot be placed over the fixture. Even worse, recessed lights leak considerable air between attics and the home.

Insulated ceiling (IC) rated fixtures have a heat sensor switch that allows the fixture to be covered—except for the top—with insulation (see Figure 20 for the proper insulation methods for these fixtures). However, these units also leak air. If you have to use recessed lights, install airtight IC-rated fixtures. There are alternatives to recessed lights, including surface-mounted ceiling fixtures and track lighting, which typically contribute less air leakage to the home.

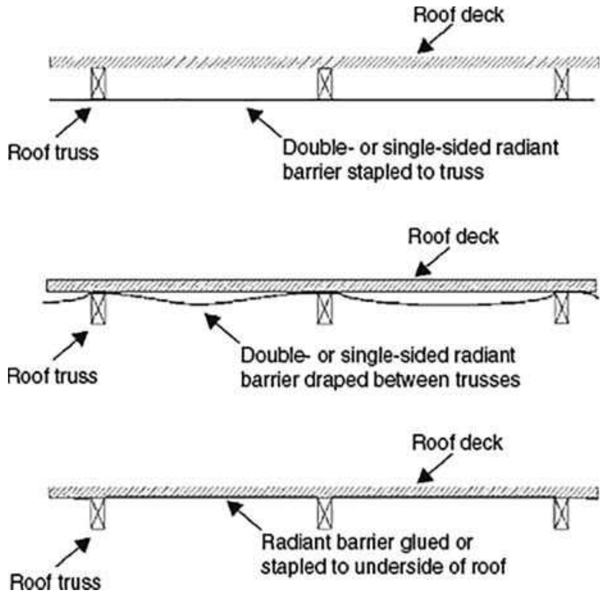


FIGURE 21 Radiant barrier configuration

Radiant Heat Barriers

Radiant heat barriers (RHB) are reflective materials that can reduce summer heat gain via the insulation and building materials in attics and walls. RHBs work two ways: first, they reflect thermal radiation well; and, second, they emit (give off) very little heat. RHBs should always face a vented airspace and be installed to prevent dust build-up. They are usually attached to the underside of the rafter or truss top chord or to the underside of the roof decking. Acceptable attic radiant barrier configurations can be found in Figure 21.

How Radiant Barrier Systems Work

A radiant barrier reduces heat transfer. Thermal radiation, or radiant heat, travels in a straight line away from a hot surface and heats any object in its path.

When sunshine heats a roof, most of the heat conducts through the exterior roofing materials to the inside surface of the roof sheathing. Heat then transfers by radiation across the attic space to the next material—either the top of the attic insulation or the attic floor. A radiant barrier, properly installed in one of many locations between the roof surface and the attic floor, will reduce radiant heat flow. Thermal insulation on the attic floor resists the flow of heat through the ceiling into the living space below. The rate at which insulation resists this flow determines the insulation's *R*-value. The amount of thermal insulation affects the potential radiant barrier energy savings. For example, installing a radiant barrier in an attic that already has high levels of insulation (*R*-30 or above) would result in much lower energy savings than an attic insulated at a low level (R-11 or less).

All radiant barriers use reflective foil that blocks radiant heat transfer. In an attic, a radiant barrier that faces an air space can block up to 95% of the heat radiating down from a hot roof. Only a single, thin, reflective surface is necessary to produce this reduction in radiant heat transfer. Additional layers of foil do little more to reduce the remaining radiant heat flow.

Conventional types of insulation consist of fibers or cells that trap air or contain a gas to retard heat conduction. These types of insulation reduce conductive and radiant heat transfer at a rate determined by their

R-value, while radiant barriers reduce only radiant heat transfer. There is no current method for assigning an *R*-value to radiant barriers. The reduction in heat flow achieved by the installation of a radiant barrier depends on a number of factors, such as ventilation rates, roof reflectivity, ambient air temperatures, geographical location, the amount of roof solar gains, and the amount of conventional insulation present.

Several factors affect the cost effectiveness of installing a radiant barrier. You should examine the performance and cost savings of at least three potential insulation options: adding additional conventional insulation, installing a radiant barrier, and adding both conventional insulation and a radiant barrier.

In 1991 (revised June, 2001), the U.S. Department of Energy (DOE) published the Radiant Barrier Attic Fact Sheet, which shows how to calculate the economics of radiant barriers and added ceiling insulation. It includes an Energy Savings Worksheet with an example. The worksheet is part of the fact sheet and can be found at http://www.ornl.gov/sci/roofs+walls/radiant/rb_05.html.[11]

Because radiant barriers redirect radiant heat back through the roofing materials, shingle temperatures may increase between 1 and 10 F (0.6 C-5.6 C). This increase does not appear to exceed the roof shingle design criteria. The overall effect on roof life, if any, is not known.

Remember, radiant barriers are most effective in blocking summer radiant heat gain and saving air-conditioning costs. Although the radiant barrier may be somewhat effective in retaining heat within a cold-climate home, it may also block any radiant winter solar heat gain in the attic.

Conclusion

Buildings are insulated to help moderate the environment that we live and work in. As utility costs rise, this aspect of our constructed environment becomes more and more important. Many new materials have entered the market in the past few years, each having their own particular advantages and disadvantages. This entry has described in detail the importance of installation practices and their effect on overall performance. Emphasis was placed on properly defining the building envelope and different methods and materials that can be used to effectively insulate and, therefore, thermally isolate the constructed environment from the daily extremes produced by local weather.

New techniques and materials will evolve; however, some guidelines will remain the same:

- Choose an insulation material with characteristics suitable for the climate region.
- Use the proper R-value for the climate region where each particular structure is being constructed.
- Apply the recommended thickness, or R-value, to each building component, such as ceiling, walls, and floors.
- Install each particular insulation product in a proper fashion, following the manufacturer's recommendations and instructions, leaving no voids, and filling each building cavity to the level required for the building site.

Insulation materials and techniques play a large part in minimizing energy consumption and maximizing human comfort. Recent advances in both materials and techniques will help further this trend.

Acknowledgments

Many thanks to Kyle Allen, an engineering student at the University of Florida, because without his persistent efforts, the publication of this entry would not have been possible.

References

1. *Energy Efficient Building Construction in Florida*, Florida Energy Extension Service, University of Florida, 2005. Adapted with permission, Gainesville, FL.
2. Jeffrey, T.; Dennis, C. *A Builder's Guide to Energy Efficient Homes in Georgia*; Georgia Environmental Facilities Authority; Division of Energy Resources: Atlanta, GA, 1999.

3. Environmental Building News. 2005. *Insulation Materials— Summary of Environmental and Health Considerations.* 122 Birge Street, Suite 30, Brattleboro, VT 05301. Phone: (800) 861–0954. Web site: http://www.BuildingGreen.com.
4. Federal Trade Commission. *Labeling and Advertising of Home Insulation (R-Value Rule) 16CFR460.* Available online at: http://www.ftc.gov/bcp/rulemaking/rvalue/16cfr460.htm Federal Register Notice (amendments) http://www3.ftc.gov/os/2005/05/05031homeinsulationfrn.pdf.
5. ToolBase Services. *Insulating Concrete Forms.* http://www.toolbase.org/techinv/techDetails.aspx?technologyIDZ97.
6. Oak Ridge National Laboratory, *Wall Insulation.* October 2000. Available online at: http://www.eere.energy.gov/buildings/info/documents/pdfs/26451.pdf.
7. Southface Energy Institute. March 2002. *Wall Insulation.* Available online at: http://www.southface.org/web/resources&services/publications/factsheets/26_insulatewalls4PDF.pdf.
8. Knowles, H. 2005. *Performance Under Pressure: Structural Insulated Panel (SIP) Walls.* http://www.energy.ufl.edu/factsheets.htm.
9. Desjarlais, A.O.; Petrie, T.W.; Stovall, T. In *Comparison of Cathedralized Attics to Conventional Attics: Where and* When do Cathedralized Attics Save Energy and Operating Costs? *Performance of Exterior Envelopes of Whole Buildings, IX International Conference ASHRAE: Clearwater, FL,* 2004.
10. Oak Ridge National Laboratory, *Ceilings and Attics.* February 2000. Available online at: http://eber.ed.ornl.gov/DOE-ceilingsattics_771.pdf.
11. U.S. Department of Energy. *Radiant Barrier Attic Fact Sheet.* Available online at: http://www.ornl.gov/sci/roofsC-walls/radiant/rb_05.html.

Bibliography

1. Air Conditioning Contractors of America. Appendix A-5, detailed infiltration estimate. In *Manual J, Load Calculation for Residential Winter and Summer Air Conditioning*, 8th Ed.; May be ordered online at: http://www.accaconference.com/Merchant2/merchant.mv?ScreenZPROD&Store_Code ZACCOA&Product_CodeZ33-8&Category_CodeZ.
2. Jeffrey, T.; Dennis, C. *Builder's Guide to Energy Efficient Homes in Louisiana*; Louisiana Department of Natural Resources: Louisiana, 2002.
3. Insulation Contractors Association of America. *Technical Bulletin No. 18, Reflective Insulations and Radiant Barriers.* Alexandria, VA, July 1997.
4. Oak Ridge National Laboratory, *Attic Access.* February 2000. Available online at: http://eber.ed.ornl.gov/Atticaccess_DOE_GO10099–768.pdf.
5. Office of Energy Efficiency and Renewable Energy (EERE) 1–877–337–3463 (EERE-INF) http://www.eere.energy.gov and http://www.eere.energy.gov/consumer/.
6. U.S. Department of Energy. *Insulation and Weatherization.* Available online at: http://www.pueblo.gsa.gov/cic_text/housing/energy-savers/insulation.html.
7. U.S. Department of Energy. *Insulation Fact Sheet.* Available online at: http://www.ornl.gov/roofsCwalls/insulation.
8. U.S. Department of Housing and Urban Development, *Volume 2—The Rehab Guide: Exterior Walls.* Washington, DC, August 1999. HUD-DU100C000005956 Available online at: http://www.huduser.org/publications/destech/walls.html.
9. U.S. Department of Housing and Urban Development, *Volume 3—The Rehab Guide: Roofs.* Washington, DC, March 1999. HUD-DU100C000005956, Available online at: http://www.huduser.org/publications/destech/roofs.html.

IV

DIA: Diagnostic Tools: Monitoring, Ecological Modeling, Ecological Indicators, and Ecological Services

24

Environmental Accounting: A Tool for Supporting Environmental Management and Nature Conservation

Pier Paolo Franzese, Elvira Buonocore, and Giovanni F. Russo

Environmental Management and Nature Conservation

Healthy ecosystems are capable of maintaining their structures and functions, ensuring the generation and maintenance of natural capital stocks and ecosystem services flows (Buonocore et al., 2018; Pauna et al., 2018).

There is a growing evidence that biodiversity increases the stability of ecosystem functioning, representing the basis for the generation of ecosystem services vital for human well-being (Teixeira et al., 2019; Vihervaara et al., 2019). Therefore, biodiversity loss represents one of the major threats to humanity with potential impacts on both nature and human economy (Cardinale et al., 2012; TEEB, 2010).

In line with the objectives of the Convention on Biological Diversity (CBD) striving to address the issue of biodiversity loss, ensure the sustainable use of natural resources, and allow an equitable share of benefits, there has been significant progress toward conservation targets (UNEP-WCMC, IUCN, and NGS, 2018). Nonetheless, efforts are still needed to increase the effectiveness of nature conservation actions (Hoffmann et al., 2018).

Many international policy processes, including the CBD and the UN 2030 Agenda for Sustainable Development, recognize protected areas as successful management tools for protecting biodiversity while ensuring the sustainable exploitation of natural resources. When effective management measures are in place, protected areas are capable of meeting the multitude of objectives they are designed for, supporting the achievement of both local and large-scale sustainability goals.

Over the past decade, there have been increasing research efforts to assess the value of natural resources, also exploring how these values can be embedded into decision making (Costanza et al., 2014; European Union, 2014; Franzese et al., 2014, 2019).

Assessing the biophysical and economic value of natural capital and ecosystem services provides a deeper understanding of ecosystems functioning while supporting managers and policy makers in charge for implementing management plans and policies rooted in the principle of sustainable development (Maes et al., 2016; TEEB, 2010).

These assessments can be particularly useful in the case of protected areas to establish a baseline to monitor changes over time, assess the consequences of management decisions and changes, and demonstrate their importance in achieving international conservation targets (Neugarten et al., 2018).

Environmental Accounting

Environmental accounting is a tool useful for exploring three main dimensions related to the exploitation of natural resources both in the context of nature and human economy: environmental costs, impacts, and benefits (Häyhä and Franzese, 2014).

In particular, when applied in the context of nature conservation, environmental accounting is useful to address multiple aspects dealing with the exploitation of natural resources, among which (1) the assessment of environmental costs sustained by ecosystems for the generation and maintenance of natural capital stocks, (2) the assessment of goods and services that humans receive from ecosystems (i.e., the ecosystem services), and (3) the assessment of the impacts generated by human activities for the exploitation of natural resources (Figure 1).

Environmental accounting takes into account both direct and indirect environmental costs as well as local and global impacts. In general, environmental accounting methods can be assigned to two broad categories: (1) upstream methods, focusing on the cumulative amount of environmental resources supporting the investigated system, and (2) downstream methods, more concerned with the potential environmental impacts due to the system's emissions.

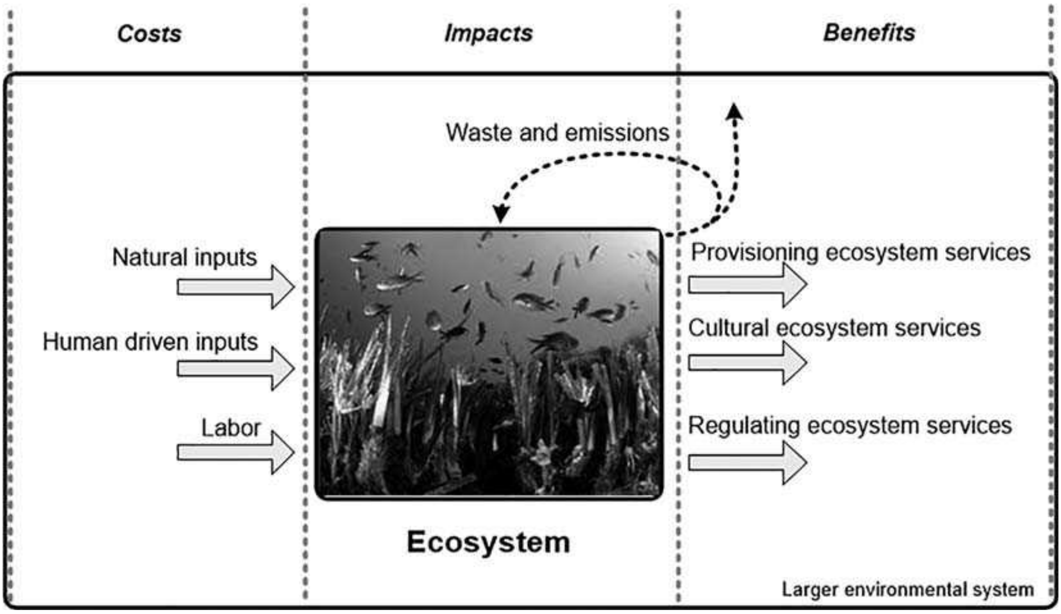


FIGURE 1 Three main dimensions of environmental accounting: assessment of environmental costs, impacts, and benefits.

The Emergy Accounting Method

The emergy accounting method (Odum, 1988, 1996) aims at calculating indicators of environmental performance and sustainability accounting for both natural and economic resources used up within ecosystems and human-dominated process.

Emergy is a "donor-side" approach assessing the value of goods and services in terms of work of biosphere invested for their generation. According to this method, the greater the cumulative environmental support to generate natural resources, the higher is their emergy value.

The emergy method takes into account: (1) natural input flows (e.g., solar radiation, wind, rain, and geothermal flow), (2) human-driven material and energy flows (e.g., machineries, fuels, chemicals), and (3) human labor and services converging to a process (Brown and Ulgiati, 2004).

According to this method, different forms of energy, materials, human labor, and economic services are accounted for in terms of their solar emergy, defined as the total amount of solar available energy (exergy) directly or indirectly required to make a given product or support a given flow, and measured in sej (solar emergy joules). The solar emergy required to generate one unit of product or service is referred to as Unit Emergy Value (UEV), expressed in sej/J in the case of energy flows or in sej/g in the case of mass flows.

The emergy method is deeply rooted in the concept of resource quality, i.e. the awareness that different energy forms have a different ability to do useful work even when their heat content is the same. Instead of only looking at what can be extracted out of a resource (exergy), the emergy accounting method focuses on what it takes for the biosphere to make and for human economy to process a given resource (Franzese et al., 2009).

The additional work provided by human activities to refine a raw resource adds up to its quality by making it more suitable to the final user. According to the emergy donor-side perspective, what makes a resource valuable is therefore both the environmental and human work investment for its production (Franzese et al., 2009).

An emergy assessment is typically implemented through the following main steps:

1. Identification of the spatial and temporal boundaries of the investigated system.
2. Modeling of the investigated system through an energy diagram drawn according to a standardized energy systems language (Odum, 1996).
3. Inventory of mass and energy flows supporting the investigated system.
4. Conversion of input flows into emergy units by using appropriate UEVs factors.
5. Calculation of the total emergy supporting the system.
6. Calculation of emergy-based indicators.

To calculate emergy-based indicators, all the input flows to a system are clustered into three main categories: the local Renewable flows (R), the local Non-renewable flows (N), and the flows Imported from outside the system (I). R, N, and I can be summed to account for the total emergy input (U) to a system.

In addition to the total emergy input (U) and the UEV of the system's output, the main emergy-based indicators used to describe the environmental performance and sustainability of the investigated system are: the Environmental Loading Ratio (ELR), the Emergy Yield Ratio (EYR), and the Emergy Sustainability Index (ESI).

The ELR compares the amount of non-renewable (N) and imported emergy (I) to the amount of locally available renewable emergy (R). In the case of a natural ecosystem, there would be no investments from outside and the system would be driven only by local renewable resources, thus having an $ELR=0$.

The EYR is a measure of the ability of a process to exploit and make available local resources by investing outside resources. The lowest possible value of the EYR is 1, situation in which the emergy converging to generate the yield does not differ significantly from the emergy invested from outside the system.

Finally, the ESI is an aggregated index calculated by the ratio of the EYR (sensitive to the outside-versus-local emergy use) and the ELR (sensitive to the non-renewable-versus-renewable emergy use) (Brown and Ulgiati, 2004).

Emergy flows can be converted into non-market "currency equivalents" by using the Emergy-to-Money Ratio (EMR) defined as the ratio between the total emergy (U) supporting a nation and its gross domestic product (GDP) in the same year (Lou and Ulgiati, 2013). This ratio represents the average amount of emergy needed to generate one unit of money in the national economy.

It is important to remark that, while the conversion of emergy units into money units can facilitate the communication of the results of emergy studies in political and socio-economic contexts, these figures expressed in money units should never be interpreted by applying a user-side perspective as they are non-market economic values.

Emergy Accounting and Marine Protected Areas

Emergy accounting has been widely applied to investigate natural, human-dominated, and man-made ecosystems. Many studies focused on the emergy assessment of natural capital and ecosystem services at local, regional, and national scales (Berrios et al., 2017; Campbell and Brown, 2012; Coscieme et al., 2014; Franzese et al., 2008, 2015; Huang et al., 2011; Lu et al., 2007, Ulgiati et al., 2011). Among them, recent studies focused on the assessment of matter and energy flows supporting the generation of natural capital stocks in marine protected areas. In particular, Vassallo et al. (2017) developed a biophysical and trophodynamic environmental accounting model based on emergy accounting and aimed at assessing the value of natural capital in marine protected areas. This model was articulated in three main steps: (1) trophodynamic analysis, providing an estimate of the primary productivity used to support the benthic trophic web within the study area; (2) biophysical accounting, providing an estimate of the biophysical value of natural capital; and (3) monetary conversion, expressing the biophysical value of natural capital into monetary units. This last step does not change the biophysical feature of the assessment, but instead it has the merit of allowing an easier understanding and effective communication of the ecological value of natural capital in socio-economic contexts.

Franzese et al. (2017) assessed the biophysical and economic value of natural capital in a Mediterranean marine protected area by using the emergy accounting method. In this study, the value of natural capital stocks in the main habitats included in the marine ecosystem was estimated in terms of work of biosphere invested for their generation and maintenance. In addition, using a GIS tool, the distribution of natural capital value in the protected area was made spatially explicit and integrated with the bionomic map and the zonation showing different protection levels.

The overlap of these different information layers allowed the identification of sites where natural capital value is more concentrated, supporting local managers and policy makers in charge for developing sustainable management strategies.

Picone et al. (2017) assessed the value of both autotrophic and heterotrophic natural capital stocks in a Mediterranean marine protected through emergy accounting. In addition, by using Marxan software, the results of the environmental accounting were integrated with spatial data on main human uses. This approach highlighted the importance of developing marine spatial planning taking into account both conservation measures and human activities devoted to resources exploitation.

Buonocore et al. (2019) jointly applied the emergy and eco-exergy methods to account for natural capital value in two marine protected areas located in Southern Italy. The emergy method allowed the assessment of natural capital in terms of direct and indirect solar energy flows invested for its generation, while the eco-exergy method accounted for the chemical energy stored in organic matter and the genetic information embodied in biomass stocks. This study showed that the sciaphilic hard bottom is characterized by a high emergy density value, confirming the high effort of nature in the formation of this habitat, suggesting the need for its proper consideration in the development of conservation strategies. Instead, *Posidonia oceanica* seagrass bed showed the highest value of eco-exergy density due to two main factors: (1) the large biomass density of *Posidonia oceanica* and (2) the high ß value of *Posidonia oceanica* reflecting its evolutionary history that involved the acquisition of key adaptations for the successful colonization of marine environments.

The integration between emergy and eco-exergy methods resulted in a useful approach to shed light on different biophysical measures of value based on different features of natural resources.

In all these studies, emergy accounting provided an alternative measure of value of natural resources complementary to more conventional economic studies. In addition, the integration of emergy accounting with other environmental accounting methods and tools allowed the generation of new scientific information useful for supporting the management of marine protected areas conceived to strike a balance between multiple human uses of the marine environment and nature conservation goals.

Concluding Remarks

Environmental accounting is a useful tool capable of addressing multiple aspects dealing with natural resources and their exploitation by human activities.

Among different environmental accounting methods, emergy accounting allows the assessment of nature's value applying a biophysical and geocentric perspective focusing on the work of biosphere supporting the generation of raw natural resources and refined products. The outcomes of emergy studies can complement those of more conventional economic analysis based on an anthropocentric perspective and on market laws.

A multicriteria approach to environmental accounting can address the plurality of nature values, supporting managers and policy makers in charge for sustaining the socio-economic development while ensuring the sustainable (long-term) management of natural resources.

References

Berrios F., Campbell D.E., Ortiz M., 2017. Emergy evaluation of benthic ecosystems influenced by upwelling in northern Chile: Contributions of the ecosystems to the regional economy. *Ecological Modelling* 359: 146–164.

Brown M.T., Ulgiati S., 2004. Emergy analysis and environmental accounting. In: *Encyclopedia of Energy*, C. Cleveland Editor, Academic Press: Elsevier, Oxford, UK, pp. 329–354.

Buonocore E., Picone F., Russo G.F., Franzese P.P., 2018. The scientific research on natural capital: A bibliometric network analysis. *Journal of Environmental Accounting and Management* 6(4): 374–384.

Buonocore, E., Picone, F., Donnarumma, L., Russo, G.F., Franzese, P.P., 2019. Modeling matter and energy flows in marine ecosystems using emergy and eco-exergy methods to account for natural capital value. *Ecological Modelling* 392: 137–146.

Campbell, E.T., Brown, M.T., 2012. Environmental accounting of natural capital and ecosystem services for the US National Forest System. *Environment, Development and Sustainability* 14: 691–724.

Cardinale, B. J., Duffy, J. E., Gonzalez, A., Hooper, D. U., Perrings, C., Venail, P., Narwani, A., Mace, G.M., Tilman, D., Wardle, D.A., Kinzig, A.P., Daily G.C., Loreau, M., Grace, J.B., Larigauderie, A., Srivastava, D.S., Naeem, S., 2012. Biodiversity loss and its impact on humanity. *Nature* 486: 59–67.

Coscieme, L., Pulselli, F.M., Marchettini, N., Sutton, P.C., Anderson, S., Sweeney, S., 2014. Emergy and ecosystem services: A national biogeographical assessment. *Ecosystem Services* 7: 152–159.

Costanza, R., de Groot, R., Sutton, P., van der Ploeg, S., Anderson, S.J., Kubiszewski, I., Farber, S., Turner, R.K., 2014. Changes in the global value of ecosystem services. *Global Environment Change* 26: 152–158.

EU, 2014. *Mapping and Assessment of Ecosystems and Their Services. Indicators for Ecosystem Assessment under Action 5 of the EU Biodiversity Strategy to 2020.* European Union. ISBN 978-92-79-36161-6.

Franzese, P.P., Brown, M.T., Ulgiati, S., 2014. Environmental accounting: Emergy, systems ecology, and ecological modelling. *Ecological Modelling* 271: 1–3.

Franzese, P.P., Buonocore, E., Donnarumma, L., Russo, G.F., 2017. Natural capital accounting in marine protected areas: The case of the Islands of Ventotene and S. Stefano (Central Italy). *Ecological Modelling* 360: 290–299.

Franzese, P.P., Buonocore, E., Paoli, C., Massa, F., Stefano, D., Fanciulli, G., Miccio, A., Mollica, E., Navone, A., Russo, G.F., Povero, P., Vassallo, P., 2015. Environmental accounting in marine protected areas: The EAMPA project. *Journal of Environmental Accounting and Management* 3(4): 324–332.

Franzese, P.P., Liu, G., Aricò, S., 2019. Environmental accounting models and nature conservation strategies. *Ecological Modelling* 397: 36–38.

Franzese, P.P., Russo, G.F., Ulgiati S., 2008. Modelling the interplay of environment, economy and resources in marine protected areas. A case study in Southern Italy. *Ecological Questions* 10: 91–97.

Franzese, P.P., Rydberg, T., Russo, G.F., Ulgiati, S., 2009. Sustainable biomass production: A comparison between gross energy requirement and emergy synthesis methods. *Ecological Indicators* 9(5): 959–970.

Häyhä, T., Franzese, P.P., 2014. Ecosystem services assessment: A review under an ecological-economic and systems perspective. *Ecological Modelling* 289: 124–132.

Huang, S., Chen, Y., Kuo, F., Wang, S., 2011. Emergy-based evaluation of peri-urban ecosystem services. *Ecological Complexity* 8: 38–50.

Hoffmann, S., Beierkuhnlein, C., Field, R., Provenzale, A., Chiarucci, A., 2018. Uniqueness of protected areas for conservation strategies in the European Union. *Scientific Reports* 8: 6445.

Lou, B., Ulgiati, S., 2013. Identifying the environmental support and constraints to the Chinese economic growth – An application of the emergy accounting method. *Energy Policy* 55: 217–233.

Lu, H., Campbell, D., Chen, J., Qin, P., Ren, H., 2007. Conservation and economic viability of nature reserves: An emergy evaluation of the Yancheng Biosphere Reserve. *Biological Conservation* 139: 415–438.

Maes, J., Liquete, C., Teller, A., Erhard, M., Luisa, M., Barredo, J. I., Grizzetti, B., Cardoso, A., Somma, F., Petersen, J., Meiner, A., Gelabert, E.R., Zal, N., Kristensen, P., Bastrup-Birk, A., Biala,K., Piroddi, C., Egoh, B., Degeorges, P., Fiorina, C., Santos-Martín, F., Naruševičius, V., Verboven, J., Pereira, H. M., Bengtsson, J., Gocheva,K., Marta-Pedroso, C., Snäll, T., Estreguil, C., San-Miguel-Ayanz, J., Pérez-Soba, M., Grêt-Regamey, A., Lillebø, A.I., Malak, D.A., Condé, S., Moen, J., Czúcz, B., Drakou, E.G., Zulian, G., Lavalle, C., 2016. An indicator framework for assessing ecosystem services in support of the EU biodiversity strategy to 2020. *Ecosystem Services* 17: 14–23.

Neugarten, R.A., Langhammer, P.F., Osipova, E., Bagstad, K.J., Bhagabati, N., Butchart, S.H.M., Dudley, N., Elliott, V., Gerber, L.R., Gutierrez Arrellano, C., Ivanić, K.-Z., Kettunen, M., Mandle, L., Merriman, J.C., Mulligan, M., Peh, K.S.-H., Raudsepp-Hearne, C., Semmens, D.J., Stolton, S., Willcock, S., 2018. *Tools for Measuring, Modelling, and Valuing Ecosystem Services: Guidance for Key Biodiversity Areas, Natural World Heritage Sites, and Protected Areas*. IUCN: Gland, Switzerland, 70 pp.

Odum, H.T., 1988. Self organization, transformity and information. *Science* 242: 1132–1139.

Odum, H.T., 1996. *Environmental Accounting: Emergy and Environmental Decision Making.* John Wiley & Sons: New York, 369 pp.

Pauna, V.H., Picone, F., Le Guyader, G., Buonocore, E., Franzese, P.P., 2018. The scientific research on ecosystem services: A bibliometric analysis. *Ecological Questions* 29 (3): 53–62.

Picone, F., Buonocore, E., D'Agostaro, R., Donati, S., Chemello, R., Franzese, P.P., 2017. Integrating natural capital assessment and marine spatial planning: A case study in the Mediterranean sea. *Ecological Modelling* 361: 1–13.

TEEB, 2010. *The Economics of Ecosystems and Biodiversity for Local and Regional Policy Makers*. TEEB: Bonn, Germany.

Teixeira, H., Lillebø, A.I., Culhane, F., Robinson, L., Trauner, D., Borgwardt, F., Kuemmerlen, M., Barbosa, A., McDonald, H., Funk, A., O'Higgins,T., Tjalling, J., der Wal, V., Piet, G., Hein, T., Arévalo-Torres, J., Iglesias-Campos, A., Barbière,J., Nogueira, A.J.A., 2019. Linking biodiversity to ecosystem services supply: Patterns across aquatic ecosystems. *Science of the Total Environment* 657: 517–534.

Ulgiati, S., Zucaro, A., Franzese, P.P., 2011. Shared wealth or nobody's land? The worth of natural capital and ecosystem services. *Ecological Economics* 70 (4): 778–787.

UNEP-WCMC, IUCN and NGS, 2018. *Protected Planet Report 2018*. UNEP-WCMC, IUCN and NGS: Cambridge UK; Gland, Switzerland; and Washington, DC.

Vassallo, P., Paoli, C., Buonocore, E., Franzese, P.P., Russo, G.F., Povero, P., 2017. Assessing the value of natural capital in marine protected areas: A biophysical and trophodynamic environmental accounting model. *Ecological Modelling* 355: 12–17.

Vihervaara, P., Franzese, P.P., Buonocore, E., 2019. Information, energy, and eco-exergy as indicators of ecosystem complexity. *Ecological Modelling* 395: 23–27.

25

Remote Sensing: Pollution

Massimo
Antoninetti

Introduction

The earth's environment is placed under constant stress by various human activities and natural processes.

Resource management activities for conservation or sustainable use rely on several critical components: accurate and up-to-date information, an understanding of "how" the environment and its components function, and an ability to continuously and constantly monitor the environment's components and processes. Consequently, there is a need for spatial information in the form of baseline mapping and inventory, monitoring programs, and predictive models. Satellite remote sensing (RS) is going to play a major role in fields related to place and space. Thanks to this innovative technology, it is possible to monitor and map rapidly changing phenomena on the surface of our planet due to its broad spectral range, affordable cost, and rapid coverage of large areas. Synergistic use of RS and ancillary data can create a geographic information system (GIS) database, which can be used to store, process, and retrieve environmental data.

This is not a comprehensive review, as many entries have been published on the topics covered. We have attempted to emphasize a few important applications, recent reviews, and recent papers that provide the reader with an overview of the main literature on each topic.

Remote Sensing

Remote sensing is generally defined as the set of techniques, instruments, and interpretative tools able to expand and improve the perceptive ability of the human eye, collecting information about an object without being in physical contact with the object itself.[1] Consequently, RS allows the obtaining of quantitative and qualitative measures from a distance of the surface of objects, and, under some specific conditions, also of the subsurface. The most frequently measured quantity is the electromagnetic energy in one or more regions of the electromagnetic spectrum reflected or emitted from the earth's surface.[2] The primary law governing radiation is the Planck radiation law, which governs the intensity of radiation emitted by unit surface area into a fixed direction (solid angle) from a blackbody as a function of wavelength for a fixed temperature. Conceptually, a blackbody is an ideal radiator of thermal radiant energy, and consequently, also an ideal absorber. The Planck law can be expressed through the following equation:

$$M_\lambda = \frac{\varepsilon C_1}{\lambda^5 \left[\exp\left(\frac{C_2}{\lambda T} \right) - 1 \right]} \tag{1}$$

where M_λ is the spectral radiant exitance (W cm^{-2} µm^{-1}); ε = emittance (emissivity), dimensionless; C_1 is the first radiation constant, 3.7413×10^8 [W (µm)^4]/m^2; C_2 is the second radiation constant, 1.4388×10^8 (µm K); λ is the radiation wavelength (µm); and T is the absolute radiant temperature (K).

Moreover, the Stefan–Boltzmann law states that the total amount of energy M emitted by a blackbody is proportional to the fourth power of its absolute temperature T, and consequently M rapidly increases with the T increasing:

$$M = \int_0^\infty M_\lambda d\lambda = \sigma T^4 \left[Wcm^{-2} \right] \tag{2}$$

where σ is the Stefan–Boltzmann constant ($\sigma = 5.67 \times 10^{-4}$ W cm^{-2} K^{-4}).

However, there is an inverse relationship between the maximum wavelength emission $\lambda_{\max}$ of a blackbody and its temperature when expressed as a function of wavelength (Wien radiation law):

$$\lambda_{max} = \frac{2898}{T} [\mu\text{m}] \tag{3}$$

The Wien law demonstrates that the $\lambda_{\max}$ moves toward lower values at the increasing of the surface temperature T of a body. For example, the maximum emission value for the sun is 0.483 µm (corresponding to the blue-green part of the visible spectrum), 0.966 µm for a incandescent bulb (near infrared [NIR]), and 9.66 µm for the surface of the earth (thermic infrared).

Consequently, for any given material, the amount of solar radiation that it reflects, absorbs, transmits, or emits varies with wavelength. When that amount (usually intensity, as a percent of maximum) coming from the material is plotted over a range of wavelengths, the connected points produce a curve called the material's spectral signature (spectral response curve) (Figure 1). For example, at some wavelengths,

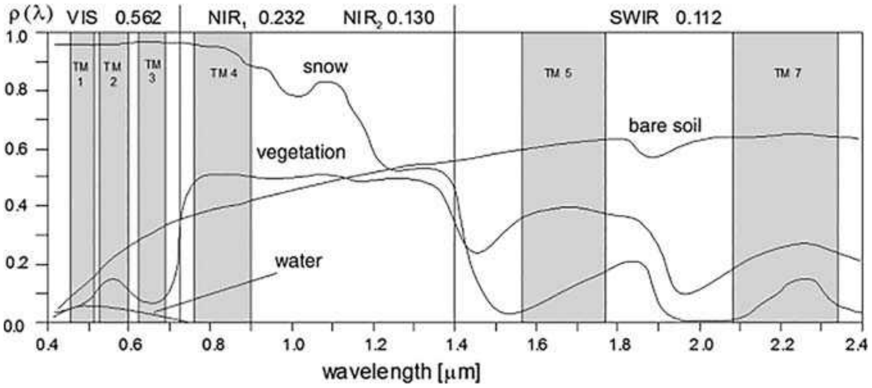


FIGURE 1 Schematization of solar radiation spectral distribution of typical surfaces' spectral reflectivity curves, and of Landsat TM spectral bands.
Source: Brivio et al.[3]

sand reflects more energy than green vegetation; however, at other wavelengths, it absorbs more (reflects less) than the vegetation.

There are two main types of RS using different sensors to register the data: passive RS and active RS. A sensor can be defined as a device that measures a physical quantity and converts it into a signal that can be read by an observer or by an instrument. This physical quantity is the radiance (W m^{-2} sr^{-1}) or the spectral radiance (W m^{-2} sr^{-1} λ^{-1}); they are radiometric measures that describe the amount of electromagnetic energy that passes through or is emitted from a particular area, and falls within a given solid angle in a specified direction. They characterize both emission from diffuse sources and reflection from diffuse surfaces.

Passive sensors detect the natural radiation emitted or reflected by the object. Reflected sunlight is the most common source of radiation measured by passive sensors. Active sensors emit energy in order to scan objects and areas, detecting and measuring the radiation that is reflected or backscattered from the target. RADAR (radio detection and ranging) is an example of active RS where the time delay between emission and return is measured, establishing the location, height, speed, and direction of an object (Figure 2). In most cases, the instruments register a raster image. The image contains a fixed number of rows and columns of pixels, as the smallest individual picture element in the image (Figure 3).

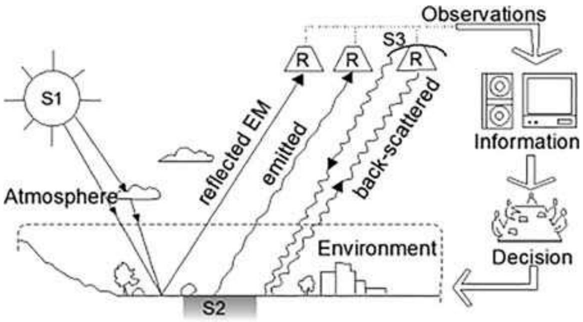
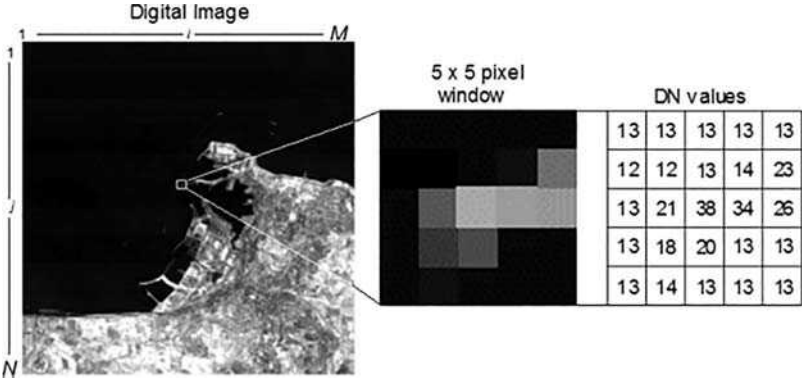


FIGURE 2 Remote sensing basic scheme. Natural energy sources are the sun (S1) and the earth (S2), while the antenna (S3) is an artificial source. Sensor R registers the information carried by electromagnetic waves. This information is stored and elaborated to be used for environmental studies.
Source: Modified from Brivio et al.[3]



13	13	13	13	13
12	12	13	14	23
13	21	38	34	26
13	18	20	13	13
13	14	13	13	13

FIGURE 3 Digital image concept and description. A digital image is formed by a matrix of $N \leftrightarrow M$ pixels, easily visible only with an enlargement in the window 5 ↔ 5. A digital number (DN) is assigned to each pixel in function of the intensity of the electromagnetic energy registered by sensors.
Source: Modified from Brivio et al.[3]

Each pixel holds a quantized value representing the amount of electromagnetic energy registered by instruments at any specific point. Consequently, the quality of RS image data consists of its spatial, spectral, radiometric, and temporal resolutions. The spatial resolution represents the size of the surface area (i.e., in square meters) being measured on the ground, determined by the sensors' instantaneous field of view, from less than 1 m to more than 1 km. A higher resolution means more image details.

The spectral resolution is the wavelength interval size (discreet segment of the electromagnetic spectrum) recorded by the instruments mounted on board, normally a radiometer, able to measure the radiant flux of electromagnetic radiation. A multispectral radiometer collects electromagnetic energy in different spectral bands, normally from the visible (0.4 μm) to the long-wave infrared (15 μm) part of the spectrum (e.g., Enhanced Thematic Mapper Plus [ETM+] mounted on board of the Landsat satellite has 8 bands recorded), while a hyperspectral imaging deals with hundreds of narrow spectral bands, producing a spectra of all pixels in the scene (e.g., Hyperion sensor on Earth Observing-1 has 220 bands with a spectral resolution of 0.10–0.11 μm per band). The radiometric resolution is the number of different intensities of radiation that the sensor is able to distinguish. Typically, this ranges from 8 to 14 bits, corresponding to the 256 levels of the gray scale and up to 16,384 intensities or "shades" of color in each band.

Finally, temporal resolution is defined as the frequency at which images are recorded/captured in a specific place on the earth's surface (National Oceanic and Atmospheric Administration Advanced Very High Resolution Radiometer [NOAA-AVHRR] capture an image every 12 hr, while Landsat ETM+ every 16 days) (Table 1).

Remote sensing data are processed and analyzed with a digital image processing system, using computer algorithms to improve the quality and quantity of the information registered. As an image is formed by an array of numbers, digital image processing uses simple or complex mathematic algorithms to obtain image restoration; to eliminate errors, noises, and distortions; to improve the information content extraction; and to recognize and classify pixels on the basis of their digital signatures.

A simple process, for example, is the "false-color" composition. Three images acquired in different spectral bands could be combined together to obtain an RGB (red–green–blue) color image where each color tone is the combination of the information contained in each band. The term "false color" identifies images whose colors represent measured intensities outside the human eye capacity.

A complex mathematical model is an artificial neural network, inspired by the structure of physically interconnected neurons in biology and able to process information using a connectionist approach through an interconnected group of artificial neurons. It provides an efficient alternative to map complex nonlinear relationships between input and output datasets without requiring a detailed knowledge of underlying physical relationships. **Table 1** List of main satellites with on-board sensors description

Finally, Geographic Information System (GIS) is a database system used for manipulating data, utilizing a data input system, data storage and retrieval, data manipulation and analysis, and data output. The data inputs are usually spatial and non-spatial data derived from a combination of existing maps, aerial photographs, and RS imagery. The data output is used to generate reports in tabular form, digital displays, or maps. Because the input data on which a GIS is based becomes obsolete quickly, it is important to update this data. Remote sensing is often the most cost-effective source for these updates. The resource database should contain as much base inventory data as possible, as well as simple layer selection and spatial and tabular query capabilities, to illustrate how this data is accessible and useful. Data manipulation is of primary concern, including the tracking of metadata that describe the legacy of the information including source, compilations scale, data of last compilation, and other information needed to qualify the usefulness of the information to support a particular application.

What goes beyond a GIS is the Spatial Data Infrastructure, tools that are interactively connected on the web for retrieving, presenting, and processing data in an efficient and flexible way, in a coordinated series of agreements on technology standards, institutional arrangements, and policies.

Remote Sensing Applications

Water

Although not all deleterious processes can be measured directly (e.g., overfishing), many environmental and ecological properties can be measured using RS. These properties include sea surface temperature (SST), chlorophyll-*a* (Chl-*a*) and suspended sediment concentration, salinity, wind speed, algal blooms, etc. Given a robust understanding of the ecosystem responses to these environmental parameters and some in situ field observations, many other biological benchmarks can often be indirectly derived (e.g., fish abundance).[4]

In the open waters, color variations chiefly depend on the presence and abundance of plankton: these waters are defined as Case I. Conversely, coastal and estuarine systems are often characterized by optically complex waters (Case II), with high concentrations of highly reflective water constituents, the color of which is also due to dissolved organic matter and suspended sediments. The constituents, characterized by variation at much smaller geographical scales, are able to modify natural equilibrium.[5]

In recent years, coastal zones, probably more than any other parts of the earth, have been exposed to pressure and processes of change caused by natural processes (waves, currents and storms, etc.), and human activities.[6] As a large percentage of the world's coastlines are poorly mapped or maps have become outdated, satellite imagery can be a useful and economic tool to overcome these constraints, producing maps containing thematic information of relevant interest in coastal environment protection: planimetric mapping of waterlines and beach and coral reef position, shallow water bathymetry, and thematic maps of cover types on adjacent land and islands. This may include mapping of bare sand, mangroves, seagrass, reef flats, wetlands, etc. The number of classes (categories) (e.g., coral reef, seagrass, sand, hard substrate) distinguishable by RS depends on many natural factors (including atmospheric clarity, surface roughness, water clarity, and water depth) and on the type of RS sensor. Very high resolution pixels (0.1–0.8 m) are required to accurately quantify the percentage of bleached corals. The most likely solution to this limitation of pixel size is the application of methods that estimate the cover of substrata within pixels. These methods known as spectral unmixing, even though they have been developed for terrestrial RS, could be adapted to aquatic environment, taking into account the interference of water column.[4]

In shallow water, the water column (including the substances dissolved or in suspension in the medium) and the seafloor contribute to the upwelling signal leaving the water body. An inverse modeling of the radiation transfer can be developed for simultaneous deriving of seafloor reflectance, depth, water-column subsurface reflectance, and diffuse attenuation coefficient from RS of coastal shallow waters.[7] This method relies on a synthetic database of subsurface irradiance reflectance-compiled field and simulated data and the use of a forward shallow-water reflectance model.[8] Uncertainties in the computation of Chl-*a* absolute values can arise primarily owing to the presence, in the water column, of optically active materials other than phytoplankton and related pigments (i.e., dissolved organic matter and suspended inorganic particles), with partially overlapping spectral signatures. Nevertheless, experience has shown that remotely sensed data can provide unique information on phytoplankton growth patterns over a range of space scales (i.e., from a few kilometers to entire basins) and of time scales (i.e., from days to years) not available by any other means.[9]

Ultraviolet (UV) radiation and high levels of photosynthetically active radiation can have a variety of negative impacts on marine phytoplankton, zooplankton, nekton, and benthos. Satellites provide global time series measurements of incident UV radiation.

Understanding the way in which cetaceans interact with the surrounding environment has become a key tool, not only in their subsequent conservation but also in the management of the ecosystems of which they form a part. Due to their high trophic status and vulnerability to anthropogenic effects, marine mammals such as cetaceans are ideally suited to act as indicator species of ecosystem change and anthropogenic threats.[10] Cetacean studies have focused on their relationships with sea temperature.

Thermal infrared SST measurements are derived from radiometric observations at wavelengths of ~3.7 μm and/or near 10 μm. Although the 3.7 μm channel is more sensitive to SST, it is primarily used only for night-time measurements because of relatively strong reflection of solar irradiation in this wavelength region, which contaminates the retrieved radiation. Both bands are sensitive to the presence of clouds and scattering by aerosols and atmospheric water vapor. For this reason, thermal infrared measurements of SST first require atmospheric correction of the retrieved signal and can only be made for cloud-free pixels. Thus, maps of SST compiled from thermal infrared measurements are often weekly or monthly composites, which allow enough time to capture cloud-free pixels over a region. Thermal infrared instruments that have been used for deriving SST include AVHRR on NOAA Polar-Orbiting Operational Environmental Satellites (POES), Along-Track Scanning Radiometer (ATSR) aboard the European Remote Sensing Satellite (ERS-2), the Geostationary Operational Environmental Satellite (GOES) Imager, and Moderate Resolution Imaging Spectroradiometer (MODIS) aboard NASA Earth Observing System (EOS) Terra and Aqua satellites.

Images supplied by the NERC Remote Sensing Data Analysis Service (RSDAS) were atmospherically corrected and provided in a Mercatore projection. Values or digital numbers (DN) associated with each pixel of the AVHRR images supplied were converted into real SST (°C) values using the following equation as supplied by the RSDAS[10]:

$$\mathrm{SST} = \mathrm{DN} \times 0.1 - 0.3 \tag{4}$$

where SST is the sea surface temperature (°C) and DN is the digital number or the value of each pixel.

Chlorophyll-*a* concentration data could be obtained from the Sea-Viewing Wide Field-of-View Sensor (Sea-WiFS) monthly composite images using the following equation, also supplied by RSDAS:

$$\mathrm{CHL} = 10^{(\mathrm{DN} \times 0.015 - 2.0)} \tag{5}$$

Where CHL is the Chl-*a* concentration (μg L^{-1}).

Higher numbers of striped dolphins were generally observed in the 21–24°C range in the Mediterranean.[11]

Environmental protection from spills of hydrocarbon compound over the sea surface is currently an important subject of increasing public concern. Oil spills, depending on the exact hydrocarbon content and type, usually involve extensive areas of film on sea surface, a fact that reduces water roughness and can therefore allow the detection by synthetic aperture radar (SAR) images, providing information on location, size, distance from the land, etc.[12,13] The automatic detection of oil slicks in SAR images is not a simple task, as objects resembling oil spills (often called look-alikes) occur frequently in SAR images, especially in low wind conditions. Most frequently, look-alikes are produced by organic film, grease, wind front areas, land, plankton formations, rain cell, current shear zones, and upwelling zones. The probability of each object extracted after image classification to be an oil spill was estimated using an artificial intelligence fuzzy logic modeling system. The fuzzy logic theory has emerged as a useful tool for modeling processes that are too complex for conventional quantitative techniques, or when the available information from the process is qualitative, inexact, or uncertain. Fuzzy logic addresses qualitative information perfectly as it resembles the way humans make inferences and decisions.[14] It fills an important gap in system design methods, which is between purely mathematical approaches (e.g., system design) and purely logic-based approaches (e.g., expert systems). While other approaches require accurate equations to model real-world behaviors, fuzzy design can accommodate the ambiguities of real-world human language and logic. It provides an intuitive method for describing systems in human terms and automates the conversion of those system specifications into effective models. Traditional set theory is based on bivalent logic, where an object is either a member of a set or not. Contrary to that, fuzzy logic allows a number or object to be a member of more than one set and most importantly it introduces the notion of partial membership.[15]

Most of the oil spill detection studies use low-resolution SAR data (quick-looks) with nominal spatial resolution of 100 m×100 m to detect oil spills. Low-resolution data is sufficient for large-scale monitoring; however, small and fresh spills cannot be detected sufficiently as they are represented in very few pixels and present brighter contrast than the bigger and older oil spills. Neural networks are able to detect and classify dark formations. A neural network's ability to successfully handle nonlinearly separable classes is a big advantage against the commonly used statistical approaches.[16] The exploitation of optical satellite images allows more frequent (sometimes daily) information if compared with SAR images. However, good weather conditions and daylight are mandatory conditions for correct detection. The physical and geometrical features of an oil spill are computed by the feature extraction module for an automatic supervised classification: geometrical features (area, perimeter, and complexity of the object) and gray level features (object standard deviation, maximum contrast, mean contrast).[17] Due to its thermal inertia (lower than the sea water), oil polluted areas have a higher brightness temperature in AVHRR thermal infrared (TIR) images collected in daytime, and the opposite during the night. In fact, materials having high thermal inertia show resistance toward a change in temperature, resulting in long-term temperature stability. In daytime, an oil spill should exhibit a relatively higher brightness temperature in AVHRR TIR images as, due to solar radiation, oil becomes warmer than the surrounding water.[18] Finally, GIS can qualitatively and quantitatively characterize not only spatial and temporal distribution of oil spills but also the environmental conditions of the sea basins as a whole. Such an environment can be created by means of integration in a GIS of different databases for seawater quality, nutrient and chemical composition, and contextual information about slick position relative to surrounding objects (ships, ship lanes, rigs, platforms, natural seeps).[19,20]

Groundwater is one of the most valuable natural resources, supporting human health, economic development, and ecological diversity. Nevertheless, it is considered to be vulnerable mainly due to anthropogenic activities such as agriculture and waste disposal. Satellite data provides quick and useful baseline information about factors controlling the occurrence and movement of groundwater, such as geology, lithology, geomorphology, soils, land use/cover, drainage patterns, and lineaments. Remote sensing helps minimize the amount of field data collection in the exploration and assessment of groundwater resources, in estimation of natural recharge distribution, and in subsurface flow and pollution modeling.[21]

A pollution risk map can be obtained combined a vulnerability map with a land use map in a GIS. The vulnerability map illustrates the potential decrease of groundwater protection as a result of the flow regime toward the aquifer, analyzing the hydrogeologic and geomorphologic conditions of the basin and the isotopic ratio$^{18}O/^{16}O$ measured in the collected water samples (springs, rivers, lakes). The updated land use map is produced by satellite image data set analysis. The socioeconomic component is introduced to integrate the whole process leading to the assessment of aquifer's pollution risk, a very important element for decision making and planning.[22]

Agricultural activities have been identified as major sources of non-point source pollutants of ground and surface waters, with nitrogen being one of the most important and problematic nutrients. The use of mathematical models is a common approach for analyzing and describing the status quo, identifying interdependencies between agriculture and hydrosphere, and investigating the effects of agricultural environmental reduction measures. Satellite imagery has been used to substitute low-resolution, outdated, and inadequately differentiated land use maps with high-resolution land cover and imperviousness maps. By applying crop-specific classified satellite imagery in a semiautomated way, the nitrogen surpluses can be transferred and spatially located to main crops on the field scale.[23]

Wetlands and lakes are very sensitive ecosystems that play a key role in maintaining ecological equilibrium and biodiversity. They provide a habitat for wildlife, support a rich biodiversity, and aid in floodwater management and water quality improvement besides having esthetic and educational benefits for humans. Moreover, local economies depend on the water and vegetation of wetlands and lakes for fisheries, reed harvesting, grazing, and recreation. Unfortunately, in many countries, the extensive loss of wetlands and trophic pollution of lakes have occurred. Imaging spectrometry is a versatile tool for assessing water quality in small lakes, monitoring the variation of submerged vegetation in coastal zones, and performing ad hoc studies on transitional ecosystems as lagoons. The Chl-a and the TSS maps, obtained by combining MIVIS hyperspectral airborne data with semi-empirical modeling, confirmed the hypereutrophic-dystrophic conditions of the lakes and the elevated load of nutrient and suspended matter transported by the tributary rivers (Figure 4). The surface temperature map also revealed the discharge of warmer waters by urbanized and industrialized areas.[24] Moreover, MIVIS data demonstrates a good capacity in evaluating variations in submerged vegetation (Figure 5). Macrophyte meadows are in fact indicators of water and sediment quality, as their disappearance from many freshwater and coastal marine environments can be attributed mainly to the phenomena of eutrophication.[25] Once the models are calibrated with in situ measurements, bio-optical models may be applicable over selected lakes. They are mathematical equations that relate radiometric variables observed above or below the water surface (e.g., satellite-derived water reflectance) to the inherent optical properties (i.e., absorption and backscattering coefficients of colored dissolved organic matter and particle).[26]

In recent decades, common reeds (*Phragmites australis*) have been monitored for various purposes, including wetland management for nature conservation. Common reeds are an ecosystem with a high value of biodiversity, ideal for fish, amphibian, and bird reproduction; moreover, they function as a cleaning filter for the lake waters as they can remove up to 10%–15% of the sediment nitrogen content and subtract significant amounts of heavy metals. Lake water eutrophication has been identified as the main reason for reed die-back, while in the hotter and drier Mediterranean area, *P. australis* seems to

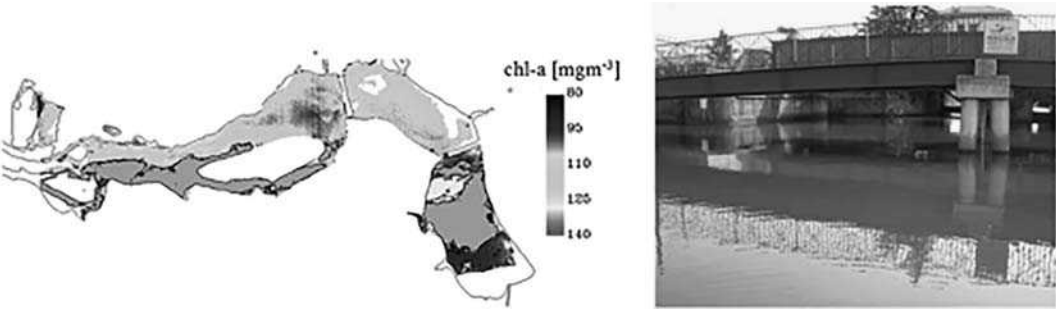


FIGURE 4 MIVIS-derived Chl-*a* maps of three lakes of Mantua, Italy. The high value concentrations are due to eutrophication caused by excessive nutrients from agricultural sources that color of green water.
Source: IREA-CNR, Milan, Italy.

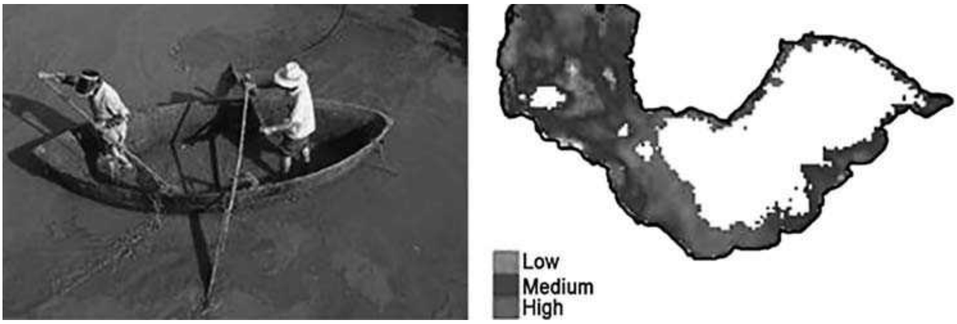


FIGURE 5 Left: Algae bloom situation at Hefei Yicheng, China, on June. Right: Probability of finding phycocyanin pigments in the algal bloom observed for the scene acquired by MERIS on May (in white, the water types in better status).
Source: IREA-CNR, Milan, Italy.

grow vigorously and to expand, even in eutrophicated areas. High normalized difference vegetation index (NDVI) values highlight greater amount of reed biomass that is the result of the higher germinative capacity in optimal environmental conditions.[27]

Land

As for land, ecosystem vegetation plays a fundamental role in the energy and mass exchange between land surface and atmosphere, and can be considered a good indicator of environmental stress. Environmental health conditions, stand structure, damage levels, and other phenomena can be determined using satellite images.[28] The spectral vegetation indices are used to investigate changes in plant stress due to soil contamination. A vegetation index combines two or more spectral bands to enhance the vegetative signal while minimizing background effects; they are commonly used to measure the sensitivity of vegetation to stress. As hyperspectral RS images are increasingly becoming available, further exploration of the increased number of spectral variables is a logical step.[29]

Canopy leaf area index (LAI) is typically defined as the one-sided area of green foliage projected onto a unit area of ground. The LAI is critical in estimating terrestrial carbon sequestration, net primary production, hydrologic watershed budgets, and pollutant deposition. The LAI retrieval method is based on an analytical solution to a multiple scattering equation. The algorithm uses different input data layers to produce a single LAI map: reflectance in the red and NIR bands, satellite zenith view angle, solar zenith altitude angle, sun sensor relative azimuth, and terrain characteristics (elevation, slope, and aspect).[30] A simplified approach uses a semi-empirical regression model between ground measurements and spectral vegetation indices (VIs), such as band ratios or normalized band difference, computed from remotely sensed images.[31,32]

Atmospheric composition affects reflectivity of the spectrum and makes it difficult to compare satellite images from different seasons. To reduce impacts from the atmosphere, the normalized ratio vegetation index (NRVI) has been applied to satellite imagery to study Chl-*a* concentration.

For Landsat ETM/TM, TM3 is a main band for absorbing the spectrum from Chl-*a* and is represented by the red band with a wavelength range varying from 630 to 690 nm. TM4 is an NIR band with a wavelength range varying from 760 to 900 nm. The band is sensitive to the category of green vegetation and is therefore often used for detecting green vegetation. Based on the characteristics of the two bands, three VIs could be selected for RS analysis: NDVI, ratio vegetation index (RVI), and NRVI, all of which are relatively simple in arithmetic and widely used.

The equation of the RVI is

$$\mathrm{RVI} = \frac{\mathrm{TM3}}{\mathrm{TM4}} \tag{6}$$

The equation of the NDVI is

$$\mathrm{NDVI} = \frac{\mathrm{TM4} - \mathrm{TM3}}{\mathrm{TM4} + \mathrm{TM3}} \tag{7}$$

The equation of the NRVI is

$$\mathrm{NRVI} = \frac{\mathrm{RVI} - 1}{\mathrm{RVI} + 1} \tag{8}$$

The NRVI has a positive correlation with Chl-*a* concentration, while NDVI has a negative correlation. Correlative degrees between NRVI and Chl-*a* are also much higher than that of NDVI, showing that the obtained regressions capture the relation between the water quality parameters (nutrient pollution loaded mainly with agricultural wastewater, phosphorous pollution loaded mainly with runoff from forest areas, pollution loaded with a mixture of domestic and industrial wastewater, and agricultural runoff and natural lake water) and the band reflectance values.[33]

Change in leaf spectral reflectance is a symptom of deteriorating forest health. Increase in red reflectance due to reduced chlorophyll absorption, decrease in NIR reflectance due to reduced cell vigor, and shifts in the red edge between these two regions have been commonly used as indicators of leaf stress.[34]

Topography also plays an important role in vegetation sensitivity to air pollution. In general, elevated and wind-ward sites appear to be more affected than depressed and leeward locations. Consequently, wind direction could apparently be more important than topography.[35] Using RS techniques to achieve a detailed analysis of the type and status of vegetation can help us understand the fate of persistent organic pollutants (POPs) in the environment. These POPs are generally bioaccumulated in the food chain and affect living organisms owing to their long persistence and toxicity. Forests play a key role in the environmental distribution of POPs since they act as filters of these chemicals, trapping them in the air compartment and transferring them to forest soils, consequently decreasing their atmospheric half-lives. Therefore, it is necessary to include the vegetation compartment in multimedia models, to understand and predict the fate of these substances in the environment. Multimedia models used to study the fate of POPs and their redistribution in different environmental compartments, from the atmosphere to soil or water via vegetation interface, require foliage biomass information. Foliage biomass (FB) is estimated with the following model:

$$\mathrm{FB}_i = \mathrm{LAI}_i \times \frac{1}{SLA_i} \times \left(\frac{\mathrm{FW}}{\mathrm{DW}}\right)_i \tag{9}$$

where $\mathrm{FB}_{i;}$ is the foliage biomass of vegetation class i (kg m^{-2}), LAI_i, is the leaf area index of vegetation class i, SLA_i,· is the specific leaf area (m^2kg^{-1}) of vegetation class i, and $(\mathrm{FW/DW})_i$, is the relative ratio between fresh and dry biomass weight of vegetation class i.[36]

Natural fires and arson greatly affect forests, vegetation, and the environment in general (Figure 6) Mercury (Hg) emissions from fires may be also detected directly using reflectance spectroscopy, field

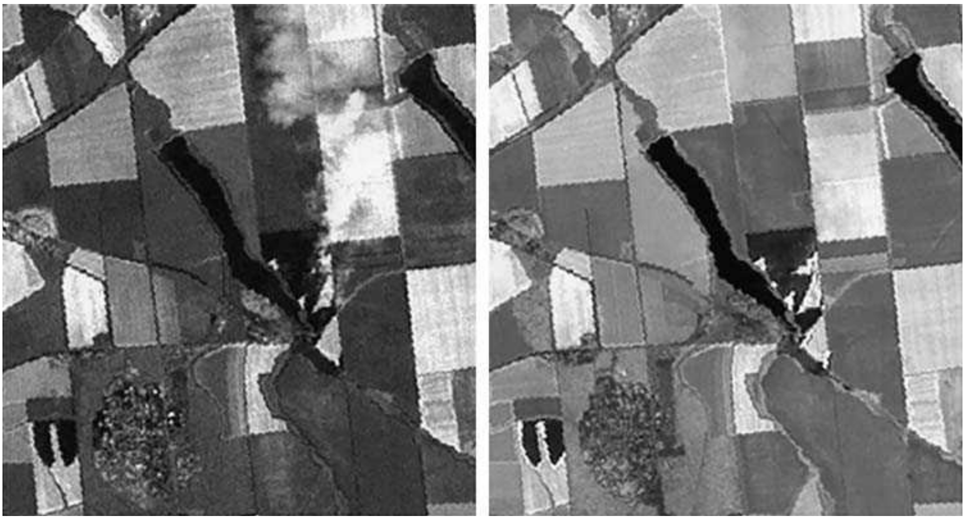


FIGURE 6 False-color Landsat-7 images taken over Ukraine on July 2001, illustrating the different possibilities in identifying agriculture fires. Left: RGB natural color composition of visible band highlights fires smoke, its intensity and direction, and burned areas. Right: RGB false color composition of visible, and infrared bands, identifies active fires and burned areas.
Source: IREA-CNR, Milan, Italy.

sampling, and laboratory analysis; however, many contaminated regions are covered by vegetation. The Hg released to the atmosphere from vegetation is primarily related to Hg concentrations in foliage. It varies with tissues of plants—predominantly leaves, bark, and root—foliar age, forest typologies, and soil characteristics.

Satellite observations were used to assess Hg emissions through the assessment of spatial and temporal distributions of forest phytomass content and forest burnt areas.[37] Plants with the greatest amount of foliar Hg showed symptoms of stress earlier than plants of other treatments. Foliar Hg appears to influence spectral reflectance at several specific regions of the electromagnetic spectrum. The red edge position is the wavelength of the maximum value of the first derivative of the spectra in the 680–740 nm region and is positively related to chlorophyll concentration and biomass of leaves and canopies.[38] Shifts in vegetation spectra, occurring in both the visible and the NIR part of the spectrum, could be metal induced due to geochemical stress or the presence of old waste deposit sites.

Mining activities, such as coal, are often responsible for environmental pollution. Consequently, the forest patches near the mining sites have been badly degraded, whereas patches away from the mining sites registered relatively less impact. The degree of degradation was based on density, canopy structure, ground flora, and soil conditions (presence of ash and pollutants). The prime reason of these changes can be attributed to the discharge of noxious/toxic gases and particulate matter during drilling, blasting, and loading/unloading operations. It has been recorded that the areas near the industrial area and affected areas hold fewer plant species than that of areas far away from these site. Variation in number of herbs per unit area was found to decrease significantly when approaching mining sites.[39]

Urban green spaces provide a variety of functions essential for improving the quality of citizen life, while urban heat island tends to aggravate the negative effects of climate over urban areas and their surroundings.

Using NDVI multitemporal image differencing it is possible to detect subtle changes and to discern "from–to" changes using a postclassification comparison.[40]

Land surface temperature (LST) retrieved by RS is an important parameter in characterizing surface energy balance and a key parameter in monitoring farming drought, and very significant in studies of global change, meteorology, hydrology, ecology, agricultural, etc. Results reveal that LST is significantly different in various types of land use, the highest is bare land; the second is built-up land; followed by cropland, grass, and forest; and at the lowest in water.[41]

City growth is increasing population and the change of land cover types from permeable to anthropogenic impervious surfaces. Impervious surface area (ISA) is defined as constructed surfaces—roofs, roads, parking lots, driveways, and sidewalks. Impervious surfaces can alter the natural hydrological condition by increasing the volume and rate of surface runoff and decreasing groundwater recharge and base flow. Pollutants either dissolved or suspended in water, or associated with sediment including nutrients, heavy metals, and oil and grease, can accumulate and wash away from ISAs. The percentage of impervious surface estimated from satellite RS data can be used to assess the spatial extent of urban land use, as well as urban development density.[42]

Building energy saving is defined as the rational use of energy and constant improvement of energy efficiency under conditions that guarantee and improve the comfort of the building. Using Landsat TM thermal data, surface radiance values for each scene pixel were calculated with the use of calibration coefficients to identify build structure, building material, vegetation, and heating methods.[43]

Night light emissions that originate mainly from large urban areas are among the main elements of environmental pollution: disturbance of biological rhythms, psychological effects, and environmental degradation. The Defense Meteorological Satellite Program, run by the NOAA, has the capability to detect faint sources of visible near infrared (VNIR) emissions on the earth's surface, making it possible to detect cities and towns. Modern GIS packages offer advanced tools for visibility analysis, including line-of-sight estimations and viewshed mapping. The viewshed identifies the cells in an input raster that can be seen from one or more observation points. Moreover, it is possible to identify areas affected with indirect light pollution.[44]

Unlike the United States, where the use of asbestos is still legal but tightly controlled, on January 1, 2005 (following the directive 76/769/CEE), the marketing and use of products containing asbestos was banned through-out the European Union. Prohibition is due to the possible diffusion of asbestos fibers into the environment constituting a health hazard leading to asbestosis (mesothelioma) cancer. Hyperspectral sensors, characterized by a high spectral and/or spatial resolution, allow reliable quantitative measurements of specific absorption features of urban materials. Spectral analyses can distinguish the asbestos spectrum shape from other roofing materials and backgrounds. Moreover, the classification of the TIR emissivity dataset permits good discrimination between buildings and open spaces and among different roof types, even though residual false positives were still present due to difficulties in the temperature/emissivity separation procedure.[45]

Around oil fields, a considerable part of lighter hydrocarbon fraction normally evaporates, while the remaining fraction in liquid form could be partially recovered. The scars of the damage caused by the spillage of oil remains as soot, tarmats, and tarcrete, which are unrecovered denser hydrocarbon fractions. The surface area covered by the "oil lakes" continues to shrink visibly, leading to the suggestion that there is a disappearance of oil-polluted surfaces. However, thermal RS data provides evidence that the spatial spread of hydrocarbon-contaminated surfaces is much more than previously reported. The addition of hydrocarbons alters soil composition, and these compositional changes are reflected as temperature variations. Landsat TM data has been processed to obtain the LST. Emissivity affects the apparent temperature due to changes in the thermal properties of materials (conductivity, density, capacity, and inertia). The effect of darkness will not affect the LST observations, so waste dumping during night-time can be easily identified for pollution control and environmental management.[46]

Kuwait's oil lakes and oil-polluted surfaces were an act of sabotage imposed on the desert environment during the 1990 to 1991 Arabian Gulf War. The oil formed networks of oil rivers and lakes that accumulated around oil wells and in relatively low areas. The majority of Kuwait's producing oil fields have high reservoir pressures that maintain natural flowing oil from the reserves to the surface. During the 10 months of oil fires, some of the oil fields, especially the Greater Burgan, were more or less inundated with crude oil. The interpretation of high-resolution satellite imagery (Landsat TM, Spot, and Indian Remote Sensing IRSI) shows that contaminated areas demonstrated remarkable resilience, in terms of vegetation growth, seven years after the burning of the oil wells, with increased rainfall from 1991 to 1995. The stabilization of the top sandy soil with tarmat and soot prevented the blowing away of plant seeds, and therefore increased the chances of vegetation growth. The soot, which later turned into black soil, did not appear to be a hindrance to vegetation growth. Oil lakes, tarmats, and black soil continue to decrease and are not always readily observed on the surface. However, harmful chemicals remain in the soil.[47]

In understanding environmental contamination/pollution, one must understand the trajectory of heavy metals in different environmental compartments, which can be seen as different morphological regions (relief forms) shown in RS images. A digital terrain model, generated from the available stereo-pair bands of the ASTER image in the VNIR system, helps aid visualization and interpretation through 3-D observation of different rock types, as well as geomorphometric information (elevation, declivity, and aspect) of the slopes, which potentially reside in, feed, and conduct the landfill wastes.[48] The GIS and RS have an important role in the linkage and analysis of changes in land use-land change (LULC) patterns data, in particular for detection (direct and indirect), extrapolation and interpretation, area calculation, and monitoring. Information on LULC changes is required to achieve environmentally sound management and decision making for future developments.[49]

Air

Over recent years, Earth observation (EO) satellite sensors sent into orbit with improved temporal and spatial resolution, together with enhanced radiometric accuracy, has led to the possibility of using satellite data within the framework of air quality analysis. In this context, the synoptic view and the daily/hourly repetition cycle of satellite observations can provide the monitoring of transboundary air pollution. Moreover, they can allow estimates of air pollutant concentration fields characterized by a more homogeneous spatial distribution and with a more complete coverage of the domain of interest with respect to those carried out by sparse in situ samplings. The processing of the radiance measured by satellite sensors in the spectral range from 0.4 to 2.1 μm supplies information on the atmospheric aerosol loading in terms of columnar aerosol optical depth (AOD) or aerosol optical thickness (AOT) as a measure of light extinction by aerosol in the atmosphere (typically at 0.55 μm), during their overpass time.[50]

Low and moderate spatial resolution satellite sensors have already shown their capability in tracking air pollution in general and aerosols particularly at a global scale. Sensors with moderate to high spatial resolution (such as MODIS and MERIS) seem also to be appropriate for aerosol retrieval at a regional scale by using the differential textural analysis (DTA) code. The DTA code quantifies the contrast reduction as local "textural degradation" on geometrically and radiometrically corrected satellite images.

The code, applied to a set of geo-corrected images, is able to retrieve and map AOT values relative to a reference image assumed to be clean of pollution with a homogeneous atmosphere.[51] As the total atmospheric optical thickness depends on molecular structure (Rayleigh scattering) and aerosols, it is possible to separate the contribution of the "natural" atmospheric molecules from "contaminants" such as aerosols. The AOT can be defined as the integrated extinction coefficient over a vertical column of unit cross section of atmosphere.

Aerosols smaller than the wavelength of satellite sensors cannot be detected, so only coarse particles more than a few micrometers can be detected by MODIS bands. Band combinations of 1, 3, and 7 of MODIS image were used to extract maps of aerosols. Bands 1 and 3, covering the optical region of the electromagnetic spectrum (0.459–0.876 μm), are used to collect information on aerosols and particulate matters, while band 7 covering the infrared region (2.105– 2.55 μm) is used for calibration purposes only.[52]

Aerosol optical thickness in the visible (or atmospheric turbidity) can be considered as an overall air pollution indicator in urban areas, among others, because during photochemical pollution episodes, light extinction is due to particles, while only the yellow-brownish coloration of smog is due to NO_2. Moreover, the presence of particles in the atmosphere always causes an increase of the extinction coefficient, which is strongly correlated with small particle concentration.

Atmospheric turbidity, which expresses overall particulate pollution levels in urban areas, can be extracted by comparing multitemporal high-resolution data. The main prerequisite to carry out this comparison is that no change has arisen in the ground intrinsic reflectance between the data sets used. The noise introduced by changes in ground reflectance is reduced by the combination of two complementary procedures based on physically independent optical effects that are induced by the aerosols in different spectral areas.[53]

Vegetation fire is a global phenomenon that affects large areas and a variety of biomes of the world, from circumboreal forests to the tropical belt through the woodland and shrubland of temperate regions.[54] Since the late 1970s, prescribed and wild vegetation fires have been recognized as a major source of atmospheric trace gases and aerosol particles that affect the composition of the atmosphere, the global climate, and air pollution.[55] In an exceptional event, large vegetation fires raged throughout the Indonesian archipelago in 1997 and 1998, causing a smog blanket that covered more than 3 million square kilometers, with economic losses estimated at over U.S. $4.5 billion.[56] Both active fires and the extent of burned areas can be observed from satellite. Fire radiative power is measured over active fires, and is provided by the MODIS sensor and by the SEVIRI (Spinning Enhanced Visible and InfraRed Imager) sensor onboard the Meteosat Second Generation satellite. Fire radiative power is a very good candidate for any assessment of emission from fires, although it relies on active fires observed by the EO systems, which are a temporal sampling of burning activity. It is consequently possible to highlight similarities and differences in the seasonality and geographical distribution of emission at the global and continental levels.[57] In addition to direct damage on vegetation, devastating fires produced large quantities of gaseous air pollutants and particulate matter (PM_{10} and $PM_{2.5}$, airborne particles smaller than 10 and 2.5 mm in size, respectively) dispersed over the surrounding region, increasing respiratory diseases, asthma, bronchitis, and eye irritation.

The MODIS sensor can measure aerosol abundance and size over land and water with nearly global coverage at moderate spatial resolutions. In addition, MISR (Multiangle Imaging SpectroRadiometer), on board the EOS satellites, is able to provide information on aerosol type and plume top heights. Particle information retrieved by satellite sensors may be suitable for monitoring the spatial and temporal trends of particle concentrations over large geographical areas.[58] The MODIS aerosol optical properties were used in a semi-empirical approach to estimate $PM_{2,5}$ content at ground level. Comparison with daily $PM_{2,5}$ sampled on the ground showed good agreement, with the satellite-based concentrations tending to underestimate the values by at most 20%.

Substantive results suggest that changes in AOD have a significant impact on infant mortality due to respiratory diseases, providing evidence that air pollution's adverse effects, although nonlinear, are not only present in large cities, but also in lower pollution settings that lack ground measures of pollution. A lack of reliable measures of air pollution across wide geographic areas hampers research on pollution's effects on health.[59] Meteorological parameters (air pressure, air temperature, relative humidity, and wind velocity) can influence the estimation of PM from MODIS AOT data. Air quality is usually monitored at fixed ground stations. Although ground measurements are able to indicate the concentration level of air pollutants and their temporal variations precisely, this method of monitoring is limited by its huge expense and sparse spatial coverage. To complement these sparse precise measurements with the estimates derived from satellite, it is more reliable to monitor PM10 from MODIS AOT data at a high temperature and relative humidity, but at low pressure and wind velocity. There is an inverse relationship between the pollution level and the accuracy at which it is monitored from the MODIS data. These findings should serve as useful guidance in selecting the appropriate meteorological conditions under which air pollution can be monitored reliably from MODIS AOT data.[60]

The problem of particulate pollution in the atmosphere has attracted new interest given the recent scientific evidence of the ill-health effects of small particles. Aerosol optical thickness in the visible (or atmospheric turbidity), which is defined as the linear integral of the extinction coefficient due to small airborne particles, can be considered an overall air pollution indicator in urban areas. Reflectance in the visible bands measured by satellite can be used to derive the PM10. In fact, the reflectance at the TOA (top of atmosphere) measured from the satellite is subtracted by the amount given by the surface reflectance to obtain the atmospheric reflectance. The atmospheric reflectance derived from Landsat TM signals can be used as independent variables in a calibration regression analysis to measure the PM10 values.[61]

Using meteorological and other ancillary datasets, an empirical relationship between AOT and $PM_{2.5}$ mass can be obtained, assessing the effects of wind speed, cloud cover, and mixing height on particulate matter air quality. The analysis shows that the $PM_{2.5}$–AOT relationship strongly depends on aerosol concentrations, ambient relative humidity, fractional cloud cover, and height of the mixing layer.[62]

Ground level observations indicated that PM concentration varies widely across different regions, which was mainly due to the difference in weather conditions and anthropogenic emissions. Results showed that MODIS AOD had a better positive correlation with the coincident hourly average PM concentration than with daily average, due to diurnal variation in PM mass measurements. After correcting AOD for relative humidity, the correlation did not improve significantly, suggesting that the humidity was not the main factor affecting the correlation of PM with AOD. The statistical regression analysis between MODIS AOD and PM mass suggested that the satellite-derived AOD is a useful tool for mapping PM distribution over large spatial domains.[63]

A recent method for multi-objective optimization of air quality monitoring systems based on satellite RS of the troposphere uses atmospheric turbidity as a surrogate for air pollution loading. The values of AOD have been extracted by images obtained by sensors on board in-polar-orbit Earth satellites (specifically SPOT 5 and Landsat 7) at the time of their daily overpass. Through an image-processing algorithm, the spatial distribution of optical depth over the whole area is obtained. Assimilating the optical depth field with the field of the mixing layer height, the scattering coefficient of the lowermost part of tropospheric aerosol was reckoned. A physico-chemical and multiphase thermodynamic equilibrium model was used to estimate the secondary aerosol formation and primary aerosol. Using non-linear multiple regression, the experiential relationship between the scattering coefficient of primary and secondary aerosol, its ambient air concentration, and relative humidity can be calculated.[64]

Ozone (O_3) is one of the most dangerous of the phytotoxic air pollutants. Its effects on plants include reductions in photosynthesis, visible leaf injury, growth limitation, and accelerated senescence. Advanced RS techniques using hyperspectral sensors demonstrated the feasibility of detecting the stress in its early phase by monitoring excess energy dissipation pathways, such as chlorophyll fluorescence and non-photochemical quenching.[65] The total ozone mapping spectrometer (TOMS) measures the reflected spectrum from the earth to estimate total column ozone thickness, including maps of erythemal (biologically damaging) UV reaching the earth's surface. Recently, TOMS has been replaced by the ozone monitoring instrument (OMI), which is on board the NASA satellite Aura. The OMI can distinguish between aerosol types such as smoke, dust, and sulfates, and can measure cloud pressure and coverage, providing data to derive tropospheric ozone.[4]

An important indicator for volcanic activity is the emission of trace gases such as sulfur dioxide (SO_2). During an eruption, SO_2 is the third most abundant gas found in volcanic plumes, after H_2O and CO_2. Changes in SO_2 flux can be a precursor for the onset of volcanic activity. In addition, SO_2 is also produced by anthropogenic activities such as power plants, refineries, metal smelting, and the burning of fossil fuels; however, its atmospheric background level is usually very low. Satellite-based instruments operating in the UV spectral region have played an important role in monitoring and quantifying volcanic SO_2 emissions. The TOMS was the first satellite instrument to detect volcanic SO_2 released during the El Chichon eruption in 1982. The detection limit to measure volcanic and anthropogenic SO_2 greatly improved for the Global Ozone Monitoring Experiment (GOME), on board the ERS-2 satellite, and the Scanning Imaging Spectrometer for Atmospheric Cartography (SCIAMACHY), on board the ENVISAT satellite. Last but not least, the Infrared Atmospheric Sounding Interferometer (IASI) on MetOp-A is also able to detect SO_2 with an excellent global coverage in combination with small footprints. Satellite-based observations provide valuable information for detecting and tracking eruption plumes, and therefore minimize the risk of aircraft encounter with hazardous volcanic clouds. GOME-2 has higher sensitivity, especially for SO_2 at lower altitudes, whereas IASI also offers night-time observations, a higher spatial resolution, and an estimation of the altitude of the plume. With commercial and freight air traffic growing globally, the risk of aircraft encounter with hazardous volcanic clouds is increasing, as many volcanoes are not regularly monitored and atmospheric winds can rapidly distribute ash and gas.[66] In fact, once aloft, winds can transport the ash and gases rapidly and in multiple directions, depending on the wind speed and wind shear (the change in wind direction with height). The ash can cause extensive damage to aircraft, stalling engines, abrading windscreens and damaging sensitive avionics equipment. Volcanic gases, specifically SO_2, may also pose a hazard to aircraft, and because gas and ash have different specific gravities they may separate and travel at different speeds and heights in a sheared atmosphere.[67] The 2010 eruptions of the volcano Eyjafjallajökull in Iceland, which, although relatively small for volcanic eruptions, caused enormous disruption to air travel across western and northern Europe. About 20 countries closed their airspace (a condition known as ATC Zero), and it affected hundreds of thousands of travelers.

Conclusion

Remote sensing is nowadays the most applicable tool not only for detection, monitoring, and tracking of pollutants, but also for damage assessment where damage is demonstrable by its effect on the environment.

There is a physical limitation that can never be overcome. Taking marine pollution as an example, it is only possible to differentiate between surfaces (i.e., oil slick and clear water) if they reflect or emit radiation in different proportions; that is, if they have different spectral signatures. This physical limitation means that toxins in water, for example, cannot be detected unless they have an effect on something that is "visible" to the sensor, such as phytoplankton. Another limitation is that electromagnetic radiation has very little power to penetrate objects; for example, it is almost entirely absorbed at the water surface for wavelengths of NIR and longer, and can only penetrate a matter of meters (approximately 15–30 m for very clear water) in the visible wavelengths. This means that RS can only operate on the surface skin of oceans and seas. Fortunately, this is where most marine pollution is focused.[68]

Recent developments in microwave RS, theory, and sensor availability have resulted in new potential and capabilities, such as the ability to extract/detect subsurface parameters and features using these techniques. More and more research is required to refine and implement these approaches. Consequently, there is a need to develop an optimal sensor system including both active and passive microwave techniques for more effective monitoring of the environment. It will allow a range of applications and the synergism of the two types of measurements to provide more useful and new information. Such studies will not only enhance and refine RS applications in environment studies, but will also significantly contribute to the sensor development program.

Finally, the integration of expert systems and spatial decision support systems with GIS is a very interesting area of research, aiding effective and timely decision making concerning the planning, design, analysis, operation, and maintenance of environmental resources systems. Moreover, these tools can greatly reduce the time and effort required in traditional approaches, automating the process of solving pollution problems and helping the selection of cost-effective management alternatives.

References

1. Sabins, F.F. Fundamental considerations. In *Remote Sensing Principles and Interpretation*; W.H. Freeman and Co.; San Francisco, 1978; 1–16.
2. Campbell, J.B. History and scope of remote sensing. In *Introduction to Remote Sensing,* 2nd Ed.; Taylor & Francis; London, 1966; 1–21.
3. Brivio, P.A.; Lechi, G.; Zilioli, E. *Principi e Metodi di Telerilevamento;* Città Studi Ed.; Milan, Italy, 2006.
4. Mumby, P.J.; Skirving, W.; Strong, A.E.; Hardy, J.T.; LeDrew, E.F.; Hochberg, E.J.; Stump, R.P.; David, L.T. Remote sensing of coral reefs and their physical environment. Mar. Pollut. Bull. **2004**, *48,* 219–228.
5. Kabbara, N.; Benkhelilb, J.; Awadc, M.; Barale, V. Monitoring water quality in the coastal area of Tripoli (Lebanon) using high-resolution satellite data. ISPRS J. Photogramm. Remote Sens. **2008**, *63,* 488–495.
6. Sesli, F.A.; Karsli, F.; Colkesen, I.; Akyol, N. Monitoring the changing position of coastlines using aerial and satellite image data: An example from the eastern coast of Trabzon, Turkey. Environ. Monit. Assess. **2009**, *153,* 391–403.
7. Dekker, A.; Brando, V.; Anstee, J.M. Retrospective sea-grass change detection in a shallow coastal tidal Australian lake. Remote Sens. Environ. **2005**, *97,* 415–433.
8. Durand, D.; Bijaoui, J.; Cauneau, F. Optical remote sensing of shallow-water environmental parameters: A feasibility study. Remote Sens. Environ. **2000**, *73*, 152–161.
9. Barale, V.; Jaquet, J.M.; Ndiaye, M. Algal blooming patterns and anomalies in the Mediterranean Sea as derived from the SeaWiFS data set (1998–2003). Remote Sens. Environ. **2008**, *112,* 3300–3313.
10. Tetley, M.J.; Mitchelson-Jacob, E.G.; Robinson, K.P. The summer distribution of coastal minke whales (*Balaenoptera acutorostrata*) in the southern outer Moray Firth, NE Scotland, in relation to co-occurring mesoscale oceanographic features. Remote Sens. Environ. **2008**, *112,* 3449–3454.
11. Panigada, S.; Zanardelli, M.; MacKenzie, M.; Donovan, C.; Mélin, F.; Hammond, P.S. Modelling habitat preferences for fin whales and striped dolphins in the Pelagos Sanctuary (Western Mediterranean Sea) with physiographic and remote sensing variables. Remote Sens. Environ. **2008**, *112,* 3400–3412.

12. Li, Y.; Li, J. Oil spill detection from SAR intensity imagery using a marked point process. Remote Sens. Environ. **2011**, *114,* 1590–1601.
13. Cheng, Y.; Li, X.; Xu, Q.; Garcia-Pineda, O.; Andersen, O.B. SAR observation and model tracking of an oil spill event in coastal waters. Mar. Pollut. Bull. **2011**, *62,* 350–363.
14. Krohling, R.A.; Campanharo, V.C. Fuzzy TOPSIS for group decision making: A case study for accidents with oil spill in the sea. Expert Syst. Appl. **2011**, *38,* 4190–4197.
15. Keramitsoglou, I.; Cartalis, C.; Kiranoudis, C.T. Automatic identification of oil spills on satellite images. Environ. Model. Softw. **2006**, *21,* 640–652.
16. Topouzelis, K.; Karathanassi, V.; Pavlakis, P.; Rokos, D. Detection and discrimination between oil spills and look-alike phenomena through neural networks. ISPRS J. Photogramm. Remote Sens. **2007**, *62,* 264–270.
17. Cococcioni, M.; Corucci, L.; Lazzerini, B. *Issues and Preliminary Results in Oil Spill Detection Using Optical Remotely Sensed Images*, Proceedings of International Workshop on the Analysis of Multi-Temporal Remote Sensing Images, 2007. MultiTemp 2007, Leuven, Belgium, 18–20 July 2007; Eds.; IEEE: New York, 2007.
18. Casciello, D.; Lacavat, T.; Pergolat, N.; Tramutoli, V. *Robust Satellite Techniques* (*RST*) *for Oil Spill Detection and Monitoring*, Proceedings of International Conference on Geoinformatics IEEE, Fairfax, USA, August 12–14, 2009; Eds.; IEEE: New York, 2009.
19. Yu, I.A.; Zatyagalova, V.V. A GIS approach to mapping oil spills in a marine environment. Int. J. Remote Sens. **2008**, *29* (21), 6297–6313.
20. Dahish, A.S.; Ahmad, A. In *An Application of Geographical Information System and Remote Sensing Techniques for Detection of Oil Spill,* Proceedings of International Conference on Geoinformatics IEEE, Fairfax, USA, August 12–14, 2009; Eds.; IEEE: New York, 2009.
21. Jha, M.K.; Chowdhury, A.; Chowdary, V.M.; Peiffer, S. Groundwater management and development by integrated remote sensing and geographic information systems: Prospects and constraints. Water Resour. Manage. **2007**, *21,* 427–467.
22. Dimitriou, E.; Zacharias, I. Groundwater vulnerability and risk mapping in a geologically complex area by using stable isotopes, remote sensing and GIS techniques. Environ. Geol. **2006**, *51,* 309–323.
23. Montzka, C.; Canty, M.; Kreins, P.; Kunkel, R.; Menz, G.; Vereecken, H.; Wendland, F. Multispectral remotely sensed data in modelling the annual variability of nitrate concentrations in the leachate. Environ. Model. Softw. **2008**, *23,* 1070–1081.
24. Bresciani, M.; Giardino, C.; Longhi, D.; Pinardi, M.; Bartoli, M; Vascellari, M. Imaging spectrometry of productive inland waters. Application to the lakes of Mantua. Ital. J. Remote Sens. **2009**, *41* (2), 147–156.
25. Giardino, C.; Bartoli, M.; Candiania, G.; Bresciani, M.; Pellegrini, L. Recent changes in macrophyte colonisation patterns: An imaging spectrometry-based evaluation of southern Lake Garda (northern Italy). J. Appl. Remote Sens. **2007**, *1*, 1–17.
26. Giardino, C.; Oggioni, A.; Bresciani, M.; Yan, H. Remote sensing of suspended particulate matter in Himalayan lakes. A case study of Alpine lakes in the Mount Everest region. Mt. Res. Dev. **2010**, *30* (2), 157–168.
27. Bresciani, M.; Stroppiana, D.; Fila, G.; Montagna, M.; Giardino, C. Monitoring reed vegetation in environmentally sensitive areas in Italy. Ital. J. Remote Sens. **2009**, *41* (2), 125–137.
28. Bochenek, Z.; Ciolkosz, A.; Iracka, M. Deteriorations of forests in the Sudety Mountains, Poland, detected on satellite images. Environ. Pollut. **1998**, *101,* 163–168.
29. Kooistraa, L.; Salasc, E.A.L.; Cleversc, G.P.W; Wehrensb, R.; Leuvena, R.S.E.W.; Nienhuisa, P.H.; Buydensb, L.M.C. Exploring field vegetation reflectance as an indicator of soil contamination in river floodplains. Environ. Pollut. **2007**, *127,* 281–290.
30. Nikolov, N.; Zeller, K. Efficient retrieval of vegetation leaf area index and canopy clumping factor from satellite data to support pollutant deposition assessments. Environ. Pollut. **2006**, *141*, 539–549.

31. Chen, J.M.; Cihlar, J. Retrieving leaf area index of boreal conifer forests using Landsat TM images. Remote Sens. Environ. **1996**, *55*, 153–162.
32. Boschetti, M.; Brivio, P.A.; Carnesale, D.; Di Guardo, A. The contribution of hyperspectral remote sensing to identify vegetation characteristics necessary to assess the fate of persistent organic pollutants (POPs) in the environment. Ann. Geophys. **2006**, *49,* 177–186.
33. Zhengjun, W.; Jianming, H.; Guisen, D. Use of satellite imagery to assess the trophic state of Miyun Reservoir, Beijing, China. Environ. Pollut. **2008**, *155,* 13–19.
34. Tuominen, J.; Lipping, T.; Kuosmanen, V. *Assessment of ENVI Forest heath Tool in Detection of Dust and Seepage*, Proceedings of IEEE International Geoscience & Remote Sensing Symposium, Boston, USA, July 06–11, 2008; Eds.; IEEE: New York, 2008.
35. Rastmanesh, F.; Moore, F.; Kharrati-Kopaei, M.; Behrouz, M. Monitoring deterioration of vegetation cover in the vicinity of smelting industry, using statistical methods and TM and ETM+ imageries, Sarcheshmeh copper complex, Central Iran. Environ. Monit. Assess. **2010**, *163,* 397–410.
36. Boschetti, M.; Brivio, P.A.; Carnesale, D.; Di Guardo, A. The contribution of hyperspectral remote sensing to identify vegetation characteristics necessary to assess the fate of Persistent Organic Pollutants (POPs) in the environment. Annals of Geophysics, **2006**, *49,* 177–186.
37. Cinnirella, S.; Pirrone, N. Spatial and temporal distributions of mercury emissions from forest fires in Mediterranean region and Russian federation. Atmos. Environ. **2006**, *40,* 7346–7361.
38. Dunagan, S.C.; Gilmore, M.S.; Varekamp, J.C. Effects of mercury on visible/near-infrared reflectance spectra of mustard spinach plants (*Brassica rapa* P.). Environ. Pollut. **2007**, *148,* 301–311.
39. Joshi, P.K.; Kumar, M.; Paliwal, A.; Midha, N.; Dash, P.P. Assessing impact of industrialization in terms of LULC in a dry tropical region (Chhattisgarh), India using remote sensing data and GIS over a period of 30 years. Environ. Monit. Assess. **2009**, *149,* 371–376.
40. Rafiee, R.; Mahiny, A.S.; Khorasani, N. Assessment of changes in urban green spaces of Mashad city using satellite data. Int. J. Appl. Earth Obs. Geoinf. **2009**, *11*, 431–438.
41. Meng, D.; Gong, H.; Li, X.; Zhao, W.; Gong, Z.; Zhu, L.; Hu, D. *Study of Urban Heat Island Based on Remote Sensing in Beijing-Capital Zone*, Proceedings of Urban Remote Sensing Joint Event, Shanghai, China, May 20–22, 2008; Eds.; IEEE: New York, 2009.
42. Xian, G.; Crane, M.; Su, J. An analysis of urban development and its environmental impact on the Tampa Bay watershed. J. Environ. Manage. **2007**, *85*, 965–976.
43. Huiping, H.; Bingfang, W.; Jingjing, Z.; Yuemin, Z. *The Method and Applications of Remote Sensing in Urban Building Heating-loss Monitoring,* Proceedings of Urban Remote Sensing Joint Event, Shanghai, China, May 20–22, 2008; Eds.; IEEE: New York, 2009.
44. Chalkias, C.; Petrakis, M.; Psiloglou, B.; Lianou B. Modelling of light pollution in suburban areas using remotely sensed imagery and GIS. J. Environ. Manage. **2006**, *79,* 57–63.
45. Bassani, C.; Cavalli, R.M.; Cavalcante, F.; Cuomo, V.; Palombo, A.; Pascucci, S.; Pignatti, S. Deterioration status of asbestos-cement roofing sheets assessed by analyzing hyperspectral data. Remote Sens. Environ. **2007**, *109,* 361–378.
46. Saif ud din; Al Dousari, A.; Literathy, P. Evidence of hydrocarbon contamination from the Burgan oil field, Kuwait—Interpretations from thermal remote sensing data. J. Environ. Manage. **2008**, *86,* 605–615.
47. Kwarteng, A.Y. Remote sensing assessment of oil lakes and oil-polluted surfaces at the Greater Burgan oil field, Kuwait. Int. J. Appl. Earth Obs. Geoinf. **1999**, *1*, 36–47.
48. de Oliveira, M.T.G.; Rolim, S.B.A; de Mello-Farias, P.C; Meneguzzi, A.; Lutckmeier, C. Industrial pollution of environmental compartments in the Sinos River Valley, RS, Brazil: Geochemical-biogeochemical characterization and remote sensing. Water Air Soil Pollut. **2008**, *192,* 183–198.
49. Doygun, H.; Alphan, H. Monitoring urbanization of Iskenderun, Turkey, and its negative implications. Environ. Monit. Assess. **2006**, *114,* 145–155.

50. Di Nicolantonio, W.; Cacciari, A.; Tomasi, C. Particulate matter at surface: Northern Italy monitoring based on satellite remote sensing, meteorological fields, and in-situ samplings. IEEE J. Select. Top. Appl. Earth Obs. Remote Sens. **2009**, *2*, 284–292.
51. Retalis, A.; Sifakis, N. Urban aerosol mapping over Athens using the differential textural analysis (DTA) algorithm on MERIS-ENVISAT data. ISPRS J. Photogramm. Remote Sens. **2010**, *65,* 17–25.
52. Sohrabiniaa, M.; Khorshiddoust, A.M. Application of satellite data and GIS in studying air pollutants in Tehran. Habitat Int. **2007**, *31,* 268–275.
53. Sifakis, N.I. Quantitative mapping of air pollution density using Earth observations: A new processing method and application to an urban area. Int. J. Remote Sens. **1998**, *19* (17), 3289–3300.
54. Brivio, P.A.; Maggi, M.; Binaghi, E.; Gallo, I. Mapping burned surfaces in Sub-Saharan Africa based on multi-temporal neural classification. Int. J. Remote Sens. **2003**, *24,* 4003–4018.
55. Crutzen, P.J.; Andreae, M.O. Biomass burning in the tropics: Image on atmospheric chemistry and biogeochemical cycles. Science **1990**, *250,* 1669–1678.
56. Stolle, F.; Chomitz, K.M.; Lambin, E.F.; Tomich, T.P. Land use and vegetation fires in Jambi Province, Sumatra, Indonesia. For. Ecol. Manage. **2003**, *179,* 277–292.
57. Stroppiana, D.; Brivio, P.A; Gregoire, G.M.; Liousse, C.; Guillaume, B.; Granier, C.; Mieville, A.; Chin, M.; Petron, G. Comparison of global inventories of monthly CO emissions derived from remotely sensed data. Atmos. Chem. Phys. Discuss. **2010**, *10,* 17657–17697.
58. Liu, Y.; Kahn, R.A.; Chaloulakou, A.; Koutrakis, P. Analysis of the impact of the forest fires in August 2007 on air quality of Athens using multi-sensor aerosol remote sensing data, meteorology and surface observations. Atmos. Environ. **2009**, *43,* 3310–3318.
59. Gutierrez, E. Using satellite imagery to measure the relationship between air quality and infant mortality: An empirical study for Mexico. Popul. Environ. **2010**, *31,* 203–222.
60. Zha, Y.; Gao, J.; Jiang, J.; Lu, H.; Huang, J. Monitoring of urban air pollution from MODIS aerosol data: Effect of meteorological parameters. Tellus **2010**, *62B,* 109–116.
61. Lim, H.S.; MatJafri, M.Z.; Abdullah, K.; Saleh, N.M.; Wong, J. *Extracting Spatial Data From Satellite Sensor to Support Air Pollution Determination Using Remote Sensing Technique,* Proceedings of International Geoscience and Remote Sensing Symposium, Barcelona, Spain, July 23–27, 2007; Eds.; IEEE: New York, 2007.
62. Gupta, P; Christophera, S.A.; Wang, J.; Gehrig, R.; Leed, Y.; Kumar, N. Satellite remote sensing of particulate matter and air quality assessment over global cities. Atmos. Environ. **2006**, *40,* 5880–5892.
63. Guo, J.P.; Zhang, X.Y.; Che, H.Z.; Gong, S.L.; An, X.; Cao, C.X.; Guang, J.; Zhang, H.; Wang, Y.Q.; Zhang, X.C.; Xue, M.; Li, X.W. Correlation between PM concentrations and aerosol optical depth in eastern China. Atmos. Environ. **2009**, *43,* 5876–5886.
64. Sarigiannis, D.A.; Saisana, M. Multi-objective optimization of air quality monitoring. Environ. Monit. Assess. **2008**, *136*, 87–99.
65. Meroni, M.; Panigada, C.; Rossini, M.; Picchi, V.; Cogliati, S.; Colombo, R. Using optical remote sensing techniques to track the development of ozone-induced stress. Environ. Pollut. **2009**, *157,* 1413–1420.
66. Rix, M.; Valks, P.; Hao, N.; van Geffen, J.; Clerbaux, C.; Clarisse, C.; Coheur, P.F.; Loyola D.G.; Erbertseder, T.; Zimmer, Z.; Emmadi, S. Satellite monitoring of volcanic sulfur dioxide emissions for early warning of volcanic hazards. IEEE J. Select. Topics Appl. Earth Obs. Remote Sens. **2009**, *2* (3), 196–206.
67. Prata, A.J. Satellite detection of hazardous volcanic clouds and the risk to global air traffic. Nat. Hazards **2009**, *51,* 303–324.
68. Clark, C.D. Satellite remote sensing of marine pollution. Int. J. Remote Sens. **1993**, *14* (16), 2985–3004.

26

Solid Waste Management: Life Cycle Assessment

Ni-Bin Chang,
Ana Pires, and
Graça Martinho

Introduction

In the end of 2005, the European Commission (EC) published a proposal for a directive on waste [COM (2005) 667,[1] which resulted in the publishing of the new Waste Framework Directive[2]] that is currently under implementation in European institutions. One element of the proposal is the establishment of a hierarchy of principles (i.e., waste hierarchy principle hereafter) to be applied for waste management. The commission proposal includes the following reference to the hierarchy:

> The Member States are to take measures, as a matter of priority, for the prevention or reduction of waste production and its harmfulness and, secondly, for the recovery of waste by means of re-use, recycling and other recovery operations.

It provides for a flexible application of the hierarchy and does not explicitly make a suggestion between the d…l/l/l/l/k,ifferent recovery options such as reuse, recycling, and recovery of energy, making the policy blurred. Later on, with the publishing of the new Waste Framework Directive 2008/98/EC,[1] resulting from COM (2005) 667, a five-step hierarchy was proposed to encourage the member states in the EC to take measures, as a matter of priority, for the following: 1) prevention; 2) preparing for reuse; 3) recycling; 4) other recovery, e.g., energy recovery; and 5) disposal. The waste hierarchy principle for prioritizing waste management options has touched the base of pollution prevention and resource conservation. Yet recent climate change impacts and resource scarcity made some of these options for solid waste management (SWM) no longer risk informed and forward looking. As a consequence, another amendment was introduced that seeks to support the flexibility of the waste hierarchy principle by stating the following:

> When applying the waste hierarchy [...], Member States shall take measures to encourage the options that deliver the best overall environmental outcome. This may require specific waste streams departing from the hierarchy where this is justified by life-cycle thinking on the overall impacts of the generation and management of such waste."

Given that extensive studies are required to deviate from the hierarchy, modern sustainable SWM strategies must be tied to an all-inclusive assessment metrics covering socioeconomic aspects, cost–benefit analysis, technical feasibility, public health requirements, environmental impacts, ecological footprint, carbon footprint, and even sociopolitical considerations. Hence, sustainable SWM has been deemed as a renewed paradigm, in which understanding how sustainable decisions in association with life cycle assessment (LCA) can be reached in a societal context becomes a critical task.

Multicriteria decision making (MCDM) processes have been helpful in translating the features of real-world problems to decision-making processes. The problems of MCDM can be broadly classified into two categories: multiple-attribute decision making (MADM) and multiple-objective decision making (MODM). The main features that differentiate both methods are the type of criteria (attributes or objectives) and the number of alternatives. Whereas MADM has a finite number of alternatives, MODM has an infinite number of alternatives. With the prescribed waste management alternatives in our study by which the salient features for sustainable SWM can be retrieved distinctly, MADM was thus selected to aid in our decision analysis. To explore the decision making with respect to a suite of criteria, this entry presents a comparative study through two different MADM approaches for the purpose of demonstration. They include the simple additive weight (SAW) and the technique for order preference by similarity to ideal solution (TOPSIS). The practical implementation was assessed by a case study in Setúbal peninsula, Portugal, where there is an acute need to improve the quality of SWM to achieve a sustainable solution.

Life Cycle Assessment for Solid Waste Management

Life cycle assessment is a compilation and evaluation of the inputs, outputs, and potential environmental impacts of a product system throughout its life cycle.[3] Thus, LCA is a tool for the analysis of the environmental burden of products at all stages in their life cycle—from the extraction of resources, through the production of materials, product parts, and the product itself, and the use of the product to the management after it is discarded, either by reuse, recycling, or final disposal (in effect, therefore, "from cradle to grave").[4]

The concept of LCA was produced in the 1960s as a way to cumulatively account for energy use leading to the projection of future resource supplies and demand.[5] Since the 1990s, LCA has been applied in a broad range of different fields including SWM. The ISO 14040 standard[3] points out the main areas of application of LCA, including the identification of improvement possibilities, decision making, selection of relevant environmental performance indicators, and marketing. The most influential momentum in the context of ISO 14040 applied for SWM is to explore how the decision-making processes in a social context can make the final option different.

The salient use of LCA for SWM was conducted in Denmark, aiming to meet the need to properly manage the packaging waste embedded in the packaging materials.[6–10] Life cycle assessment may also be applicable in conducting intercomparisons among waste treatment alternatives for specific waste streams, such as recycling of cardboard[11] and recycling vs. incineration of scrap paper.[12] Earlier applications of LCA for SWM focused on the evaluation of waste hierarchy principle in some industrialized countries during 1990.[13] In the literature,[14] it is clear that waste hierarchy principle cannot be used to substitute thorough assessment in most cases. Once LCA is capable of promoting a holistic view of

SWM, several combinations of different technologies can be meaningfully applied to support essential SWM against the contradictory suggestions in early waste hierarchy paradigm. For this reason, LCA has been recommended by the European Union Waste Framework Directive[2] to verify if waste hierarchy principle is the best solution, particularly from the environmental point of view.

The applications of LCA tools for waste management appeared mainly in Europe and United States for the screening of technologies under the umbrella of the integrated solid waste management (ISWM) systems. Several LCA models were tailored specifically for ISWM systems, and examples may include but are not limited to WISARD,[15] IWM,[16] WASTED,[17] and EASEWASTE.[18] Specifically, the IWM versions 1 and 2[19,20] for SWM systems provide life cycle inventory (LCI). The models enabled decision makers and waste managers to use an LCA for assessing their specific waste management configurations without in-depth knowledge of the theory and methodology and allowed them to learn how changes in the system could affect the environmental impacts through scenario analysis.[21] However, generic models, like UMBERTO, Gabi, and SimaPro, were applied to SWM, with raising potentials concerning SWM specificities.[22] Besides, extended tools focusing specially on the possible impact of economic features on decision-making process were developed. In United States, for example, the municipal solid waste – decision support tool (MSW-DST) developed by the Research Triangle Institute and United States Environmental Protection Agency,[23,24] where the environmental methodology is based on the use of LCA and the cost methodology with respect to the full-cost accounting, became available in the 2000s.

Application of LCA in SWM systems has also been promoted through the combination with other systems analysis tools to reach a sustainable decision. In this field, LCA was combined with site-dependent and qualitative approaches to evaluate waste-to-energy taxation.[25] Life cycle assessment and strategic environmental assessment were integrated to assess economic and environmental impacts for weight-based tax in waste incineration.[26] The ORWARE model, developed by Dalemo et al.,[27] Björklund, Dalemo, and Sonesson,[28] and Eriksson et al.[29,30] combines LCA with a simulation tool and material flow analysis. Solano et al.[31,32] developed a model for ISWM to obtain the best solution through LCA and an optimization model for balancing economic and environmental considerations. From the previous review,[22] it can be summarized that LCA results have been capable of changing packaging and packaging waste management, influencing the selection of waste treatment technologies, affecting regulation assessment, challenging waste hierarchy paradigm, increasing knowledge, and developing tools and methods for decision support. Those systems analysis models can be flexibly woven to deal with SWM issues with varying features, and more discussion can be seen in the work of Harrison et al.[33] and Chang, Pires, and Martinho.[22]

Regardless of these cases above where LCA was applied in a variety of SWM systems, it is important to understand how LCA could influence the decision making in many SWM systems and, especially, the decision-making process. The methodology capable of linking different criteria is MADM, such as the applications in the work of Kijak and Moy,[34] Contreras et al.,[35] and Skordilis,[36] all of which can pro mote the understanding in decision making for SWM.

Study Area and System Description

The Setúbal peninsula is located in the district of Setúbal, covers an area of 1522 km^2, and has 714,589 inhabitants.[37] The area is divided into nine municipalities, as shown in Figure 1. The nine municipalities have associated to manage their municipal solid waste (MSW), on a regional scale. AMARSUL company is owned by the local municipalities and is responsible for managing the MSW system since 1997. The SWM system is composed of nine recycling centers, two material recovery facilities (MRFs), two landfills, one transfer station, and one aerobic mechanical biological treatment (MBT).

Nowadays, this SWM system keeps on promoting the separation of paper/cardboard, glass, and light packaging (plastics, metals, and composite packaging) waste by means of curbside recycling systems. Each type of waste is collected separately in three specific containers and then directly sent to the MRF

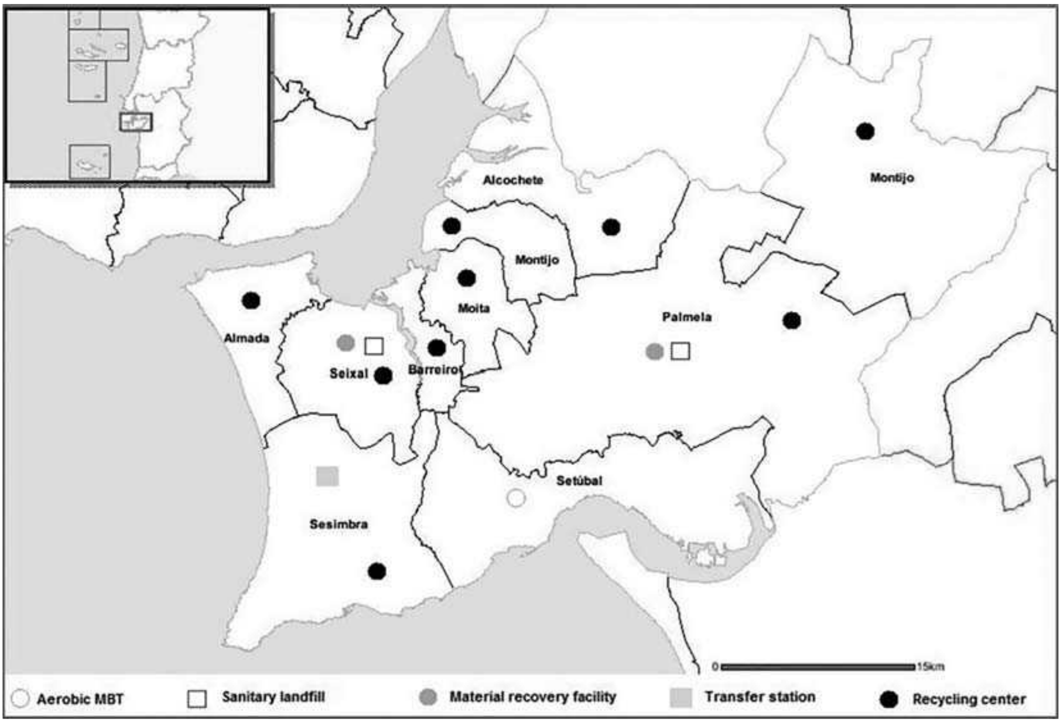


FIGURE 1 The geographical location of the Setúbal peninsula SWM system.

for recycling, recovery, and reuse. The remaining waste fractions are then collected through a door-to-door and/or bin collection scheme, which is destined for final disposal at landfills. In the case of the Sesimbra municipality, the waste stream is first sent to the transfer station, which is then followed by the final disposal at sanitary landfills. Yet the residual waste after waste separation and recycling collected from the Setúbal municipality is transported to an aerobic MBT plant, where the "stabilized residue" can be produced as fertilizer to be applied as agriculture soil-amendment materials.

Within this MSW system, there is a recent need to make some changes in order to comply with the Packaging and Packaging Waste Directive[38] and Landfill Directive.[39] The National Plan for MSW (i.e., designated as PERSU II) has decided to pursue the construction of several more MBT units. An anaerobic digestion (AD) MBT unit, with a mechanical treatment to separate recyclables and high-calorific material to produce refuse derived fuel (RDF), is predicted. It is expected that the unit will work with two separate lines, in which one is related to the biodegradable municipal solid wastes (BMW) and the other is for the residual waste streams. The RDF may be combusted in an incinerator to generate electricity. The existing aerobic MBT plant will be maintained as usual. It is expected that both MRF plants (manual sorting) will be substituted by two automated units.

The schematic of the SWM to be analyzed can be shown in Figure 2, which generally covers all stages of SWM involved, from raw waste pickup to the delivery to bins, to some intermediate processing units, and to the final disposal at landfills. Both AD MBT lines are represented as two separated processes.

These SWM processes include collection and transportation of residual waste and recyclables, waste treatment, waste transport from waste treatment facilities to final destination, energy-from-waste or waste energetic recovery (ER), and landfilling. Several final destinations for recyclables are located in Spain rather than Portugal, specifically for composite packaging and ferrous and nonferrous metal packaging materials.

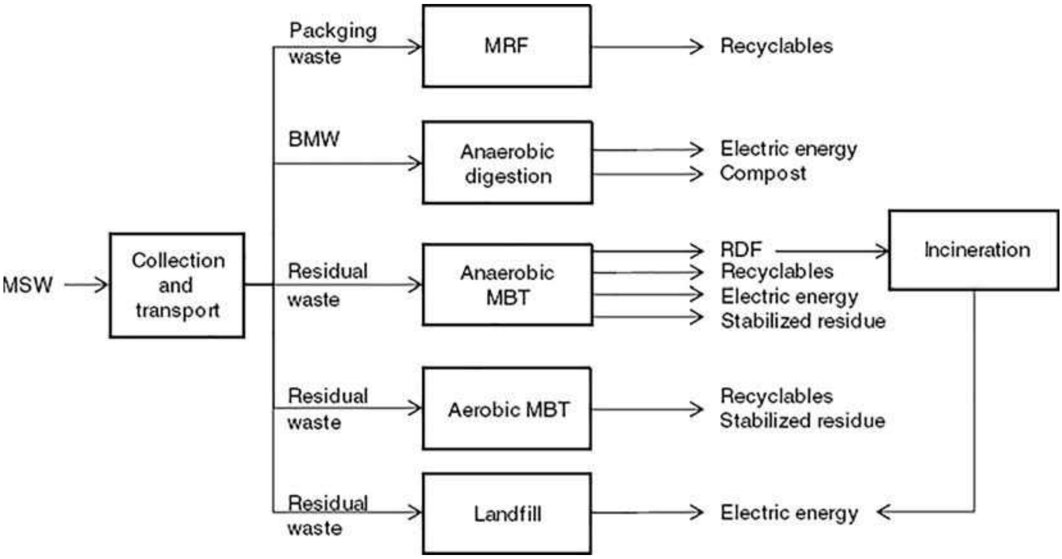


FIGURE 2 The schematic of the SWM system at Setúbal Peninsula.

Based on this system, Table 1 presents the 18 management alternatives for assessment plus the actual situation (base scenario). Those alternatives include waste collection and separate recycling of the three packaging materials through bin systems, which handle 12.4% of the current MSW in the study area. This MRF system is responsible for the compliance with the prescribed target in the Packaging Waste Directive.

Alternative 0 refers to the predicted change that will take place in the Setúbal peninsula waste management system. The remaining alternatives were designed to examine some special options for complying with the Landfill Directive. For example, alternative 1 emphasizes the inclusion of aerobic MBT; alternative 4 signifies the use of AD MBT; alternative 6 examines the specific case of using a BMW AD line. In general, alternatives 0, 3, and 5 are options for differing intermediate processing. Separation of the high-calorific fraction of waste for ER was considered through the production of RDF and the direct burning of the high-calorific fraction in municipal incinerators.

The 18 alternatives of SWM were assessed by considering two scenarios:

- *Baseline scenario*: Targets may be reached without systematic involvement and evolution, meaning that targets fulfillment can be promoted by several external agents such as government, the Green Dot Society (Sociedade Ponto Verde), and promotion campaigns that motivate a better environmental consciousness. The system may be financed by using water consumption tax for waste management to be included in the water billing system.

TABLE 1 The Distribution of Waste Streams Associated with Each Alternative in the SWM System

	Alternatives							
Fraction (%) Option	0/0[a]/0[b]	1/1[a]	2/2[a]/2[b]	3/3[a]	4/4[a]/4[b]	5/5[a]/5[b]	6/6[a]	Base
MRF	12.4	12.4	12.4	12.4	12.4	12.4	12.4	4.8
Anaerobic digestion BMW	5.4	0	0	13.3	0	7.5	28.7	0
Anaerobic digestion MBT	28.2	0	33.9	0	49.6	38.9	0	0
Aerobic MBT	13.2	49.7	15.8	32.6	0	0	0	13.8
Landfill with ER	40.8	37.9	37.9	41.7	38.0	41.2	58.9	81.4

[a] Alternatives considering RDF production plus incineration of high-calorific fraction.

[b] Alternatives not considering RDF production but considering incineration of high-calorific fraction from MBT.

- *Pay-as-you-throw (PAYT) scenario*: Targets can be reached by imposing an economic instrument— PAYT—to be implemented by various levels of MSW system managers.

The renewed interest in studying both instruments rests upon the actual billing system in Portugal, which does not comply with the philosophy of the new Waste Framework Directive due to the fact that the water billing system is dependent on the water consumption rate rather than the waste production rate. The PAYT system, being designed based on a pay-per-bag system, would fully comply with the polluter-pays principle.

Methodology

Multiattributive Decision Making

As we know, MADM is one type of multicriteria decision-making category for problem solving in decision making, by which a finite number of alternatives identified by a group of experts can be analyzed. Several MADM methods have been applied in waste management, including ELECTRE,[40] PROMETHEE[41] and GAIA,[42] AHP,[43] TOPSIS,[44] and SAW.[45] To proceed in this study, SAW and TOPSIS were selected for demonstration, both of which are well-proven techniques and make it possible to rank the solutions via a preference order for SWM with a logical procedure and a set of criteria allowing a better appliance for nonexperts, according to Cheng et al.[45] Good performance when compared with more sophisticated methods may be anticipated.[46,47]

Even though criteria of different natures might have different units, this would not be a problem in implementing SAW and TOPSIS methods, once a normalization procedure may be used to perform the criteria aggregation. In detail, SAW is a well-defined method, which is intuitive and easy to understand.[48] In SAW, the role of the decision maker is to assign weights to each attribute. The decision maker can then obtain a total score for each alternative simply by multiplying the scale rating for each attribute value by the weight assigned and summing these products over the attributes. The alternative with the highest score is the one prescribed to the decision maker. The mathematical formula of the SAW is given by Cheng et al.:[46]

$$U_j = \sum_{i=1}^{n} w_i r_{ij}, \; j = 1,\ 2,\ \ldots,\ m \tag{1}$$

where w_i is the weight of the attribute and r_{ij} is the normalized impact matrix, by which the normalization may be provided through a linear scale transformation.

TOPSIS, developed by Hwang and Yoon[44] based upon the concept that the chosen alternative should have the shortest distance from the positive ideal solution and the farthest from the negative ideal solution, was applied in this study too. A utility value $D(i)$ for each alternative i can be obtained by calculating the relative distance from i to the ideal solution, which can be described as follows:[45]

Step 1. Calculate the normalized decision matrix. The normalized value n_{ij} is calculated as

$$n_{ij} = x_{ij} / \sqrt{\sum_{j=1}^{m} x_{ij}^2}, \qquad j = 1,\ldots,m,\ i = 1,\ldots,n \tag{2}$$

Step 2. Calculate the weighted normalized decision matrix. The weighted normalized value v_{ij} is calculated as

$$v_{ij} = w_i n_{ij}, \qquad j = 1,\ldots m, I = 1,\ldots,n \tag{3}$$

where w_i is the weight of the ith attribute or criterion, and

$$\sum_{i=1}^{n} w_i = 1c$$

Step 3. Determine the positive ideal and negative ideal solution:

$$A^+ = \{v_1^+, \ldots, v_n^+\} = \left\{\left(\max_j v_{ij} \mid i \in I\right), \left(\min_j v_{ij} \mid i \in J\right)\right\}.$$
$$A^- = \{v_1^-, \ldots, v_n^-\} = \left\{\left(\min_j v_{ij} \mid i \in I\right), \left(\max_j v_{ij} \mid i \in J\right)\right\} \tag{4}$$

where i is associated with benefit criteria, and j is associated with cost criteria.

Step 4. Calculate the separation measures, using the n-dimensional Euclidean distance. The separation of each alternative from the ideal solution and for the negative ideal solution are given as, respectively,

$$d_j^+ \left\{\sum_{i=1}^{n} \left(v_{ij} - v_i^+\right)^2\right\}^{\frac{1}{2}}$$

$$d_j^- \left\{\sum_{i=1}^{n} \left(v_{ij} - v_i^-\right)^2\right\}^{\frac{1}{2}} \quad j = 1, \ldots m \tag{5}$$

Step 5. Calculate the relative closeness to the ideal solution. The relative closeness of the alternative j with respect to ideal solution is defined as

$$R_j = d_j^- / \left(d_j^+ + d_j^-\right), \qquad j = 1, \ldots, m \tag{6}$$

Since $d_j^- \geq 0$ and $d_j^+ \geq 0$, then $R_j \in [0,1]$.

Step 6. Rank the preference in descending order.

Life Cycle Assessment

According to ISO 14040,[3] LCA consists of four major stages: goal and scope definition, LCI, life cycle analysis, and interpretation of the results. Below is a detailed description of each stage to enhance comprehension.

Goal and Scope Definition

The purpose of the LCA is to assess the 18 alternatives mentioned in Table 1. The same framework was applied to both scenarios as defined by the water billing system and PAYT, respectively. The LCA provided in this study is of an attributional type. We applied the "zero burden assumption," which suggests that waste management carries none of the upstream environmental burdens into the SWM system.[49]

In an LCA system with multiple products, which is the case in our study, it is necessary to set up the methodological framework. According to ISO 14044,[50] the system boundary should be geared toward expanding the product system to include the additional functions related to the coproducts to avoid allocation. In this LCA, the material recycling, energy recovery, and fertilizer application (i.e., stabilized residue waste) of MSW were included in the LCA, which collectively resulted in an expansion of the system boundary. In this LCA, the emissions resulting from the referred operations were included as the baseline information as the emissions of those competing products and energy recovery potential resulting from those alternative operations were also included for the purpose of comparison. In this context, the system can be expanded to include additional burdens of coproduct processing and the avoided burdens of any avoided processes (i.e., substitution or avoided burden method).[4,51–53]

To ensure correct implementation of the avoided burden method through successful MSW recycling and reuse, the expanded system products should have the same function as the raw products. The substitution ratios were then applied considering closed-loop and open-loop procedures. Table 2 presents the substitution ratios for recovered materials and energy consumed. In the cases where substitution ratio is 1:1, they were considered as a closed-loop procedure with a hypothesis that no changes occurred in the inherent properties of the recycled material.[54] For example, 1 kg of recycled glass can replace 1 kg of virgin glass without considering degradation of the material during the recycling so that the quality of the secondary material may not be worse than that of the primary material.[54] The materials included in this situation in our study were glass, metals, polyethylene plastics, plastic wood, fertilizers, and

TABLE 2 Criteria Applied in Decision Making

Evaluation Criteria	Description
	Technical criteria
Landfill deviation rate	Ratio between waste not landfilled and total waste generated in a year (in percentage).
	Environmental criteria
Abiotic depletion (AbD)	Extraction of natural non-living resources. It is the difference between resources consumed during waste life cycle and resources consumption avoided from materials and energy substituted (in kg Sb eq).
Acidification	Referent to acidifying pollutants emitted during waste life cycle. The calculation is the difference between impacts from waste life cycle and the avoided impact from substituted materials and energy (in kg SO_2 eq).
Eutrophication	It is the consequence of high levels of macronutrients, such as nitrogen and phosphorus. It is the difference between eutrophication substances' potential impact during waste life cycle and avoided impacts from substituted materials and energy (in kg PO_4^{3-} eq).
Global warming potential (GWP)	Represents the impact of greenhouse gas (GHG) emissions on the radiative forcing of the atmosphere, inducing climate change. It is obtained from GHG potential impact from waste life cycle less the GHG impact from substituted materials (in kg CO_2 eq).
Human toxicity (HT)	It is the difference between impacts on human health of toxic substances emitted and the impacts from substituted materials and energy life cycle [in kg para-dichlorobenzene (p-DCB) eq].
Photochemical oxidation (PO)	Represents the formation of reactive chemical compounds, such as ozone, by action of sunlight on certain primary air pollutants. The calculation is provided from impact difference between waste life cycle and materials and energy substituted life cycles (in kg C_2H_2 eq).
Gross energy requirement (GER)	Amount of commercial energy that is required directly and indirectly by the process of making a good or service. It is the difference between energy consumed and energy produced (in kJ).
	Economic criteria
Investment costs	Represents the amount to be expended to implement the alternative (in infrastructure, equipment, vehicles, and land) (in millions €).
Operational costs	Related to the amount to be expended during alternative operation (in material, electricity, maintenance, and labor) and to financial costs like annuity (in €).
Operational revenues	The amount related to the profit obtained from selling products (energy, recyclables, compost) or
Social criteria	with the avoidance of landfilling products (RDF, recyclables) (in €).
Fee	It is the amount paid by the population to finance MSW management system (in €/t).

electricity. Specifically, 15% of the electricity consumed in Portugal was purchased from Spain, and the ratio can be taken into account too, with a proportion of 85:15, for carrying out the LCA. Polyethylene, expandable polystyrene, and plastic wood are specific cases having a 1:1 substitution ratio, since they appeared only one time in the market, such that degradation of the material was not considered.

In the cases where substitution ratio is less than 1, an open-loop allocation procedure was applied since degradation of the material should be considered, such as the cases of polyethylene terephthalate (PET), paper/cardboard, and paper from composite packaging. The calculation of substitution ratios was based on the limit number of times that a specific material can be recycled and reused repeatedly.[55] For PET, the limit number of recycling with respect to losing physical properties considered was five times.[56] Concerning paper from composite packaging, the same limit may be applied given that the proportion of paper in the packaging (0.75%) may be assumed, and the calculation procedure adopted by Rigamonti et al.[54] was applied. The substitution ratio adopted for PET was collected from the Institute for Prospective Technological Studies.[57]

Life Cycle Inventory

The LCI is the second phase of the LCA. It is an inventory of input/output data related to the SWM system that is being studied. It involves the collection of the data, which is necessary to meet the goals of the defined study.[50] In accordance with the scope of the study, an LCI was prepared for the waste management activities specified in Figure 2. The Umberto 5.5 software package was used to support the LCA.

Concerning each operational unit analyzed in the AMARSUL system, a short description of the data and assumptions considered for prescribed scenarios were provided. First of all, some of the information applied for our systems analysis was provided by the Empresa Geral do Fomento, co-owner of the SWM system at AMARSUL, which is responsible for the management of this MSW system, and the Portuguese Environment Agency. The rest of information was drawn from the Umberto software library and from the selected data sources such as the vendors of machinery features.

Life Cycle Analysis and Interpretation

Our LCA was then carried out using the Umberto 5.5 software package with the aid of the entire LCI as described in the previous section. Following the methodology suggested by the ISO 14040-44 standard,[3,50] environmental indicators were obtained for covering different impact categories. The impact categories being studied include abiotic depletion, acidification, eutrophication, global warming, human toxicity (HT), and photochemical oxidation. The characterization factors applied to each impact category are proposed by the CML 2000 method.[4]

Determination of Weighting Factors

The criteria applied to assess the alternatives in this study are partially based on some traditional criteria, which concern investment costs, operational costs, and possible revenues. Besides, the fee for waste disposal to be applied to the general public is an important criterion, which is considered as a social concern. Such a fee system is designed to increase the landfill life span, though. Due to this reason, the target of landfill space saving was considered as a technical criterion. Environmental criteria related to environmental impact categories in the context of LCA are those to be particularly tested for sensitivity analysis in this study (Table 2).

To determine the appropriate weighting factors for this application, all relevant criteria associated with different stakeholders involved in the decision-making processes were taken into account. The stakeholders involved in such decision-making processes of SWM are composed of governmental agencies, municipalities, technical groups, academics, and environmentalists. In general, governmental agencies would favor these economic criteria, whereas academia might prefer to weigh all criteria

TABLE 3

	Weights Options			
Criteria	Overall with Equal Weight	Technical	Economic	Social
Investment costs, operational costs, operational revenues	0.2	0.01	0.33	0.01
Fee	0.2	0.01	0.01	0.96
Landfill deviation rate	0.2	0.96	0.01	0.01

TABLE 4 Weight Sets for Traditional and LCA Criteria

	Weights Options				
	Overall with Equal Weight	Technical	Environmental	Economic	Social
	Criteria				
AbD, Acid., Eutr., GWP, HT, PO, GER	0.08	0.01	0.14	0.01	0.01
Investment costs, operational costs, operational revenues	0.08	0.01	0.01	0.30	0.01
Fee	0.08	0.01	0.01	0.01	0.89
Landfill deviation rate	0.08	0.89	0.01	0.01	0.01
	LCA criteria selected				
AbD/GWP/HT[a] or Acid./Eutr./PO[a] or AD/Acid./GER[a]	0.125	0.01	0.317	0.01	0.01
Investment costs, operational costs, operational revenues	0.125	0.01	0.01	0.317	0.01
Fee	0.125		0.01		0.93
Landfill deviation rate	0.125	0.93	0.01		

[a] The remaining environmental impact categories will have zero weight; their contribution to decision making will not be considered.

equally. Yet municipalities would have a closer look into a delicate balance between economic and social criteria simultaneously, whereas technical groups may focus on technical criteria. With such observations, a regional survey waste conducted to obtain stakeholders' weighting factors, being the results summarized in Table 3.

As a few more environmentalists were brought into the discussion to investigate the differential views, the LCA criteria were configured, signified, and summarized in such a decision-making process. To assess in detail the effect due to the inclusion of those LCA criteria in the SWM decision-making processes, three combinations of impact categories from LCA were tested. This combination will result in the application of three environmental impact categories (Table 4).

Alternative Screening and Ranking

Ranking Based on Traditional Criteria

When choosing the traditional criteria in SAW, the best option to be applied is conformed to the current one. That is mainly due to the revenues from selling the electric power generated at the sanitary landfill since the investment made for hardware construction has been present already, resulting in a reduced fee. However, there is a need to implement some essential measures to com-ply with Landfill and Packaging Directives in the future, which drives us to look at some more management options. Alternative P.A3* is the best when considering the same weight to all criteria. From a technical point of view, the option of treating MSW in MBT plants with AD treatment and ER of high-calorific fraction is the best option. In regard to the consideration of economic criteria only, the option of collection and treatment of separate BMW streams in an AD plant, including RDF production, is the best option. In having only the social criteria, the best option is to maximize the aerobic MBT plant. The results are collectively presented in Table 5. When applying the same practices based on the TOPSIS method, the best options associated

TABLE 5 Complete Ranking of the Alternatives from SAW Methodology

	Overall with Equal Weight		Technical		Economic		Social	
Ranking	Alt.	SAW	Alt.	SAW	Alt.	SAW	Alt.	SAW
1	Base	0.786	P.A4[b]	0.989	Base	0.878	Base	0.989
2	P.A3[a]	0.781	A4[b]	0.989	P.A6[a]	0.831	P.A1	0.595
3	A3[a]	0.780	P.A5[b]	0.968	A6[a]	0.831	P.A1[a]	0.593
4	P.A1[a]	0.776	A5[b]	0.968	P.A6	0.829	A1[a]	0.591
5	A1[a]	0.776	P.A4[a]	0.963	A6	0.828	A1	0.588
6	P.A3	0.774	A4[a]	0.963	P.A3[a]	0.801	P.A4[b]	0.564
7	A3	0.774	P.A3[a]	0.952	A3[a]	0.801	A4[b]	0.563
8	P.A1	0.773	A3a	0.952	P.A3	0.800	P.A4[a]	0.560
9	P.A4[b]	0.772	P.A2[b]	0.946	A3	0.800	A4[a]	0.555
10	A4[b]	0.772	A2[b]	0.946	P.A1	0.790	P.A3[a]	0.548
11	A1	0.771	A0[b]	0.946	P.A1[a]	0.789	P.A3	0.546
12	P.A6[a]	0.770	P.A0[b]	0.945	A1[a]	0.788	A3[a]	0.546
13	A6[a]	0.769	P.A5[a]	0.937	A1	0.788	A3	0.545
14	P.A4[a]	0.765	A5[a]	0.937	P.A5[b]	0.775	P.A4	0.542
15	P.A5[b]	0.765	P.A0[a]	0.925	A5b	0.775	A4	0.536
16	A5[b]	0.764	A0[a]	0.925	P.A5[a]	0.773	P.A5[b]	0.532
17	P.A6	0.764	P.A3	0.925	A5[a]	0.772	A5[b]	0.531
18	A4[a]	0.763	A3	0.925	P.A4[b]	0.769	P.A5[a]	0.529
19	A6	0.763	P.A1[a]	0.922	A4[b]	0.769	A5[a]	0.528
20	P.A5[a]	0.757	A1[a]	0.922	P.A4[a]	0.767	P.A6[a]	0.521
21	A5[a]	0.756	P.A2[a]	0.920	A4[a]	0.766	A6[a]	0.518
22	P.A2[b]	0.741	A2[a]	0.920	P.A4	0.756	P.A6	0.518
23	A2[b]	0.740	P.A1	0.902	A4	0.754	A6	0.514
24	A0[b]	0.736	A1	0.902	P.A5	0.754	P.A2[b]	0.510
25	P.A2[a]	0.734	P.A6[a]	0.834	A5	0.754	A2[b]	0.509
26	P.A0[b]	0.733	A6[a]	0.834	A0[b]	0.750	P.A2[a]	0.508
27	A2[a]	0.733	P.A6	0.813	P.A2[b]	0.750	A2[a]	0.504
28	P.A0[a]	0.731	A6	0.813	A2[b]	0.749	P.A5	0.502
29	A0[a]	0.731	P.A0	0.644	P.A0[a]	0.749	A5	0.501
30	P.A5	0.663	A0	0.644	A0[a]	0.749	P.A2	0.488
31	A5	0.663	P.A5	0.552	P.A2[a]	0.748	A2	0.487
32	P.A0	0.656	A5	0.552	A2[a]	0.747	P.A0[b]	0.486
33	P.A4	0.654	P.A2	0.502	P.A0[b]	0.745	A0[b]	0.485
34	A4	0.652	A2	0.502	P.A2	0.733	P.A0[a]	0.484
35	P.A2	0.638	P.A4	0.459	A2	0.733	A0[a]	0.483
36	A2	0.638	A4	0.459	P.A0	0.726	P.A0	0.457
37	A0	0.071	Base	0.308	A0	0.726	A0	0.456

Note: Alt., alternatives.

[a] Alternatives considering RDF production plus incineration of high-calorific fraction.

[b] Alternatives not considering RDF production but considering incineration of high-calorific fraction from MBT.

with the prescribed weighted criteria selected are similar to those obtained from the case of SAW except when all the criteria present the same weight. These results are collectively presented in Table 6. Comparing the scenarios with PAYT (denoted as P in the context) and with water billing indicates that the best is the PAYT system except when the base alternative is taken into account.

TABLE 6 Complete Ranking of the Alternatives from TOPSIS Methodology

	Overall with Equal Weight		Technical		Economic		Social	
Ranking	Alt.	TOPSIS	Alt.	TOPSIS	Alt.	TOPSIS	Alt.	TOPSIS
1	P.A1[a]	0.623	P.A4[b]	0.997	Base	0.600	P.A1	0.992
2	A1[a]	0.621	A4[b]	0.997	P.A6[a]	0.581	P.A1[a]	0.986
3	P.A1	0.617	P.A5[b]	0.960	A6[a]	0.579	A1[a]	0.977
4	A1	0.611	A5[b]	0.960	P.A6	0.577	A1	0.962
5	P.A3[a]	0.610	P.A4[a]	0.952	A6	0.575	P.A4[b]	0.813
6	A3[a]	0.609	A4[a]	0.952	P.A3[a]	0.507	A4[b]	0.808
7	P.A3	0.599	P.A3[a]	0.929	A3[a]	0.506	P.A4[a]	0.794
8	A3	0.598	A3[a]	0.929	P.A3	0.504	A4[a]	0.760
9	P.A4[b]	0.596	P.A2[b]	0.921	A3	0.503	P.A4	0.705
10	A4[b]	0.596	A2[b]	0.921	P.A1	0.474	P.A3[a]	0.703
11	P.A4[a]	0.586	A0[b]	0.921	P.A1[a]	0.471	P.A3	0.694
12	A4[a]	0.581	P.A0[b]	0.921	A1[a]	0.470	A3[a]	0.692
13	P.A5[b]	0.573	P.A5[a]	0.904	A1	0.469	A3	0.684
14	A5[b]	0.572	A5[a]	0.904	P.A5[b]	0.434	A4	0.665
15	P.A6[a]	0.562	P.A0[a]	0.883	A5[b]	0.433	Base	0.637
16	P.A5[a]	0.560	A0[a]	0.883	P.A5[a]	0.430	P.A5[b]	0.599
17	A6[a]	0.559	P.A3	0.878	A5[a]	0.429	A5[b]	0.591
18	A5[a]	0.559	A3	0.878	P.A4[b]	0.418	P.A5[a]	0.580
19	P.A6	0.549	P.A2[a]	0.873	A4[b]	0.418	A5[a]	0.570
20	A6	0.546	A2[a]	0.873	P.A4[a]	0.414	P.A6[a]	0.515
21	P.A2[b]	0.538	P.A1[a]	0.873	A4[a]	0.411	P.A6	0.491
22	A2[b]	0.537	A1[a]	0.873	P.A4	0.406	A6[a]	0.491
23	P.A2[a]	0.526	P.A1	0.835	A4	0.403	A6	0.465
24	A2[a]	0.523	A1	0.835	P.A5	0.402	P.A2[b]	0.439
25	A0[b]	0.522	P.A6[a]	0.704	A5	0.402	A2[b]	0.430
26	P.A0[b]	0.519	A6[a]	0.704	A0[b]	0.376	P.A2[a]	0.425
27	P.A0[a]	0.513	P.A6	0.666	P.A0[a]	0.374	P.A5	0.409
28	A0[a]	0.512	A6	0.666	A0[a]	0.373	A5	0.401
29	Base	0.481	P.A0	0.353	P.A2[b]	0.369	A2[a]	0.397
30	P.A0	0.347	A0	0.353	P.A0[b]	0.369	P.A2	0.304
31	A0	0.347	Base	0.188	A2[b]	0.369	A2	0.295
32	P.A5	0.345	P.A5	0.177	P.A2[a]	0.366	P.A0[b]	0.240
33	A5	0.345	A5	0.177	A2[a]	0.364	A0[b]	0.232
34	P.A4	0.324	P.A2	0.084	P.A2	0.351	P.A0[a]	0.226
35	A4	0.320	A2	0.084	A2	0.350	A0[a]	0.214
36	P.A2	0.288	P.A4	0.004	P.A0	0.328	P.A0	0.014
37	A2	0.288	A4	0.004	A0	0.328	A0	0.009

Note: Alt., alternatives.

[a] Alternatives considering RDF production plus incineration of high-calorific fraction.

[b] Alternatives not considering RDF production but considering incineration of high-calorific fraction from MBT.

Ranking Based on Traditional and LCA Criteria Together

By looking into the results obtained through SAW and TOPSIS with the inclusion of LCA criteria, the results vary considerably. In most cases, the option of treating MSW through AD MBT followed by ER of waste is the best option, as shown in Table 7. Only when the economic criteria were taken into

TABLE 7 Complete Ranking of the Alternatives from SAW Methodology—All Impact Categories

	Overall with Equal Weight		Technical		Environmental		Economic		Social	
Ranking	Alt.	SAW	Alt.	SAW	Alt.	SAW	Alt.	SAW	Alt.	SAW
1	P.A4[b]	0.734	P.A4[b]	0.968	P.A4[b]	0.709	Base	0.805	Base	0.908
2	A4[b]	0.733	A4[b]	0.968	A4[b]	0.709	P.A6[a]	0.790	P.A4[b]	0.575
3	P.A4[a]	0.697	P.A4[a]	0.940	P.A4[a]	0.654	A6[a]	0.789	A4[b]	0.574
4	A4[a]	0.696	A4[a]	0.940	A4[a]	0.654	P.A6	0.786	P.A4[a]	0.567
5	P.A5[b]	0.661	P.A5[b]	0.940	P.A5[b]	0.596	A6	0.786	A4[a]	0.562
6	A5[b]	0.661	A5[b]	0.940	A5[b]	0.596	P.A4[b]	0.765	P.A1[a]	0.552
7	P.A5[a]	0.634	P.A2[b]	0.909	P.A5[a]	0.556	A4[b]	0.765	P.A1	0.552
8	A5[a]	0.634	A2[b]	0.909	A5[a]	0.556	P.A5[b]	0.762	A1[a]	0.550
9	P.A2[b]	0.564	P.A5[a]	0.909	P.A2[b]	0.453	A5[b]	0.761	A1	0.545
10	P.A0[b]	0.564	A5[a]	0.909	A2[b]	0.453	P.A4[a]	0.759	P.A5[b]	0.537
11	P.A2[a]	0.546	A0[b]	0.907	P.A0[b]	0.426	A4[a]	0.758	A5[b]	0.536
12	A2[b]	0.545	P.A0[b]	0.907	P.A2[a]	0.426	P.A5[a]	0.757	P.A5[a]	0.531
13	A0[b]	0.545	P.A3[a]	0.888	A0[b]	0.426	A5[a]	0.756	A5[a]	0.530
14	P.A0[a]	0.545	A3[a]	0.888	A2[a]	0.426	P.A3[a]	0.748	P.A3[a]	0.513
15	A2[a]	0.517	P.A0[a]	0.885	P.A0[a]	0.382	A3[a]	0.748	A3[a]	0.511
16	A0[a]	0.516	A0[a]	0.885	A0[a]	0.382	P.A3	0.745	P.A3	0.510
17	P.A6[a]	0.463	P.A2[a]	0.884	P.A6[a]	0.270	A3	0.745	P.A4	0.509
18	A6[a]	0.463	A2[a]	0.884	A6[a]	0.270	P.A1[a]	0.733	A3	0.508
19	P.A6	0.446	P.A3	0.861	P.A6	0.247	A1[a]	0.733	P.A2[b]	0.506
20	A6	0.446	A3	0.860	A6	0.247	P.A1	0.732	A2[b]	0.505
21	P.A3[a]	0.350	P.A1[a]	0.857	P.A4	0.100	A1	0.731	A4	0.504
22	P.A3	0.349	A1[a]	0.857	A4	0.100	P.A2[b]	0.728	P.A6[a]	0.503
23	A3[a]	0.331	P.A1	0.837	P.A5	0.083	A2[b]	0.728	P.A2[a]	0.502
24	P.A1[a]	0.331	A1	0.837	P.A3[a]	0.083	A0[b]	0.726	A6[a]	0.500
25	A3	0.324	P.A6[a]	0.792	A3[a]	0.079	P.A2[a]	0.724	A2[a]	0.498
26	P.A1	0.324	A6[a]	0.792	A5	0.079	A2[a]	0.723	P.A6	0.498
27	A1[a]	0.314	P.A6	0.772	P.A3	0.054	P.A0[b]	0.722	A6	0.495
28	P.A4	0.313	A6	0.772	P.A0	0.054	P.A0[a]	0.722	P.A0[b]	0.481
29	P.A5	0.307	P.A0	0.601	A3	0.053	A0[a]	0.722	A0[b]	0.481
30	P.A0	0.307	A0	0.601	A0	0.053	P.A4	0.707	P.A0[a]	0.476
31	A1	0.306	P.A5	0.517	P.A1[a]	0.040	A4	0.706	A0[a]	0.475
32	A4	0.305	A5	0.517	A1[a]	0.040	P.A5	0.705	P.A5	0.471
33	A5	0.286	P.A2	0.467	P.A2	0.030	A5	0.705	A5	0.470
34	P.A2	0.286	A2	0.467	A2	0.030	P.A2	0.681	P.A2	0.455
35	A0	0.265	P.A4	0.432	P.A1	0.012	A2	0.681	A2	0.454
36	A2	0.265	A4	0.432	A1	0.012	P.A0	0.676	P.A0	0.427
37	Base	0.231	Base	0.276	Base	−0.118	A0	0.676	A0	0.426

Note: Alt., alternatives.

[a] Alternatives considering RDF production plus incineration of high-calorific fraction.

[b] Alternatives not considering RDF production but considering incineration of high-calorific fraction from MBT.

account, applying the AD to treat a source separated from BMW followed by RDF production remains the best option, if the base case is not considered. When the TOPSIS was applied, the results, as presented in Table 8, have changed, being preferable options with AD and incineration of high-calorific fraction. However, concerning social criteria, the option of maximizing aerobic MBT is considered as the best solution. Also, at almost criteria cases, the option of applying PAYT is better than the same technological option with the water billing system.

TABLE 8 Complete Ranking of the Alternatives from TOPSIS Methodology—All Impact Categories

	Overall with Equal Weight		Technical		Environmental		Economic		Social	
Ranking	Alt.	TOPSPS	Alt.	TOPSPS	Alt.	TOPSPS	Alt.	TOPSPS	Alt.	TOPSPS
1	P.A4[b]	0.795	P.A4[b]	0.975	P.A4[b]	0.807	Base	0.575	P.A1[a]	0.881
2	A4[b]	0.795	A4[b]	0.975	A4[b]	0.807	P.A6[a]	0.574	P.A1	0.879
3	P.A4[a]	0.789	P.A5[b]	0.953	P.A4[a]	0.801	A6[a]	0.572	A1[a]	0.879
4	A4[a]	0.789	A5[b]	0.953	A4[a]	0.801	P.A6	0.570	A1	0.871
5	P.A5[b]	0.777	P.A4[a]	0.946	P.A5[b]	0.791	A6	0.568	P.A4[b]	0.809
6	A5[b]	0.777	A4[a]	0.946	A5[b]	0.791	P.A3[a]	0.511	A4[b]	0.805
7	P.A5[a]	0.765	P.A2[b]	0.915	P.A5[a]	0.779	A3[a]	0.510	P.A4[a]	0.791
8	A5[a]	0.765	A2[b]	0.915	A5[a]	0.779	P.A3	0.507	A4[a]	0.759
9	P.A2[b]	0.732	A0[b]	0.915	P.A2[b]	0.743	A3	0.506	P.A4	0.689
10	A2[b]	0.732	P.A0[b]	0.915	A2[b]	0.743	P.A1	0.485	P.A3[a]	0.683
11	P.A2[a]	0.724	P.A3[a]	0.909	P.A2[a]	0.735	P.A1[a]	0.484	P.A3	0.673
12	A2[a]	0.724	A3[a]	0.909	A2[a]	0.735	A1[a]	0.483	A3[a]	0.672
13	A0[b]	0.717	P.A5[a]	0.900	A0[b]	0.729	A1	0.480	A3	0.664
14	P.A0[b]	0.716	A5[a]	0.900	P.A0[b]	0.729	P.A5[b]	0.460	A4	0.652
15	P.A0[a]	0.703	P.A0[a]	0.879	P.A0[a]	0.715	A5[b]	0.459	Base	0.634
16	A0[a]	0.703	A0[a]	0.879	A0[a]	0.715	P.A5[a]	0.455	P.A5[b]	0.604
17	P.A1[a]	0.578	P.A2[a]	0.870	P.A1[a]	0.577	A5[a]	0.455	A5[b]	0.596
18	A1[a]	0.578	A2[a]	0.870	A1[a]	0.577	P.A4[b]	0.454	P.A5[a]	0.585
19	P.A1	0.567	P.A3	0.865	P.A4	0.570	A4[b]	0.454	A5[a]	0.575
20	A1	0.566	A3	0.865	A4	0.570	P.A4[a]	0.449	P.A6[a]	0.506
21	P.A4	0.560	P.A1[a]	0.863	P.A1	0.565	A4[a]	0.446	A6[a]	0.484
22	A4	0.560	A1[a]	0.863	A1	0.565	P.A4	0.417	P.A6	0.483
23	P.A3[a]	0.547	P.A1	0.827	P.A2	0.545	A4	0.415	A6	0.459
24	A3[a]	0.547	A1	0.827	A2	0.545	P.A5	0.411	P.A2[b]	0.450
25	P.A2	0.536	P.A6[a]	0.699	P.A3[a]	0.544	A5	0.410	A2[b]	0.442
26	A2	0.536	A6[a]	0.699	A3[a]	0.544	A0[b]	0.405	P.A2[a]	0.436
27	P.A0	0.535	P.A6	0.662	P.A0	0.543	P.A2[b]	0.405	A2a	0.410
28	A0	0.535	A6	0.662	A0	0.542	A2[b]	0.404	P.A5	0.406
29	P.A3	0.534	P.A0	0.354	P.A3	0.532	P.A0[a]	0.402	A5	0.399
30	A3	0.534	A0	0.354	A3	0.532	A0[a]	0.401	P.A2	0.309
31	P.A6[a]	0.528	Base	0.192	P.A5	0.529	P.A2[a]	0.401	A2	0.299
32	A6[a]	0.528	P.A5	0.179	A5	0.529	A2[a]	0.399	P.A0[b]	0.265
33	P.A5	0.522	A5	0.179	P.A6[a]	0.527	P.A0[b]	0.398	A0[b]	0.257
34	A5	0.522	P.A2	0.091	A6[a]	0.527	P.A2	0.366	P.A0[a]	0.251
35	P.A6	0.519	A2	0.091	P.A6	0.518	A2	0.366	A0[a]	0.241
36	A6	0.519	P.A4	0.037	A6	0.518	P.A0	0.346	P.A0	0.072
37	Base	0.290	A4	0.037	Base	0.278	A0	0.346	A0	0.071

Note: Alt., alternatives.

[a] Alternatives considering RDF production plus incineration of high-calorific fraction.

[b] Alternatives not considering RDF production but considering incineration of high-calorific fraction from MBT.

TABLE 9 Best-Five Ranking for the Three Different LCA Impact Categories from SAW

	Overall with Equal Weight		Technical		Environmental		Economic		Social	
	Alt.	SAW	Alt.	SAW	Alt.1	SAW	Alt.	SAW	Alt.	SAW
					Ranking of AbD/GWP/HT					
1	P.A6[a]	0.701	P.A4[b]	0.968	P.A6[a]	0.595	Base	0.833	Base	0.941
2	A6[a]	0.700	A4[b]	0.968	A6[a]	0.595	P.A6[a]	0.824	P.A1	0.560
3	P.A6	0.689	P.A5[b]	0.948	P.A6	0.575	A6[a]	0.823	P.A1[a]	0.559
4	A6	0.688	A5[b]	0.948	A6	0.575	P.A6	0.821	A1[a]	0.557
5	P.A4[b]	0.600	P.A4[a]	0.941	P.A5[b]	0.348	A6	0.820	P.A4[b]	0.557
					Ranking of Acid./Eutr./PO					
1	P.A4[b]	0.858	P.A4[b]	0.989	P.A4[b]	0.989	Base	0.855	Base	0.962
2	A4[b]	0.857	A4[b]	0.989	A4[b]	0.989	P.A6[a]	0.801	P.A1	0.588
3	P.A4[a]	0.832	P.A4[a]	0.962	P.A4[a]	0.934	A6[a]	0.800	P.A1[a]	0.587
4	A4[a]	0.831	A4[a]	0.962	A4[a]	0.934	P.A6	0.798	A1[a]	0.585
5	P.A2[b]	0.759	P.A5[b]	0.960	P.A2[b]	0.786	A6	0.797	A1	0.582
					Ranking of AbD/ Acid./GER					
1	P.A4[b]	0 858	P.A4[b]	0.989	P.A4[b]	0.989	Base	0.858	Base	0.965
2	A4[b]	0.857	A4[b]	0.989	A4[b]	0.989	P.A6[a]	0.816	P.A1	0.586
3	P.A4[a]	0.841	P.A5[b]	0.965	P.A4[a]	0.958	A6[a]	0.815	P.A1[a]	0.586
4	A4[a]	0.840	A5[b]	0.965	A4[a]	0.958	P.A6	0.813	A1[a]	0.584
5	P.A5[b]	0.809	P.A4[a]	0.963	P.A5[b]	0.877	A6	0.812	P.A4[a]	0.580

Note: Alt., alternatives.

[a] Alternatives considering RDF production plus incineration of high-calorific fraction.

[b] Alternatives not considering RDF production but considering incineration of high-calorific fraction from MBT.

Another concern when applying LCA criteria is that the total number and type of impact category used in the SAW and TOPSIS could impact the final options. If we reduce the consideration from seven to three criteria, the results would be changed in both SAW and TOPSIS. When looking into the case of SAW, such a change is salient as only AbD/global warming potential (GWP)/HT impact categories were applied, as shown in Table 9. In this case, the AD of BMW separated from source with the inclusion of RDF ER in an incineration plant would be the best option for more criteria. For the remaining environmental impacts the results are mostly in accordance with the total LCA criteria applied (Table 7). Such a finding suggests that there are some LCA criteria that would have a bigger influence on the ranking outcome.

The phenomenon mentioned before has also been verified via TOPSIS (Table 10). When looking into the case of TOPSIS, for LCA criteria AbD/GWP/HT, the results were changed even more than those in the counterpart. When only the technical criterion was taken into account, the option favored the use of an anaerobic MBT plant with ER from the high-calorific fraction, without producing RDF. Such an option is always advantageous no matter how many LCA criteria were included for screening and ranking. Options like the adoption of a full aerobic MBT plant to divert organic waste from landfill as well as selective collection of BMW plus RDF production with burning in an incineration plant were selected as the best options when considering the other combinations of environmental impact categories in LCA (Table 10). Yet the inclusion of a varying number of criteria in LCA might impact the preference for PAYT. It obviously leads to the conclusion that the PAYT can be a better instrument to meet the environmental targets from a sustainable point of view.

TABLE 10 Best-Five Ranking for the Three Different LCA Impact Categories from TOPSIS

	Overall with Equal Weight		Technical		Environmental		Economic		Social	
	Alt.	TOPSIS	Alt.	TOPSIS	Alt.	TOPSIS	Alt.	TOPSIS	Alt.	TOPSIS
					Ranking of AbD/GWP/HT					
1	P.A6[a]	0.735	P.A4[b]	0.976	P.A6[a]	0.753	P.A6[a]	0.587	P.A1	0.903
2	A6[a]	0.734	A4[b]	0.976	A6[a]	0.753	Base	0.586	P.A1[a]	0.903
3	P.A6	0.723	P.A5[b]	0.956	P.A5[b]	0.742	A6[a]	0.585	A1[a]	0.901
4	A6	0.723	A5[b]	0.956	A5[b]	0.742	P.A6	0.583	A1	0.894
5	P.A5[b]	0.722	P.A4[a]	0.946	P.A6	0.741	A6	0.581	P.A4[b]	0.807
					Ranking of Acid./Eutr./PO					
1	P.A4[b]	0.887	P.A4[b]	0.997	P.A4[b]	0.996	Base	0.591	P.A1[a]	0.938
2	A4[b]	0.887	A4[b]	0.997	A4[b]	0.996	P.A6[a]	0.569	P.A1	0.937
3	P.A4[a]	0.882	P.A5[b]	0.958	P.A4[a]	0.977	A6[a]	0.567	A1[a]	0.935
4	A4[a]	0.881	A5[b]	0.958	A4[a]	0.977	P.A6	0.565	A1	0.926
5	P.A2[b]	0.833	P.A4[a]	0.952	P.A2[b]	0.893	A6	0.563	P.A4[b]	0.816
					Ranking of AbD/Acid./GER					
1	P.A4[b]	0.799	P.A4[b]	0.983	P.A4[b]	0.992	Base	0.595	P.A1[a]	0.930
2	A4[b]	0.799	A4[b]	0.983	A4[b]	0.992	P.A6[a]	0.576	P.A1	0.928
3	P.A4[a]	0.789	P.A5[b]	0.957	P.A4[a]	0.961	A6[a]	0.574	A1[a]	0.928
4	A4[a]	0.787	A5[b]	0.957	A4[a]	0.961	P.A6	0.572	A1	0.918
5	P.A5[b]	0.740	P.A4[a]	0.950	P.A5[b]	0.847	A6	0.570	P.A4[b]	0.814

Note: Alt., alternatives.

[a] Alternatives considering RDF production plus incineration of high-calorific fraction.

[b] Alternatives not considering RDF production but considering incineration of high-calorific fraction from MBT.

Final Remarks on the Importance of LCA

The inclusion of environmental criteria into the MADM process can result in a change of the best option when compared with the cases where only all or part of the traditional criteria are taken into account. This proves that a sustainable decision must include all elements of sustainability, or else the decision would end up a biased option. Even though LCA impact category can be considered, the results would vary when differing numbers of LCA impact category were assessed in the MADM process. Impact categories are mutually exclusive, however. When all impact categories were included, the normalization process in the MADM process might dilute the relative importance or effect of some impact categories and, therefore, distort the outcome to some extent. In any circumstance, it is not possible to implement weighted criteria to restore such an outcome. To reach a sociopolitical sustainability, the selection of an impact category should be the stakeholders' responsibility. Regardless of some discrepancies present in the selection of impact categories and normalization, the final results obtained from both SAW and TOPSIS allow improvements for SWM in Portugal complying with the new guidelines and mandatory items with respect to those sustainable principles reinforced by European policies. Overall, the option of applying PAYT has been shown to be a good option relative to the water billing system.

Conclusions

In this entry, we analyze how LCA could have an impact on the decision-making processes of SWM. A case study in Portugal clearly indicates that the inclusion of LCA for screening and ranking has shown the potential of helping SWM decision making. The inclusion of LCA impact categories, together with traditional criteria associated with economic considerations, is capable of producing relatively sustainable decisions, which otherwise would not be reached. Within the analysis, both SAW and TOPSIS were proven capable of showing which alternatives should be favored. Comparing traditional criteria, these two methods have quite-consistent outputs in terms of several weighted criteria applied. When considering only the traditional criteria, the base case is the best option, without regard to environmental regulations. With the inclusion of LCA criteria, the best solution is to adopt the AD MBT plant followed by the energy recovery at incineration plant. Overall, the option of applying PAYT has been shown to be a good option relative to the water billing system. The selection of the number and type of LCA criteria (environmental impact categories) to be included in the MADM process is flexible, depending on the stakeholders' preference during the decision analysis. Such selection must be careful, however, since it ends up making an obvious difference in the final decision analysis results.

References

1. Council of European Parliament. *Proposal for a Directive of the European Parliament and of the Council on Waste— COM(2005) 667;* European Union: Brussels, Belgium, 2005.
2. Council of European Parliament. *Directive 2008/98/EC of the European Parliament and of the Council of 19 November 2008 on Waste and Repealing Certain Directives,* Official Journal L312/3; European Union: Strasbourg, France, 2008; 3–30.
3. ISO. *ISO 14040—Environmental Management—Life Cycle Assessment—Principles and Framework;* International Standard Organization: Switzerland, 2006.
4. Guinée, J.B.; Gorree, M.; Heijungs, R.; Huppes, G.; Kleijn, R.; van Oers, L.; Wegener Sleeswijk, A.; Suh, S.; Udo de Haes, H.A.; de Bruijn, J.A.; van Duin, R.; Huijbregts, M.A.J., Eds. *Handbook on Life Cycle Assessment—Operational Guide to the ISO Standards;* Kluwer Academic Publisher: Dordrecht, Netherlands, 2002.
5. Available at http://www.epa.gov/nrmrl/lcaccess/pdfs/600r06060.pdf (accessed December 2010).
6. Available at http://www2.mst.dk/Udgiv/publications/1998/87–7909–026-5/pdf/87–7909–026-5.pdf (accessed January 2010).
7. Available at http://www2.mst.dk/Udgiv/publications/1998/87–7909–025–7/pdf/87–7909–025–7.pdf (accessed January 2010).
8. Available at http://www2.mst.dk/Udgiv/publications/1998/87–7909–024–9/pdf/87–7909–024–9.pdf (accessed January 2010).
9. Available at http://www2.mst.dk/Udgiv/publications/1998/87–7909–023-0/pdf/87–7909–023-0.pdf (accessed January 2010).
10. Available at http://www2.mst.dk/Udgiv/publications/1998/87–7909–014-1/pdf/87–7909–014-1.pdf (accessed January 2010).
11. Finnveden, G.; Person, L.; Steen, B. *Förpackningar i kret-sloppet: Atervinning av mjölkkartong—En LCA-studie av skillnader i miljöbelastning (Packaging in Circulation: Recycling of Cardboard for Milk—An LCA Study of the Differences in Environmental Load),* Report 4301; Swedish Environmental Protection Agency: Stockholm, Sweden, 1994.
12. Finnveden, G.; Ekvall, T. *Energieller materialåtervinning av pappersförpakningar? (Energy Recovery or Material Recycling of Paper Packaging?);* Svensk Kartongåtervinning AB: Stockholm, Sweden, 1998.
13. Sakai, S.; Sawell, S.E.; Chandler, A.J.; Eighmy, T.T.; Kos-son, D.S.; Vehlow, J.; van der Sloot, H.A.; Hartlén, J.; Hjel-mar, O. World trends in municipal solid waste management. Waste Manage. **1996**, *16* (5/6), 341–350.
14. Klöpffer, W. Conference reports: LCA in Brighton. Int. J. Life Cycle Assess. **2000**, *5* (4), 249.
15. Available at http://www.ecobilan.com/wisard/index.php (accessed December 2008).
16. Available at http://www.iwm-model.uwaterloo.ca/iswm_booklet.pdf (accessed December 2009).
17. Diaz, R.; Warith, M. Life-cycle assessment of municipal solid wastes: Development of the WASTED model. Waste Manage. **2006**, *26* (8), 886–901.

18. Kirkeby, J.; Christensen, T.; Bhander, G.; Hansen, T.; Birgisdottir, H. LCA modelling of MSW management system: Approach and case study. In *Proceedings Sardinia 2005,* Tenth International Waste Management and Landfill Symposium, S. Margherita di Pula, Cagliari, Italy, October 3–7, 2005; Cossu, R., Stegmann, R. Eds.; CISA, Environmental Sanitary Engineering Centre, 2005.
19. McDougall, F.; White, P.; Franke, M.; Hindle, P. *Integrated Solid Waste Management: A Life Cycle Inventory;* Blackwell Science Ltd.: Oxford, 2001.
20. White, P.; Franke, M.; Hindle, P. *Integrated Solid Waste Management: A Life-Cycle Inventory;* Blackie Academic & Professional: Glasgow, U.K., 1995.
21. Winkler, J.; Bilitewski, B. Comparative evaluation of life cycle assessment models for solid waste management. Waste Manage. **2007**, *27* (8), 1021–1031.
22. Chang, N.B.; Pires, A.; Martinho, G. Empowering systems analysis for solid waste management: The tools and perspectives in European countries. Crit. Rev. Environ. Sci. Technol. **2011**, *41* (16), 1449–1530.
23. Weitz, K.; Barlaz, M.; Ranji, R.; Brill, D.; Thorneloe, S.; Ham, R. Life cycle management of municipal solid waste. Int. J. Life Cycle Assess. **1999**, *4* (4), 195–201.
24. Thorneloe, S.; Weitz, K.; Jambeck, J. Application of the US decision support tool for materials and waste management. Waste Manage. **2007**, *27* (8), 1006–1020.
25. Nilsson, M.; Bjorklund, A.; Finnveden, G.; Johansson, J. Testing a SEA methodology for the energy sector: A waste incineration tax proposal. Environ. Impact Assess. Rev. **2005**, *25* (1), 1–32.
26. Björklund, A.; Finnveden, G. Life cycle assessment of a national policy proposal—The case of a Swedish waste incineration tax. Waste Manage. **2007**, *27* (8), 1046–1058.
27. Dalemo, M.; Sonesson, U.; Bjorklund, A.; Mingarini, K.; Frostell, B.; Nybrant, T.; Jonsson, H.; Sundqvist, J.-O.; Thyselius, L. ORWARE—A simulation model for organic waste handling systems. Part 1: Model description. Resour., Conserv. Recycl. **1997**, *21* (1), 17–37.
28. Björklund, A.; Dalemo, M.; Sonesson, U. Evaluating a municipal waste management plan using ORWARE. J. Cleaner Prod. **1999**, *7* (4), 271–280.
29. Eriksson, O.; Frostell, B.; Bjorklund, A.; Assefa, G.; Sundwvist, J.O.; Granath, J.; Carlsoon, M.; Baky, A.; Thyselius, L. ORWARE—A simulation tool for waste. Resour., Conserv. Recycl. **2002**, *36* (4), 287–307.
30. Eriksson, O.; Reich, M.; Frostell, B.; Bjorklund, A.; Assefa, G.; Sundqvist, J.O.; Granath, J.; Baky, A.; Thyselius, L. Municipal solid waste management from a systems perspective. J. Cleaner Prod. **2005**, *13* (3), 241–252.
31. Solano, E.; Dumas, R.; Harrison, K.; Ranjithan, S.; Barlaz, M.; Brill, E. Life-cycle–based solid waste management. II: Illustrative applications. J. Environ. Eng. **2002**, *128* (10), 993–1005.
32. Solano, E.; Ranjithan, S.; Barlaz, M.; Brill, E. Life-cycle–based solid waste management. I: Model development. J. Environ. Eng. **2002**, *128* (10), 981–992.
33. Harrison, K.; Dumas, R.; Solano, E.; Barlaz, M.; Brill, E.; Ranjithan, S. Decision support tool for life-cycle–based solid waste management. J. Comput. Civ. Eng. **2001**, *15* (1), 44–58.
34. Kijak, R.; Moy, D. A decision support framework for sustainable waste management. J. Ind. Ecol. **2004**, *8* (3), 33–50.
35. Contreras, F.; Hanaki, K.; Aramaki, T.; Connors, S. Application of analytical hierarchy process to analyze stakeholders preferences for municipal solid waste management plans, Boston, USA. Resour., Conserv. Recycl. **2008**, *52* (7), 979–991.
36. Skordilis, A. Modelling of integrated solid waste management systems in an island. Resour., Conserv. Recycl. **2004**, *41* (3), 243–254.
37. Available at http://www.amarsul.pt/relatorios_e_contas.php (accessed July 2009).
38. Council of European Parliament. *Directive 2004/12/EC of the European Parliament and of the Council of 11 February 2004 Amending Directive 94/62/EC on Packaging and Packaging Waste,* Official Journal L047; European Union: Strasbourg, France, 2004; 26–32.

39. Council. *Council Directive 99/31/EC of April 1999 on the Landfill of Waste,* Official Journal L182; European Union: Luxembourg, 1999; 1–19.
40. Roy, B. The outranking approach and the foundations of ELECTRE methods. Decis. Theory **1991**, *31* (1), 49–73.
41. Brans, J.P.; Mareschal, B.; Vincke, P.H. PROMETHEE: A new family of outranking methods in multicriteria analysis. Oper. Res. **1984**, *84,* 447–490.
42. Mareschal, B. Geometrical representations for MCDA: The GAIA procedure. Eur. J. Oper. Res. **1988**, *34,* 69–77.
43. Saaty, T.L. *The Analytic Hierarchy Process;* McGraw-Hill: New York, 1980.
44. Yoon, K.; Hwang, C.L. Manufacturing plant location analysis by multiple attribute decision making: Part I—Singleplant strategy. Int. J. Prod. Res. **1985**, *23* (2), 345–359.
45. Zanakis, S.H.; Solomon, A.; Wishart, N.; Dublish, S. Multiattribute decision making: A simulation comparison of select methods. Eur. J. Oper. Res. **1998**, *107* (3), 507–529.
46. Cheng, S.; Chan, C.W.; Huang, G.H. Using multiple criteria decision analysis for supporting decision of solid waste management. J. Environ. Sci. Health A **2002**, *37* (6), 975–990.
47. Chang, Y.H.; Yeh, C.H. Evaluating airline competitiveness using multiattribute decision making. Omega **2001**, *29* (5), 405–415.
48. Hwang, C.L.; Yoon, K. *Multiple Attribute Decision Making—Lecture Notes in Economics and Mathematical Systems;* Springer: Berlin, 1981.
49. Ekvall, T.; Assefa, G.; Björklund, A.; Eriksson, O.; Finnveden, G. What life-cycle assessment does and does not do in assessments of waste management. Waste Manage. **2007**, *27* (8), 989–996.
50. ISO. *ISO 14044—Environmental Management—Life Cycle Assessment—Requirements and Guidelines;* International Standard Organization: Switzerland, 2006.
51. Tillman, A.M.; Ekvall, T.; Baumann, H.; Rydberg, T. Choice of system boundaries in life cycle assessment. J. Cleaner Prod. **1994**, *2* (1), 21–29.
52. Thomassen, M.A; Dalgaard, R.; Heijungs, R.; de Boer, I. Attributional and consequential LCA of milk production. Int. J. Life Cycle Assess. **2008**, *13* (4), 339–349.
53. Finnveden, G.; Hauschild, M.Z.; Ekvall, T.; Guinée, J.; Heijungs, R.; Hellweg, S.; Koehler, A.; Pennington, D.; Suh, S. Recent developments in life cycle assessment. J. Environ. Manage. **2009**, *91* (1), 1–21.
54. Rigamonti, L.; Grosso, M.; Sunseri, M.C. Influence of assumptions about selection and recycling efficiencies on the LCA of integrated waste management systems. Int. J. Life Cycle Assess. **2009**, *14* (5), 411–419.
55. Rigamonti, L.; Grosso, M.; Giugliano, M. Life cycle assessment for optimising the level of separated collection in integrated MSW management systems. Waste Manage. **2009**, *29* (2), 934–944.
56. Comieco. Carta: Introduzione (Cardbord: Introduction). Matrec. 2008. Available at: http://www.matrec.it/ (accessed January 2010).
57. Delgado, C.; Barruetabeña, L.; Salas O. *Assessment of the Environmental Advantages and Drawbacks of Existing and Emerging Polymers Recovery Processes,* JRC Scientific and Technical Reports; European Communities: Luxembourg, 2007; 1–278.

27

Sustainable Development: Ecological Footprint in Accounting

Simone Bastianoni, Valentina Niccolucci, Elena Neri, Gemma Cranston, Alessandro Galli, and Mathis Wackernagel

Introduction

Since the onset, sustainable development has called for a set of suitable tools and sustainability indicators that are able to monitor the relevant aspects of systems under scrutiny. In general, sustainability indicators are required to highlight the major "hot spots" of a system and provide the most relevant elements upon which to carry out a "sustainability diagnosis." This is necessary to determine the most appropriate "sustainability therapy," which may include future action plans, programs, and strategy, to support the development of environmental management policies and the decision-making process.

In recent decades, the global increase in resource consumption, waste production, and environmental pollution has become a core aspect within the sustainable development framework. The mismatch between the earth's resource regeneration rates and humanity's consumption rates is fundamental to understanding sustainability; the importance of this discrepancy was highlighted by H. Daly through the formulation of his famous "sustainability principles." Such principles can be expressed as follows: 1) renewable resources must be used no faster than the rate at which they regenerate; 2) nonrenewable resources must be used no faster than the time it takes for renewable substitutes to be implemented; and 3) pollution and wastes must be emitted no faster than the speed with which natural systems can absorb them, recycle them, or render them harmless. Assessments are therefore needed to account for both the ecosystem's capacity to provide ecological assets to support human activities and the subsequent effects human use has on the environment. A solution could be to focus upon careful management of natural capital with the long-term aim to preserve the integrity of ecosystems and improve human well-being.

The "life-supporting" natural capital upon which humans depend is generated by the negentropic capacity of the planet to convert, via photosynthesis, low-quality forms of energy (e.g., solar energy) into high-quality forms of energy that can be used by living organisms among which are humans. As such, the Ecological Footprint Analysis (EFA) can be considered a resource accounting framework, with a biophysical and thermodynamic basis, providing reliable measures of both (biosphere) supply of and (human) demand for the photosynthetic capacity of the planet.

Ecological Footprint analysis is not able to answer all aspects of sustainability but instead provides an account of the flows and overuse of resources by humanity. The imbalance between human demand and nature's capacity is one part of the sustainability issue, and the Ecological Footprint is able to answer the key question of how much of the earth's regenerative capacity does humanity demand.

The Ecological Footprint is a highly versatile tool with a wide range of applications and functions ranging from territorial levels (cities, regions, nations, etc.), to a wide set of systems. Ecological Footprint analyses help governments, businesses, and individuals to manage resources available and plan on how to use them in a sustainable way.

This entry consists of seven main sections. After a general introduction, the first section details the conceptual background underpinning the Ecological Footprint methodology. There are two fundamental aspects to the methodology—Ecological Footprint and biocapacity—both of which are introduced and described. In the section "The Mathematics of the Ecological Footprint," the mathematics behind Ecological Footprint and biocapacity calculations is explored. This section also discusses the topic of ecological balances (the comparison between Ecological Footprint and biocapacity) and focuses on the concept of biocapacity remainder and deficit. The section "Toward a Multiindicator Approach" offers a joint discussion of the Ecological Footprint with other indicators like human development index (HDI), gross domestic product (GDP), and index of sustainable economic welfare (ISEW). The section "New Insight in Footprint Theory: Toward a ThreeDimensional Ecological Footprint Geography" presents a new theoretical insight: a three dimensional representation of the Ecological Footprint based on the distinction between depletion of natural capital stocks and use of natural capital flows. An overview of existing applications of the Ecological Footprint to environmental management practices is then provided in the section "Applications of Ecological Footprint," from territorial to products and services. The section "Weakness and Limitations of EFA" acknowledges the limitations of the methodology and provides final conclusions.

Fundamentals of EFA

The EFA was introduced as an environmental accounting and management tool able to provide either static snapshots or temporal trends of human demand on ecological assets (i.e., natural resources and ecosystem services).[1,2] The EFA introduced two relevant spatial parameters:[3–5] Ecological Footprint and biocapacity. The Ecological Footprint tracks the amount of biologically productive area directly and indirectly demanded on a continuous basis by humanity to provide the energy and material resource flows used and to assimilate the wastes generated (i.e., CO_2 emissions), given prevailing technologies and resource management practices.

Biocapacity represents the maximum regenerative capacity of the biosphere, which is annually available for human use. Therefore, biocapacity is an intrinsic property for any given area (where the population lives or the product is produced). Ecological Footprint and biocapacity have some common properties—both can be divided into six major land-use categories according to the World Conservation Union classification:[6] cropland, pasture land, forest, fishing ground, built-up land, and carbon uptake land. The first four land types are required to make animal-, plant-, and fish-based food and fiber and wood products available for human consumption. The built-up land is required for shelter and other physical infrastructure such as roads and cities. The last land type represents a sink, the bioproductive space required to absorb the wastes emitted from human economies. Given global data limitations, national footprint assessments currently consider only carbon dioxide emissions from fossil fuel burning as a waste. Other waste streams would also be included, if the data became available. CO_2 emissions from fossil fuel burning are converted into an equivalent forested area needed to remove the excess CO_2 from the atmosphere, given forest sequestration rates. On the demand side, this area is called the carbon footprint; on the biocapacity side, land for CO_2 sequestration is forest (forest can serve either timber and fuel wood production, or carbon sequestration—but these services are typically in competition for space). Due to data limitation about forests with long-term carbon sequestration commitments, current Ecological Footprint accounts do not distinguish between forests for forest products, long-term carbon uptake (or "carbon uptake land"), or biodiversity reserves.

It should be noted that the term "carbon footprint" used in the context of the Ecological Footprint methodology is not to be confused with the "Carbon Footprint" methodology itself. The latter term refers to a different methodology and aims to measure the total amount of all greenhouse gas emissions that are directly and indirectly caused by an activity or are accumulated over the life stages of a product or an activity. Further information on the differences between the carbon footprint component of the Ecological Footprint methodology and the Carbon Footprint methodology can be found in the work of Galli et al.[7]

Ecological Footprint and biocapacity are both expressed in a standardized unit, global hectares (gha);[3,8] the global hectare represents a hectare with world-average biological productivity.[9] Global hectares are normalized so that the number of actual hectares of biologically productive land and sea on the planet is equal to the total worldwide budget of global hectares in any given year.[10] Because of international trade and the dispersion of carbon dioxide wastes, hectares demanded can be physically located anywhere in the world.[11] While the global hectare is the unit recommended by Global Footprint Network, world-average hectares (*w*ha) and nation-specific hectares (*n*ha) are also possible units,[9,12] and results can be easily converted from one unit of measure to the other. Global hectares are needed to measure bioproductivity rather than surface area. Each global hectare has world-average productivity for all land types and provides more information than simply weight (which does not capture the extent of land and sea area used) or physical area (which does not capture how much ecological production is associated with that land). Global hectares are particularly useful for ranking different products based on their total ecological demands. World-average hectares are areas of a specific land type with world-average productivity for that land type (e.g., 1ha of forestland with the ecological production of the average forest hectare globally). Nation-specific actual hectares are physical areas of a specific land type located within a specific country and characterized by the bioproductivity of that country.

The choice of the unit of measure depends on the posed research question. The main advantage of using global hectares lies in the possibility to compare the ecological demands of nations and/or products across the globe. A productivity-based normalization is necessary to convert actual hectares into global hectares: yield and equivalence factors (EQFs) are used for this purpose. The yield factor (YF) captures the difference between local and global (world-average) productivity.[9,10] It is calculated as the ratio between the yield for the production of each product in a specific nation and the average yield for the production of the same product in the world. In each year, every country has its own set (one for each of the six land types) of YFs.

The EQF captures the difference between the productivity of a given land type and the world-average productivity of all biologically productive land types.[9,10] Equivalence factors are currently calculated by Global Footprint Network by using the Global Agro-Ecological Zones model, which provides data on potential agricultural suitability. The EQF for a land type depends on its level of potential agricultural suitability relative to world-average suitability. A set of EQFs is calculated on a yearly basis. The EQF for marine area is calculated such that a single global hectare of pasture will produce an amount of calories of beef equal to the amount of calories of salmon that can be produced by a single global hectare of marine area. The EQF for inland water is set equal to the EQF for marine area. The set of YFs and EQFs is annually updated and released by Global Footprint Network.

Mathematics of the Ecological Footprint

The EFA may have various purposes and be applied in different contexts. Here we refer to the Ecological Footprint within a geographically delineated area, e.g., a nation and the Ecological Footprint of a product.

Ecological Footprint of a Nation

The Ecological Footprint of a nation, EF_N, is a function of the number of inhabitants, Pop, and the footprint of consumption of a single inhabitant, $EF_{C,\,i}$, associated with the ith product and/or waste:

$$\mathrm{EF_N} = \mathrm{Pop}^* \sum_{i=1}^{n} \mathrm{EF}_{C,\,i} \tag{1}$$

where the per capita footprint of consumption is the sum of the per capita footprint of production ($EF_{P,\,i}$) and imports ($EF_{I,\,i}$) minus the footprint of exports ($EF_{E,\,i}$) for the ith product and/or waste.

$$\mathrm{EF}_{C,\,i} = \mathrm{EF}_{P,\,i} + \mathrm{EF}_{I,\,i} - \mathrm{EF}_{E,\,i} \tag{2}$$

In calculating footprints of production, imports, and exports for each ith product and/or waste. the reference formula is

$$\mathrm{EF}_i = \sum_{j=1}^{6} \frac{T_i}{Y_{N,\,i}} * YF_{N,\,j} * EQF_j = \sum_{j=1}^{6} \frac{T_i}{Y_{W,\,i}} * EQF_j \tag{3}$$

where

- T_i is the produced, imported, or exported amount of each product i (in tonnes).
- $Y_{\mathrm{N},\,i}$ is the annual yield (t ha^{-1} yr^{-1}) for the production of each product, i, in the nation, N. This is calculated as the tonnes of product, i, produced annually divided by all areas in the nation on which this product is grown.
- $Y_{\mathrm{W},\,i}$ is the world-average (W) annual yield (t ha^{-1} yr^{-1}) for the production of each product, i, given by the tonnes of product, i, produced annually across the world divided by all areas in the world on which this product is grown.
- $YF_{\mathrm{N},\mathrm{j}}$ is the YF specific to nation, N, and land type, j.
- EQF_j is the EQF for land type, j.

From these equations, it can be derived that the Ecological Footprint of a population is a function of four main factors:[9,10] 1) population size; 2) average standard of living; 3) the average productivity of land and water ecosystems; and 4) the efficiency of resourcing, harvesting, processing, and use.

Biocapacity is an aggregate measure of the amount of bioproductive land available, which is weighted by the productivity of that land. It represents the ability of the biosphere to produce crops, livestock (pasture), timber products (forest), and fish as well as to take up carbon dioxide in forests. It is calculated as reported in Eq. 4:

$$\mathrm{BC_N} = \sum_{j=1}^{6} A_{N,\,j} * YF_{N,\,j} * EQF_j \tag{4}$$

where $A_{N,\,j}$ represents the estimated bioproductive area expressed in nation-specific hectares that is available. $YF_{N,\,j}$ and EQF_j have been defined above.

Biocapacity represents the maximum theoretical rate of resource supply that can be sustained by a territory under prevailing technology and management schemes. Thus, biocapacity depends on natural conditions as well as on prevailing land-use practices in, for example, farming and forestry.[10]

The Ecological Footprint of a population can be directly compared with the biocapacity of the area where the population resides in the same way that expenditure is compared against income in financial terms.[4,10] The resulting "biocapacity balance" reveals whether that population is living, in net terms, within or beyond its local ecological means. This ecological balance has significance at both the global and national level, though it has to be noted that living within one's ecological means does not yet guarantee ecological sustainability.[13,14] Additional criteria such as biodiversity conservation, soil preservation, decreased release of toxic elements, and the like should be tracked in a comprehensive and multidisciplinary sustainability assessment. Further, countries with a biocapacity reserve (where biocapacity > Ecological Footprint) may use their available biocapacity to satisfy their own domestic consumptions or export ecological resources to other nations. This is generally the case for low-income countries that use only a fraction of their locally available resources, like some African and Latin America countries.[13,15,16]

Conversely, countries with a biocapacity deficit (where Ecological Footprint > biocapacity) must rely on biocapacity from outside their own borders or draw down their own natural budget. Many high-income countries (like the United States, Canada, and some western European countries such as Italy, the United Kingdom, and France) have footprints several times larger than their domestic biocapacity.[13,15,16] Such a biocapacity deficit is becoming an increasing economic risk for countries, particularly in a world of growing global overshoot. It highlights a country's dependence on additional external goods and ecological services, which are provided through one or more of the three following mechanisms:[10,13] 1) the biocapacity trade deficit, which consists of net import of biocapacity from other regions of the world; 2) the biocapacity deficit due to depletion, due to an overuse of local resources; and 3) the demand on biocapacity due to occupation of global commons, such as emissions of greenhouse gases into the global atmosphere (rather than domestic absorption) or fishing in international waters.[17]

The analysis of biocapacity and Ecological Footprint trends reveals how human consumption is changing over time. At the global level, the latest data released[13,16] show that humanity is currently operating in a state of overshoot: in other words, demand for natural resources exceeds the regenerative capacity of existing natural capital by at least 50% according to calculations of Global Footprint Network.[15] Furthermore, the gap between Ecological Footprint and biocapacity globally has been continuously increasing since the mid-1970s (Figure 1).

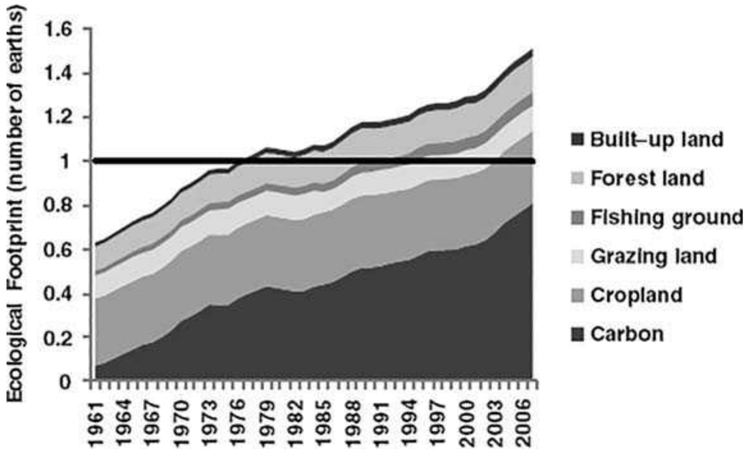


FIGURE 1 Humanity's Ecological Footprint, 1961–2007.

In 2007, humanity's total Ecological Footprint worldwide was 18 billion gha; with world population at 6.7 billion people, the average person's footprint was 2.7 gha.[13,16] However, there was only 11.9 billion gha of biocapacity available that year, equivalent to 1.8 gha per person. This overshoot of approximately 50% means that in 2007, humanity used the equivalent of 1.5 earths to support its consumption. In other words, it would have taken the earth approximately a year and a half to regenerate the resources used by humanity in that year. The largest Ecological Footprint component was the carbon footprint. This has increased by 35% since 1961 and currently accounts for more than half of the global Ecological Footprint.[13,16]

Even if the earth has a high resilience, prolonged biocapacity deficit is not possible since vital ecosystems and nonrenewable stocks would be depleted due to insuperable ecological and thermodynamic constraints. Also, it is not obvious that high-input agriculture can maintain its yields in the long run, particularly in the face of soil loss and potential phosphate limitations. It has become an urgent task to reduce our consumption levels back within the limits of our ecological budget.[14]

The growing global trends, however, hide significant regional variation (Figure 2).

Both demand on and supply of biocapacity are unevenly distributed across the world. Ecological Footprint and biocapacity values can therefore be used to develop new criteria for distinguishing among world nations. For instance, an alternative approach can be used to look at countries, based on their "biocapacity balances" (Figure 3), which helps to identify where resources are located and who uses what and to what extent. While in 1961, approximately 80% of the world population was living in countries characterized by a biocapacity remainder, in 2007, most of the world population was living in countries running a biocapacity deficit situation.

The total Ecological Footprint demanded by a country is strongly related to GDP[18,19] and changes accordingly, in both its extent and its composition, among high-, middle-, and low-income countries (Figure 4).[13]

Generally, high-income countries have per-capita Ecological Footprint values nearly 3 times higher than the world average, the majority of which is from the carbon footprint (approximately 65% of the total value). Conversely, middle- and low-income countries have average Ecological Footprint values that are equal to and lower than the world average, respectively. These countries are frequently characterized by transition economies, in which the carbon footprint component, although increased over the last decades, still constitutes less than 50% of the overall demand (Figure 5).

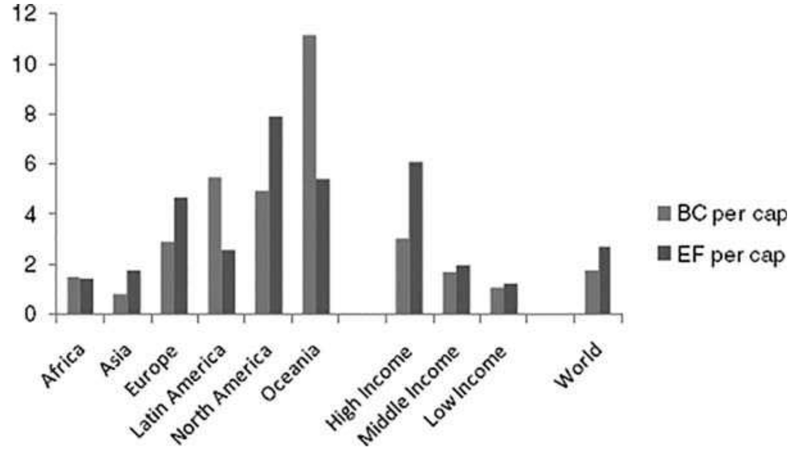


FIGURE 2 Ecological footprint (EF) and biocapacity (BC) by income level and country. The unequal distribution of human demand on bioproductive lands was also investigated by White.[20]

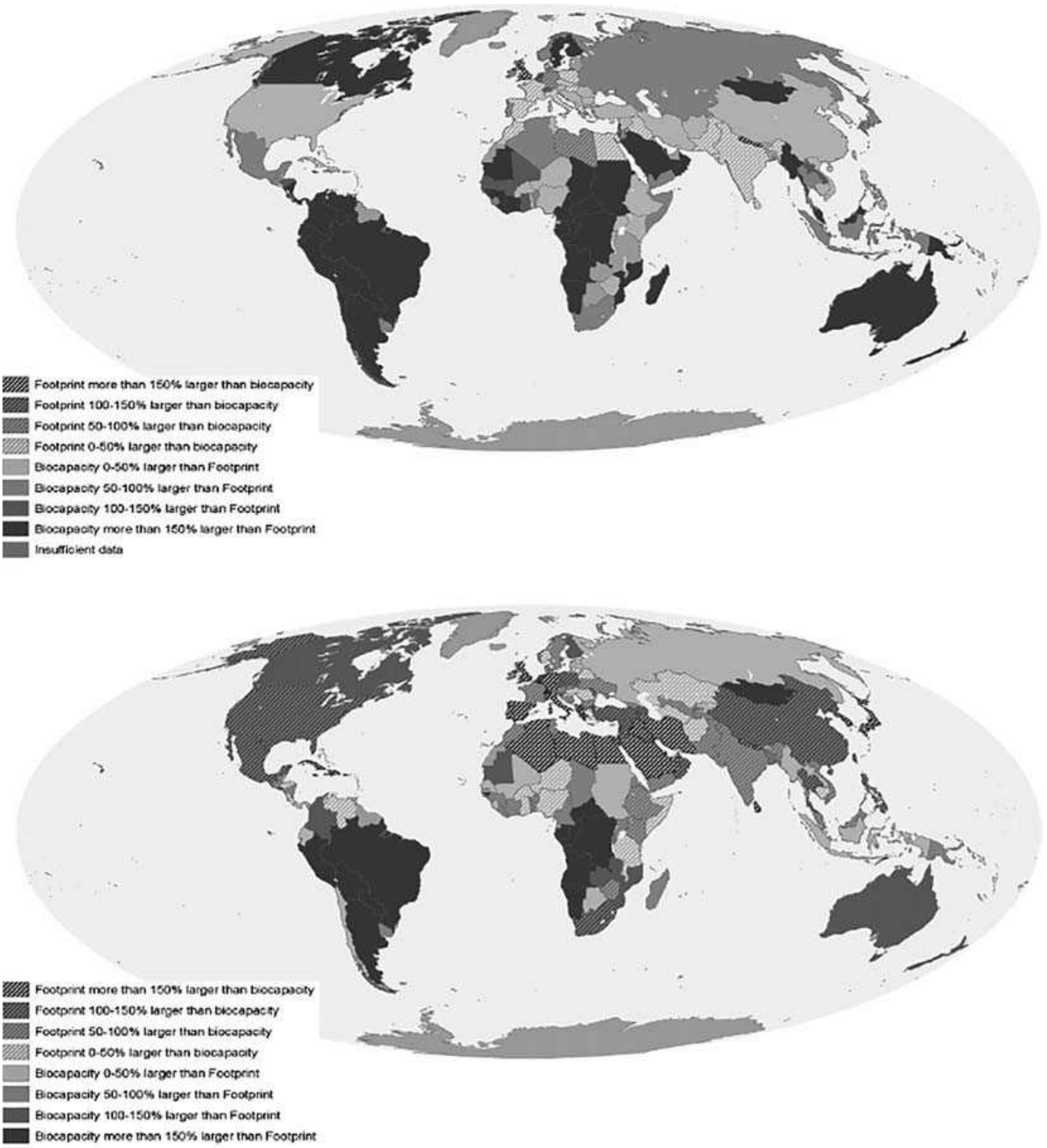


FIGURE 3 Biocapacity reminder/deficit status for world countries in 1961 and 2007. Green nations represent countries where the local biocapacity is greater than their residents' footprint (biocapacity reminder countries); red represents countries where the footprint is greater than local biocapacity (biocapacity deficit countries).

Ecological Footprint of a Product

The Ecological Footprint of a final or intermediate product is defined as the total amount of resources and waste assimilation capacity required in each of the phases required to produce, use, and/or dispose of that product.[8] The lifecycle boundaries can be flexible and changed according to the aim and scope of the analysis i.e., from cradle to gate (production to distribution) or cradle to grave (production to destruction). The Ecological Footprint is evaluated by considering all the direct and indirect inputs that are associated with the analyzed system for its entire life cycle. Each of these inputs is converted in terms

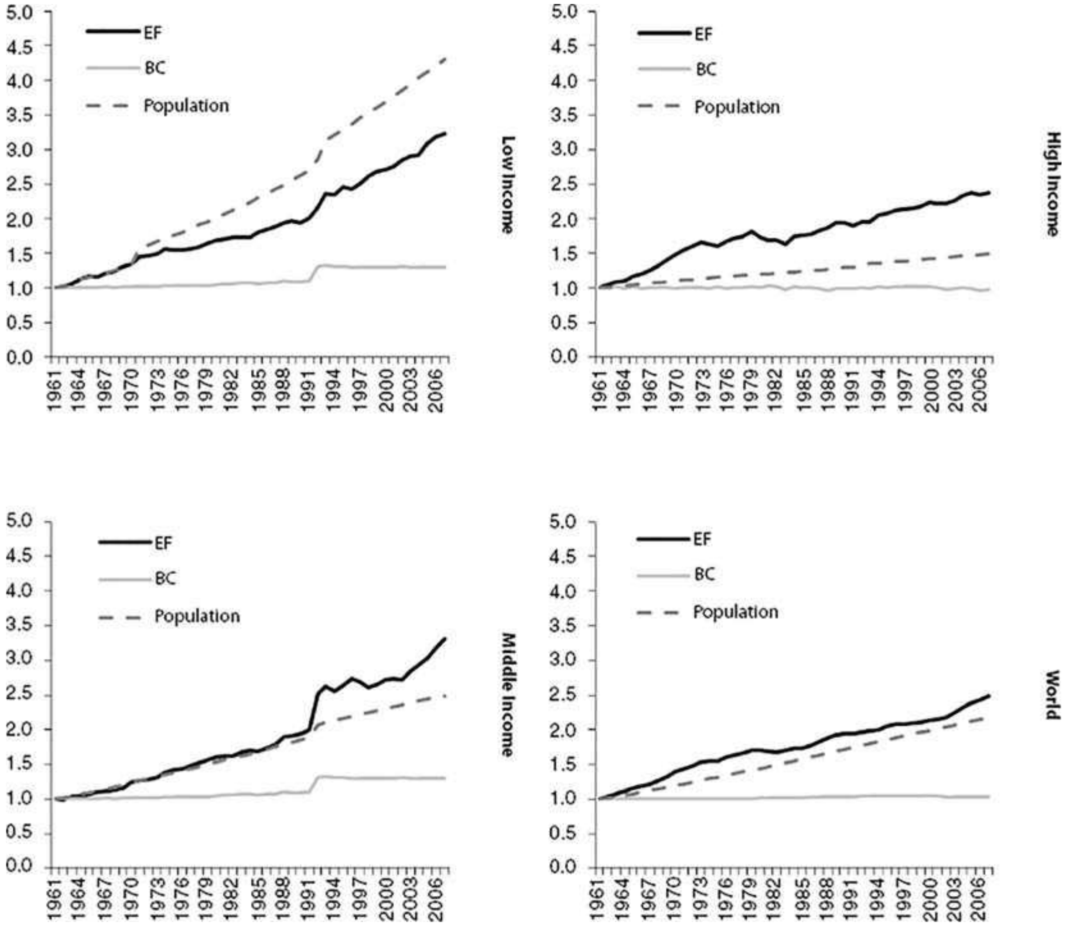


FIGURE 4 Variations on Ecological Footprint (EF), biocapacity (BC), and population for the world and low-income, middle-income, and high-income countries, indexed to 1961.

of the global hectares needed to support their production The EFp is expressed in units of global hectare years (gha yr), not just global hectares.[8] As the Ecological Footprint is strictly related to the production system, the way a product is produced should be clear and identifiable. The functional unit for the analyzed system should be also defined as well as the temporal and spatial boundaries.

There are two widely used approaches for calculating the Ecological Footprint of a product, both standards compliant: process-based life-cycle assessment (P-LCA) and environmentally extended input-output life-cycle assessment.[8] Process-based life-cycle assessment has the advantage of a large amount of detail, as individual product types and even brands can be analyzed, with the general disadvantage of lacking complete upstream coverage of the production chain (e.g., truncation error). Extended input-output life-cycle assessment has the advantage of full upstream coverage but the disadvantage of generality, as input-output tables typically do not disaggregate down to the level of individual product types (e.g., homogeneity assumption).

Following the P-LCA, Ecological Footprint of product ($\mathrm{EF_P}$) is given by the sum of the footprints of each input consumed and disposed within the life cycle of the production process as reported below.

$$\mathrm{EF_P} = \sum_{j=1}^{6}\sum_{i=1}^{n}\frac{T_i}{Y_{N,i}} * \mathrm{YF}_j * \mathrm{EQF}_j$$

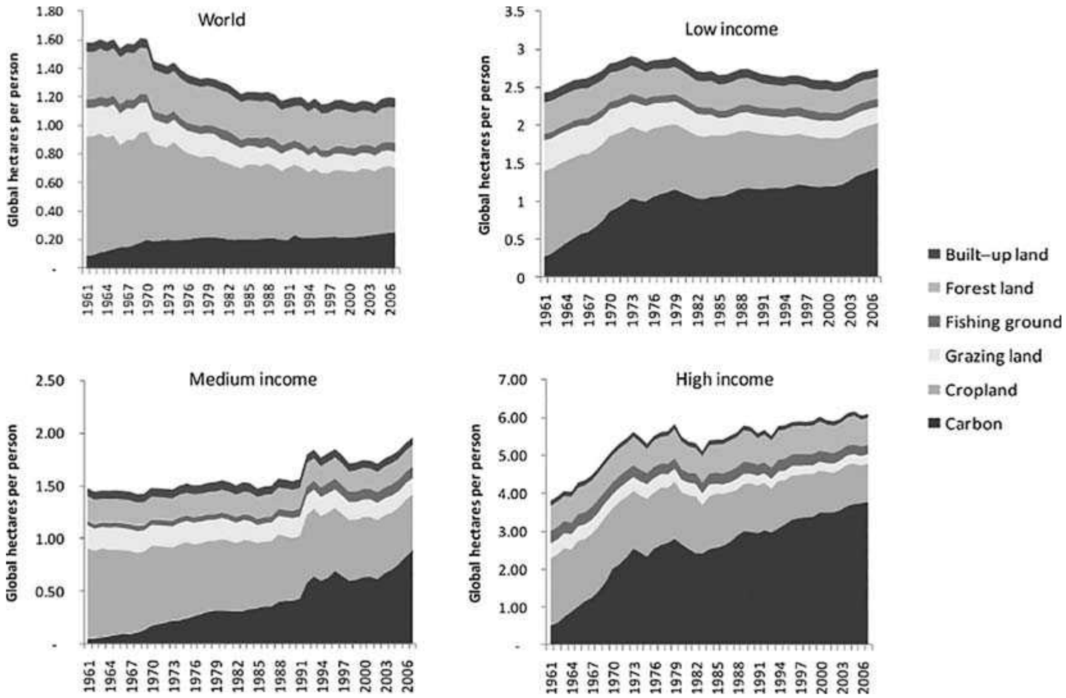


FIGURE 5 Per-capita Ecological Footprint for the world and high-, middle-, and low-income countries by land type, 1961–2007.

$$= \sum_{j=1}^{6}\sum_{i=1}^{n} \frac{T_i}{Y_{W,i}} * \mathrm{EQF}_j \tag{5}$$

where the variables are as follows:

- i refers to the n-input needed
- j refers to the six different land-use types (cropland, grazing land, fishing grounds, forest area, built-up land, and carbon uptake land)
- YF_j is the YF of the jth land type
- EQF_j is the EQF of the jth land type

Toward a Multi-Indicator Approach

Building on the premise that no single indicator is able to provide a full sustainability diagnosis and indicators should rather be used and interpreted jointly (i.e., the joint use of more than one indicator), this section reports some of the most interesting applications.

For instance, the HDI[21] can be used together with the Ecological Footprint to provide important insights on whether a high level of consumption is necessary for a high level of human development.[15] The HDI is a composite indicator used to rank countries by level of “human development” and then of well-being. It is a comparative measure of life expectancy, literacy, education, and standards of living for countries worldwide. The relationship between Ecological Footprint and HDI has two different categories (Figure 6).

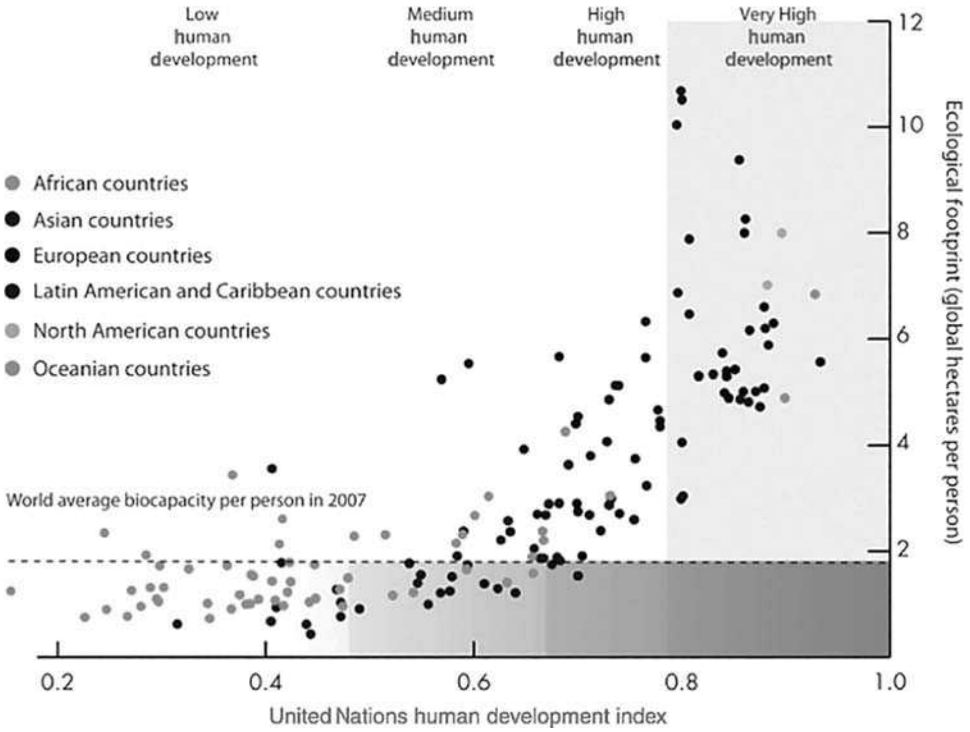


FIGURE 6 Human development index vs. Ecological Footprint, 2007.

While countries with a low level of development (HDI <0.8) report small variations in Ecological Footprint per capita, countries with a higher HDI value than the threshold show significant variations in Ecological Footprint per capita (for example, from Peru with 1.5 gha per person to Luxembourg with over than 9 gha per person). However, as development increases beyond a certain level, so does per-person Ecological Footprint; this is up to a point where small gains in HDI come at the cost of very large Ecological Footprint increases. Moreover, several countries with a high level of development have a similar per-capita Ecological Footprint to countries with a much lower level of development. Together with the breakdown in connection between wealth and well-being above a certain level of GDP per capita, this indicates that a high level of consumption is not necessarily required for a high level of development or well-being.

Ecological Footprint and biocapacity temporal trends have been compared with economic indicators such as GDP and the ISEW.[22] In particular, it has been shown that some (western) countries are characterized by increased economic wealth (measured by GDP) and growing environmental pressure (documented by Ecological Footprint trend) at the expense of a decrease in environmental sustainability (as shown by the increasing gap of Ecological Footprint and biocapacity) and a stagnation or even decrease in welfare (as ISEW demonstrates; Figure 7).[22]

A comparison between Ecological Footprint and life satisfaction—used as a subjective measure of well-being— was performed to understand the nature of their relationship and to test whether it could be possible for humans to reduce their levels of resource consumption without compromising well-being and life satisfaction and, at the same time, protecting the environment.[23,24] The relationship between Ecological Footprint and life satisfaction has two different paths. While countries with an Ecological Footprint of less than 4 gha per capita report significant variations in life satisfaction, countries with higher values (Ecological Footprint >4 gha per capita) report small variations in life satisfaction (Figure 8).

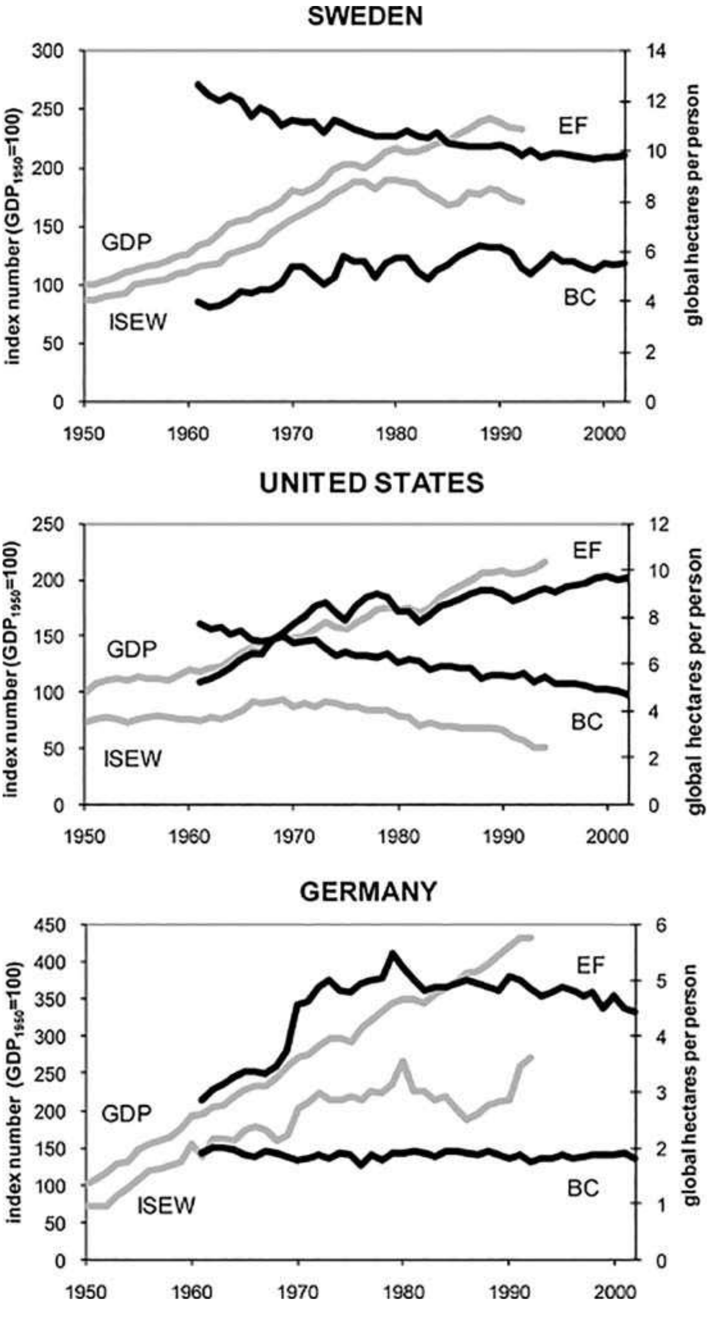


FIGURE 7 The comparison among temporal trends of GDP vs. ISEW (left scale) and Ecological Footprint (EF) vs. biocapacity (BC) (right scale) for Sweden, United States, and Germany.

In general, people with high consumption and income levels are more satisfied with life than people with lower consumption and income levels despite the differing levels and patterns of consumption that are necessary to obtain the same level of satisfaction. On the contrary, people with low consumption and income levels are less satisfied, but the same unit of consumption produces different perceptions of satisfaction (life satisfaction values are more unpaired).

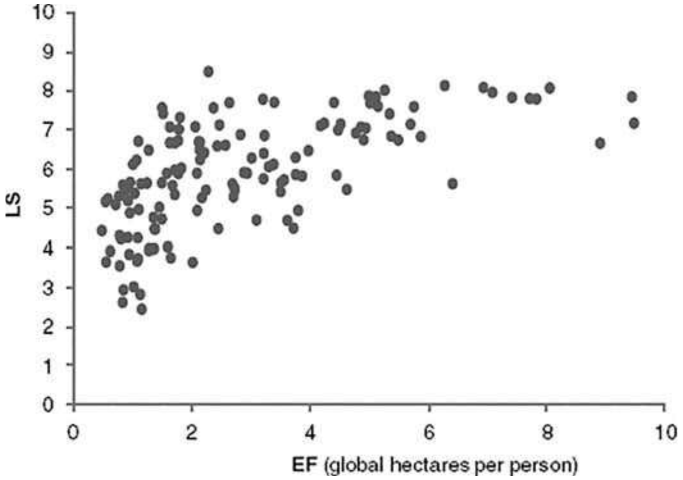


FIGURE 8 Life satisfaction (LS) vs. Ecological Footprint (EF) for 130 world countries.

Biocapacity and Ecological Footprint temporal trends have recently been used to develop alternative interpretations of the geopolitical context of nations around the world.[25,26] Based on their development paths, Ecological Footprint and biocapacity trends for nations around the world can be grouped into four main dynamic typologies: parallel, scissor, wedge, and descent.[25] Each typology corresponds to a particular environmental situation, the implications of which could have extreme relevance for environmental management, economic, and social prosperity as well as the development of sustainability policies. In particular, the role of biocapacity is highlighted in maintaining healthy economies, in offering an acceptable quality of life, and as an essential asset to ensure national competitiveness.[27]

Finally, under the recent EU-funded One Planet Economy Network Europe (OPEN:EU) project, three indicators have been identified as useful and complementary in assessing environmental issues—ecological, carbon, and water footprints—and therefore grouped together to form a suite of indicators called "footprint family."[7] Although not yet comprehensive, this suite provides a quantifiable platform for discussions regarding the limits to biotic resource and freshwater consumption and greenhouse gas emissions, as well as how to address the sustainability of natural capital use across the globe, thus enabling decision makers to more easily understand the environmental consequences of economic activities.

New Insight in Footprint Theory: Toward a Three-Dimensional Ecological Footprint Geography

The presence of global overshoot proves that the current human economy partially relies on natural capital depletion rather than just on sustainable flow consumption.[4] Considering natural capital and its limits, differentiating between these two components is fundamental for environmental planning and management. Recently, a variant of the classical Ecological Footprint model has been proposed, where a distinction between depletion of natural capital stocks and use of natural capital flows is operated.[28] The Ecological Footprint was redesigned as a three-dimensional model (^{3D}EF) with two relevant components, called size (EF_{size}) and depth (EF_{depth}), related to the two different uses of natural capital. The EF_{size} regards the appropriation of the so-called "income" of natural capital, i.e., the consumption of flows of resource yearly regenerated by natural cycles. It is the spatial component expressed in global hectares and plotted in the (x, y) plane. By definition, its value ranges from zero to biocapacity.

On the other side, the EF_{depth} regards the depletion of natural capital, which is the use of stocks of resources requiring a regeneration time longer than the flows. It is the intensity component plotted on the z-axis. In particular, the EF_{depth} component arises when overshoot is present and expresses the number of years it would theoretically take to regenerate the natural capital used in one year. By definition, its value ranges from 1 (the reference value called "natural depth") upward (additional depth).

It should be remarked that the two approaches, classical and three-dimensional, are simply two different ways of representing the same footprint values. The ^{3D}EF originates from the fact that flow and stock are technically incommensurable and cannot be summed up because the former is consumed each year and regenerated the following year, whereas the latter represents the irreversible erosion of natural stocks that add up from year to year into an accumulated "environmental debt."

Footprint size and depth have been characterized by opposite trends:[17] 1) EF_{size} grew continuously until mid-1970s, when it reached the asymptote (i.e., the earth biocapacity), and has remained constant ever since; and 2) EF_{depth} has remained equal to the natural depth until the appearance of overshoot, and it has subsequently been growing. Recently, the ^{3D}EF model has been theoretically applied to national case studies with the aim of enhancing the significance and potential usefulness of the Ecological Footprint in tracking relevant issues in the sustainability debate, such as the differentiation between resource stocks and flows.[17] Moreover, EF_{size} can be used as a proxy to highlight the existing intragenerational (in)equity in the appropriation of resources and ecological services by the residents of different nations. At the same time, EF_{depth} enables the relationships between current and future generations to be examined.

Although several questions remain to be addressed, the implementation of both a multilateral trade framework and the ^{3D}EF model in the National Footprint Accounts could form the basis for a new Ecological Footprint-based geography able to differentiate pressures on flows and stocks and identify the spatial/geographical location of such pressures.[4]

Applications of Ecological Footprint

Territorial Systems

Examples of national Ecological Footprint studies can be found in the literature,[29,42] although the most comprehensive set of national Ecological Footprint assessments is represented by Global Footprint Network's National Footprint Accounts.[16]

As of today, more than 35 nations have engaged with the Global Footprint Network directly, 17 nations have completed reviews of the Ecological Footprint, and a few have formally adopted it. Wales has adopted the Ecological Footprint as its headline indicator for sustainability. The Swiss government has incorporated the footprint into its national sustainable development plan. Japan includes the footprint as a measure in its environmental plan, the United Arab Emirates is using the Ecological Footprint as a tool to recommend and assist in the development of longterm science-based policies,[43] and Ecuador has set official footprint reduction targets in its 2009–2013 National Development Plan. There are several other countries that are currently collaborating with Global Footprint Network.[44]

Among nongovernmental organizations, WWF (World Wide Fund for Nature) International, one of the world's most influential conservation organizations, uses the Ecological Footprint in its communication and policy work for advancing conservation and sustainability. WWF has recently established a target of bringing humanity out of overshoot by 2050 and is actively pursuing this goal through its One Planet programs.

Numerous applications have also been performed at various subnational scales.[45–53] Currently, there are two suggested methodologies for subnational Ecological Footprint evaluations: component (or bottom-up) and compound (or top-down) approaches. The component method starts from specific individual consumption and waste production data to then calculate the total Ecological Footprint. While the method is detailed and flexible, several problems, including double counting, the lack of detailed data, and specific conversion factors, make this approach less acceptable. The compound approach evaluates the subnational Ecological Footprint by scaling the national Ecological Footprint value according to differences in consumption and life style. This is the widest used and more complete method.[54,8]

Products and Services

Despite its diffusion and popularity, product Ecological Footprint applications are still scarce, though a few interesting case studies exist. Studies on cultivation of tomatoes,[55] conventional vs. organic wine farming,[56] nectarine production,[57] shrimp and tilapia aquaculture,[58] marine aquaculture of reef fish,[59] and fisheries products[60] have been performed to highlight appropriation of natural capital, efficiency of natural resource use, and environmental pressure. Evaluations of the environmental impact of farms[61] and dairy production[62] as well as assessment of economic and ecological carrying capacity of crops[63] have been proposed via the combined use of the Ecological Footprint with other methods, such as life-cycle assessment, emergy analysis and economic cost, and return estimation.

In the context of product, Ecological Footprint has been used as a basis for the elaboration of the model of the double pyramid, which flanked the food pyramid with the environmental pyramid.[64] This model highlights that, in general, modern lifestyles produce a growing impact on the planet; it tries to promote eco-sustainable life and eating styles.

In the context of industrial processes, the potential of using EFA as an environmental indicator for the textile sector has been considered, although the contribution of wastes other than carbon dioxide should be included in the footprint methodology.[65]

The EFA has also been applied to the tourism sector. Starting from some pioneering studies,[66,67] the EFA has received attention as a key environmental indicator of sustainable tourism,[68] and several studies have been published at different scales.[69–72]

Over time, several tools and software have been developed for Ecological Footprint assessment for both territorial systems and products, although only a few are standard compliant. Most of them are freely downloadable from Web sites such as the footprint calculator (available at http://www.footprintnetwork.org/en/index.php/GFN/page/personal_footprint/). The REAP (Resource and Energy Analysis Programme) software (available at http://www.sei.se/reap) is a scenario based, integrated resource–environment modeling tool developed by the Stockholm Environment Institute to help local authorities in the U.K. make decisions about how to reduce their Ecological Footprint.

The Footprinter software (available at http://www.foot-printer.com), developed by Best Foot Forward, is comprehensive and powerful analytical software for carbon and Ecological Footprint assessment for products and organizations. It is based on the use of EcoIndex database.

Role of Business

The Ecological Footprint has been used to evaluate the environmental pressure of production processes; this type of investigation is becoming increasingly important to integrate sustainability issues (as natural capital consumption) into industrial and business decision-making processes. In a recent World Business Council for Sustainable Development project (WBCSD), several companies and industries came together to assess their role in helping to shape a future sustainable society for mankind.[73] With the help of Global Footprint Network, the consequences, in Ecological Footprint and biocapacity terms, of the hypothetical scenarios (up to the year 2050) envisioned by WBCSD were estimated. Results from this study showed that humanity will likely require the equivalent of 2.3 planets' worth of resources upon following a business-as-usual scenario or, conversely, 1.1 planets' worth of resources by implementing all the envisioned actions and activities, such as 50% reduction in CO_2 emissions compared with 2005 levels; enhanced forest productivity using better management techniques and an extension of their acreage between 2030 and 2050; increased crop productivity (+2% each year over past trends) due to technological advances and the diffusion of best practices; and changes to the average global nutritional regime, in terms of both diet and calorie content.[74]

Weakness and Limitations of EFA

Despite its popularity, EFA, like most indicators, is not exempt from criticisms regarding philosophical as well as methodological issues. In this section, a brief summary of the main weaknesses and limitations is offered. For further details on this topic see, for example, Best et al.,[75] Fiala,[76] Kitzes et al.,[77] and van der Bergh and Verbruggen.[78]

The first criticism is that the EFA cannot be fully defined as a measure of sustainability. The EFA research question is limited to identify the extent to which humanity is consuming bioproductive land compared with the available biocapacity.[21] We believe consuming resources within the capacity of the planet is a first necessary although not sufficient criterion for sustainability; as such, in order to depict a comprehensive picture of the system analyzed, it is strongly recommended to combine the Ecological Footprint with other complementary indicators (environmental as well as social and economic).

The use of a spatial unit makes some impacts difficult to determine.[75] Ecological Footprint is not able to directly account for all resources that cannot be referred to in spatial terms. This is the case for the depletion of nonrenewable deposits, such as metals, minerals, or fossil fuel reserves. For processes of extraction and refining, only the CO_2 emissions related to these processes are accounted for. The use of fossil fuels is evaluated in an indirect way, considering the amount of forestland required for the absorption of the CO_2 that is emitted. To date, carbon dioxide is the only greenhouse gas accounted for, and its associated footprint relies on the assumption that all emissions are absorbed only by forests and the oceans, neglecting carbon uptake by other biomes.

Other missing elements in EFA are freshwater consumption and soil erosion, even if the latter could be accounted for, at least theoretically. A possible way to include the overexploitation phenomena possible in agricultural land as well as in other land types into the classical Ecological Footprint framework was proposed by Bastianoni et al.[79] As EFA is unable to show unsustainable practices and their consequences, when agricultural EFA is performed, there can be some misunderstanding and misinterpretation of the results. Sometimes, it seems that EFA encourages more intensive farming, as this increases agricultural intensities, resulting in a higher biocapacity. The Ecological Footprint shows pressures that could lead to degradation of natural capital (e.g., reduced quality of land or reduced biodiversity) but does not predict this degradation.

Furthermore, multifunctional land-use patterns are not considered, in order to prevent double counting.[79] Each hectare is counted only once, even though it might provide multiple services. Counting them multiple times would produce an overestimation on Ecological Footprint.

Finally, EFA also is not able to capture the impacts due to the release of long-life toxic materials (e.g., pollution in terms of waste generation, toxicity, eutrophication, etc.), for which no regenerative capacity exists.

Conclusion

This entry offers a comprehensive insight to the EFA. Based on its simple logic and unit of measure, the Ecological Footprint has become a very popular sustainability indicator preserving scientific rigor on an ecological and thermodynamic basis.

By tracking a wide range of human activities, the Ecological Footprint is able to monitor the combined impact of anthropogenic pressures that are more typically evaluated independently and can thus be used to understand, from multiple angles, the environmental consequences of human activities.

The main strength of this methodology is its ability to explain, in simple terms, the concept of ecological limits, thus helping to safeguard the long-term capacity of the biosphere to support mankind and understand how resource issues are linked with economic and social issues.

One of the positive characteristics of this methodology is its ability to make visible aspects that are traditionally invisible for conventional economic analyses. For instance, the choice of an area as a unit of measure reflects the fact that many basic ecosystem services and ecological resources are provided by surfaces where photosynthesis takes place. Unfortunately, these surfaces are limited by physical and planetary constraints: reporting results in terms of an area helps to better communicate the existence of physical limits to the growth of human economies.

Currently, the Ecological Footprint is a robust method widely used to give a measure of the (un)sustainability of consumption patterns at different scales as well as to establish the natural capital requirement of products, services, and activities.

Due to the growing number of applications, Global Footprint Network has released the Ecological Footprint Standards[8] in order to enhance the consistency and quality of footprint assessments. This standard contains a list of mandatory requirements for standards compliance. The document also suggests the best way to present the results avoiding distortion and misinterpretations. The National Footprint Accounts are updated and improved every year, and the annual release of the newest version ensures that the method is more robust, reliable, and detailed than previous versions, though some shortcomings still exist and remain to be addressed. The Ecological Footprint Standards are also periodically updated. The information derived from Ecological Footprint assessments could be included in the environmental management and future planning of territories to promote more competitive lifestyles, resource-efficient strategy, and a more effective management of our ecological assets.

References

1. Rees, W.E. Ecological Footprints and appropriated carrying capacity: what urban economics leaves out. Environ. Urban. **1992**, *4* (2), 121–130.
2. Wackernagel, M. Ecological Footprint and Appropriated Carrying Capacity: A Tool for Planning Toward Sustainability. Ph.D. Thesis. School of Community and Regional Planning, The University of British Columbia, 1994.
3. Wackernagel, M.; Rees, W.E. *Our Ecological Footprint: Reducing Human Impact on the Earth,* 1st Ed.; New Society Publishers: Gabriola Island, Canada, 1996.
4. Wackernagel, M.; Rees, W.E. Perceptual and structural barriers to investing in natural capital: Economics from an Ecological Footprint perspective. Ecol. Econ. **1997**, *20*, 3–24.
5. Wackernagel, M.; Schulz, N.B.; Deumling, D.; Linares, A.C.; Jenkins, M.; Kapos, V.; Monfreda, C.; Loh, J.; Myers, N.; Norgaard, R.; Randers, J. Tracking the ecological overshoot of the human economy. Proc. Natl. Acad. Sci. **2002**, *99*, 9266–9271.
6. The World Conservation Union, United Nations Environment Programme and the World Wide Fund for Nature. *Caring for the Earth: A Strategy for Living Sustainably;* IUCN, UNEP and WWF: Gland, Switzerland, 1991.
7. Galli, A.; Wiedmann, T.; Ercin, E.A.; Knoblauch, D.; Ewing, B.; Giljum, S. Integrating ecological, carbon, and water footprint into a “footprint family” of indicators: Definition and role in tracking human pressure on the planet. Ecol. Indic. **2012**, *16*, 100–112.
8. Global Footprint Network, 2009. *Ecological Footprint Standards 2009;* Global Footprint Network: Oakland, available at http://www.footprintstandards.org. (accessed July 2011).
9. Galli, A.; Kitzes, J.; Wermer P.; Wackernagel M.; Niccolucci V.; Tiezzi E. An exploration of the mathematics behind the Ecological Footprint. Int. J. Ecodyn. **2007**, *2* (4), 250–257.
10. Monfreda, C.; Wackernagel, M.; Deumling, D. Establishing national natural capital accounts based on detailed Ecological Footprint and biological capacity assessments. Land Use Policy **2004**, *21*, 231–246.
11. Kitzes, J.; Peller, A.; Goldfinger, S.; Wackernagel, M. Current methods for calculating national Ecological Footprint accounts. Sci. Environ. Sustainability Soc. **2007**, *4* (1), 1–9.
12. Wiedmann, T.; Lenzen, M. On the conversion between local and global hectares in Ecological Footprint analysis, Ecol. Econ., **2007**, *60* (4), 673–677.
13. Ewing, B.; Reed, A; Galli, A.; Kitzes, J.; Wackernagel, M. *Calculation Methodology for the National Footprint Accounts,* 2010 Ed.; Global Footprint Network: Oakland, CA, 2010.

14. Kitzes, J.; Wackernagel, M.; Loh, J.; Peller, A.; Goldfinger, S.; Cheng, D.; Tea, K. Shrink and share: Humanity's present and future Ecological Footprint. Phil. Trans. R. Soc. Lond. B **2008**, *363* (1491), 467–475.
15. WWF International; Global Footprint Network; ZSL. *Living Planet Report 2010*. WWF: Gland, Switzerland, 2010, available at http://www.footprintnetwork.com. (accessed July 2011).
16. Ewing, B.; Goldfinger, S.; Oursler, A.; Reed, A.; Moore, D.; Wackernagel. M. The *Ecological Footprint Atlas 2010*; Global Footprint Network: Oakland, 2010.
17. Niccolucci, V.; Galli, A.; Reed, A.; Neri, E.; Wackernagel, M.; Bastianoni S. Towards a 3D national Ecological Footprint geography. Ecol. Modell. **2011**, *222* (16), 2939–2944.
18. Jorgenson, A.K. Unpacking international power and the Ecological Footprint of nations: A quantitative cross national study. Sociol. Perspect. **2005**, *48*, 383–402.
19. Jorgenson, A.K.; Burns, T.J. The political–economic causes of change in the Ecological Footprints of nations, 1991–2001: A quantitative investigation. Soc. Sci. Res. **2007**, *36*, 834–853.
20. White, T.J. Sharing resources: The global distribution of the Ecological Footprint. Ecol. Econ. **2007**, *64*, 402–410.
21. UNDP (United Nation Development Programme). *Annual Report 2009;* New York, 2009.
22. Niccolucci, V.; Pulselli, F.M.; Tiezzi, E. Strengthening the threshold hypothesis: Economic and biophysical limits to growth. Ecol. Econ. **2007**, *60* (4), 667–672.
23. Abdallah, S.; Thompson, S.; Michaelson, J.; Marks, N.; Steuer, N. *The (Un)Happy Planet Index 2.0: Why Good Lives Don't Have to Cost the Earth;* New Economics Foundation: London, 2009; 61 pp.
24. Patrizi, N.; Capineri, C.; Rugani, B.; Niccolucci, V. *"SocioEconomic Design and Nature": A Possible Representation through Ecological Footprint. Design and Nature* v. Brebbia, C.A., Carpi, A., Eds.; WIT Press: Southampton, U.K. WIT Trans. Ecol. Environ. **2009**, 527–534.
25. Niccolucci, V.; Tiezzi, E.; Pulselli, F.M.; Capineri, C. Biocapacity vs. Ecological Footprint of world regions: A geopolitical interpretation. Ecol. Indic. **2012**, *16*, 23–30.
26. Global Footprint Network. *The Ecological Wealth of Nations. Earth's Biocapacity as a New Framework for International Cooperation,* 1st Ed.; Goldfinger, S., Poblete, P., Eds.; Global Footprint Network: Oakland, CA, 2010.
27. Moore, D., Brooks, N., Cranston, G., Galli, A. *The Future of the Mediterranean. Tracking Ecological Footprint Trends.* Galli, A., Ed.; Global Footprint Network, available at http://www.footprintnetwork.org. (accessed July 2011).
28. Niccolucci, V.; Bastianoni, S.; Tiezzi, E.B.P.; Wackernagel, M.; Marchettini, N. How deep is the footprint? A 3D representation. Ecol. Modell., **2009**, *220*, 2819–2823.
29. Barrett, J.; Simmons, C. *An Ecological Footprint of the U.K.: Providing a Tool to Measure the Sustainability of Local Authorities*; Stockholm Environment Institute (SEI): Stockholm, Sweden, 2003, available at http://www.lande-con.cam.ac.uk/up211/EP04/readings/footprint_UK.pdf. (accessed July 2011).
30. Galli, A.; Kitzes, J.; Niccolucci, V.; Wackernagel, M.; Wada, Y.; Marchettini, N. Assessing the global environmental consequences of economic growth through the Ecological Footprint: A focus on China and India. Ecol. Indic. **2012**, *17*, 99–107.
31. Wackernagel, M., Lewan, L., Borgström Hansson, C. Evaluating the use of natural capital with the Ecological Footprint. Ambio, **1999**, *28*, 604–612.
32. Bicknell, K.B.; Ball, R.J.; Cullen, R.; Bigsby, H. New methodology for the Ecological Footprint with an application to the New Zealand economy. Ecol. Econ. **1998**, *27*, 149–160.
33. Fricker, A. The Ecological Footprint of New Zealand as a step towards sustainability. Futures **1998**, *30*, 559–567.
34. Van Vuuren, D.P.; Smeets, E.M.W. Ecological Footprints of Benin, Bhutan, Costa Rica and the Netherlands. Ecol. Econ. **2000**, *34*, 115–130.
35. Simpson, R.W.; Petroeschevsky, A.; Lowe I. An Ecological Footprint analysis for Australia. Aust. J. Environ. Manage. **2000**, *7*, 11–18.

36. Ferng, J.J. Using composition of land multiplier to estimate Ecological Footprints associated with production activity. Ecol. Econ. **2001**, *37*, 159–172.
37. Haberl, H.; Erb, K.H.; Krausmann, F.; Loibl, W.; Schulz, N.B.; Weisz. H. Changes in ecosystem processes induced by land use: Human appropriation of net primary production and its influence on standing crop in Austria. Global Biogeochem. Cycles **2001**, *15*, 929–942.
38. Lenzen, M.; Murray, S.A.; A modified Ecological Footprint method and its application to Aust. Ecol. Econ. **2001**, *37*, 229–255.
39. Lenzen, M., Murray, S.A. The Ecological Footprint—Issue and Trends. ISA Research Paper 01–03. The University of Sidney, 2003, available at http://www.isa.org.usyd.edu.au/publications/documents/Ecological_Footprint_Issues_and_Trends.pdf. (accessed July 2011).
40. McDonald, G.W.; Patterson, M.G. Ecological Footprint and interdependencies of New Zealand regions. Ecol. Econ. **2004**, *50*, 49–67.
41. Erb, K.H. Actual land demand of Austria 1926–2000: A variation on Ecological Footprint assessments. Land Use Policy, **2004**, *21*, 247–259.
42. Medved, S. Present and future Ecological Footprint of Slovenia: The influence of energy demand scenarios. Ecol. Modell. **2006**, *192*, 25–36.
43. Abdullatif, L., Alam, T., 2011. *The UAE Ecological Footprint Initiative. Summary Report 2007–2010.* Available at: http://awsassets.panda.org/downloads/en_final_report_ecological_footprint.pdf (accessed July 2011).
44. Global Footprint Network (GFN), 2010. 2010 Annual Report, available at http://www.footprintnetwork.org/im-ages/uploads/2010_Annual_Report.pdf. (latest access: July 2011). (accessed July 2011).
45. Folke, C.; Jansson, A.; Larsson, J.; Costanza, R. Ecosystem appropriation by cities. Ambio **1997**, *26* (3), 167–72.
46. Bagliani, M.; Galli, A.; Niccolucci, V.; Marchettini, N. Ecological Footprint analysis applied to a sub-national area: The case of the Province of Siena (Italy). J. Environ. Manage. **2008**, *86*, 354–364.
47. Niccolucci, V.; Galli A.; Bastianoni, S. Deriving environmental management practices with the Ecological Footprint analysis: A case study for the Abruzzo Region. In *Ecosystems And Sustainable Development 7*; Brebbia, C.A., Tiezzi, E., Eds.; WIT press, 2009; 195–204.
48. Bagliani, M.; Ferlaino, F.; Procopio, S. The analysis of the environmental sustainability of the economic sectors of the Piedmont Region (Italy). In *Ecosystems and Sustainable Development;* Tiezzi, E., Brebbia, C.A., Uso, J.L., Eds.; WIT Press: Southampton, U.K., **2003**; 613–622.
49. Barrett, J.; Vallack, H.; Jones, A.; Haq, G. A material flow analysis and Ecological Footprint of York. Technical Report. Stockholm Environment Institute: Stockholm, Sweden, 2002.
50. Lenzen, M.; Lundie, S.; Bransgrove, G.; Charet, L.; Sack, F. Assessing the Ecological Footprint of a large metropolitan water supplier: Lessons for water management and planning towards sustainability. J. Environ. Plann. Manage. **2003**, *46*, 113–141.
51. Birch, R.; Wiedmann, T.; Barret, J. *The Ecological Footprint of Greater Nottingham and Nottinghamshire—Results and Scenarios.* Stockholm Environment Institute (SEI), University of York: York, U.K., 2005.
52. Vergoulas, G.; Simmons, C. An Ecological Footprint analysis of Essex-East England. Best Foot Forward Ltd., Commissioned by Essex County Council, 2004.
53. Collins, A.; Flynn, A.; Wiedmann, T.; Barrett, J. The environmental impacts of consumption at a sub-national level: The Ecological Footprint of Cardiff. J. Ind. Ecol. **2009**, *10*, 9–24.
54. Kitzes, J.; Galli, A.; Bagliani, M.; Barrett, J.; Dige, G.; Ede, S.; Erb, K-H.; Giljum, S.; Haberl, H.; Hails, C.; Jungwirth, S.; Lenzen, M.; Lewis, K.; Loh, J.; Marchettini, N.; Messinger, H.; Milne, K.; Moles, R.; Monfreda, C.; Moran, D.; Nakano, K.; Pyhälä, A.; Rees, W.; Simmons, C.; Wackernagel, M.; Wada, Y.; Walsh, C.; Wiedmann, T. A research agenda for improving national Ecological Footprint accounts. Ecol. Econ. **2009**, *68* (7), 1991–2007.

55. Wada, Y. The appropriated carrying capacity of tomato production: The Ecological Footprint of hydroponic greenhouse versus mechanized open field operations. M.A. Thesis. School of Community and Regional Planning, University of British Columbia: Vancouver, Canada, 1993.
56. Niccolucci, V.; Galli, A.; Kitzes, J.; Pulselli, R.M.; Borsa, S.; Marchettini, N. Ecological Footprint analysis applied to the production of two Italian wines. Agric. Ecosyst. Environ. **2008**, *128*, 162–166.
57. Cerutti, A.K.; Bagliani, M.; Beccaro, G.L.; Bounous, G. Application of Ecological Footprint analysis on nectarine production: methodological issues and results from a case study in Italy. J. Cleaner Prod. **2010**, *18* (8), 771–776.
58. Kautsky, N.; Berg, H.; Folke, C.; Larsson, J.; Troell, M. Ecological Footprint for assessment of resource use and development limitations in shrimp and tilapia aquaculture. Aquacult. Res. **1997**, *28*, 753–766.
59. Warren-Rhodes, K.; Sadovy, Y.; Cesar, H. Marine ecosystem appropriation in the Indo-Pacific: A case study of the live reef fish food trade. Ambio **2003**, *32*, 481–488.
60. Parker, R.; Tyedmers P. Uncertainty and natural variability in the Ecological Footprint of fisheries: A case study of reduction fisheries for meal and oil. Ecol. Indic. **2012**, *16*, 76–83.
61. Van der Werf, H.M.G.; Tzilivakis, J.; Lewis, K.; Basset-Mens, C. Environmental impacts of farm scenarios according to five assessment methods. Agric. Ecosyst. Environ. **2006**, *118* (1–4), 327–338.
62. Thomassen, M.A.; de Boer, I.J.M. Evaluation of indicators to assess the environmental impact of dairy production systems. Agric. Ecosyst. Environ. **2005**, *111*, 185–199.
63. Cuandra, M.; Bjorklund, J. Assessment of economic and ecological carrying capacity of agricultural crops in Nicaragua. Ecol. Indic. **2007**, *7* (1), 133–149.
64. Barilla Center for Food and Nutrition. Double Pyramid 2011: Healthy Food for people, Sustainable for the Planet; Parma, IT, 2011, available at http://www.barillacfn.com/en/dp-doppiapyramid/position-paper. (accessed July 2011).
65. Herva, M.; Franco, A.; Ferreiro, S.; Alvarez, A.; Roca, E. An approach for the application of the Ecological Footprint as environmental indicator in the textile sector. J. Hazard. Mater. **2008**, *156* (1–3), 478–487.
66. Gossling, S.; Borgstrom, C.; Horstmeier, H.O.; Saggel, S. Ecological Footprint analysis as a tool to assess tourism sustainability. Ecol. Econ. **2002**, *43* (2–3), 199–211.
67. Hunter, C. Sustainable tourism and the tourist Ecological Footprint. Environ. Dev. Sustainability **2002**, *4*, 7–20.
68. Hunter, C.; Shaw, J. The Ecological Footprint as a key indicator of sustainable tourism. Tourism Manage. **2007**, *28* (1), 46–57.
69. Peeters, P.; Schouten, F. Reducing the Ecological Footprint of inbound tourism and transport to Amsterdam. J. Sustainable Tourism **2006**, *14*, 157–171.
70. Patterson, T.M.; Niccolucci, V.; Bastianoni, S. Beyond "more is better": Ecological Footprint accounting for tourism and consumption in Val di Merse, Italy. Ecol. Econ. **2007**, *62* (3–4), 747–756.
71. Patterson, T.M.; Niccolucci, V.; Marchettini, N. Adaptive environmental management of tourism in the Province of Siena, Italy using the Ecological Footprint. J. Environ. Manage. **2008**, *86* (2), 407–418.
72. Castellani, V.; Sala, S. Ecological Footprint and life cycle assessment in the sustainability assessment of tourism activities. Ecol. Indic. **2012**, *16*, 135–147.
73. WBCSD. *Vision 2050: The New Agenda for Business.* WBCSD: Geneva, Switzerland, 2010.
74. Moore, D.; Cranston, G.; Reed, A.; Galli, A. Projecting future human demand on the Earth's regenerative capacity. Ecol. Indic. **2012**, *16*, 3–10.
75. Best, A.; Giljum, S.; Simmons, C.; Blobel, D.; Lewis, K.; Hammer, M.; Cavalieri, S.; Lutter S.; Maguire C. Potential of the Ecological Footprint for monitoring environmental impacts from natural resource use: Analysis of the potential of the Ecological Footprint and related assessment tools for use in the EU's Thematic Strategy on the Sustainable Use of Natural Resources. Report to the European Commission, DG Environment, 2008.

76. Fiala, N. Measuring sustainability: Why the Ecological Footprint is bad economics and bad environmental science. Ecol. Econ. **2008**, *67* (4), 519–525.
77. Kitzes, J.; Moran, D.; Galli, A.; Wada, Y.; Wackernagel, M. Interpretation and application of the Ecological Footprint: A reply to Fiala. Ecol. Econ. **2009**, *68* (4), 929–930.
78. van den Bergh, J.C.J.M.; Verbruggen, H. Spatial sustainability, trade and indicators: An evaluation of the 'Ecological Footprint'. Ecol. Econ. **1999**, *29*, 61–72.
79. Bastianoni, S.; Niccolucci, V.; Pulselli, R.M.; Marchettini N. Indicator and indicandum: "Sustainable way" vs. "prevailing conditions" in Ecological Footprint definition. Ecol. Indic. **2012**, *16*, 47–50.

28

Environmental Legislation: Asia

Wanpen Wirojanagud

Introduction

The 21st century has been dubbed "The Asian Century." The region's economic growth is unprecedented, with the rise of China and India as the next economic superpowers.[1] China is the second largest economy and is predicted to take over after the United States as the largest economy in the world by 2020. India is one of the fastest-growing economies and is expected to become the world's third largest economy in the near future.[2–3] More than 58% of the world's population live in Asia and the Pacific region, and Asia is categorized both as the factory of the world as well as the booming market with tremendous growth potential. As a result, natural resources, which are the capital for development of both economic and social sectors, have been remarkably exploited, with a consequent substantial increase in environmental pollution. Without proper environmental management, the ecosystem continues to suffer, resulting in loss of biodiversity, depletion of ecosystem services, desertification, loss of fertile land, atmospheric pollution, aquatic and marine pollution, etc. The consequences of environmental degradation have become more and more acute and chronic over the years. Moreover, natural disasters are more frequent and more severe. As indicated in the Asia-Pacific Disaster 2010 Report of the United Nations (UN), "while the region generated one quarter of the world's GDP, it accounted for a staggering 85 per cent of deaths and 38 per cent of global economic losses due to natural disasters over the last three decades." The report concludes that Asia is the most disaster-prone region in the world.[4] The change in climate patterns has also become more severe.

Environmental status in Southeast Asia, as summarized in Table 1, evidently indicates that over the last decades pressure mostly from anthropogenic activities (rapid population growth, urbanization, economic growth, and consumptive lifestyle) has had a substantial impact on the plentiful natural resources and the environment. Natural resources depletion results in impact on agricultural

TABLE 1 Key Environmental Issues and Causes in ASEAN

Country	Shared Issues	Key Causes
Brunei	Seasonal smoke and haze; solid wastes	Transboundary pollution from land and forest fires; inadequate waste management facilities and practices
Cambodia	Soil erosion; sedimentation; water pollution; deforestation; loss of biodiversity; and threats to natural fisheries	Unmanaged waste and effluent discharge into Tonle Sap Lake; destruction of mangrove wetlands through extensive industrial and aquaculture development
Indonesia	Deforestation; loss of biodiversity; water pollution; air pollution in urban areas; national and transboundary seasonal smoke and haze; land degradation; pollution of Malacca Straits	Deficiencies in urban infrastructure—unmanaged industrial wastes and municipal effluents and waste; vehicle congestion and emissions; extensive land clearance and forest fires for pulp wood and oil palm production; extensive and unmanaged mining activities; national and transboundary industrial pollution; tourist developments in coastal regions beyond carrying capacity
Laos	Deforestation; loss of biodiversity; soil erosion; limited access to potable water; water-borne diseases	Land clearance; shifting cultivation; inadequate water supply and sanitation infrastructure
Malaysia	Urban air pollution; water pollution; deforestation; loss of biodiversity; loss of mangrove habitats; national and transboundary smoke/haze	Vehicle congestion and emissions; deficiencies in urban infrastructure industrial and municipal effluents; extensive land clearance and forest fires for pulp wood and oil palm production; unmanaged coastal developments; tourist developments in coastal regions beyond existing carrying capacity
Myanmar	Deforestation; loss of biodiversity; urban air pollution; soil erosion; water contamination and water-borne diseases	Land clearance; excessive mineral extraction; vehicle congestion and emissions; deficiencies in urban infrastructure—unmanaged industrial and municipal effluents
Philippines	Deforestation in watershed areas; loss of biodiversity; soil erosion; air and water pollution in Manila leading to waterborne diseases; pollution of coastal mangrove habitats; natural disasters (earthquakes, floods)	Illegal forest cutting; land clearance; rapid urbanization and deficiencies in urban infrastructure—unmanaged industrial and municipal effluents, inadequate water supply and sanitation; tourist developments in coastal regions beyond existing carrying capacity
Singapore	Industrial pollution; water shortages; waste disposal problems	Seasonal smoke/haze; limited natural fresh water resources; limited land available for waste disposal
Thailand	Deforestation; loss of biodiversity; land degradation and soil erosion; shortage of water resources in dry season and flooding in rainy season; conflict of water users; coastal degradation and loss of mangrove habitat; urban air pollution; pollution from solid waste, hazardous materials and hazardous waste	Sporadic development and destruction of watersheds; unmanaged aquaculture; tourist growth exceeding growth in carrying capacity; deficiencies in urban and rural infrastructure; freshwater resources polluted by domestic/industrial wastes, sewage, and contaminated runoff
Vietnam	Deforestation and soil degradation; loss of biodiversity; loss of mangrove habitat; water pollution and threats to marine life; groundwater contamination; limited potable water supply; natural disasters (e.g., floods)	Land clearance for industry; forest clearance and chronic impact of Agent Orange; extensive aquaculture and overfishing; growing urbanization and infrastructure deficiencies; inadequate water supply and sanitation (particularly in Hanoi and Ho Chi Minh City)

Source: Data from Nguyen.[5]

productivity; increased frequency of disasters such as floods, landslides, and soil erosion[5]; depletion of aquifers; deterioration of diversity; increased loads from municipal, industrial, and hazardous wastes; atmospheric pollution; and marine and aquatic pollution.[6]

Hence, the key policy question facing Asian governments is how to reconcile economic development and environmental protection. In other words, how to make a transition from the "grow first, clean later" approach to the policy of "sustainable development"—a holistic integration of economic, social, and environmental dimensions.[7] More importantly, the main driving forces in the development of environmental management tools in Asia are international agreements, the occurrence of environmental disasters, and non-harmonized legal framework. Asian countries have shown increasing interest in environmental management by which environmental legislation has been used as one of the important tools in this complex task of integrating environmental protection and economic development. Environmental legislation is among the most determined elements in environmental management toward sustainable development.[8] "Legislation is an important element of the institutional framework for environmental management. The role of legislation is to implement and enforce policy and to provide effective administrative and regulatory mechanisms."[9]

Relevantly, development and implementation of environmental law may involve interaction with legislation and administrative practices and institutions. Thus, environmental laws are undeniably increasingly important in Asia. Environmental law is defined as a body of state and federal statutes intended to protect the environment, wildlife, land, and beauty; prevent pollution and overcutting of forests; save endangered species; conserve water; develop and follow general plans; and prevent damaging practices.[9] The particular law gives individuals and groups the right to bring legal actions or seek court orders to enforce the protection, or demand revisions of private and public activity that may have detrimental effects on the environment.

In terms of the evolution of environmental legislation, two different types of statutes can be distinguished, referred as to the "first generation" and the "second generation" of environmental laws. Although there is no total agreement about how to characterize each generation, generally, the first generation of environmental legislation refers to "command-and-control statutes and regulations administered with technology-based standards and enforced by rule-of-law litigation."[10] Then arrived the "second generation" in the late 1980s with the new concept of "sustainable development" aimed at reconciling environmental protection and economic development.[11] The role of the people, particularly local communities, is highlighted, and compliance incentives and market-based mechanisms have been developed to encourage compliance and provide flexibility.[12]

In the Asia-Pacific region, environmental legislation began to emerge in the 1970s, the same period as the growing global interest in environmental protection. Examples of the early framework laws and regulations for environmental protection are presented in Table 2. Noticeably, there is no lack of environmental legislation in Asia. However, there is a wide gap between law and practice. The countries have to put an emphasis on ways and means to increase the effectiveness of the implementation and enforcement of environmental law.

This entry aims to provide a comparative overview of environmental law and legislation in Asia. It should be noted, however, that the development of environmental legislation in Asia has been uneven and reflects the equally uneven socioeconomic development in the region. In some cases, differences may not be significant in written laws, but lie in the effectiveness of their implementation and enforcement. Due to a wide range of environmental issues and the number of countries in Asia, it is impossible to cover all of them. The entry spotlights four themes—biodiversity conservation, electronic waste (e-waste) management, environmental assessment, and climate change—because of their significance in the Asian context. The choice of countries as case studies aims to demonstrate the uneven development of environmental legislation. Thus, Japanese law serves as a reference of a well-developed body of environmental law, with the second-generation of legislation. China, Korea, Malaysia, Indonesia, Singapore, Hong Kong, Cambodia, Thailand, Laos, and Vietnam are at various stages of development ranging from the first-generation statutes, to development of second-generation environmental laws, and implementation and enforcement.

TABLE 2 Early Laws and Regulations on Environment in the Region

Country	Law or Regulation	Year
Cambodia	Environmental Protection and Natural Resource Management Law	1996
China	Environmental Protection Law	1978
Hong Kong	Water Pollution Control Ordinance	1980
Indonesia	Environmental Management Act No. 4	1982
Japan	Cabinet Directive	1972
Korea	Environmental Preservation Act	1977
Lao PDR	Lao PDR Constitution	1991
Mongolia	Environmental Protection Law	1996
Philippines	Environmental Policy Presidential Decree No. 1151	1977
Singapore	Environmental (Public Health) Act	1969
Thailand	Enhancement and Conservation of the National Environmental Quality Act	1992
Vietnam	Environmental Protection Law	1994

Source: Data from World Bank.[42]

Law on Biodiversity Conservation

Asia-Pacific is one of the richest regions in terms of biodiversity and 60% of the world's species are found in this region. As of 2008, the Asian and Pacific regions had the highest number of threatened species in any of the world's regions, almost one-third of all threatened plants and more than one-third of all threatened animal species.[13] The first generation of biodiversity law consisted mainly in establishing protected areas where human intervention is curtailed or prohibited. However, this exclusionary approach has caused notable socioeconomic impacts, in particular to local communities who used to live in the areas before the establishment of national parks. Therefore, more inclusive approaches have been proposed, such as community-based conservation (comanagement of a protected area by the government and local communities) and payments-based conservation (individuals are paid for their activities to conserve biodiversity). Biodiversity conservation capability is thus a function of a sound legislation in both international and national levels.

Recent Development in International Law and Its Potential Consequences on National Legislations

The international legal instruments, together with agreements that initiate the legislative framework of individual countries for biodiversity conservation, include The Convention on Biological Diversity (CBD) that entered into force on December 29, 1993, the Convention on Wetlands of International Importance Especially as Waterfowl Habitat (Ramsar 1971), the Convention Concerning the Protection of the World Cultural and Natural Heritage (Paris 1972), the Convention on International Trade in Endangered Species of Wild Fauna and Flora (Washington 1973), the Convention on the Conservation of Migratory Species of Wild Animals (Bonn 1979), etc.[14] However, an effective implementation for the international legal agreements is largely dependent on the actions taken by individual countries. Consequently, the most important of all is to establish the legal systems for biodiversity conservation at state level.

Very recently, on February 2, 2011, the Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization was opened for signature by Parties to the Convention of Biological Diversity. Many Asian countries are in the process of amending their legislation to implement the provisions under this new protocol. As most Asian countries are resource-providing countries, the Nagoya Protocol will allow better control of their genetic resources and ensure fair financial compensation among various stakeholders, including local communities.[15]

As for forestry, the new market-based mechanism of "Reducing Emission from Deforestation and Degradation (REDD)," which is under discussion in the UN Climate Change negotiations, should provide additional impetus to the improvement of forest law in Asian developing countries, particularly in the areas of monitoring, reporting, and data collection. Cambodia, Indonesia, Papua New Guinea, the Philippines, and Vietnam are among the first countries to receive support from the UN REDD Program.[16]

Development of Biodiversity Laws at National Level

At present, the biodiversity conservative laws of Asian countries like Japan, Thailand, Indonesia, and Laos contain more or less elements of public participation that characterize the second generation of legislations. Except for Japan, however, communities' rights are still very limited and difficult to enforce. The payment-based conservation approach has generally only been implemented through small-scale projects and not adopted by Asian legislators yet.

China. China covers an enormous land area of 9,600,000 km^2, including complex and varied geomorphology, climate, and natural conditions. That creates a country rich in ecosystems, which can be categorized into five types, namely forest, grassland, desert, inland wetland and other freshwater ecosystems, and ocean and coast. Due to a massive and distinct diversity of flora and fauna, China is regarded as one of the most important biodiversity countries. It is ranked among the top 10 nations in the world diversity of its mammal, bird, amphibian, and plant species. China has been considered to have one of the most important stocks of genetic diversity in the world. It is very important to protect and conserve this biodiversity for the national and international heritage.

China has promulgated a series of laws and regulations. The main domestic laws are Forest Law (1984), Grassland Law (1985), Fishery Law (1986), and Wild Animal Conservation Law (1988). Examples of regulations related to biodiversity include Reproduction and Conservation of Aquatic Resources (1979), Regulation on Forest and Wild Animal Nature Reserves Management (1985), Regulation on Forest Fire Prevention and Control (1988), Regulation on Seed Management (1989), Regulation on Conservation of Terrestrial Wild Animals (1992), Regulation on Nature Reserves (1994), etc.

Regarding the enforcement of the statutes, great progress for in situ and ex situ biodiversity conservation has been achieved. There are, however, still some gaps in the legislation. Based on the current status of conservation legislation in China and in accordance with the Convention on Biological Diversity, more attention should be paid to the conservation legislation for genetic resources, wild plant species, and various natural ecosystems.

Korea. Ecosystems in Korea comprise forest, mountain, freshwater, coastal and marine, and agriculture ecosystems. The total forest area covers 6.394 million hectares, estimated as about 64% of the country's land area. Forests are mainly coniferous, deciduous, and mixed forests. The variety of habitats creates a rich biodiversity of plants, animals, and other living organisms (fungi, protista, prokaryotes, etc.). Some species are considered to be extinct, such as the tiger and Siberian leopard, fox, wolf, and sitka deer. The decline of biodiversity in Korea is associated with its economic development. The main threats to biodiversity include overexploitation of land and biological resources, and environmental pollution.

Under the guiding principles of the Framework Act on Environmental Policy 1990 and the Constitution, the Natural Environment Conservation Act 1991 administered by the Ministry of Environment is Korea's basic law for biodiversity and nature protection. It defines categories of protected areas and provides for species and habitat protection. The law serves as a common framework for nature conservation and strengthens the provisions of other nature laws administered by government agencies. Several government agencies share the responsibility of conservation and sustainable use of biodiversity, in accordance with various laws. The Ministry of Environment is responsible for general biodiversity conservation under the Law of Natural Environment Conservation, Law of Wildlife Protection and Hunting, Law of Wetland Conservation, Law of Natural Parks, and Law of Ecosystem Conservation for Uninhabited Islands. The Ministry of Environment is also responsible for preventing inappropriate uses of natural resources through the Environmental Impact Assessment (EIA) process by the Law of Environmental Impact Assessment. The Forestry Administration, part of the Ministry of Agriculture and Forests, which manages forests under the Law of Forests.

Japan. The country is 3000 km long in the north–south direction, with a vertical range from coasts to mountains, with thousands of islands, and a geological history of intermittent connection to and separation from the continent, and various disturbances such as eruption of volcanoes, flooding of precipitous rivers, and typhoons. Those geological characteristics, together with four definite seasons due to the monsoon climate, create diverse habitats. It is such rich biodiversity that makes Japan one of the 34 biodiversity hotspots identified worldwide. A biodiversity hotspot refers to a region that is originally rich in biological diversity and endemic species but is now exposed to a serious threat of loss of such diversity. Besides the original local geohistorical and natural conditions, the tradition of wet-paddy rice agriculture and the rural lifestyle, which rely on a secondary natural environment known as "satochi-satoyama," or simply "satoyama," as well as the way the land has been used for agricultural purpose, have also contributed to the area's biodiversity richness. As in other countries, biodiversity loss in Japan is due to high economic growth with industrial development, and also natural disasters. It can, however, be said that the Japanese law on biodiversity is regarded as one of the best laws in the field. It goes beyond the protected area–based approach by mainstreaming biodiversity conservation into the daily life of the people. The law defines not only the responsibilities of national government and local governments but also those of businesses, citizens, and private bodies. The elaboration of the National Biodiversity Strategy and regional biodiversity strategies are mandatory. The 4th National Biodiversity Strategy, which was adopted in 2010, sets a long-term goal of 100 years, mid-and short-term targets for 2020 and 2050, and indicates about 720 measures with 35 targets.[17] The results of the implementation have to be reported to the Diet every year in the Annual Report on the State of Biodiversity. The law puts emphasis on preventive and adaptive approaches, including land use planning, research and technology development, EIA, and prevention of global warming. The role of the public is also highlighted with mandatory public consultation before formulation of policies and support of voluntary activities by businesses and citizens for the conservation of biodiversity.

Other relevant legislations are the Nature Conservation Law, Natural Parks Law, Law for the Promotion of Nature Restoration, Law for the Promotion of Biodiversity Conservation Activities, Law for the Conservation of Endangered Species of Wild Fauna and Flora, Wildlife Protection and Proper Hunting Law, Invasive Alien Species Act, and Law Concerning the Conservation and Sustainable Use of Biological Diversity through Regulations on the Use of Living Modified Organisms.

Malaysia. Malaysia has been identified as one of the world's mega-diversity areas with extremely rich biodiversity. Covering much of the country are the tropical forests, the oldest and most biologically diverse ecosystem on Earth. With ratification of the CBD, Malaysia is working toward incorporating into its national policies and planning a set of commitments under the treaty as well as setting the goal to become a world leader in conservation, research, and sustainable utilization of tropical biodiversity by 2020. To accomplish the ratification and goal, Malaysia has enacted a spectrum of legislation aimed at protecting biodiversity. Examples of law relevant to biodiversity conservation are as follows: Environment Quality Act 1974, Fisheries Act 1985, Pesticides Act 1974, Plant Quarantine Act 1976, Protection of Wildlife Act 1972, National Parks Act 1980, National Forestry Act 1984, Parks Enactment 1984, Forest Enactment 1992, Fauna Conservation Ordinance 1963, etc.[18,19]

Malaysia has enacted a number of laws and regulations to protect the nation's environment but, while adequate, there was no single overarching statute (or policy?) that relates to biodiversity conservation and management until the just approved National Policy on Biodiversity. Much of the present legislation is sector based. As stated above, the attainment of biodiversity conservation is significantly dependent on implementation and enforcement of the legislation. The effectiveness of legislation can be accomplished by the dedication of government agencies as well as public participation for accountability.[18,19]

Indonesia. Indonesia is a rich and diversified archipelagic nation. With the topographical characteristics of approximately 13,500 islands and extensive reef system, Indonesia has a wide range of natural habitats, with a wealth of fauna and flora, corals, fish, and other reefs; thus, it is recognized as a major world center for biodiversity. However, that substantial biodiversity is decreasing in the country owing to illegal clearing and deforestation, including large-scale burning in oil palm plantations and small-scale slash and burn for shifting farming, as well as illegal logging and trade in timber. Indonesia has one of the world's worst deforestation rates. In addition, illegal poaching, trade in protected species, and illegal and unsustainable fishing have threatened the country's biodiversity.[20,21]

Regarding the domestic legal framework, as the result of 1972 UN Stockholm Conference on the Environment, Indonesia promptly established of the Office of the State Minister for the Environment and enacted the Environmental Management Act (EMA) No. 4 of 1982 replaced in 1997 by the EMANo. 23 of 1997. This act and its implementing regulations are set in the broader context of the state policies passed every 5 years by the People's Consultative Assembly, which have since the early 1970s progressively entrenched the concepts of sustainable development and natural resources management. The EMA must also be read in the context of other natural resources management acts, such as the Forestry Act No. 41 of 1999, and the Fisheries Act No. 31 of 2004, and associated regulations and decrees transferred to the local governments' agencies. Laws and regulations related to biodiversity conservation are exampled as follows: Act on the Conservation of Biological Resources and their Ecosystems (Act No. 5 of 1990), Decree No. 1 of the Minister of Agriculture on the Conservation of the Riches of the Fish Resources of Indonesia, and Decree of the Ministry of Forestry No. 424/ KPTS-VI/1994 on the Guidelines on Crocodile Management in Indonesia, Fisheries Law (No. 9 of 1985). Similarly to other countries, law enforcement is still not an effective function. Indonesia has no specialized environmental law courts. Environmental cases are heard by the general and administrative courts as well as the Supreme Court on appeal.[21]

Furthermore, a multitude of biodiversity laws in Indonesia tend to be conflicting and uncoordinated. Also, in the forestry sector, the power to manage forests is shared between the central government and regional governments with no clear division of powers and responsibilities. It is believed that the Regional Autonomy Law, which entered into force in 1999, has been abusively used by local authorities to issue their own logging concessions, thus resulting in massive deforestation and forest fires.[22] The Indonesian case shows that a legislation that transfers power from center to periphery has to be carefully drafted to ensure that the power is put in the hands of local communities whose livelihood depends on the forests and biodiversity, and not in the hands of local officials known for corruption and nepotism. Some main legislations of Indonesia relating to biodiversity are the Conservation of Biodiversity and Ecosystems Law (1990) and the Basic Forestry Law (2000). These laws include provisions on participatory forestry planning, people's economic empowerment, partial transfer of authority to regional governments, and community-based forest monitoring.

Philippines. The Philippines is a tropical archipelago of 7100 islands located off the southeast coast of mainland Asia. It occupies a land area of 299,400 km^2, and territorial waters covering around 2,200,000 km^2, that create precious terrestrial and aquatic ecosystems and habitat types. As a consequence of rapid loss of biodiversity, as well as widespread destruction of the country's environment, a strong effort has been put into biodiversity conservation, including the formation of the multisector Philippine Council for Sustainable Development (1992), ratification of the Convention on Biological Diversity (1993), and preparation of the Philippine Biodiversity Assessment Report and the National Biodiversity Strategy and Action Plan (NBSAP) (1995–1997). The NBSAP proposes a wide range of strategies and actions, including information generation, in situ and ex situ conservation, legislative and policy development, institutional capability building, information, education and communication, and strengthened international cooperation.[23]

Regarding ecosystem and habitat conservation in the Philippines, it was innovated through the National Integrated Protected Areas System Act of 1992, a landmark piece of legislation that provides the framework for a decentralized, community-based reserve management strategy. The legislation relevant to biodiversity conservation include the following: Act No. 2590 (1916), An Act for the Protection of Game and Fish; RA 7586 National Integrated Protected Areas System Act of 1992; RA 7900, High-Value Crops Development Act of 1995; PD 1433, Plant Quarantine Decree of 1978; Proc. No. 926, Establishing Subic Watershed Forest Reserve; DAO 20, s 1996, Implementing Rules and Regulations on the Prospecting of Biological and Genetic Resources; DAO 24, s 1991, Shift in Logging from the Old Growth (Virgin) Forests to the Second Growth (Residual) Forests; DAO 20, s 1996, Implementing Rules and Regulations on the Prospecting of Biological and Genetic Resources; etc.[24]

Singapore. The Republic of Singapore, situated off the southern tip of the Malay Peninsula, comprises one major and more than 50 adjacent islands, with a total area of 648 km^2. The main island is separated from Malaysia by the narrow Johor Strait on the north, and from Indonesia's Riau Archipelago by the Singapore Strait on the south. It is a small country with an urbanized character. Forest land and coastal areas have been cleared to provide land for residential and commercial sites and other developments, resulting in a substantial depletion of flora/fauna and deterioration of natural habitats. At present, the environmental policy states that 5% of the land area should be set aside for nature reserves, national parks, catchment areas, bird sanctuaries, and gardens. Some of the environmental laws related to biodiversity conservation are Fisheries Act 1966, Wild Animals and Birds Act 198, Parks and Trees Act (for parks not gazetted as national parks) 1985, Endangered Species Act 1989, National Parks Act 1990, Animals and Birds Act (Revised) 2002.[25]

To accomplish the vision of *Singapore Today,* which is to be "A Garden City, A Haven for Biodiversity," the Singapore Green Plan (SGP) 2012 was established to provide the direction for protected area management in the next decade. One of the SGP objectives is to ensure the quality of the living environment, including the enhancement of the country's environmental heritage. SGP 2012 also recaps the state's commitment to maintain the 5% of land set aside for nature areas, and provides the direction.[26]

Hong Kong Special Administration Region. Like Singapore, Hong Kong SAR (or Hong Kong in short) experienced very rapid growth and development, to become urbanized and industrial in character. This would have jeopardized the biodiversity of the area, which is recognized as part of the natural heritage of Hong Kong. The most common causes of biodiversity loss in Hong Kong now are habitat destruction associated with infrastructure development, population growth, hillside fires, unsustainable exploitation of wild species, introduction of alien species, pollution, and global environmental change. Possible relevant factors contributing to such causes are lack of a comprehensive legal framework to protect areas of high conservation value and poor enforcement of existing conservation, environmental, and planning laws.[27]

Environmental protection in Hong Kong is constituted by 16 ordinances. Nonetheless, overall environmental laws in Hong Kong are still short of an unambiguous conservation objective. Only the 1995 Marine Parks Ordinance (Cap. 476) and the 1997 Protection of the Harbour Ordinance (Cap. 531) enclose evidently expressed conservation principles. Other legislations protecting flora and fauna are focused on conserving particular species; however, they fail to address the values and principles that lie beneath these objectives. Regarding the Country Parks Ordinance (Cap. 208) originally enacted in 1976, a revision is needed, "To provide for the designation, control and management of country parks and special areas for the purposes of conservation of biological diversity, countryside recreation and education." Additionally, the Environmental Impact Assessment Ordinance (EIAO) is the newest piece of legislation that seeks to protect the environment. The EIAO requires certain designated projects (generally major infrastructure projects) to undertake an EIA before they can be granted an environmental permit for development to proceed. A Technical Memorandum contains guidance on the criteria and guidelines to use for ecological, fisheries, landscape, and visual impact assessment.[27]

As well as domestic laws, Hong Kong is obligated to protect its natural and cultural heritage by international treaties, including the 1973 Convention on International Trade in Endangered Species of Wild Fauna and Flora, the 1979 Convention on the Conservation of Migratory Species of Wild Animals, the 1971 (Ramsar) Convention on Wetlands of International Importance, and the 1972 Convention for the Protection of World Cultural and Natural Heritage. Seemingly, application of the 1992 Convention on Biodiversity in Hong Kong remains limited as there is no sign that the treaty will be formally applied to accomplish the CBD's biodiversity objectives after the government's endorsement.[27,28]

Thailand. Thailand has a total land area of 513,115 km^2, lying in a hot and humid climatic zone in the middle of Southeast Asia. Much of Thailand is situated in the Mekong River basin, and it is one of the Greater Mekong Subregion (GMS) countries. Thailand also has an extensive coastline. With such a location, it is enriched with biodiversity associated with terrestrial and aquatic (both freshwater and marine) ecosystems. It also covers agricultural ecosystems (about one-fifth of the country), which include biodiversity components of rice, farm crops, and livestock. Similarly to other countries, the major threat to biodiversity is human disturbance through overexploitation of natural resources/ habitat, illegal logging and trading of animals, overhunting of wildlife, deforestation, urban expansion and pollution, etc. Such disturbances cause an adverse reduction of bio-diversity.[29]

Biodiversity in Thailand is safeguarded by a number of laws and regulations. Some of the important ones are the National Park Act 1961, National Forest Reserve Act 1964, Wild Animal Reservation and Protection Act 1992, Plant Quarantine Act of 1964 and Plant Quarantine Act (second issue) 1994, Animal Species Maintenance Act 1966, Importing and Exporting of Goods Act 1979, Enhancement and Conservation of National Environmental Quality Act 1992, and Plant Varieties Protection Act 1999.[30]

In addition, Section 66 Paragraph 1 of the Thai Constitution of 2007 guarantees the right of an individual and communities to participate in the preservation and exploitation of natural resources and biological diversity. However, out of 35 pieces of legislation relating to natural resources, biodiversity and environment, only 4 provides for public participation, namely, the National Promotion and Conservation of Environment Quality Act, B.E. 2535 (1992); the Private Irrigation Act, B.E. 2482 (1939); the Plant Varieties Protection Act, B.E. 2542 (1999); and the Protection and Promotion of the Thai Traditional Medicine Act, B.E. 2542 (1999). Moreover, the rights given to individuals are still very limited, and the authorities retain most of the control in natural resources management.

Lao People's Democratic Republic. Lao PDR, situated in the Mekong River basin, is counted as one of the GMS countries. For Southeast Asia, Lao PDR is one of the countries with a large proportion of land covered with undisturbed forest. It covers a land area of 236,800 km^2, where the topography is largely mountainous, with elevations above 180 m typically characterized by steep terrain, narrow river valleys, and low agricultural potential. Lao PDR is rich with natural resources such as forestry, minerals, and hydroelectric power. Due to the country's still abundant forestry resources, it thus generates a prosperous biodiversity.[31,32] Similarly to other GMS countries, Lao PDR has enacted laws for biodiversity conservation, but the enforcement is ineffective. Illegal logging and wildlife hunting have been frequently found. Mining and hydropower developments might also cause loss of biodiversity.[31,32]

In Lao PDR, biodiversity conservation is provided for in the Forest Law of 1996, a 1993 decree that designated the first national biodiversity conservation areas, logging ban, decree on (PM Decree No. 67/PM, 1991) protecting trees by logging, decision on adoption of (PM Decree No. 66/PM, 1991) forest conservation, and seed material import regulation (quarantine regulation) that controls importation of plant material.[32]

Cambodia. Cambodia is a tropical country in mainland Southeast Asia with a territory of 181,035 km^2. It is adjacent to the Gulf of Thailand on the south and shares borders with Thailand (west and north), Laos (north), and Vietnam (east). Cambodia is in the Mekong River basin. The country is dominated by the Mekong River, known as the Tonle Thom or "great river," and the Tonle Sap or "fresh water lake." The Mekong River basin is one of the most biodiverse regions in the world. As situated in the Mekong River Basin, Cambodia has an abundance of terrestrial and aquatic ecosystems that are significant habitat for plants and aquatic organism, fishes in particular. Such ecosystems are invaluable resources for economic development and human well-being.[33,34]

To protect and sustain the country's biodiversity, Cambodia enacted the framework for the Law on Environmental Protection and Natural Resources Management in 1996. Subsequently, the Ministry of Environment was created in 1998, which manages natural resources along with the Ministry of Water Resources and Meteorology and the Ministry of Land Use Management, Urbanization, and Construction. Cambodia has continued to enact more environmental and conservation laws. Examples of the relevant laws are as follows: Environmental Protection and Natural Resources Law 1996, Law on Commune

Administration (part of the decentralization process) 2001, Land Law 2001, Forestry Law 2002, and Wildlife Law 2002. Important subdecrees include Subdecree on Concession Management (moratorium on logging and log transport), Subdecree on Community Forestry, Subdecree on Environmental Impact Assessment, Subdecree on Industrial Agricultural Concessions, Subdecree on Social Concessions, and Royal Decree on Protected Areas 1993.

Cambodia is party to a number of important international conventions of which those stated here are relevant to biodiversity conservation. The Ramsar Convention on Wetlands has been ratified, and Cambodia has identified three wetland sites for recognition: BoengChhmar in the Tonle Sap floodplain, KohKapik on the coast, and a portion of the middle Mekong river north of Stung Treng. In addition, the Tonle Sap is recognized as a UNESCO Biosphere Reserve. Cambodia ratified the Convention on Biological Diversity in 1995. In 1997, the government prepared a biodiversity prospectus and in 2002 completed a National Biodiversity Strategy and Action plan.[19,33]

Vietnam. Vietnam is situated at the eastern side of a peninsula that protrudes into the Eastern Sea, which is a bay of the Pacific Ocean. Within the Mekong River Basin, Vietnam is at the most downstream length of the Mekong River. Due to the country occupied with land, river, and sea, Vietnam is enriched with a variety of ecosystems, including tropical rainforests and monsoon savannah, marine life, and mountainous subalpine scrubland. Additionally, a specific feature of Vietnam is its length of more than a thousand miles from north to south, with a width of only 30 miles from east to west at its narrowest point, thus generating an abundance of natural resource along the coast and sea, and a richness of biodiversity. The threats tobiodiversity are mainly due to transformation of forest and wetland areas to other uses, infrastructure construction, urbanization, industrialization, and environmental pollution.

Vietnam's environmental law is based on its constitution. The Law on Environmental Protection was initially promulgated in 1993; subsequently, Vietnam has enacted a variety of laws and decrees on conservation issues. These laws affect, directly or indirectly, the conservation of biodiversity. Examples of the direct laws are the Decree on the Conservation and Development of Wetlands, the Decree on Protection of the Environment (which details rare and precious flora and fauna), and a related decree that determines methods for regulating their protection and management. The indirect laws include decrees regulating wastewater, controls on businesses creating environmental damage, the 2003 Land Law, etc. Regarding the protected areas, Vietnam has two laws concerning the establishment and management of protected areas. The two statutes are the Law on Forest Protection and Development of December 2004 (revised from the 1991 original), and the Law on Biodiversity that became effective starting July 2009. The Law on Forest Protection and Development provides the guiding principles for the development and use of special-use forests, while the Law on Biodiversity focuses on protected area concerns such as categorization and decentralization of protected area management.

Challenges to Effective Implementation and Enforcement of Biodiversity Law

Evidently, the countries presented here have enough laws and regulations to protect their environment; however, much of the present legislation is sector based. Moreover, the achievement of biodiversity conservation is not a function of the number of laws and regulations but of the implementation and enforcement of such legislation. The effectiveness of legislation can be accomplished by the dedication of government agencies as well as public participation for accountability.[19]

As partly demonstrated through the case studies mentioned above, challenges facing Asian governments in the implementation and enforcement of biodiversity law lie in corruption, weak institutional capacity, lack of reliable data and budget, high demand for alternative land use, and traditional beliefs.

With regard to weak institutional capacity, in most Asian countries, forestry staffs are not adequately trained and are underpaid. To address this shortcoming, governments such as China and Vietnam have raised the pay and living conditions for forest personnel.[19] Inter-agency coordination is another serious institutional problem. There are usually many agencies involved in forest management with overlapping areas of responsibilities or, on the contrary, areas where no agency is in charge.

Economic development has put pressure on forest protection. The construction of hydropower dams and roads, and agricultural activities have often encroached on protected lands. In an effort to reconcile forest protection and economic development, Lao PDR has piloted hydropower levies that support the management of protected areas affected.[34]

Forests and wildlife statistics are often out of date, inaccurate, and incomplete. The remoteness of the forest areas and the reluctance of officials to report forest crimes (as they may be seen as a sign of failed forest management on their part) are among factors that contribute to the lack of reliable data and information sharing among relevant agencies. The use of satellite imagery has helped improve the data collection system to a certain extent.

Traditional beliefs may sometimes come in the way of forest and wildlife protection. In many Asian countries such as China, Vietnam, Thailand, Laos, and Cambodia, the consumption of rare wildlife, such as bear's paw and monkey brains, is seen as status symbols or medicines.

Regional Cooperation

Regional cooperation has played an important role in promoting sustainable forest management. Examples of regional initiatives are the East Asia Forest Law Enforcement and Governance process (EAFLEG) and the Asia Forest Partnership (AFP). Both the EAFLEG and the AFP aim to bring together various stakeholders in forest management, including governments, non-governmental organizations (NGOs), and the private sector. They serve as an informal forum for information sharing, dialogue, and joint action. Within the Southeast Asian Nations Association (ASEAN), the ASEAN Senior Officials on Forestry (ASOF) was entrusted with the task of policy coordination and decision making on regional cooperation in the forest sector.

Law on E-Waste Management

The quantity of e-wastes in Asia has exploded in recent years with the exponential growth in the use of electronic equipment (computers, mobile phones, televisions, refrigerators, etc.) coupled with the consumers' behavior of regularly replacing their devices in order to stay up-to-date with the latest technology. E-wastes often end up in incinerators and/or landfills. Toxic substances such as mercury and lead that are commonly used in electronic products can contaminate the environment, including land, water, and air. E-waste is commonly characterized as hazardous waste. As of July 2008, among 46 countries in the Asia-Pacific region, there are 32 countries that have ratified the Basel Convention on the Control of Transboundary Movements of Hazardous Wastes and Their Disposal and some of them have also ratified the Ban amendment.[35] However, there are some countries in Asia that still lack regulations for controlling hazardous wastes including e-waste. An effective regulatory framework on the management of such wastes is urgently needed. A new legal concept created to deal with e-waste management is the Producer Extended Responsibility (PER) where producers are held liable for the costs of managing their products to end of life of the product. In this way, producers are encouraged to design environmentally friendly products to reduce disposal costs. Only a few countries, such as Japan, Korea, and Taiwan, have a well-established legislation on e-waste management based on the concept of PER and many years of experience in implementation. Some Asian countries such as China have recently enacted a law on the matter; however, the effectiveness of the implementation remains to be seen. Other countries have already passed similar laws, but the laws have yet to enter into force (Thailand and India) or are still at the drafting stage. An adaptation period is usually granted for businesses before the entry into force of the law. Other actors in a product's life cycle such as distributors, repair and customer service providers, consumers, and recyclers may also be required to bear some of the treatment and disposal costs. E-waste discussed herein includes used and waste electrical and electronic equipment (UEEE/WEEE).

China. The Ministry of Environment and Protection is responsible for regulating and controlling e-waste management. At the end of February 2006, China promulgated the law entitled "Administration on the Control of Pollution Caused by Electronic Information Products," which is simply called as China RoHS in the industry.[36] Import of WEEE has taken place since 2002 under the existing laws, including Law of the People's Republic of China on Prevention of Environmental Pollution Caused by Solid Waste, Interim Provisions on Administration of Environmental Protection on Import of Waste and its supplementary provisions, List of Wastes Prohibited against Import (Notice No. 25, 2002), and Catalogue of Solid Waste Forbidden to Import in China (Announcement No. 11, 2008). Regarding UEEE, it is allowed except used TVs; UEEE requires 3C certification. Applicable laws for UEEE are as follows: Administrative Method on Inspection and Supervision of Imported Used Mechanical and Electrical Products, Measures for Administration of Import of Specified Used Mechanical and Electrical Products (Order No. 5, 2008), and Catalogue of Import of Specified Used Mechanical and Electrical Products (Announcement No. 37, 2008). A new law on the management of e-waste has established higher standards for recycling processes and allows only certified recyclers to engage in the e-waste recycling business. To support recyclers to improve their equipment and facilities, a centralized mandatory fund has been established with contributions from domestic producers and sellers of imported electronic devices. The law also places responsibility on manufacturers, distributors, repair and consumer service providers, and recycling companiesto collect and responsibly handle e-waste. One shortcoming of the law, however, is that the scope of their responsibility and the penalizing measures for non-compliance remains vague.[35]

Hong Kong, China. The regulation of e-waste management is under the responsibility of the Environment Protection Department. Hong Kong has begun its waste import and export control through the Waste Disposal Ordinance in 1996. For the purpose of import, WEEE and UEEE (that is classified as WEEE) have been controlled through a permit system in accordance with guidelines on "import and export of hazardous waste including electrical and electronic appliances containing hazardous constituents or components."[35]

Republic of Korea. The Resource Recirculation Policy Division, Ministry of Environment, is responsible for regulating and controlling e-waste management. Specific regulations applicable to UEEE and WEEE do not currently exist; however, in general, e-waste management falls under the "Waste Control Act" of December 1986 and later amendments. Import control of WEEE is performed through application for a license from the Ministry of Environment in accordance with the Act on the Control of Trans-Boundary Movement of Hazardous Wastes and Their Disposal (Basel Convention) and Act on Resource Recycling of Electrical and Electronic Equipment and Vehicles. Import of UEEE is allowed; no specific law is applicable.[35]

Japan. The Ministry of Environment is responsible for regulating e-waste management. Since 2001, Japan has enforced the Fundamental Law for Establishing a Sound Material-Cycle Society to promote comprehensively and systematically the policies for realizing a Sound Material-Cycle Society, providing an umbrella framework for the relevant waste management laws of the country. Regarding the import and export control of WEEE, the Law for the Control of Export, Import, and Others of Specified Hazardous Wastes and Other Wastes was entered into force in 1993. This law stipulates the necessary import/export procedures of hazardous waste to comply with the requirements of the Basel Convention. The Waste Management and Public Cleansing Law of 1970 was amended in 1993 to regulate import and export of waste. None of these laws is applicable for the management of UEEE. With the specific purpose to complement to the Basic Act on Establishing a Sound Material-Cycle Society (2000), the following two specific e-waste recycling legislations have been established: the Law for the Promotion of Effective Utilization of Resources (LPUR), and the Law for Recycling of Specified Home Appliances (LRHA) (1991). The LPUR applies to used computers and small-sized secondary batteries and encourages manufacturers' voluntary efforts to take part in collection and recycling. Recycling costs are borne by both manufacturers and consumers. The LRHA is a stricter regulation. It covers television sets, refrigerators, washing machines, and air conditioners. Manufacturers and retailers have an obligation to take back used products and recycle them. Consumers are required to pay for the cost of transportation and recycling. The LRHA also sets up a procedure that allows tracking a product from the beginning until the end of its life cycle.[35]

Malaysia. The Department of Environment, Ministry of Natural Resources and Environment, is responsible for regulating e-waste management. Under the Environmental Quality Act, 1974, several regulations for the control of scheduled wastes (hazardous wastes) management in Malaysia have been enacted. The principal regulation on e-waste management is the Environmental Quality (Schedule Waste) Regulation of 2005, enforced by Department of Environment, and in which specific categories of e-waste are defined and coded. The Guidelines for the Classification of Used Electrical and Electronic Equipment entered into force in January 2008, which prohibits the import of WEEE and export for the purpose of disposal. Waste generators are allowed to export waste for recycling, recovery, or treatment provided prior written consent are obtained from the importing state. The Ministry of Local Government and Housing has the jurisdiction over households and business entities/institutions and has enacted the Solid Waste Management and Public Cleansing Act of 2007. The Royal Malaysian Customs enforces transboundary movements of hazardous waste under the Customs Act 1967, Customs (Prohibition of Import) Order 2008, and Customs (Prohibition of Export) Order 2008.[35]

Indonesia. The Environmental Impact Management Agency, Ministry of Environment, is responsible for regulating e-waste management; however, neither specific criteria on e-waste nor specific regulations on e-waste management have been established. The existing laws have been employed for WEEE and UEEE. WEEE is only allowed for export, but prohibited for import to Indonesia under Act No. 23 of 1997 on Environmental Management, Articles 20 and 21, Presidential Decree No. 61/1993 Basel Convention Ratification, and Ministerial Decree No. 231/ MPP/Kp/07/1997 Regarding Import Procedure of Waste. Import of UEEE and e-waste for direct consumption by consumers is prohibited under Decree No. 756/MPP/ Kep/12/2003 on Import of Non-New Capital Goods and Decree No. 610/MPP/Kep/10/2004 Regarding Amendment of No. 756/MPP/Kep/12/2003.[35]

Philippines. The Department of Environment and Natural Resources (DENR) is responsible for regulating e-waste management. Import of WEEE and UEEE requires permit. Laws applicable to WEEE are the Toxic Substances and Hazardous and Nuclear Wastes Control Act of 1990 (Republic Act No. 6969), DENR Administrative Order 200–6 (Implementing Rules and Regulations for RA 6969), DENR Administrative Order 1994-28 (Interim Guidelines for the Importation of Recyclable Materials Containing Hazardous Substances), DENR Administrative Order 1997-28 (Amending Annex A of DAO 1994-28), and DENR Administrative Order 2004–27 (Amending Annex A of DAO 1994-28). The law applicable to UEEE is DENR Administrative Order 1994-28 (Interim Guidelines for the Importation of Recyclable Materials Containing Hazardous Substances). DAO-94-28 allows the import of electronic assemblies and scrap with the condition that residuals from recycling of materials that contain hazardous substances without any acceptable method of disposal in the Philippines must be shipped back.[35]

Singapore. Export, import, or transit waste requires a permit from the Pollution Control Department of Singapore in accordance with the Hazardous Waste (Control of Export, Import, and Transit) Act. Import/export of UEEE are allowed if there are documents to support that the appliances for import/export are in working condition and suitable for reuse. Export of UEEE that are not suitable for reuse is prohibited. Import of UEEE for the purpose of dismantling and re-export of the dismantled components are prohibited.[35]

Thailand. The Ministry of Natural Resources and Environment and the Ministry of Industry are the administrative authorities for hazardous waste and e-waste management. Both UEEE and WEEE are controlled under the Hazardous Substance Act B.E. 2535 (AD 1992) in Thailand. UEEE can be imported only under a subordinate law for import control of UEEE. Import of UEEE is allowed only for reuse, repair/refurnish back to its original purposes, disassembly and recycle/recovery under certain conditions. Thirty-two UEEE items require import permits from the Ministry of Industry. WEEE can be imported and exported under a subordinate law and following Basel Convention procedures. The e-waste management act will only enter into force in 2014 and will seek to regulate the entire life cycle of an electronic product. At the beginning of its life cycle, importers and producers will be taxed, and the tax money collected will be used for e-waste management. At the end of the product's life cycle, the law will promote construction of an integrated waste management facility at production sites and other areas throughout the country.[35]

Lao People's Democratic Republic. Lao PDR does not have legislation discretely mentioning e-wastes; however, e-waste is considered hazardous waste by defining that hazardous and toxic wastes include batteries, old paint cans, aerosol, and other refuse. Such wastes are mixed with municipal solid wastes that are disposed at landfills. Accordingly, there are no specific laws or regulations directed to e-wastes. For solid waste disposal itself, a decree on waste management is planned in connection with the finalization of the revision of the Environmental Protection Law of 1999. Nonetheless, some laws have implications for solid waste management in Lao PDR, for example, the Environmental Protection Law (Article 22), the Decree on Implementing the Environmental Protection Law (Article 9.4), and in addition some of the provinces and the Capital City of Vientiane have issued specific regulations on urban environmental management including solid waste management. Thus, Lao PDR is encountering solid waste problems due to several reasons, as follows: inadequate legal framework, ambiguous institutional responsibilities and lack allocation of responsibilities on solid waste management to specific institutions, insufficient budget allocation to carry out functions in accordance with the law, etc.[37]

Cambodia. At present, Cambodia does not produce EEE products. Regulation on e-wastes is the responsibility of the Ministry of Environment. Specific law and/or regulation to properly manage, recycle, and dispose EEW does not yet exist. Cambodia, however, uses the following existing laws for WEEE management: Subdecree on Solid Waste Management and the Inter-Ministerial Declaration on SWM in Cities and Provinces. Import of WEEE to Cambodia is banned, while import of both new EEE and UEEE is allowed for domestic consumption. Export of household waste and hazardous waste from Cambodia requires approval from the Ministry of Environment, export license from the Ministry of Trade, and permit from the import country (BCRC China, 2009).

Vietnam. Regulating e-waste management is the responsibility of Hazardous Waste Management Division, Waste Management and Environment Promotion Agency, Vietnam Environment Administration, Ministry of Natural Resources and Environment. In January 2006, Vietnam promulgated the Implementation Rules for the Law on Trade (No. 12/2006/ND CP), which bans import of waste materials (both WEEE and UEEE), toxic chemical substances, and second-hand commodities, including electronic, cooling, and home appliances. Other applicable laws are the Regulation of Management of Hazardous Waste (Decision No. 155/1999/QD-TTg), and Decision No. 23/2006/QD-BTNMT on the List of Hazardous Waste. In Circular No. 12/2006/TT-BTNMT, export of hazardous waste shall follow Basel Convention procedures.[35]

Transboundary Movements of Hazardous Waste

Since the entry into force of the Basel Convention on the Control of Transboundary Movements of Hazardous Wastes in 1992, most Asian countries have acceded to the Convention and enacted national laws banning import of waste into their territories. However, despite the legal prohibition, China and India continue to be the world's largest waste dumping yards. One of the reasons is that there are loopholes in the laws that provide for some exceptions to the ban. For example, in India, the law allows for imports of secondhand computers and laptops if they are intended for donations to educational institutions or NGOs. Due to several arrests for e-waste smuggling recently, the government is considering a complete ban on e-waste.[38] In China, some recyclable wastes such as poly-silicon and artificial fiber continue to be allowed into the country. To prevent the smuggling of prohibited waste, all solid waste imports are required to undergo electronic inspections.[39]

Challenges to Effective Implementation and Enforcement of E-Waste Management Law

Barriers to effective implementation and enforcement of e-waste management laws, like other pollution-control laws, are as follows: 1) lack of specific legal frameworks; 2) complexity of institutional arrangements; and 3) lack of technological, financial, and human resources.

Lack of specific legal frameworks. As the above-mentioned case studies show, many Asian countries have yet to enact a specific e-waste management law.

Complexity of institutional arrangements. Agencies involved in pollution control are often numerous with overlapping areas of responsibilities as well as gray areas where the institutional responsibilities are unclear or lacking. Lack of coordination is not only an issue between ministries but also between central governments and local authorities, as well as between the administration and the industrial sector.

Lack of technological, financial, and human resources. E-waste management requires considerable investments in planning, staff training, purchasing technology, and building new facilities. The laws need to provide funding or economic/fiscal incentives to manufacturers and local authorities in order to allow them to comply with environmental standards. Or else, many would prefer to pay a little extra money to officials who would turn a blind eye on their polluting practices. How to raise sufficient resources for pollution control remains a big challenge for Asian governments. Thus far, taxation is the most commonly adopted method in resource raising (property taxes, sewerage charges, and vehicle taxes, for instance). However, taxation is never popular. International assistance by means of finance, technology transfer, and capacity building is also essential to help improve the capacity of developing countries in addressing pollution problems.[40]

Law on Environmental Assessment

Environmental assessment considered herein includes EIA and strategic environmental assessment (SEA). EIA and SEA are related methods with the purpose to prevent, mitigate, and compensate adverse environmental impacts that may be caused by a proposed activity. EIA focuses on projects such as construction of a dam, industry, mining, etc., whereas SEA is applied to policies, plans, programs, and macro projects (in some countries) such as a transportation development plan, energy development plan, international airport project in Hong Kong, etc. EIA regulations were established in most Asian countries in the 1980s and 1990s. In terms of legislation, they are varied ranging from none (Myanmar), to very recent and not widely applied legislation (Laos and Cambodia), to moderately to highly applied legislation (Thailand, Malaysia, the Philippines, Indonesia), and to extensively applied EIA regulation within a broader planning framework (Japan, Hong Kong, South Korea, China).[41,42] Accordingly, EIA has been practiced throughout Asia with varying degrees of rigor and effectiveness. However, SEA is a relatively new method and, thus far, few Asian countries have incorporated SEA into their legislation.[42]

In 2006, the World Bank released a report called "Environmental Impact Assessment Regulations and Strategic Environmental Assessment Requirements: Practices and Lessons Learned in East and Southeast Asia." The report divides countries in the region into three categories: 1) Hong Kong SAR, Japan, and Korea, which have the most advanced EIA/SEA legislations and effective implementation; 2) China, the Philippines, Indonesia, and Thailand, which have less strict EIA/SEA legislations and have encountered difficulties in their implementation; and 3) Vietnam, Mongolia, Lao PDR, and Cambodia, which have recently established EIA/SEA legislations and are at an early stage of implementation. Some countries are exampled for discussion herein with the content is mostly drawn from such mentioned report.

China. The EIA system has been in place since 1979, which was the year that China enacted the Environmental Protection Law. This law contained broad elements requiring EIA, particularly for construction projects. Later in 1986, the first legal document on EIA in China was issued by the

Ordinance of Environmental Protection of Construction Projection (1986). A series of regulations on construction projects were issued, including Environmental Protection Procedures for Construction Project (SEPA, 1990), Regulation of Environmental Protection of Construction Projects (State Council No. 253, 1998), and Environmental Management Catalogue for Construction Projects (SEPA, 1999). To broaden environmental assessment, the current EIA law has been modified and extended to cover plans, and has become SEA inclusive. A new law on EIA was approved by the National People's Congress in 2002, and has functioned since September 1, 2003. The new EIA law incorporates the concept of SEA for plans and programs, but not for policies. Subsequently, the EIA law covers two large areas: plans and construction projects. Plans are divided into two categories: 1) plans for land use, regional, watershed, and offshore development; and 2) "specific plans," which include agriculture, industry, livestock breeding, forestry, natural resources, cities, energy, transportation, tourism, etc. However, enforcement in legislation, public participation, and capacity building should be undertaken for applications of policy-and plan-based SEA.[42]

Hong Kong SAR. To solve environmental problems encountered in the country, the Hong Kong government has put forth effort to mitigate, control, and prevent such problems. Environmental assessment is considered an important tool in preventing environmental pollution. It is applied not only to individual projects but also to strategic policies and proposal that facilitate the country moving toward sustainability. The EIA process has been applied to projects since 1986, to plans since 1988, and to strategies and policies since 1992. Hong Kong's EIAO was enacted in 1997 in order to formalize the 15 years experience with EIA, environmental monitoring, and auditing, and came into force in April 1, 1998. The EIA in Hong Kong is considered SEA inclusive. With a successful application of EIA/SEA tools with proven records in legal provision, technical capacity, training and implementation, Hong Kong has become one of the most transparent EIA systems in the world.[42]

Japan. The EIA concept was introduced in Japan in the 1960s and implemented through various administrative guidelines, sector legislation (such as the Public Water Area Reclamation Law), and ordinances and guidelines issued by local authorities. The unified law called "the Environmental Impact Assessment Law" was finally adopted in 1997 and took effect in 1999. The law adopts a listing method by scale to identify projects for which environmental impact statement (EIS) is required.[42] Legal requirements that make the Japanese EIA system more strict and comprehensive than those of many other countries in the region are, for instance, EIA requirements for small-scale projects with potential adverse impacts on the environment and emphasis on public participation. Therefore, public opinion is requested at both the scoping stage and the EIA conduction stage, and a period of 100 days is provided for public hearings and information display before submission of EIA report.

Although Japan has yet to make SEA into law, there is a strong political will at both the national and regional level to integrate SEA in the policy-making process. The Ministry of Environment and other ministries have adopted SEA guidelines such as the Ministry of Environment's preliminary guideline on SEA in the formulation of municipal waste management plans. Local governments are also active in implementing SEA.

Republic of Korea. The development of the EIA system was initiated in 1977 through the Environmental Preservation Act, and put into effect by the legislation of "Regulations on the Preparation of EIA" enacted in February 1981. After the Environmental Administration was upgraded to the ministerial level in 1990, the previous Environmental Preservation Act was divided into a number of separate laws. One of those is the Basic Environmental Policy Act enacted in August 1990. The Environmental Impact Assessment Act was enacted as a separate law on June 11, 1993, and was put into effect on December 12, 1993. To increase efficiency of the system, the EIA Act was further revised in 1997, and became the EIA law of 1997. The current EIA system is considered as SEA exclusive. Nonetheless, an SEA type of system was applied in the late 1990s known as the Prior Environmental Review System (PERS), which is mainly implemented for various plans and programs. The current PERS has been amended as an SEA type in general, but not to cover policy level. Thus, EIA in Korea includes two types, the PERS conducted at a planning stage and EIA carried out at the project-development stage. In this system, a decision on whether to execute a development project will be made at the planning stage, taking into account environmental concerns.[42]

Malaysia. The Environmental Quality Act (EQA) was enacted in 1974 as the major federal environmental statute. It was not until 1987 that EIA procedures were introduced under the EQA as a control preventative mechanism. The EIA is well established in Malaysia under the responsibility of the federal government.[43] The situation is currently changing by delegation of EIA powers to the state level. The states of Sarawak and Sabah have adopted independent impact assessment procedures for natural resource management, and it is possible that other states may follow. However, the EIA at state level has faced some problems due to insufficiency of skilled staffs, low institutional capacity, and an absence of effective monitoring of mitigation measures.[43] Therefore, there is a need to strengthen the state capability on EIA implementation. EIA in Malaysia is considered SEA exclusive. However, there is now evidence of an up-and-coming commitment to SEA in the country. Government objectives in environmental protection and management are moving forward, although the regulatory framework to achieve these objectives is not, as yet, fully developed.[44]

Some major infrastructure projects, such as roads and power facility development, could be subject to SEA procedures in the future. Integration of SEA into policy, plan, and programs is necessary to secure a more environmentally sustainable development in the country.

Indonesia. The development of the EIA system was initiated by Government Regulation No. 29 (1986) in compilation with the provisions of Article 15 of the former Environmental Management Act No. 4/1982. Later, Government Regulation No. 51 (1993) concerning EIA imposed significant revisions to the assessment system. Currently, Regulation No. 27/1999 is a revision of EIA regulation No. 51/1993. The new regulation is expected to be improved and provide a more democratic basis. The EIA system is the responsibility of the Environmental Impact Management Agency. The EIA in Indonesia is project based and SEA exclusive. However, the government has realized the importance of SEA in the decision making process, but its application is not compulsory. The Ministry of Environmental published a book on Strategic Environmental Assessment that provides the fundamentals, procedures, and benefits of applying SEA in the policy, plan, and program process.[42]

Philippines. Originally, the EIS system, which is equivalent to the EIA system, was conceived in Philippines Environmental Policy (P, D, No 1151). The actual establishment of the EIA system began with Presidential Decree (P.D.) No. 1586 in 1978. After issuance of some respective decrees, the EIS was adopted in the document *DAO No. 30 of 2003 Implementing Rules and Regulations (IRR) for the Philippine Environmental Impact Statement (EIS) System,* which was issued in 2003. The EIS system is the responsibility of the DENR. The EIS system is well established in the Philippines, including a legal mandate, administration, procedure, and guidelines. It is regarded as extremely comprehensive and perhaps entails the most stringent requirements in the whole Southeast Asia region.[45] The current EIA system in the Philippines is still project based and SEA exclusive. However, some SEA initiatives have been undertaken. In DAO 30/2003, it was stated that, “The EMB shall study the potential application of EIA to policy-based undertakings as a further step toward integrating and streamlining the EIS system” (Article II, Section 7). The SEA covering policy and plan are being considered to be contained in a new EIA Act.[46]

Singapore. Implementation of EIA in Singapore has been operated through the Environmental Pollution Control Act 2000 and the Land Planning process. Regarding EIA through Pollution Control, an EIA may be required for particular projects specified by the Ministry of Environment and Water Resource that have potential to cause pollution affecting public health; for example, petrochemical works, gasworks, and refuse–incineration plants; foreign investment projects using or storing large quantities of hazardous substances; etc. This is likely a project-based EIA. By EIA through Land Planning, Singapore established a document called the *Concept Plan,* which broadly outlines land-use policies in the country, of which the policies are translated into detailed proposals for local areas called “Development Guide Plans” (DGPs). The basic environmental concerns are considered in the DGPs. This is an SEA-like approach for spatial planning. With the provision of relatively effective laws and efficient centralized planning mechanisms, the lack of an EIA law does not appear to obstruct environmental management endeavors. Not only the comprehensive planning and effective pollution control mechanisms but also, more importantly, a stringent enforcement system makes it possible for Singapore to move toward sustainable development.[42]

Thailand. The EIA system was established in Thailand through the Improvement and Conservation of the National Environmental Act (1975), followed by the Enhancement and Conservation of Environmental Quality Act (1992). Although the EIA system is well established in Thailand with a good number of qualified personnel, the law remains vague on many issues and public hearing is optional for some projects. The recent EIA procedures for projects identified as having potential adverse impacts to natural resources, environment, and health legally require public involvement through a strict and specific process. EIA for such listed projects is named as Environmental Health Impact Assessment.[47,48] Regarding SEA implementation by law, Thailand has not made SEA mandatory yet. However, SEA has been performed for some projects specified by line agencies for certain areas of interest, such as SEA for Economic Zone at Border Territory at Chiang Rai Province by the Office of Natural Resource and Environment Policy and Planning (2005).[49] SEA legislation is now under development. At this moment, it can be stated that EIA in Thailand is project based and SEA exclusive.

Lao PDR. Lao PDR enacted the regulations on EIA back in 2000, and these regulations were revised and upgraded to decree with the Prime Minister Decree on Environmental Impact Assessment No. 112 of 2010. The Ministry of Natural Resource and Environment is responsible for administrating the EIA system, approving EIAs, and for issuing Environmental Compliance Certificates Related legislation on social impacts, including the Decree on the Compensation and Resettlement of Development Projects, 192/PM of 2005, and the Regulations for Implementing Decree 192/PM on Compensation and Resettlement of People Affected by Development Projects.

Since the EIA system of Lao PDR is quite recent, it sets relatively high legal standards and requirements, particularly focusing on social aspects. EIA reports in Lao PDR are called Environmental and Social Impact Assessment. Despite a relatively good legal basis, many barriers to effective implementation persist in Lao PDR, particularly the lack of qualified professionals in EIA. Thus, many capacity building programs for EIA personnel have been funded by international organizations and donor countries most prominently the long-term development cooperation with Sweden (Strengthening Environmental Management Phase I and II) and the assistance from the Government of Finland (Environmental Management Support Programme). At this time, EIA is legally implemented as project based and SEA exclusive; however, through the above-mentioned support, the Ministry of Natural Resources and Environment is revising the Environmental Protection Law and developing a decree and guidelines on SEA. This is combined with comprehensive capacity building and case studies on SEA.[42]

Cambodia. In 1996, the National Assembly of Cambodia enacted the Law on Environmental Protection and Natural Resource Management (EPNRM) as a framework law governing environmental protection and natural resources management. The law requires the Royal Government to prepare national and regional environmental plans. In addition, there are subdecrees concerning a wide range of environmental issues, including EIAs, pollution prevention and control, public participation, and access to information (SIDA). Cambodia has subsequently established an EIA system under the EPNRM law through the EIA Subdecree on Environmental Impact Assessment issued in 1999, which mandates general requirements, procedures, and responsibilities. The subdecree instructs the Ministry of Environment to formulate implementing rules and guidelines.[50] This EIA system covers only projects and is SEA exclusive.

Vietnam. EIA was first mentioned in the Law on Environmental Protection (LEP) 1994 Article 18, which stipulates that organizations and individuals must submit EIA reports to be appraised by the state management agency for environmental protection. The *Government Decree on Providing Guidance for the Implementation of the Law on Environmental Protection* (Government Decree No. 175/CP, 1994) is an important legal document on EIA in Vietnam.

The implementation of the EIA system is the responsibility of the Ministry of Natural Resources and Environment. The current EIA system in Vietnam is basically consistent with international practice.[51] SEA has already been adopted conceptually in the Vietnamese legislative framework, for example, in the LEP, GD 175/CP, and Circular No. 490/TT-BKHCNMT, where "EIA not only must be carried out at project level, but also for master plans for development of regions, sectors, provinces, cities and industrial zones." There are several cases of applying SEA in Vietnam in recent years. As plans are covered by the EIA system, the system is conceptually SEA inclusive. The government is considering accommodating SEA in the new environmental legislation.

Conclusion

Development of environmental law in the region began about four decades ago. However, it has been uneven depending on the economic development of each country. According to the level of development of environmental legislation and effectiveness of the implementation, Asian countries may be divided into three categories: 1) countries with advanced economy, comprehensive environmental legislations, capable and well-coordinated institutional framework, and relatively effective implementation, such as Japan, Korea, Hong Kong, and Singapore; 2) countries with a developing economy, relatively well-established legislations but ineffective implementation due to problems such as institutional complexity and lack of qualified personnel, such as China, Malaysia, Indonesia, the Philippines, and Thailand; 3) countries that have recently emerged from conflict and are rebuilding, and therefore are at an early stage of developing environmental legislations and institutions, such as Laos, Cambodia, and Vietnam.

Environmental legislation in Asia is incorporating more and more elements of the "second generation of environmental legislation," which are public participation, compliance incentives, and market-based mechanism. However, regarding the role of the people, in many cases, the law recognizes rights of local communities in natural resource management but does not provide for procedure or institutional framework that would enable the effective exercise of such rights. In addition, environmental legislations in Asia are struggling to keep pace with the rise of the emerging environmental problems and the level of environmental degradation. The problem of e-waste is one good example. For environmental assessment including EIA and SEA, all the countries presented herein have enacted with Environmental Protection Law that include EIA. However, SEA implemented by law is limited. The EIA systems of most countries by law are project based and SEA exclusive. Only some countries such as China, Hong Kong, and Vietnam have environmental assessment system by law for EIA with SEA inclusion.

To ensure effective implementation, a good environmental legislation must be complemented with a capable and coordinated institutional framework, qualified personnel, and an adequate budgetary allocation. Regional and international cooperation are of utmost importance in the development and implementation of environmental legislations in Asia, particularly in the areas of capacity building, technical cooperation, and funding.

Acknowledgments

This entry could not have been completed without the great help of Miss JinjutaManotham, Second Secretary, Ministry of Foreign Affairs, Thailand, for searching some meaningful environmental legislation-related topics; Mr. Brian J D'Arcy, Environmental Expert, Scotland, U.K.; Mr. Peter Gammelgaard Jensen, Chief Technical Advisor, Grontmij, Denmark, for editing the manuscript; and Miss SriratSuwannakom, my research assistant, for formatting the document. I would like to thank the Research Center for Environmental and Hazardous Substance Management, KhonKaen University, National Excellence Center for Environmental and Hazardous Substance Management for supporting me on the relevant research and academic services. I would also like to thank the *Encyclopedia of Environmental Management,* Taylor & Francis Group, for giving me the opportunity to write this entry.

References

1. Kim, M.; Hodges, H.J. Is the 21st century an “Asian Century”? Raising more reservations than hopes. Pac. Focus **2010**, *25* (2), 161–180.
2. Kastner, S.L. The global implications of China’s rise. Int. Stud. Rev. **2008**, *10* (4), 786–794.
3. Browne, A. Why China is the REAL master of the universe, available at http://www.dailymail.co.uk/news/article-559133/Why-China-REAL-master-universe.html (accessed January 2012).
4. UNESCAP. *Asia Pacific Disaster Report 2010*; Protecting Development Gains, Information and Communications Technology and Disaster Risk Reduction Division: Bangkok, Thailand, 2010.
5. Nguyen, L.D. *Environmental Indicators for ASEAN: Developing an Integrated Framework*; UNU-IAS Working Paper No. 109; United Nations University Institute of Advanced Studies: Yokohama, Japan, 2004.
6. UNESCAP. *State of the Environment in Asia and the Pacific 2000, ST/ESCAP/2087;* United Nations: New York, 2000.
7. King, P.; Annandale, D.; Bailey, J. A conceptual framework for integrated economic and environmental planning in Asia—A literature review. J. Environ. Assess. Policy Manag. **2000**, *2* (3), 297–315.
8. Hezri, A.A.; Hasan, M.N. Towards sustainable development? The evolution of environmental policy in Malaysia. Nat. Resour. Forum **2006**, *30* (1), 37–50.
9. Chapter 2.3: National Legislation, Module 2: Policy and Legislation, Mekong River Commision’s Watershed Management Resource Kit. Available at http://www2.gtz.de/snrd/wmrk/2PolLeg/203_National_Legislation.pdf
10. Arnold, C.A. Fourth-Generation Environmental Law: Integrationist and Multimodal. Express O. Available at http://works.bepress.com/anthony_arnold/1.
11. Kheng-Lian, K. The first generation of environmental laws in Asia. KohKheng-Lian. In *Towards a “Second Generation” in Environmental Laws in the Asian and Pacific Region: Select Trends;* Lin-Heng, L., Manguait, M., Socorro, Z., Eds.; IUCN: Gland, Switzerland, 2003; 15–25.
12. Arnold, C.A. Fourth-generation environmental law: Integrationist and multimodal, http://works.bepress.com/anthony_arnold/1 (accessed May 2011).
13. ESCAP; ADB; UNEP. *Preview Green Growth, Resources and Resilience Environmental sustainability in Asia and the Pacific;* ST/ESCAP/2582,2010; United Nations: Bangkok, Thailand, 2010; 5.
14. Sustainable Hunting Project. *Regional Action Plan for Moving Toward Sustainable Hunting and Conservation of Migratory Birds in Mediterranean Third Countries;* Bird-life International: Cambridge, U.K., 2007.
15. Secretariat of the Convention on Biological Diversity. The Nagoya Protocol on Access and Benefit-Sharing of Benefits Arising from their Utilization to the Convention on Biological Diversity: text and annex, Convention on Biological Diversity; United Nations, Quebec, Canada, 2011.
16. Meridian Institute. *Reducing Emissions from Deforestation and Forest Degradation (REDD): An Options Assessment Report*; Prepared for the Government of Norway; Angelsen, A., Boucher, D., Brown, S., Merckx, V., Streck C., Zarin, D., Eds.; Meridian Institute: Washington, D.C., 2009.
17. O’Neill, H.; Wandel, A. *Future Policy Award 2010: Celebrating the World’s Best Biodiversity Policies;* World Future Council: Hamburg, Germany, 2010.
18. UNCTAD. Latiff, A. and A.H. Zakri (UniversitiKebang-saan Malaysia). Protection of Traditional Knowledge, Innovations and Practices: The Malaysian Experience (paper presented at the UNCTAD Expert Meeting on Systems and National Experiences for Protecting Traditional Knowledge, Innovations and Practices). 30 October –1 November 2000. Available at http://r0.unctad.org/trade_env/docs/malaysia.pdf (accessed May 2012).
19. Nagle, J.C. The effectiveness of biodiversity law. J. Land Use Environ. Law **2009**, *24,* 203.

20. Maxim, S.; Hadad, I.; Sitorus, S. Biodiversity conservation in Indonesia: The case of KEHATI, available at http://www.synergos.org/knowledge/03/asiafinancingkehati.htm (accessed June 2011).
21. International Development Law Organization. Strengthening Environmental Law Compliance in Indonesia, Towards Improved Environmental Stringency and Environmental Performance. Development Law Update 2006,6, available at http://www.idlo.int/publications/30.pdf (accessed June 2011).
22. Tan, A.K. Environmental laws and institutions in Southeast Asia: A review of recent developments. SYBIL **2004**, *VIII,* 177–192.
23. Meniado, A.P.; Garcia, J.L.; Madamba, E.J. The Philippines. In *Biodiversity Planning in Asia. A Review of National Biodiversity Strategies and Action Plans*; Carew-Reid, J., Ed.; IUCN: Gland, Switzerland, 2002; 216–236.
24. Senga, R.G. Establishing protected areas in the Philippines: Emerging trends, challenges and prospects. George Wright Forum **2001**, *18* (1), 56–65.
25. Bugna, S.C. A profile of the protected area system in Singapore. Asian Biodivers. **2002**, *2* (2): 30–33.
26. National Parks Board. *Conserving Our Biodiversity: Singapore's National Biodiversity Strategy and Action Plan*; National Parks Board: Singapore, 2009.
27. Wan, J.; Telesetsky, A. *Creating Opportunities: Saving Hong Kong's Natural Heritage*; Civic Exchange: Hong Kong, China, 2002, 1–53.
28. Sharma, C. Enforcement mechanisms for endangered species protection in Hong Kong: A legal perspective. VJEL **2003–2004**, *5* (1), 1–34.
29. Office of Environmental Policy and Planning. *National Report on the Implementation of Convention on Biological Diversity;* Ministry of Science, Technology and Environment: Bangkok, Thailand, 2002; 1–60.
30. Office of Natural Resources and Environmental Policy and Planning. *Thailand: National Report on the Implementation of the Convention on Biological Diversity*; Ministry of Natural Resources and Environment: Bangkok, Thailand, 2009; 1–76.
31. Country Report on the State of Plant Genetic Resources for Food and Agriculture, Lao PDR, available at http://www.fao.org/docrep/013/i1500e/Lao%20Peoples%20Democratic%20Republic.pdf (accessed August 2011).
32. Clarke, J.E. Protected area management planning. Oryx **2000**, *34* (2), 85–89.
33. FAA 118/119 Analysis, Conservation of Tropical Forests and Biological Diversity in Cambodia, available at http://www.oired.vt.edu/sanremcrsp/documents/team-room/usaid-info/USAID-Cambodia-Forest-and-Biodiversity-Report.pdf (accessed July-August 2011).
34. ICEM. Lessons learned in Cambodia, Lao PDR, Thailand and Vietnam: Review of protected areas and development in the Lower Mekong River Region, Indooroopilly, Queensland, Australia. ICEM **2003**, 1–104.
35. Basel Convention Coordinating Center for Asia and the Pacific. *Report of the Project on the Import/Export Management of E-waste and Used EEE, Asia-Pacific Regional Centre for Hazardous Waste Management Training and Technology Transfer;* Department of Environmental Science and Engineering, Tsinghua University: Beijing, China, 2009.
36. Kirschner, M. RoHS in China, available at http://www.conformity.com/A0725 (accessed July 2011).
37. Wittmaier, M.; Langer, S.; Wolff, S.; Bilitewski, B.; Werner, P.; Stefan, C.; Schingnitz, D.; Wiesmeth, H.; Parthasarathy, R.; Wooldridge, C.; Green, J.; Quynh, D.N.; Viet, L.H.; Ngan, N.V.C.; Hoang, N.X.; Trang, N.T.D.; Minh, P.H.; Touch, V.; Samell, K.; Khouangv-ichit, S.; Daladone, P.; Tia, S.; Songkasiri, W.; Commins, T.: Framework conditions for waste management in Lao PDR, Vietnam, Cambodia and Thailand (Ch. 2). In *INVENT—Innovative Education Modules and Tools for the Environmental Sector, Particularly in Integrated Waste Management, Part I Curricula and Modules*; Handbook/e-book; University of Applied Sciences: Bremen, Germany, 2009; 45–87.

38. Thakur, P. Govt may ban import of e-waste, *Times of India,* September 6, 2010.
39. Hurst, C. *China's Rare Earth Elements Industry: What Can the West Learn?*; Institution of the Analysis of Global Security (IAGS): Potomac, MD, 2010; 1–42.
40. UNEP. *Asia-Pacific Environmental Outlook* 2; United Nations Environment Programme, Regional Resource Centre for Asia and the Pacific: Pathumthani, Thailand, 2001.
41. Li, C.J. *Environmental Impact Assessments in Developing Countries: An Opportunity for Greater Environmental Security?* Working Paper No. 4; USAID and Foundation for Environmental Security and Sustainability (FESS): Falls Church, VA, 2008.
42. World Bank. *Environmental Impact Assessment Regulations and Strategic Environmental Assessment Requirements Practices and Lessons Learned in East and Southeast Asia, Safeguard Dissemination Note No.* 2; Environment and Social Development East Asia and Pacific Region: Washington, D.C., 2006.
43. Briffett, C.; Obbard, J.P.; Mackee, J. Towards SEA for the developing nations of Asia. Environ. Impact Assess. Rev. **2003**, *23* (2), 171–196.
44. ADB. Country assistance plan (2000–2002) Malaysia. Asian Development Bank, 1999. Cited by Briffett, C.; Ob-bard, J.P.; Mackee, J. Towards SEA for the developing nations of Asia. Environ. Impact Assess. Rev. **2003**, *23* (2), 171–196.
45. Tan, A.K.J. APCEL Report: Environmental Law (ASEAN-10), Faculty of Law, National University of Singapore, 2000. Cited by World Bank. *Environmental Impact Assessment Regulations and Strategic Environmental Assessment Requirements Practices and Lessons Learned in East and Southeast Asia, Safeguard Dissemination Note No. 2*; Environment and Social Development East Asia and Pacific Region: Washington, D.C., 2006.
46. Villaluz, M.G. Advancing the EIA system in the Philippines, UNEP EIA Training Source Manual, 2003. Cited by World Bank. *Environmental Impact Assessment Regulations and Strategic Environmental Assessment Requirements Practices and Lessons Learned in East and Southeast Asia, Safeguard Dissemination Note No. 2*; Environment and Social Development East Asia and Pacific Region: Washington, D.C., 2006.
47. Section 4. Constitution of the Kingdom of Thailand, B.E. 2550 (2007), Government Gazette, Vol. 124, Part 27a, dated 24th August B.E. 2550 (2007).
48. The Enhancement and Conservation of National Environmental Quality Act B.E. 2535 (NEQA 1992). Government Gazette, Vol. 109, Part 37, dated 4th April B.E. 2550 (2007); Government Gazette, Vol. 119, Part 102, dated 8th October B.E. 2550 (2007).
49. ONEP. *SEA in the Border Area in Chiang Rai Province, Final Report*; Office of Natural Resources and Environmental Policy and Planning: Bangkok, Thailand, 2005.
50. Government of Cambodia. *Cambodia's Report to WSSD—National Assessment of Implementation of Agenda 21—Progress, Challenges and Directions*; Government of Cambodia: Phnom Penh, Cambodia, 2002.
51. Obbard, J.P.; Lai, Y.C.; Briffett C. Environmental assessment in Vietnam: Theory and practice. J. Environ. Assess. Policy Manag. **2002**, *4* (3), 267–295.

Bibliography

1. Aoki-Suzuki, C. *Trade of Second Hand Electrical and Electronic Equipment (SH-EEE) in Asia: Focusing on Actors in Reuse Markets and the Need for Deepened Actor Analysis and Integrated Sustainability Assessment,* ISIE Asia-Pacific Meeting and ISIE MFA-Con Account Meeting, Tokyo, Nov 7–9, 2010; Institute for Global Environmental Strategies: Tokyo, Japan, 2010.
2. Asian Development Bank (ADB). *Emerging Asia: Changes and Challenges;* ADB: Manila, Philippines, 1997.
3. Brandon, C.; Ramesh, R. *Toward an Environmental Strategy for Asia*, World Bank Discussion Paper No. 224; World Bank: Washington, D.C., 1993.

4. Bruch, C.; Mrema, E. *UNEP Guidelines and Manual on Compliance with and Enforcement of Multilateral Environmental Agreements*, Proceedings of the Seventh International Conference on Environmental Compliance and Enforcement, Marrakech, Morocco, Apr 9–15, 2005; Gerardu, J., Jones, D., Markowitz, K., Zaelke, D., Eds.; International Network for Environmental Compliance and Enforcement: Washington, D.C., 2005.
5. Chongrak, P.; Indra, G. Asian environmental status: Emerging issues and future scenarios. In *Environmental Management Tools. A Training Manual*.Routray, J.K., Mohanty, A., Eds.; UNEP: Asian Institute of Technology, Pathumthani, Thailand, 2006.
6. Clive, B. Environmental Impact Assessment in Southeast Asia: Fact and Friction. Geo J. **1999**, *49* (3), 333–338.
7. Daniel, E.; Marie, P. *Globalization and The Environment in Asia, United States–Asia Environment Partnership Framing Paper;* USAID Development Experience Clearinghouse: Washington, D.C., 1999; 1–53.
8. Department for Economic and Social Information and Policy Analysis. *Glossary of Environment Statistics,* Studies in Methods, Series F, No. 67; United Nations: New York, 1997; 1–83.
9. Desai, B.H. Multilateral environmental agreements: Legal status of the secretariats. J. Environ. Law **2010**, *23* (1), 155–157.
10. Evans, P.J. Industry and the environment in Asia. TDRI Q. Rev. **1998**, *13* (3), 9–27.
11. Available at http://ewasteguide.info/stepping-efforts-con (accessed July 2011).
12. Available at http://lawprofessors.typepad.com/environmental_law/2007/01/what_is_environ.html (accessed July 2011).
13. Available at http://timesofindia.indiatimes.com/india/Govt-may-ban-import-of-e-waste/articleshow/6501864.cms (accessed August 2011).
14. Available at http://www.cecphils.org/node/55 (accessed August 2011).
15. Available at http://www.chinaenvironmentallaw.com/wp-content/uploads/2009/03/regulations-on-waste-electric-and-electronic-products-chn-eng.pdf (accessed September 2011).
16. Available at http://www.env.go.jp/en/laws/recycle/10.pdf (accessed September 2011).
17. Available at http://www.ide.go.jp/English/Publish/Down-load/Spot/pdf/30/007.pdf (accessed August 2011).
18. Available at http://www.indiaenvironmentportal.org.in/files/DraftE-waste-Rules303.10.pdf (accessed September 2011).
19. Available at http://www2.kankyo.metro.tokyo.jp/anmc21_WM/legislation.htm (accessed September 2011).
20. Iqbal, M.T. Environmental law and multilateral environmental agreements (MEAs). In *Environmental Management Tools, A Training Manual*; Routray, J.K., Mohanty, A., Eds.; School of Environment, Resources and Development, Asian Institute of Technology: Pathumthani, Thailand, 2006; 75–82.
21. Jones, S. Highlights of waste control laws and regulations in China. In *A China Environmental Health Project Fact Sheet*; China Environment Forum's Partnership with Western Kentucky University on the USAID-Supported China Environmental Health Project: Bowling Green, KY, 2007.
22. Kaniaru, D.; Kurukulasuriya, L. UNEP's role in capacity building in environmental law. In *International Capacity Building*, Fourth International Conference on Environmental Compliance and Enforcement. Chiang Mai, Thailand, Apr 22–26, 1996; International Network for Environmental Compliance and Enforcement: Washington, D.C., 1996.
23. Lee, S.; Na, S. E-Waste recycling systems and sound circulative economies in East Asia: A comparative analysis of systems in Japan, South Korea, China and Taiwan. Sustainability **2010**, *2*, 1632–1644.
24. Memon, A. Devolution of environmental regulation: EIA in Malaysia. In *Case Studies for Developing Country*; UNEP EIA Training Resource Manual 2002. Sadler, B., Fuller, K., Ridgway, B., McCabe, M., Baily, J., Saunders, R., Eds.; UNEP: Geneva, Switzerland, 2002; 45–61.

25. Millennium Ecosystem Assessment. *Ecosystems and Human Well-Being: Synthesis*; Island Press: Washington, D.C., 2005.
26. Ministry of the Environment. *Environmental Impact Assessment in Japan*; Environmental Policy Bureau, Ministry of the Environment, Government of Japan: Tokyo, Japan. Available at http://www.env.go.jp/en/policy/assess/pamph.pdf
27. O'Connor, D. *Grow Now/Clean Later, or Pursuit of Sustainable Development?;* Research Programme on Economic Opening, Technology Diffusion, Skills and Earnings, Working Paper No. 111; OECD Development Centre: Paris, France, 1996.
28. Pescott, M.J.; Durst, P.B. Leslie, R.N. Forest law enforcement and governance: Progress in Asia and the Pacific; RAP PUBLICATION 2010/05; FAO: Bangkok, Thailand, 2010; 1–205.
29. Peter, H.A. Asian cultural influences on environmental legal norms: RodaMushkat, International Environmental Law and Asian Values, Toronto, UBC Press. Rev. Québécoise Droit **2004**, *17* (1), 283–286.
30. Pulhin, J.M. Trends in forest policy in the Philippines policy. In *Trend Report 2002;* Inoue, M., Ed.; The Institute for Global Environmental Strategies (IGES); Forest Conservation Project; Soubun Printing Co. Ltd: Tokyo, Japan, 2003; 29–41.
31. Richardson, B.J.; Wood, S., Eds. *Environmental Law for Sustainability;* Hart Publishing: Oxford, U.K., 2006; 1–18.
32. Schluep, M.; Hagelueken, C.; Kuehr, R.; Magalini, F.; Maurer, C.; Meskers, C.; Mueller, E.; Wang, F. *Recycling—From E-Waste to Resources;* Final Report, Sustainable Innovation and Technology Transfer Industrial Sector Studies; UNEP and UNU: Paris, France, 2009.
33. Suzuki, K. Sustainable and environmentally sound land use in rural areas with special attention to land degradation, APFED Third Substantive Meeting (APFED3), Guilin, People's Republic of China, Jan 23, 2003; Asia-Pacific Forum for Environment and Development: APFED3/EM/03/Doc.4.
34. UNEP. *Global Environment Outlook 2000;* Earthscan Publications Ltd.: London, 1999.
35. UNEP. *Global Environment Outlook-3;* Earthscan Publications Ltd.: London, 2002.
36. World Bank. *World Development Indicators 2001;* Development Data Cebter, The World Bank: Washington, D.C., USA, 2001.
37. Yang, H.; Innes, R. Economic incentives and residential waste management in Taiwan: An empirical investigation, environmental and resource economics. Eur. Assoc. Environ. Resour. Econ. **2007**, *37* (3), 489–519.

V

ELE: Focuses on the Use of Legislation or Policy to Address Environmental Problems

29

Environmental Policy

Sanford V. Berg

Introduction

When economic activity leads to pollution and over-use of common property resources, government intervention can improve social welfare. Pollution involves a market failure in which damages caused by a producer or consumer are imposed on third parties. These damages can involve personal health, the physical deterioration of buildings, and foregone options for the future. Of course, if transaction costs are low, those that are causing pollution damage can be taken to court if the liability rules are clear. Destruction of common property resources such as losing unique ecological habitats, endangering particular species, or destroying valued scenic vistas is another form of market failure affecting the environment. Because there may not be clear property rights to such elements of the environment, these common property resources can be overutilized. Given the lack of well-defined property rights, government enacts environmental laws to address these market failures. However, identifying and quantifying the damage caused by pollution sources or inflicted upon sensitive ecosystems can be difficult, making any determination of the benefits and costs a contentious exercise. Consequently, choosing policies that define the extent of the environmental protection can be both contentious and problematic.

The next section describes the multidisciplinary inputs that are incorporated into environmental policy analysis, selection, and implementation. Other topics addressed here include policy impacts, the burden of proof, economic evaluation, and the strengths and limitations of policy options.

Multidisciplinary Approach to Developing Environmental Policy

Environmental economics is the study of how economic and environmental issues interact. Issues addressed by environmental economists include but are not limited to evaluating ways to reduce pollution, analyzing the trade-offs between using renewable and nonrenewable resources, or estimating monetary values for ecosystems or habitat. While no single field of study contains all the insights needed to develop and implement sound environmental policies, the focus here will be on economics because it provides a system for incorporating many perspectives and it is the framework by which environmental policy is designed and evaluated. Depending on the burden of proof, the resulting policies might be excessively stringent (costly relative to their benefits) or inadequate for the protection and preservation of environmental features that affect human health and welfare and have intrinsic value.

We know from materials balance that human activity does not create matter but only changes its form, concentration, and location, thus there is a need for physical sciences such as chemistry, physics, and biology to help inform environmental policy. While all societies affect natural systems, the scale of potential impacts has grown with economic development. There is evidence that as incomes rise, citizens are willing to devote relatively more resources to controlling environmental impacts. Moreover, many citizens would like to see much more attention given to reducing current damages and limiting the risks for future harm, hence an understanding of societal and political dynamics is also important for informing environmental policy.

The development and implementation of sound environmental policy draws upon information and procedures from many fields of study. Here, economics is utilized as the framework for integrating the concepts, measurements, and values required for the steps:

1. Determine appropriate regulatory objectives (through citizen participation in political processes and community consensus-building)
2. Balance those objectives to determine regulatory priorities
3. Identify and legislate oversight responsibilities for environmental agencies
4. Develop (a) mechanisms for monitoring environmental impacts (such as ambient air and water quality) and (b) methodologies for integrating new scientific understandings of environmental impacts into the policy prioritization process
5. Define the appropriate targets for different types of pollutants and the protection of biodiversity
6. Determine (and then apply) the appropriate policies for meeting objectives
7. Analyze environmental indictors on a regular basis, checking for noncompliance
8. Evaluate the impacts, recognizing potential biases in the measures and the ways impacts are valued
9. Establish an effective process for monitoring and reviewing the framework, including the penalties and sanctions applied when there is noncompliance

These steps require input from a number of disciplines that shape the way we see things. Although technical training allows analysts to delve deeply into subjects in a consistent manner, awareness of other disciplines' perspectives can be important for constructive environmental policy-making, including:

- Engineers look to technology for solutions to environmental problems. They are able to incorporate new (often expensive) control technologies into energy extraction, production, consumption (energy efficiency), and pollutant disposal and storage (as with nuclear waste).
- Meteorologists and hydrologists analyze pollution transport in air and water systems. They have a deep understanding of the impacts of discharges under different conditions. In conjunction with demographers and epidemiologists, they can estimate the doses received by different population groups.
- Medical scientists and toxicologists analyze the dose response relationships for citizen health, conducting exposure and risk assessments.

- Ecologists study the impacts of pollutants on the local and global environment, assess the value of ecosystem services, and track invasive species and biodiversity. Climate scientists help assess the causes and consequences of changes in local and global temperatures and other weather patterns.
- Materials scientists look at damages caused by air and water pollution. The associated impacts include cleaning and painting buildings, treatment costs, and shorter life spans for affected equipment.
- Political scientists focus on issues of power, legitimacy, social cohesion, and the roles of different stakeholder groups in influencing environmental policies. Consensus is critical because ultimately, in a democratic system, there needs to be widespread agreement on the desired outcomes if the system is to avoid instability.
- Economists emphasize the importance of efficiency in resource allocation. They apply benefit-cost analysis and tend to depend on price signals to provide incentives for the adoption of appropriate control technologies and conservation measures.
- Planners deal with land-use and zoning issues, given population growth projections. Planners integrate legal constraints with historical experience, bringing topological, aesthetic, and geographical elements to the analysis.
- Archeologists and anthropologists provide insights on the impacts of dams, mines, and their related economic activities on unique historical sites, local populations, and indigenous groups. Such impacts create difficult valuation issues.[1,2]
- Lawyers spotlight the institutions of policy implementation. For example, rules and regulations attempt to pay significant attention to procedural fairness. Due process contributes to the legitimacy of outcomes. If different parties perceive that there is no transparency and no opportunity for participation, environmental policy will be perceived as unreasonable and the laws will either be changed or they will be disobeyed in a variety of ways.
- Environmentalists advocate sustainability and environmental equity. The by-products of energy production affect public health and have environmental outcomes. Those impacts have economic value, but often that value is nonmonetary or difficult to quantify. For example, generation and transmission siting decisions incorporate impacts on biodiversity and sustainability.
- Ethicists help society understand personal values and notions of stewardship. Humans have a clear responsibility to leave future generations with a legacy of sound institutions and a clean environment, though the best means to this end are often not obvious.

Thus, physical, biological, and social scientists attempt to uncover patterns and identify lessons to help us improve policy. Given the complexity of environmental issues, most environmental problems are managed, not solved.

Impacts

Energy production and consumption impact people and the environment in a number of ways. For example, activities can damage ecosystems in the extraction phase (oil drilling or coal mining) or involve cross-media emissions in the consumption phase that can lead to further ecosystem damage. Emissions can be from a single point or a mobile source. In addition, they can be continuous or intermittent (with exposure and impacts depending on wind and other weather conditions or the presence of other chemicals). The transport mechanism can be complicated and involve multiple jurisdictions (as with SO_2 and NOx—emissions lead to "acid rain" or ozone problems in downwind areas).

Air

Issues range from local concentrations of particulate matter in the atmosphere to concerns over anthropogenic climate change. Consequences for health, ecosystems, agriculture, coastal settlements, species survival, and other impacts make atmospheric change a serious policy issue. For example, long-range transport means pollutants cross national boundaries and require coordination. Other pollutants—such as greenhouse gas emissions of CO_2—require coordination not due to transport, but because the effects are global in nature regardless of where emissions occur.

Water

Effects of contaminants vary in surface waters and groundwater. The United States has primary standards to protect public health (with maximum contamination levels [MCLs] for toxic wastes). Secondary standards and associated MCLs are meant to protect public welfare (for example, ensuring that the taste, odor, and appearance of groundwater do not result in persons discontinuing water use). Other environmental issues include species loss and dealing with nonindigenous, invasive species.

Land Use

Siting is an issue for electricity generators, transmission lines, and distribution systems (other aspects of land use include urban sprawl and availability of land for agriculture. The focus here is on the environmental impact of energy systems. For example, social investments in mass transit affect emissions from mobile sources (autos). However, environmental policy addresses many other issues, such as the use of pesticides and fertilizers by agriculture or deforestation). The problem of not in my back yard (NIMBY) is universal: we like the convenience of electricity but do not want its production or transport to affect our own property. Surface coalmines are an eyesore, but restoration can be costly. Hydroelectric dams can affect fisheries, flood unique canyons (causing a loss of scenic vistas), damage ecosystems (as in the Amazon), or displace human populations (as with China's Three Gorges Project). Solar collection stations and wind generators require space and have impacts on aesthetics. For some, viewing large windmills along the crest of a lovely mountain range is an eyesore. For others, the same scene is a symbol of hope.

Environmental policy-makers must be aware of the relationship between changes in impacts in one medium and changes in impacts in other media. For example, reducing airborne emissions of mercury will also lead to reduced mercury concentrations in rivers and lakes. However, it may also be the case that reducing ozone precursors from auto emissions by using methyl tert-butyl ether (MTBE) leads to increasing harm to bodies of water as the MTBE precipitates out in rain. Finally, there may be policy and impact trade-offs that must be evaluated with reducing CO_2 emissions through a greater use of nuclear energy. The policy reduces greenhouse gases but raises issues and associated risks of waste storage and protection. In all cases, the links between different environmental media and different environmental policy must be understood for society to properly evaluate the trade-offs.

Burden of Proof

Because environmental issues tend to be complex, delays in responding to citizen concerns and new scientific information can lead to negative impacts or a local crisis. What is more problematic: erring on the side of environmental protection or erring on the side of development? When science is unclear or when studies yield conflicting outcomes, the issue of burden of proof arises. Two types of errors are possible. In a Type I error, a hypothesis is rejected when it is in fact true (e.g., deciding that a pollutant causes no health damages when in fact it does). Rejecting the hypothesis of a health link would lead to more emissions (and citizen exposure) than otherwise would be the case.

A Type II error occurs when the decision maker fails to reject a hypothesis that is in fact false (e.g., not rejecting the hypothesis that low doses of a pollutant have no damaging side effects for certain types of citizens, such as asthmatics, who are viewed as potentially sensitive to a particular pollutant). If in fact at low doses the pollutant does not have negative health impacts, environmental regulators might have imposed standards that induced costly compliance strategies that were based on the Type II error.

Dose-response models that do not reject linear functions when the actual relationships are non-linear would fall into this category.

Both types of errors have costs. However, the political implications may depend on the type of error, leading decision-makers to prefer making errors that are difficult to detect. Thus, it can be argued that environmental regulators will tend to avoid making Type I errors. When evidence accumulates and shows conclusively that a pollutant has health impacts, those responsible for environmental policy do not want to be blamed for acting too slowly. Furthermore, citizens might prefer excessive caution (labeled a "precautionary bias"). On the other hand, Type II errors can result in regulators imposing high abatement costs onto polluters (and those purchasing associated products) in a manner that is not cost effective.

A related issue is whether or not the environmental impact is irreversible. If it is not reversible, a case can be made that the burden of proof should be assigned to those who assert that relatively higher levels of pollution are not problematic. On the other hand, if abatement costs are systematically underestimated and the benefits of pollution reduction are overestimated, it is possible to devote excessive resources to limiting environmental impacts.

Economic Framework

Economists are aware that it is difficult to place monetary values on many impacts of pollution but argue that environmental amenities must be balanced against other valued goods and services.[3] Some view economists as overemphasizing the efficacy of market incentives to the exclusion of other instruments. However, because economics offers a consistent framework for integrating insights from other fields, it will be described here.

Cost-Benefit Analysis (CBA)

The most fundamental economic analysis looks at how pollution impacts (reflected in "external costs") cause excessive consumption of polluting goods in the absence of government intervention. These external costs are the negative spillover effects of production or consumption for which no compensation is paid (e.g., a polluted stream that damages the health of those living along the stream). Producers consider the environment to be a free input; hence they only minimize private costs. If these external costs are added to the private costs (reflected in the supply curve), this is the total social cost.

Figure 1 shows how a competitive product market yields an equilibrium price ($4) and quantity (80 units per week). However, in the absence of public intervention, the price only reflects the private costs of production, not damages imposed on others (amounting to $2 when the 80th unit is produced, but this is assumed to be less if only 65 units of the good are produced. The external costs are higher at

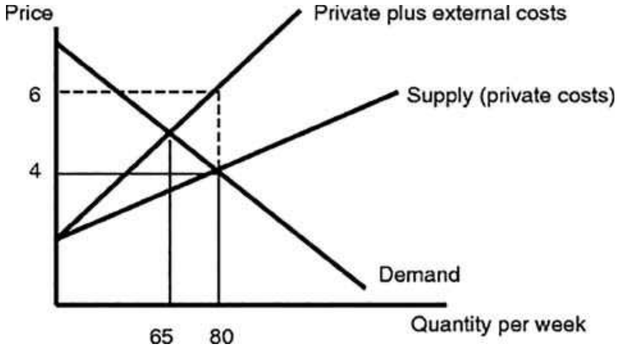


FIGURE 1 Private costs and external costs.

higher levels of output presumably because damages rise dramatically when there are very high concentrations of the pollutant in the atmosphere). Determining the extent of those damages requires some valuation metric.

For now, let us assume that the analysts "got it right" in estimating both benefits and costs. This is a strong assumption because environmental services are notoriously hard to price. This problem can limit the ultimate effectiveness of CBA because the abatement costs tend to be short-term and quantifiable, but the benefits (avoided damages) are often long-term and difficult to quantify. For now, consider the impacts of environmental regulation within the CBA framework. Regulation requires pollution abatement activity, raising production costs but reducing the pollution and associated damages (as shown in Figure 2).

The imposition of environmental regulation raises production costs (shifting the supply curve up) and reduces equilibrium consumption of the polluting good (from 80 to 75 per week) because the price has risen (from $4.00 to $4.40). In addition, external costs are reduced (so the sum of private and external costs is now $5 when 75 units of the good are produced). Emissions are reduced (though this particular figure only indicates the reduction in damages, not the precise reduction in emissions).

The next question is how much pollution abatement makes economic sense, since control costs rise rapidly as emissions are cut back towards zero. Continuing with our illustrative example, Default three depicts the total benefits of abatement and the total cost of abatement. The latter depends on the abatement technology and input prices and the interdependencies among production processes (for retrofitting control technologies). It is relatively easy to compute abatement costs from engineering cost studies, although predicting future control costs is not easy because innovations will create new control technologies. The benefits from abatement (or the reduction of pollution damages—the cost of pollution) depend on the size of the affected population, incomes (indicating an ability to pay for environmental amenities), and citizen preferences (reflecting a willingness to pay). The benefits can be very difficult to estimate. Consider, for example, the health benefits of reduced particulates in the atmosphere, habitat values, and citizen valuations of maintaining a habitat for a particular species. Physical benefits can be found from dose-response studies. Various survey methods and market proxies for computing willingness to pay to avoid experiencing the impacts of pollution have methodological problems. However, if the dollar metric is to be used for determining the benefits of environmental improvements, techniques can at least establish rough benchmarks (as shown in Figure 3) (Some argue strongly against the use of CBA.[4]).

The total benefits and costs to the marginal benefits and costs can be related because, for economists, the issue is not zero emissions versus unlimited emissions. In the former situation, if 80 units of the good results in 80 tn of emissions, then zero emissions reduced would be characterized as having no abatement costs (but also, no benefits from abatement). When the total benefits equal the total cost of abatement (at 65 tn of emissions reduced per week in Figure 3), the last reductions in emissions were very

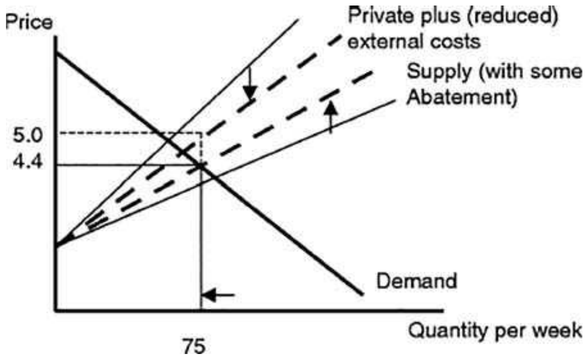


FIGURE 2 Reducing external costs.

costly and the additional benefits were fairly small. Zero abatement activity would also be inefficient in this example because the marginal damages are very high and marginal abatement costs are quite low. Economics tries to determine the optimal amount of emissions. In the hypothetical example, the "optimal" quantity of reduced emissions is about 25, where the marginal abatement cost is just equal to the marginal benefits of $2. These are depicted in Figure 4.

This outcome means that there are still 55 tn of emissions per week. If the estimated benefits and costs of pollution abatement are correct in this illustration, economic efficiency would be violated if additional resources were devoted to abatement activity. For example, if 50 tn of emissions were reduced (so only 30 tn of the pollutant are released), the marginal benefit would be about $1, but the marginal cost would be greater than $5. From the standpoint of economic efficiency, those resources could be used to create greater value in other activities.

Of course, the difficulty of obtaining a common dollar metric for all the impacts of different pollutants means that benefit-cost analysis must incorporate a range of values. The range could be incorporated in Figure 4 as a band around the marginal benefit curve indicating one standard deviation from the calculated values. A conservative approach would recognize that the marginal benefit function could be above that depicted in Figure 4, which would lead to optimal emission reduction of more than 25 tn per week (improving the ambient air quality).

Further complicating the analysis are production and exposure interdependencies. For example, the marginal cost of abatement associated with one type of emission may depend on the level of treatment (or abatement) for another contaminant. A joint optimization problem results, with the basic principles

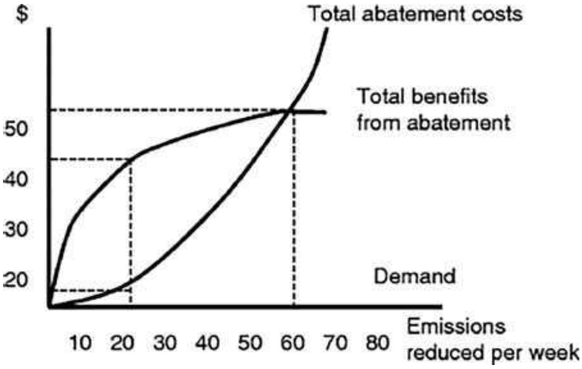


FIGURE 3 Total benefits and total cost of abatement.

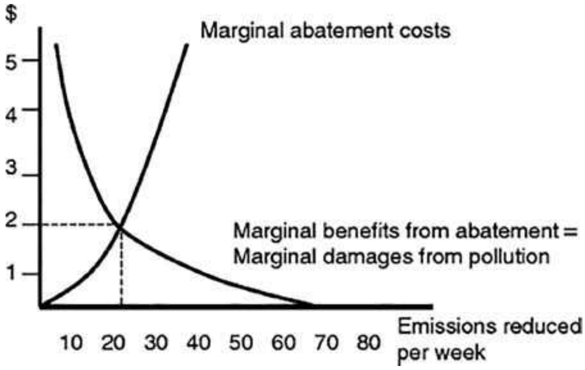


FIGURE 4 Marginal benefits and marginal cost of abatement.

unchanged. Many investments in abatement equipment have this characteristic: once one set of contaminants is being reduced from a discharge flow and the cost of dealing with additional contaminants can be relatively low. For example, in the case of water discharges, if iron or manganese is removed via the precipitation method, total dissolved solids (TDS) is reduced and there may be an improvement in water clarity.

Interdependencies can also arise on the benefit side when the dose-response relationship for a particular contaminant is influenced by the presence of other contaminants. Again, in the case of secondary groundwater standards, perceptions of odor and color will be affected by whether or not they occur in combination. Such considerations must be factored into the analysis when comparing the benefits and costs of different treatment options.

Cost-Effective Analysis

Instead of trying to estimate the dollar benefits of saving a human life (or reducing the incidence of asthmatic attacks), one can compare the number of lives saved per dollar spent in abatement activity across programs. Thus, cost-effective analysis involves finding the least-cost method of achieving a given economic or social objective such as saving lives or retaining unique ecological settings. No dollar value (or explicit measure of avoided damages) is placed on that objective.[5] One advantage of this approach is that the focus is on minimizing the cost of meeting the (politically determined) target. It promotes consistency across a range of programs that might be designed to address a particular problem, whether that involves health impacts or a loss of habitat. Cost-effective analysis facilitates comparisons across programs, leading to reallocations of resources devoted to meeting such targets as new information is gathered over time.

Policy Instruments

Political systems have passed legislation and created agencies to apply laws to improve environmental performance. For example, in the United States, the Water Pollution Control Act of 1956 and the Clean Air Act of 1963 and subsequent amendments to both pieces of legislation have focused on achieving ambient standards. The U.S. Environmental Protection Agency is responsible for implementing these laws, and in other nations agencies have also been established to reduce emissions and improve environmental outcomes. A number of policy options can lead to emission reductions.[6,7] These instruments have different economic efficiency implications. In addition, some of these approaches are difficult to implement (due to being information-intensive), some are not cost effective (in that other approaches achieve the same outcome at lower cost), and the distributional implications can differ across these approaches (tax burdens differ or some groups obtain valuable assets).

Tax on the Polluting Good

An excise tax could be imposed on the good, cutting back consumption to 65 units per week (Figure 1). Of course, the problem is not with the product but with the emissions associated with its production. Thus, this option does not provide incentives for developing new technologies that reduce abatement costs—it represents a static approach to the problem because it does not promote technological innovation.

Tax on Emissions

A penalty or charge for each ton of emissions would lead suppliers to cut back on emissions—to the extent that the abatement is less expensive than the tax. Thus, in Figure 4, a tax of $2/tn would lead to the optimal reduction of pollutants. In addition, it provides incentives for innovation in the control technology industry. Firms will seek ways to reduce abatement costs, thus reducing their pollution taxes. This strategy is likely to be opposed by polluters who will be passing the taxes on to customers (where the ultimate incidence depends on supply and demand elasticities in the product market).

Tradable Emissions Permits

The same result (and incentive) is obtained if "allowances" of 25 tn are allocated to polluting firms, limiting emissions (the situation is not completely identical—a tax has certain costs to firms but yields uncertain overall abatement because regulators will not have precise estimates of abatement costs; the allowances have certainty in terms of overall abatement but uncertain cost. Of course, with monitoring, the tax can be varied over time to achieve a desired ambient condition). This approach provides an incentive for those with low abatement costs to reduce emissions and sell their permits (allowances) to others whose abatement costs would be very high. This places entrants at a disadvantage because incumbent firms are "given" these valuable allowances. The SO_2 regime in the United States has this feature. Of course, the initial allocations raise political issues (because permits represent wealth). In establishing a tradable permit regime, an environmental agency must determine the allowed level of emissions (here, 25 tn) and whether additional constraints might be applied to local areas with particular circumstances. In addition, the energy sector regulator has to make decisions regarding the treatment of cost savings from the regime. For example, savings might be passed on to consumers or retained by firms. The latter situation provides an incentive for utilities to participate in the emissions trade markets. A sharing plan can also be utilized so customers benefit as well.

Tighten Liability Rules

An alternative approach would utilize a court-based system, where fees would be assessed against those responsible for damaging the health of others, for reducing the economic value of assets, or for reducing the amenity values of ecosystems. Of course, this approach requires a well-specified set of property rights and clear causal links between specific emitters and affected parties. The transaction costs of such a system (resources devoted to negotiations and legal activity) could be prohibitive for many types of pollutants.

Emission Reduction Mandates (Quantity-Based Command-and-Control)

Although equal percentage cutbacks sound "fair," this strategy is not cost-effective because abatement costs can differ widely across pollution sources. If there are scale economies to emission reductions, it would be most efficient to have a few firms reduce emissions. The least-cost way to achieve a given overall reduction in emissions will involve differential cutbacks from different firms.

Mandate a Specific Control Technology (Technology-Based Command-and-Control)

This "command and control" strategy is not cost-effective because production conditions and retrofitting production processes differ across firms (based on the age of the plant and other factors). However, this policy option has been utilized in a number of situations as a "technology-forcing" strategy.

Other Policy Issues

The above instruments have been utilized in different circumstances. Additional issues include intrinsic benefits, income distribution, sustainability, and renewable resources.

Intrinsic or Nonuse Benefits

Some people take a more expansive view of environmental amenities as they attempt to separate economic values from inherent values. However, this might be partly accounted for in terms of the perceived benefits to future generations. Intrinsic benefits from environmental programs include option values, existence values, and bequest values.[8] The first value represents a form of insurance so future access to a potential resource is not eliminated due to current consumption. The rationale behind option values is closely related to the "margin for error" argument noted earlier. Existence values reflect a willingness to pay for the knowledge that the amount of contaminant in the environment does not exceed particular levels or that a particular species (or level of biodiversity) is retained. The resource or ecological system is available for others. The bequest values can be interpreted as the willingness to pay for preserving a resource (or a geographic site) for future generations.

Redistributive Effects

It is important to note that citizens being harmed by emissions are not necessarily the same as those who are consuming the polluting good (such as electricity). Even if a particular program has positive net benefits, some parties are likely to be losers. They are seldom compensated and left better off, raising concerns about the distributional consequences of alternative policies. Furthermore, those harmed may have lower incomes (and thus, a lower willingness to pay to avoid damages due to the lower ability to pay). This point underscores the role of fairness as a factor that might outweigh efficiency considerations in some circumstances. Some agencies have been forbidden to use CBA on the grounds that the numbers are too speculative and that social concerns should be given priority. Intergenerational concerns can be interpreted as reflecting redistributive considerations.

Sustainable Development

Some of the issues associated with energy involve the use of nonrenewable resources (irreversibility). Some citizens argue that sustainability requires development that can be supported by the environment into the future. These people wish to ensure that resources are not depleted or permanently damaged. However, since sustainability depends on technology and innovations change resource constraints, defining the term with precision is quite difficult.

Renewable Energy Resources

Generating electricity without fossil fuels (e.g., hydro, wind, solar, biomass) is sometimes referred to as using green options. Green options are often limited in the amount (and reliability) of energy produced in a given time period. Utility applications for renewable resources include bulk electricity generation, on-site electricity generation, distributed electricity generation, and non-grid-connected generation. A number of regulatory commissions have required utilities to meet renewable portfolio standards. Such strategies reduce dependence on a particular energy source (to reduce the region's vulnerability to supply disruptions or rapid price run-ups). In addition, such requirements imply that managers are not making the most efficient investments in long-lived assets. Also, note that demand reduction through energy-efficient technologies is a substitute for energy, whatever the source.

Conclusions

The three main trends in environmental regulation in recent years have been shifting from command-and-control regulation towards a greater use of economic instruments (such as emissions trading), seeking more complete information on the monetary value of environmental costs and benefits, and a tendency for addressing environmental objectives in international meetings, as with the Kyoto Protocol.[9]

The interactions between economic and environmental regulation raise important policy issues. If energy sector regulation and environmental regulation remain separate, some means of harmonization may be necessary to promote improved performance. Collaboration would involve clarifying the economic duties of the environmental regulator and the environmental duties of the economic regulator. To avoid regulatory competition, agencies sometimes establish task forces or other mechanisms for identifying and resolving issues that might arise between jurisdictional boundaries (across states or between state and federal authorities). Such cooperation can serve to clarify the division of responsibilities and identify regulatory instruments that will most effectively meet economic and social objectives.

In summary, policy-makers respond to domestic political pressures by devising institutions and instruments to address pollution and environmental sustainability.[10] Although no single field of study contains all the tools necessary for sound policy formulation, economics does provide a comprehensive framework for evaluating the strengths and limitations of alternative policy options. Because of the pressures brought to bear by powerful stakeholders, adopted policies and mechanisms are not necessarily cost minimizing. The resulting inefficiencies may partly be due to considerations of fairness, which places constraints on whether, when, how, and where environmental impacts are addressed. As emphasized in this survey, citizens want to be good stewards of the land. We appreciate the adage: “The land was not given to us by our parents; it is on loan to us from our children.” How to be good stewards—through the development and implementation of sound environmental policies—has no simple answer given the complexity of the issues that need to be addressed.

References

1. Maler, K.-G.; Vincent, J.R. In *Handbook of Environmental Economics: Environmental Degradation and Institutional Responses*; North-Holland: Amsterdam, 2003; Vol. 1.
2. Maler, K.-G.; Vincent, J.R. *Handbook of Environmental Economics: Valuing Environmental Changes*; North-Holland: Amsterdam, 2005; Vol. 2.
3. Viscusi, W.K.; Vernon, J.M.; Harrington, J.E., Jr. *Economics of Regulation and Antitrust;* MIT Press: Cambridge, MA, 2000.
4. Ackerman, F.; Heinzerling, L. *Priceless: On Knowing the Price of Everything and the Value of Nothing*; New Press: New York, 2004.
5. Freeman, A.M., III *The Measurement of Environmental and Resource Values: Theory and Methods,* 2nd Ed. Resources for the Future: Washington, DC, 2003.
6. Portney, P.; Stavins, R. *Public Policies for Environmental Protection*; Resources for the Future: Washington, DC, 2000.
7. Vig, N.; Kraft, M.E. *Environmental Policy and Politics: New Directions for the Twenty-first Century,* 5th Ed.; Congressional Quarterly Press: Washington, DC, 2003.
8. Krutilla, J. Conservation reconsidered. Am. Econ. Rev. **1967**, *57* (4), 777–786.
9. Busch, P.-O.; Jörgens, H.; Tews, K. The global diffusion of regulatory instruments: the making of a new international environmental regime. In *The Annals of the American Academy of Political and Social Science,* Sage Publications: Thousand Oaks, CA, 2005, Vol. 598, 146–167.
10. Kraft, M.E. *Environmental Policy and Politics,* 3rd Ed.; Pearson/Longman: New York, 2004.

30

Environmental Policy: Innovations

Alka Sapat

Introduction

Over the past three decades, American environmental policy management has often entailed the search for new solutions and innovative ways of managing environmental problems. These solutions have taken the form of environmental policy innovations (EPIs), which have frequently been pioneered and adopted by local and state entities, at times with the help of citizen groups and non-profit organizations. They have become an important means of managing environmental problems, particularly for intractable, multimedia, multijurisdictional environmental issues. While policy innovations have been studied in a number of areas at the state level,[1–19] less attention has been paid to policy innovations in environmental management.[20–24] The focus of this entry is to shed light on these environmental policy initiatives, termed here as EPIs.

To provide an understanding of these policy initiatives, I begin by defining and discussing EPIs and supply some understandings of the concept. Next, I discuss the nature and type of a large majority of EPIs; in particular, I focus on their reliance and inclusion of collaborative forms of environmental management and provide some examples. The fourth section of the entry analyzes some of the main factors affecting the adoption of EPIs, including determinants such as institutional commitment, resources, the severity of the problem, and the role of interest groups. The effect of neighboring entities in spurring innovation adoption and diffusion is also considered in this section. The entry ends with a conclusion that summarizes the main points of the entry, provides some considerations for further research, and discusses the policy implications of adopting environmental policy initiatives as a strategy for managing the environment.

Environmental Policy Innovations

EPIs here are defined in this entry as government-sponsored initiatives to protect the environment. This is based in part on the original definition of a policy innovation originally espoused by Walker[1] who examined the adoption of new policies by states, which he termed as "innovations." Walker's work also differentiated between states that were "leaders" and those that are deemed to be "laggards," along with examining the diffusion or the spread of these innovations from leading states to other states. Thus, he analyzed the amount of time that elapsed between the invention of an innovative program and a state's adoption of it. Based on 88 different state programs, he concluded that larger, wealthier, and more industrialized states adopt new programs more rapidly than do small, less-developed states. He also found strong regional relationships among states that were thought to reflect conditions of emulation and competition among proximate states. Walker's research lead to further investigation into state policy innovations. Gray[3] undertook a more detailed analysis and rejected Walker's contention that leading states lead no matter what the policy area; in a study of education, welfare, and civil rights policies, Gray found that innovativeness was not a pervasive factor. She concluded that based on the issue and the era, different states would comprise the leading, "middle-adopting," and lagging clusters.

During the next two decades, there were numerous other studies of state innovation that were published in a variety of policy areas. These included studies of innovation in juvenile corrections,[12] consumer affairs,[13] technology,[14] energy,[15] tort law,[16] judicial administration,[17] and human services.[18] These studies explored various determinants of innovativeness in different policy areas. However, while they expanded the scope of policy areas subject to innovation analysis, they did not lead to major advances in the conceptualization of state innovations or in the empirical approaches to its investigation.

In the 1990s, however, significant advances were made in the field of state innovations research. Research on state innovation was tested with sophisticated empirical methods in the area of state lottery adoptions[4] and state taxation.[5] In both these studies, the authors extended theories of innovation to different policy areas and made a substantial methodological contribution to the literature by refining the empirical analysis of state innovation with the employment of event history analysis (EHA), a pooled cross-sectional time-series technique. Berry and Berry[4,5] also analyzed the importance of internal determinant explanations (which posit that the factors causing a state government to innovate are the political, economic, and social characteristics of a state) and regional diffusion explanations (which point toward the role of policy adoptions by neighboring states in prompting a state to adopt).

An interesting twist on the diffusion aspect of state innovations was presented in a study of state living will laws.[19] In this study, Glick and Hays argued that as policy innovations diffuse across states, a process of "policy reinvention" occurs.[19] Policy innovations, in their view, are not adopted in their original form; rather, they are changed or selectively adapted as innovations by other states. The contribution of this study lay in refinements that added to earlier understandings of innovation diffusion.

Another variation on state innovation research was the study undertaken by Berry.[6] In this study, Frances Stokes Berry[6] extended the original research undertaken by Berry and Berry[4,5] to state innovations in strategic planning by various regulatory agencies. This extension was interesting as it was one of the first studies to empirically evaluate an innovation adopted by state bureaucracies as opposed to legislatures. This research was followed by more sophisticated understandings of the catalysts and actors behind the adoption of innovations; in his analysis, Mintrom[7,8] focused specifically on the importance of policy entrepreneurs at the state level. Using both secondary data sources and primary data gathered from a 50-state survey of state education officials in an EHA, Mintrom examined both the consideration of school choice and its adoption by states across time. He focuses in particular on the role of policy entrepreneurs by developing a theory of policy entrepreneurship and testing that theory with respect to school choice. Using EHA, he concluded that policy entrepreneurs act as catalysts in the policy innovation process by promoting ideas for dynamic policy change, persuading legislators, and networking with others in government.

In the area of environmental policy, state policy innovations has been the subject of research by Rabe,[20–23] Sapat,[24] and others. For instance, in a comparison of Canada and the United States, Rabe[21] found that despite the relatively decentralized Canadian institutional framework and more centralized and "bureaucratic" framework in the United States, the latter had been much more active and effective in devising innovative policy approaches. Comparing four U.S. states and four Canadian provinces, he concluded that within those sample jurisdictions, the United States was clearly ahead of Canada in cross-media permit integration, pollution prevention, disclosure of information on toxic substances, and achieving greater refinement and use of environmental outcome indicators. Similarly, in later research analyzing state experimentation in the realm of climate change, Rabe[22] analyzed the factors prompting successful state initiatives in reducing greenhouse gases. He finds that some states did succeed in cutting through traditional partisan divides and that agent- based entrepreneurs helped to develop policy ideas and form viable coalitions.

State EPIs adopted by administrative agencies (rather than by legislatures) have also been the subject of past research. In a study of state policy innovations adopted by agencies in the areas of hazardous waste and groundwater contamination, Sapat[24] found that state agencies are more likely to adopt innovations to deal with problems created by hazardous waste contamination than for groundwater contamination. Further, she finds that state environmental managers are not directly influenced by interest groups and that the inclusion of all stakeholders is likely to lead to greater support for new policy initiatives.

EPIs as defined here are drawn from and are based on this past research. An EPI as defined here can be one that is initiated by an agency or government institution (such as a state legislature, county government, or city) or it can be one that is borrowed from another government entity/ institution. Thus, EPIs can be adopted by legislatures or administrative agencies. Also, as defined in prior research by Sapat,[24] EPIs are defined as initiatives that protect the environment. The emphasis on the positive aspects of EPIs is important, because initiatives can have both positive and negative benefits. The term EPI as used in this entry is positive, i.e., it refers to laws, regulations, and policies that are beneficial to the environment. In other words, the goals of these EPIs are to regulate pollution problems and protect the environment.

At this point, a caveat is necessary. Clearly, policy choices can be subjectively interpreted; they may be viewed as being positive for the environment by some and as bad for the environment by others. This is particularly true within the realm of environmental politics.[25–27] For instance, the use of economic incentives, such as fees, pollution taxes, or offsets, is viewed as being beneficial by some and negative by others.[25] Even environmental organizations are divided on this issue. For instance, large environmental interest groups such as the Environmental Defense Fund, which utilize the services of economists and scientists, view economic incentives as being beneficial for the environment.[28] However, a number of smaller environmental interest groups are more suspicious about the environmental benefits of adopting market-based incentives.[29] Moreover, the problem of clearly defining an environmentally "positive" policy solution is also compounded by the fact that for some environmental issues, even scientific experts cannot come to an agreement on what the best technical solution is to a pollution problem.[30] Hence, there is extensive debate about the benefits of particular environmental policies, because of conflicting perspectives regarding policy outcomes and because of incomplete and inconclusive technical knowledge of solutions to environmental problems. To overcome some of these problems, the EPIs discussed here are those that are deemed as being beneficial for the environment by the United States Environmental Protection Agency (USEPA).[31,32] These assessments are arguably less subjective than those carried out by environmental or industry interest groups.

The Nature of EPIs: Conflict to Collaboration

Following the definition given above, EPIs can be categorized in various ways. For instance, they may be categorized according to their function; the USEPA provides a portfolio for instance, in which various environmental innovations are characterized by the core agency function that they address, such as helping in improving service delivery, enhancing regulatory outcomes, supporting superior environmental performance, designing targeted geographic solutions, etc. While it is useful to categorize environmental innovations thus, other ways of classifying environmental innovations may provide a broader and deeper understanding of these initiatives.

In this entry, I contend that rather than categorizing EPIs purely by function, it is also useful to classify such innovations by the underlying regulatory and managerial premises of such policy innovations. More specifically, it is instructive to understand the extent to which such innovations rely on collaborative as compared to non-cooperative, more conflict-based approaches to environmental management. This perspective would be useful because there has been growing evidence in both practice and theory about the usefulness of collaborative approaches to managing the environment and natural resources.

Reasons for the emergence of these collaborative approaches have been many and varied, but the primary reasons for the growth in collaborative forms of environmental management have been ascribed to the problems stemming from command and control measures.[33–39] These traditional forms of regulation have often resulted in failure to achieve key goals and/or protracted conflict. Traditional forms of environmental regulation that rely on bureaucratic, adversarial, and often technology-based regulatory approaches are also seen as being far too rigid and, more importantly, ineffective, in that they fail to address multimedia and multijurisdictional environmental hazards, particularly those stemming from non-point pollution.[33,36] Furthermore, traditional hierarchical approaches do not allow for more democratic forms of participation and often rely on unrealistic models of individual and administrative rationality.[37,40]

To overcome such problems, a search for alternative solutions to managing environmental problems began to evolve and grow. For instance, a study by John in 1994[40] found that states and local governments were increasingly turning away from “top–down command-and-control” regulation to a bottom–up style, which he termed “civic environmentalism.” Civic environmentalism relies primarily on the use of non-regulatory tools, which encompass market-based incentives to regulation.

Civic environmentalism or collaborative environmental approaches or similar interchangeable terms that can be used to describe a more cooperative, less adversarial form of environmental management are to be found in a number of EPIs that are undertaken by environmental agencies at different levels of government. Indeed, some EPIs were initially pioneered and adopted because of the failure of conflict-based approaches to manage environmental problems. For instance, environmental permitting processes engender a high level of stakeholder frustration and conflict due to permitting backlogs, long lead times, costs, and uncertainty. To improve these processes and reduce conflict and backlog times, some state and local agencies adopted EPIs to shift away from media-specific permitting for individual facilities and also to improve internal agency permitting processes. For instance, Massachusetts shifted away from a facility permitting approach to a multimedia, sector-based regulatory approach, targeting sectors with large numbers of small sources, as an alternative to facility-specific state permits with industry-wide environmental performance standards and annual self-certifications of compliance.[31,41]

Similarly, other EPIs take a collaborative approach by helping to enhance partnerships to resolve environmental problems that cannot be effectively solved without the participation and collaboration of multiple actors. Leveraging such partnerships enables environmental agencies to solve complex problems by harnessing energies and resources of stakeholders to achieve mutually desirable outcomes. An example of an EPI that does this is the effort undertaken in South Carolina to reduce neighborhood contamination. This initiative enlists the support of numerous community organizations and local businesses to provide education and outreach to reduce community exposure to lead and other hazardous substances as part of the Charleston–North Charleston Community-Based Environmental Partnership (CBEP).[42]

Other examples of collaboration-based EPIs[43] are provided in Table 1.

To summarize, EPIs are initiatives adopted to protect the environment and a number of them have been adopted at the local, state, and federal level by both legislative and administrative agencies. These EPIs are characterized by more collaborative approaches to environmental management. To understand EPIs further, we can try to ascertain which factors are likely to affect the adoption of these innovations. It is to this issue that I turn to in the next section.

TABLE 1 Examples of EPIs

Neighborhood Contamination Reduction—South Carolina	Enlists numerous community organizations and local businesses to support education and outreach to reduce community exposure to lead and other hazardous substances as part of the Charleston–North Charleston Community- Based Environmental Partnership (CBEP). (http://www.epa.gov/Region4)
Northeast Ohio Initiatives—Ohio	Responds to regional economic, sprawl, ecosystem and infrastructure challenges through a 15-county, community-based approach. (http://www.epa.gov/glnpo/lakeerie/leneohio.htm)
Urban Environmental Program—Boston, Massachusetts	Adopts a community-based approach including the city of Boston, Massachusetts DEP, and community organizations to improve the quality of life in urban settings by targeting issues such as asthma and indoor air quality, lead poisoning, vacant lots and green spaces, and pollution prevention. (http://www.epa.gov/boston/eco/uep/boston/)
Pollution Complaint Response—Indiana	Coordinates an agency-wide, multimedia response to citizen inquiries and complaints using Web-based information, enabling the agency to reduce costs and increase public trust. (http://www.in.gov/idem/5274.htm)
Ford Good Neighbor Dialogue—Illinois	Brings together stakeholders, academics, and agency representatives in a collaborative process to periodically discuss a large manufacturing facility's environmental management and performance. (http://www.delta-institute.org/)
Community Environmental Awareness Project—Michigan	Develops an approach to improve the way environmental information is presented and made available to the public; the goal of the CEAP is to improve the public's access to and understanding of how major industries are performing under environmental laws and regulations. (http://www.deq.state.mi.us/ceap)

Resources, Needs, Politics, and Other Determinants

The factors that lead to the adoption of EPIs can be many and varied. Prior research on innovation adoption in other areas, as well as on EPI adoption is useful here. The following factors are some of the most influential determinants of innovation adoption as seen in extant research.

Need/Problem Severity

Keeping in with the adage that "necessity is the mother of invention," researchers have traditionally regarded problem severity as a significant influence on the adoption of innovations.[1,3,44,45] For innovations adopted at the state and local level, the expectation is that states and local governments rather than the federal government are likely to understand local problems more clearly and are hence likely to be responsive to the needs and problems present in their jurisdictions.[27,45] For instance, in a study of state innovation, ranging from state testing on teacher competency and rail passenger service to state regulation on sodomy, Nice[45] found that the problem environment was prominent in explaining five of eight policy innovations analyzed in this study. For EPIs, the severity of the pollution problem itself has been regarded as one of the most important reasons for innovation adoption. To deal with intractable environmental problems, several entities (local, state, non-profit, and even private institutions) have come up with new ways to solve old problems.[20,31]

Institutional Factors: The Importance of Commitment and Capacity

While necessity can often drive innovation, institutional commitment and capacity have also been found to be important for the adoption of policy innovations.[24,40,46] The theoretical origins of the role of institutional actors may be found in ideas related to roles of institutional elites and theories of institutionalism and neo-institutionalism.

Elites: The importance of elites in the policy-making process is theoretically supported by the elite perspective of policy analysis. Elite theorists argue that power is concentrated in the hands of elites who use the resources of their respective organizations to manage and impose order on society.[47] Societal stability, according to this perspective, rests not on a common political culture and a set of values, but on a forced consensus created and reinforced by the elite. Pluralist politics can coexist with the governmental elite; however, key decisions regarding policy are made by elites. Popular and electoral politics are, for the most part, mainly symbolic and concerned with middle-level policy issues, according to the elite perspective.[47]

The elite theory of the policy process is close to the neopluralist view: neo-pluralists, such as Lindblom,[48] challenged the pluralist notion that power was diffuse and argued instead about the privileged position of business. Similarly, others such as E.E. Schattschneider pointed out, “the flaw in the pluralist heaven is that the heavenly chorus sings with a strong upper-class accent.”[49]

Institutionalism and Neo-institutionalism: The motivation of institutional actors and elites is also stressed in theories that emphasize the importance of institutions in policy and governance.[50–54] Institutional theories hold that government actors can act independently of interest group pressures and other factors. According to this view, government actors are not just simply “pawns” of various interest groups; rather, the perceptions and attitudes of these actors shape the way they process information and affect independently the choices they make.[52,54] Institutional theories also recognize that informational constraints and computational limitations of political actors prevent actors from making purely “rational decisions” that are independent of the actor’s subjective representation of the decision problem.[52,55] Thus, the attitudes and ideological views of institutional actors can influence their choices in innovation adoption.

Based on the theories discussed above, it is likely that pro-environmental policy actors will likely push for adoption of EPIs. Related to that issue is the importance of resources. Adoptions of EPIs require both the motivation and the commitment of key institutional actors in terms of time, money, and other resources. Typically for most policy adoptions, resources are required. For this reason, wealthier and larger states and local governments are more likely to be innovation adopters.[11,56] Resource provision by third- party institutions, such as non-profits, or by higher levels of government, such as by the federal government to state and local governments, can also help spur innovation adoption and diffusion. For instance, the USEPA provides State Innovation Grants to spur innovation adoption; these grants are typically provided to states with projects in three areas: environmental results programs, environmental management systems and permitting, and environmental leadership programs such as EPA’s Performance Track.[31]

However, while institutional resources may be required for innovation adoption, some EPIs are adopted to economize on resources or to raise revenues. For instance, EPIs in pesticide regulation adopted by certain states helped raise fees, adding to revenue sources.[24] The role of resources or the importance of economic factors can also be played out in other ways. That is, objectives of economic development by states could deter their willingness to adopt strict environmental regulations.[57–60] For instance, states with a higher percentage of economic activity relying on manufacturing or industrial activities with negative environmental consequences are less likely to favor the adoption of innovative laws and programs that would negatively affect revenue-generating industries. This raises the issue then of interest group support and its importance for EPI adoption.

Interest Group Support

Interest groups have always played an important role in the American policy process, due in part to constitutional provisions for pluralism. Theoretically, the pluralist perspective of policy making argues that public policy is the product of democratic participation by individuals who are represented by organizations such as interest groups. Theories of interest group influence on legislators and bureaucrats range from narrow views of interest group influence[61–63] to those that posit interest groups as not exercising any more influence over regulatory policy than other actors or bureaucrats.[64,65]

Pluralist theorists, such as Truman[ZS] and Dahl[67] argued that group exchange (pluralism) was the dominant (and desired) method of political decision making. In their view, this was possible because the decentralization of American institutions and the relative openness of the political system guaranteed competing groups access and some degree of power. Pluralism thus presents a view of the political system in which multiple centers of power compete to shape policy. Power is diffuse, fluid, multifaceted, and dispersed among a number of groups. Political participation is largely a goal- oriented activity in which citizens take part in order to obtain some benefit from the government.[68]

Pluralist group theory was questioned by a number of scholars, mainly neo-pluralists, who argued that dominant groups and interests had the ability to control the political agenda, preventing pluralist discourse over a full range of policy options.[47,48,69] Group theory was also evaluated in part in analyses of regulatory policies.[62,70,71] A number of scholars argued that regulatory policies emerged from a political equilibrium produced by coalitions of regulated industries and their customer groups, leading to "agency capture."[62,72,73] This narrow interest group perspective is based on the assumption that certain groups of people, who are organized and powerful in terms of possessing economic resources, will have the capability to dominate policy at the subnational level.

While this view is compatible with the Madisonian perspective of private parochial interests modifying policy outcomes, it has been modified considerably to account for characteristics of interest groups such as size and density that could affect their capabilities to exert influence over state regulations.[74] Moreover, the motivations and actions of other political actors, bureaucrats, and other interest groups themselves have also been found to be important in influencing policy outcomes.[75] Further, the narrow view of interest groups adduced by scholars such as Stigler[62] and Posner[73] have been more successful in explaining economic regulation as opposed to social regulation, such as consumer protection and regulation of the environment. The traditional "iron triangle" or "agency capture" theory does not adequately account for the emergence and power of newly organized and motivated consumer, environmental, and other "public interest" groups, which seek social benefits.

With respect to environmental policy, past studies have found environmental interest groups to be important in influencing policy outcomes. For instance, Ringquist[76] and Hird[27] analyzed the influence of environmental interest groups that countered the regulatory priorities advocated by industry groups. Ringquist[76] found environmental interest groups to have a significant influence on state water quality regulation in a group influence model of state policy influence.

With regard to EPIs, the presence and activities of environmental interest groups are likely to be important as well. However, industry interest group opposition to EPIs may be tempered by the nature of these policy innovations. As discussed above, a number of EPIs rely on collaborative nonadversarial approaches to solve environmental problems. Past research has also shown that some EPIs provide new ways to work with such industries to improve compliance with regulations and do not burden them with any additional regulatory costs.[24] Interest group opposition may also be mitigated by public support for certain EPIs, which provide broader benefits than they do costs to the public at large.

Regional Diffusion

Previous empirical studies on the diffusion of innovations have found that external influences by other neighboring entities are important. For instance, adoption of innovations by neighboring states can affect a state's adoption of a policy innovation.[1,4,11,56,77] Policy adoptions by neighboring states can decrease the information costs regarding the possible consequences, including the electoral consequences, of the adoption of a policy.[4] State officials can view the relative success/failure of programs in these adjacent states and decide whether such programs are suitable for their own state. Moreover, if a policy has been successful in a neighboring state, then state officials, in particular state legislators, can, by using the experiences of neighboring states as an example, boost public and legislative support for similar legislation in their own state. This allows them then to selectively utilize the experiences of neighboring states for their own political gain. The decrease in information costs and political uncertainty can thus increase the motivation of state policy makers to initiate or adopt a policy innovation.[24]

While early research on policy innovation diffusion focused on regional influences primarily with respect to policy innovations adopted by states,[8,11,24] later work has also studied factors affecting diffusion at the country level[78] and the local level.[11] On a related note, there has also been research on examining the mechanisms of policy diffusion,[79–82] which has moved away from looking simply at the effects of neighboring states to analyzing the more complex and multiple factors affecting diffusion, such as the presence of competition, learning, imitation, and coercion; this research has found that policy mimicry is often tempered by the size of the state and by the presence of other third-party institutions that may aid diffusion.[11] Thus, while regional effects or "external determinants" may be important, it is also necessary to understand the mechanisms of diffusion, which may be varied and complex.

There are thus, a number of factors that can influence the adoption and implementation of EPIs; need or the type of problem alone may be a necessary but not a sufficient determinant of innovation adoption. Institutional commitment and capacity, the role played by interest groups, and the effects of innovation adoption by neighboring institutions and entities are also important influences affecting the espousal of EPIs.

Conclusion

Complex environmental problems require innovative solutions for effective management. The focus of this entry has been on such initiatives or EPIs. In doing so, this entry began by defining and explaining the nature of such initiatives, including a discussion on past research on state policy innovations. I contend that a number of environmental initiatives have pioneered and are symptomatic of a shift in environmental management to more collaborative ways of solving environmental problems. Need or the severity of the problem has been one of the reasons affecting the adoption of EPIs, but institutional commitment and capacity are also critical to adoption and implementation of policy innovations in general and for EPIs. Given that interest groups are critical in the American policy process, the support or opposition of key groups is also vitally important to innovation adoption. Since a number of EPIs provide new ways to work in partnership rather than in conflict with industries to improve compliance with regulations and do not burden them with any additional regulatory costs,[24] industry opposition to such EPIs may be lessened. Interest group opposition may also be mitigated by public support for certain EPIs, which provide broader benefits than they do costs, to the public at large. Finally, innovation adoption may also be affected by the presence or adoption of innovations by neighboring entities, the external determinants. However, the mechanisms of such diffusion need further study and research.

Future research could also address the policy implications and long-term outcomes of innovation adoption. For instance, do the adoptions of EPIs lead to better environmental outcomes in terms of solving environmental problems and improve environmental quality? Do EPIs foster better outcomes in terms of increasing stakeholder participation and provide for more democratic and innovative means of achieving environmental goals? Preliminary evidence does suggest that EPIs have positive outcomes, but more research is needed on this issue. Given that the environmental problems facing us in the coming century are only likely to grow and be exacerbated by the slow-moving yet devastating effects of more large-scale problems such as climate change, the need for innovative environmental solutions is likely to be critically important.

Acknowledgments

This research was funded in part by the National Science Foundation, NSF Grant No. SBER-9510308. The author is responsible for any omissions or errors.

References

1. Walker, J. The diffusion of innovation among the American states. Am. Polit. Sci. Rev. **1969**, *63*, 880–899.
2. Mohr, L. Determinants of innovations in organizations. Am. Polit. Sci. Rev. **1969**, *63*, 111–126.

3. Gray, V. Innovation in the States: A diffusion study. Am. Polit. Sci. Rev. **1973**, *67,* 1174–1185.
4. Berry, W.D.; Berry, F.S. State lottery adoptions as policy innovations: An event history analysis. Am. Polit. Sci. Rev. **1990**, *84,* 395–415.
5. Berry, W.D.; Berry, F.S. Tax innovation in the States: Capitalizing on political opportunity. Am. J. Polit. Sci. **1992**, *36,* 715–742.
6. Berry, F.S. Innovations in public management: The adoption of strategic planning. Public Adm. Rev. **1994**, *54,* 322–330.
7. Mintrom, M. Policy entrepreneurs and the diffusion of innovation. Am. J. Polit. Sci. **1997**, *41* (3), 738–770.
8. Mintrom, M. Policy Entrepreneurship in Theory and Practice: A Comparative State Analysis of the Rise of School Choice as a Policy Idea. Unpublished Ph.D. dissertation. State University of New York at Stony Brook, 1994.
9. Mooney, C.Z.; Mei-Hsien, L. Legislating morality in the American states: The case of the pre-Roe abortion regulation reform. Am. J. Polit. Sci. **1995**, *39* (3), 599–628.
10. Berry F.S.; Berry, W.D. Innovation and diffusion models in policy research. In *Theories of the Policy Process;* Sabatier, P., Smith, H.-J., Eds.; Westview Press: CO, 2007; 169–180.
11. Shipan, C.R.; Volden, C. The mechanisms of policy diffusion. Am. J. Polit. Sci. **2008**, *52* (4), 840–857.
12. Downs, G.W., Jr.; Mohr, L.B. Conceptual issues in the study of innovation. Adm. Sci. Q. **1976**, *21* (December), 700–714.
13. Siegelman, L.; Smith, R. Consumer regulation in the American states. Soc. Sci. Q. **1980**, *61,* 58–76.
14. Menzel, D.; Feller, I. Leadership and interaction patterns in the diffusion of innovation among the American states. West. Polit. Sci. Q. **1977**, *30* (4), 528–536.
15. Regens, J.L. State policy responses to the energy issue: An analysis of innovation. Soc. Sci. Q. **1980**, *61* (June), 44–57.
16. Canon, B.C.; Baum, L. Patterns of adoption of tort law innovations: An application of diffusion theory to judicial doctrines. Am. Polit. Sci. Rev. **1981**, *75,* 975–990.
17. Glick, H.R. Innovation in state judicial administration: Effects on court management and organization. Am. Polit. Q. **1981**, *9*, 49–69.
18. Sigelman, L.; Roeder, P.W.; Sigelman, C.K. Social service innovation in the American states: Deinstitutionalization of the mentally retarded. Soc. Sci. Q. **1982**, *62* (3), 503.
19. Glick, H.R.; Hays, S.P. Innovation and reinvention in state policy-making: Theory and the evolution of living will laws. J. Polit. **1991**, *53* (3), 835–850.
20. Rabe, B. Power to the states: The promise and pitfalls of decentralization. In *Environmental Policy in the 1990s*; Vig, N., Kraft, M.E., Eds.; Congressional Quarterly Press: Washington, DC, 1995; 31–52.
21. Rabe, B.G. Federalism and entrepreneurship: Explaining American and Canadian innovation in pollution prevention and regulatory integration. Policy Stud. J. **1999**, *27* (2), 288–306.
22. Rabe, B. *Statehouse and Greenhouse: The Stealth Politics of America Climate Change Policy*; Brooking Institution Press: Washington, DC, 2004.
23. Rabe, B. Racing to the top, the bottom, or the middle of the pack? The evolving state government role in environmental protection. *In Environmental Policy: New Directions for the Twenty-First Century;* Vig, N., Kraft, M.E., Eds.; Congressional Quarterly Press: Washington, DC, **2010**; 27–51.
24. Sapat, A. Devolution and innovation: The adoption of state environmental policy innovations by administrative agencies. Public Adm. Rev. **2004**, *64* (2), 141–151.
25. Davis, C.E. *The Politics of Hazardous Waste;* Prentice- Hall: Englewood Cliffs, NJ, 1993.
26. Lester, J.P.; Bowman, A. Subnational hazardous waste policy implementation: A test of the Sabatier-Mazmanian model. Polity **1989**, *21* (4), 731–753.
27. Hird, J. *Superfund: The Political Economy of Environmental Risk;* John Hopkins University Press: Baltimore, MD, 1994.

28. Environmental Defense Fund. Using Economics to Solve Eco-Challenges, available at http://www.edf.org/approach/markets (accessed November 2011).
29. Ingram, H.; Colnic, D.; Mann, D.E. Interest groups and environmental policy. In *Environmental Politics and Policy: Theories and Evidence,* 2nd Ed.; Lester, J.P., Ed.; Duke University Press: Durham, NC, 1995.
30. Barke, R.P.; Jenkins-Smith, H.C.; Silva, C. Consistency and Controversy in the Translation of Scientific Knowledge into Policy Recommendations. Paper Presented at the Annual Meeting of the Midwest Political Science Association, Chicago, IL, April 6–8, 1995.
31. United States Environmental Protection Agency. *Environmental Innovation Portfolio.* U.S. EPA, National Center for Environmental Innovation, Washington, DC, available at http://www.epa.gov/osem/pdf/portfolio.pdf (accessed October 2011).
32. Morandi, L. *Groundwater Protection Legislation: Survey of State Action 1988–1992;* National Conference of State Legislatures: Washington, DC, 1994.
33. Multi-State Working Group on Environmental Performance and Regulatory Policy Program, Mossavar-Rahmani Center for Business and Government. *Environmental Innovation: A Dialogue on the Role of Government, Law, and Regulatory Approaches*. University of Massachusetts, Lowell and the John F. Kennedy School of Government, Harvard University, January 2006.
34. National Academy of Public Administration. *Resolving the Paradox: EPA and the States Focus on Results*; NAPA: Washington, DC, 1997.
35. O'Leary, R.; Durant, R.F.; Fiorino, D.; Weiland, P.S. *Managing for the Environment: Understanding the Legal, Organizational, and Policy Challenges*; Jossey-Bass: San Francisco, 1999.
36. Durant, R.F.; Fiorino, D.; O'Leary, R., Eds. *Environmental Governance Reconsidered: Challenges, Choices, and Opportunities*; The MIT Press: Cambridge, MA, 2004.
37. Koontz, T.M.; Thomas, C.W. What do we know and need to know about the environmental outcomes of collaborative management? Public Adm. Rev. **2006**, *66*, 111–121.
38. Weber, E.P. *Pluralism by the Rules: Conflict and Cooperation in Environmental Regulation*; Georgetown University Press: Washington, DC, 1998.
39. Rogers, E.; Weber, E.P. Thinking harder about outcomes for collaborative governance arrangements. Am. Rev. Public Adm. **2010**, *40* (5), 546–567.
40. John, D. *Civic Environmentalism: Alternatives to Regulation in States and Communities*; Congressional Quarterly Press: Washington, DC, 1994.
41. Massachusetts Department of Environmental Protection. MassDEP Environmental Results Program, available at http://www.mass.gov/dep/service/envrespr.htm (accessed November 2011).
42. United States Environmental Protection Agency. About EPA Region 4, available at http://www.epa.gov/Region4 (accessed November 2011).
43. United States Environmental Protection Agency. Environmental Innovation Portfolio, available at http://www.epa.gov/osem/portfolio/protection.htm (accessed November 2011).
44. Savage, R.H. When a policy's time has come: Cases of rapid policy diffusion, 1983–1984. Publius **1985**, *15* (3), 111–125.
45. Nice, D.C. *Policy Innovation in State Government;* Iowa State University Press: Ames, Iowa, 1994.
46. Karch, A. National intervention and the diffusion of policy innovations. Am. Polit. Q. **2006**, *34* (4), 403–426.
47. Mills, C.W. *The Power Elite*. Oxford University Press: New York, 1956.
48. Lindblom, C. *Politics and Markets: The World's Political Economic Systems*; Basic Books: New York, 1977.
49. Schattschneider, E.E. *The Semisovereign People; A Realist's View of Democracy in America;* Holt, Rinehart and Winston: New York, 1960.
50. Noll, R.; Owen, B.M. *The Political Economy of Deregulation: Interest Groups in the Regulatory Process*; American Enterprise Institute for Public Policy Research: Washington, DC, 1983.

51. Moe, T. The politics of bureaucratic structure. In *Can the Government Govern?* Chubb, J.E., Petersen, P.E., Eds.; Brookings Institution: Washington, DC, 1989.
52. North, D.C. *Institutions, Institutional Change and Economic Performance*; Cambridge University Press: New York, 1990.
53. Chubb, J.E.; Petersen, P.E. *Can the Government Govern?* Brookings Institution: Washington, DC, 1989.
54. Van Horn, C.E. *The State of the States,* 3rd Ed.; Congressional Quarterly Press: Washington, DC, 1996.
55. Herbert A. Decision Making and Problem Solving. Research Briefings 1986: Report of the Research Briefing Panel on Decision Making and Problem Solving. National Academy Press: Washington, DC, 1986.
56. Shipan, C.R.; Volden, C. Bottom-up federalism: The diffusion of antismoking policies from U.S. cities to states. Am. J. Polit. Sci. **2006**, *50* (4), 825–843.
57. Getz, M.; Walter, B. Environmental policy and competitive structure: Implications for the hazardous waste management program. Policy Stud. J. **1980**, *9* (Winter).
58. Feiock, R.C.; Rowland, C.K. Environmental regulation and economic development: The movement of chemical production among states. West. Polit. Q. **1991**, 561–576.
59. Lowry, W. *The Dimensions of Federalism;* Duke University Press: Durham, NC, 1992.
60. Feiock, R.C.; Davis, C. Can state hazardous waste regulation be reconciled with economic development: A test of the Rowland–Goetze model. Am. Polit. Q. **1992**, 19.
61. Bernstein, M. *Regulating Business by Independent Commission*; Princeton University Press: Princeton, NJ, 1955.
62. Stigler, G. The theory of economic regulation. Bell J. Econ. Manage. Sci. **1971**, *2*, 3–21.
63. Peltzman, S. Toward a more general theory of regulation. J. Law Econ. **1974**, *19*, 211–240.
64. Meier, K.J. *The Political Economy of Regulation: The Case of Insurance*; SUNY Press: Albany, NY, 1988.
65. Derthick, M.; Quirk, P.J. *The Politics of Deregulation;* Brookings Institution: Washington, DC, 1985.
66. Truman, D. *The Governmental Process*; Knopf: New York, 1951.
67. Dahl, R. *A Preface to Democratic Theory*; University of Chicago Press: Chicago, IL, 1956.
68. Verba, S.; Nie, N.; Kim, J. *The Modes of Democratic Participation: A Cross-National Comparison;* Sage Publications: Beverly Hills, CA, 1971.
69. Bachrach, P.; Baratz, M.S. Two faces of power. Am. Polit. Sci. Rev. **1962**, *56* (4), 947–952.
70. Becker, G. A theory of competition among pressure groups for political influence. Q. J. Econ. **1983**, *98* (3), 371–400.
71. Wilson, J.Q. *The Politics of Regulation;* Basic Books: New York, 1980.
72. Bernstein, M. *Regulating Business by Independent Commission*; Princeton University Press: Princeton, NJ, 1955.
73. Posner, R. Theories of economic regulation. Bell J. Econ. Manage. Sci. **1974**, *5* (3), 337–352.
74. Aggarwal, V.K.; Keohane, R.O.; Yoffie, D.B. The dynamics of negotiated protectionism. Am. Polit. Sci. Rev. **1987**, *81* (2).
75. Wilson, J.Q. *Bureaucracy: What Government Agencies Do and Why They Do It*; Basic Books: New York, 1989.
76. Ringquist, E. *Environmental Protection at the State Level: Politics and Progress in Controlling Pollution*; M.E. Sharpe: Armonk, NY, 1993.
77. Rogers, E.M. *Diffusion of Innovations,* 5th Ed.; The Free Press: New York, 2003.
78. Simmons, B.A.; Dobbin, F.; Garrett, G. Introduction: The international diffusion of liberalism. Int. Organ. **2006**, *60* (4), 781–810.
79. Berry, W.D.; Baybeck, B. Using geographic information systems to study interstate competition. Am. Polit. Sci. Rev. **2005**, *99* (4), 505–519.

80. Boehmke, F.J.; Witmer, R. Disentangling diffusion: The effects of social learning and economic competition on state policy innovation and expansion. Polit. Res. Q. **2004**, *57* (1), 39–51.
81. Weyland, K. Theories of policy diffusion: Lessons from Latin American pension reform. World Polit. **2005**, *57*, 262–295.
82. Weyland, K. *Bounded Rationality and Policy Diffusion*; Princeton University Press: Princeton, NJ, 2007.

31

Food Quality Protection Act

Christina D. DiFonzo

Tolerances

FQPA fundamentally changed the way EPA sets tolerances for pesticide residues in food. EPA must review all (nearly 10,000) pesticide tolerances under new FQPA guidelines. The tolerance assessment schedule developed by EPA called for examining 33% within 3 years after August 1996, 66% within 6 years, and 100% within 10 years. EPA initially took a "worst-first" approach, to review the pesticides it considered to be of greatest risk, particularly to children, by August 1999. Three major pesticide groups, organophosphates (OPs), carbamates, and probable human carcinogens (B2s), were targeted under the worst-first approach. OPs and carbamates, the majority of which are insecticides, are neurotoxins structurally related to nerve gas. They affect the enzyme acetylcholinesterase in animals, including humans. B2 carcinogens are pesticides classified by EPA as having sufficient evidence for causing cancer in lab animals (usually at very high dose levels), but human evidence is lacking. Several important fungicides, plus a few herbicides and insecticides, are classified in this category.

Before FQPA, a single tolerance was established for each pesticide/crop combination, based only on dietary exposure to pesticide residue. Under FQPA, EPA must consider the combined (aggregate) exposure to a pesticide through dietary, drinking water, and nondietary sources (for example, structural, turf, garden, and pet uses) as well as the cumulative exposure to related pesticides with a common mechanism of toxicity. Furthermore, FQPA directs EPA to consider sensitive subpopulations, especially children, when setting tolerances. To insure that sensitive groups are adequately protected, EPA can require a safety factor of up to 10-fold on existing tolerances.

The Risk Cup

An analogy of a "risk cup" is used by EPA to explain changes in the establishment of tolerances under FQPA. Before FQPA, there was a separate risk cup for each pesticide/crop combination, containing only dietary exposure to residue. FQPA creates a separate risk cup for *each group* of related pesticides with common toxicity. Multiple pesticides, as well as multiple residues from all sources—food, water, and nonfood—of each pesticide, go into the same cup. Under this scenario, the cup gets crowded, and individual tolerances for each pesticide/crop combination in the group must get smaller. Furthermore, safety factors for children may reduce the overall size of each cup, potentially by a factor of 10.

Endocrine Disruption

Under FQPA, all pesticides and pesticide additives must be tested for effects on the endocrine system. This may require in vitro and in vivo screening for three different types of endocrine effects: estrogenic (mimics or blocks estrogen), androgenic (mimics or blocks androgens), and thyroid. Of the potential targets of a screening program, these three hormone groups are important in human development, are fairly well studied, and some laboratory methodology is already available to detect changes in level and function. Estimates are that up to 70,000 pesticides and other chemicals will be screened under FQPA and a second law, the Safe Drinking Water Act.

Consumer Right-to-Know

Another issue addressed in FQPA is consumer right-to-know about pesticide residues in food. FQPA mandated that EPA create a brochure to inform consumers about pesticide risks and benefits, and ways to remove residues from food they purchase. The brochure was completed and distributed to supermarkets in early 1999. However, FQPA did not mandate that stores actually display the publication.

Potential Impacts of FQPA

Pesticides that do not meet FQPA standards must either be mitigated (use patterns changed) or eliminated (some or all uses dropped). Thus, as FQPA is implemented, it potentially will have a tremendous impact on American agriculture.

- Changes in labeling or use patterns (number, frequency, and timing of applications) of pesticides to mitigate residue.
- Loss of critical pesticide uses, particularly for so-called minor (specialty) crops. These commodities represent smaller markets for pesticide manufacturers and thus are often “expendable.”
- Increases in production costs. Traditional broad-spectrum products might be replaced by more expensive, reduced-risk alternatives that control a narrower range of pests.
- Increased complexity of production and pest management systems. Broad-spectrum pesticides will be replaced by narrower spectrum tactics that require better knowledge and more intense management of the production system on the part of the producer.
- Potential for pesticide resistance. Loss of certain classes of pesticides could lead to resistance to remaining products, which are being relied on too heavily.

Bibliography

1. Colborn, T.; Dumanoski, D.; Myers, J. P. *Our Stolen Future*; Penguin Books U.S.A.: New York, 1996; 316.
2. *Public Law 104–170 to Amend the Federal Insecticide, Fungicide, and Rodenticide Act and the Federal Food, Drug, and Cosmetic Act, and for Other Purposes,* H.R. 1627; Federal Register: Washington, DC, Aug. 3, 1996.
3. Proceedings National Pesticide Impact Assessment Program Workshop, USDA CSREES NAPIAP Program, Sacramento, CA, May 5–7, 1998; Melnicoe, R., Ed.: Washington, DC, 1998; 76.
4. National Research Council. *Pesticides in the Diets of Infants and Children*; National Academy Press: Washington, DC, 1993; 386.
5. U.S. Environmental Protection Agency, Office of Pesticide Programs. *The Food Quality Protection Act (FQPA) of 1996.* http://www.epa.gov/oppfead1/fqpa/

32

Food: Cosmetic Standards

David Pimentel
and Kelsey Hart

Introduction

The American marketplace features nearly perfect fruits and vegetables. Gone are apples with an occasional blemish and fresh spinach with a leaf miner. This increase in the "cosmetic standards" of fruits and vegetables has resulted from the efforts of the Food and Drug Administration (FDA) and the U.S. Department of Agriculture (USDA) to limit the levels of insects and mites in produce, and new standards established by food wholesalers, processors, and retailers. Meeting more stringent standards has led to significant increases in the amounts and toxicity of pesticides used in crops. Increased pesticide use has negative environmental and public health consequences. In comparison, the health risks from consuming herbivorous insects/insect parts in food do not exist and certainly do not justify the increase in pesticide use and the associated problems. Recent research indicates that pesticide use can be reduced by 35% to 50% without any substantial increase in food prices or loss of crop yields.[1] Surveys suggest that the public would support relaxation of cosmetic standards if it decreases pesticide residues in its food.

History of Cosmetic Standards

The FDA sets defect action levels (DALs) for insects and mites allowed in fruits and vegetables and in products made from them. These DALs were established to reduce insect and mite infestation in foods to a reasonable and safe level, because their presence in food products was thought to indicate that crops had insufficient insect and mite control, were improperly washed, were unsatisfactorily inspected, and contained insects and mites harmful to health. Besides visual prejudice against insects in food, there is the well- placed concern that insects such as nonherbivorous houseflies and cockroaches may transmit disease.

During the past 40 years, the FDA has steadily lowered DALs.[2] For example, a fivefold decrease in the number of leaf miners permitted in spinach occurred from 1930 to 1974. As tolerance levels for insects in food have fallen, wholesalers, processors, and retailers have increased their "cosmetic standards" for produce and other food products so that most marketed U.S. produce is visually perfect. Produce distributors encourage high cosmetic standards because their contracts enable them to visually inspect produce before buying and reject it when the supply is excessive. Growers are motivated to produce cosmetically perfect produce to ensure its sale.

However, to meet these increasingly stringent regulations, farmers have had to use greater amounts of increasingly toxic pesticides and implement other pest control strategies. Synthetic pesticide use in

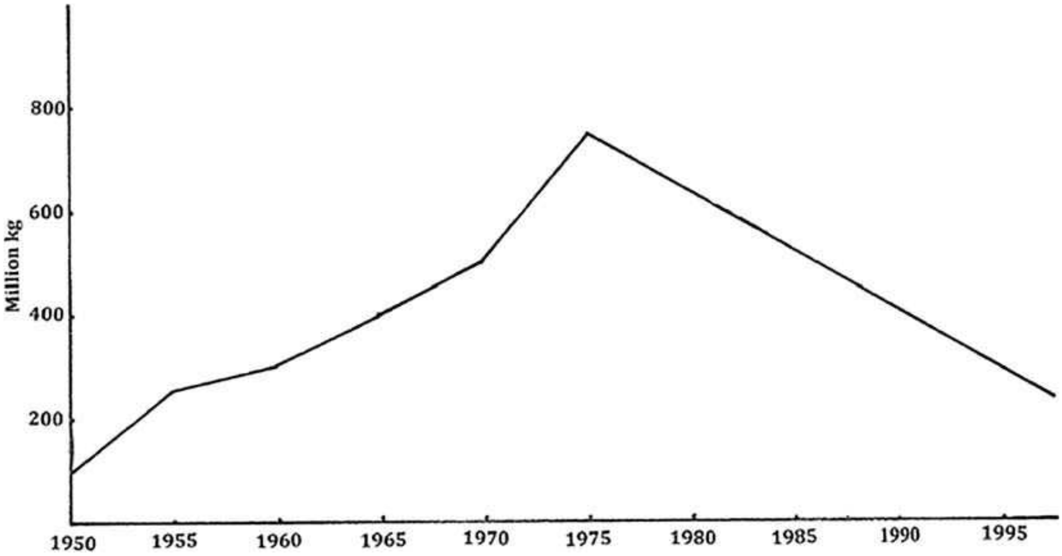


FIGURE 1 The amount of synthetic pesticides produced in the United States. About 90% is sold in the United States. The decline in total amount produced since 1975 is in large part due to the 10- to 100-fold increased toxicity and effectiveness of the newer pesticides.[1,3]

the United States has increased about 33-fold since 1945, and the toxicity of pesticides used has increased 10- to 100-fold in the past 25 years (Figure 1). More pesticides need to be used to produce the blemish-free produce distributors and consumers expect.

There is little evidence that eating herbivorous insects or insect parts is hazardous to human health.[4] However, solid data suggest that the adverse health and environmental impacts of pesticide exposure are substantial.[5] Given the direct correlation between increases in cosmetic standards and increases in pesticide use, why are cosmetic standards and DALs growing increasingly severe and perpetuating further increases in pesticide use?

The increase in cosmetic standards and more stringent DALs are based on the premise that consumers demand unblemished, insect-free food. Clearly, cosmetic appearance of produce is a primary factor consumers use in assessing the quality of produce. Unfortunately, this assessment is often made without more substantive quality information, such as nutritional values or pesticide residue levels. Recent evidence suggests that when consumers are aware of the trade-offs between blemish-free produce and pesticide use, they will purchase produce that is not cosmetically perfect because it has less pesticide residue.[6]

Environmental and Health Effects of Pesticide Exposure

An estimated 617,000 tons of more than 600 different kinds of pesticides are used annually in the United States, at a cost of approximately $9 billion.[7] Still, pests such as insects, plant pathogens, and weeds destroy 37% of all potential food and fiber crops.[8] Typically, each dollar invested in pesticides returns about $4 in crops saved.

However, this economic evaluation does not take into account the impacts of pesticide use on public health and the environment. Approximately 0.1% of applied pesticides reach target pests, leaving 99.9% of the pesticides to impact the environment.[8] Environmental effects of pesticides can be significant: Domestic animals and wildlife can be poisoned or adversely affected by pesticide exposure; beneficial natural enemies of harmful pests can be destroyed by pesticide use; heavy pesticide use can result in pesticide resistance and subsequently even heavier or more toxic pesticide use; and already limited natural resources such as soil, groundwater, and surface water can be contaminated by pesticide residues or drift.[5]

The human health effects of pesticide exposure through food are also diverse and significant. About 35% of foods purchased by American consumers have detectable levels of pesticides, and about 1%–3% of these foods have residue levels over the legal tolerance level 8. These estimates are conservative because detection methods currently detect only about one-third of the pesticides now in use in the United States. The contamination rate is undoubtedly higher for fruits and vegetables because they receive the highest levels of pesticides. One USDA study indicates that some pesticide residue remains in produce even after it is washed, peeled, and cored.[9]

Both the acute and chronic health effects of pesticide exposure are significant. Worldwide, about 26.5 million acute pesticide poisonings occur each year, resulting in about 3 million hospitalizations, approximately 220,000 fatalities, and 750,000 cases of chronic pesticide-related illness.[10] Chronic effects can adversely affect most systems of the human body. U.S. data indicate that 18% of all insecticides and about 90% of all fungicides are carcinogenic.[11] Many pesticides are also estrogenic, linked to increased breast cancer among some women in the United States. Pesticide exposure can also damage the respiratory and reproductive systems, leading to conditions like asthma and infertility.[12–14] In the United States, EPA[15] reports that 300,000 non-fatal pesticide poisonings occur each year.

The negative health effects that pesticides can have are more significant in children. Children have higher metabolic rates than adults, and their ability to detoxify and excrete toxic compounds is different. Also, because of their smaller size, children are typically exposed to higher levels of pesticides than adults. Finally, certain types of pesticides, such as carbamates and organophosphates, are more dangerous for children than adults.[10]

Given the significant environmental and public health impacts that pesticides can have, it appears desirable to limit pesticide exposure to minimize these adverse effects. However, the increasingly stringent DALs and cosmetic standards have resulted in considerable increases in pesticide use. Do the health effects of eating herbivorous insects, insect parts, or blemished produce warrant the risks and the substantial health consequences of increasing pesticide exposure to meet these standards?

Health Effects of Eating Insects/Insect Parts in Food

Even under the current stringent DAL regulatory guidelines, a few insects and mites do remain in or on produce. For instance, the DAL for apple butter is an "average of 5 whole insects or equivalents per 100 grams not counting mites, aphids, thrips, or scale insects." DALs for many other food products are similar. Many insects commonly found in foods and food products are so minute in size that they are practically impossible to eliminate. Although the numbers of insects are strictly limited by FDA regulations, some do remain and are eaten.

This, however, is not a cause for concern. In contrast with the well-documented acute and chronic negative health effects resulting from pesticide exposure, there is not one known case of human illness from ingesting insects and mites in or on foods. In addition, though some insects do carry disease or present health risks—houseflies, for example—all herbivorous insects/mites found on harvested produce are harmless to humans.

While ingesting insects or insect parts in our food may seem distasteful to many Americans, many cultures eat insects by choice. Insects are a substantial source of protein, with digestible protein content ranging from 40% to 65%,[16] and insects, shrimp, lobster, and crawfish are all arthropods; the latter three are often considered food delicacies.

Given that herbivorous insects found on produce are not a health hazard, consumers must decide whether they are willing to tolerate the presence of a few insects rather than insisting on visually "perfect" produce that requires high levels of pesticides.

Conclusions and Future Directions

If the health effects of herbivorous insects found on or in food products are not cause for concern, then the need for strict DALs might be relaxed. Relaxing cosmetic standards for some fruits and vegetables might be feasible. Approximately 10% to 20% of pesticides applied to fruits and vegetables are used only to comply with the current strict cosmetic standards established by the FDA, USDA, wholesalers, and retailers that result in blemish-free pro-duce. Rigorous cosmetic standards are probably unnecessary, since surface blemishes on fruits and vegetables generally do not affect nutritional content, storage life, or flavor. However, will the American public purchase produce that appears less than perfect?

Research on public preferences shows that 97% of Americans prefer food without pesticide residues. In addition, 50%–66% are willing to pay more for food with less pesticide residue.[17] It is estimated that in the United States, pesticide use can be reduced by about 50% without reducing crop yields. The estimated increase in the consumer's food costs would be only 0.6%.[10] This marketplace cost increase does not take into account the positive environmental and health benefits that would be realized if pesticide use were reduced. Sweden, for example, has reduced pesticide use 68% and reduced pesticide poisonings 77%.[18] The small increase in consumer cost would be more than offset by these benefits.

Therefore, given the environmental and health tradeoffs related to high cosmetic standards for produce, it appears that human health and the environment would be best protected by less stringent DALs and relaxed cosmetic standards for produce, to minimize unnecessary pesticide use and related adverse effects.

References

1. Pimentel, D.; Kirby, C.; Shroff, A. The relationship between "cosmetic standards" for foods and pesticide use. In *The Pesticide Question: Environment, Economics, and Ethics;* Pimentel, D., Lehman, H. Eds.; Chapman and Hall: New York, 1993; 85–105.
2. FDA. *The Food Defect Action Levels: Levels of Natural or Unavoidable Defects in Foods That Present No Health Hazards for Humans;* ed.; Department of Health and Human Services, Public Health Service, Food and Drug Administration, Center for Food Safety and Applied Nutrition: Washington, DC, 1995.
3. KRS Network. *U.S. Pesticide Industry Report Executive Summary.* Covington, GA, 2005; 15 pp. available at http://www.knowtify.net/2005USPestIndReptExecSum.pdf.
4. Defoliart, G.R. The human use of insects as food and as animal feed. Bull. Entomol. Soc. Am. 1986, *35,* 22–35.
5. Pimentel, D. Environmental and economic costs of the application of pesticides primarily in the United States. Environ. Dev. Sustainability 2005, *7,* 229–252.
6. Pimentel, D.; Terhune, E.; Dritschilo, W.; Gallahan, D.; Kinner, N.; Nafus, D.; Peterson, R.; Zareh, N.; Misiti, J.; Haber-Schaim, O. Pesticides, insects in food, and cosmetic standards. BioScience 1977, *27,* 178–185.
7. Environmental Protection Agency. Pesticide Market Estimates: Usage, 2000–2001; 3; available at http://www.epa.gov/opp bead1/pestsales/usage2001.htm (accessed February 11, 2010).
8. Pimentel, D.; Greiner, A. Environmental and socioeconomic costs of pesticide use. In *Techniques for Reducing Pesticide Use*; Pimentel, D., Ed.; John Wiley and Sons: Chichester, U.K., 1997; 51–78.
9. Wiles, R.; Campbell, C. *Washed, Peeled—Contaminated: Pesticide Residues in Ready-To-Eat Fruits and Vegetables;* Environmental Working Group: Washington, DC, 1994.
10. Pimentel, D.; Hart, K.A. Pesticide use: ethical, environmental, and public health implications. In *New Dimensions in Bioethics: Science, Ethics and the Formation of Public Policy;* Galston, A.; Shurr, E., Eds.; Kluwer Academic Publishers: Boston, 2001; 79–108.
11. Eastmond, D.A. Mechanisms of Carcinogenesis of the Fungicide and Rat Bladder Carcinogen o-Phenylphenol. EPA Grant No.: R826408 Project Period: May 1, 1998 through April 30, 2001; 5; available at http://cfpub.epa.gov/ncer_abstracts/index.cfm/fuseaction/display.abstractDetail/abstract/876.

12. Weisenburger, D.D. Human health effects of agrichemical use. Hum. Pathol. 1993, *24,* 571–576.
13. Schneider, E.P.; Dickert, K.J. Health costs and benefits of fungicide use in agriculture. J. Agromed. 1993, *1,* 19–37.
14. Yang, L.; Kemadjou, J.R.; Zinsmeister, C.; Bauer, M.; Legradi, J.; Muller, F.; Pankratz, M.; Jakel, J.; Strahle, U. Transcriptional profiling reveals barcode-like toxicoge-nomic responses in the zebrafish embryo. Genome Biol. 2007, *8,* R227.
15. Environmental Protection Agency. *Hired Farm Workers Health and Well-Being at Risk*. United States General Accounting Office Report to Congressional Requesters. February, 1992.
16. Gorham, J.R. Foodborne filth and human disease. J. Food Prot. 1989, *52*, 674–677.
17. Anon. Appearance of produce versus pesticide use. Chem-ecology 1991, *20* (4), 11.
18. Pesticide News. Persistence pays—lower risks from pesticides in Sweden. Pesticide News *34,* 10–11; available at http://www.pan-uk.org/pestnews/Issue/pn54/pn54p10.htm (accessed February 11, 2010).

33

Laws and Regulations: Food

Ike Jeon

Tolerance Limits for Pesticide Residues

The responsibility for ensuring that pesticide residues in foods are not present above the limits is shared by three major government agencies.[1] The Environment Protection Agency (EPA) determines the safety of pesticide products and sets tolerance levels for pesticides. The Food and Drug Administration (FDA) enforces the tolerances in all foods except meat and poultry products. The U.S. Department of Agriculture's Food Safety and Inspection Service (FSIS) regulates commercially processed egg, meat, and poultry products including combination products (e.g., stew, pizza). In addition, any products containing 2% or more poultry or poultry products, or 3% or more red meat or red meat products are also under jurisdiction of the FSIS. The pesticides of concern usually include insecticides, fungicides, herbicides, and other agricultural chemicals. Table 1 illustrates examples of tolerance levels for pesticide residues in several food categories.[2,3] These tolerance levels are extremely low, usually below parts per million, but do not represent permissible levels of contamination where it is avoidable. In addition, blending of a food (or feed) containing a substance in excess of an action level or tolerance with another food (or feed) is not permitted, and the final product from blending is unlawful, regardless of the level of the contaminant.

Regulatory Inspection and Enforcement

The FDA monitors the levels of pesticide residues in processed foods. For imported products, the FDA checks a sample of the food at entry into the United States and can stop shipments at the entry. If illegal residues are found in domestic samples, FDA can take regulatory actions, such as seizure or injunction.

The U.S. Department of Agriculture also monitors pesticide residues in food.[4] The Department was charged in 1991 with implementing a program to collect data on pesticide residues on various food commodities. The program has become a critical component of the Food Quality Protection Act of 1996 and currently is known as the Pesticide Data Program. The data on pesticides in selected commodities are used by the EPA to support its dietary risk assessment process and pesticide registration and by the FDA to refine sampling for enforcement of tolerances.

If a product is in violation of the tolerance limits, it is *adulterated* under the food law. The product may be destroyed or recalled from the market by the manufacturer or shipper. The recall may be initiated voluntarily by the manufacturer (or shipper) or at the request of the regulatory agency. The responsible agency also may seize the product on orders obtained from the Federal courts and may prosecute persons or firms responsible for the violation.

Tolerance Limits for Insect Fragments

Many food materials may contain natural but unwanted debris that cause no health hazards for humans. These debris may include insects, insect fragments, and rodent hairs and are considered unavoidable defects in foods with the current agricultural practices. In fact, the use of chemical substances to control insects, rodent, and other contaminants has little, if any, impact on natural and unavoidable defects in foods. The FDA contends that the use of pesticides does not effectively reduce the presence of these food defects. This has led the regulatory agencies to establish maximum levels of natural or unavoidable defects allowable in foods for human use. The FDA currently lists over 100 products from fruits to fish,[5] and Table 2 shows only several examples. If no defect action level exists for a product, the FDA evaluates and decides on a case-by-case basis using criteria of reported findings such as length of hairs and size of insect fragments.

The FDA sets these action levels under the premise that it is economically impractical to grow, harvest, or process raw products that are totally free of nonhazardous, naturally occurring, unavoidable defects. It is incorrect, however, to assume that because the FDA has an established defect action level for a food, the manufacturer needs only keep defects just below that level. The defect levels do not represent averages of the defects that occur in any of the products. The levels represent limits at which FDA will regard the food product as *adulterated* and, therefore, subject to enforcement action. Like pesticide residues, blending of food with a defect at or above the current defect action level with

TABLE 1 Examples of Tolerance Limits for Pesticide Residues in Human Food

		Action Level	
Substance	Commodity	(Parts per Million)	Remark
Aldrin and dieldrin	Asparagus	0.03	
	Fish	0.3	Edible portion
	Peanuts	0.05	
Chlordane	Carrots	0.1	
	Fish	0.3	Edible portion
	Lettuce	0.1	
	Poultry	0.3	Fat basis
DDT[a]	Carrots	3.0	
	Citrus fruits	0.1	
	Tomatoes	0.05	
Lindane	Beans	0.5	
	Corn	0.1	
	Milk	0.3	Fat basis
	Beef	7.0	Fat basis

Source: FDA[2] and USDA.[3]

[a] Dichlorodiphenyltrichloroethane.

another lot of the same or another food is not permitted. That practice renders the final food unlawful regardless of the defect level of the finished food.

Responsibility of Food Manufacturers

Food manufacturers are required to follow the standard manufacturing procedures under a federal regulation, known as good manufacturing practice (GMP), during food production.[6] The GMP guidelines imply that all food materials used must not exceed the tolerance limits set for pesticide residues or any other poisonous or deleterious substances. The GMP also calls for the same regulatory requirement for natural or unavoidable defects in all food materials. The food materials susceptible to contamination may be tested for compliance or relied on a supplier's guarantee or certification that they are in compliance. In addition, the GMP regulation stipulates that food manufacturers and distributors must utilize at all times quality control operations that reduce natural or unavoidable defects to the lowest level feasible with the current technology.

Potential Consumer Benefits

Through conducting a monitoring program, the federal government agencies work together to improve consumer protection. The EPA will continue to review scientific data on all pesticide products, while the FDA and U.S. Department of Agriculture will closely monitor levels of pesticide residues in all foods including both domestic and imported products. The U.S. Department of Agriculture's data for 1998 suggest that violation of the pesticide tolerance limits was very low in all raw products including fruit and vege, wheat, and milk samples. In 1993, the FDA reported that no pesticide residues were found in infant formulas, and no residues over EPA tolerances or FDA action levels were found in any of the foods that were prepared as consumers normally would prepare them at home.[7]

Acknowledgments

Contribution No. 00–231-B, Kansas Agricultural Experiment Station, Manhattan, Kansas 66506, U.S.A.

TABLE 2 Examples of Tolerance Limits for Natural or Unavoidable Defects in Foods

Product	Defect	Action Level
Sweet corn, canned	Insect larvae	2 or more 3 mm or longer larvae
Macaroni	Insect filth	225 insect fragments or more per 225 g
	Rodent filth	4.5 rodent hairs or more per 225 g
Peaches, canned and frozen	Mold/insect damage	Wormy or moldy on 3% or more fruits
	Insects	1 or more larvae and/or larval fragments whose aggregate length exceeds 5 mm in 12 one-pound cans
Peanut butter	Insect filth	30 or more insect fragments per 100 g
	Rodent filth	1 or more rodent hairs per 100 g
Popcorn	Rodent filth	1 or more rodent excreta pellets or rodent hairs in 1 or more subsamples
Tomato juice	Drosophila fly	10 or more fly eggs per 100 g
	Mold	24% of mold counts in 6 subsamples
Wheat flour	Insect filth	75 or more insect fragments per 50 g
	Rodent filth	1 or more rodent hairs per 50 g

Source: FDA.[5]

References

1. FDA. *FDA's Food and Cosmetic Regulatory Responsibilities*; U.S. Food and Drug Administration: Washington, DC, 1998; 1–5, http://vm.cfsan.fda.gov/~dms/regresp.html (accessed June 2000).
2. FDA. *Action Levels for Poisonous or Deleterious Substances in Human Food and Animal Feed*; U.S. Food and Drug Administration: Washington, DC, 1998; 1–17, http://vm.cfsan.fda.gov/~lrd/fdaact.html (accessed June 2000).
3. USDA. *Domestic Residue Book (Appendix I)*; U.S. Department of Agriculture, Food Safety and Inspection Service: Washington, DC, 1998; 1–30, http://www.fsis.usda.gov:80/OPHS/redbook1/appndx1.htm (accessed June 2000).
4. USDA. *Pesticide Data Program Annual Summary—Calendar Year of 1998;* U.S. Department of Agriculture, Agricultural Marketing Service: Washington, DC, 2000; 1–19.
5. FDA. The Food Defect Action Levels—Levels of Natural or Unavoidable Defects in Foods that Present No Health Hazards for Humans. In *FDA/CFSAN Food Defect Action Level Handbook;* U.S. Food and Drug Administration: Washington, DC, 1998; 1–36, http://vm.cfsan.fda.gov/~dms/dalbook.html (accessed June 2000).
6. CFR. Current good manufacturing practice in Manufacturing, Packing, or Holding Human Food. In *Code of Federal Regulations, Title 21, Part 110;* U.S. Government Printing Office: Washington, DC, 1999; 206–215.
7. FDA. *FDA Reports on Pesticides in Foods*; U.S. Food and Drug Administration: Washington, DC, 1993; 1–5, http://vm.cfsan.fda.gov/~lrd/pesticid.html (accessed June 2000).

34

Laws and Regulations: Pesticides

Praful Suchak

Introduction

Why Regulate Pesticides?

Chemical or biological pesticides have target specific toxicity that controls or eradicates pests falling under different groups. These products, though developed for specific usage, could have adverse effects on living beings and the environment and unchecked use can cause havoc. Regulating pesticides, therefore, would assure reasonable safety in use of these toxic substances and ensure that risks from pesticides to humans and their environment are minimized and are consistent with the benefits achieved by their use in terms of reduced losses.

Regulating pesticides at the international and national level should consider social costs in line with social benefits. Pesticides impose costs on society, such as health risks and environmental degradation, which are not borne by the user. The available policy remedies include bans on individual or classes of chemicals that prohibit the introduction of hazardous compounds into the environment, and economic instruments such as taxes, registration fees, and import duties that work to redistribute and adjust the social costs occurring for pesticide use and also provide the government with revenues that can be used to cover health costs and environmental clean-up activities.

History

The United States in 1910 introduced the Federal Pesticide Act that underwent complete metamorphosis to become the Federal Insecticide, Fungicide and Rodenticide Act (FIFRA) in 1947, which since 1970 is under the auspices of the Environmental Protection Agency.

Australia initiated pesticide legislation with one state in 1925 and by 1945 all states had their individual laws. The Industry Association brought law common to all states in 1995. By the end of 1999 about 95% of the countries in the world had adopted full/partial regulatory systems.

Early in-depth studies were not carried out on the long-term effects of: 1) repeated exposures, 2) residual toxicity, 3) accumulated toxicity, and 4) the impact on environment. With additional knowledge on the cumulative toxicity of chlorinated hydrocarbons such as DDT having come to light, the regulating authorities have started demanding the generation of additional critical toxicological data to assess short-term, long-term, and environmental toxicity of earlier registered pesticides. The European Union has already undertaken reviews of 90 molecules in the first phase by a Commission regulation dated December 11, 1992, to be completed in 12 years, and a further 148 molecules in the second phase effective March 1, 2000. The remaining substances in the European Union would be included in third phase.

Regulatory requirements for pesticides have undergone a change over the past half a century. With the advent of highly sophisticated testing equipment, more knowledge about harmful effects of the toxic chemicals has come to light. Consistent watch by environmentalists and organizations like the Pesticides Action Network (PAN), Greenpeace, Save the Planet groups, and other nongovernmental organizations has resulted in added awareness resulting in hosts of data requirements for registration/reregistration of pesticides.

Although all developed countries and most of the developing countries have their own legislation to regulate pesticides, there have been vast variations in data requirements for registrations between these countries. With globalization it has became imperative to have harmonized data requirements so that the registrant can hope for faster registration in different (pesticide consuming) countries.

Available International Guidelines

1. Agenda 21 of the United Nations Conference on Environment and Development (UNCED)
2. The *Codex Alimentarius*
3. The FAO International Code of Conduct and Prior Informed Consent (PIC)
4. WTO and International Trade with respect to pesticides
5. Agreement on Persistent Organic Pollution (POP)
6. Guidelines of Minor Donor Institutions on the purchase of pesticidess

Implementation Problems

Although FAO took the lead to harmonize data requirements in participating nations for registrations of pesticides, certain problems and practical difficulties have occurred such as

7. The original registrant, having invested huge amounts in data generation, is unable to protect the data
8. Absence of confidentiality assurance by the registering country, creating difficulties in multiple country registrations
9. Recommended uses differ from country to country, resulting in difficulties
10. Unchecked dumping of unsafe or banned pesticides in less-developed countries
11. New registrations by a company other than the original registrant by providing data generated by such a company could not be checked

Steps Undertaken

Although though PIC entry of banned pesticides could be prevented, this instrument has not been fully effective. Once it becomes fully operational legally things should improve.

With the United States implementing the Food Quality Protection Act and fixing maximum residue limits for 3000 toxic compounds, countries worldwide would need to harmonize their registrations on toxic chemicals so as to meet the residue levels in food.

The formation of the European Union with 15 member countries, OECD with 29 members, and the Technical Working Group having EPA, Canada, and Mexico, has accelerated the pace towards harmonization. However, since a vast disparity exists between developed countries on one side and developing countries on the other side, it is rather difficult to have a unified data requirement, particularly in case of risk assessment.

Acceptance of electronic data submission and dossier/ monograph submissions and joint reviews by EU would also pave the way toward harmonization and would address questions in the nondietary exposure area.

Apart from studies related to bioefficiency of the product, the toxicological studies of the toxicant, its analogues, impurities and breakdown products, residual toxicity, etc., as listed in Appendix 1 would help understanding and regulating pesticides.

Present Scenario and Probable Remedies

Substantial evidence exists that pesticides are being applied in a technically and economically inefficient manner. Many developing countries subsidize pesticides and equipment, resulting in excessive use of pesticides.

Also in developing countries, the current legal environment and enforcement capabilities have been inadequate and dysfunctional, thus exerting a significant impact on current levels of pesticide use. This is partly due to lack of resources and partly due to manipulation by vested interests.

The inadequacies of the existing regulatory framework, institutional rigidities, and a bias in favor of pesticide-dependent paths also contribute to improper use of pesticides.

A major problem confronting many countries is the absence of well-established procedural mechanisms for public involvement in the decision making process including crop protection policy. Competing interests with a stake in the process, including farmers, the pesticide industry, and policy makers responsible for food security, argue for a more liberal regulatory stance. On the other hand, environmentalists, public health workers, and consumers demand strict regulation and reduced pesticide volumes.

To be more effective, pesticide regulation and implementation should be handled by a neutral agency like the Ministry of Environment or similar organization and not the Ministry of Agriculture or other interested ministry.

Pesticide policy needs to be integrated into the broader public policy debate concerning the nations' agricultural, environmental, and health strategies.

Nevertheless, two general principles should apply. First, dispassionate analysis of the costs and benefits of pesticide use would provide a useful tool for the formulation of normal policies; and second, the broader and more inclusive the debate, the more likely it is that the outcome will serve the public rather than specific private interests.

Future Global Policy

A uniform global regulatory system needs to ensure

1. Agricultural chemical use increases agricultural output
2. Food supplies are safe from harmful toxicants/ residues
3. Reduced-risk chemical pesticides, biopesticides, and nonchemical alternatives are encouraged
4. Uniform MRLs to eliminate trade barriers
5. Uniform health-based safety standards for pesticide residues
6. Special provisions for certain groups of the population including infants and children

Appendix 1

Toxicological and Other Data Requirements for Pesticide Registration

7. Identity of active substance
   Chemical name
   Empirical and structural formula
   Molecular mass
   Method of manufacture (synthesis pathways)
   Purity
   Identity and content of isomers
   Impurity and additives

8. Physical and chemical properties
   Melting point
   Boiling point and relative density
   Vapor pressure
   Volatility
   Appearance
   Absorption spectra-molecular extinction at relevant wavelength
   Solubility in water/organic solvents
   Partitioning coefficient N-octanol/water
   Stability and hydrolysis rate in water
   Photochemical degradation on surface, in water, and in air
   Thermal stability and stability in air

9. Analytical method
   Analytical method for the determination of the pure active substance in the technical grade.
   For breakdown products and additives in plant products, soil, water, animal body fluids, and tissues.

10. Toxicological and metabolism studies
    Studies on acute toxicity—oral, percutaneous, inhalation, intraperitoneal, skin and, where appropriate, eye irritation, and skin sensitization.
    Short-term toxicity—oral, cumulative toxicity, and other routes inhalation or dermal.
    Chronic toxicity—oral, long-term toxicity, and carcinogenicity.
    Mutagenicity—reproductive toxicity-teratogenicity and multigeneration studies in mammals. Metabolism studies in mammals—absorption, distribution, and excretion studies, elucidation of metabolic pathways.
    Supplementary studies—neurotoxicity studies— toxic effects of metabolites from treated plants and toxic effects on livestock and pests.
    Medical data—medical surveillance on manufacturing plant personnel, clinical cases, poisoning incidents from industry and agriculture sensitization/ allergenicity observations, observations on exposure of the general population, and epidemiological studies if appropriate. Diagnosis and specific signs of poisoning, clinical tests, and prognosis of expected effects of poisoning. Proposed treatment: first aid measures, antidotes, and medical treatment. Summary of toxicological studies and conclusions, critical scientific evaluation with regard to all toxicological data, and other information concerning the active substance.

11. Residues in or on treated products, food and feed metabolism in plants and livestock
    In treated plants (distribution, metabolism, binding constituents, etc.).
    In livestock (uptake, distribution, metabolism, binding constituents, etc.).

12. Fate and behavior in the environment
    Studies on aerobic and anaerobic degradation under laboratory conditions in different soil types. Adsorption and desorption in different soil types including metabolites.
    Mobility of the active ingredients in different soil types.
    Behavior in water and air, rate and route of degradation.
13. Ecotoxicological studies
    Effects on birds, fish, aquatic organisms such as Daphnia magna, algae, honeybees, earthworms, other nontarget macroorganisms and microorganisms.
14. Information concerning the labeling including indication of danger and safety measures.

Bibliography

1. Pesticides Policies in Developing Countries—Do They Encourage Excessive Use? In *World Bank Discussions Paper No. 238;* 1994.
2. Asian Development Bank. In *Handbook on the Use of Pesticides in Asia Pacific Region*; ADB: Manila, Philippines, 1987.
3. *Pesticide Policy Project Hannover*; Publication Serial No. 1, January 1995; No. 2, November 1995; No. 3, December 1995; No. 4, December 1996; No. 5, December 1996; No. 6, 1998; No. 7, April 1999; No. 8, April 1999.
4. EC Directives 91/414/EEC and Subsequent Directives Including 1999/80/EC.
5. Proceedings of Asia Pacific Crop Protection Conference 1997 and 1999, PMFAI: Mumbai, India.
6. *Global Pesticides Directory,* 2nd Ed.; Suchak's Consultancy Services: Mumbai, India, 1997. suchakgr@vsnl.net
7. *Pesticides News*; No. 20–47, Pesticides Action Network (PAN): London, 1993 to 1999.
8. Guidelines on the Operation of Prior Informed Consent (PIC) Rome FAO 1990, Guidance to Government in PIC Rome 1991, and Other FAO Publications.
9. U.S. EPA Pesticides Information Network. http://www.cdpr.ca.govt/docs/epa/epachim.htm.

35

Laws and Regulations: Rotterdam Convention

Barbara Dinham

Introduction

When chemical pesticides were introduced 50 years ago, little attention was paid to the environmental and health impacts. With the rapid expansion of use in the 1950s, understanding gradually increased of the consequences of exposure to certain chemicals. Wide-ranging impacts began to be identified, including: environmental persistence and effects on birds and wildlife; residues in soil, water, and air; residues in food; human poisonings from acutely toxic pesticides or long-term health impacts such as cancer; and pest resistance, often leading to dramatic crop losses.

With almost 1000 different pesticides and thousands of formulations on the market to control insects, diseases, weeds, and other pests, action was clearly needed to protect human health and the environment. International standards recommended that governments establish a registration system to authorize each formulation of a pesticide for each specific crop or other use. Concern with some pesticides led governments to ban or restrict them to a limited number of uses. Few developing countries can fully implement a registration scheme, and they are often unaware of bans imposed elsewhere. Recognizing these problems, in the early 1980s, governments, international organizations, and public interest groups began to demand action to provide a warning system to help developing countries regulate or ban the use of hazardous pesticides.

The Rotterdam Convention on Prior Informed Consent Procedure for Certain Hazardous Chemicals and Pesticides in International Trade[1] is the outcome of 15 years of activity on trade in hazardous chemicals. Adopted on 10 September 1998 in Rotterdam, the Netherlands, the Convention was signed by 73 countries[2] and by June 2001 had been ratified by 14 parties. It will become legally binding after 50 countries have ratified.

The Convention takes an important step toward protecting humans and the environment from highly toxic chemicals. For the first time, it will help monitor and control trade in dangerous substances, circulate better information about health and environmental problems of chemicals, and prevent unwanted imports of certain hazardous chemicals.

Central to the Rotterdam Convention is the system of Prior Informed Consent (PIC), a means of obtaining and disseminating decisions of importing countries about their willingness to receive shipments of certain chemicals, and ensuring compliance to these decisions by the exporter. To be included in PIC, a pesticide must be banned or severely restricted for health or environmental reasons by two countries in two different regions of the world—indicating that its adverse effects are a "global concern."

But focusing on banned or severely restricted pesticides may only touch the tip of the iceberg. Industrialized countries rely on trained and informed users able to apply good practice as safeguards: in developing countries where pesticides are often used under conditions of poverty, these measures cannot be applied. Furthermore, older—and often more hazardous—pesticides are often cheaper, making them attractive to poorer farmers. The Convention recognizes that "severely hazardous pesticide formulations" should be included in PIC if they cause health or environmental problems in developing countries or in Eastern Europe—termed "countries with economies in transition"— in the Convention.

History of PIC

A PIC system was first proposed in the early 1980s as part of the International Code of Conduct on the Distribution and Use of Pesticides, negotiated by governments in the Food and Agriculture Organization (FAO) of the UN. Some governments resisted the concept, and the Code was adopted in 1985 without any reference to PIC. But intense pressure from nongovernmental organizations (NGOs) and others won support, and the principle was accepted in 1987. It took until 1989 to establish the wording and issue a revised version of the Code.[3] That same year, the UN Environment Programme (UNEP) included an identical provision in the London Guidelines on the Exchange of Information on Chemicals in International Trade, and a voluntary system was put in place with the FAO acting as the Secretariat for pesticides and UNEP for industrial chemicals. The first pesticides were added in 1991, and by 1995, 22 pesticides and five industrial chemicals were included.

From Voluntary to Legally Binding

The issue of transforming the voluntary scheme into a legally binding international Convention was first mooted in 1992 at the United Nations Conference on Environment and Development (UNCED).[4] In November 1994, the FAO Council meeting agreed to proceed, and this was followed in May 1995 by a decision of the UNEP Governing Council. The two organizations convened an Intergovernmental Negotiating Committee (INC) to draft and agree international legally binding instrument.

Banning Exports of Banned Pesticides

An alternative to PIC strongly advocated at the time was to stop all exports of banned pesticides. However, unless action to limit the market for a banned pesticide could be taken, banning exports could encourage companies to relocate production, possibly in a country with less stringent controls. Preventing the export of banned pesticides would have no effect on severely restricted chemicals. Without a PIC system, a developing country could unwittingly allow the import of banned or severely restricted pesticides, ignorant of action taken by some governments. Many developing countries maintained that an export ban could limit their development, as alternatives were more expensive, and that import decisions should rest with them. PIC does not prevent individual countries from deciding that their banned pesticides should not be exported, but does ensure that regulatory actions are widely shared.

How the Convention Is Operated

In negotiating the text of the Rotterdam Convention, governments built on the experience gained in the voluntary PIC. As a mark of its importance, the Convention began immediately on a voluntary basis, with FAO and UNEP continuing as an interim Joint Secretariat.

Designated National Authorities

To participate in PIC, governments must appoint a Designated National Authority (DNA). By December 2000, 170 governments had appointed a DNA or a focal point. When ratifying the Convention, DNAs must be authorized to carry out administrative functions such as receiving, transmitting, and circulating information.

Notifying Regulatory Actions

When a government bans or severely restricts a pesticide, it must notify the Joint Secretariat within 90 days. Governments need to demonstrate that their action is final and that it was based on a risk evaluation, including a review of scientific data, and the Secretariat will validate the notification. Once two valid notifications from different PIC regions have been received for the same pesticide, it becomes a candidate for PIC.

Chemical Review Committee

The Convention set up a Chemical Review Committee to consider notifications, and advise the Conference of the Parties (CoP—this will replace the INC after ratification). A parallel structure operates in the voluntary phase, with an Interim Chemical Review Committee (ICRC). The Committee will review PIC notifications, and—when they meet the agreed criteria—draft a Decision Guidance Document (DGD).

Two Routes to Be "PIC-ed"

Pesticides in the voluntary PIC were carried forward, and new pesticides continue to be added. By June 2001, the process included 26 pesticides and five industrial chemicals (Table 1).

There are two routes for adding pesticides to the Convention. Under Article 5, a ban or severe restriction in any two regions triggers PIC if the action is taken for health or environmental reasons. Governments have decided that the PIC regions would be: Africa (48 countries), Latin America, and the Caribbean (33 countries), Asia (23 countries), Near East (22 countries), Europe (49 countries), North America (2 countries: Canada and US), Southwest Pacific (16 countries).

The second route is covered in Article 6, and addresses "severely hazardous pesticide formulations." This category applies only to pesticide formulations found to be causing health or environmental problems under conditions of use in developing countries, or countries with economies in transition. These pesticides may not have been banned, but—generally because of high toxicity— cause poisonings and deaths when used without extreme caution. Governments must submit evidence based on a "clear description of incidents related to the problem, including the adverse effects and the way in which the formulation was used." Nevertheless, this kind of evidence is rare, and collecting information is difficult: incidents take place far from medical facilities; many farmers are unaware of the active ingredients of pesticides they use; and it is common to use mixtures of several pesticides. The ICRC is investigating how to deal with these problems.

TABLE 1 Pesticides Covered by the Interim PIC Procedure, November 2000

Banned or severely restricted pesticides[a]
2,4,5-T (dioxin contamination)
Aldrin
Binapacryl (INC6)[a]
Captafol
Chlordane
Chlordimeform
Chlorobenzilate
DDT
Dieldrin
Dinoseb and dinoseb salts
1,2-Dibromoethane (EDB, or ethylene dibromide)
Ethylene dichloride (INC7)[a]
Ethylene oxide (INC7)[a]
Fluoroacetamide
HCH, mixed isomers
Heptachlor
Hexachlorobenzene
Lindane
Mercury compounds
mercuric oxide
mercurous chloride, Calomel
other inorganic mercury compounds
alkyl mercury compounds
alkoxyalkyl/aryl mercury compounds
Pentachlorophenol
Toxaphene (INC6)[a]
Severely hazardous pesticide formulations[b]
Monocrotophos
Methamidophos
Phosphamidon
Methyl parathion
Parathion

[a] Indicates that these four pesticides were added to the PIC list at the 6th and 7th International Negotiating Committee meetings.

[b] Only certain formulations of these severely hazardous pesticides are included.

Source: http://www.pic.int/[5]

Import Decisions, Information, and Website

Once a pesticide is included in PIC, the DGD is circulated to all governments who must decide whether to consent to or prohibit its import. Import decisions are posted on the PIC website, and circulated biannually. Governments in exporting countries must ensure that their exporters comply. Of course, many countries are both importers and exporters and under the rules of international trade, a country cannot ban the import of a pesticide that is manufactured and used nationally.

An important tool is the PIC Circular, updated every six months by the Secretariat. Circulated in hard copy and on the website,[5] it includes new bans and severe restrictions, importing country responses, and general progress reports. For the first time, it is easy to access sound information on government regulatory actions, even if these do not meet all the full PIC criteria.

The Convention—More than PIC

Information exchange is an important principle promoted under Article 14 of the Convention. Developing countries lack resources to undertake extensive evaluations of pesticides and governments are encouraged to share scientific, technical, economic, and legal information on chemicals within the scope of the Convention, as well as other information on their regulatory actions.

Building Capacity/Improving Regulations

The process of identifying problem pesticides through PIC will be slow, and there are limitations. In some cases, for example, governments will have no easy substitute, although this may increase the incentive to seek safer and more appropriate alternatives, including Integrated Pest Management strategies.

Financial resources are needed, not only to allow the Secretariat to meet its obligations, but also to ensure that regulators in developing countries can participate in workshops and training sessions. In poorer countries, with competing demands on scarce resources, chemical regulation is not always a priority. The status of an international Convention gives PIC the attention it requires to be effective, and should help attract the necessary funds.

PIC is just one tool, although an important one, in the regulation of pesticides. With good training and additional resources, PIC can play a central role as part of capacity building initiatives to help governments improve their ability to regulate pesticides, and to look for products and strategies that reduce the dependence on hazardous chemicals.

References

1. Rotterdam Convention on the Prior Informed Consent Procedure for Certain Hazardous Chemicals and Pesticides in International Trade, UNEP and FAO, Text and Annexes, January 1999.
2. The signatory countries can be found on the PIC website: http://www.pic.int/. The Convention closed for signatures in September 1999: countries which have not signed accede to, rather than ratify, the Convention, to the same effect.
3. *International Code of Conduct on the Distribution and Use of Pesticides* (*Amended Version*)*; FAO, 1989. The Code is currently being revised and updated.*
4. United Nations Conference on Environment and Development, Agenda 21, Chapter 19, Environmentally Sound Management of the Toxic Chemicals, Including Prevention of International Illegal Traffic in Toxic and Dangerous Products, UNEP, Nairobi, 1992.
5. Convention text and PIC website (http://www.pic.int/).

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

36

Laws and Regulations: Soil

Ian Hannam
and Ben Boer

Introduction

At a national level, soil law means a body of law to promote soil conservation enacted by a legislature, e.g., an act, decree, regulation, or other formal legal instrument that is legally enforceable. Soil law, or "soil legislation" as it may also be referred, includes those laws that have primary responsibility for soil conservation, soil and water conservation, and land rehabilitation. They are generally characterized by provisions to mitigate and manage soil erosion and soil degradation and methods to conserve soil resources. Internationally, the legal framework for the conservation of soil can include conventions, protocols, agreements, and covenants, which are expressed to be legally binding. Worldwide, soil law is managed by a variety of legal and institutional systems, which are the individual organizational and operational regimes that have the administrative authority over soil.

Why Law for Soil?

Soil bodies are effectively large ecosystems and comprise fundamental components of the earth's biodiversity. Soil is thus seen as the basis for the conservation of terrestrial biological diversity and the sustenance of all terrestrial organisms, including people. The ongoing and widespread soil degradation as a result of human use of soil provides the imperative for enactment of soil law. The ever-increasing demand for food by rapidly growing populations in many countries in the past few decades has exerted increasing environmental stress on the soil leading to widespread soil degradation.[1] The following definitions provide the context for soil law.

Soil

Soil forms an integral part of the earth's ecosystems and is situated between the earth's surface and bedrock. It is subdivided into successive horizontal layers with specific physical, chemical, and biological characteristics. From the standpoint of history of soil use, and from an ecological and environmental point of view, the concept of soil also embraces porous sedimentary rocks and other permeable materials together with the water that these contain and the reserves of underground water.[2]

Soil Degradation

Soil degradation is a loss or reduction of soil functions or soil uses. It includes aspects of physical, chemical, and biological deterioration, including loss of organic matter, decline in soil fertility, decline in structural condition, erosion, adverse changes in salinity, acidity, or alkalinity, and the effects of toxic chemicals, pollutants, or excessive flooding.[1]

Sustainable Use of Soil

The sustainable use of soils preserves the balance between the processes of soil formation and soil degradation while maintaining the ecological functions and needs of soil. In this context, the use of soil means the role of soil in the conservation of biological diversity and the maintenance of human life.[3]

International Law and Soil

International environmental law is an essential component for setting and implementing global, regional, and national policy on environment and development. There is an increasing recognition of the role of international environmental law to overcome the global problems of soil degradation, including its ability to provide a juridical basis for action by nations and the international community.[4] A number of international and regional instruments introduced in the past 10 years contain elements that can contribute to achieving sustainable use of soil. None are sufficient on their own. Some of the instruments could assist by promoting the management of some of the activities that can control soil degradation. However, this role is not readily apparent except for those that include provisions specifically directed to soil (e.g., see Article IV "Soil"— 1968 *African Convention on the Conservation of Nature and Natural Resources,* final revision text adopted by the African Union Assembly on July 11, 2003).

Declarations

A number of nonbinding declarations and charters draw attention to the fact that soil degradation and desertification are reaching alarming proportions and seriously endangering human survival. They call on states to cooperate and develop the tools to conserve soils. Key declarations relevant to soil include the 1972 *Stockholm Declaration on the Human Environment,* the 1981 FAO World Soil Charter, the 1982 *World Charter for Nature*, the 1982 *Nairobi Declaration,* the 1992 *Rio Declaration on Environment and Development,* and the 2002 *Johannesburg Declaration on Sustainable Development*. Also of relevance is the Programme for the Development and Periodic Review of Environmental Law for the First Decade of the 21st Century, known as the Montevideo Programme; this program includes provisions to improve the conservation, rehabilitation, and sustainable use of soils.[5]

International Conventions, Covenants, Treaties, and Agreements

Many multilateral agreements include provisions that could be used to promote sustainable use of soil, but the provisions are generally tangential to the needs of soil as such. Key global instruments relevant to soil include the 1992 *Convention on Biological Diversity,* the 1992 *United Nations Framework Convention on Climate Change* and the 1997 *Kyoto Protocol,* and the 1994 *United Nations Convention to Combat Desertification*. Relevant regional instruments include the 1968 *African Convention on the Conservation of Nature and Natural Resources* (Revised July 2003), the 1985 *ASEAN Agreement on the Conservation* *of Nature and Natural Resources*, the 1986 *Convention for the Protection of the Natural Resources and Environment of the South Pacific Region,* the 1986 *European Community Council Directive,* the 1995 *Convention Concerning the Protection of the European Alps,* and the 1998 *Protocol for the Implementation of the Alpine Convention of 1991 in the Area of Soil Protection.*[6]

National Soil Law

Legislation has been used for some 60 years in many countries to control soil degradation problems and to manage soil. A worldwide examination of national legal and institutional frameworks indicates that most countries approach the management of soil in a fragmented manner. The term "soil law" also covers those situations where comprehensive provisions for soil protection and management have been integrated in legislation that protects other aspects of the environment, such as forests, water, biodiversity, and desertification. In general, soil law thus provides for farm planning, implementation of soil erosion control measures, establishing community groups, planning catchment schemes, and compliance and enforcement. Some jurisdictions, such as the United Kingdom, have multiple soil legislation mechanisms that cover a broad range of functions including soil planning, access to sensitive land types, organic farming practices, nitrate sensitive areas, and soil restoration. On the other hand, federally organized countries often have a system where each state or province has its own soil legislation and supportive legal mechanisms. Hybrid situations also exist, such as in the People's Republic of China, which has enacted the *Water and Soil Conservation Law 1991* and the *Desertification Law 2002* at a national level, but causes them to be implemented through a comprehensive provincial system of law and regulations.

There is a wide variety of types of legal mechanisms used to protect and manage soil, including acts, decrees, resolutions, ordinances, codes, regulations, circulars, decisions, orders, and bylaws. Whereas these are generally appropriate, many need to be applied in more inventive ways to effectively manage the soil in an ecosystem context.[3]

Effectiveness of Soil Law

The effectiveness of international and national soil law is generally dependent on two matters: first, the capacity of a legal and institutional framework to manage soil—which is measured by the ability of a legislative and institutional system to achieve sustainable use of soil—and second, by the number and type of essential legal and institutional elements present in a soil statute in a format that enables soil degradation issues to be identified. These need to be backed by the legal, administrative, and technical capability in the particular instrument as a basis of some form of effective action. Capacity is also represented in the form of legal rights, the type of legal mechanisms, and importantly, the number and comprehensiveness of the essential elements and their functional capabilities. Legal and institutional "elements" for soil are the basic, essential components of a legal and institutional system. An individual law can include a number of legal mechanisms in a well thought-out structure that gives an organization the power it needs, through its executive and administrative structure, to address soil degradation. It is also possible that the necessary elements may be distributed among a number of individual laws within a comprehensive national legal and institutional system.[7]

Most key soil management issues are multifactorial (i.e., many include a sociological, a legal, and a scientific component), so it is obvious that generally more than one piece of environmental legislation (along with detailed regulations) and many types of legal and institutional elements will be needed to effectively manage soil degradation issues.[7] Legal and institutional elements can be used to assist in the evaluation of an existing law or legal instrument to determine its capacity to meet certain prescribed standards of performance for the sustainable use of soil. They can also be used to guide the reform of an existing soil law or to develop new legislation for the sustainable use of soil. The manner and degree in which an "essential element" is applied will vary according to the particular type of legal mechanism concerned and its expected role in a particular jurisdiction. For example, an international legal instrument may include a provision for dispute resolution, but the actual implementation of this provision between states might not rely on, or be influenced by, the existence of similar provisions within a law of either of the disputing states.[7]

IUCN Commission on Environmental Law

The Commission on Environmental Law of IUCN (The World Conservation Union) has carried out extensive investigation into the options for a new international instrument focusing on soil. The commission has also identified a variety of ways available for states to approach the task of a detailed legal and institutional analysis and the design of appropriate legal and institutional systems that provides for the effective management of soil. Arising from this work, in which the authors have been centrally involved, two principal strategies can be considered for the development of legal and institutional arrangements for soil. These are:

- A nonregulatory strategy which is characterized by elements for education, participatory approaches, soil management, and incentive schemes
- A regulatory approach that is characterized by statutory soil use plans that prescribe legal limits and targets of soil and land use, issue of licenses or permits to control soil use, and the use of restraining orders and prosecution for failure to follow prescribed standards of sustainable soil use.

These strategies can be approached on a short-term time frame for implementation or a longer-term time frame, which involves substantial reform of existing laws, policies, and institutional and sectoral change.[7]

Conclusions

Soil law in the past has been neglected at the international level and, in many of the world's regions, at the domestic level. However, the growing recognition of soil degradation as a major international environmental issue in the context of the conservation of biological diversity is gradually being addressed, and this is starting to change attitudes toward the benefits of improved international and national legal and institutions for soil.[8] Soil bodies represent complex terrestrial ecosystems. They require careful management of their ecological characteristics through the medium of soil law at a national and international level. This approach is essential for the long-term sustainable use of soil and to meet the food production requirements of the expanding human population of the world, as well as to meet the needs of all flora and fauna that depend on the soil for sustenance.

References

1. Bridges, E.M., Hannam, I.D., Oldeman, L.R., Penning deVries, F., Scherr, S.J., Sombatpanit, S., Eds.; *Response to Land Degradation;* Science Publishers Inc.: Enfield, NH, 2001.
2. Council of Europe European conservation strategy. Recommendations for the 6th European Ministerial Conference on the Environment; Council of Europe: Strasbourg, 1990.
3. Hannam, I.D.; Boer, B.W. *Legal and Institutional Frameworks for Sustainable Soils;* The World Conservation Union: Gland, U.K., 2002.
4. Khan, R. International law of land degradation. In *International Studies. 30:3;* Sage Publications: New Delhi, India, 1993.
5. UNEP. The Montevideo Programme III—The Programme for the Development and Periodic Review of Environmental Law for the First Decade of the 21st Century; 2001 Decision 21/23 of the Governing Council of UNEP, UNEP: Nairobi, February 2001.

6. Sands, P. *Principles of International Environmental Law;* Cambridge University Press: Cambridge, 2003.
7. Boer, B.W.; Hannam, I.D. Legal aspects of sustainable soils: International and National. Rev. Eur. Commun. Int. Environ. Law **2003**, *12* (2), 149–163.
8. WSSD (World Summit on Environment and Development). *A Framework for Action on Agriculture*; WEHAB Working Group: New York, 2002.

37

LEED-EB: Leadership in Energy and Environmental Design for Existing Buildings

Rusty T. Hodapp

Introduction

Existing buildings comprise a significant proportion of the total building stock in the United States and building operations consume large amounts of resources (energy, water, building materials, land, etc.), while generating great amounts of waste. For example, in the United States, commercial and industrial buildings alone are estimated to be responsible for the following[1]:

- 38.9% of primary energy use
- 38% of all CO_2 emissions
- 72% of all electricity use
- 13.6% of all potable water use
- 170 million tons of construction and demolition debris
- Using 40% of raw materials globally

Furthermore, because the average person spends 90% of their time indoors, the quality of a building's interior environment impacts virtually everyone. This suggests a very personal interest in better buildings in addition to the national implications of large-scale resource consumption. Issues such as these have driven government, corporate, and personal interest in sustainability and “green” topics. Applied to the building industry, this interest is forcefully seen in the rise of green building certification programs and, in particular, the Leadership in Energy and Environmental Design (LEED) Green Building Rating System™ of the United States Green Building Council (USGBC). Intended to guide the development and verify the performance of green buildings, LEED rating systems have become well accepted as a national standard. Consisting of a number of products, the LEED Green Building Rating System largely focuses on design and construction of new buildings. One rating system, however, known as LEED for Existing Buildings: Operations and Maintenance (LEED-EB O&M) is oriented at the operation, maintenance, and management of existing buildings. LEED-EB O&M is of particular interest to facility managers, energy managers, owners, or others interested in reducing operating costs, improving indoor environmental quality, and minimizing the environmental impact of buildings as a growing body of case study and research evidence suggests that these outcomes are linked to green building practices.

This entry presents an overview of the LEED-EB O&M Green Building Rating System, its benefits and distinctions from other LEED rating systems, how it is organized and implemented, and the value of high-performance green buildings.

Introduction to LEED

LEED stands for the USGBC’s family of standards for rating “green buildings.” The LEED Green Building Rating System is USGBC’s effort to provide a national standard to define a green building. Used as guideline for design, construction, operation, and maintenance and with third-party certification, LEED provides a consistent, credible means of developing and operating high-performance, environmentally sustainable buildings.

United States Green Building Council

Formed in 1993, the USGBC has become perhaps the most prominent “green building” organization in America. With more than 18,000 member organizations, a network of 78 local affiliates, and more than 140,000 LEED Professional Credential holders, the non-profit organization works through leaders in all sectors of the building industry to advance buildings that are environmentally responsible, profitable, and healthy places to live and work. Driving its mission to transform the building marketplace to sustainability is the Council’s LEED Green Building Rating System and related training and professional accreditation programs. USGBC also supports education, research, and advocacy programs as well as strategic alliances with key industry, research organizations, and federal, state, and local government agencies to transform the building market.[2]

LEED Green Building Rating System

The LEED Green Building Rating System is a voluntary, consensus-based, market-driven building rating system. LEED assesses sustainability from a whole-building perspective by evaluating five key areas of a building’s performance in terms of economic, environmental, and human health impact: sustainable site development, water savings, energy efficiency, materials selection, and indoor environmental quality. Buildings are awarded different levels of certification (Certified, Silver, Gold, Platinum) based upon the amounts of credits satisfied and points earned. These credits are performance oriented and intended to address specific impacts inherent in the design, construction, operation, and maintenance of buildings.

The initial LEED rating system (referred to as LEED Version 1.0) was released in August 1998. LEED 1.0 was extensively modified and released as Version 2.0 in March 2000 with LEED Version 2.1 following in 2002 and Version 2.2 in 2005. LEED has continued to evolve, undertaking new initiatives and expanding into a family of products.

As of April 2009, the portfolio of LEED rating systems consists of several products targeting specific sectors of the buildings market:

- LEED 2009 for New Construction and Major Renovation
- LEED 2009 for Commercial Interiors
- LEED 2009 for Core and Shell Development
- LEED 2009 for Existing Buildings: Operations and Maintenance
- LEED 2009 for Schools
- LEE 2009 for Retail
- LEED for Healthcare
- LEED for Homes
- LEED 2009 for Neighborhood Development

With new products, technologies, and design innovations coming into the green building marketplace daily, the Rating Systems and Reference Guides will continue to evolve as necessary to stay current and relevant.[3]

Benefits of Green Building Certification

LEED provides a guidebook for the design, construction, operation, and maintenance of green buildings, the general benefits of which will be described in more detail later in this entry. The LEED rating system is flexible in order to provide owners and design teams the ability to accommodate circumstances or goals specific to their project. The rigorous and independent certification process provides firm and compelling proof that the building has achieved the sustainable goals established for it and is performing as intended. The credible assurance that a building is in fact green can be valuable to owners, occupants, investors, and other key stakeholders in the industry as well as the public at large.

LEED and Existing Buildings

With the exception of LEED-EB O&M, the LEED family of products is intended to address the design and construction phases of buildings. The primary users of these products are architects, engineers, construction contractors, and building owners. A building designed and constructed to LEED standards has verifiably incorporated green or sustainable features and, therefore, should perform better in the key impact areas (economic returns, environmental impact, and occupant health and comfort) than a typical building built to basic code standards. Addressing the design and construction phase of a building's life is extremely important because it is in these phases that many irreversible decisions with long-term impacts on the building's performance are made.

Figure 1 depicts graphically what has become well established regarding the life-cycle cost of buildings: 1) the majority (75% or more) of total life-cycle costs occur after construction (i.e., during the O&M phase); and 2) many of the decisions that drive long-term cost occur during programming and design.[4]

However, the fact is that new buildings represent a very small percentage of the total commercial building stock in the United States. According to some sources, new construction amounts to 2% of the total stock of buildings in any one year.[5] No doubt it is safe to assume that the majority of these existing buildings were not designed or are being operated and maintained to green standards. In order to improve the building sector's performance in terms of sustainability, clearly the existing stock must be addressed in addition to new construction. Similarly, since the post-construction phase of a building's life cycle contributes disproportionately to its total cost, resource consumption, and impact on users, standards for operations and maintenance are necessary to maximize the benefits of green practices in the building sector. This is exactly the focus of the LEED-EB O&M Rating System.

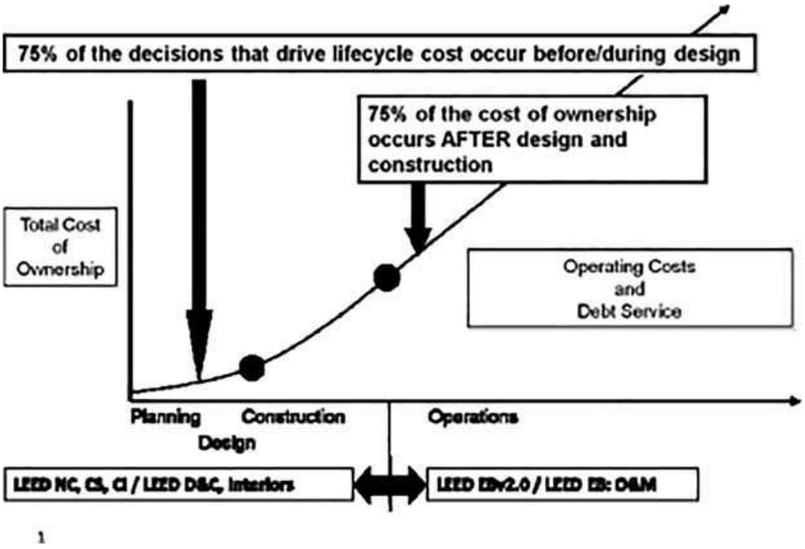


FIGURE 1 Building life-cycle cost curve.

LEED for Existing Buildings: Operations and Maintenance

The LEED-EB O&M Rating System is a set of performance standards for sustainable operations and maintenance of existing buildings of various types and all sizes. It is intended to advance high-performance, healthy, durable, affordable, and environmentally sound practices in existing buildings. LEED-EB O&M provides an entry point into the LEED certification process for the existing building stock. It can be used for buildings new to LEED certification as well as those previously certified under LEED-NC. The USGBC began developing LEED-EB in 2000 and it was tested in a pilot phase involving more than 100 buildings in 2002. The final version (Version 2.0) was released in October 2004. The current version, LEED-EB O&M, was released in April 2009 under the suite of 2009 LEED rating systems and has been further updated in April 2010. The introduction to the LEED 2009 for Existing Buildings: Operations and Maintenance Green Building Rating System states “LEED for Existing Buildings Operations and Maintenance encourages owners and operators of existing buildings to implement sustainable practices and reduce the environmental impacts of their buildings over their functional life cycles.” To achieve this, LEED-EB O&M addresses exterior building site maintenance, water and energy use, environmentally preferred products and practices for cleaning and alternations, waste stream management, and ongoing indoor environmental quality.[6]

Issues Addressed by LEED-EB O&M

LEED-EB O&M addresses all the key facets of building operations and maintenance that impact total cost of ownership, the environment, and building occupants. Some examples include the following:

- Energy use
- Water use
- Building operations and maintenance
- Building systems (e.g., mechanical, electrical, plumbing) performance
- Maintenance of building exterior and site
- Ventilation and indoor air quality
- Lighting quality

- Thermal comfort of spaces
- Green cleaning
- Recycling programs
- Green product purchasing programs
- Management of indoor pollutants and toxic substances
- Systems upgrades[7]

LEED-EB O&M seeks to address sustainability on an ongoing basis. This largely falls under the scope of those involved in managing and operating buildings, and clearly, their involvement and expertise are necessary to successfully certify a building under LEED-EB O&M. To the extent the benefits of green buildings and popularity of standards like LEED continue growing, market forces will create new opportunities for facility managers, energy managers, etc., demand for their services, and highlight the overall value of their contributions.

Key Distinctions between LEED-EB O&M and Other LEED Products

Although sharing many common features in terms of structure and process with other LEED products, LEED-EB O&M is fundamentally distinct in three key ways. LEED-NC (and the other new building-oriented products) is essentially a onetime event whereas LEED-EB O&M represents an ongoing process. Second, with LEED-NC, the green building process ends after the design and construction phase. For LEED-EB O&M, the green building phase is a continuous process that deals with ongoing operations, maintenance, and upgrades of a building over its life cycle. Buildings certified under LEED-EB O&M require recertification at least once every 5 years. Finally, given their focus on different phases of a building's life cycle, LEED-NC is primarily a capital budget event while LEED-EB O&M deals with operating budgets.[8]

Overview of LEED-EB O&M

In the same manner as all LEED products, the LEED-EB O&M Rating System is based on evaluations of a building in seven categories:

1. Sustainable sites
2. Water efficiency
3. Energy and atmosphere
4. Materials and resources
5. Indoor environmental quality
6. Innovation (in this case, in operations)
7. Regional priority

Minimum Program Requirements

All projects must meet certain minimum program requirements (MPRs) to be eligible for certification under the LEED 2009 rating systems. MPRs define the minimum characteristics that a project must possess in order to be certified and are intended to: 1) provide clear guidance to users; 2) protect the integrity of LEED program; and 3) reduce challenges during the certification process. The LEED 2009 MPRs for EB O&M are as follows:

1. Must comply with environmental laws
2. Must be a complete, permanent building or space
3. Must use a reasonable site boundary
4. Must comply with minimum floor area requirements

5. Must comply with minimum occupancy rates
6. Must commit to sharing whole-building energy and water usage data
7. Must comply with a minimum building area-to-site area ratio

The ongoing performance data from buildings required as part of the certification will be compiled and used to establish benchmarks for building performance and provide operators an idea of how their building compares on water and energy use. To further its commitment to improving building performance, the USGBC launched the Building Performance Initiative (BPI) in August 2009 to complement the MPR for ongoing performance data. The BPI will make the data collected available to building owners for analysis and feedback.

Prerequisites

Also consistent with other LEED rating systems, LEED-EB O&M requires every project to meet certain prerequisites in order to be considered for certification (see the list of prerequisites by category in Table 1). All prerequisites must be satisfied for a project to be eligible for certification.

The prerequisites include such items as minimum levels of water and energy efficiency, building commissioning, no CFC refrigerants, no-smoking policy, and other basic elements of a high-performance, green building operation. A key prerequisite involves a minimum performance period for the building. LEED-EB O&M requires buildings to be in operation for a minimum of 12 continuous months before certifying (3 months for all prerequisites and credits except Energy and Atmosphere Prerequisite 2 and Credit 1, which require a minimum of 12 months).

Credits and Points

Buildings achieve certification under all LEED products by accumulating a certain number of credit points. Points can be obtained in any combination within and among the credits and categories (see Table 2 for the credit and point breakdown for LEED-EB O&M). All LEED rating systems have 100 base points, and up to 10 bonus points can be earned through Innovation and Regional Priority credits.

TABLE 1 LEED 2009 for EB O&M Prerequisites

Category	Prerequisites
Sustainable sites	0
Water efficiency	1
Energy and atmosphere	3
Materials and resources	2
Indoor environmental quality	3
Total	9

TABLE 2 LEED-EB O&M Credits and Points

Category	Credits	Points
Sustainable sites	9	26
Water efficiency	4	14
Energy and atmosphere	9	35
Materials and resources	9	10
Indoor environmental quality	15	15
Innovation in operations	3	6
Regional priority	1	4
Total	50	110

Each LEED rating system uses the same format for prerequisites and credits. The sections include the following:

- *Intent*—describes the main goal of the prerequisite or credit.
- *Requirements*—specifies the criteria needed to satisfy the prerequisite or credit.
- *Submittals*—specifies the documentation required to demonstrate compliance with the prerequisite or credit.
- *Potential Technologies and Strategies*—identifies means and methods that project teams may consider to achieve the prerequisite or credit.

Certification Levels

LEED rating systems allow buildings to achieve various levels of certification based on points achieved (see Table 3 for the certification levels for LEED-EB O&M).

Registration Process

With the launch of LEED Version 3 in 2009, USGBC implemented a new certification model. LEED v3 consists of three components:

- LEED 2009 rating systems.
- An upgrade to LEED Online to make it faster and easier to use.
- New building certification model—an expanded infrastructure based on ISO standards, administered by the Green Building Certification Institute (GBCI) for improved capacity, speed, and performance.

LEED Online is the primary resource for managing the LEED documentation process. It allows project teams to manage project details, complete documentation requirements for LEED credits and prerequisites, upload supporting files, submit applications for review, receive reviewer feedback, and ultimately earn LEED certification. The GBCI is an independent, third-party organization that has assumed administration of LEED certification for all commercial and institutional projects registered under any of the LEED rating systems.

The process of certifying a building under any of the LEED rating systems is essentially the same—the project is first registered with GBCI using LEED Online. Once a project is registered, access to software tools, supplemental resources, sample documentation, credit interpretation rulings, and other essential information is provided. For LEED-EB O&M, the initial application (application for standard review) must be submitted within 60 calendar days of the performance periods used. The application has to include complete documentation for all prerequisites and enough points for certification. GBCI reviews the application and designates each credit and prerequisite as anticipated pending or denied. This preliminary standard review is targeted (but not guaranteed) for completion within 25 business days of receipt of the application. Within 25 days of receiving GBCI's preliminary standard review, the owner may submit a response including any revised documentation. GBCI will then review and return comments for all credits and prerequisites in response to the preliminary Standard Review and

TABLE 3 LEED-EB O&M Certification Levels

Certification Level	Points Required
Certified	40–49
Silver	50–59
Gold	60–79
Platinum	80+

designate each as awarded or denied. This final Standard Review is targeted for completion within 15 business days of receipt of the completed application. The owner then accepts or appeals the final review. Following acceptance of the final certification review, LEED projects:

1. Will receive a formal certificate of recognition
2. Will receive information on how to order plaque and certificates, photo submissions, and marketing
3. May be included (at the owner's discretion) in the online LEED Project Directory of registered and certified projects
4. May be included (along with photos and other documentation) in the U.S. Department of Energy High Performance Buildings Database[9]

Implementation Process

From a practical standpoint, the process of implementing LEED-EB O&M should generally involve the following steps:

1. Become familiar with the LEED-EB O&M Rating System
2. Gain the support of key decision makers and stakeholders
3. Form a project team
4. Conduct a preliminary building audit and identify corrective actions required to meet prerequisites and/ or opportunities
5. Establish project goals related to target certification level, credits to be pursued, and budget
6. Register the project
7. Create and adopt policies and procedures, implement upgrades, make operational changes, etc., in accordance with the project goals
8. Track performance
9. Assemble and submit required documentation (preliminary and any required corrections or resubmittals)
10. Achieve certification

The minimum performance period for initial certifications under LEED-EB O&M is 12 months. During this period, actual operational performance must be tracked and reported. The performance tracking period can be as long as 2 years depending upon the project goals and/or implementation strategy.

The USGBC provides project teams with numerous resources including the LEED Online system for managing and preparing the certification application, credit templates that define supporting documentation needed and compliance calculations, credit interpretation rulings that can help answer questions on credits and implementation strategies, and the LEED Reference Guides.[10,11]

Benefits of Green Buildings

The premise inherent in the LEED rating systems is that "green" buildings provide superior value to owners, occupants, and other stakeholders. Typically, the value proposition construct for sustainability, green buildings, etc., is the well-known triple bottom line of economic returns, environmental impact, and social benefit. Green buildings, also known (perhaps more accurately) as high-performance buildings, are premised as providing superior economic returns, reduced environmental impact, and enhanced social benefits. Such buildings in theory:

- Were properly built and/or are well operated and maintained
- Use resources (e.g., energy, water, building materials, O&M supplies) more efficiently
- Provide a safer, more comfortable, and productive working environment for occupants

In fact, there is a robust and growing body of evidence in research and case study that supports these claims. The following examples are illustrative.

A report commissioned by California's Sustainable Building Task Force found that energy savings alone exceeded the average cost increase associated with 33 different LEED buildings studied. When adding the life-cycle cost benefits of water savings, reduced emissions, operations and maintenance efficiencies, and improved occupant productivity and health, the 20 years net present value of the financial benefits of green buildings exceeded the implementation costs by as much as 10–15 times.[12] Case studies on commissioning alone show that construction and operating costs can be reduced from 1 to 70 times the initial cost of commissioning.[13] Improved thermal comfort, reduced indoor pollutants, enhanced ventilation rates, and other characteristics of green buildings have been found to have positive impacts on occupant productivity, student test scores, absenteeism, and incidences of various sicknesses.[14]

Other benefits continue to be demonstrated in case studies and research, including the following:

- Increased building value
- Risk mitigation
- Employee loyalty and recruitment
- Brand image and public relations
- Environmental stewardship[15]

Energy Efficiency Potential of Green Buildings

In LEED 2009, points are allocated among credits based on the potential environmental impacts and human benefits of each credit. As a result, the allocation of points significantly changed in comparison to previous versions of the LEED rating systems. These changes increased the relative importance of reducing energy consumption and building-related greenhouse gas emissions. Reflecting this, one credit in LEED 2009 EB O&M comprises the largest potential amount of points—Energy and Atmosphere Credit 1 Optimize Energy Efficiency Performance provides the opportunity for 18 possible points. Furthermore, with its emphasis (and associated requirements) on demonstrated performance, LEED-EB O&M presents tremendous potential to reduce energy consumption throughout the commercial building sector. When considering the fact that existing buildings comprise some 95% of the commercial building stock in the United States, the magnitude of potential reduction is immense. To put this potential in context, consider U.S. Energy Information Administration (EIA) projections of the impact of energy efficiency on per capita commercial energy consumption. In their *Annual Energy Outlook 2010,* EIA estimates per capita commercial energy consumption in 2035 could be decreased by as much as 17.5% depending upon the degree to which technology-based efficiency improvements are deployed throughout the sector.[16] McKinsey and Company estimates that, by 2020, the United States could reduce annual energy consumption by 9.1 quadrillion British thermal units (BTUs) (23%) of end-use energy (18.4 quadrillion BTUs in primary energy) from a business-as-usual baseline by deploying an array of energy efficiency measures with the commercial sector accounting for 25% of this potential. At full potential, the projected efficiency improvements could reduce greenhouse gas emissions by as much as 1.1 gigatons of CO_2 by 2020 and serve as a bridge to low-carbon energy sources.[17] Similarly, the Electric Power Research Institute (EPRI) estimates that the combination of energy efficiency and demand response programs has the potential to reduce summer peak electric demand in the United States by 157 GW to 218 GW (14% to 20%) by 2030.[18] Finally, consider the potential national impact of LEED energy savings presented in the Green Building Market and Impact Report 2009. This report projects that given the acceleration of LEED adoption, energy savings in the United States could reach 1.75 quadrillion BTUs by 2020 and 3.9 quadrillion BTUs by 2035 (8.3% and 17.3%, respectively, of national annual commercial building energy use). A best-case scenario could see those savings rise to 22.3% by 2030.[19]

Conclusion

The USGBC's LEED Green Building Rating System has become a well-recognized standard for guiding the development of green, high-performance buildings and for verifying that established green building goals have been accomplished. The LEED family of rating systems is focused on the design and construction process for new buildings. However, one rating system—LEED-EB O&M— focuses on operations and maintenance of existing buildings. LEED-EB O&M is of particular interest to facility managers, energy managers, and other professionals involved in building operation and management. LEED-EB O&M provides a guidebook for those interested in "greening" their existing building stock. Implementing these green processes and practices can be an effective means of reducing a building's life-cycle costs, reducing its environmental impact, and improving occupant health and productivity.

References

1. United States Green Building Council. Green Building Facts. Available at http://www.usgbc.org/ (accessed August 2007).
2. United States Green Building Council. About USGBC. Available at http://www.usgbc.org/ (accessed August 2007).
3. United States Green Building Council. *LEED Reference Guide for Green Building Design and Construction 2009 Edition;* USGBC: Washington, 2009; xii.
4. National Research Council. *Investments in Federal Facilities, Asset Management Strategies for the 21st Century;* The National Academies Press: Washington, 2004; 27.
5. Architecture2030. Available at http://architecture2030.org/the_ solution/buildings_solution_how (accessed November 2010).
6. United States Green Building Council. *LEED 2009for Existing Buildings: Operations and Maintenance Rating System* (*Updated April 2010*)*;* USGBC: Washington, 2010; xvi.
7. Opitz, M. What LEED-EB Is and Why to Use It, available at http://www.fmlink.com/ProfResources/Sustainability/Articles/article.cgi?USBGC:200604-01.html (accessed August 2007).
8. United States Green Building Council. LEED-EB Presentation. Available at http://www.usgbc.org/ (accessed August 2007).
9. Green Building Certification. GBCI LEED Certification Manual. Available at http://www.gbci.org/main-nav/building-certification/leed-certification.aspx (accessed July 2010).
10. Opitz, M. Starting and Managing Your LEED-EB Project, available at http://www.fmlink.com/ProfResources/Sustainability/Articles/article.cgi?USBGC:200609-20.html (accessed September 2007).
11. United States Green Building Council. *LEED 2009 for Existing Buildings: Operations and Maintenance Rating System (Updated April 2010);* USGBC: Washington, 2010; xvi–xxiv.
12. Kats, G.; Alevantis L.; Berman, A.; Mills, E.; Perlman, J. The Costs and Financial Benefits of Green Buildings: A Report to California's Sustainable Building Task Force, available at http://www.cap-e.com/ewebeditpro/items/O59F3259.pdf (accessed September 2006).
13. ASHRAE. *ASHRAE GreenGuide;* Elsevier: Burlington, 2006; 14.
14. Callan, D. Green Building Report: Studies Relate IAQ and Productivity. Build. Operating Manage. **2006**, *52* (11), 6–8.
15. Yudelson, J. *Marketing Green Buildings, Guide for Engineering, Construction and Architecture;* The Fairmont Press: Lilburn, 2006; 5–7.
16. U.S. Energy Information Administration, U.S. Department of Energy. *Annual Energy Outlook 2010,* 2010, 59.
17. Choi Granade, H.; Creyts, J.; Derkach, A.; Farese, P.; Nyquist, S.; Ostrowski, K. Unlocking Energy Efficiency in the U.S. Economy, available at http://www.mckinsey.com/USenergyefficiency (accessed July 2010).

18. Faruqui, A.; Hledik, S.; Rohmund, I.; Sergici, S.; Siddiqui, O.; Smith, K.; Wikler, G.; Yoshida, S. *Assessment of Achievable Potential from Energy Efficiency and Demand Response Programs in the U.S. (2010–2030).* EPRI: Palo Alto, CA, 2009; xi.
19. Watson, R. Green Building Market and Impact Report, 2009, available at http://www.greenbiz.com/business/research/report/2009/11/05/green-building-market-and-impact-report-2009 (accessed July 2010).

38

LEED-NC: Leadership in Energy and Environmental Design for New Construction

Stephen A. Roosa

Introduction

Land development practices have yielded adverse environmental consequences, urban dislocation, and changes in urban infrastructure. Urban development in particular has long been associated with reduced environmental quality and environmental degradation.[2] The rate at which undeveloped land is being consumed for new structures—and the growing appetite of those structures for energy and environmental resources—has contributed to ecosystem disruption and has fostered impetus to rethink how buildings are sited and constructed. While urban developmental patterns have been associated with environmental disruptions at the local and regional scales, the scientific assessments of global impacts have yielded mixed results. In part as a reaction to U.S. development patterns that have traditionally fostered suburbanization and subsidized automobile-biased transportation infrastructure, design alternatives for structures with environmentally friendly and energy efficient attributes have become available.

According to the United Nations Commission on Sustainable Development, "air and water pollution in urban areas are associated with excess morbidity and mortality … Environmental pollution as a result of energy production, transportation, industry or lifestyle choices adversely affects health. This would include such factors as ambient and indoor air pollution, water pollution, inadequate waste management, noise, pesticides and radiation."[3] It has been demonstrated that a relationship exists between the rates at which certain types of energy policies are adopted at the local level and select indicators of local sustainability.[4] As more urban policies focus on the built environment, buildings continue to be the primary building blocks of urban infrastructure. If buildings can be constructed in a manner that is less environmentally damaging and more energy efficient, then there is greater justification to label them as "green" buildings.

The concept of sustainability has evolved from considerations of land development, population growth, fossil fuel usage, pollution, global warming, availability of water supplies, and the rates of resource use.[5] Thankfully, a vocabulary of technologies and methodologies began to develop in the 1970s and 1980s that responded to such concerns. Driven by ever increasing energy costs, energy engineers began to apply innovative solutions, such as use of alternative energy, more efficient lighting systems and improved electrical motors. Controls engineers developed highly sophisticated digital control systems for heating, ventilating and air conditioning systems. With growing concerns about product safety and liability issues regarding the chemical composition of materials, manufacturers began to mitigate the potential adverse impacts of these materials upon their consumers. Resource availability and waste reduction became issues that began to influence product design. In the span of only 25 years, local governments made curbside recycling programs in larger U.S. cities nearly ubiquitous. Terms and phrases such as "mixed use planning," "brownfield redevelopment," "alternative energy," "micro-climate," "systems approach," "urban heat island effect," "energy assessments," "measurement and verification," and "carrying capacity" created the basis for a new vocabulary which identifies potential solutions. All of these concerns evolved prior to the 1992 U.N. Conference on the Environment and Development, which resulted in the Rio Agenda 21 and clarified the concept sustainability.

In regard to the built environment, architectural designers renewed their emphasis on fundamental design issues, including site orientation, day lighting, shading, landscaping, and more thermally cohesive building shells. Notions of "sick building syndrome" and illnesses like Legionnaires' disease, asthma and asbestosis, jolted architects and engineers into re-establishing the importance of the indoor environmental conditions in general and indoor air quality (IAQ) in particular when designing their buildings.

The decisions as to what sort of buildings to construct and what construction standards to apply are typically made locally. Those in the position to influence decisions in regard to the physical form of a proposed structure include the builder, developer, contractors, architects, engineers, planners, and local zoning agencies. In addition, all involved must abide by regulations that apply to the site and structure being planned. The rule structure may vary from one locale to another. What is alarming is that past professional practice within the U.S. building industry has only rarely gauged the environmental or energy impact of a structure prior to its construction. Prior to the efforts of organizations like the U.S. Green Building Council (USGBC) (established in 1995), the concept of what constituted a "green building" in the United States lacked a credible set of standards.

Concept of Green Buildings

Accepting the notion that sustainable, environmentally appropriate, and energy efficient buildings can be labeled "green," the degree of "greenness" is subject to multiple interpretations. The process of determining which attributes of a structure can be considered "green" or "not green" is inconclusive and subjective. Complicating the process, there are no clearly labeled "red" edifices with diametrically opposing attributes. While it is implied that a green building may be an improvement over current construction practice, the basis of attribute comparison is often unclear, subjective, and confusing. It is often unclear as to what sort of changes in construction practice, if imposed, would lead the way to greener, more sustainable buildings. If determinable, the marketplace must adjust and provide the technologies and means by which materials, components, and products can be provided to construction sites where greener buildings can arise. Since standards are often formative and evolving, gauging the degree of greenness risks the need to quantify subjective concepts.

There are qualities of structures, such as reduced environmental impact and comparatively lower energy usage, which are widely accepted as qualities of green construction practices. For example, use of recycled materials with post-consumer content that originates from a previous use in the consumer market and post-industrial content that would otherwise be diverted to landfills is widely considered an issue addressable by green construction practices. However, evaluation of green building attributes or standards by organizations implies the requirement that decisions be based on stakeholder consensus.

This process involves input to the decision-making processes by an array of representative stakeholders in often widely diverse geographic locations. For these and other reasons, developing a rating system for green buildings is both difficult and challenging.

Rating Systems for Buildings

Rating systems for buildings with sustainable features began to emerge in embryonic form in the 1990s. The most publicized appeared in the United Kingdom, Canada, and the United States. In the United Kingdom, the Building Research Establishment Environmental Assessment Method (BREEAM) was initiated in 1990. BREEAM™ certificates are awarded to developers based on an assessment of performance in regard to climate change, use of resources, impacts on human beings, ecological impact, and management of construction. Credits are assigned based on these and other factors. Overall ratings are assessed according to grades that range from pass to excellent.[6]

The International Initiative for a Sustainable Built Environment, based in Ottawa, Canada, has its Green Building Challenge program with more than 15 countries participating. The collaborative venture is geared toward the creation of an information exchange for sustainable building initiatives and the development of "environmental performance assessment systems for buildings."[7] In the United States, agencies of the central government co-sponsored the development of the Energy Star™ program, which provides "technical information and tools that organizations and consumers need to choose energy-efficient solutions and best management practices."[8] Expanding on their success, Energy Star™ developed a building energy performance rating system which has been used for over 10,000 buildings.

Entering the field at the turn of the new century, the USGBC grew from an organization with just over 200 members in 1999 to 3500 members by 2003.[9] The LEED™ rating system is a consensus-developed and reviewed standard, allowing voluntary participation by diverse groups of stakeholders with interest in the application and use of the standard. According to Boucher, "the value of a sustainable rating system is to condition the marketplace to balance environmental guiding principles and issues, provide a common basis to communicate performance, and to ask the right questions at the start of a project."[10] The first dozen pilot projects using the rating system were certified in 2000.

LEED-NC Rating System

The USGBC's Green Building Rating System is a voluntary, consensus-developed set of criteria and standards. This rating system evolved with a goal of applying standards and definition to the idea of high-performance buildings. The use of sustainable technologies is firmly established within the LEED project development process. LEED loosely defines green structures as those that are "healthier, more environmentally responsible and more profitable."[1]

LEED-NC 2.1 is the USGBC's current standard for new construction and major renovations. It is used primarily for commercial projects such as office buildings, hotels, schools, and institutions. The rating system is based on an assessment of attributes and an evaluation of the use of applied standards. Projects earn points as attributes are achieved and the requirements of the standards are proven. Depending on the total number of points a building achieves upon review, the building is rated as Certified (26–32 points), Silver (33–38 points), Gold (39–51 points) or Platinum (52 or more points).[11] Theoretically, there are a maximum of 69 achievable points. However, in real world applications, gaining certain credits often hinders the potential of successfully meeting the criteria of others. While achieving the rating of Certified is relatively easily accomplished, obtaining a Gold or Platinum rating is rare and requires both creativity and adherence to a broad range of prescriptive and conformance-based criteria.

The LEED process involves project registration, provision of documentation, interpretations of credits, application for certification, technical review, rating designation, award, and appeal. Depending on variables such as project square footage and USGBC membership status, registration fees can range up to $7500 for the process.[12]

LEED Prerequisites Categories and Criteria

To apply for the LEED labeling process, there are prerequisite project requirements which earn no points. For example, in the Sustainable Sites category, certain procedures must be followed to reduce erosion and sedimentation. In the category of Energy and Atmosphere, minimal procedures are required for building systems commissioning. Minimal energy performance standards must be achieved (e.g., adherence to ANSI/ASHRAE/IESNA Standard 90.1–1999, Energy Standard for Buildings Except Low-Rise Residential Buildings, or the local energy code if more stringent), and there must be verification that CFC refrigerants will not be used or will be phased out. In addition, there are prerequisite requirements outlining mandates for storage and collection of recyclable material, minimum IAQ performance (the requirements of ASHRAE Standard 62–1999, Ventilation for Acceptable Indoor Air Quality must be adhered to), and the requirement that non-tobacco smokers not be exposed to smoke.

In addition to the prerequisite requirements, the LEED process assigns points upon achieving certain project criteria or complying with certain standards. The total points are summed to achieve the determined rating. Projects can achieve points from initiatives within the following sets of categories: Sustainable Sites (14 points), Water Efficiency (5 points), Energy and Atmosphere (17 points), Materials and Resources (13 points), and Indoor Environmental Quality (15 points). Use of a LEED Accredited Professional (1 point) to assist with the project[13] earns a single point. Additional points are available for Innovation and Design Process (maximum of 4 points).

Within each category, the specific standards and criteria are designed to meet identified goals. In the category of Sustainable Sites, 20.2% of the total possible points are available. This category focuses on various aspects of site selection, site management, transportation and site planning. The goals of this category involve reducing the environmental impacts of construction, protecting certain types of undeveloped lands and habitats, reducing pollution from development, conserving natural areas and resources, reducing the heat island impacts, and minimizing light pollution. Site selection criteria are designed to direct development away from prime farmland, flood plains, habitat for endangered species and public parkland. A development density point is awarded for projects that are essentially multistory. If the site has documented environmental contamination or is designated by a governmental body as a brownfield, another point is available. In regard to transportation, four points are available for locating sites near publicly available transportation (e.g., bus lines or light rail), providing bicycle storage and changing rooms, provisions for alternatively fueled vehicles and carefully managing on-site parking. Two points in this category are obtained by limiting site disturbances and by exceeding "the local open space zoning requirement for the site by 25%."[14] In addition, points are available by following certain storm water management procedures, increasing soil permeability, and attempting to eliminate storm water contamination. Potential urban heat island effects are addressed by crediting design attributes such as shading, underground parking, reduced impervious surfaces, high albedo materials, reflective roofing materials, or vegetated roofing. Finally, a point is available for eliminating light trespass.

Water efficiency credits comprise 7.2% of the total possible points. With the goal of maximizing the efficiency of water use and reducing the burden on water municipal systems, points are credited for reducing or eliminating potable water use for site irrigation, capturing and using rainwater for irrigation, and using drought tolerant or indigenous landscaping. This section of the LEED standard also addresses a building's internal water consumption. Points are available for lowering aggregate water consumption and reducing potable water use. Reducing the wastewater quantities or providing on-site tertiary wastewater treatment also earns points.

Energy and Atmosphere is the category that offers the greatest number of points, 24.6% of the total possible. The intents of this category include improving the calibration of equipment, reducing energy costs, supporting alternative energy, reducing the use of substances that cause atmospheric damage, and offering measurement and verification criteria. Optimizing the design energy cost of the regulated energy systems can achieve a maximum of ten points. To assess the result, project designs are modeled against a base case solution which lacks certain energy-saving technologies. Interestingly, the unit of measure for evaluating energy performance to achieve credits is not kilocalories or million Btus, but dollars. Points are awarded in whole units as the percentage of calculated dollar savings increases incrementally. In addition to the ten points for energy cost optimization, a maximum of three additional points is available for buildings that use energy from onsite renewable energy generation. Purchased green power is allocated a single point if 50% of the electrical energy (in kWh) comes from a two year green power purchasing arrangement. This category provides points for additional commissioning and elimination of the use of HCFCs and halon gases. Measurement and Verification (M&V) is allowed a point, but only if M&V options B, C, and D, as outlined in the 2001 edition of the International Measurement and Verification Protocol (IPMVP), are used.

The Materials and Resources category represents 18.8% of the total possible points. This category provides credit for material management; adaptive reuse of structures; construction waste management; resource reuse; use of material with recycled content; plus the use of regionally manufactured materials, certain renewable materials and certified wood products. A point is earned for providing a space in the building for storage and collection of recyclable materials such as paper, cardboard, glass, plastics and metals. A maximum of three points is available for the adaptive reuse of existing on-site structures and building stock. The tally increases with the extent to which the existing walls, floor, roof structure, and external shell components are incorporated into the reconstruction. LEED-NC 2.1 addresses concerns about construction waste by offering a point if 50% of construction wastes (by weight or volume) are diverted from landfills and another point if the total diversion of wastes is increased to 75%. A project that is composed of 10% recycled or refurbished building products, materials, and furnishings gains an additional two points. Another two points are available in increments (one point for 5%, two points for 10%) if post-consumer or post-industrial recycled content (by dollar value) is used in the new construction. To reduce environmental impacts from transportation systems, a point is available if 20% of the materials are manufactured regionally (defined as being within 500 miles or roughly 800 km of the site), and an added point is scored if 50% of the materials are extracted regionally. A point is available if rapidly renewable materials (e.g., plants with a ten year harvest cycle) are incorporated into the project, and yet another point is earned if 50% of the wood products are certified by the Forest Stewardship Council.

The category of Indoor Environmental Quality allows 21.7% of the possible total points available. The goals include improving IAQ, improving occupant comfort, and providing views to the outside. With ASHRAE Standard 62–1999 as a prerequisite, an additional point is available for installing CO_2 monitoring devices in accordance with occupancies referenced in ASHRAE Standard 62–2001, Appendix C. A point is also available for implementing technologies that improve upon industry standards for air change effectiveness or that meet certain requirements for natural ventilation. Systems that provide airflow using both underfloor and ceiling plenums are suggested by LEED documentation as a potential ventilation solution. Points are available for developing and implementing IAQ management plans during construction and prior to occupancy. The requirements include using a Minimum Efficiency Reporting Value (MERV) 13 filter media with 100% outside air flush-out prior to occupancy. There are points available for use of materials that reduce the quantity of indoor air pollutants in construction caused by hazardous chemicals and by volatile organic compounds in adhesives, sealants, paints, coatings, composite wood products, and carpeting. A point is offered for provision of perimeter windows and another for individual control of airflow, temperature, and lighting for half of the non-perimeter spaces. Points are available for complying with ASHRAE Standard 55–1992 (Thermal Environmental Conditions for Human Occupancy), Addenda 1995, and installing permanent temperature and humidity control systems. Finally, points are gained for providing 75% of the spaces in the building with some form of daylighting and for providing direct line of-sight vision for 90% of the regularly occupied spaces.

In the category of Innovation and Design Process, 7.2% of the total possible points are available. The innovation credits offer the opportunity for projects to score points as a result of unusually creative design innovations, such as substantially exceeding goals of a given criteria or standard.

Assessing LEED-NC

The LEED-NC process has numerous strengths. Perhaps the greatest is its ability to focus the owner and design team on addressing select energy and environmental considerations early in the design process. The LEED design process brings architects, planners, energy engineers, environmental engineers, and IAQ professionals into the program at the early stages of design development. The team adopts a targeted LEED rating as a goal for the project. A strategy evolves based on selected criteria. The team members become focused on fundamental green design practices that have often been overlooked when traditional design development processes were employed.

Furthermore, the LEED program identifies the intents of the environmental initiatives. Program requirements are stated and acceptable strategies are suggested. Scoring categories attempt to directly address certain critical environmental concerns. When appropriate, the LEED-NC program defers to engineering and environmental standards developed outside of the USGBC. The components of the program provide accommodation for local regulations. Case study examples, when available and pertinent, are provided and described in the LEED literature. To expedite the process of documenting requirements, letter templates and calculation procedures are available to program users. The educational aspects of the program, which succinctly describe select environmental concerns, cannot be understated. A Web site provides updated information on the program with clarifications of LEED procedures and practice. The training workshops sponsored by the USGBC are instrumental in engaging professionals with a wide range of capabilities.

These considerations bring a high degree of credibility to the LEED process. Advocates of the LEED rating system have hopes of it becoming the pre-eminent U.S. standard for rating new construction that aspires to achieve a “green” label. To its credit, it is becoming a highly regarded standard and continues to gain prestige. Nick Stecky, a LEED accredited professional, firmly believes that the system offers a “measurable, quantifiable way of determining how green a building is.”[15]

Despite its strengths, the LEED-NC has observable weaknesses. The LEED-NC registration process can appear to be burdensome, and has been perceived as slowing down the design process and creating added construction cost. Isolated cases support these concerns. Kentucky’s first LEED-NC school, seeking a Silver rating, was initially estimated to cost over \$200/ft^2 (\$2152/m^2) compared to the local standard costs of roughly \$120/ft^2 (\$1290/m^2) for non-LEED construction. However, there are few comparative studies available to substantially validate claims of statistically significant cost impact. Alternatively, many case studies suggest that there is no cost impact as a result of the LEED certification process. It is also possible that the savings resulting from the use of certain LEED standards (e.g. reduced energy use) can be validated using life-cycle costing procedures. Regardless, LEED-NC fails as a one-size-fits-all rating system. For new construction, Kindergarten to 12th-grade (K-12), school systems in New Jersey, California, and elsewhere have adopted their own sustainable building standards.

There are other valid concerns in regard to the use of LEED-NC. In an era when many standards are under constant review, standards referenced by LEED are at times out of date. The ASHRAE Standard 90.1–1999 (without amendments) is referenced throughout the March 2003 revision of LEED-NC. However, ASHRAE 90.1 was revised, republished in 2001, and the newer version is not used as the referenced standard. Since design energy costs are used to score Energy and Atmosphere points, and energy use comparisons are baselined against similar fuels, cost savings from fuel switching is marginalized. In such cases, the environmental impact of the differential energy use remains unmeasured, since energy units are not the baseline criteria. There is no energy modeling software commercially available that has been specifically designed for assessing LEED buildings. LEED allows most any energy modeling software to be used, and each has its own set of strengths and weaknesses when used for LEED energy modeling purposes. It is possible for projects to comply with only one energy usage prerequisite, applying a standard already widely adopted, and still become LEED certified.

In fact, it is not required that engineers have specialized training or certification to perform the energy models. Finally, LEED documentation lacks System International (SI) unit conversions, reducing its applicability and exportability.

A number of the points offered by the rating system are questionable. While indoor environmental quality is touted as a major LEED concern, indoor mold and fungal mitigation practices, among the most pervasive indoor environmental issues, are not addressed and are not necessarily resolvable using the methodologies prescribed. It would seem that having a LEED-accredited professional on the team would be a prerequisite rather than an optional credit. Projects in locations with abundant rainfall or where site irrigation is unnecessary can earn a point by simply documenting a decision not to install irrigation systems. The ability of the point system to apply equally to projects across varied climate classifications and zones is also questionable and unproven.

While an M&V credit is available, there is no requirement that a credentialed measurement and verification professional be part of the M&V plan development or the review process. Without the rigor of M&V, it is not possible to determine whether or not the predictive preconstruction energy modeling was accurate. The lack of mandates to determine whether or not the building actually behaves and performs as intended from an energy cost standpoint is a fundamental weakness. This risks illusionary energy cost savings. Finally, the M&V procedures in the 2001 IPMVP have undergone revision and were not state-of-the-art at the time that LEED-NC was updated in May 2003. For example, there is no longer a need to exclude Option A as an acceptable M&V alternative.

The LEED process is not warranted and does not necessarily guarantee that in the end, the owner will have a "sustainable" building. While LEED standards are more regionalized in locations where local zoning and building laws apply, local regulations can also preempt certain types of green construction criteria. Of greater concern is that it is possible for a LEED certified building to devolve into a building that would lack the qualities of a certifiable building. For example, the owners of a building may choose to remove bicycle racks, refrain from the purchase of green energy after a couple of years, disengage control systems, abandon their M&V program, and remove recycling centers—yet retain the claim of owning a LEED certified building.

Conclusion

The ideal of developing sustainable buildings is a response to the environmental impacts of buildings and structures. Developing rating systems for structures is problematic due to the often subjective nature of the concepts involved, the ambiguity or lack of certain standards, and the local aspects of construction. While there are a number of assessment systems for sustainable buildings used throughout the developed world, LEED-NC is becoming a widely adopted program for labeling and rating newly constructed "green" buildings in the United States. Using a point-based rating system, whereby projects are credited for their design attributes, use of energy, environmental criteria, and the application of select standards, projects are rated as Certified, Silver, Gold, or Platinum.

The LEED-NC program has broad applicability in the United States and has been proven successful in rating roughly 150 buildings to date. Its popularity is gaining momentum. Perhaps its greatest strength is its ability to focus the owner and design team on energy and environmental considerations early in the design process. Today, there are over 1700 projects that have applied for LEED certification. Due to the program's success in highlighting the importance of energy and environmental concerns in the design of new structures, it is likely that the program will be further refined and updated in the future to more fully adopt regional design solutions, provide means of incorporating updated standards, and offer programs for maintaining certification criteria. It is likely that the LEED program will further expand, perhaps offering a separate rating program for K-12 educational facilities. Future research will hopefully respond to concerns about potential increased construction costs and actual energy and environmental impacts.

References

1. U.S. Green Building Council. *LEED green building rating system.* 2004.
2. Spirn, A.W. *The Granite Garden: Urban Nature and Human Design;* Basic Books: New York, 1984.
3. United Nations. *Indicators of Sustainable Development: Guidelines and Methodologies;* United Nations: New York, 2001; 38.
4. Roosa, S.A. *Energy and Sustainable Development in North American Sunbelt Cities;* RPM Publishing: Louisville, KY, 2004.
5. Koeha, T. What is Sustainability and Why Should I Care? *Proceedings of the 2004 world energy engineering congress,* Austin, TX, Sept 22–24, 2004; AEE: Atlanta 2004.
6. URS. http://www.urseurope.com/services/engineering/engineering-breeam.htm (accessed Feb 2005).
7. iiSBE. International Initiative for a Sustainable Built Environment http://iisbe.org/iisbe/start/iisbe.htm (accessed Feb 2005).
8. United States Environmental Protection Agency. *Join energy star—Improve your energy efficiency.* http://www.energystar.gov. (accessed June 2003).
9. Gonchar, J. Green building industry grows by leaps and bounds; Engineering News-Record; 2003.
10. Boucher, M. Resource efficient buildings—Balancing the bottom line. *Proceedings of the 2004 world energy engineering congress,* Austin, TX, Sept 22–24, 2004; AEE: Atlanta 2004.
11. U.S. Green Building Council. *Green Building Rating System for New Construction and Major Renovations Version 2.1 Reference Guide;* May 2003; 6.
12. U.S. Green Building Council. *LEED—Certification Process.* http://www.usgbc.org/LEED/Project/certprocess.asp. (accessed Oct 2004).
13. U.S. Green Building Council. *Green Building Rating System for New Construction and Major Renovations Version 2.1:* Nov 2002.
14. U.S. Green Building Council. *Green Building Rating System for New Construction and Major Renovations Version 2.1:* Nov 2002; 10.
15. Stecky, N. Introduction to the ASHRAE greenguide for LEED, *Proceedings of the 2004 world energy engineering congress,* Austin, TX, Sept 22–24 2004; AEE: Atlanta 2004.

39

Nanomaterials: Regulation and Risk Assessment

Steffen Foss Hansen,
Khara D. Grieger,
and Anders Baun

Introduction

The topics of regulation and risk assessment of nanomaterials have never been more relevant and controversial in Europe than they are at this point in time. As the first major piece of legislation to be amended in Europe, the cosmetics legislation was adopted in 2009 requiring all nanomaterial-containing cosmetics to be labeled after 2013 and producers to provide a safety assessment of the nanomaterial used.[1,2]

Concurrently with the recasting of various pieces of legislation, such as the Novel Foods Regulation, the European Commission has commissioned an expert-/multistakeholder investigation of whether nanospecific amendments are needed to the current technical guidelines on substance identification and chemical safety assessment, which lie at the core foundations of the European Chemical legislation known as REACH—Registration, Evaluation, Authorization, and Restriction of Chemicals.[3,4] It is the major piece of legislation concerning regulating the manufacturing and applications of nanomaterials, although the text in REACH itself has only been subject to minor changes thus far. A number of other pieces of legislation relevant to the manufacturing, use, and disposal of nanomaterial and products have furthermore not been subject to any nanospecific changes, although they might be revised in the future.

In the following, some of the major pieces of legislation relevant for the regulation of nanomaterials in Europe will be presented. Examples of both horizontal regulation as well as subject-specific legislation will be given. Some of these have yet to take nanospecific issues into consideration, and the focus will therefore be at explaining the limitations of these in handling nanomaterials. For others, nanospecific aspects have recently been taken into consideration and for these the focus will be at explaining how this has been done and what kinds of challenges still need to be addressed.

Chemical risk assessment plays a crucial role in many of these pieces of legislation, and hence a short introduction and discussion of the applicability of chemical risk assessment to nanomaterials will be included.

Registration, Evaluation, and Authorization of Chemicals (Reach)

One of the key pieces of European legislation affecting nanomaterials is the European chemical regulation known as REACH, which went into force in mid-2007.[5] REACH prescribes

1. The registration of chemicals commercialized by manufacturers and importers in Europe as well as the collection of data on their use and toxicity.
2. The evaluation and examination by governments of the need for additional testing and regulation of chemicals.
3. That authorization has to be sought and given to manufacturers in order for them to use chemicals of high concern.
4. European Union (EU)-wide restrictions or complete ban of certain chemicals that cannot be used safely.

The REACH regulation replaced more than 40 other directives and subsequently shifted the responsibility in the registration and authorization process of REACH onto manufacturers and importers (including downstream users of chemicals) to provide data of uses and hazard information. Industry, furthermore, has to show that chemicals of high concern can be used safely. The evaluation and restriction process is still the responsibility of the national authorities, the newly established European Chemical Agency, and the European Commission.

Registration of all the commercialized chemicals in the European market is a tremendous task that is expected to occur gradually. Substances produced or imported in the highest volumes or of the greatest (known) concern are to be registered first. Substances produced or imported in more than 1000 tonnes per year per manufacturer or importer had to be registered by November 30, 2010, by the latest date. This was also the case for substances marketed in 100 tons/yr that have been classified as very toxic to aquatic organisms and for substances produced/imported in more than 1 ton/yr and which have been classified as Category 1 or Category 2 carcinogens, mutagens, or reproductive toxicants. Furthermore, substances entering the European market in yearly quantities above 100 tons, and 10 tons per producers or importers have to be registered by June 1, 2013, and June 1, 2018, respectively.[5]

REACH does not specifically mention nanomaterials, but does cover chemicals in all their physical–chemical states, using the following definition of a substance: "a chemical element and its compounds in the natural state or obtained by any manufacturing process, including any additive necessary to preserve its stability and any impurity deriving from the process used, but excluding any solvent which may be separated without affecting the stability of the substance or changing its composition."[5] Therefore, REACH is formally the relevant legislative frame for industrially used nanomaterials, and the exemption registration of carbon and graphite was redrawn in 2008 to address concerns raised about carbonaceous nanomaterials.[6,7] Companies will now have to register these materials if produced in quantities above 1 ton per producer or manufacturer per year. However, for a number of nanomaterials, it is not evident whether a nanoequivalent of a substance with different physicochemical and (eco)toxicological properties from the bulk substance would be considered as the same or as another substance under REACH.[8]

If a nanomaterial is considered to be a different substance under REACH, hazard information specifically related to the nanoform of the substance would have to be generated for the registration, if produced in more than 1 ton/yr. On the other hand, if a nanomaterial is considered to be the same as a registered bulk material, hazard information data generated for the registration might not be directly relevant for the nanoform of the substance and hence open to discussion.[8,9] In response to these concerns, the European Commission has launched a multistakeholder project on nanomaterials to look into substance identification under REACH in order to get recommendations on whether the nanoform of a substance should be considered different from the bulk form of the substance.[3,4]

If manufacturers and importers produce or import nanomaterials in volumes of more than 10 tons/yr and if it meets the criteria for classification as dangerous or a PBT (persistent, bioaccumulative, and toxic) or vPvB (very persistent and very bioaccumulative), a chemical safety assessment is required that includes information about uses, (eco)toxicological information, exposure assessment, and risk characterization(s).

Thus far, no nanomaterial has been classified as PBT or as vPvB, but if it was to be it is highly unclear how companies should do a chemical safety assessment. Both the Commission of the European Communities[10] as well as the its Scientific Committee on Emerging and Newly Identified Health Risks (SCENIHR)[11] have pointed out that current test guidelines in REACH are based on conventional methodologies for assessing chemical risks and may not be appropriate for assessing risks associated with nanomaterials.

It should be noted that a chemical safety assessment can also be required if a nanomaterial is selected for further evaluation by a member state or by the European Chemicals Agency due to specific concerns; or if a substance is a CMR (carcinogenic, mutagenic, or toxic for reproduction), PBT, vPvB, ED (endocrine disrupting), or substance of equivalent concern.

EU Water Framework Directive

Whereas REACH deals with the manufacturing and import of chemicals, the EU Water Framework Directive (WFD) deals with improving water quality and reducing dangerous chemicals in European river basins. The key aim of the WFD, which was adopted in 2000, is to promote long-term sustainable water use, preventing further deterioration of surface waters, transitional waters, coastal waters, and groundwater, and to protect and enhance the status of aquatic ecosystems with regard to their water needs, terrestrial ecosystems, and wetlands directly depending on the aquatic ecosystems,[12] Article 1.

The WFD establishes water management by a river basin approach with cooperation and joint objective setting across member state borders and even in some cases beyond the EU territory. Geographical and hydrological formation of each river basin determines which member states need to establish and implement a so-called river basin management plan. The river basin management plan, which needs to be updated every 6 years, specifies the measures to be taken to meet the environmental objectives for surface waters, for groundwater, and for protected areas. The WFD prescribes the setting of the environmental quality standards ensuring the general protection of the aquatic ecology, specific protection of unique and valuable habitats, and protection of drinking water resources and bathing water. For instance, for surface waters, member states shall implement necessary measures to prevent deterioration, and promote restoration of artificial and heavily modified water bodies with the aim of achieving "good ecological potential" and "good surface water chemical status" in 2015 by the latest. This has to be done along with a progressive reduction of pollution from a set of "priority substances" and discontinue emissions of priority hazardous substances,[12] Article 4.

For all surface waters, the WFD set a number of "general requirements for ecological protection" as well as a "general minimum chemical standard" and defines "good ecological status" and "good chemical status" in terms of the quality of the biological community, the hydrological characteristics, and the chemical characteristics,[12] Article 4.

The definition of "good chemical status" is especially relevant in regard to nanomaterials as it is defined in terms of compliance with all the quality standards established for chemical substances at the European level. For "priority substances," member states are required to set environmental quality standards (EQSs) to monitor the chemical status of a water body (European Parliament and the Council of the European Union (EP & CEU),[12] Article 16). Thus, the EQS is taken as concentration below which the chemical status is referred to as "good" in the WFD terminology (European Parliament and the Council of the European Union (EP & CEU),[12] Article 2). Even for the so-called priority hazardous substances, only a few EQSs have been set, but more substances will follow with a specific focus at substances that are toxic toward humans and/or aquatic organisms, compounds with a widespread environmental distribution, and those that are discharged in significant quantities.

A key question in regard to WFD is whether nanomaterials are possible candidates as priority substances.[13] In favor of this speaks the widespread and diffuse use of nanoparticles in a range of consumer products along with the hazard characteristics of some nanomaterials such as functionalized carbon nanotubes (CNTs), nanoscale silver, and zinc oxide. Some applications of nanomaterials furthermore involve direct contact with the water cycle, e.g., in relation to their use for water disinfection[14] and wastewater treatment,[15] as well as in regard to the direct use for treating soil and groundwater contamination.[16]

If a given type of nanoparticles is included in the list of priority substances in the future based on environmental occurrence or hazard information, an EQS will have to be defined.[12] To derive an EQS for a priority substance, the WFD outlines that test results from both acute and chronic ecotoxicological standard tests should be used for the "base set" organisms, i.e., algae and/or macrophytes, crustacean, and fish. Estimating EQS for nanoparticles is currently hampered by lack of ecotoxicological data even for the most tested nanoparticles such as C60, CNTs, TiO_2, ZnO, and Ag. For instance, the degradability of C60 and CNTs and their ability to bioaccumulate in the aquatic environment remains to be studied, making it virtually impossible to set an EQS for these two kinds of nanoparticles.[13]

Not only are the number of studies very limited, but the number of tested taxa is also too few to be used in the context of setting an EQS. The reliability and interpretation of the available ecotoxicity data is furthermore impeded as a result of factors such as particle impurities, suspension preparation methods, release of free metal ions, and particle aggregation.[13,17,18]

Besides these issues, mainly related to the lack of relevant data, it is also questionable whether the principles for deriving EQSs for chemicals can be directly transferred to nanoparticles. The setting of EQS is based on a chemical safety assessment similar to the one required under REACH and, as noted above, the European Commission's SCENIHR have pointed out that amendments have to be made to the guidelines for chemical safety assessment.[11,19]

Another manner in which nanomaterials could meet the criteria to be included in a WFD list of priority substances is if there is "evidence from monitoring of widespread environmental contamination".[12] However, when it comes to nanomaterials, monitoring in natural waters represents some profound challenges.[20,21] While applicable methods for in situ monitoring remain to be developed and refined,[22] it is also challenging to set up a reliable monitoring program for nanoparticles since a number of issues still remain to be resolved, e.g., choice of suitable sampling materials, preconcentration/fractionation methods, and analytical methods to characterize and quantify collected particles.[21] Despite significant progress in recent years, reliable methods are not yet available to determine nanoparticle identity, concentrations, and characteristics in complex environmental matrices, such as water, soil, sediment, sewage sludge, and biological specimens.

Pharmaceutical Regulation

Liposomes, polymer–protein conjugates, polymeric substances, or suspensions are examples of well-described and understood medicinal products containing nanoparticles and have been given marketing authorizations within the EU under the existing regulatory framework.[23–26]

As in the case of REACH, nanomedicine and nanomaterials are not specifically mentioned in the EU legislation on medicinal products and devices, tissue engineering, and other advanced therapies. Although the scope of the various EU regulations and directives that constitutes this framework covers nanomedicine, they have been accused of being too general, non-specific, and fraught with difficulties in case of complex drugs.[27,28] Given this, it does seem that it is generally believed that the regulatory framework for medicine covers medical products based on nanotechnology, and that the extensive pre-market safety assessment of medicine in general is sufficient to ensure that the benefits outweigh any identified risks or the adverse side effects.[29,30]

Concerns have, however, been raised that the risk assessment, safety, and quality requirements for medicine may not be designed to address nanomedicine and medical devices based on nanotechnology, as these have to be fulfilled by conformity to established quality systems and published product standards. This might be especially true for novel applications such as nanostructure scaffolds for tissue replacement, nanostructures enabling transport across biological barriers, remote control of nanoprobes, integrated implantable sensory nanoelectronic systems, and chemical structures for drug delivery and targeting of disease.[31]

Currently, the mechanism of action is key to decide whether a product should be regulated as a medicinal product or as a medical device. This could be problematic when it comes to many novel applications of nanomedicinal products as they are likely to span regulatory boundaries between medicinal products and medical devices.[29,31] This is due to the notion that they may exhibit a complex mechanism of action combining mechanical, chemical, pharmacological, and immunological properties, and combining diagnostic and therapeutic functions.

For new marketing authorization applications of pharmaceuticals, an environmental risk assessment has to be provided, which involves a rough calculation of the predicted environmental concentration (PEC) for surface water. Actions have to be initiated if the PEC is predicted to surpass 0.01 ppb.[32] However, this threshold cannot be interpreted as a safe concentration, and it is not based on a scientific evaluation.[32] It could furthermore be problematic when it comes to nanomedicine, as concentration in terms of mass per volume might not be the relevant metric to characterize the environmental hazard of nanomaterials.[33–35]

Nanofood Regulation

Food and food packaging are regulated by a number of directives and regulations in the EU, such as the EU Food Law Regulation and the EU Novel Foods Regulation.[36] As an overarching principle, all food are required to be safe and this overarching principle of safety applies to all foods and food packaging that contain nanomaterials. This has, however, been criticized for being too loose.[37]

During the recent discussion related to the update of regulation regarding food additives, the European Parliament's Committee on Environment, Public Health, and Food Safety stated that it wanted separate limit values for nanotechnologies and that the permitted limits for an additive in nanoparticle form should not be the same as when it is in traditional form.[38] This demand, however, never made into the actual regulation and the final adopted regulation on food additives is limited to requiring that food additives that have been produced via nanotechnology or consist of/or include materials fulfill a number of criteria before it can be include on the list of approved food, food additives, food enzymes, and food flavorings. Nanotechnology and nanomaterials are not defined in the regulation, but these criteria include what use does not pose a safety concern to the health of the consumer at the level of use proposed on the basis of available scientific evidence. Furthermore, there has to be a reasonable technological need that cannot be achieved by other economically and technologically practicable means, and using the food additives should entail consumer advantages and benefits.[39]

Another important piece of legislation in regard to food regulation in the EU is the Novel Foods Regulation. This regulation requires mandatory premarket approval of all new ingredients and products. In 2008, the European Commission adopted a proposal to revise the Novel Foods Regulation with the purpose of improving the access of new and innovative foods to the EU market.[40] The definition of novel foods was broadened to include those modified by new production processes, such as nanotechnology and nanoscience, which might have an impact on the food itself. This proposal is currently being discussed in the European Parliament and is going through what is known as a "third reading," and has to be adopted after a co-decision wherein both the Council of the European Union and the Parliament has to agree on the final text of the regulation. If agreement cannot be reached, it goes to conciliation.

There are a number of areas on which the European Parliament and the Council of the European Union disagree in regard to nanomaterials. The requirement of having mandatory labeling is also controversial and so is the issue of whether to have premarket safety testing of nanotechnology and nanomaterials in food and packaging.[41]

In the first line of revisions suggested to the Novel Foods Regulation, both the Council and the Parliament mention the lack of adequate information and lack of test methods for assessing the risks of nanomaterials.[41] Once the European Commission receives an application for authorization of a novel food, the European Food Safety Authority (EFSA) is responsible for the evaluation of whether a novel food and its use as an ingredient presents a danger to or misleads consumers. By regulation, the EFSA is required to provide assessment on the composition, nutritional value, metabolism, intended use, and the level of microbiological and chemical contaminants. Studies on the toxicology, allergenicity, and details of the manufacturing process may also be considered. No distinction is, however, made in regulation in regard to particle size, and hence nanoparticles will not require new safety assessments if the substance has already been approved in bulk form.

Risk Assessment of Nanomaterials

Three different kinds of limitations have been identified in various independent analysis of the applicability of existing regulatory frameworks when it comes to nanomaterials.

The first category of limitations are related to the limitations of definitions of what qualifies as a "substance," "novel food," etc., when it comes to nanomaterials. For instance, does the definition of a chemical substance cover both the bulk from as well as the nanoform of the substance, and does any given application of nanotechnology to manufacture a given food fall under the definition of a novel food? This issue is currently being discussed in a multistakeholder expert working group; however, this has failed to reach a consensus.[42]

In the second category fall requirements triggered by thresholds values not tailored to the nanoscale, but based on bulk material. For instance, for pharmaceuticals, the environmental concentration of medical products has to be estimated before marketing, and if it is below 0.01 ppb and "no other environmental concerns are apparent," no further actions are to be taken for the medical product in terms of environmental risk assessment.[32] Such a predefined action limit could potentially be problematic since the new properties of nanobased products are expected to also affect their environmental profiles, and this problem has yet to be addressed.[35]

The third category of limitations are related to lack of metrological tools, (eco)toxicological data, and environmental exposure limits as required by, e.g., REACH, the pharmaceuticals regulation, and the recast of the Novel Foods Regulation. The availability of (eco)toxicological data and chemical risk assessments is necessary to support existing legislation.

In regard to REACH, companies are urged to use already existing guidelines when performing chemical risk assessments, despite the fact that both the European Commission[10] and its SCENIHR,[11] as well as others,[9,10] have pointed out that current test guidelines supporting REACH are based on conventional methodologies for assessing chemical risks and may not be appropriate for the assessment of risks associated with nanomaterials.

Chemical risk assessment consists of four elements i.e., hazard identification, dose–response assessment, exposure assessment, and risk characterization. In Europe, legislation for controlling the production, use, and release of chemical substances is based on chemical safety assessment or risk assessment, as described in detail in the "Guidance on Information Requirements and Chemical Safety Assessment".[43] The guidance totals a staggering number of pages and is issued by the ECHA to help companies carry out chemical safety assessments. It includes extensive technical details for conducting hazard identification, dose (concentration)–response (effect) assessment, exposure assessment, and risk characterization in relation to human health and the environment.[43] Each of these four elements holds a number of limitations that are not easily overcome despite the fact that a lot of effort is being put into investigating the applicability of each of these four elements.

Hazard Identification of Nanomaterials

Toxicity and ecotoxicity have been reported on for multiple nanoparticles (metal and metal oxide nanoparticles, carbonaceous nanomaterials, and quantum dots) in scientific studies; however, many of these need further confirmation. Univocal hazard identification is currently impossible as it is hard to systematically link reported nanoparticle properties to the observed effects.

For instance, in regard to multiwalled CNTs (MWCNTs), Poland et al.[44] compared the toxicity of four kinds of MWCNTs of various diameters, lengths, shape, and chemical composition by exposing the mesothelial lining of the body cavity of three mice with 50 mg MWCNT for 24 hr or 7 days. This method was used as a surrogate for the mesothelial lining of the chest cavity. It was found that long MWCNTs "produced length-dependent inflammation, FBGCs, and granulomas that were qualitatively and quantitatively similar to the foreign body inflammatory response caused by long asbestos." Only the long MWCNTs caused significant increase in polymorphonuclear leukocytes or protein exudation. The short MWCNTs failed to cause any significant inflammation at 1 day or giant cell formation at 7 days. Poland et al.[44] also found that the water-soluble components of MWCNT did not produce significant inflammatory effects 24 hr after injection, which rules out that residue metals were the cause of the observed effects, as others previously had speculated on the basis on in vitro studies.[45,46] The findings by Poland et al.[44] have since then been supported by Ma-Hock et al.[47] and Pauluhn et al.[48] in 90-day inhalation toxicity studies.

Less work has been done in regard to exploring the ecotoxicological aspects of nanomaterials, but a number of significant studies have been published.

In 2004, Oberdorster[49] published the first ecotoxicological study and reported observed significant increase in lipid peroxidation of the brain of juvenile largemouth bass after exposure to uncoated fullerenes (99.5%) in concentrations of 0.5 and 1 ppm after exposure for 48 hr. C60 was dissolved in tetrahydrofuran (THF), which have since then led to some discussion about whether C60 or the THF was responsible for the effects observed.[50,51] The use of THF is no longer recommended.[18]

In regard to CNTs, Templeton et al.[52] compared "as prepared" single-walled CNTs (SWCNTs) with electro- phorectically purified SWCNTs and the fluorescent fraction of nanocarbon by-products. They observed an average cumulative life cycle mortality of 13 ± 4%, while mean life cycle mortalities of 12 ± 3%, 19 ± 2%, 21 ± 3%, and 36 ± 11% were observed for 0.58, 0.97, 1.6, and 10 mg/L. Exposure to 10 mg/L showed: 1) significantly increased mortalities for the naupliar stage and cumulative life cycle; 2) a dramatically reduced development success to 51% for the nauplius to copepodite window, 89% for the copepoddite to adult window, and 34% overall for the nauplius to adult period; and 3) a significantly depressed fertilization rate averaging only 64 ± 13%.

A number of studies have furthermore highlighted the need to investigate the potential interactions with existing environmental contaminants or what has become known as the "Trojan horse effect." For instance, Baun et al.[53] found that the toxicity of phenanthrene was increased toward algae and crustaceans following sorption to C60 aggregates. In contrast, Baun et al.[53] found that the toxic effect of pentachlorophenol decreased when C60 was added. After studying the ecotoxicity of cadmium to algae in the presence of 2 mg/L TiO_2 nanoparticles of three different sizes, Hartmann et al.[17] found that the presence of TiO_2 in algal tests reduced the toxicity of cadmium. This is thought to be due to decreased bioavailability of cadmium resulting from sorption/complexation of Cd^{2+} ions to the TiO_2 surface. However, the observed growth inhibition was, however, greater for the 30 nm TiO_2 nanoparticles than could be explained by the concentration of dissolved Cd(II) species alone, which indicates a possible carrier effect, or combined toxic effect of TiO_2 nanoparticles and cadmium.

Dose–Response Relationship in Regard to Nanomaterials

In regard to the second element of chemical risk assessment, it is fundamental that a dose–response relationship can be established so that no-effect concentrations or no effect levels need to be predicted or derived. It is unclear whether a no-effect threshold can be actually be established and what the best hazard descriptor(s) of nanoparticles is, and what the most relevant dose metrics and the what the most sensitive endpoints are. Several studies have reported observing a dose–response relationship. This goes for, especially, in vitro studies on, among others, C60, SWCNTs and MWCNTs, and various forms of nanometals. Normally, dose refers to "dose by mass"; however, based on the experiences gained in biological tests of nanoparticles, it has been suggested that biological activity of nanoparticles might not be mass dependent, but dependent on physical and chemical properties not routinely considered in toxicity studies.[54] For instance, Oberdorster and col-leagues[55,56] and Stoeger and colleagues[57,58] found that the surface area of the nanoparticles is a better descriptor of the toxicity of low-soluble, low-toxicity particles, whereas Wittmaack[59,60] found that the particle number worked best as dose metrics. Warheit et al.[61,62] found that toxicity was related to the number of functional groups in the surface of nanoparticles.

Exposure Assessment

Completing a full exposure assessment requires extensive knowledge about, among others, manufacturing conditions, level of production, industrial applications and uses, consumer products and behavior, and environmental fate and distribution. Such detailed information is not available, and thus far no full exposure assessment has been published for any one or more nanomaterials. This may partly be due to difficulties in monitoring nanomaterial exposure in the workplace and the environment, and partly due to the fact that the biological and environmental pathways of nanomaterials are still largely unexplored.[63] Some efforts have been made to assess occupational, consumer, and environmental exposure, however, both to assess the level of exposure and to assess the applicability of current exposure assessment methods and guidelines.

These are, however, hampered by the paucity of knowledge, lack of access to information, difficulties in monitoring nanomaterial exposure in the workplace and the environment, and by the fact that the biological and environmental pathways of nanomaterials are still largely unexplored. Hence, they should be seen as "proof of principle" rather than actual assessment of the exposure.[64–67]

Risk Characterization

All the information from the first three elements of the risk assessment come together in the fourth and final element of chemical risk assessment, namely risk characterization.[63] In the risk characterization process, exposure levels are compared with quantitative or qualitative hazard information, then suitable predicted no-effect concentrations or derived no-effect levels are determined in order to decide if risks are adequately controlled.[43]

Often, risk characterization boils down to the estimation of a risk quotient. For the environment, this is, for instance, defined as the PEC/predicted no effect concentration (PNEC). If the risk quotient is <1, no further testing or risk reduction measures are needed according to the European Chemical Agency.[43] If it is >1, further testing can be initiated to lower the PEC/PNEC ratio. If that is not possible, risk reduction could be implemented.

A number of studies reported having completed—or attempted to complete—risk assessments of various nanomaterials such as CeO_2, TiO_2, C60, and CNTs.[18,67–69] For instance, in regard to the use of CeO2-based diesel fuel additive in the United Kingdom, Park et al.[67] assessed the risk of CeO_2 causing pulmonary inflammation. First, they estimated an internal dose of $3.8 \times 10^{-7}\,cm^2/cm^2$ by converting the retained dose into surface area units and then dividing by the area of the proximal alveolar region of the lung. Then, they compared this value to the highest noobserved-effect level found in a number of in vitro toxicity studies. This value was $26.75\,cm^2/cm^2$. Assuming that in vitro exposure data can be accurately projected to the in vivo situation, Park et al.[67] concluded that "it is highly unlikely that exposure to cerium oxide at the environmental levels (from both monitored and modeled experimental data) would elicit pulmonary inflammation."

Mueller and Nowack[69] reported having completed the first quantitative risk assessment of nanoparticles in the environment. In a first attempt to derive PEC values, Mueller and Nowack used the threshold concentrations of 20 and 40 mg/L reported in the literature for nano-Ag on *Bacillus subtilis* and *Escherichia coli,* and considered it to be equivalent to a no-observed-effect concentration. For nano-TiO_2 and CNT, the lowest value found in the literature was <1 mg/L for algae, daphnia, and fish.[69] Applying the assessment factor of 1000, the PNEC in water was found to be 0.04, <0.001, and <0.0001 mg/L for nano- Ag, nano-TiO_2, and CNT, respectively. Combining these PNEC values with the predicted exposure, Mueller and Nowack[69] calculated the environmental concentrations in Switzerland for nano-Ag, nano-TiO_2, and CNTs stemming from textiles, cosmetics, coatings, plastics, sports gear, electronics, etc. Assuming worse-case exposure levels, Mueller and Nowack[69] found that the risk quotient for nano-Ag and CNT is less than one-thousandth, and they state that their modeling suggests that currently little or no risk is to be expected from nano-Ag and CNT to organisms in water and air. Nano-TiO_2, on the other hand, might pose a risk to organism in water—according to Mueller and Nowack[69]—with risk quotients ranging from >0.7 to >16. The PNEC for soil could not be determined due to lack of information.

Despite the preliminary risk characterizations by Park et al.,[67] Mueller and Nowack,[69] Shinohara et al.,[68] and Stone et al.,[18] it is important to realize that risk characterization critically involves reflection of the data behind each step and determining what the overall risk will be.[63] As elaborated on previously, each of the three first steps of risk assessment holds a number of challenges, and since risk characterization is the fourth and final step where all the information is to come together, the sum or maybe even the power all of these limitations are conveyed to calculating risk quotients for nanomaterials under REACH.[9]

Revisions of the Technical Guidance of Risk Assessment

The European Commission has commissioned an expert/multistakeholder investigation of whether nanospecific amendments are needed to the current technical guidelines on chemical safety assessment. This has to develop, among others, specific advice on

1. How REACH information requirements on intrinsic properties of nanomaterials can be fulfilled
2. The appropriateness of the relevant test methods for nanomaterials
3. The possible specific testing strategies, if relevant
4. Information needed for safety evaluation and risk management of nanomaterials (especially, information beyond the current information requirements under REACH)
5. How to do exposure assessment for nanomaterials, hazard, and risk characterization for nanomaterials[3,4]

The latter will involve threshold/non-threshold considerations, analysis of existing evidence related to setting limit values for nanomaterials, identification of critical items for dose description (mass, number concentration, surface area, particle size(s) etc.), whether and how no-effect-levels for health and the environment could be established, and finally development of recommendations on the feasibility of whether categorization of nanomaterials (e.g., different types of CNTs) in the hazard assessment is compatible with the exposure assessment parameters/metrics in order to prepare a meaningful risk characterization.[3,4]

In regard to novel food, EFSA published a scientific opinion in 2008 on the potential risks arising from the use of nanotechnology in food, concluding that nanotechnology aspects shall be considered when risk assessment guidance documents in the food and feed area are reviewed, and, among others, recommend that risk assessment of nanomaterials in the food and feed areas should consider the specific properties of nanomaterials in addition to those common to the equivalent non-nanoforms.[70] Recently, EFSA closed for public comments on a draft guidance on risk assessment concerning potential risks arising from applications of nanoscience and nanotechnologies to food and feed. This guidance holds practical advice on how to complete risk assessments of nanomaterials used in food and food products.[71]

In the light of the limitations of chemical risk assessment, a number of alternative or complementary tools and methods, such the precautionary matrix[72] and multicriteria decision analysis,[73] have been proposed recently. Many of them hold great promises, but they need further evaluation and validation.[74]

Conclusion

In this entry, we presented a number of major pieces of legislation such as REACH, the WFD, and the Novel Foods Regulation, and discussed their relevance and limitations in regard to nanomaterials. Only a limited number of EU regulations, directives, etc., actually mention nanotechnology and/nanomaterials. In general, there seem to be three overall challenges when it comes to current regulation of nanomaterials: 1) limitations in regard to terminology and definitions of key terms such as a "substance," "novel food," etc.; 2) safety assessment requirements triggered by thresholds values not tailored to the nanoscale, but based on bulk material and; 3) limitations related to lack of metrological tools, (eco) toxicological data, and environmental exposure limits as required by, e.g., REACH, WFD, the pharmaceuticals regulation, and the recast of the Novel Foods Regulation. Chemical risk assessment provides a fundamental element in support of existing legislation. Risk assessment is normally said to consist of four elements, i.e., hazard identification, dose–response assessment, exposure assessment, and risk characterization. Each of these four elements holds a number of limitations that are not easily overcome, although a lot of effort is being put into investigating the applicability of each of them. However, political decisions to revise substance definition and the current thresholds that trigger safety evaluation are still needed as these are not tailored to the nanoscale, but based on bulk material.

References

1. European Parliament and Council of the European Union (EP & CEU). Regulation (EC) No 1223/2009 of the European Parliament and of the Council of 30 November 2009 on cosmetic products (1). Off. J. Eur. Union L342, vol. 52, 22 December 2009. ISSN 1725–2555. L 342/59 L 342/209.
2. Bowman, D.M.; Calster, G.v.; Friedrichs, S. Nanomaterials and regulation of cosmetics. Nat. Nanotechnol. **2009**, *5*, 92.
3. Safenano. REACH-NanoInfo: Rip-oN 2, 2010a. Available at http://www.safenano.org/REACH nanoInfo.aspx (accessed October 2010).
4. Safenano. REACH-NanoHazEx: Rip-oN 3, 2010b. Available at http://www.safenano.org/REACHnanoHazEx.aspx (accessed October 2010).
5. European Parliament and the Council of the European Union (EP & CEU). Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/45/ EC and repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC. 30.12.2006. Off. J. Eur. Union L 2006, 396/1-L 396/849.
6. Führ, M.; Hermann, A.; Merenyi, S.; Moch, K.; Möller, M. Legal Appraisal of Nanotechnologies, 2007. Available at http://www.umweltdaten.de/publikationen/fpdf-l/3198.pdf (accessed December 2010). Umwelt Bundes Amt UBA-FB 000996. ISSN 1862–4804.
7. C&EN. Carbon losses its exemption statues under REACH. Chem. Eng. News 2008, June 23, p.9.
8. Chaundry, Q.; Blackburn, J.; Floyd, P.; George, C.; Nwaogu, T.; Boxall, A.; Aitken, R. *A Scoping Study to Identify Gaps in Environmental Regulation for the Products and Applications of Nanotechnologies;* Department for Environment, Food and Rural Affairs: London, 2006.
9. Hansen SF. *Regulation and Risk Assessment of Nanomaterials—Too Little, Too Late?*, PhD Thesis; Technical University of Denmark: Kgs. Lyngby, Denmark, 2009. Available at http://www2.er.dtu.dk/publications/fulltext/2009/ENV2009–069.pdf (accessed April 2009).

10. Commission of the European Communities (CEC). *Communication from the Commission to the European Parliament, the Council and the European Economic and Social Committee Regulatory Aspects of Nanomaterials, [Sec(2008) 2036] Com(2008), 366 final;* Commission of the European Communities: Brussels, Belgium, 2008a.
11. Scientific Committee for Emerging and Newly Identified Health Risks (SCENIHR). *The Appropriateness of the Risk Assessment Methodology in Accordance with the Technical Guidance Documents for New and Existing Substances for Assessing the Risks of Nanomaterials*; European Commission of Health and Consumer Protection Directorate-General: Brussels, Belgium, 2007.
12. European Parliament and the Council of the European Union (EP & CEU). Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for Community action in the field of water policy. Off. J. Eur. Commun. **2000**, *L 327,* 1–72, Brussels, Belgium.
13. Baun, A.; Hartmann, N.B.; Greiger, K.D.; Hansen, S.F. Setting the limits for engineered nanoparticles in European surface waters. J. Environ. Monit. **2009**, *11,* 1774–1781.
14. Li, Q.; Mahendra, S.; Lyon, D.Y.; Liga, M.V.; Li, D.; Alvarez, P. Antimicrobial nanomaterials for water disinfection and microbial control: Potential applications and implications. Water Res. **2008**, *42,* 4591–4602.
15. Nano Iron. *Technical Data Sheet.* NANOFER 25S. Ra- jhrad, Czech Republic, 2009. http://www.nanoiron.cz/en/?f1/4technical_data (accessed February 2011).
16. Li, X.; Elliott, D.W.; Zhang, W.X. Zero-valent iron nanoparticles for abatement of environmental pollutants: Materials and engineering aspects. Crit. Rev. Solid State Mater. Sci. **2006**, *31,* 111–122.
17. Hartmann, N.B.; Von der Kammer, F.; Hofmann, T.; Baalousha, M.; Ottofuelling, S.; Baun, A. Algal testing of titanium dioxide nanoparticles—Testing considerations, inhibitory effects and modification of cadmium bioavailability. Toxicology **2010**, *269* (2010), 190–197.
18. Stone, V.; Hankin, S.; Aitken, R.; Aschberger, K.; Baun, A.; Christensen, F.; Fernandes, T.; Hansen, S.F.; Hartmann, N.B.; Hutchinson, G.; Johnston, H.; Micheletti, G.; Peters, S.; Ross, B.; Sokull-Kluettgen, B.; Stark, D.; Tran, L. Engineered Nanoparticles: Review of Health and Environmental Safety (ENRHES), 2010. Available at http://nmi.jrc.ec.europa.eu/project/ENRHES.htm (accessed February 2010).
19. Scientific Committee for Emerging and Newly Identified Health Risks (SCENIHR). *Risk Assessment of Products of Nanotechnologies*; Scientific Committee on Emerging and Newly Identified Health Risks, European Commission of Health and Consumer Protection Directorate-General: Brussels, Belgium, 2009.
20. Tiede, K. *Detection and Fate of Engineered Nanoparticles in Aquatic Systems, PhD Thesis, University of York,* Environment Department and Central Science Laboratory: York, U.K., 2008.
21. Hassellöv, M.; Readman, J.W.; Ranville, J.F.; Tiede, K. Nanoparticle analysis and characterization methodologies in environmental risk assessment of engineered nanoparticles. Ecotoxicology **2008**, *17,* 344–361.
22. Lead, J.R.; Wilkinson, K.J. Natural aquatic colloids: Current knowledge and future trends. Environ. Chem. **2006**, *3,* 159–171.
23. European Parliament and the Council of the European Union (EP & CEU).Regulation (EC) No 726/2004 of the European Parliament and of the Council of 31 March 2004 Laying Down Community Procedures for the Authorisation and Supervision of Medicinal Products for Human and Veterinary Use and Establishing a European Medicines Agency, 2004.
24. Council of the European Communities (CEC). Council Directive 90/385/EEC of 20 June 1990 on the approximation of the laws of the Member States relating to active implantable medical devices. Off. J. Eur. Commun. **1990**, *L 189,* 17–36.
25. Council of the European Communities (CEC). Council Directive 93/42/EEC of 14 June 1993 concerning medical devices. Off. J. Eur. Commun. **1993**, *L 169,* 1–43.
26. Council of the European Communities (CEC). Directive 2001/83/EC of the European Parliament and of the Council of 6 November 2001 on the Community code relating to medicinal products for human use. Off. J. Eur. Commun. **2001**, *L 311,* 67–128

27. Editorial. Regulating nanomedicine. Nat. Mater. **2007**, *6*, 249.
28. D'Silva, J.; Van Calster, G. Regulating Nanomedicine: A European Perspective, 2008. Available at http://ssrn.com/abstract=1286215 (accessed November 2008).
29. European Group on Ethics in Science and New Technologies (EGE). Opinion 21—On the Ethical Aspects of Nanomedicine, The European Group on Ethics in Science and New Technologies (EGE), 2007. Retrieved March 15, 2008, available at http://ec.europa.eu/european_group_ethics/publications/docs/final_publication_%20op21_en.pdf.
30. N&ET Working Group. Report on Nanotechnology to the Medical Devices Expert Group: Findings and Recommendations, 2007. Available at http://ec.europa.eu/enterprise/medical_devices/net/entr-2007-net-wg-report-nanofinal.pdf (accessed December 2008).
31. European Medicines Agency (EMEA). *European Medicines Agency: Reflection Paper on Nanotechnology-Based Medicinal Products for Human Use, EMEA/CHMP/70769/2006*; European Medicines Agency: London, 2006a.
32. European Medicines Agency (EMEA). *Guideline on the Environmental Risk Assessment of Medicinal Products for Human Use,* EMEA/CHMP/SWP/4447/00; European Medicines Agency: London, 2006b.
33. Zhang, X.; Sun, H.; Zhang, Z.; Niu, Q.; Chen, Y.; Crittenden, J.C. Enhanced bioaccumulation of cadmium in carp in the presence of titanium dioxide nanoparticles. Chemo- sphere **2007**, *67,* 160–166.
34. Baun, A.; Hartmann, N.B.; Grieger, K.; Kusk, K.O. Ecotoxicity of engineered nanoparticles to aquatic invertebrates—A brief review and recommendations for future toxicity testing. Ecotoxicology **2008a**, *17* (5), 387–395.
35. Baun, A.; Hansen, S.F. Environmental challenges for nanomedicine. Nanomedicine **2008**, *2* (5), 605–608.
36. European Parliament and the Council of the European Union (EP & CEU). Regulation (EC) No 178/2002 of the European Parliament and of the Council of 28 January 2002 laying down the general principles and requirements of food law, establishing the European Food Safety Authority and laying down procedures in matters of food safety. Off. J. Eur. Commun. **2002**, L 31/1-L 31/24.
37. Friends of the Earth (FOE). *Out of the Laboratory and on to Our Plates Nanotechnology in Food and Agriculture*; Friends of the Earth: Australia, Europe, and U.S.A, 2008.
38. Halliday J. EU Parliament Votes for Tougher Additives Regulation, Foodnavigator.com, July 12, 2008. Available at http://www.foodnavigator.com/Legislation/EU-Parlia-mentvotes-for-tougher-additives-regulation (accessed November 2008).
39. European Parliament and the Council of the European Union (EP & CEU). Directive 2008/98/EC of the European Parliament and of the Council of 19 November 2008 on waste and repealing certain directives. Off. J. Eur. **2008**, Union L 312/3-L 312/30.
40. Commission of the European Communities (CEC). *Proposal for a Regulation of the European Parliament and of the Council on Novel Foods and Amending Regulation* (*EC*) *No. XXX/XXXX#* [common procedure] (presented by the Commission) [SEC(2008) 12] [SEC(2008) 13]. COM(2007), 872 final 2008/0002 (COD). 14-1 2008; Commission of the European Communities: Brussels, Belgium, 2008b.
41. European Parliament. Draft Recommendation for Second Reading on the Council Position at First Reading for Adopting a Regulation of the European Parliament and of the Council on Novel Foods, Amending Regulation (EC) No 1331/2008 and Repealing Regulation (EC) No 258/97 and Commission Regulation (EC) No 1852/2001 (11261/2/2009–C7-0000/2010–2008/0002(COD)), 2010.
42. C&EN. Wrangling Over Substance ID Hits REACH Nano Project. Lack of Consensus will Lead to Decision Driven by Policy Not Science, 2011. Available at http://chemicalwatch.com/6324/wrangling-over-substance-id-hits-reach-nano-project?q=Wrangling%20over%20substance%20ID%20hits%20REACH%20nano%20project (accessed January 2011).

43. ECHA. Guidance Documents. Guidance on Information Requirements and Chemical Safety Assessment, 2010. Available at http://guidance.echa.europa.eu/docs/guidance_document/information_ requirements_en.htm?time=1288375681 (accessed October 2010).
44. Poland, C.A.; Duffin, R.; Kinloch, I.; Maynard, A.; Wallace, W.A.H.; Seaton, A.; Stone, V.; Brown, S.; Macnee, W.; Donaldson, K. Carbon nanotubes introduced into the abdominal cavity of mice show asbestos-like pathogenicity in a pilot study. Nat. Nanotechnol. **2008**, *3*, 423–428.
45. Shvedova, A.A.; Kisin, E.R.; Mercer, R.; Murray, A.R.; Johnson, V.J.; Potapovich, A.I.; Tyurina, Y.Y.; Gorelik, O.; Arepalli, S.; Schwegler-Berry, D.; Hubbs, A.F.; Antonini, J.; Evans, D.E.; Ku, B.K.; Ramsey, D.; Maynard, A.; Kagan, V.E.; Castranova, V.; Baron, P. Unusual inflammatory and fibrogenic pulmonary responses to single-walled carbon nanotubes in mice. Am. J. Physiol. Lung Cell. Mol. Physiol. **2005**, *289,* 698–708.
46. Kagan, V.E.; Tyurina, Y.Y.; Tyurin, V.A.; Konduru, N.V.; Potapovich, A.I.; Osipov, A.N.; Kisin, E.R.; SchweglerBerry, D.; Mercer, R.; Castranova, V.; Shvedova, A.A. Direct and indirect effects of single walled carbon nanotubes on RAW 264.7 macrophages: Role of iron. Toxicol. Lett. **2006**, *165* (1): 88–100.
47. Ma-Hock, L.; Treumann, S.; Strauss, V.; Brill, S.; Luizi, F.; Mertler, M.; Wiench, K.; Gamer, A.O.; van Ravenzwaaay, B.; Landsiedel, R. Inhalation toxicity of multi-walled carbon nanotubes in rats exposed for 3 months. Toxicol. Sci. **2009**, *112,* 273–275.
48. Pauluhn, J. Subchronic 13-week inhalation exposure of rats to multiwalled carbon nanotubes: Toxic effects are determined by density of agglomerate structures, not fibrillar structures. Toxicol. Sci. **2010**, *113* (1), 226–242.
49. Oberdorster E. Manufactured nanomaterials (fullerenes, C60) induce oxidative stress in juvenile largemouth bass. Environ. Health Perspect. **2004**, *112,* 1058–1062.
50. Zhu, S.Q.; Oberdorster, E.; Haasch, M.L. Toxicity of an engineered nanoparticle (fullerene, C-60) in two aquatic species, *Daphnia* and fathead minnow. Mar. Environ. Res. **2006**, *62*, S5–S9.
51. Henry, T.B.; Menn, F.M.; Fleming, J.T.; Wilgus, J.; Compton, R.N.; Sayler, G.S. Attributing effects of aqueous C_{60} nano-aggregates to tetrahydrofuran decomposition products in larval zebrafish by assessment of gene expression. Environ. Health Perspect. **2007**, *115* (7), 1059–1065.
52. Templeton, R.C.; Ferguson, P.L.; Washburn, K.M.; Scrivens, W.A.; Chandler, G.T. Life-cycle effects of single-walled carbon nanotubes (SWNTS) on an estuarine meiobenthic copepod. Environ. Sci. Technol. **2006**, *40* (23), 7387–7393.
53. Baun, A.; Sorensen, S.N.; Rasmussen, R.F.; Hartmann, N.B.; Koch, C.B. Toxicity and bioaccumulation of xenobiotic organic compounds in the presence of aqueous suspensions of aggregates of nano-C60. Aquat. Toxicol. **2008b**, *86,* 379–387.
54. Oberdorster, G.; Maynard, A.; Donaldson, K.; Castranova, V.; Fitzpatrick, J.; Ausman, K.; Carter, J.; Karn, B.; Kreyling, W.; Lai, D.; Olin, S.; Monteiro-Riviere, N.; Warheit, D.; Yang, H. Principles for characterizing the potential human health effects from exposure to nanomaterials: Elements of a screening strategy. Part. Fibre Toxicol. **2005**, *2*, 8.
55. Oberdorster, G. Significance of particle parameters in the evaluation of exposure dose–response relationships of inhaled particles. Part. Sci. Technol. **1996**, *14,* 135–151.
56. Oberdorster, G.; Stone, V.; Donaldson, K. Toxicology of nanoparticles: A historical perspective. Nanotoxicology **2007**, *1* (1), 2–25.
57. Stoeger, T.; Reinhard, C.; Takenaka, S.; Schroeppel, A.; Karg, E.; Ritter, B.; Heyder, J.; Schultz, H. Instillation of six different ultrafine carbon particles indicates surface area threshold dose for acute lung inflammation in mice. Environ. Health Perspect. **2006**, *114*(3), 328–333.
58. Stoeger, T.; Schmid, O.; Takenaka, S.; Schulz, H. Inflammatory response to TiO_2 and carbonaceous particles scales best with BET surface area. Environ. Health Perspect. **2007**, *115* (6), A290–A291.
59. Wittmaack, K. In search of the most relevant parameter for quantifying lung inflammatory response to nanoparticle exposure: Particle number, surface area, or what? Environ. Health Perspect. **2007a,** *115,* 187–194.

60. Wittmaack, K. Dose and response metrics in nanotoxicology: Wittmaack responds to Oberdoerster et al. and Stoeger et al. Environ. Health Perspect. **2007b**, *115* (6), A290–291.
61. Warheit, D.B.; Webb, T.R.; Colvin, V.L.; Reed, K.L.; Sayes, C.R. Pulmonary bioassay studies with nanoscale and fine-quartz particles in rats: Toxicity is not dependent upon particle size but on surface characteristics. Toxicol. Sci. **2007a**, *95* (1), 270–280.
62. Warheit, D.B.; Webb, T.R.; Reed, K.L.; Frerichs, S.; Sayes, C.M. Pulmonary toxicity study in rats with three forms of ultrafine-TiO_2 particles: Differential responses related to surface properties. Toxicology **2007b**, *230,* 90–104.
63. CCA. *Small Is Different: A Science Perspective on the Regulatory Challenges of the Nanoscale;* The Council of Canadian Academies: Ottawa, Canada, 2008.
64. Boxall, A.B.A.; Chaudhry, Q.; Sinclair, C.; Jones, A.; Aitken, R.; Jefferson, B.; Watts, C. *Current and Future Predicted Environmental Exposure to Engineered Nanoparticles;* Central Science Laboratory: York, U.K., 2008.
65. Hansen, S.F.; Michelson, E.; Kamper, A.; Borling, P.; Stuer-Lauridsen, F.; Baun, A. Categorization framework to aid exposure assessment of nanomaterials in consumer products. Ecotoxicology **2008**, *17* (5), 438–447.
66. Luoma, S.N. *Silver Nanotechnologies and the Environment Old Problems or New Challenges?* PEN 15. Project on Emerging Nanotechnologies; Woodrow Wilson International Center for Scholars: Washington, D.C., 2008.
67. Park, B.; Donaldson, K.; Duffin, R.; Tran, L.; Kelly, F.; Mudway, I.; Morin, J-P.; Guest, R.; Jenkinson, P.; Samaras, Z.; Giannouli, M.; Kouridis, H.; Martin, P. Hazard and risk assessment of a nanoparticulate cerium oxide-based diesel fuel additive—A case study. Inhal. Toxicol. **2008**, *20* (6), 547–566.
68. Shinohara, N. *Risk Assessment of Manufactured Nanomaterials—Fullerene* (*C60*)*,* Interim Report issued on October 16, 2009, Executive Summary. The Research Institute of Science for Safety and Sustainability, AIST available at http://www.aistriss.jp/main/modules/product/nano_rad.html.
69. Mueller, N.; Nowack, B. Exposure modeling of engineered nanoparticles in the environment. Environ. Sci. Technol. **2008**, *42*, 4447–4453.
70. European Food Safety Authority (EFSA). Draft Opinion of the Scientific Committee on the Potential Risks Arising from Nanoscience and Nanotechnologies on Food and Feed Safety (Question No EFSAQ-2007–124). European Food Safety Authority: Parma, Italy, 2008.
71. European Food Safety Authority (EFSA). Endorsed for Public Consultation Draft Scientific Opinion Guidance on Risk Assessment Concerning Potential Risks Arising From Applications of Nanoscience and Nanotechnologies to Food and Feed. EFSA Scientific Committee, European Food Safety Authority (EFSA): Parma, Italy, 2011.
72. Höck, J.; Epprecht, T.; Hofmann, H.; Höhner, K.; Krug, H.; Lorenz, C.; Limbach, L.; Gehr, P.; Nowack, B.; Rie- diker, M.; Schirmer, K.; Schmid, B.; Som, C.; Stark, W.; Studer, C.; Ulrich, A.; von Götz, N.; Wengert, S.; Wick, P.; Guidelines on the Precautionary Matrix for Synthetic Nanomaterials. Federal Office of Public Health and Federal Office for the Environment, Berne 2010, Version 2. Available at http://www.bag.admin.ch/themen/chemika-lien/00228/00510/05626/index.html?lang=en (accessed 17 May 2010).
73. Linkov, I.; Satterstrom, F.K.; Steevens, J.; Ferguson, E.; Pleus, R.C. Multi-criteria decision analysis and environmental risk assessment for nanomaterials. J. Nanopart. Res. **2007**, *9* (4), 543–554.
74. Grieger, K.D.; Linkov, I.; Hansen, S.F.; Baun, A. A review of alternative frameworks and approaches for assessing environmental risks of nanomaterials. Nanotoxicology, **2011**, *in press.*
75. Harrington, R. Nano Risk Assessment: A Work in Progress, 2010. Available at http://www.foodproductiondaily.com/Quality-Safety/Nano-risk-assessment-a-work-in-progress/?c=lIxHi8W7vkWyj1EO6E28rw%3D%3D&utm_source=newsletter_daily&utm_medium=email&utm_campaign=Newsletter%2BDaily (accessed June 2010).

76. Höck, J.; Epprecht, T.; Hofmann, H.; Höhner, K.; Krug, H.; Lorenz, C.; Limbach, L.; Gehr, P.; Nowack, B.; Riediker M. et al. *Guidelines on the Precautionary Matrix for Synthetic Nanomaterials,* Version 2; Swiss Federal Office for Public Health and Federal Office for the Environment: Berne, Switzerland, 2010.
77. Lam, C.W.; James, J.T.; McCluskey, R.; Hunter R.L. Pulmonary toxicity of single-wall carbon nanotubes in mice 7 and 90 days after intratracheal instillation. Toxicol. Sci. **2004**, *77*, 126–134.
78. Paik, S.Y.; Zalk, D.M.; Swuste, P. Application of a pilot control banding tool for risk level assessment and control of nanoparticle exposures. Ann. Occup. Hyg. **2008**, *52* (6), 419–428.
79. Royal Society and the Royal Academy of Engineering. *Nanoscience and Nanotechnologies: Opportunities and Uncertainties;* Royal Society: London, 2004.

VI

ENT: Environmental Management Using Environmental Technologies

40

Industrial Waste: Soil Pollution and Remediation

W. Friesl-Hanl,
M.H. Gerzabek,
W.W. Wenzel, and
W.E.H. Blum

Introduction

Anthropogenic activities such as mining and smelting, combustion of waste and fossil fuels, as well as the use of organic and inorganic chemicals and radionuclides in agriculture and industry entail risks for the overall environment and human beings. The awareness of these risks due to industrial wastes emerged through numerous cases of severe environmental impacts many years after their disposal. Notorious examples include hazardous waste disposal in the Love Canal, New York, United States[1]; Lekkerkerk near Rotterdam, Netherlands[2]; the collapsed tailing dams containing acidic pyrite sludge in Donana, Spain[3]; cyanide- and metal-rich liquid waste disposals near Baia Mare, Romania[4]; and many other accidents.[5] A report describing the world's worst-polluted industrial sites was updated in 2016 with the focus on the top polluting industries identified as (1) used lead–acid battery recycling, (2) mining and ore processing, (3) lead smelting, (4) tanneries, (5) artisanal small-scale gold mining, (6) industrial dumpsites, (7) industrial estates, (8) chemical manufacturing, (9) product manufacturing, and (10) dye industry. These industries put over 32 million people at risk and account for 7 million to 17 million Disability-Adjusted Life Years (DALYs) in low- and middle-income countries.[6] Heavily polluted sites are distributed across all continents, and a former report highlighted only the tip of an iceberg, listing sites such as Dzerzinsk (chemical weapons, Russia), Linfen (coal industry, China), Ranipet (tanning industry, India), Kabwe (mining and smelting of Pb, Zambia), Haina (recycling of Pb, Dominican Republic), and La Oroya (polymetallic smelter, Peru).[7] While full remediation of these sites is economically and technically impossible, low-cost in situ technologies for managing and monitoring aiming at risk reduction are becoming increasingly available. More importantly, we need to learn our lessons and avoid similar pollution in the future.

Besides these negative effects of industrial waste, their reuse potential for soil amelioration, remediation, and recultivation also should be considered. Future perspectives and visions for handling industrial waste must be discussed in terms of the potential environmental risks and how to minimize them. The proclaimed Sustainable Development Goals (SDGs) by the United Nations (UN) should be a guidance for these next steps.[8]

Industrial Waste Overview

Generally, industrial activities generate three categories of waste: (1) gaseous emissions, (2) wastewater, and (3) solid waste. Gaseous pollutants or particulates are generated by combusting fossil fuels and wastes; they are found both close to the emitter and far away, especially in forest ecosystems, due to their filter capacity.[9] Dispersion and long-term transport of heavy metals, radionuclides, and persistent organic pollutants occur.[10] Acidification and eutrophication are due to the transboundary distribution of nutrients often involved in combustion processes (e.g., S, N compounds). Wastewater derives from domestic and industrial establishments; combined, they are known as municipal wastewater. Surface water and groundwater are the most threatened sinks of pollutants, besides irrigated areas. Upon removal of water, wastewater turns to sludge and later to solid waste. Generally, solid waste generation starts with mining and petroleum production and with the subsequent processing of these raw materials to goods for consumption. Emerging waste includes mineral ores, tailings, slags, and oil-contaminated soils and residues that stem from the processing of these materials.

Industrial activities generate mostly mixed wastes, yielding transient waste types. Pollutants can be released in gaseous, particulate, aqueous, or solid form, depending on the industrial processes involved. They emanate from point[11] or diffuse[12] sources. Generally, pollution tends to be highest close to emitters and declines with increasing distance. However, elevated concentrations of persistent organic chemicals have been found in environments far from pollution sources, such as in the Arctic, in distant mountain regions, and in remote forest soils.[13]

Soil quality is prone to degradation through physical, chemical, and biological processes, resulting in soil erosion, reduced productivity, and soil and groundwater contamination. Detrimental effects of industrial waste on soil comprise the following: (1) siltation and compaction through application of fine-grained and salt-rich materials; (2) any type of pollution, acidification, or landfilling; and (3) loss of macro- and microfaunal diversity by intoxication. Heavy metals (Pb, Hg, Cd, Cr), metalloids (As, Sb), and organic compounds (e.g., dioxins and other persistent organic pollutants) are the most relevant soil pollutants. Since 1950, more than 140,000 new chemicals and pesticides have been synthesized whereof 5,000 are produced in large quantities, dispersed in the environment, and responsible for human exposure.[14] Waste generation has dramatically increased from chemical manufacturing, wastewater treatment, agriculture, mining and smelting, food processing, energy production, pulp and paper production, and other industries. The subsequent waste–environmental interactions have potential consequences on soil and water qualities. Estimates show that up to 75% of the materials used in industrial processes do not end up in final products.[15]

Prevention of industrial waste production, enhanced by optimization of processes reducing the amount of generated waste and the concentration of pollutants in waste streams, has a higher priority than end-of-pipe technologies such as soil remediation. However, emerging technologies to filter and retain pollutants in solid industrial waste materials require close monitoring as this may limit the re-use of such materials, for example, in soil remediation, due to increased pollutant concentrations. For instance, novel clean coal technologies lead to changed coal combustion residues (CCR) composition as volatile elements (e.g., Hg and Se) are more efficiently retained while due to concurrent changes in pH the solubility (Hg, Cd, Cu) and volatilization (Hg, NH_3) of certain pollutants are likely enhanced.[16]

Potential Use of Industrial Waste

Soil Amelioration

In the case of metals, humankind has imbalanced biogeochemical cycles through the use of metals. Considerable amounts are lost in technical processes into the above-mentioned waste categories. Electronic waste is one of the most rapidly growing problems.[17] E-waste volume estimation for 2013 is given in Table 1. Instead of exporting this waste[18] from "developed" into "developing" countries and thereby contaminating soil and water due to the lack of legal regulations there, the strategy should be to promote recycling and reuse of metals. Besides heavy metal-contaminated arable land, nutrient-deficient areas are widespread over the world. Here, the reuse of Cu, Mn, Mo, Fe, and Zn from electronic waste as nutrients could contribute toward more sustainable systems.

Gravel sludge, a fine-grained by-product of the gravel industry—applied on sandy soils poor in nutrients—can improve soil physical properties and crop production.

On the other hand, the use of industrial waste materials as soil amendments has been estimated to contaminate thousands of hectares of productive agricultural land in countries throughout the world. This is a particular problem in the Asia Pacific region, where land-based disposal of untreated wastes is still being practiced due to the lack of regulatory guidelines.[19] If crops are cultivated on contaminated soils, they may become less competitive on the world market, provided efficient quality control is implemented. Figure 1 gives a schematic overview of waste reuse possibilities.

Soil Remediation

In situ fixation of labile metals in soils by addition of amendments is one option to remediate heavy metal-contaminated soils. The aim is to reduce the risk of contamination by balancing ecological and economical needs. Several industrial wastes or by-products have been screened for their ability to immobilize heavy metals. Red mud is produced by the alumina industry around the world in countries like Jamaica, Australia, the United States, and Germany[20] (Table 1). Results of greenhouse, outdoor, and field experiments indicate immobilization of many heavy metals in soil after red

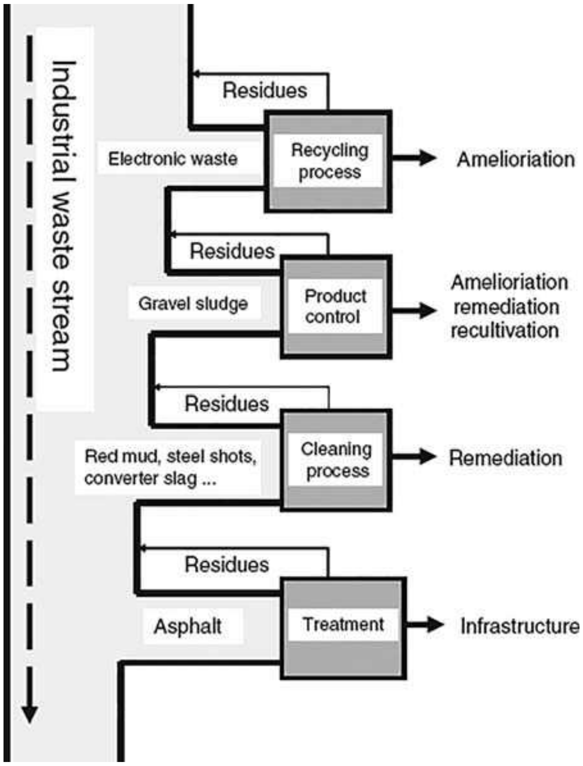


FIGURE 1 A schematic overview of waste reuse possibilities.

TABLE 1 Industrial Waste Quantities

Industry	Industrial Waste	Produced Amount (Mg/a)
Electronic industry	Electronic waste	20–50 Mio, worldwide[16]
Alumina industry	Red mud	120 Mio, worldwide[20]
Pulp and paper industry	Paper mill sludge	16 Mio, US[24], CA[25]
Iron	Slags (25%–30% of crude iron)	300–360 Mio,[a] worldwide[29]
Steel industry	Slags (10%–15% of crude steel)	190–290 Mio,[b] worldwide[29]
Energy production	CCP	40 Mio, EU15, (2016)[30]
	CCP	110 Mio, US (2017)[31]

[a] Calculated from produced amount of iron.

[b] Calculated from produced amount of steel. CCP, coal combustion products (including fly ash, flue-gas desulfurization gypsum, bottom ash); US, United States; CA, Canada.

mud addition.[21–23] Similar results were obtained for steel shots, a waste product of the metal shaping industry; converter and basic slag from the steel and iron industry; and "Iron Rich," generated by TiO_2 production.[26] Fly ash, a coal combustion residue, and beringite, a modified aluminosilicate that originates from the fluidized bed burning of coal refuse from a former coal mine in Beringen (Belgium), also have potential for metal immobilization.[27] A further example could be the reuse of the tailings of the world's largest siderite mining site (Erzberg, Austria).[28] The tailings consist of high amounts of fine-textured iron and carbonates, and it has been demonstrated in pot and field experiments that they hold promise for in situ inactivation of heavy metals in polluted soils and associated improvement of crop yield and quality if applied to arable land.[32] The composition of these amendments varies considerably, and the immobilizing effect is mostly due to Al-, Fe-, Mn-oxides, phosphates, silicates, and alkaline materials.

An emerging amendment for soil remediation/amelioration could be biochar. These materials derive from different sources. If organic residues/waste by-products will be pyrolyzed and applied on soils, different functions could be improved. Besides the storage of carbon in soil,[33] also soil physical (e.g., water holding capacity[34]) or chemical (e.g., immobilization of contaminants) properties can be improved.[35]

Application rates of these mentioned amendments up to 5% (w/w) are practicable and result in establishing a vegetation cover on bare land, reduction of visual symptoms of metal phytotoxicity, and a significant reduction of the metal accumulation in aboveground biomass. An example of a field application is given in Friesl-Hanl et al. (2017).[23]

Advantages of these easily available materials, including contaminant inactivation and often low costs,[36] are opposed by some disadvantages or failures that may occur due to indigenous contamination.[37] Monitoring of the sustainability of the applied measures is necessary and should be mandatory.[24] The recommendation is to adjust the amendment and application rate on a case-by-case basis. In certain cases, removal of contaminants may be required.

Conclusions

Pollution by industrial waste is widespread throughout the world, adversely affecting soil and environment qualities. Management of these wastes is necessary to achieve a more sustainable system. After exhausting the two approaches of preventing waste generation and recycling suitable materials, the use of wastes, by-products, or mixtures of different materials for amelioration, remediation, or recultivation is one option to regulate material fluxes. Besides considering the advantages of using waste materials, we also have to take into account the disadvantages and assess them on a case-by-case basis.

References

1. Manhan, S.E. *Environmental Chemistry*, 6th Ed.; Lewis Publisher: New York, 1994.
2. Alloway, B.J.; Ayres, D.C. *Chemical Principles of Environmental Pollution*; Chapman and Hall: London, 1993.
3. Meharg, A.A.; Osborn, D.; Pain, D.J.; Sanchez, A.; Naveso, M.A. Contamination of Doñana food chain after the Aznalcollar mine disaster. *Environ. Pollut.* 1999, *105* (3), 387–390.
4. Premysl, S.; Pavonic, M.; Boucek, J.; Kokes, J. Baia Mare accident – Brief ecotoxicological report of Czech experts. *Ecotoxicol. Environ. Safety* 2001, *49*, 255–261.
5. Johnson, J. Industrial safety – The uneven world of chemical accident investigation. *Chem. Eng. News* 2016, 18–20.
6. Pure Earth Blacksmith Institute and Green Cross Switzerland, World's Worst Pollution Problems. The Toxics beneath Our Feet, 2016, 54, available at http://www.pureearth.org (accessed March 2019).
7. The Blacksmith Institute. The World's Worst Polluted Places—The Top Ten. P 57, available at http://www.blacksmithinstitute.org (accessed March 2019).
8. UN SDGs (Sustainable development Goals); https://www.un.org/sustainabledevelopment/news/communications-material/ (accessed on 27 March 2019). Transforming our world. The 2030 Agenda for sustainable development. A/RES/70/1; p 37.
9. Schulte, A.; Blum, W.E.H. Schwermetalle im Waldökosystem. In Geochemie und Umwelt; Matschullat, J., Tobschall, H.J.; Voigt, H.J., Eds.; Springer: Berlin, Heidelberg, 1997; 53–74.
10. Blum, W.E.H. Soil degradation caused by industrialization and urbanization. In *Towards Sustainable Land Use*; Blume, H.P.; Eger, H.; Fleischhauer, E.; Hebel, A.; Reij, C.; Steiner, K.G., Eds.; Catena: Reiskirchen, 1998; Vol. I, Advances in Geoecology 31, 755–766.
11. Naidu, R.; Megharaj, M.; Dillon, P.; Kookana, R.; Correll, R.; Wenzel, W.W. Soils and point source pollution. In *Encyclopedia of Soil Science*; Lal, R., Ed.; Marcel Dekker, Inc.: New York, 2002; 1012–1017.
12. Naidu, R.; Megharaj, M.; Dillon, P.; Kookana, R.; Correll, R.; Wenzel, W.W. Soils and non-point source pollution. In *Encyclopedia of Soil Science*; Lal, R., Ed.; Marcel Dekker, Inc.: New York, 2002; 872–874.
13. Wania, F. On the origin of elevated levels of persistent chemicals in the environment. *Environ. Sci. Pollut. Res.* 1999, *6* (1), 11–19.
14. Landrigan, P.J.; Fuller, R.; Acosta, N.J.R.; Adeyi, O.; Arnold, R.; Baldé, A.B.; Bertollini, R.; Bose-O'Reilly, S.; Boufford, J.I.; Breysse, P.N.; Chiles, T. The Lancet Commission on pollution and health. *The Lancet* 2018, *391*, 462–512.
15. Onida, M. Innovation in prevention of quantitative and qualitative waste production. In *Innovation in Waste Management*, Proceedings of the IV European Waste Forum, Milano, Italy, Nov 30–Dec 1, 2000; Centro de Ingegneria per la Protezione dell' Ambiente (C.I.P.A.), Eds.; European Waste Forum: Milano, 2000; 91–105.
16. Dellantonio, A.; Reepmann, F.; Fitz, W.J.; Wenzel, W.W. Disposal of coal combustion residues in terrestrial systems: Contamination and risk management. *J. Environ. Qual.* 2010, *39*, 761–775. doi:10.2134/jeq2009.0068
17. Wang, F.; Huisman, J.; Stevels, A.; Baldé, C.P. Enhancing e-waste estimates: Improving data quality by multivariate input-output analysis. *Waste Manag.* 2013, *33*, 2397–2407.
18. Puckett, J.; Byster, L.; Weservelt, S.; Gutierrez, R.; Davis, S.; Hussain, A.; Dutta, M. *Exporting Harm: The High-Tech Trashing of Asia*; The Basel Action Network: Washington, DC, 2002.
19. Ullah, S.M.; Nuruzzaman, M.; Gerzabek, M.H. Heavy metal and microbiological pollution of water and sediments by industrial wastes, effluents and slums around Dhaka City. In *Tropical Limnology III*; Timotius, K.H.; Goltenboth, F., Eds.; Satya Waeann University Press: Salatiga, Indonesia, 1995; 179–186.

20. Hua, Y.; Heal, K.V.; Friesl-Hanl, W. The use of red mud as an immobiliser for metal/metalloid-contaminated soil: A review. *J. Hazard. Mater.* 2017, *325*, 17–30; doi:10.1016/j.jhazmat.2016.11.073
21. Friesl, W.; Lombi, E.; Horak, O.; Wenzel, W.W. Immobilization of heavy metals in soils using inorganic amendments in a greenhouse study. *J. Plant Nutr. Soil Sci.* 2003, *166*, 191–196.
22. Friesl, W.; Horak, O.; Wenzel, W. Immobilization of heavy metals in soils by the application of bauxite residues: Pot experiments under field conditions. *J. Plant Nutr. Soil Sci.* 2004, *167* (1), 54–59.
23. Friesl-Hanl, W.; Platzer, K.; Riesing, J.; Horak, O.; Waldner, G.; Watzinger, A.; Gerzabek, M.H. Non-destructive soil amendment application techniques on heavy metal-contaminated grassland: Success and long-term immobilizing efficiency. *J. Environ. Manag.* 2017, *186*, 167–174; doi:10.1016/j.jenvman.2016.08.068.
24. Geng, X.; Zhang, S.Y.; Deng, J. Characteristics of paper mill sludge and its utilization for the manufacture of medium density fiberboard. *Wood Fiber Sci. 39* (2), 2007, 345–351.
25. Bonomo, L.; Higginson, A.E. *International Overview on Solid Waste Management*; Academic Press: London, 1988.
26. Knox, A.S.; Seaman, J.C.; Mench, M.J.; Vangronsveld, J. Remediation of metal- and radionuclides-contaminated soils by in situ stabilization techniques. In *Environmental Restoration of Metal-Contaminated Soils*; Iskandar, I.K., Ed.; Lewis Publishers: Boca Raton, 2000; 21–60.
27. Vangronsveld, J.; Van Assche, F.; Clijsters, H. Reclamation of a bare industrial area contaminated by non-ferrous metals: In situ metal immobilization and revegetation. *Environ. Pollut.* 1995, *87*, 51–59.
28. Siderit, available at http://de.wikipedia.org/wiki/Siderit (accessed March 2019).
29. U.S. Geological Survey, Mineral commodity summaries 2019: U.S. Geological Survey, 2019, 200, https://doi.org/10.3133/70202434.
30. ECOBA European Coal Combustion Products Association, available at http://www.coal-ash.co.il/docs/ECOBA_Stat_2016_EU15.pdf (accessed March 2019).
31. ACAA American Coal Ash Association, available at https://www.acaa-usa.org/publications/productionusereports.aspx (accessed March 2019).
32. Touceda-Gonzalez, M.; Brader, G.; Antonielli, L.; Ravindran, V.B.; Waldner, G.; Friesl-Hanl, W.; Corretto, E.; Campisano, A.; Pancher, M.; Sessitsch, A. Combined amendment of immobilizers and the plant growth-promoting strain Burkholderia phytofirmans PsJN favours plant growth and reduces heavy metal uptake. *Soil Biol. Biochem.* 2015, *91*, 140–150.
33. Schmidt, H.-P.; Anca-Couce, A.; Hagemann, N.; Werner, C.; Gerten, D.; Lucht, W.; Kammann, C. Pyrogenic carbon capture and storage; *GCB Bioenergy.* 2018, 1–19. doi:10.1111/gcbb.12553
34. Ojeda, G.; Mattana, S.; Àvila, A.; Alcañiz, J.M.; Volkmann, M.; Bachmann, J. Are soil–water functions affected by biochar application? *Geoderma* 2015, *249–250*, 1–11, doi:10.1016/j.geoderma.2015.02.014
35. Uchimiya, M.; Wartelle, L.H.; Klasson, K.T.; Fortier, C.A.; Lima, I.M. Influence of pyrolysis temperature on biochar property and function as a heavy metal sorbent in soil. *J. Agric. Food Chem.* 2011, *59*, 2501–2510. doi:10.1021/jf104206c
36. Puschenreiter, M.; Horak, O.; Friesl, W.; Hartl, W. Low-cost agricultural measures to reduce the heavy metal transfer into human food chain – A review. *Plant Soil Environ.* 2005, *51* (1), 1–11.
37. Friesl-Hanl, W.; Platzer, K.; Horak, O.; Gerzabek, M.H. Immobilising of Cd, Pb, and Zn contaminated arable soils close to a former Pb/Zn smelter: A field study in Austria over 5 years. *Environ. Geochem. Health* 2009, *31*, 581–594.

41

Pest Management: Crop Diversity

Maria R. Finckh and
Jan Henrik Schmidt

Introduction

Improvement in crop management using modern machinery equipment, fertilizers, and pesticides and progress in plant breeding within a few selected crops has led in recent decades to highly specialized agricultural practices on farms. Out of more than 300,000 plant species existing, about 7,000 crops are known to be cultivated for human or animal diet; however, 60% of the word's energy source is based on only three crops, namely wheat, maize, and rice [1]. As a consequence, a general impoverishment of plant diversity and a high degree of genetic erosion in several important crops for human nutrition is documented since the end of World War II [2].

Simplification of the ecosystem by using one-sided crop rotations, monoculture, and crop plants of uniform genotype and the elimination of weeds with herbicides results in a strong selection for adapted pests, pathogens, and weeds leading to frequent resistance breakdown and severe weed infestations [3–5] and, importantly, a general loss of biodiversity [6].

Increasing the diversity of selection pressures acting on pests and weeds can reduce their ability to develop resistances to pesticides. Diverse crop ecosystems with adapted crops in sequential cropping including cover crops and living mulches, crop populations, and species mixtures are generally less damaged by pests, pathogens, and weeds compared to monocultures [2]. Interestingly, weeds are less abundant but weed species diversity is often higher within such diverse crop rotations. Hence, outbreaks of weed, pest, and disease epidemics and the probability of losses are reduced [4,7,8].

In this chapter, diversity by planting sequence (crop rotation), crop-border diversity, and crop-weed diversity are discussed.

Influence of Crop Rotation on Diseases, Insect Pests, and Weeds

The positive effects of a diverse crop rotation on yield and agricultural stability have been well known since ancient times. Awareness has built that much of the rotation effects lie in the plants themselves as they release considerable amounts of organic compounds into the soil in order to influence the microbiome, the solubility of nutrients and to interact with other plants. Most pathogens and pests are weak competitors outside their hosts and barely survive. If soil life is abundant, then it will outcompete and often directly live on the resting stages of pests and pathogens and weeds [9].

Crop rotation is most effective in the case of specialized pathogens and pests that are dependent on a host crop or have a narrow host range. For example, many soil-borne pathogens or insects survive on root or plant residues and require the cultivation of a susceptible crop for continuous survival. The length of host crop interval for insect or disease control will depend on how quickly the insect pest or pathogen can be destroyed by starvation and/or by other antagonistic effects [10,11].

With respect to weeds, important criteria for their control are variation in planting time and timing as well as techniques of seed bed preparation [12,13]. To prevent weeds from germinating simultaneously with the cash crop or at least from seed shed, the whole cropping system needs to be considered [14]. Main effects are caused by the sowing time of crops (winter or spring crops), sequence and placement of crops in the crop rotation, competitive ability of different crops against weeds, and direct weed control (herbicides and/or mechanical control) [15]. Allelopathic effects, that is, the exudation of specific compounds from the roots of certain crops may reduce germination of different weeds [15,16]. For perennial weeds, methods that deplete the weed storage organs through competition, for example, due to living mulches and perennial forage crops, and/or frequent mowing can be useful to reduce number and biomass of weeds [17–19]. These methods combined with fragmentation of weed rhizomes that can reduce the viability of storage organs have proven a high efficacy in Quackgrass (*Elymus repens*) control in Sweden [20]. In monoculture and pure grain rotations, weed infestation can reach a high level and weeds are more difficult to control compared to a rotation in alternation with dicotyledonous plants, for example, field beans, potatoes, or canola [21,22]. Especially the degree of infestation of problem weeds, for example, *Apera spica venti*, *Viola tricolor arvensis*, and *Matricaria* spp., increases considerably [23,24].

When developing a crop rotation, certain criteria should thus be considered: adequate change in tillage practices and timing, maintenance of high levels of organic matter in the soil, inclusion of crops which do not stimulate subsequent growth of the pathogen or crops that have direct negative effects on pests and pathogens (e.g., producing toxins) [11]. For example, oats and brassicas, such as various mustards, are known to exude substances that are directly suppressive to many pathogens, pests, and weeds and therefore justifying their recent popularity in rotations [25–27]. However, they are also hosts to some broad-range pathogens such as *Fusarium avenaceum*. Such broad host range pathogens need to be managed through the enhancement of soil suppressiveness [28,29]. To achieve this, soil protection from damaging compaction through heavy machinery, radiation, drought, and heavy rainfall by high aboveground living and dead biomass, such as occurring in minimum tillage systems, is mandatory [28,30]. In addition, this will foster the microbial activity in the topsoil layers and result, in the long term, in sustainable production systems.

Influence of Decoy and Trap Crops on Pests

Decoy crops are non-host crops that are sown to stimulate the activation of dormant propagules of the pathogen in the absence of the host. In this way, the soil-borne pathogens waste their inoculum potential. For example, *Lolium* spec., *Papaver rhoeas*, and *Reseda odorata* can act as decoy crops for the pathogen *Plasmodiophora brassicae* in Brassica [31]. In the case of trap crops, the crop is host to the pathogen often nematodes. The trap crop attracts nematodes to infect, but the crop is harvested or destroyed before the nematode can complete its life cycle. A famous, still effective method had been developed by Julius Kühn in the 1840s in Germany when certain crucifers are sown and plowed before the beet cyst nematode (*Heterodera schachtii*) can fully develop its life cycle [26]. Recent breeding efforts resulted in *H. schachtii*-resistant oilseed radish varieties that are used as winter cover crops prior to growing sugar beets. Root exudates of resistant varieties induce hatching of young nematode larvae from eggs and cysts with subsequent penetration of the radish roots. However, due to the resistance mechanism, *H. schachtii* juveniles are not capable of building large feeding sites within the radish roots, which prevents the forming of new cysts [32].

Influence of Weed and Border Diversity on Pests and Diseases

Weeds not only compete with crops; they may also be intermediate hosts for diseases and parasites while at the same time offering food and refuge to beneficial insects within the agricultural ecosystem. Within monocropped fields, weeds are important sources of biodiversity and may be useful to improve the stability of the agricultural ecosystem. They play an important ecological role by supporting a complex of beneficial arthropods that aid in suppressing pest populations and thus the pest damage [33–36]. Strip management with weeds or flowering crops is by now standard in many sustainable perennial cropping systems such as orchards (e.g., [37–39]). Weed strips or living mulches, such as grasses or clover species, sown between winter cereals increase ground beetle densities and the number of species considerably by providing these beneficial arthropods with better food supplies and more suitable overwintering sites, from which they can colonize cereals in spring [40].

The abundance and diversity of entomophagous insects within a field are closely related to the character of the surrounding vegetation but also to the field size [6,41]. There are many examples that indicate that crops cultivated near hedgerows, grassy margins, or uncultivated fields with flowering weeds sustain less damage by insect pests than crops cultivated in the absence of such flowering vegetation [33–36]. Nevertheless, it is important to consider the complete life cycle and feeding habit of pathogens and insect pests. Certain weeds and structure elements will serve as alternate hosts to crop pathogens. Carrot flies (*Psila rosae*) need hedges nearby to protect themselves from heat and many aphids overwinter in woody plants. There, adequate distances to fields or the inclusion of specific border strips as trap crops need to be considered [7]. In most cases, field margins, consisting of hedges, sown grasses, and flower strips, for example, provide more beneficial effects through natural enemies regulating pest populations than detrimental effects through pest habitat provision [42].

Future Concerns

Diversity provides an essential key to reduce the risk of losing crop yield to pest damage. Of immediate concern are the effects of climate vagaries and change that will lead to the permanent change and invasions of new pests to new areas [43]. Many pesticides may disappear from the market in the near future due to their detrimental effects. Thus, massive insecticide use in combination with increasing field sizes and simplified cropping sequences were made responsible for the drastic insect decline in the past decades [44], while glyphosate and other herbicides are antibiotics and suspected of being responsible for the development of multi-resistance with direct implications on human health [45–48] including in bee decline [46]. The realization of greater crop diversity by crop rotation and trap crops and surrounding vegetation to stabilize overall crop yields will be of even greater importance in the face of climate change and the need to reduce pesticide use. The most effective way to increase general biodiversity is to simply reduce overall field sizes, though [6]. The implementation of concepts based on crop diversity will preserve long-term stability and productivity of agricultural land and minimize environmental problems caused by intensive agriculture, for example, biodiversity loss, soil erosion, groundwater and air pollution with nutrients and pesticides.

See also *Intercropping for Pest Management.*

References

1. Li, X. and K.H.M. Siddique, *Future Smart Food: Rediscovering Hidden Treasures of Neglected and Underutilized Species for Zero Hunger in Asia*. 2018, Bangkok: FAO. 242.
2. Storkey, J., et al., The future of sustainable crop protection relies on increased diversity of cropping systems and landscapes, in *Agroecosystem Diversity*, G. Lemaire, et al., Editors. 2019, London: Academic Press. pp. 199–209.

3. Heap, I. *The International Survey of Herbicide resistant Weeds*. https://www.weedscience.com, 2019.
4. Lamichhane, J.R., et al., Advocating a need for suitable breeding approaches to boost integrated pest management: a European perspective. *Pest Management Science*, 2018. **74**(6): pp. 1219–1227.
5. Ostergaard, H., et al., Time for a shift in crop production: embracing complexity through diversity at all levels. *Journal of the Science of Food and Agriculture*, 2009. **89**: pp. 1439–1445.
6. Sirami, C., et al., Increasing crop heterogeneity enhances multitrophic diversity across agricultural regions. *Proceedings of the National Academy of Sciences*, 2019. **116**(33): pp. 16442–16447.
7. Finckh, M.R. and M.S. Wolfe, Biodiversity enhancement, in *Plant Diseases and Their Management in Organic Agriculture*, M.R. Finckh, A.H.C.V. Bruggen, and L. Tamm, Editors. 2015, St. Paul, MN: APS. pp. 153–174.
8. Bedoussac, L., et al., Grain legume–cereal intercropping systems, in *Achieving Sustainable Cultivation of Grain Legumes*, S. Sivasankar, et al., Editors. 2018, Cambridge: Burleigh Dodds Science Publishing. pp. 243–255.
9. Cook, R.J. and K.F. Baker, *The Nature and Practice of Biological Control of Plant Pathogens*. 1983, St. Paul, MN: APS Press. pp. 1–539.
10. Finckh, M.R., et al., Organic temperate legume disease management, in *Plant Diseases and Their Management in Organic Agriculture*, M.R. Finckh, A.H.C. van Bruggen, and L. Tamm, Editors. 2015, St. Paul, MN: APS Press. pp. 293–310.
11. Leoni, C., W. Rossing, and A.H.C. van Bruggen, Crop rotation, in *Plant Diseases and Their Management in Organic Agriculture*, M.R. Finckh, A.H.C. van Bruggen, and L. Tamm, Editors. 2015, St. Paul, MN: APS Press. pp. 127–140.
12. Bond, W. and A.C. Grundy, Non-chemical weed management in organic farming systems. *Weed Research*, 2001. **41**(5): pp. 383–405.
13. Brandsæter, L.O., et al., Control of perennial weeds in spring cereals through stubble cultivation and mouldboard ploughing during autumn or spring. *Crop Protection*, 2017. **98**: pp. 16–23.
14. Liebman, M. and A.S. Davis, Integration of soil, crop and weed management in low-external-input farming systems. *Weed Research*. 2000. **40**: pp. 27–47.
15. Liebman, M.L. and E. Dyck, Crop rotation and intercropping strategies for weed management. *Ecological Applications*, 1993. **3**: pp. 92–122.
16. Belz, R. and K. Hurle, Weed suppression by crops – Which role play allelochemicals? *Mitteilungen aus der Biologischen Bundesanstalt für Land- und Forstwirtschaft*. 2000. p. 487.
17. Kolberg, D., et al., Effect of rhizome fragmentation, clover competition, shoot-cutting frequency, and cutting height on quackgrass (Elymus repens). *Weed Science*, 2017: pp. 1–11.
18. Schipanski, M.E., et al., Balancing multiple objectives in organic feed and forage cropping systems. *Agriculture, Ecosystems & Environment*, 2017. **239**: pp. 219–227.
19. Tautges, N.E., et al., Competitive ability of rotational crops with weeds in dryland organic wheat production systems. *Renewable Agriculture and Food Systems*, 2017. **32**(1): pp. 57–68.
20. Bergkvist, G., et al., Control of *Elymus repens* by rhizome fragmentation and repeated mowing in a newly established white clover sward. *Weed Research*, 2017. **57**(3): pp. 172–181.
21. Blackshaw, R.E., et al., Tillage intensity and crop rotation affect weed community dynamics in a winter wheat cropping system. *Canadian Journal of Plant Science*, 2001. **81**(4): pp. 805–813.
22. Ruisi, P., et al., Weed seedbank size and composition in a long-term tillage and crop sequence experiment. *Weed Research*, 2015: pp. 1–9.
23. Schmidt, J.H., S. Junge, and M.R. Finckh, Cover crops and compost prevent weed seed bank buildup in herbicide-free wheat–potato rotations under conservation tillage. *Ecology and Evolution*, 2019. **9**(5): pp. 2715–2724.
24. Davis, A.S., K.A. Renner, and K.L. Gross, Weed seedbank and community shifts in a long-term cropping systems experiment. *Weed Science*, 2005. **53**(3): pp. 296–306.

25. Chellemi, D.O., A.H.C. van Bruggen, and M.R. Finckh, Direct control of soilborne diseases, in *Plant Diseases and their Management in Organic Agriculture*, M.R. Finckh, A.H.C. van Bruggen, and L. Tamm, Editors. 2015, St. Paul, MN: APS Press. pp. 217–226.
26. Hallmann, J. and S. Kiewnick, Diseases caused by nematodes in organic agriculture, in *Plant Diseases and their Management in Organic Agriculture*, M.R. Finckh, A.H.C. van Bruggen, and L. Tamm, Editors. 2015, St. Paul, MN: APS Press. pp. 91–105.
27. Kato-Noguchi, H., J. Mizutani, and K. Hasegawa, Allelopathy of oats. II. Allelochemical effect of L-tryptophan and its concentration in oat root exudates. *Journal of Chemical Ecology*, 1994. **20**(2): pp. 315–319.
28. Schlatter, D., et al., Disease suppressive soils: new insights from the soil microbiome. *Phytopathology*, 2017. **107**(11): pp. 1284–1297.
29. Weller, D.M., et al., Microbial populations responsible for specific soil suppressiveness to plant pathogens. *Annual Review of Phytopathology*, 2002. **40**:pp. 309–348.
30. Kinkel, L.L., M.G. Bakker, and D.L. Schlatter, A coevolutionary framework for managing disease-suppressive soils. *Annual Review of Phytopathology*, 2011. **49**: pp. 47–67.
31. Chaube, H.S. and U.S. Singh, Adjustment of crop culture to minimize disease, in *Plant Disease Management: Principles and Practices*, H.S. Chaube and U.S. Singh, Editors. 1991, CRC Press: Boca Raton, FL. pp. 199–214.
32. Müller, J., Der Einfluß von Larvenschlupf und Weibchenentwicklung bei Heterodera schachtii auf die Resistenzbewertung kreuzblütiger Zwischenfrüchte. *Nachrichtenblatt des Deutschen Pflanzenschutzdienstes*, 1989. **41**(4): pp. 56–56.
33. Altieri, M.A., *Biodiversity and Pest Management in Agroecosystems*. 1994, Haworth Press: New York.
34. Andow, D.A., Vegetational diversity and arthropod population responses. *Annual Review Of Entomology*, 1991. **36**: pp. 561–586.
35. Tooker, J.F. and S.D. Frank, Genotypically diverse cultivar mixtures for insect pest management and increased crop yields. *Journal of Applied Ecology*, 2012. **49**(5): pp. 974–985.
36. Letourneau, D.K., et al., Does plant diversity benefit agroecosystems? A synthetic review. *Ecological Applications*, 2011. **21**: pp. 9–21.
37. Tamm, L., I. Pertot, and D. Gubler, Organic grape disease management, in *Plant Disease Management in Organic Agriculture*, M.R. Finckh, A.H.C. van Bruggen, and L. Tamm, Editors. 2015, St. Paul, MN: APS Press. pp. 335–350.
38. Holb, I. and B. Heijne, Organic apple disease management, in *Plant Disease Management in Organic Agriculture*, M.R. Finckh, A.H.C. van Bruggen, and L. Tamm, Editors. 2015, St. Paul, MN: APS Press. pp. 316–334.
39. Tamm, L., H. Willer, and A.H.C. van Bruggen, Organic perennial crop production, in *Plant Diseases and their Management in Organic Agriculture*, M.R. Finckh, A.H.C. van Bruggen, and L. Tamm, Editors. 2015, St. Paul, MN: APS Press. pp. 33–41.
40. Hartwig, N.L. and H.U. Ammon, Cover crops and living mulches. *Weed Science*, 2002. **50**(6): pp. 688–699.
41. Martínez, E., et al., Habitat heterogeneity affects plant and arthropod species diversity and turnover in traditional cornfields. *PLoS One*, 2015. **10**(7): p. e0128950.
42. Marshall, E.J.P. and A.C. Moonen, Field margins in northern Europe: their functions and interactions with agriculture. *Agriculture, Ecosystems & Environment*, 2002. **89**(1): pp. 5–21.
43. Juroszek, P. and A.V. Tiedemann, Plant pathogens, insect pests and weeds in a changing global climate: a review of approaches, challenges, research gaps, key studies, and concepts. *Journal of Agricultural Science*, 2013. **151**: pp. 163–188.
44. Hallmann, C.A., et al., More than 75 percent decline over 27 years in total flying insect biomass in protected areas. *PLoS One*, 2017. **12**(10): p. e0185809.

45. Van Bruggen, A.H.C., et al., Environmental and health effects of the herbicide glyphosate. *Science of the Total Environment*, 2018. **616–617**: pp. 255–268.
46. Motta, E.V.S., K. Raymann, and N.A. Moran, Glyphosate perturbs the gut microbiota of honey bees. *Proceedings of the National Academy of Sciences*, 2018. **115**: pp. 10305–10310.
47. Kurenbach, B., et al., Agrichemicals and antibiotics in combination increase antibiotic resistance evolution. *Peer-reviewed Journal*, 2018. **6**: p. e5801.
48. Kurenbach, B., et al., Sublethal exposure to commercial formulations of the herbicides dicamba, 2,4-dichlorophenoxyacetic acid, and glyphosate cause changes in antibiotic susceptibility in *Escherichia coli* and *Salmonella enterica serovar typhimurium*. mBio, 2015. **6**(2): pp. e00009–e00015.

42

Pest Management: Intercropping

Maria R. Finckh

Introduction

Until the past few hundred years, agricultural systems were based on large numbers of different crops, crop varieties, and landraces that were heterogeneous in their genetic make-up [1]. In addition, farming systems included both animals and plants further increasing diversity. As a result of increasing specialization, mechanization, and modern plant breeding, diversity on the farming system level and crop level has been drastically reduced worldwide at an ever-accelerating speed especially over the past 100 years. Fewer and fewer varieties that are genetically homogeneous are being grown in ever-larger fields [2]. The large-scale monoculture agricultural practices relying on pesticide and mineral fertilizer inputs have also led to a general decline in soil organic carbon and with this in soil microbial diversity and activity [3,4] with consequent effects on soil fertility and health [5].

Monoculture is usually understood to be the continuous use of a single crop species over a large area. However, with respect to plant pathogens and pests, it is important to differentiate between monoculture at the level of *species*, *variety*, or *resistance genes* (see Table 1) [1,6]. For example, within a species, there may be many different genotypes with different resistances to a specific pest or pathogen and great variation with respect to competitiveness with weeds and other crops. Depending on the breeding system, within a variety, diversity for resistance or morphological traits can be non-existent or high. Especially clonally reproduced or strictly inbreeding species may contain no diversity within a given

TABLE 1 Possibilities for Intercropping at Three Levels of Uniformity on Which Monocultures Are Commonly Practiced [1]

Level of Uniformity	Intercropping Possibilities
Species: Different individuals may differ in genetic make-up (resistance, morphology, etc.)	Arrangements among and within species, varieties, and resistances using intercropping
Variety: Usually genetically uniform, the same gene(s) in the same genetic background	Arrangements among and within varieties and resistances—includes variety mixtures, multilines, and populations
Resistance gene: The same gene may exist in different genetic backgrounds	Arrangements among resistances—multilines and populations

variety. Resistance gene monocultures can arise if different varieties all possess the same resistance (or susceptibility) gene(s). For example, in the late 1960s, virtually all hybrid maize cultivars in the southeastern United States possessed the cytoplasmically inherited Texas male sterility (*Tms*). Unfortunately, *Tms* is closely linked to susceptibility to certain strains of the pathogen *Cochliobolus carbonum* (syn. *Helminthosporium maydis*). The monoculture for susceptibility (while different varieties had been planted) led to selection for these strains and in 1970 the pathogen caused more than $1 billion (= 10^9) losses [7]. Currently, Europe is experiencing the decline of the European ash (*Fraxinus excelsior*) due to the invasion of a novel pathogen (*Hymenoscyphus fraxineus*) for which there is almost no resistance present, that is, a monoculture of susceptibility [8].

Intercropping [9] can be practiced at the species, variety, and gene level (Table 1) with effects on pathogens [1,6,10], insect pests [11,12], and weeds [13–15] (Table 2). One of the most important considerations for the successful design of intercropping systems for pest control is the achievement of *functional diversity*, that is, diversity that limits pathogen and pest expansion and that is designed to make use of knowledge about host–pest/pathogen interactions to direct host and pathogen evolution [1,10]. Moreover, functional diversity is also a matter of complementary use of resource niches, for example, deep versus shallow rooting, legume versus non-legume crops. A famous example, not only in terms of functional diversity but also of human diet, is the successful intercropping of the “Three sisters”, namely maize, bean, and winter squash, in Central and North America since 3500 B.C. [16].

Protection Mechanisms Acting in Intercropped Systems

Pathogens, insect pests, and weeds differ fundamentally in their biology and their effects on crops, and different protection mechanisms act with respect to these organisms (Table 2).

Pathogens are mostly dispersed through wind, water splash, soil, and animals (vectors). In intercropped systems, the most important mechanisms for disease control are mechanical distance and barrier effects and changes in microclimatic conditions due to differences in plant architecture. In addition, resistance reactions induced by avirulent pathogen strains may prevent or delay infection by virulent strains. A large percentage of the reduction of airborne diseases such as the powdery mildews and rusts in cereal cultivar mixtures has been shown to be due to induced resistance [17,18]. The protection mechanisms are universal with respect to airborne, splashborne, and some soilborne, diseases and they may be enhanced by pathogen diversity that, in turn, is enhanced by plant diversity [19]. Mixtures of plants varying in reaction to a range of diseases will lead to a multitude of additional interactions and the overall response in such populations will tend to correlate with the disease levels of the components that are most resistant to these diseases. In addition, less affected plants may compensate for yield losses due to reduced competition from diseased neighbors [1].

In contrast to pathogens for which passive or vectored dispersal is the norm, insects often search actively for their hosts. Thus, behavioral, visual, and olfactory cues play an important role. While environmental factors and landing on a non-host is likely the most important mortality factor for pathogens, natural enemies are at least as important for insect population dynamics [11,12]. Host dilution may affect an insect’s ability to see and/or smell its hosts. Predators and parasitoids are dependent on the constant presence of prey and alternative food sources, such as pollen and nectar, in the absence of the hosts and for effective control of insect pests, the presence of sufficient numbers of natural enemies is critical. The importance of natural enemies was often only recognized after insecticide applications induced pest resurgence due to the destruction of natural enemy populations. Intercrops and weeds therefore can play an important role in regulating insect pests. Plant–insect communication also plays a role. For example, plants may signal their neighbors about insect attacks leading to the production of antinutritive compounds or attractants for natural enemies [20,21].

Weeds usually are early successional plants adapted to colonize open, nutrient-rich spaces. Intercrops, especially cover and mulch crops, directly compete with weeds for these spaces and also for light. As many weeds are adapted to certain crops and cropping patterns, changing these patterns (e.g., rotations)

TABLE 2 Mechanisms Affecting Pathogens, Insect Pests, and Weeds in Intercropped Systems and Selected Additional Interactions of Importance [1,13,19,29]

Mechanisms Reducing Disease
Increased distance between susceptible plants
Barrier effects of intercrop
Variation in plant architecture may lead to less humid microclimate
Induced resistance
Selection for the most resistant and/or competitive genotypes
Interactions among pathogen strains on host plants
Increased microbial diversity and activity above and below ground
Mechanisms Reducing Insect Pests
Enhancement of natural enemies
Reduction of host density (reduced resource concentration)
Reduction of plant apparency (visual or olfactory cues reduced)
Alteration of host quality (with respect to the insect pest) through plant–plant and plant–microbe interactions
Increased microbial diversity and activity above and below ground
Mechanisms Reducing Weeds
Reduction of bare soil and layering of crops (increased competition for light, water, and nutrients)
Variation in tillage needs and operations of intercrops may disturb weeds
Other Beneficial Interactions
Yield enhancement through niche differentiation of hosts
Compensation for yield losses by less affected hosts
Better soil cover with intercrop (soil and water conservation, microclimatic effects)
Possible Unwanted Interactions
Weeds may serve as alternate hosts for pathogens and insects
Interactions among virus vectors and weeds
Greater difficulty to specifically reduce weeds with herbicides or mechanically
Overall denser intercropped stands may produce a more humid microclimate that may enhance certain problems

and management operations (e.g., sowing time) connected with different kinds of crops within the same field impede the adaptation and dominance of (problem) weeds. Also, filling the spaces that usually would be taken up by weeds with useful or more neutral plants will reduce weed habitat and help out-competing them [13,14,22]. An important consideration is that plants may be weeds only during certain phases of crop development. At other stages, the presence of the same "weeds" may be beneficial because they may provide food and habitat for beneficial insects and erosion control.

Besides the many positive effects of intercrops, it is important to keep in mind that weeds may serve as alternative hosts for insect pests and pathogens and that insects often are disease vectors, especially for viruses that may reside symptomless in certain weeds [1].

In order to understand the many interactions in intercropped systems, it is indispensable not only to consider plants, insects, and pathogens in the system but also the whole microbiome that interacts with them [23]. Plants not only take up nutrients from the soil, but they also actively release organic carbon-based chemicals such as organic acids and sugars into the soil. Roughly speaking about 30%–50% of the carbon that plants assimilate through photosynthesis is released into the soil [24]. These compounds provide the energy (carbon) and organic acids necessary for the microorganisms in order to function as well as for supporting the weathering of soil minerals making them available as plant nutrients. In addition, most plants live in symbiosis with root infecting mycorrhizal fungi that usually are very specific and genetically diverse [25]. The diversity of root exudates and in turn the diversity of the soil microbiome is greatly enhanced by plant diversity including intraspecific diversity (e.g., [26]), and by this, many of the interactive processes between crops, weeds, insects, and pathogens are influenced. For example, high microbial diversity and activity in the soil usually supports resistance induction and direct disease suppression [19,27]. Also, suppression of the production of certain secondary metabolites by plants can be triggered by certain rhizosphere bacteria. In turn, this may reduce their attractiveness to certain insects. In the case of cucumbers, this is helpful in suppressing the spotted and striped cucumber beetles (*Diabrotica undecimpunctata* and *Acalymma vittata*, respectively) that act as vectors (transmitting agents) of bacterial wilt of cucumber due to *Erwinia tracheiphila* [28].

Intercropping in Practice

Variety mixtures and multilines are used mainly to control diseases. For example, they are used in cereals on a commercial scale in the United States, Denmark, Finland, Poland, and Switzerland to control rusts, mildews, and certain soilborne diseases (e.g., *Cephalosporium* stripe). When barley cultivar mixtures were used on more than 360,000 ha in the former German Democratic Republic, powdery mildew of barley and consequently fungicide input was reduced by 80% within 5 years [1]. Wheat cultivar mixtures and multilines are grown on several hundred thousand hectares in the United States in the Pacific Northwest and in Kansas to protect against diseases and abiotic stresses. In Colombia, coffee multilines are grown on more than 700,000 ha to control coffee rust [1].

Attention has also been called to possible beneficial effects of greater intravarietal diversity in the oat-frit fly (*Oscinella frit* L.) system. The flies can attack the host plants only at a particular growth stage and a higher degree of variability within an oat crop could allow for escape from attack and subsequent compensation.

Cereal species mixtures for feed production have been reported as important practice in Poland and have been shown consistently to restrict diseases. In Switzerland, the “maize-ley” system (i.e., maize planted without tillage into established leys), which is being promoted to reduce soil losses and nutrient leaching, has been shown to reduce smut disease and attacks by European stem borer and aphids [30].

The deliberate planting or maintenance of flowering weeds and grass in established vineyards and orchards in Switzerland and Germany greatly increases natural enemies while reducing soil erosion. This practice is becoming increasingly popular world-wide if water is not the limiting factor in the system.

The required reduction in insect populations for effective reduction of insect-transmitted diseases may be beyond that which can be achieved by diversity alone. However, simultaneous reduction of insect vectors, for example, by enhancing natural enemies and plant resistance and diversity and disease inoculum, for example, by dilution of the susceptible hosts through intercropping with resistant varieties or non-host species can be effective [31,32]; in addition, interactions with the soil microbiome making hosts less attractive to their insect pests as described above [28] may play a role.

Future Concerns

There are many reasons why intercropping is not practiced more widely. First, modern crops are bred to be grown in monoculture and may not necessarily be well adapted to intercropping. Efforts of breeders to produce breeding lines adapted to intercropping need to be strengthened. Breeding for intra-and inter-specific diversity has recently gained momentum [10,33]. EU legislation has been put into place to support these activities at least for the organic sector [34], and there is a need to expand this into the agricultural sector as a whole. Second, while intercropping clearly provides a means for reducing pesticide needs, there is a lack of adapted machinery allowing for efficient management of intercropped cultures. Third, successful intercropping strategies have to be carefully designed as preventive measures that include the whole growing system and rotational cycle [23], while monoculture cropping is very simple due to the availability of (very) cheap fertilizer and pesticides that can be applied once a problem occurs. However, pesticides are more and more criticized for their unintended side-effects causing insect decline [35] and affecting human and environmental health [36,37] and alternative approaches based on applied agroecology are needed [38]. Fourth, a concern often raised is the quality of products raised as intercrops such as varietal mixtures of cereals. In some countries, there is resistance in the food processing industry to such products. However, such problems could be overcome if breeders, producers, and processors work together. For example, in the 1980s, in the German Democratic Republic, first-quality malting barley was produced in large-scale mixtures in collaboration with breeders, growers, and processors [39]. Also, some of the highest-quality coffee in the world from Colombia is produced from multilines demonstrating that quality produce is not dependent on genetic uniformity. For the use of products grown from species mixtures, technological advances (e.g., color- or shape-based optical and/or mechanical separation devices during harvest or processing) need to be implemented.

References

1. Finckh, M.R. and M.S. Wolfe, Biodiversity enhancement, in *Plant Diseases and Their Management in Organic Agriculture*, M.R. Finckh, A.H.C.V. Bruggen, and L. Tamm, Editors. 2015, St. Paul, MN: APS. pp. 153–174.
2. Agardy, T., et al., *Ecosystems and Human Well-Being: Biodiversity Synthesis. Millennium Ecosystem Assessment*. 2005, Washington, DC: World Resources Institute.
3. Bellamy, P.H., et al., Carbon losses from all soils across England and Wales 1978–2003. *Nature*, 2008. **437**: pp. 245–248.
4. Capriel, P., Trends in organic carbon and nitrogen contents in agricultural soils in Bavaria (south Germany) between 1986 and 2007. *European Journal of Soil Science*, 2013. **64**: pp. 445–454.
5. Cook, R.J., Mangement of resident plant growth-promoting rhizobacteria with the cropping system: a review of experience in the US Pacific Northwest. *European Journal of Plant Pathology*, 2007. **119**: pp. 255–264.
6. Finckh, M.R. and M.S. Wolfe, Diversification strategies, in *The Epidemiology of Plant Diseases*, B.M. Cooke, D. Gareth Jones, and B. Kaye, Editors. 2006, Heidelberg: Springer-Verlag. pp. 269–307.
7. Ullstrup, A.J., The impacts of the southern com leaf blight epidemics of 1970–1971. *Annual Review Phytopathology*, 1972. **10**: pp. 37–50.
8. Landolt, J., et al., Ash dieback due to *Hymenoscyphus fraxineus*: what can be learnt from evolutionary ecology? *Plant Pathology*, 2016. **65**(7): pp. 1056–1070.
9. Vandermeer, J.H., *The Ecology of Intercropping*. 1989, Cambridge, New York, Melbourne: Cambridge University Press. pp. 1–235.
10. Finckh, M.R., Integration of breeding and technology into diversification strategies for disease control in modern agriculture. *European Journal of Plant Pathology*, 2008. **120**: pp. 399–409.
11. Tooker, J.F. and S.D. Frank, Genotypically diverse cultivar mixtures for insect pest management and increased crop yields. *Journal of Applied Ecology*, 2012. **49**(5): pp. 974–985.
12. Letourneau, D.K., et al., Does plant diversity benefit agroecosystems? A synthetic review. *Ecological Applications*, 2011. **21**: pp. 9–21.
13. Liebman, M.L. and E. Dyck, Crop rotation and intercropping strategies for weed management. *Ecological Applications*, 1993. **3**: pp. 92–122.
14. Liebman, M.L. and E.R. Gallandt, *Many little hammers: ecological management of crop-weed interactions*, in *Ecology in Agriculture*, L.E. Jackson, Editor. 1997, San Diego, CA, New York, London: Academic Press. pp. 291–343.
15. Poggio, S.L., Structure of weed communities occurring in monoculture and intercropping of field pea and barley. *Agriculture Ecosystems & Environment*, 2005. **109**: pp. 48–58.
16. Hart, J.P., Evolving the three sisters: the changing histories of maize, bean, and squash in New York and the Greater Northeast, in *Current Northeast Paleoethnobotany II*, J.P. Hart, Editor. 2008, Albany, NY: The University of the State of New York. pp. 87–99.

17. Calonnec, A., H. Goyeau, and C. Devallavieillepope, Effects of induced resistance on infection efficiency and sporulation of *Puccinia striiformis* on seedlings in varietal mixtures and on field epidemics in pure stands. *European Journal of Plant Pathology*, 1996. **102**: pp. 733–741.
18. Walters, D.R., Are plants in the field already induced? Implications for practical disease control. *Crop Protection*, 2009. **28**: pp. 459–465.
19. Finckh, M.R. and L. Tamm, Organic management and airborne diseases, in *Plant Diseases and their Management in Organic Agriculture*, M.R. Finckh, A.H.C.V. Bruggen, and L. Tamm, Editors. 2015, St. Paul, MN: APS Press. pp. 53–66.
20. Ninkovic, V., R. Glinwood, and J. Pettersson, Communication between undamaged plants by volatiles: the role of allelobiosis, in *Communication in Plants*, F. Baluka, S. Mancuso, and D. Volkmann, Editors. 2006, Berlin: Springer. pp. 421–434.
21. Ninkovic, V., U. Olsson, and J. Pettersson, Mixing barley cultivars affects aphid host plant acceptance in field experiments. *Entomologia Experimentalis et Applicata*, 2002. **102**(2): pp. 177–182.
22. Finckh, M.R. and A.H.C.V. Bruggen, Organic production of annual crops, in *Plant Diseases and their Management in Organic Agriculture*, M.R. Finckh, A.H.C.V. Bruggen, and L. Tamm, Editors. 2015, St. Paul, MN: APS Press. pp. 25–32.
23. Bruggen, A.H.C.V., et al., One health—Cycling of diverse microbial communities as a connecting force for soil, plant, animal, human and ecosystem health. *Science of the Total Environment*, 2019. **664**: pp. 927–937.
24. Marschner, H., *Marschner's Mineral Nutrition of Higher Plants (Third Edition)*, P. Marschner, Editor. 2012, San Diego, CA: Academic Press.
25. Smith, S.E. and D.J. Read, *Mycorrhizal Symbiosis*. 2010, New York: Elsevier Science.
26. Mazzola, M. and Y.-H. Gu, Wheat genotype-specific induction of soil microbial communities suppressive to disease incited by *Rhizoctonia solani* anastomosis group (AG)-5 and AG-8. *Phytopathology*, 2002. **92**: pp. 1300–1307.
27. Schlatter, D., et al., Disease suppressive soils: new insights from the soil microbiome. *Phytopathology*, 2017. **107**(11): pp. 1284–1297.
28. Zehnder, G.W., et al., Application of rhizobacteria for induced resistance. *European Journal of Plant Pathology*, 2001. **107**: p. 39–50.
29. Andow, D.A., Vegetational diversity and arthropod population responses. *Annual Review of Entomology*, 1991. **36**: pp. 561–586.
30. Finckh, M.R., et al., Cereal variety and species mixtures in practice, with emphasis on disease resistance. *Agronomie*, 2000. **20**: pp. 813–837.
31. Power, A.G., *Competition between viruses in a complex plant-pathogen system*. *Ecology*, 1996. **77**: pp. 1004–1010.
32. Finckh, M.R. and M.S. Wolfe, *The use of biodiversity to restrict plant diseases and some consequences for farmers and society*, in *Ecology in Agriculture*, L.E. Jackson, Editor. 1997, San Diego, CA: Academic Press. pp. 199–233.
33. Baćanović-Šišić, J., D. Dennemoser, and M.R. Finckh. *Symposium on breeding for diversification. A joint meeting of the EUCARPIA section, organic and low-input agriculture*, ECO-PB, LIVESEED, INSUSFAR, DIVERSify, HealthyMinorCereals, ReMIX, and Wheatamix University of Kassel, 19th–21st February 2018, Witzenhausen, Germany. 2018. Kassel University Press.
34. EC, Commission implementing decision of 9 October 2018 amending implementing decision 2014/150/EU on the organisation of a temporary experiment providing for certain derogations for the marketing of populations of the plant species wheat, barley, oats and maize pursuant to Council Directive 66/402/EEC. *Official Journal of European Union*, 2018. **L**: p. L256/65–L256/66.
35. Hallmann, C.A., et al., More than 75 percent decline over 27 years in total flying insect biomass in protected areas. *PLoS One*, 2017. **12**(10): p. e0185809.
36. Bruggen, A.H.C.V., et al., Environmental and health effects of the herbicide glyphosate. *Science of the Total Environment*, 2018. **616**: pp. 255–268.

37. Kurenbach, B., et al., *Agrichemicals and antibiotics in combination increase antibiotic resistance evolution*. *Peer-Reviewed Journal*, 2018. **6**: p. e5801.
38. Finckh, M.R., et al., Disease and pest management in organic farming: a case for applied agroecology, in *Improving Organic Crop Cultivation*, Vol. 1, U. Köpke, Editor. 2018, Cambridge: Burleigh Dodds. pp. 271–301.
39. Wolfe, M.S., Barley diseases: maintaining the value of our varieties, in *Barley Genetics VI*, L. Munk, Editor. 1992, Copenhagen: Munksgaard International Publishers, Ltd. pp. 1055–1067.

43

Precision Agriculture: Water and Nutrient Management

Robert J. Lascano,
Timothy S. Goebel,
and J.D. Booker

Introduction

Management of agronomic inputs, such as water and fertilizers, to cropping systems requires information on when and how much of each input to apply. The correct management decision is essential to maintain the productivity of the cropping system, and a management strategy is to use crop production functions that describe (mathematically) maximum economic yield (profit) as a function of agronomic inputs. Essentially, the strategy is to apply the least amount of input to produce the maximum economic crop yield in the farming operation. In conventional cropping systems, inputs are generally applied uniformly across a field regardless of their need, and the amount applied is normally based on average responses of these inputs to crop yield across the field. Since the 1970s, increased costs of crop production and emphasis on production efficiency revolutionized production agriculture. Environmental concerns, including quality and safety of harvested products and impact of the cropping procedures on the ecosystem, are also considered in the decision-making process. Developments in spatial statistics and computer hardware (i.e., increased microprocessor speed and decreased cost), and increased availability of soil, elevation, and weather data have contributed to improved concepts and procedures to address spatial and temporal variability in cropping systems.

Precision agriculture (PA) and precision farming are generic terms that describe the way production management inputs (e.g., water, nutrients, harvest aids, and herbicides) are managed in response to cropping system variability. In contrast to a uniform blanket application of an input across the field, each input is applied according to specific needs across the field. PA is an integrated crop management system that attempts to match the kind and amount of inputs with the actual crop needs for small areas within a farm field. Perhaps a better descriptor for this type of farming is site-specific management, which manages an agricultural crop at a spatial scale smaller than the entire field by dividing it into management zones, defined by topography, soil type, and level of a particular nutrient, such as nitrogen and phosphorus.

Driving forces that have contributed to producers implementing and adopting PA procedures are advances in computer hardware and software, electronics, and equipment technology (sensors); decreased profit margins due to increased costs of production; and environmental awareness. For example, advances in crop growth simulation models, variable-rate application equipment, adoption of soil sampling for nutrients across the field and as a function of soil depth, and integration of crop yield monitors with global positioning satellite systems have contributed to the use of PA concepts to manage crops. The cost of agronomic inputs continues to increase, for example, petroleum-based inputs such as fertilizer, diesel, insecticides, and herbicides. Awareness of environmental concerns related to nutrient contamination of groundwater and surface water and the quality and safety of food and fiber are factors that impact how crops are produced and delivered to the consumer. Operating a farm requires management strategies that consider both economic and environmental consequences, and PA provides the concepts that can be used to achieve this goal.

In recent years, the development of sensors to measure soil properties[1,2] and of using remote sensing techniques[3,4] has increased the availability of georeferenced digital data used in PA. These large databases are generically referred to as "Big Data", and examples of their application to PA are given by Sonka[5] and by Fulton and Port.[6] Further, the availability of real-time weather information via networks known as Mesonets[7] and the use of software applications that are accessed via smart phones[2,8,9] has both modernized and revolutionized crop production and gives producers information to manage their farming operation more efficiently. Examples of these applications for nutrient and water management are given by Hedley,[10] Abit et al.[11] and by Neupane and Gao.[12] In PA, the rate of adoption and use of this technology will continue to increase. The factors that contribute to this implementation are the decreased cost of computer hardware, increased availability of input data, and development of computer software providing tools to manage the farming enterprise more efficiently. Also, tailoring production inputs to site-specific locations contributes to improve the sustainability of the food supply.[13]

This chapter gives a general overview of PA with emphasis on crop water and nutrient use, and crop yield using the state-space analysis to describe, for example, how cotton lint yield varies at a landscape scale. This entry is divided into five parts. First, we give a general overview of PA. Second, we describe the relation between crop yield and water use. Third, as examples, we show measurements of cotton water use along two 700 m transects on a 60-ha field and the use of a large-scale model to calculate the temporal and spatial variability of soil water content in a 70-ha field. Fourth, we use geostatistical tools to quantify cotton lint yield as a function of nitrogen, topography, and soil water. Fifth, we describe how the future of PA will leverage knowledge from PA research, such as that presented here, to further support the development and use of large-scale cropping system models.

Precision Agriculture

PA, also known as site-specific management,[13,14] refers to the practice of applying agronomic inputs across a farm, mainly fertilizers and other chemicals, at variable rates based on soil nutrients or chemical tests, soil textural changes, weed pressures, and/or yield maps for each field in the farm.[13–15] In large fields (e.g., >40 ha), crop yield and thus crop water and nutrient use are notoriously variable.[12,14,16] The sources of this variation are related to soil physical and chemical properties, pests, microclimate, genetic and phenological responses of the crop and their interactions.[16] The technology for crop yield mapping is more advanced than current methodologies for determining and understanding the causes of yield variability. Prevailing and traditional management practices treat fields uniformly as one unit. However, reports[15–18] show that to understand underlying soil processes that explain crop yield variability, research must be done at the landscape level and using appropriate statistical tools for large-scale studies.[18]

Crop Yield and Water Use

There is a linear relation between crop yield and water use when the only limiting factor is water[19]; however, root water uptake is synergistically related to nutrient uptake, and the two processes cannot be separated. Precision farming has the potential to improve water and nutrient use efficiency on large fields provided there is quantitative understanding of what factors affect crop water and nutrient use and how they vary across the field.[16,20,21] It is known that crop water and nutrient use are a function of many biotic and abiotic variables, including managed inputs, and harvestable yield is a manifestation of how these variables and inputs interact and are integrated during the growing season. However, it is difficult to determine a hierarchy on the contribution of each input and variable to the measured crop yield using classical statistics.[22–25] Often, variables that affect water and nutrient supply to the plant contribute to crop yield at a high level assuming an adequate plant stand and weed control. The cause-and-effect relation between a single state variable and crop yield is site specific and is difficult to establish without considerable sampling of the soil and/or crop. The establishment of response functions, that is, crop water and nutrient use as a function of variable x_i, gives only a partial answer to explain crop water use, nutrient use, and yield based on inputs. The general idea of PA is to optimize input application to the measured crop yield at each sampling location using the law of the optimum formulated by Liebscher, which states that a production factor that is in minimum supply contributes more to production, the closer other production factors are to their optimum.[26] This is a simple premise; however, the decisions for variable-rate application of any agronomic input must consider temporal and spatial variability of the soil's properties affecting crop growth, water and nutrient use, and yield. Soil factors that affect water storage, such as depth of the root-restricting layer and soil textural differences, must be considered in any precision farming operations that attempt to improve crop water use and yield related to agronomic inputs. Similarly, to improve the use of any micro- and macronutrient by the crop, the overall cycle of the nutrient must be considered, including its availability in the soil and demand by the crop. Examples of managing nitrogen fertilization and irrigation at a site-specific and farm scale for cotton production is given by Bronson et al.,[17] Li et al.,[18] and Booker et al.[27]

Precision farming must incorporate the inherent spatial and temporal variability of soil physical, chemical, and biological factors within a field for input management. Accurate representation of spatial and temporal variability in a field requires taking and analyzing many samples. Sampling is normally done on a grid with a scale that can vary from one to several hundred meters.[22,23] Once properties are measured, geostatistical tools (e.g., semivariogram, kriging, cokriging), and other spatial statistical tools (e.g., autocorrelation, cross-correlation, state-space analysis), can be used to establish statistical relations in space and to minimize the number of soil samples to characterize and map fields.[22,23,25] The number of samples required *a priori* to determine spatial and temporal variability is perhaps the single largest deterrent in the application of precision farming practices to manage and improve crop water and nutrient use. Collection of field data to characterize the spatial variability must remain a priority for any study that attempts to understand how to maximize crop yield across the landscape. The trend of using data generated from pedotransfer functions[28] rather than field measured data is of concern and results thus obtained are preliminary at best and should be verified using measured field data. Further, input data generated from pedotransfer functions and used to assess the spatial variability of soil properties can be used as a guide to establish a sampling scheme to measure the properties of interest. Input data generated from pedotransfer functions are not a substitute to field measured data.

There is not much information published on combined crop water and nutrient use across large fields at the landscape level and in the context of precision farming.[17,29–31] An exception is the study by Li et al.[18] where cotton water and total nitrogen uses were measured along a 700-m transect to illustrate the landscape pattern of cotton water and total nitrogen uses and to determine the underlying soil processes governing cotton lint yield variability. In this study, state-space analysis[18,22,25] was used to formulate management decisions that may improve crop water and nitrogen use and lint yield using precision farming practices. An additional study regarding variable-rate nitrogen at different locations within a

48-ha field is given by Bronson et al.,[17] and in a 14-ha field, by Booker et al.[27] A global-scale assessment showing that global yield crop variability is a function of fertilizer use, irrigation, and climate is given by Mueller et al.[32] and a review of integrated nutrient management to sustain crop productivity while minimizing the environmental impact is given by Wu and Ma.[33]

Landscape Crop Water Use

The concept of crop water and nitrogen use in a 60-ha field study is given by Li et al.[18] and Li and Lascano.[34] In 1999, a field experiment was conducted near Lamesa, Texas, on a research farm of Texas A&M University on the southern edge of the High Plains of Texas. The soil was classified as an Amarillo sandy loam. The field was 60 ha with slopes ranging between 0.3% and 6.3%.[18] To evaluate the effect of soil water, nitrate-nitrogen (NO_3-N), and topography on cotton lint yield across the landscape, two irrigation levels were used. The irrigation treatments consisted of water applications at a 50% and 75% grass reference evapotranspiration (ET_o) with a center-pivot low-energy precision application irrigation system.[35] At each irrigation level, one transect was established following the circular pattern of the center pivot. The two transects were instrumented with 50 neutron access tubes, each 15 m apart, and soil volumetric water content (θ_v) was measured periodically throughout the growing season. At each point, θ_v was measured in 0.3 m depth increments to 2.0 m depth using a neutron probe calibrated for this soil. In addition, at each transect point soil texture, soil and plant NO_3-N, leaf area index, cotton lint yield, slope, plant density, and other parameters were measured.[18,34]

A comprehensive study to evaluate site-specific management of cotton production systems at a landscape scale was done by Booker et al.[36,37] using the Precision Agricultural Landscape Modeling System (PALMS),[38] which was integrated with the cotton simulation model GOSSYM.[39] The combined model, PALMSCot, was applied to simulate a cotton crop irrigated with a 400-m center pivot system and covering a ¼-section of land (~65 ha). The PALMSCot model is grid-based (10-m resolution) and defines up to 26 soil depth layers and for each "cell" the water, energy, nutrient, and carbon balance is calculated on a ¼-h time interval for the length of the growing season. This model provides a detailed calculation of the water and nitrogen use for each cell across the field. For example, for an irrigated field with a 400-m-long center pivot sprinkler system, PALMSCot calculates the cotton lint yield, evapotranspiration, and nitrogen use for each of the 5,026 cells defined by the 10 × 10 m grid system (length and width) and the assigned soil depth.

Statistical Calculations

It has been shown that the use of classical statistics, such as regression analysis and analysis of variance, is designed to describe the strength of the covariance structure between variables and fails to completely explain the cause and effect between, for example, crop yield and measured soil variables in precision farming experiments.[18,22,23,25,31] These techniques, in general, account for spatial and temporal variability through blocking and do not describe the spatial and temporal structure. Instead, there are other more appropriate statistical tools for relating the variability of soil and plant parameters measured in space and time. For example, the structure of the spatial (or less often, temporal) variance between measurements may be derived from the sample semivariogram, which is the average variance between neighboring measurements separated by the same distance. Spatial or temporal structure between variables is often determined using autocorrelation and cross-correlation functions. Although these techniques can be used to evaluate the temporal variability structure, they are most often used in PA to analyze spatial variables. Autocorrelation measures the linear correlation of a variable in space along a transect. Cross-correlation is the comparison of two variables measured along a transect and is used to describe the spatial correlation between two landscape variables, that is, where one variable, the tail variable, lags behind the head variable by some distance. The spatial association between several variables can be described using state-space analysis, which is a multivariate autoregressive technique.[18,22]

Spatial Analysis of Crop Water Use

To illustrate the variability of crop water use or crop evapotranspiration (ET_c), values measured along the 50% irrigation transect were selected.[18] In Figure 1, the relation between the scaled ET_c and elevation, both as a function of distance along the transect, is shown. The ET_c data are scaled to the maximum value of 426 mm of water, which was measured 210 m from the south end of the transect. These results show, as expected and, that in general higher ET_c was measured at lower elevations and that ET_c tended to decrease at the higher elevations. Higher elevations are eroded and the depth of the root zone is shallower holding less water. In contrast, lower elevations tend to have more clay and thus hold more water.

Calculated values of soil volumetric water content for the surface 0.05 m for a center pivot irrigated cotton field in Floydada, Texas, are shown in Figure 2.[36,37] The four circles shown in Figure 2 illustrate the corresponding wetted soil area (blue) by the center pivot, with values of 45% soil water content over a 6-day period, starting on the 10 July and ending on the 16 July. This is an example of the spatial resolution, 20 × 20 m, and temporal variability that can be used to analyze crop water use across the landscape that can be obtained with a model such as PALMSCot.

Spatial cross-correlation between cotton lint yield and soil water, cotton lint yield and site elevation, and soil water and site elevation are shown in Figure 3. For a 95% confidence interval, the cotton lint yield was positively cross-correlated with soil θ_v across a lag distance of ±30 m. Cotton lint yield and θ_v were negatively cross-correlated with elevation at a lag distance of ±30 m. These results show the effect of topography on the θ_v and of crop water use measured along the transect. Similar results are given in other reports.[17,30,31,40] In this example, the cross-correlation between θ_v and elevation shows the spatial structure of measured variables and, further, that more water was stored in lower elevations, resulting in higher ET_c.

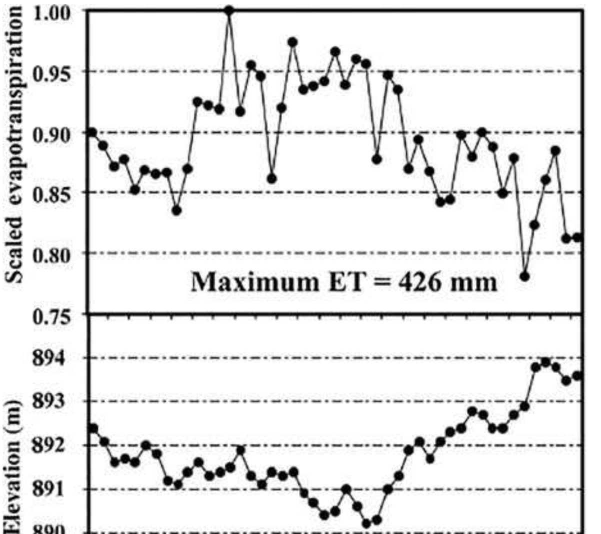
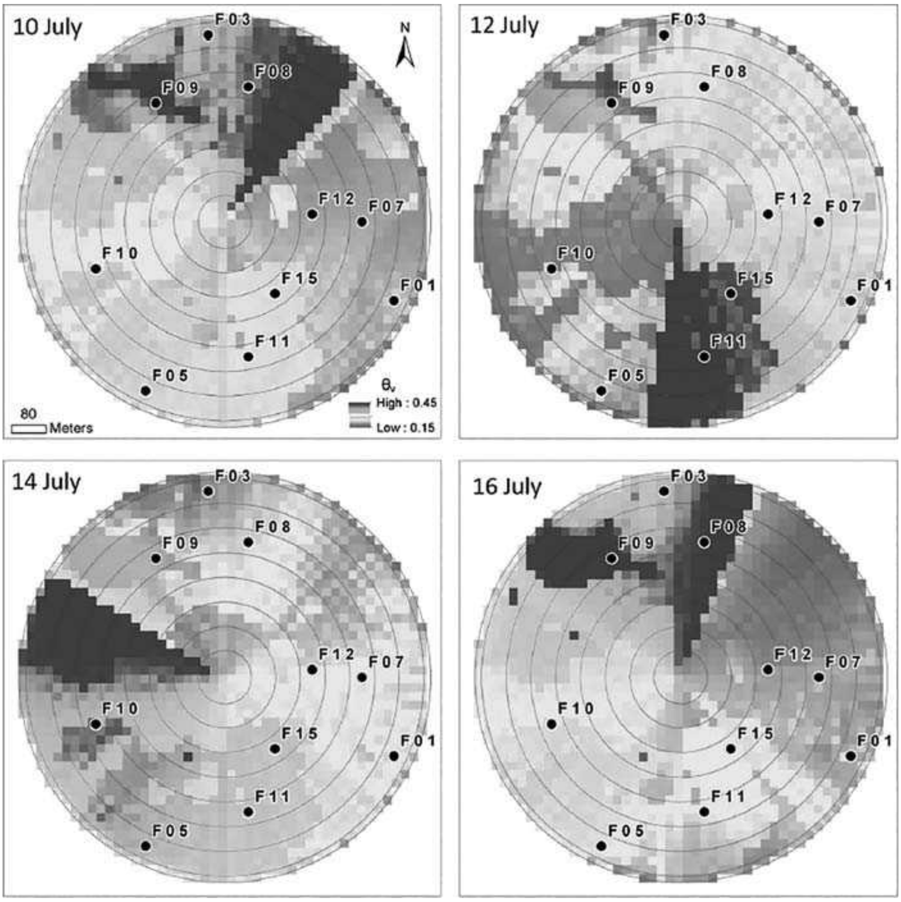


FIGURE 1 Scaled crop evapotranspiration and elevation as a function of distance along a 700-m transect.


10 July

12 July

14 July

16 July

FIGURE 2 Calculated values of soil volumetric water content for the surface 0.05 m obtained with the PALMSCot model using 20 × 20 m grid cells for the period from 10 to 16 July 2011. This is a center pivot irrigated cotton production field located in Floydada, Texas. The soils at the site are classified as a Pullman clay loam with 0%–1% slope and an Olton clay loam with a 1%–3% slope. The blue zones correspond to irrigation water applied by the center pivot sprinkler system.
Source: Booker et al.[37]

Linear regression analysis between θ_v and cotton lint yield and relative site elevation is shown in Figure 4, and the state-space analysis for the relation between cotton lint yield and three measured parameters is shown in Figure 5. Results in Figure 4 show the shortcomings of using an inappropriate statistical tool to understand underlying processes explained with the state-space analysis. This analysis (Figure 5) quantified how cotton lint yields varied as a function of distance and showed that by using θ_v, soil NO_3-N and elevation the variation in cotton lint yield can be explained with a high level of confidence. While studies like those of Li et al.,[18] Bronson et al.,[17] and Booker et al.[27] are empirical in design, the relationships that are evaluated provide important validation and field testing of the more mechanistic mass and energy balance accounting provided by models such as PALMS.[38] These studies also provide foundational information for developing PA management strategies at the crop production scale.

Benefits of PA to improve crop water and nutrient use may be obtained by an economic analysis of maximizing crop yield as a function of application of nitrogen fertilizer and irrigation water as given by the state-space equation.[18,34] In the example given, the decision can be made to apply more

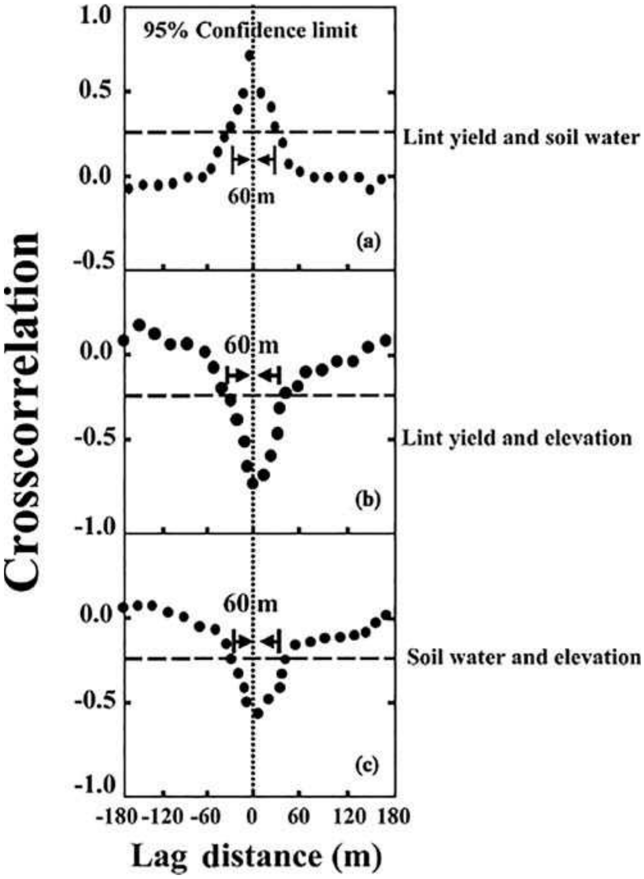


FIGURE 3 Cross-correlation as a function of lag distance. (a) Lint yield and soil water, (b) lint yield and elevation, and (c) soil water and elevation. Shown is the 95% confidence for the cross-correlation distance.
Source: Li et al.[18]

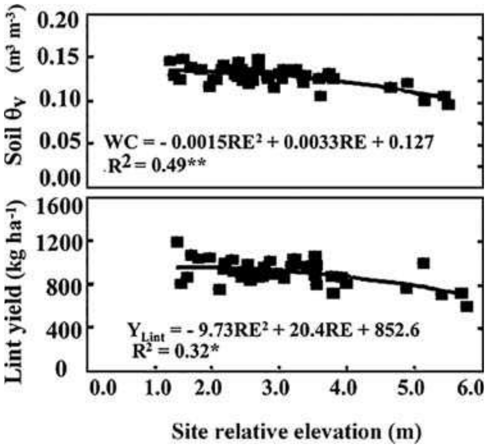


FIGURE 4 Soil water content (θ_v) and cotton lint yield as a function of site relative elevation.

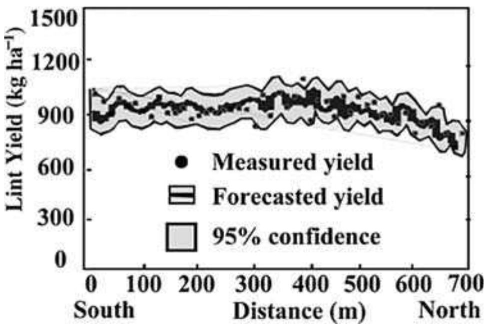


FIGURE 5 State-space equation relating cotton lint yield (*Y*) to water content (*W*), nitrogen level (*N*), and elevation (*E*) as a function of distance and location (*i*) along a 700-m transect.
Source: Li et al.[18]

nitrogen fertilizer to the lower areas of the field that also hold more water and increase crop water use and nitrogen and lint yield. With the introduction of variable rate planters, it is possible to discriminate site locations and plant more "drought"-tolerant varieties or change the seeding rate in areas that are prone to have less soil water. This implies the delineation of management zones[14,15] within a field that are defined based on potential crop water use and their interaction with other input variables to maximize economic yield across the field. This type of precision farming is slowly being adopted, and wide use remains within the realm of possibilities that this type of farming has to offer. The introduction and use of computer models of cropping systems will likely expedite and facilitate the adoption of PA management techniques.[38]

$$Y_{(50\%ET)i} = -0.201\ Y_{i-1} + 1.107\ W_{i-1} + 0.332\ N_{i-1} D49.54\ E_{i-1} + \varepsilon_i$$

A final consideration is the cost/benefit of PA practices and its impact on agriculture. Currently, hardware for variable-rate application of agronomic inputs is relatively expensive and in many cases unavailable; however, with increased adoption and use of these practices, the cost will be reduced. For example, tractor guidance systems[41,42] were quickly adopted by producers, and high demand reduced its cost. Further, environmental and material cost concerns for a given area will probably place limits on the amount of certain nutrients, for example, nitrogen fertilizer, used for crop production. This will force producers to apply nitrogen and other nutrients across the field according to site-specific needs and position along the landscape. These practices will be beneficial from both an environmental and an economical point of view.[10,11,20,33]

Future of PA

Considerable PA-related research, similar to that presented above, has been conducted over the past decade, studying empirical relationships and attempting to better understand the underlying processes controlling crop yield variability. Much of this research has focused on grid soil and crop sampling,[43] surface characterization (e.g., apparent electrical conductivity),[44] and ground- or aerial-based remote sensing.[3,4,45,46] Such research has described numerous useful process relationships but has been somewhat less successful in providing broad-based tools to support production-scale management. The lack of development of decision-support systems to implement precision decisions is the major impediment to the adoption of PA.[47]

Recent advances in the availability of soil data provided by the U.S. Department of Agriculture, elevation data provided by the U.S. Geological Survey, and weather data from weather Mesonets (networks) provide the necessary input to model the water, energy, nutrient, and carbon balance of large-scale agricultural fields. An example of such a model is the Precision Agricultural-Landscape Modeling System (PALMS) given by Molling et al.[38] The integration of PALMS with crop growth models[36,37] provides a framework whereby site-specific management of crops is an achievable goal.

The concept of using simulation models to manage crops was introduced in the 1970s. Many of the theoretical algorithms related to model crop photosynthesis and transpiration were formulated by C.T. de Wit and coworkers at Wageningen University, Netherlands.[48] An example of such model is the simulation of field water use and crop yield given by Feddes et al.[49] An example of a crop-specific model, that is, cotton, known as GOSSYM/COMAX was developed by McKinion et al.[39] This model was used by crop consultants in the Texas High Plains to provide services on irrigation scheduling and application of nitrogen fertilizer, growth regulators, and chemicals to terminate the crop.[50] The biggest drawback ofthe application of these models was that the required inputs, soil and weather, were both difficult and expensive to obtain. Furthermore, these models provided only an average estimate of crop yield for the entire field regardless of size. The models could be run separately for different parts of a field, but this increased the demand on limited computer and input resources and even then did not represent the interaction between various parts of the field. In retrospect, we now recognize that these models were ahead of their time. Given the current availability of soil and weather data that is required by these models, along with the increased computer speed and reduced cost, a resurgence in the application of simulation models to manage crops and cropping systems is anticipated.

In coming years, a likely scenario to emerge to manage cropping systems will be based on the combination of three factors. First, is the realization from site-specific management that shows that crop yield varies temporally and spatially and that increases in crop yield are possible by targeting different amounts of an input, for example, irrigation water and fertilizer, to specific parts of the field. Second, crop management, from planting to harvesting, is complex, and simulation models can be used as a decision aid. Use of crop simulation models, developed in the 1970s–1980s, is facilitated due to the increased availability of required soil, elevation, and weather input data to execute the models and reduced cost of computer hardware. The third factor is an increased awareness of producers on production efficiency and environmental concerns. For example, in many agricultural areas, the amount of nitrogen fertilizer that can be applied is restricted and linked to the residual nitrogen in the soil and its potential effect on contamination of surface water and groundwater.[11,13,14,33,51] Advances in management information systems, development of computer software, and communication via the Internet provide us with the tools to manage a crop in real time.[1,6,36,52] We are currently working toward the development of a PA model that includes all of the above factors.[52,53]

The integration of a landscape-scale model such as PALMS[38] with a cotton growth model[39] using site-specific management of water and nitrogen[17,18] can give us the tools to manage, for example, a 50-ha irrigated cotton field. The model provides three key features important to real-time production-scale management: (1) it represents the variability in space and time within the entire field and accounts for hydrologic interactions between areas within the field; (2) it provides water, energy, nutrient, and carbon balance information without reliance on field-installed hardware that must be avoided during management operations; and (3) it can provide predictive information that can support various what-if scenario evaluations (something that physical field measurements do not provide). For example, this field can be divided into 5,000 (100 m^2) or 20,000 (25 m^2) cells, and the model will calculate a cotton lint yield value for each cell, using weather data collected at or near the field and previously collected soil and elevation data, both of which are stable and once developed can be used for many growing seasons. Further, the estimate of cotton lint yield is based on interactions of soil–plant–weather parameters, and the model itself can be used to explore what combination of inputs (e.g., water and nitrogen) would give the highest economic yield while minimizing leaching of nitrogen below the root zone. This is a current topic of research of our cropping system research unit.[52,53]

The input and output terms of the water balance of a cropping system are usually well quantified except for information on the input variable rain, that is, frequency and rate of rainfall events across the landscape.[53] Rainfall events are normally measured at a single point in space and seldom is the rate of individual rain events measured and/or recorded. Determining, the amount of water from a rain event that is stored in the soil, that is, effective rain, and runs off is key to correctly model the water balance across the landscape. For example, in the semiarid Texas High Plains, about 86% of annual rain events are <13 mm[54] and a large portion of this rain is lost as evaporation from the soil before it can be used by the crop as transpiration.[55] To correctly calculate the water balance across the landscape we need to determine the amount for any rain event that is stored in the soil and is used by the crop as transpiration.[55,56] Rainfall is variable in space and an appropriate cluster of rain gauges is needed across the landscape to measure effective rain and correctly determine the water balance across the landscape.

Conclusion

PA is a generic term that describes the way that agronomic inputs to a farming operation are managed and each input is applied according to specific needs across the field. An outcome of PA is the recognition that crop yields vary temporally and spatially. Crop management, from planting to harvest, is complex and requires management strategies that consider economic and environmental consequences. Recently, nutrient contamination and quality and safety of food delivered to consumers are also factors that have received consideration. A summary of general concepts learned from PA experiments follows.

First, agronomic experimental work needs to be done at the scale of application. For example, results from small research plots are normally not transferable to large fields.[22,25,40] The underlying principle is to take advantage of the inherent spatial variability of soil properties across the landscape. In PA experiments, the spatial variability of a given property is used as the source of variation instead of imposing treatments (e.g., levels of nitrogen fertilizer) to obtain variation. Second, the use of classical statistics fails to explain the cause and effect between variables. Spatial structure between variables can be quantified using specialized statistical tools such as autocorrelation, cross-correlation, and state-space analysis.[22,25] Third, given the complexity of current cropping systems, crop models[39,48] introduced in the 1980s are being combined with geographic information data of soils and elevation, real-time weather, and management information systems to provide a framework to manage crops. These models can provide three key features important to real-time production-scale management: (1) representation of variability in space and time within the entire field and hydrologic interactions between areas within the field; (2) water, energy, nutrient, and carbon balance information without reliance on field-installed hardware that must be avoided during management operations; and (3) predictive information to support various what-if scenario evaluations. As a result, we now have the capability of tracking different layers of data (soil, plant, and weather) and, for example, a 50-ha field may be divided into 5,000 cells, each cell 10 m wide, 10 m long, and 2.0 m deep.[36,37,52,53] The water, energy, carbon, and nitrogen balance (input equal to output) of each cell is calculated, and the change of any one variable on crop yield can be evaluated. This is a powerful management tool that will assist producers on how to maximize economic crop yield while minimizing nutrient losses to surface or groundwater.[52,53] Adoption of PA practices will continue to rise given the increased constraints of production, advances in technology, and demand for safe food and fiber of high quality.[13,33,46,47]

References

1. Kumar, S.A.; Ilango, P. The impact of wireless sensor network in the field of precision agriculture: A review. *Wireless Pers. Commun.* **2018**, *98*, 685–698.
2. Rogovska, N.; Laird, D.A.; Chiou, Chien-Ping; Bond, L.J. Development of field mobile soil nitrate sensor technology to facilitate precision fertilizer management. *Precis. Agric.* **2019**, *20*, 40–55.
3. Mulla, D.J. Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps. *Biosyst. Eng.* **2013**, *114*, 358–371.

4. Ferguson, R.; Rundquist, D. Remote sensing for site-specific crop management. In *Precision Agriculture Basics*; Shannon, D.K.; Clay, D.E.; Kitchen, N.R., Eds.; ASA-CSSA-SSSA: Madison, WI, **2018**, *8*, 103–117.
5. Sonka, S.T. Big data: Fueling the next evolution of agricultural innovation. *J. Innov. Manag.* **2016**, *4*(1), 114–136.
6. Fulton, J.P.; Port, K. 830. In *Precision Agriculture Basics*; Shannon, D.K.; Clay, D.E.; Kitchen, N.R., Eds. ASA-CSSA-SSSA: Madison, WI, **2018**, *12*, 169–188.
7. Mahan, J.R.; Lascano, R.J. Irrigation analysis based on long-term weather data. *Agriculture* **2016**, *6*, 42, doi:10.3390/agriculture6030042
8. Pongnumkul, S.; Chaovalit, P.; Surasvadi, N. Applications of smartphone-based sensors in agriculture: A systematic review of research. *J. Sensors* **2015**, Article ID 195308, 18, doi:10.1155/2015/195308
9. Mesas-Carrascosa, F.J.; Castillejo-González, I.L.; Sánchez de la Orden, M.; García-Ferrer, A. Real-time mobile phone application to support land policy. *Comput. Electron. Agric.* **2012**, *85*, 109–111.
10. Hedley, C. The role of precision agriculture for improved nutrient management on farms. *J. Sci. Food Agric.* **2015**, *95*, 12–19.
11. Abit, M.J.M.; Arnall, D.B.; Phillips, S.B. Environmental implications of precision agriculture. In *Precision Agriculture Basics*; Shannon, D.K.; Clay, D.E.; Kitchen, N.R., Eds.; ASA-CSSA-SSSA: Madison, WI, **2018**; *14*, 210–220.
12. Neupane, J.; Gao, W. Agronomic basis and strategies for precision water management: A review. *Agronomy* **2019**, *9*, 87, doi:10.3390/agronomy9020087
13. Gebbers, R.; Adamchuk, V.I. Precision agriculture and food security. *Science* **2010**, *327* (5967), 828–831.
14. Bongiovanni, R.; Lowenberg-DeBoer, J. Precision agriculture and sustainability. *Precis. Agric.* **2004**, *5*, 359–387.
15. Plant, R.E. Site-specific management: The application of information technology to crop production. *Comput. Electron. Agric.* **2001**, *30*, 9–29.
16. Monzon, J.P.; Calviño, P.A.; Sadras, V.O.; Zubiaurre, J.B., Andrade, F.H. Precision agriculture based on crop physiological principles improves whole-farm yield and profit: A case study. *Eur. J. Agron.* **2018**, *99*, 62–71.
17. Bronson, K.F.; Booker, J.D.; Bordovsky, J.P.; Keeling, J.W.; Wheeler, T.A.; Boman, R.K.; Parajulee, M.N.; Segarra, E.; Nichols, R.L. Site-specific irrigation and nitrogen management for cotton production in the Southern High Plains. *Agron. J.* **2006**, *98*, 212–219.
18. Li, H.; Lascano, R.J.; Booker, J.; Wilson, L.T.; Bronson, K.F. Cotton lint yield variability in a heterogeneous soil at a landscape scale. *Soil Tillage Res.* **2001**, *58*, 245–258.
19. Sinclair, T.R.; Tanner, C.B.; Bennett, J.M. Water-use efficiency in crop production. *BioScience* **1984**, *34*(1), 36–40.
20. Fixen, P.; Brentrup, F.; Bruulsema, T.W.; Garcia, F.; Norton, R.; Zingore, S. Nutrient/fertilizer use efficiency: Measurement, current situation and trends. In *Managing Water and Fertilizer for Sustainable Agricultural Intensification*; Drechsel, P.; Heffer, P.; Magen, H.; Mikkelsen, R.; Wichel, D. Eds.; International Fertilizer Industry Association, International Water Management Institute, International Plant Nutrition Institute, and International Potash Institute: Paris, France, First edition, **2015**, *2*, 8–38.
21. Sharma, B.; Molden, D.; Cook, S. Water use efficiency in agriculture: Measurement, current situation and trends. In *Managing Water and Fertilizer for Sustainable Agricultural Intensification*; Drechsel, P.; Heffer, P.; Magen, H.; Mikkelsen, R.; Wichel, D. Eds.; International Fertilizer Industry Association, International Water Management Institute, International Plant Nutrition Institute, and International Potash Institute: Paris, France, First edition, **2015**, *3*, 39–64.
22. Wendroth, O.; Al-Oman, A.M.; Kirda, C.; Reichardt, K.; Nielsen, D.R. State-space approach to spatial variability of crop yield. *Soil Sci. Soc. Am. J.* **1992**, *56*, 801–807.

23. Nielsen, D.R.; Wendroth, O.; Pierce, F.J. Emerging concepts for solving the enigma of precision farming research. In *Precision Agriculture, Proceedings of the Fourth International Conference*, Minneapolis, MN, July 19–22, 1998; Robert, P.C., Rust, R.H., Larson, W.E., Eds.; **1999**; 303–318.
24. Shumway, R.H.; Stoffer, D.S. *Time Series Analysis and Its Application*; Springer Verlag: New York, 2000; 549 pp.
25. Nielsen, D.R.; Wendroth, O. *Spatial and Temporal Statistics*. Catena Verlag: Reiskirchen, Germany, 2003; 398 pp.
26. Kho, R.M. On crop production and the balance of available resources. *Agric. Ecosyst. Environ.* **2000**, *80*, 71–85.
27. Booker, J.D.; Bronson, K.F.; Keeling, J.W.; Bordovsky, J.P.; Segarra, E.; Velandia-Parra, M. Farm-scale testing of site-specific irrigation and nitrogen fertilization for cotton production in the Southern High Plains of Texas, USA. In *Precision Agriculture '05*; Stafford, J.V., Ed.; Wageningen Academic Publishers: Wageningen, Netherlands, 2005; 951–957.
28. Van Looy, K.; Bouma, J.; Herbst, M.; Koestel, J.; Minasny, B.; Mishra, U.; Montzka, C.; Nemes, A.; Pachepsky, Y.A.; Padarian, J.; Schaap, M.G.; Tóth, B.; Verhoef, A.; Vanderborght, J.; Van der Ploeg, M.J.; Weihermüller, L.; Zacharias, S.; Zhang, Y.; Vereecken, H. Pedotransfer functions in earth system science: Challenges and Perspectives. *Rev. Geophys.* **2017**, *55*, 1199–1256.
29. Halvorson, G.A.; Doll, E.C. Topographic effects on spring wheat yield and water use. *Soil Sci. Soc. Am. J.* **1991**, *55*, 1680–1685.
30. Hanna, A.Y.; Harlan, P.W.; Lewis, D.T. Soil available water as influenced by landscape position and aspect. *Agron. J.* **1982**, *74*, 999–1004.
31. Timlin, D.J.; Pachepsky, Y.A.; Snyder, V.A.; Bryant, R.B. Spatial and temporal variability of corn grain yield on a hillslope. *Soil Sci. Soc. Am. J.* **1998**, *62*, 764–773.
32. Mueller, N.D.; Gerber, J.S.; Johnston, M.; Ray, D.K.; Ramankutty, N.; Foley, J.A. Closing yield gaps through nutrient and water management. *Nature* **2012**, *490*, 253–257.
33. Wu, W.; Ma, B. Integrated nutrient management (INM) for sustaining crop productivity and reducing environmental impact: A review. *Sci. Total Environ.* **2015**, *512–513*, 415–427.
34. Li, H.; Lascano, R.J. Deficit irrigation for enhancing sustainable water use: Comparison of cotton nitrogen uptake and prediction of lint yield in a multivariate autoregressive state-space model. *Environ. Exp. Bot.* **2011**, *71*, 224–231.
35. Lyle, W.M.; Bordovsky, J.P. Low energy precision application (LEPA) irrigation system. *Trans. ASAE.* **1981**, *24*, 1241–1245.
36. Booker, J.D.; Lascano, R.J.; Evett, S.R., Zartman, R.E. Evaluation of a landscape-scale approach to cotton modeling. *Agron. J.* 2014, *106*, 2263–22794.
37. Booker, J.D.; Lascano, R.J.; Molling, C.C.; Zartman, R.E.; Acosta-Martinez, V. Temporal and spatial simulation of production-scale irrigated cotton systems. *Precis. Agric.* **2015**, *16*, 630–653.
38. Molling, C.C.; Strikwerda, J.C.; Norman, J.M.; Rodgers C.A.; Wayne R.; Morgan, C.L.S.; Diak, G.R.; Mecikalski, J.R. Distributed runoff formulation designed for a precision agricultural landscape modeling system. *J. Am. Water Res. Assoc.* **2005**, *41*, 1289–1313.
39. McKinion, J.M.; Baker, D.N.; Whisler, F.D.; Lambert, J.R. Application of the GOSSYM/COMAX system to cotton crop management. *Agric. Syst.* **1989**, *31*, 55–65.
40. Cassel, D.K.; Wendroth, O.; Nielsen, D.R. Assessing spatial variability in an agricultural experiment station field: Opportunities arising from spatial dependence. *Agron. J.* **2000**, *92*(4), 706–714.
41. Wilson, J.N. Guidance of agricultural vehicles—A historical perspective. *Comput. Electron. Agric.* **2000**, *25*, 3–9.
42. Bell, T. Automatic tractor guidance using carrier-phase differential GPS. *Comput. Electron. Agric.* **2000**, *25*, 5366.
43. Bronson, K.F.; Keeling, J.W.; Booker, J.D.; Chua, T.T.; Wheeler, T.A.; Boman, R.K.; Lascano, R.J. Influence of landscape position, soil series, and phosphorus fertilizer on cotton lint yield. *Agron. J.* **2003**, *95*, 949–957.

44. Bronson, K.F.; Booker, J.D.; Officer, S.J.; Lascano, R.J.; Maas, S.J.; Searcy, S.W.; Booker, J. Apparent electrical conductivity, soil properties and spatial covariance in the U.S. Southern High Plains. *Precis. Agric.* **2005**, *6*, 297–311.
45. Stafford, J.V. Implementing precision agriculture in the 21st century. *J. Agric. Eng. Res.* **2000**, *76*, 267–275.
46. Zhang, N.; Wang, M.; Wang, N. Precision agriculture—A worldwide overview. *Comput. Electron. Agric.* **2002**, *36*, 113–132.
47. McBratney, A.; Whelan, B.; Ancev, T.; Bouma, J. Future directions of precision agriculture. *Precis. Agric.* **2005**, *6*, 7–23.
48. Bouman, B.A.M.; Van Keulen, H.; Van Laar, H.H., Rabbinge, R. The "School of de Wit' crop growth simulation models: A pedigree and historical overview. *Agric. Syst.* **1996**, *52*(2/3), 171–1793.
49. Feddes, R.A.; Kowalik, P.J.; Zaradny, H. *Simulation of Field Water Use and Crop Yield*. Simulation Monographs; Centre for Agricultural Publishing and Documentation: Wageningen, Netherlands, 1978; 188 pp.
50. Staggenborg, S.A.; Lascano, R.J.; Krieg, D.R. Determining cotton water use in a semiarid climate with the GOSSYM cotton simulation model. *Agron. J.* **1996**, *88*, 740–745.
51. Bronson, K.F.; Malapati, A.; Booker, J.D.; Scanlon, B.R.; Hudnall, W.M.; Schubert, A.M. Residual soil nitrate in irrigated southern High Plains cotton fields and Ogallala groundwater nitrate. *J. Soil Water Conserv.* **2009**, *64*, 98–104.
52. Nelson, J.R.; Lascano, R.J.; Booker, J.D.; Zartman, R.E.; Goebel, T.S. Evaluation of the precision agricultural landscape modeling system (PALMS) in the semiarid Texas Southern High Plains. *Open J. Soil Sci.* **2013**, *3*, 169–181.
53. Lascano, R.J.; Nelson, J.R. Circular planting to enhance rainfall capture in dryland cropping systems at a landscape scale: Measurement and simulation. *Adv. Agric. Syst. Models* **2014**, *5*, 85–111.
54. Jones, O.R.; Eck, H.V.; Smith, S.J.; Coleman, G.A.; Hauser, V.L. Runoff, soil, and nutrient losses from rangeland and dry-farmed cropland in the Southern High Plains. *J. Soil Water Conserv.* **1985**, 40(1), 161–164.
55. Kool, D.; Agam, N.; Lazarovitch, N.; Heitman, J.L.; Sauer, T.J.; Ben-Gal, A. A review of approaches for evapotranspiration partitioning. *Agric. Forest Meteorol.* **2014**, *184*, 56–70.
56. Goebel, T.S.; Lascano, R.J. Rainwater use by cotton under subsurface drip and center pivot irrigation. *Agric. Water Manage.* **2019** *215*, 1–7.

VII

PRO: Basic Environmental Processes

44

Green Processes and Projects: Systems Analysis

Abhishek Tiwary

Introduction

Given the level of inadvertent impacts generated from human breakthroughs in science and technology post industrial revolution, be it from the combustion engines based on coal/fossil fuel or the use of plastic in modern day commodities, it is imperative to develop a systems perspective on the sustainability of a process or project. This chapter introduces a holistic approach to environmental appraisal for minimizing the impacts from transboundary interactions between the material flows of a system, thereby ensuring efficient management of any potential environmental impacts. This approach combines application of "dynamic science-based" geo-spatial analysis with "static inventory-based" life cycle assessment (LCA) to understand the full-scale of impacts generated from a proposed process or project at the systems level, thereby ascertaining their true "greening" potential. Hereafter a "process" implies a set of activities targeting a specific output (e.g., 1 MW electricity, 1 ton of a material), whereas a "project" implies a combination of cross-cutting processes to achieve a specific objective (e.g., constructing a new building, greening of transport network). Two case studies are presented, one each for a green process and a green project, to enable the readers to get a grasp of the usefulness of this tool in global environmental management, more so in the context of supply chain spread diffusively over numerous geographical regions.

Systems Approach in Environmental Management

In the context of fool-proofing human existence from untoward climate change feedbacks, management of both built-up and natural environments has gained center stage in recent years. It is widely agreed now that effective environmental management involves striking a fine balance between the technological innovations underpinning its viability (essential for economic growth) and the constraints to long-term sustainability (Figure 1). This chapter deals with how a systems approach can be applied to achieve this goal.

Traditionally, systems approach has its roots in industrial (or laboratory) processes that simply comprise a series of related activities aimed at optimizing production (and hence the profit). However, in the face of climate change, much of the environmental research conducted over the last decade has changed our notion of the scale of management problem we face. Take, for instance, the case of cheap biofuel provisioning in the first decade of the 21st century—the demand for palm oil in the Western world has caused massive deforestation in Indonesia and Amazonia. Having a wider scope to environmental management, beyond what can be considered as local activities within a country's borders, has given way to the systems approach in environmental management. Although simple in principle and rigorously tested in processing industries, its adaptation to the wider environmental problems comes with extreme operational challenges. The latter is mainly highlighted in terms of data availability, essential to the success of this vision. Despite this, it is considered as an efficient framework to facilitate a more holistic management of material, energy, and pollution across a range of related activities, usually spread over large geographical areas.

The main advantage of using a systems approach is to figure out the hotspots of environmental concerns (greenhouse gas emissions, ecotoxicity, air pollution, etc.) within a system. For this purpose, the scope of environmental management has now crossed local boundaries to a systems scale, encompassing the whole range of supply-chain spreading over several countries to ensure global sustainability. These associated activities form what is known as a "life cycle" in terms of environmental appraisal. It facilitates mapping of the "stocks" and "flows" of emissions (also called environmental burdens) through different stages. The concept has been used to develop the LCA and Material Flow Analysis (MFA) methods. Both these techniques have proven to be useful diagnostic tools. Detailed texts describing the LCA/MFA methods exist in the literature. Interested readers can find some useful sources listed in different parts of this chapter; in particular, *The Hitch Hiker's Guide to LCA*[1] serves as a good starting point to develop fundamental understanding of these concepts.

The main objective of applying LCA/MFA is to assess the critical environmental burdens and impacts contributing to the adverse effects on human and ecological health. These critical pollutants can then

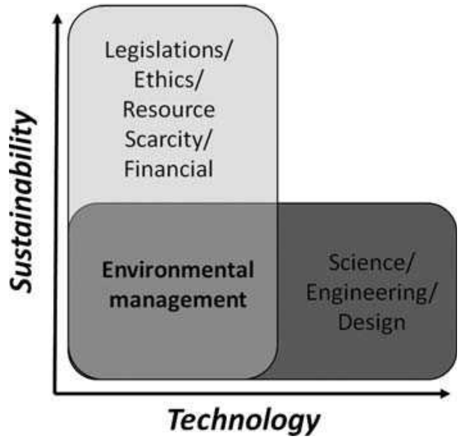


FIGURE 1 Constraints to environmental management at a systems scale.

be dealt with, and their impacts mitigated, through implementation of effective control measures. This ensures a solution to an environmental problem at the local scale while considering its footprints at the global scale, literally over the whole life cycle. As a consequence, this approach offers a robust framework for implementing holistic sustainability.

Best Practice Guidelines in Systems Analysis

Standard Protocols

In essence, the International Organization for Standardization ISO 14044[2] specifies four mandatory steps for quality assurance in LCA, namely—goal and scope definition, inventory analysis, impact assessment, and interpretation. To ensure that the environment as a whole is protected, the process chain forming the system can be scoped in two ways. The first comprises pre-chains involved in excavation of the resources and their transportation to the industrial site, the end-of-pipe emissions from the plants, and the disposal of the wastes at the end of the cycle. This approach in life cycle thinking is usually referred to as “cradle-to-gate” since it follows an activity from the extraction of raw materials (i.e., cradle) to the delivery of the product or service (i.e., the exit gate). The other, in addition to accounting for all the above, also includes one-off construction and demolition of the infrastructures or end-of-life processing and disposal of the equipment/commodities used. Appropriately so, it is then called a “cradle-to-grave” system.

As a standard practice, the resources utilized and the emissions added to the environment in all these activities are usually modeled on a unit basis, typically annual turnover for the industrial products and annual usage for the services. This is termed as the “functional unit” of analysis in LCA. Typical functional unit for an assessment can be “activities over 1 year,” representative of the quantitative metric of the output of products or services that the system is expected to deliver. Likewise, the typical timescale for an assessment could range from hours to 100 years. The latter is specifically relevant for assessing long-term environmental impacts such as global warming potential and acidification.

The spatial context of setting up a life cycle model is also essential in order to establish its system boundaries. Typical spatial scales range from local (e.g., urban) to wider environment (e.g., remote excavation site, mines) and possibly representing a global scale. This aspect of the analysis is useful for consistency checks, comparing different LCA results, and ensuring they have been conducted on comparable system boundaries. Given that life cycle approach facilitates assessment of a full spectrum of process chains involved in a process for all the activities, it is recommended to group the burdens and impacts obtained from an analysis into two distinct categories: (1) arising in the direct environment from the main activity/ies under investigation (commonly known as “foreground”) and (2) arising from a whole series of linked pre- and post-chain activities in the wider global environment (known as “background”). This is clearly shown in Figure 2, which shows a series of activities associated with the production of electricity from biofuel through a schematic representing a power generation system. In this figure, the main activity (i.e., the cogeneration plant) is shown as the foreground (shaded region), whereas the peripheral activities, associated primarily with the sourcing of the raw materials, shipment, and final disposal are shown in the background. On the one hand, this approach enables clear accounting of the environmental impacts and, on the other, it offers insight into the hotspots at each stage of process chain, facilitating effective management of the problem through visualizing the entire system at the same time. This is considered superior to resource- and cost-intensive piecemeal solutions encountered in traditional management approaches, specifically so in the context of achieving global sustainability.

The environmental impacts from LCA are usually calculated from mass balances of the input/output flows on the basis of the problem-oriented (midpoint) approach. It is a baseline method that provides a list of impact assessment categories grouped into obligatory impacts, used in most LCAs. A baseline indicator is considered suitable for simplified studies since it utilizes the principle of best available practice when several methods for obligatory impact categories are available.

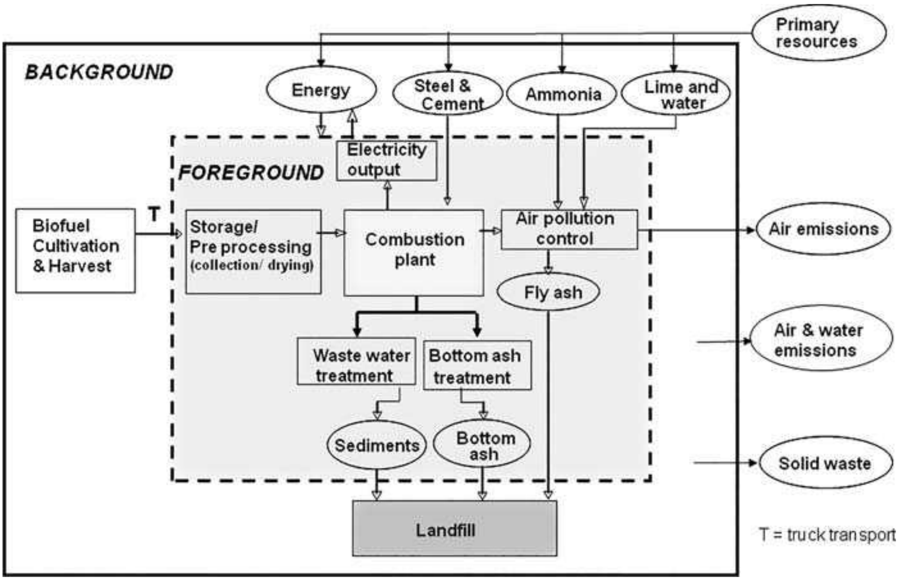


FIGURE 2 Split between foreground and background activities from a systems perspective.
Source: Tiwary and Colls.[3]

Uncertainty Assessment

To decide on the future courses of action based purely on results of LCA analysis, one needs to account for the uncertainty in these results—this enables reaching better decisions. The information on uncertainty in LCA is usually contained in the assessment of probability of realistic representativeness of the results. This is expressed in terms of the confidence bounds on LCA results, which illustrate the region within which the true values have an estimated likelihood of falling.

This section covers only the fundamentals of uncertainty inherent in LCA. Readers will be guided to use dedicated literature to pursue a more in-depth uncertainty assessment exercise.

A number of approaches have been suggested in the literature for the integrated consideration of both technical and valuation uncertainties involved in LCA. The former is associated with the uncertainties in data collection, while the latter is inherent to the impact assessment method used. The reason we need to care about uncertainty in LCA is because the statements or assertions we tend to make about the world on the basis of direct LCA outputs may be wrong. These errors have been mainly associated with either uncertainty or variability in the outputs. Whereas variability is considered to be inherent in the real world, uncertainty is mainly associated with inaccurate measurements, lack of data, and/or model assumptions.[4] The following can be considered as a rigorous (although not an exhaustive) list of categories of uncertainties identified in the literature:

- *Parameter uncertainty*: This is due to the uncertainty within the large number of parameters used in LCA models. It leads to uncertainty in the final output from the LCA exercise. Empirical inaccuracy, non-representativeness (incomplete or outdated measurements), and lack of data are common sources of parameter uncertainty.
- *Epistemological uncertainty*: The use of the information in databases for life cycle inventory (LCI) introduces epistemological uncertainty since the system where the data is to be used to model a process may differ from the system where the data was generated.[5]
- *Model uncertainty*: In situations where an LCA model suffers from uncertainty in the underlying model assumptions and the basic model calculations (departure from the default linear

programming), the results of a parameter uncertainty analysis can be misleading and hence provide no useful information.

- *Spatial variability*: Application of spatially averaged data to model specific processes in certain parts of the world leads to these discrepancies.
- *Temporal variability*: These are mainly attributed to time-dependent variations in emissions and other technical process characteristics.

Handling Uncertainty in Systems Analysis

It has been suggested that not all the methods of statistical analyses of uncertainty can be applied to LCA, primarily because the underlying LCA data are not based on random samples, that is, we are not strictly dealing with random variables that follow some known frequency distribution. However, one can apply subjective probability distributions to quantify the uncertainties involved. The established procedure to achieve this consists of the following five steps: selection of essential parameters, determination of probability distributions, Monte Carlo simulation, significance analysis, and interpretation of the results.[4,6]

Recent approaches have tried to integrate consideration of both technical and valuation uncertainties during decision making on the basis of the results provided by LCA. These are to be used in conjunction with established decision support tools based on multiple criteria decision analysis since it has been demonstrated that the structure of LCA has parallels with a decision analysis approach to decision making.[7] Key elements of this approach include "distinguish ability analysis" to determine whether the uncertainty in the performance information is likely to make it impossible to distinguish between the activities under consideration, and the use of a multivariate statistical analysis approach, such as principal component analysis. The latter enables rapid analysis of large numbers of parallel sets of results, thereby allowing for the identification of choices (options) that lead to similar and/or opposite evaluation of activities.

Application of Subjective Probability Distributions Approach

This method of uncertainty analysis in LCA has inherent attributes since they are an alternative to simple point estimates, which allow for the use of a range within which we expect the true value to lie rather than using a single number to estimate the results of some real-world quantity (e.g., tons of sheet of steel required to make 1 ton of an output product). In this manner, by developing subjective probability distribution, the methods of uncertainty analysis can be used in LCA to account for the uncertainty in the results. This further enables accounting for the reliability of the decisions reached on the basis of the LCA outcomes, and if the reliability of the conclusions is not sufficient for our decision-making needs, then uncertainty analysis helps identify which data uncertainties are most significant, that is, influential in the process chain. Furthermore, applying the inverse method, this step can help the decision maker determine the levels of reduction in data uncertainty required to reach a specific level of confidence in the results. For example, to attain "90% confidence" in the results, one would need to assure that using the subjective probability distribution approach if the LCA analysis of a process is repeated several times, each time using new and equally probable randomly selected values for the uncertain input quantities, the conclusions would be correct 90% of the time.

Uncertainty Analysis versus Data Quality Characterization

The LCA is very dependent on data of good quality. In most cases, LCA practitioners rely on generic databases provided by different sources. The commonly available databases to date are the European Union's (EU) European Reference Life Cycle Data System (ELCD), Swiss National LCI Database (ecoinvent), the U.S. LCI Database created by the National Renewable Energy Laboratory and its partners, the Canadian Raw Materials Database (CRMD), the Swedish National LCA Database (SPINE@ CPM), the Danish Food Database, and the Korean National LCI Database (KNCPC). Apart from these, currently Australia and Japan also have ongoing initiatives to generate national LCA databases.

Owing to a limited number of sources providing information while generating the LCI data, quite often the uncertainty related to the amount of a specific input or output cannot be derived from the available information. For such circumstances Frischknecht et al.[8] have developed a simplified standard procedure to quantify the uncertainty. This simplified approach includes a qualitative assessment of data quality indicators on the basis of a pedigree matrix from published literature. Basic uncertainty factors are used for the kind of input and output considered. For example, it is assumed that CO_2 emissions generally show a much lower uncertainty as compared with CO emissions since the former is usually calculated from fuel input, whereas the latter depends on numerous parameters such as boiler characteristics, engine maintenance, and load factors. Table 1 provides a list of proposed uncertainty factors in a pedigree matrix, which have been based on expert judgments.

TABLE 1 Examples of Basic Uncertainty Factors (Dimensionless) Applied for Technosphere Inputs and Outputs and for Elementary Flows

Input/Output Group	c	p	a
Demand of			
Thermal energy, electricity, semifinished products,	1.05	1.05	1.05
working material, waste treatment services	2.00	2.00	2.00
Transport services (t km)	3.00	3.00	3.00
Infrastructure			
Resources			
Primary energy carriers, metals, salts	1.05	1.05	1.05
Land use, occupation	1.50	1.50	1.50
Land use, transformation	2.00	2.00	2.00
Pollutants Emitted to Air			
CO_2	1.05	1.05	
SO_2	1.05		
NMVOC total	1.50		
NO_x, N_2O	1.50		1.40
CH_4, NH_3	1.50		1.20
Individual hydrocarbons	1.50	2.00	
PM>10 micron	1.50	1.50	
PM_{10}	2.00	2.00	
$PM_{2.5}$	3.00	3.00	
Polycyclic aromatic hydrocarbons (PAH)	3.00		
CO, heavy metals	5.00		
Inorganic emissions, others		1.50	
Radionuclides (e.g., radon-222)		3.00	
Pollutants emitted to water			
BOD, COD, DOC, TOC, inorganic compounds		1.50	
(NH_4, PO_4, NO_3, Cl, Na)		3.00	
Individual hydrocarbons, PAH		5.00	1.80
Heavy metals			1.50
Pesticides			1.50
NO_3, PO_4			
Pollutants emitted to soil			
Oil, hydrocarbon total		1.50	
Heavy metals		1.50	1.50
Pesticides			1.20

Source: Data from Frischknecht et al.[8].

c, combustion emissions; p, process emissions; a, agricultural emissions.

Case Study I—System Analysis of Biofuel-Based Electricity Generation Process

This example demonstrates the application of a systems approach in assessing the merits of biofuels as a green process. The LCA generates a profile of all the steps involved in the process chain—both onsite and upstream—to provide quantitative information on potential impacts of an industrial activity (i.e., a system). Usually, this is done in terms of the released emissions (i.e., burdens), using a linear model. It takes into account the emission factors of all the known pollutants, inventoried during controlled analysis following standard monitoring protocols. These are then scaled by the volume of the industrial activity to provide their actual burdens. Also, in the case that LCA data are available for multiple similar installations, their respective operational performances can be benchmarked and links between operational efficiency and environmental impacts can be established.[9] A number of LCA studies available in the literature compare the environmental impacts of energy production (heat and/or electricity) from co-firing different biofuels in an existing coal-fired power plant, showing their overall greenhouse gas benefits. These cover the use of agricultural residues such as straw and residual wood, short rotation coppiced (SRC) wood,[10] and perennial rhizomatous grasses.[11,12] A more recent study quantifies the airborne emissions from different biomass-based electricity production systems using different technologies, feedstocks, and scales in order to establish the extent to which offsite emissions may contribute to overall environmental impact.[13]

In LCA, the emission factors are adjusted according to the pollution abatement technologies used in the industry to compensate for the release of acidic gases, although they do not reflect the fate of the emissions once they are out of the stack. Thus, this approach remains capable of providing realistic emission scenarios as long as the fate of the released emissions is not altered significantly by the dispersion and chemical transformation in the surrounding media (air, water, or soil). Whereas it allows successful prediction of greenhouse gas emissions in a fairly straightforward manner (assuming minimal phase alterations), the modeling of the gas–particle interactions leading to quantification of total particulate matter (PM) loading is far from complete. For example, the PM emissions calculated in the LCA from a biofuel combustion plant represent mainly the dust emissions from fly ash.[14] Any additional aerosols generated from gas-phase interactions of the resulting emissions, either during biofuel cultivation or combustion, would not be adequately quantified within this approach.

This case study utilizes the power of LCA as a diagnostic tool to track the pollutants and carbon emissions over the whole life cycle and mainly focuses on feasible management options for mitigating secondary aerosol generation potential from photochemical neutralization of the acidic emissions with ammonia using the following precursor chemistry[15]:

$$NH_{3(g)} + HCl_{(g)} \leftrightarrow NH^{+}_{4(aq)} + Cl^{-}_{(aq)}$$

A cradle-to-gate system is applied to all the energy systems modeled in this study, accounting for all the relevant flows involved in the extraction of resources (renewable/non-renewable) and cultivation of biofuels leading up to production of the required amount of energy outputs. As shown in Figure 2, the system comprises a series of foreground and background activities. The foreground activities are considered as the focal point of the system. Shown in the central part as shaded region in the figure, it involves storage and preprocessing of the fuel and its combustion to produce electricity. The background activities primarily include the cultivation and harvesting of the biofuel, production of required chemicals, and their transport to the power plant. The other end of the process chain accounts for the disposal of wastes generated. The atmospheric emissions from all these stages have been accounted for in the models.

Electricity production from different renewable biofuel sources has been modeled using scenario analysis. Representative biofuel options currently feasible have been considered. The analysis in this study is based on the functional unit defined as "1 terajoule (TJ) electricity produced from biomass and delivered to the grid." It is achieved by using available LCI data for a 50-MW electricity steam turbine/back pressure cogeneration plant firing solid biomass.[16] It is important to note that, in this study, the modeled system has not been credited for the cogenerated heat in the process. The following five different electricity production scenarios, representative of the technology for 2010 (base scenarios), have been considered, each using different types of biofuels:

- *Scenario A*: Perennial rhizomatous grass (*Miscanthus giganteas*).
- *Scenario B*: SRC chips.
- *Scenario C*: Residual/waste wood.
- *Scenario D*: SRC chips–grass blend (by energy); SRC chips (80%) and perennial rhizomatous grass (20%).
- *Scenario E*: Waste wood–grass blend (by energy); waste wood (80%) and perennial rhizomatous grass (20%).

In scenarios D and E, a fuel mix of perennial grass and wood in 1:4 ratio (by energy) has been considered. This is meant to improve the combustion quality and minimize excessive atmospheric emissions. In all the base scenarios, the biofuels have been assumed to be sourced locally (50 km distance to the combustion plant) and transported from their origin to the energy production site using 40-ton payload trucks. This is mainly aimed to investigate potential local air quality degradation from interactions of emissions from the power plant and the cultivation sites.

The atmospheric burdens for almost all the base scenarios (A–E) are dominated by pollutant emissions from the power plant (Figure 3). Overall, there is no clear winner among the scenarios. For example, scenarios C and E (waste wood) are best for CO_2, NH_3, non-methylated volatile organic carbons (NMVOCs), and CH_4, and scenario B (SRC chips) has the lowest HCl emissions. Scenario A (miscanthus) is best for N_2O and PM; however, it is worst for HCl and SO_2. For both miscanthus and SRC wood (scenarios A, B, and D), only biomass cultivation is considered to be the main source of NH_3 release to the local environment. However, it can be noted that a considerable amount of CO_2 is emitted during biomass production, transport, storage, and drying processes for these scenarios, which are mainly associated with the use of fossil-based energy sources in these stages. It has been reported that the forest logging operations involved in harvesting SRC wood contribute to significant atmospheric emissions

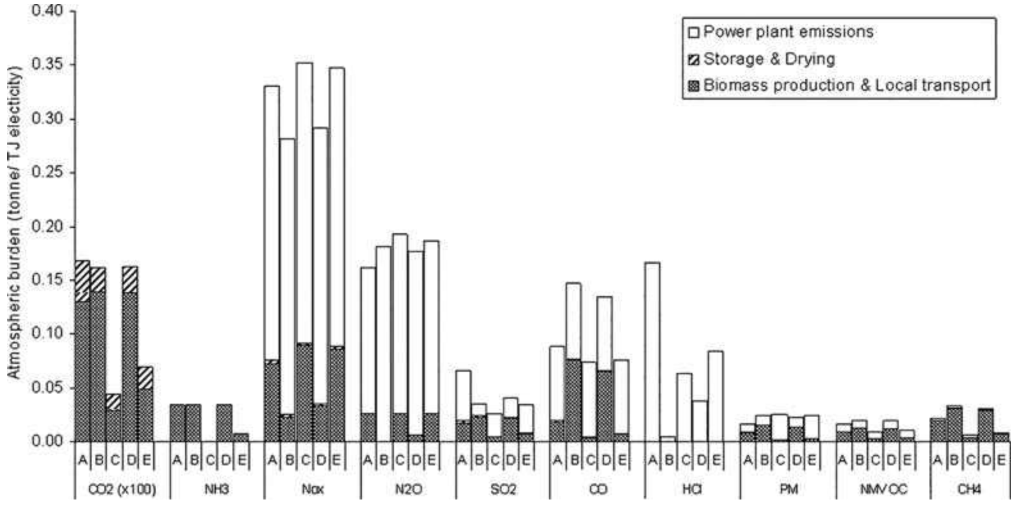


FIGURE 3 Environmental burdens of atmospheric emissions for base scenarios A–E.
Source: Tiwary and Colls.[3]

of CO_2, CO, NMVOCs, and PM due to fuel, chainsaw, and hydraulic oil consumption by heavy-duty diesel engine vehicles (Athanassiadis, 2000). This is reflected for the base scenarios B and D whose emissions of these pollutants from the "biomass production and local transport" stage are much higher compared with the rest of the scenarios. In case of waste wood, no mechanical chipping was assumed to be involved and hence the corresponding emissions from scenarios C and E have been relatively much lower. However, the power plant loadings of NO_x, SO_2, HCl, and PM in these two scenarios are found to be much higher than in scenarios B and D. This could be due to their incomplete combustion, which has been reported in earlier studies to result in highly variable emissions.[17,18]

Emissions of CO and NO_x per terajoule electricity output from power plant alone show comparable values for all the base scenarios but on a life cycle basis, that is, including the biofuel sourcing and storage stages, waste wood (scenarios C and E) seems to have up to 25% higher NO_x burdens, whereas SRC wood (scenarios B and D) seems to have up to 100% higher CO burdens. The corresponding emissions of SO_2 and HCl from miscanthus plant (scenario A) are higher by as much as 120% and 350%, respectively, compared with the rest of the scenarios. Therefore, scenario A poses the maximum likelihood of secondary aerosol generation potential through interactions of the acidic gas emissions from the power plant with the ammonia released from nearby harvest fields. On the other hand, SRC wood combustion plant (scenario B) has much lower HCl emissions (<5 kg T/J electricity) and, therefore, despite having considerable contributions to NH_3 release from the harvesting stage, its secondary aerosol generation potential is over 35-fold lower. Results indicate that co-firing waste wood sourced locally (scenarios C and E) would have the least secondary aerosol generation potential from photochemical neutralization, despite having higher HCl loadings than SRC wood, as it shows negligible NH_3 emissions over its fuel life cycle. Nevertheless, this is balanced by higher direct PM releases and life cycle NO_x releases from waste wood combustion. Although in this study heavy metals have not been included in the analysis owing to limited information, separate studies have reported wood combustion to be enhancing emissions of heavy metals.[19,20] The latter highlights additional air quality problems associated with large-scale implementation of wood-burning installations in the future.

The LCA model offers the flexibility to assess reduction in aerosol formation from different mitigation options, both for material (and cost) and energy, which is an invaluable screening exercise in order to come up with a viable and cost-effective management solution. For example, in this case study, the following three mitigation measures have been assessed.

Biomass Gasification and Direct Firing

Gasification technology is considered to provide efficient and clean power generation from biomass.[21] In this study, gasification of three different types of biomass (miscanthus, SRC, and waste wood) through air injection into the fuel in a circulating fluidized bed has been considered. The synthesis gas generated is compressed and injected into the combustion chamber of a cogeneration plant to produce steam that subsequently drives a turbine. As in the base scenarios, the system has not been credited for the cogenerated heat in the process for consistency.

Relative changes in the environmental burdens from application of biomass integrated gasification/combined cycle technology have been estimated with reference to the power plant emissions from the base scenarios (Figure 4). Results suggest significant reductions (up to 100% in some cases) in the emissions of acidic gases in the direct vicinity of the power plant. This is mainly due to much lower release of HCl and SO_2 from the plant for all scenarios considered. Likewise, PM emission in the foreground is also reduced as a consequence of gasification. In addition, N_2O emission is also lowered in all cases, whereas little change is observed in NO_x emission, except for a prominent decrease in the case of residual waste wood combustion (scenario C). However, while achieving the significant reduction in acidic gas burdens, this approach apparently increases end-of-pipe emissions of CH_4 and CO. Both these pollutants are produced during gasification of the biofuel, and their atmospheric emissions are associated with unintended (fugitive) release from the gasifier plant.[22] The main consideration for no obvious relative change (%) in

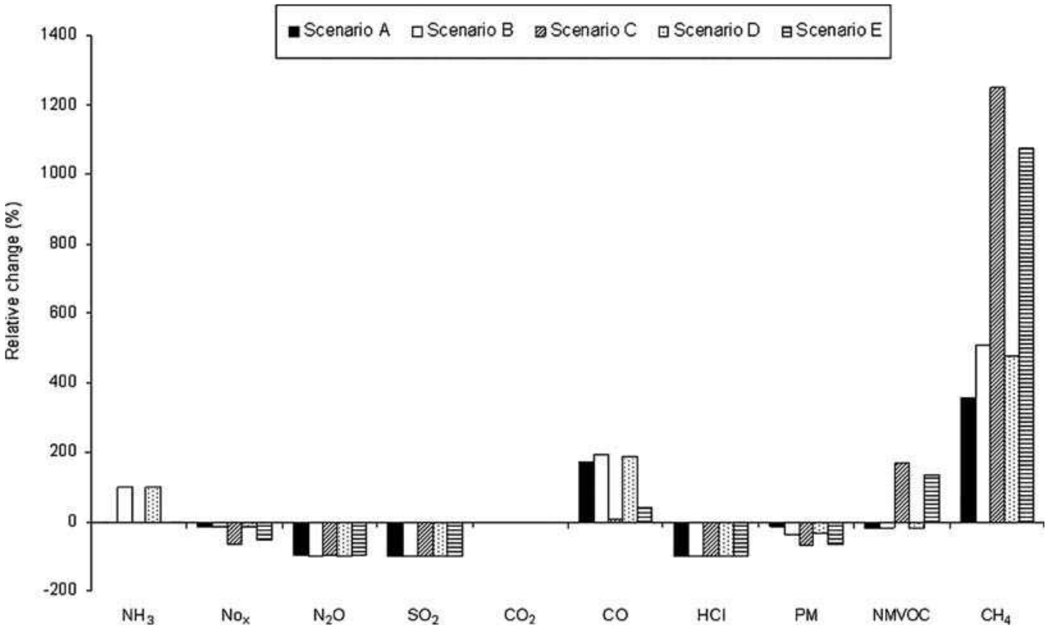


FIGURE 4 Relative change in environmental burdens by application of gasification option to base scenarios A–E.
Source: Tiwary and Colls.[3]

CO_2 emission in gasification options relative to base case is due to offsetting of the emissions enhancement from processing of additional feedstock with the increased efficiency in combustion of the gasified fuel. The increase in CH_4 emissions for scenarios A, B, and D ranges between 200% and 500%; however, for scenarios C and E, it appears to be exceptionally high (up to 1,250%). The latter two scenarios also show increased emissions of NMVOCs. As both CH_4 and NMVOCs contribute to global warming, the two scenarios (C and E) do not seem to present sustainable alternatives. Excessive release of methane could be potentially minimized by optimizing the gasification temperature and circulating the fugitive methane into the combustion process.[22] It is also noticed that power production from SRC wood biomass from this approach leads to production of NH_3, which makes scenarios B and D unviable too. Therefore, although gasification for scenario A enhances its CO emissions (up to 200% increase) relative to the corresponding base scenario, it presents the best option for reducing secondary aerosol generation potential.

Delaying the Harvest of Perennial Energy Crop

Delaying the harvest of miscanthus from late autumn (current common practice) to early spring has been reported to foster prolonged soil–vegetation interactions by allowing the stems to dry during the winter months. In effect, it allows for leaching of a substantial portion of the ash, chlorine, and potassium contents from the biomass through roots,[12] thereby reducing the adverse environmental impacts later during combustion. Delaying the harvest also allows for extended recycling of nutrients by their rhizome systems, which enables them to have low demand for nutrient inputs, especially for nitrogen. This lowers their fertilization needs. However, it has been reported that this delaying approach inadvertently leads to substantial yield losses (of up to 35%) and reduction in calorific value of the biofuel. The latter entails additional emissions from extra biomass combustion for the same energy throughput.

This study used the data provided by recent studies on reduction in dry matter yield, water content, and concentrations of specific toxic components, including ash, nitrogen, chlorine, and sulfur, as a consequence of delaying the harvest.[12] It mainly focused on assessing the influence of delaying the harvest of the biomass and its linked effects on storage and drying, whereas additional effects of topography and soil properties for differences in site locations could not be quantified at this stage. The ideal late harvest is recommended at a water content of 20% or less, as this minimizes the cost of harvesting and drying while keeping a high biomass quality of the crop. For the purpose of estimating the environmental burdens from drying of harvested biomass during storage, we have used literature data (Table 2).

Although the emission per unit biomass consumption is reduced in this manner, the resulting biofuel has deteriorated in calorific value. The requirement of additional biomass supply to compensate the lost calorific value of the fuel leads to considerable increase in the life cycle environmental burdens for NH_3, N_2O, and CO and slight increments for SO_2, PM, NMVOCs, and CH_4 (Figure 5). Delaying the harvest leads to reduction in the water content of the biomass. Interestingly, this allows for mitigating CO_2 emissions at two stages of the fuel life cycle: (1) reduced demands for fossil fuel in the storage and drying process and (2) increased energy conversion efficiency from enhanced combustion in the power plant. However, despite offering some reduction in emissions of CO_2, NO_x, and HCl, this technique does not seem to provide a definitive and reliable option to mitigate the secondary aerosol generation potential from chemical interactions of the precursor gases.

TABLE 2 Model Parameters for Technical Drying of Energy Crop in Storage

Harvest Type	Energy Use in Drying	Water Content
Autumn harvest (>40 wt.%)[a]	68 kWh heating oil + 9.7 kWh electricity per ton of water removal	15 wt.%[b]
Early winter harvest (20–40 wt.%)[a]	144 kWh electricity per ton of water removal	15 wt.%[b]
Late winter/early spring harvest (<20 wt.%)[a]	No drying is carried out	<20 wt.%[b]

Source: Data from Tiwary and Colls.[3]

[a] Water content of freshly harvested biomass.

[b] Water content of biomass before combustion.

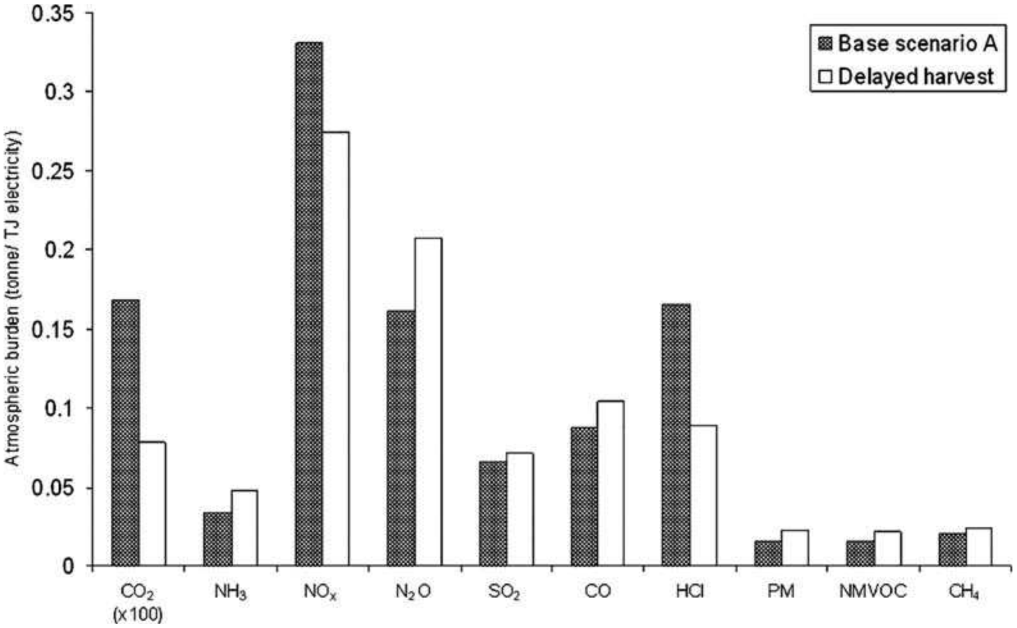


FIGURE 5 Comparison of the life cycle environmental burdens for base scenario A and delayed harvest of miscanthus.
Source: Tiwary and Colls.[3]

Increasing the Separation of Biomass Plant from Harvest Site (Importing the Biofuel)

As preliminary results from base scenarios (as shown in Figure 3) suggest, ammonia emissions from harvesting miscanthus and SRC wood are high. Assuming that the harvest sites in the new locations would be remotely rural with low acidic releases to the ambient air, a third possible mitigation option of reducing the impact of ammonia by displacing the cultivation sites has been considered. This option may also become necessary in the case of an eventual shift to larger production and use of biomass, as the EU is going to need large areas of agricultural land for its production.[23] In the United Kingdom alone, an estimated 125,000 ha of energy cropping would be required by 2010 to meet the government's target for electricity generation from biomass (DEFRA, 2001). In such situations, importing biomass from other European countries that offer favorable climatic conditions, mainly higher solar radiation and soil water contents, has been recommended as a more sustainable option.[24]

As considering all probable transportation means to move biofuels from one country to another gives rise to numerous possible scenarios,[25] some practical assumptions, based on the geographical location of the electricity generation site (here assumed to be in South England), have been made. Each option includes transport from production sites to a central gathering point, followed by a combination of international transport and local transport to a storage location near the generation plant. It is noteworthy that this assessment does not account for the additional emissions from the storage requirements of the biofuel crops, either in the country of their origin or during their long-haul transport.

As shown in Figure 6, the corresponding transport-related emissions for all the pollutants considered from this transport chain are much higher, which makes it less favorable from environmental perspectives. To a larger extent, this is solely dependent on the assumption used in this study to source miscanthus from land-locked Czech Republic and transport it overland across the stretch of European mainland on a freight train to the nearest ferry port. Had the biofuel been sourced from a coastal country and shipped directly to a port, the emission regimes would be very different. Therefore, one message that comes out of this study is that additional emissions from transporting the biofuel are very much site dependent, and this can be optimized.

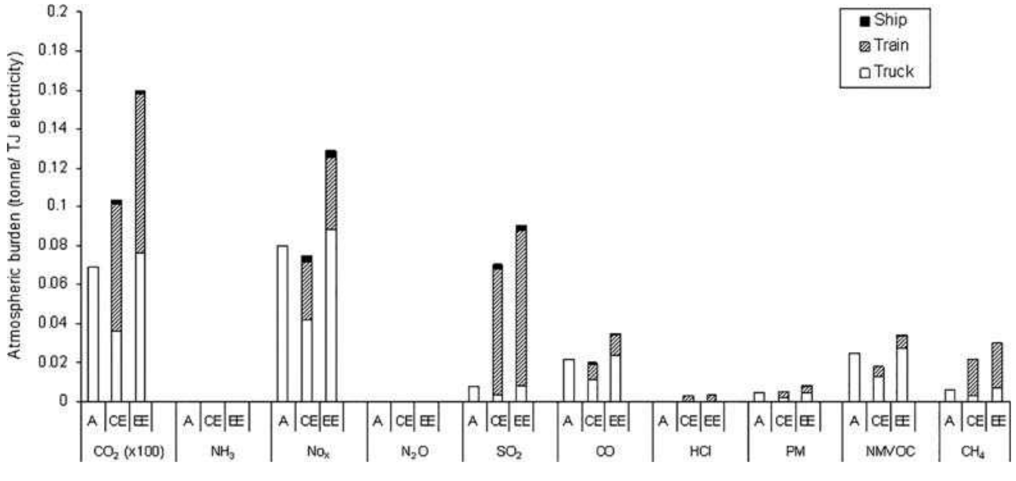


FIGURE 6 Transport-related burdens for different options associated with the transport of miscanthus (A—miscanthus in base scenario, CE—miscanthus imported from Central Europe, EE—miscanthus imported from Eastern Europe).

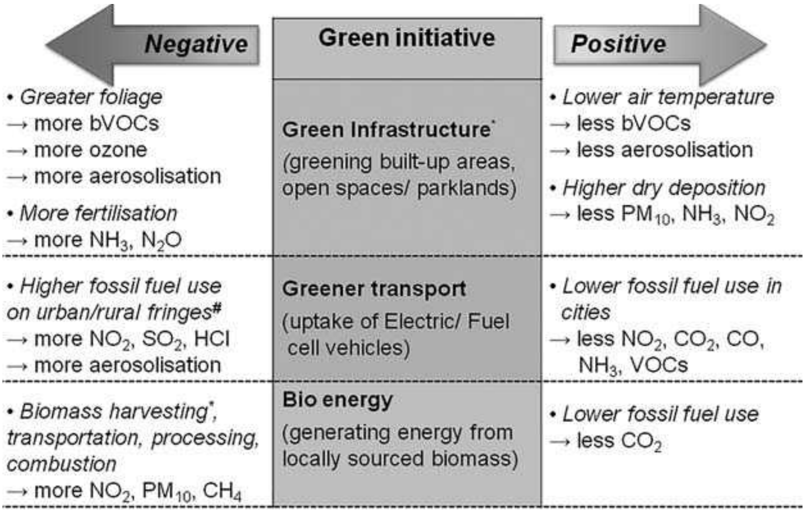
Source: Tiwary and Colls.[3]

Case Study II—System Analysis of a Project Integrating Green Initiatives

This section demonstrates the application of a systems approach in assessing the environmental impacts of low-carbon urban planning as a green project. Interactions between a suite of cross-sectoral green initiatives have been considered for two scenarios—status quo and aggressive, including decarbonizing of road transport, decentralizing energy production through biomass plants, and increasing the urban green vegetation cover. The study utilized real and projected information for the North East region of the United Kingdom to assess air quality implications on a systems scale. Using a geo-spatial analysis framework, it captures both the direct and the second-order environmental impacts arising from the interplay between different anthropogenic and biogenic components of green initiatives to ensure sustainable development through amelioration of local (and regional) air quality while minimizing climate change impacts. The assessment is mainly confined to pollutants that are currently of particular concern, including both primary and secondary PM (considered here as the combined pool of PM_{10}, $PM_{2.5}$, i.e., particles with aerodynamic diameters <10 and <2.5 μm, respectively), ozone (O_3) and nitrogen dioxide (NO_2).

Cross-Sectoral Assessment Framework for a Green Project

A systems scale assessment framework, encompassing plausible combinations of emerging green initiatives that will be implemented over the next 10–20 years (around 2020/2030 horizons) is presented in Figure 7. It draws together the evidence-base from available literature on cross-disciplinary climate change and urban sustainability research, applying a cross-sectoral approach to three broad categories of green initiatives, including (a) use of vegetation, (b) low-emission personal transport, and (c) renewable energy from biomass. Each initiative (shown in boldface text) is characterized by a set of positive and negative environmental impacts (shown in italicized text) with their resulting air quality implications. These depend on the activities involved and their influence on either formation or removal of air pollutants. This was considered as an essential first step toward scoping the systems framework of landscape interactions between biogenic emissions (primarily biogenic volatile organics, bVOCs) and the anthropogenic emissions from fossil fuel combustion (mainly from transport and energy sources) in



Negative	**Green initiative**	**Positive**
• *Greater foliage* → more bVOCs → more ozone → more aerosolisation • *More fertilisation* → more NH_3, N_2O	**Green Infrastructure*** (greening built-up areas, open spaces/ parklands)	• *Lower air temperature* → less bVOCs → less aerosolisation • *Higher dry deposition* → less PM_{10}, NH_3, NO_2
• *Higher fossil fuel use on urban/rural fringes*$^{\#}$ → more NO_2, SO_2, HCl → more aerosolisation	**Greener transport** (uptake of Electric/ Fuel cell vehicles)	• *Lower fossil fuel use in cities* → less NO_2, CO_2, CO, NH_3, VOCs
• *Biomass harvesting***, transportation, processing, combustion* → more NO_2, PM_{10}, CH_4	**Bio energy** (generating energy from locally sourced biomass)	• *Lower fossil fuel use* → less CO_2

FIGURE 7 Theoretical framework illustrating the positive and negative implications for air quality from emerging green initiatives (2020/2030 horizons).
Source: Tiwary et al.[26]

future Green Cities. This step incorporates both the positive and the negative effects arising from such interactions, in order to assess the overall sustainability implications of the green initiatives.

Uptake of greener transportation technologies, through a combination of low-emitting internal combustion and electric/fuel-cell traction, is projected to reduce primary emissions of CO, NO_x and PM from vehicle use. However, such initiatives can be considered green only to the extent that the source of energy supply is renewable. Fuel cell-powered vehicles may still be associated with pollutant emissions in peri-urban regions if the hydrogen is generated by fossil fuel sources, which would contribute to additional aerosols from atmospheric reactions of SO_2, NO_x, NH_3 and VOCs, originating from the refineries. Incorporating biomass into the future energy mix is meant to keep the decarbonized energy generation affordable but at the cost of air quality. Systems-level assessments of different fuel mixes have found increased N_2O, NH_3 and primary PM from the harvest phase and enhanced NO_x, CH_4 and secondary aerosol (SA) formation potentials from the combustion phase.

The effect of land cover modification on ambient temperature is another key driver to energy demand (for cooling/heating) and corresponding pollutant emissions (both in terms of primary components from associated activities as well as the photochemical precursors, i.e., chemicals that lead to tropospheric ozone production). While climate change effects are projected to contribute to aggravation of the urban heat island effects (and an overall increase in ambient temperature), lower air temperatures (resulting from vegetation cover modifications and retrofitting initiatives to enhance evapotranspiration and albedo-effects) have shown reduction in cooling electricity demand. The implications appear to be varied at global levels depending on the green initiatives pursued and more deterministic scenario modeling would be required to inform the outcomes.

Conclusion

This chapter introduced some basic concepts of system analysis of “greener” processes and projects, specifically in the context of holistic environmental management. It explained the implementation of transboundary material flows and life cycle thinking in terms of ensuring global sustainability, by avoiding unacceptable implications for interventions in wider process chains—both in the supply streams and in the waste streams.

It is noted that uncertainty within systems analysis can be neither neglected nor simply addressed. This is mainly because the information about uncertainty in science-driven data cannot be fully captured within the inventoried databases. This chapter also provides an overview of different sources of uncertainly and the available methods to resolve them. A significant share of this uncertainty arises in practice, based on the relationship between the data and the intended reality being modeled. Software and algorithms should be applied rigorously by researchers on a breadth of real-world case studies in order to closely identify which parameters are the major sources of uncertainty in LCA results.

Two example case studies illustrate the role of LCA and geo-spatial approaches in assessing the greenhouse gases and air quality emissions across system boundaries, respectively, for a process and a project. It shows the merit of this approach in revealing the hotspots contributing to potential environmental impacts while devising green solutions viable at systems scale. Some discussion is presented to develop possible mitigation for transboundary environmental management of the problem, including all associated activities in the process chain. Overall, this chapter provides insight into the role of systems approach in efficient environmental management, through assessment of the “true green credentials” of a process or a project in terms of its greenhouse gas emissions, ecotoxicity, air pollution, etc. within the system.

References

1. Baumann, H.; Tillman, A. *The Hitch Hiker’s Guide to LCA*; Studentlitteratur AB: Lund, Sweden, 2004.
2. ISO 14044. *Environmental Management—Life Cycle Assessment—Requirements and Guidelines*; International Organisation for Standardisation: Geneva, Switzerland, 2006.

3. Tiwary, A.; Colls J. Mitigating secondary aerosol generation potentials from biofuel use in the energy sector. *Sci. Total Environ*. **2010**, *408*, 607–616.
4. Sonnemann, G.W.; Schuhmacher, M.; Castells, F. Uncertainty assessment by a Monte Carlo simulation in a life cycle inventory of electricity produced by a waste incinerator. *J. Clean. Prod*. **2003**, *11*, 279–292.
5. von Bahr, B.; Steen, B. Reducing epistemological uncertainty in life cycle inventory. *J. Clean. Prod*. **2004**, *12*, 369–388.
6. Maurice, B.; Frischknecht, R.; Coelho-Schwirtz, V.; Hungerbuhler, K. Uncertainty analysis in life cycle inventory. Application to the production of electricity with French coal power plants. *J. Clean. Prod*. **2000**, *8*, 95–108.
7. Hertwich, E.G.; Hammitt, J.K. A decision-analytic framework for impact assessment. Part 1: LCA and decision analysis. *Int. J. Life Cycle Assess*. **2001**, *6* (1), 5–12.
8. Frischknecht, R., et al. The ecoinvent database: Overview and methodological framework. *Int. J. Life Cycle Assess*. **2005**, *10* (1), 3–9.
9. Lozano, S.; Iribarren, D.; Moreira, M.T.; Feijoo, G. The link between operational efficiency and environmental impacts: A joint application of life cycle assessment and data envelopment analysis. *Sci. Total Environ*. **2009**, *407* (5), 1744–1754.
10. Heller, M.C.; Keoleian, G.A.; Mann, M.K.; Volk, T.A. Life cycle energy and environmental benefits of generating electricity from willow biomass. *Renew. Energy* **2004**, *29* (7), 1023–1042.
11. Kaltschmitt, M.; Reinhardt, G.A.; Stelzer, T. Life cycle analysis of biofuels under different environmental aspects. *Biomass Bioenergy* **1997**, *12*, 21–34.
12. Lewandowski, I.; Heinz, A. Delayed harvest of miscanthus—Influences on biomass quantity and quality and environmental impacts of energy production. *Eur. J. Agron*. **2003**, *19* (1), 45–63.
13. Thornley, P. Airborne emissions from biomass based power generation systems. *Environ. Res. Lett*. **2008**, *3*, 6–8.
14. Faaij, A., et al. Externalities of biomass based electricity production compared with power generation from coal in the Netherlands. *Biomass Bioenergy* **1998**, *14*, 125–147.
15. Ryu, C., et al. Effect of fuel properties on biomass combustion: Part I. Experiments—Fuel type, equivalence ratio and particle size. *Fuel* **2006**, *85* (7–8), 1039–1046.
16. Öko-Institut. *Global Emission Model of Integrated System (GEMIS) Database*; Öko-Institut: Darmstadt, Germany, 2006; available at http://www.oeko.de.
17. McDonald, J.D., et al. Fine particle and gaseous emission rates from residential wood combustion. *Environ. Sci. Technol*. **2000**, *34* (11), 2080–2091.
18. Rivela, B.; Moreira, M.T.; Muñoz, I.; Rieradevall, J.; Fei-joo, G. Life cycle assessment of wood wastes: A case study of ephemeral architecture. *Sci. Total Environ*. **2006**, *357* (1–3), 1–11.
19. Damen, K.; Faaij, A. *A Life Cycle Inventory of Existing Biomass Import Chains for "Green" Electricity Production, NW&S-E-2003-01*; Utrecht University: Utrecht, The Netherlands, 2003.
20. Wierzbicka, A., et al. Particle emissions from district heating units operating on three commonly used biofuels. *Atmos. Environ*. **2005**, *39* (1), 139–150.
21. Rodrigues, M.; Faaij, A.; Walter, A. Techno-economic analysis of co-fired biomass integrated gasification/combined cycle systems with inclusion of economies of scale. *Energy* **2003**, *28* (12), 1229–1258.
22. Lucas, C.; Szewczyk, D.; Blasiak, W.; Mochida, S. High-temperature air and steam gasification of densified biofuels. *Biomass Bioenergy* **2004**, *27* (6), 563–575.
23. Faaij, A. Bio-energy in Europe: Changing technology choices. *Energy Policy* **2006**, *34* (3), 322–342.
24. Lewandowski, I., et al. The potential biomass for energy production in the Czech Republic. *Biomass Bioenergy* **2006**, *30* (5), 405–421.
25. Hamelinck, C.N.; Suurs, R.A.; Faaij, A. *International Bioenergy Transport Costs and Energy Balance*; Utrecht University, Copernicus Institute Science Technology Society: Utrecht, The Netherlands, 2003.
26. Tiwary, A., et al. Systems scale assessment of the sustainability implications of emerging green initiatives. *Environ. Pollut*. **2013**, 183, 213–223.

45

Green Products: Production

Puangrat Kajitvichyanukul,
Jirapat Ananpattarachai,
and Apichon Watcharenwong

Introduction

Since 1990, there has been a wide concern for environmental product development and how to achieve green products. Products affect the environment in several ways. Once a product has been designed and generated from the manufacturing process, its environmental attributes are largely fixed along its life cycle. Environmental requirements have been made by several attempts to ultimately obtain environmentally friendly products. For example, design processes are aimed to minimize raw material and energy consumption, waste generation, health and safety risk, and ecological degradation. Several terminologies of product design integrating environmental concerns were introduced in the late 1990s. These terms include design for the environment,[1–3] life cycle analysis,[2] green chemistry,[4–6] pollution prevention,[7–10] environmentally conscious manufacturing,[11,12] and sustainable development.[13] The aggregation of these terms is called "industrial ecology."[14] In 1996, the Office of Pollution Prevention and Toxics of the U.S. Environmental Protection Agency (EPA) set up two programs, Design for the Environment (DfE) and green chemistry, under the industrial ecology scope. The DfE has been defined as follows: "Design for environment is the systematic process by which firms design products and processes in an environmentally conscious way."[15] The broad definition of green chemistry has been also provided as follows: "Green chemistry is the design of chemical products and processes that reduce or eliminate the use and generation of hazardous substances and seek to reduce and prevent pollution at source."[16] Since 1995, green chemistry has played an important role in green product development, especially chemical products. Nowadays, green chemistry has become an important tool in achieving sustainable development. The implementation of green chemistry—the design of chemical products and processes that reduce or eliminate the use and generation of hazardous substances—has become essential for increasing the standard of human living without having a negative impact on the environment. In addition, green chemistry provides solutions to critical global challenges such as climate change, sustainable agriculture, energy, toxics in the environment, and depletion of natural resources.

In this entry, the concepts in green chemistry for producing green products, especially involving the chemical industrial process, are reviewed. It is noted that chemical products are largely created using energy-intensive processes and non-renewable, petroleum-based resources as feedstocks.[17] Many of these industrial processes use hazardous materials or generate wastes that are harmful to human health and the environment. Thus, introducing green chemistry concepts through these chemical products is a particularly good example of developments in this field that may inspire future sustainable development.

Principles of Green Chemistry

The principles of green chemistry have been refined from a diverse set of practices and emerging research. The EPA defines green chemistry as the use of chemistry for pollution prevention at the molecular level.[18] The mission of green chemistry is to promote innovative chemical technologies that reduce or eliminate the use or generation of hazardous substances in the design, manufacture, and use of chemical products.[19] They address alternative starting and target materials; alternative reagents, solvents, and catalysts; and improved processes and process control in producing chemical products. Green chemistry consists of chemicals and chemical processes designed to reduce or eliminate negative environmental impacts.

The principles of green chemistry were originally published by Pual Anastas and John Warner in *Green Chemistry: Theory and Practice* (Oxford University Press: New York, 1998). The 12 principles of green chemistry provide a road map to implement green chemistry. Following Anatas and Warner,[20] the principles are as follows:

1. *Prevention.* It is better to prevent waste than to treat or clean up waste after it has been created.
2. *Atom economy.* Synthetic methods should be designed to maximize the incorporation of all materials used in the process into the final product.
3. *Less hazardous chemical syntheses.* Wherever practicable, synthetic methods should be designed to use and generate substances that possess little or no toxicity to human health and the environment.
4. *Designing safer chemicals.* Chemical products should be designed to effect their desired function while minimizing their toxicity.
5. *Safer solvents and auxiliaries.* The use of auxiliary substances (e.g., solvents, separation agents) should be made unnecessary wherever possible and innocuous when used.
6. *Design for energy efficiency.* Energy requirements of chemical processes should be recognized for their environmental and economic impacts and should be minimized. If possible, synthetic methods should be conducted at ambient temperature and pressure.
7. *Use of renewable feedstocks.* A raw material or feedstock should be renewable rather than depleting whenever technically and economically practicable.
8. *Reduce derivatives.* Unnecessary derivatization (use of blocking groups, protection/deprotection, temporary modification of physical/chemical processes) should be minimized or avoided if possible because such steps require additional reagents and can generate waste.
9. *Catalysis.* Catalytic reagents (as selective as possible) are superior to stoichiometric reagents.
10. *Design for degradation.* Chemical products should be designed so that at the end of their function, they break down into innocuous degradation products and do not persist in the environment.
11. *Real-time analysis for pollution prevention.* Analytical methodologies need to be further developed to allow for real-time, in-process monitoring and control before the formation of hazardous substances.
12. *Inherently safer chemistry for accident prevention.* Substances and the form of a substance used in a chemical process should be chosen to minimize the potential for chemical accidents, including releases, explosions, and fires.

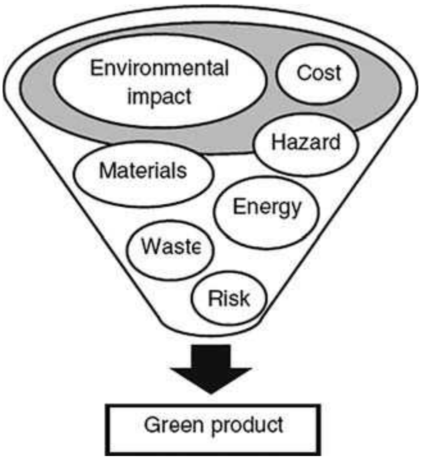


FIGURE 1 Objectives of green chemistry.

Green chemistry addresses hazards, whether physical (flammability, explosivity), toxicological (carcinogenicity, endocrine disruption), or global (ozone depletion, climate change) as an inherent property of a molecule.[21] Green chemistry also addresses all aspects of chemical designing manufacturing.[22,23] The four general objectives derived from the green chemical philosophy are as follows:[24]

- Reduction of use and generation of polluting chemicals in the chemical process
- Reduction of use of dangerous chemicals in the chemical process
- Reduction of the harmful effects of final products.
- Reduction of the use of exhaustible feedstock materials and of scarce resources

The objectives of green chemistry are shown in Figure 1. These concepts lead the movement on the front lines of research, education, and incentive and research programs of the EPA, National Science Foundation, U.S. Department of Energy, U.S. Department of Agriculture, and international organizations worldwide.

Approaching Sustainability through Principles

During the latter half of the 20th century, an exponentially increasing number of environmental regulations were enacted in developed countries. In the United States, the Clean Air Act of 1970, the Clean Water Act of 1972, the Resource Conservation and Recovery Act of 1976, and their amendments, have set the fundamental regulations covering air, water, and hazardous waste pollutants, respectively. In 2001, the EPA promulgated pharmaceutical maximum achievable control technology standards that regulate 187 hazardous air pollutants, including solvents commonly used in the pharmaceutical industry. In Europe, the European Union has a set of common rules for permitting and controlling industrial installations in the Integrated Pollution Prevention and Control (IPPC) Directive of 1996.[25,26] The IPPC Directive is about minimizing pollution from industrial sources throughout the European Union. In 2007, REACH, the new Regulation on Registration, Evaluation, Authorization, and Restriction of Chemicals, was launched and considered a tool for green chemistry. This regulation improves the former legislative framework on chemicals of the European Union.[27] One of the main aims of REACH is to improve the protection of human health and the environment from the risks that can be posed by chemicals. The major concern of REACH is chemical safety and in that it is one of the green chemistry metrics.[28]

As environmental regulation has become more stringent owing to climate concern, industries have to look for innovative ways to obtain purer final products using a reduced amount of raw materials and energy and generating a smaller amount of waste. Concepts of green chemistry have become a solution for the industrial requirement. Many "green" terminologies have emerged and been defined for use in environmental product development. Green engineering has been defined as follows: "Green engineering is the design, commercialization and use of processes that are feasible and economic, reduce the generation of pollution at the source, and minimize the risk to human health and the environment."[16] Green technology is also widely used as it involves using science to create technologies that conserve natural resources and lessen the human impact on the environment.

Green chemistry, green engineering, and green technology have been focused on and proposed by several scientists and researchers worldwide to respond to the need of industry. Collaboration among industries, academic institutes, and government agencies were initiated to push green chemistry to the manufacturing process. In 2003, the collaboration of industry, the EPA, and the U.S. Department of Commerce's Manufacturing Extension Partnership, called the Green Suppliers Network, was officially launched in the automotive industry, and is currently in the aerospace, pharmaceutical and health-care, and office furniture industries. The established outcomes of this program are the effective processes and products with higher profits while minimizing environmental impact. In 2005, the Committee on Grand Challenges for Sustainability in the Chemical Industry arranged a workshop on identifying research needs to enable the industry to develop products and processes for achieving the goal of sustainability. In addition, the pharmaceutical industry and the American Chemical Society Green Chemistry Institute collaborated to encourage innovation while catalyzing the integration of green chemistry and green engineering in the pharmaceutical industry. To support the collaboration in green chemistry, the Green Chemistry Research and Development Act of 2005[29] was proposed and passed by the House Science Committee in April 2005. This bill promotes green chemistry by authorizing a coordinated green chemistry research and development program.

The emergence of these collaborations is the driving force behind the development of green business. It is reported that green chemistry is drawing many investors from different companies to invest in producing green product lines. For example, Wilmar International, a Singaporean company, plans to launch production of derivatives of glycerol through a green chemistry process. Mitsubishi Chemical and Thai Company PTT plc are jointly conducting a study focused on the production of biosuccinic acid and biopolybutylene succinate. Moreover, the Japanese trading house Mitsui & Co. and Toyota Tsusho are also involved in green chemistry investments in the Asian regions.[30]

Nowadays, concepts of green chemistry have widely spread to the manufacturing process with the understaning that it is a superior, innovative chemistry that is cost-effective and has minimal impact on the environment. The applications of the concepts through the manufacturing process are shown in Figure 2. The types of products and processes adopting green chemistry principles include medicine, food, and energy production; packaging materials; household and commercial cleaning products; electronics and automotive chemicals; and a wide range of consumer goods.[31]

Advances in Green Chemistry to Produce Green Products

Catalysts

Catalysts play a significant role in green chemistry by decreasing energy requirements, increasing selectivity, and permitting the use of less hazardous reaction conditions. The first example in applying green chemistry to the catalytic process is the synthesis of maleic anhydride.[32–34] Originally, the benzene process was used; however, it was substituted by the butane process for the following reasons:

- The loss of two carbon atoms (starting from benzene) is avoided.
- Using butane, the by-products are carbon oxides and a small amount of acetic acid. With benzene, several byproducts are formed.
- The toxicity aspects related to the use of benzene are avoided, reducing costs related to safety systems and benzene handling.

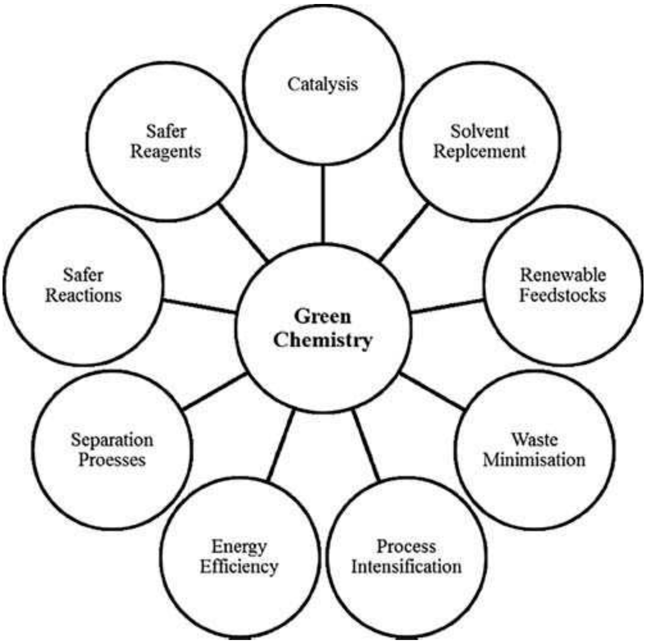


FIGURE 2 Applications of concepts through the manufacturing process.

The benefits of using the butane process following the green chemistry concept are as follows: 1) use of non-toxic reactants; 2) improved atom economy; 3) a complex multistep transformation (the reaction is a 14 e^- oxidation, with abstraction of 8 H atoms from the butane molecule and insertion of 3 O atoms) is realized in one single step, without using solvents; and 4) waste formation is minimized.

Synthesizing adipic acid directly from cyclohexene using an aqueous 30% hydrogen peroxide solution, a catalytic amount of Na_2WO_4, and a phase-transfer catalyst is another green chemistry application.[35] Previously, adipic acid was derived from the reduction of benzene to cyclohexene under high temperature and pressure resulting in nitrous oxide, a greenhouse gas and ozone-depleting substance, as an undesirable product. Using the new process, adipic acid is produced with a high yield without the use of organic solvents and without the generation of nitrous oxide.[36] Using the concept of on-site production is another alternative way in green chemistry that is already used in some new production processes to avoid the storage of toxic or dangerous chemicals. The on-site production of HCN in the manufacturing process of toluene diisocyanate has already been implemented. This was made possible by the combined development of new catalysts and new reactor engineering solutions that allow an economic production even in a small size plant.[34]

Nowadays, biocatalysts, both isolated-enzyme and whole-cell systems, are increasingly being used to assist in synthetic routes to complex molecules. The biggest role for biocatalysis is found in the pharmaceutical sector.[37–39] Biocatalysis is one of the greenest technologies for the synthesis of chiral molecules due to exquisite regioselectivity and stereoselectivity in water under mild conditions. Recently, Tao and Xu[40] reviewed the development of novel biocatalytic pharmaceutical processes to replace chemical routes with poorer process efficiency and higher manufacturing costs. It is emphasized that to empower the green chemistry feature of biocatalysis, it is essential to integrate enzymatic transformations strategically into chemical transformations at the retrosynthetic level. Enzymatic catalysis can provide a new dimension for route redesign to meet both the process and green chemistry metrics. It is underscored that recent advances in large-scale DNA sequencing and enzyme-directed evolution rendered biocatalysis a much more practical technology to provide green chemistry solutions for the industrial production of chemicals.

Nanomaterials

In the search for environmentally friendly materials, recently green chemistry has shown its role in nanomaterials, especially in the preparation and synthesizing steps. A solvent is considered an important chemical that is widely used in synthesizing nanomaterials. It is an important parameter that determines the green nature of the nanoparticle catalysis, regarding the fact that the solvent accounts for 50% of the greenhouse gas emissions during post-treatment and 60% of the energy used in pharmaceutical processes.[41] Recently, there have been several attempts to use other chemicals to substitute these harmful reducing agents and solvents to obtain engineered nanoparticles. For solvent substitution, ionic liquids are very promising replacements for the traditional volatile organic solvents owing to their high mobility, low melting points, negligible vapor pressure, thermal stability, low toxicity, large electrochemical window, non-flammability, and ability to dissolve a variety of chemicals.[42–45] Ionic liquids are versatile solvents that retain their liquid states over a wide range of conditions, but have a minimal vapor pressure. They also possess interesting properties that are a combination of the characteristics of water and organic solvents. The use of ionic liquids as stabilizers for nanoparticles was first demonstrated by Dupont et al. in 2002.[46] Ionic liquids can be functionalized and attached with weak coordination groups to further enhance the stability as well as catalytic activity in the synthesis reaction. Zhao et al.[47] used nitrile- functionalized ionic liquids to protect Pd nanoparticles with an average diameter of 5 nm. The resulting Pd nanoparticles were excellent recyclable catalysts, and an optimum balance between stability and reactivity can be achieved. Ionic liquids widely used in nanomaterial synthesis are BMI(BF_4), BMI(NTf_2), BMI(PF_6), and C_2OHMIM(BF_4). Nowadays, various nanostructured materials, such as iridium,[46] palladium,[47,48] gold,[49] tellurium,[50] TiO_2,[51,52] ZnO,[53–56] and CoPt[57] have been synthesized in ionic liquids.

One of the concerns in the process of preparation of nanomaterials is the choice of the reducing agent. zIt is reported in the literature that in the preparation of super-paramagnetic nanoparticles, the reductants used to date include hydrazine,[58] sodium borohydride,[59] carbon monoxide,[60] and dimethylformamide.[61] All of these are highly reactive chemicals and pose potential environmental and biological risks.[62] As appeared in recent literature reviews, many interesting methods are currently being applied to the green preparation of nanoparticles, called biosynthesis of nanoparticles by glucose,[62] starch,[63–65] plant leaf broth,[66] edible mushroom extract,[67] apiin,[68] and by latex of *Jatropha curcas*.[69] These biological extracts are used as reducing agents in the green synthesis of nanoparticles, especially in metal nanoparticle synthesis. Raveendran et al.[70] reported a completely "green" synthesis of Ag nanoparticles with sizes of 1–8 nm using starch as a capping agent and α-D-glucose as a reducing agent. Mallikarjuna and Varma[71] described a simple method for the shape-controlled synthesis of nanostructures (with size 2–15 nm) of noble metals such as Au, Ag, Pd, and Pt by microwave-assisted spontaneous reduction of noble metal salts using α-D-glucose, sucrose, and maltose in aqueous solution. Fructose, glucose, and sucrose were used as reducing and capping agents for the preparation of Au, Ag, Pd, Pt, and Au–Ag nanoparticles by heating to dryness for 2 h the solutions in a hot water bath.[72] The effect of two different sugars on the sizes and surface morphology for Ag nanoparticle synthesis was also reported by Filippo et al.[73] Furthermore, Nersisyan et al.[74] developed an effective way of preparing a nanosized (10–50 nm) colloidal dispersion of Ag using glucose as a reducing agent. Lu et al.[62] also reported a facile and green method to synthesize superparamagnetic Fe_3O_4 nanoparticles with α-D-glucose as the reductant, which is a mild, renewable, inexpensive, and non-toxic reducing agent, and without any additional stabilizer and dispersant.

Solvents

Solvents play an important role in chemical synthesis. Considering green chemistry requirements, water is an extremely attractive solvent choice. It is inexpensive and readily available, non-toxic, non-carcinogenic, and nonflammable. In addition, water is probably the least expensive among all available solvents.[75]

Besides water, supercritical fluids (SCFs) draw attention from chemists for their several commercial applications. These SCFs are gases that are compressed until their density approaches that of liquids. Compression can occur only above the "critical temperature" of the fluid, since at lower temperatures the fluid liquefies upon compression. Therefore, an SCF is a gas that displays some of the properties of a liquid. The advantages of using SCFs are related to the absence of toxic residues, the relatively low temperature needed for the extraction process, and their higher solvent power.[34] The SCFs are characterized by high diffusivity, low viscosity, and intermediate density. Among the SCFs, supercritical CO_2 (SC-CO_2) is the most widely used solvent in many applications. It is non-toxic, noncarcinogenic, and non-flammable. Oakes et al.[76] investigated the catalysis of the Diels–Alder reaction between *n*-butyl acrylate and cyclopentadiene in SC-CO_2. Scandium tris (trifluoromethanesulfonate) was selected as the Lewis acid catalyst. By varying the pressure of the solvent, a significant improvement over selectivity is achieved in conventional solvent. It was reported that rate enhancement was observed under SC-CO_2 conditions.

High boiling alcohols, including glycerol and polyethylene glycol, are also considered as "green solvents." They have recently attracted much attention as reaction media.[77,78] These high boiling alcohols are of low toxicity and volatility, and are biodegradable, non-expensive, and easily functionalized. These properties represent important environmentally benign characteristics. Glycerol is a natural product widely used as a cosmetic and food additive. Polyethylene glycol is also non-toxic and is widely used in the pharmaceutical,[79] semiconductor,[80] and food industries.[81]

Application of Green Chemistry in the Manufacturing Process

There are several ways to apply green chemistry to the manufacturing process for industries. The concepts can be applied through waste minimization, waste or material recycling, solvent selection, atom utilization, intensive processing, and clean synthesis. According to the Presidential Green Chemistry Challenge from the EPA, three basic methods are currently used in industrial processes:

1. Applications of greener synthetic pathways—Many green pathways for a new chemical product are widely used. It involves using a novel, green pathway to redesign the synthesis of an existing chemical product. The synthetic pathways include
   - Use greener feedstocks that are innocuous or renewable (e.g., biomass, natural oils)
   - Use novel reagents or catalysts, including biocatalysts and microorganisms
   - Apply natural processes, such as fermentation or biomimetic synthesis
   - Use atom-economical systems
   - Use convergent syntheses

An example of applying the synthetic pathway in the manufacturing process is from Dow and BASF, winners of the 2010 Greener Synthetic Pathways Award (EPA). Both companies have jointly developed a new route to make propylene oxide with hydrogen peroxide that eliminates most of the waste and greatly reduces water and energy use.

2. Use of greener reaction conditions—This application is mainly about improving conditions other than the overall design or redesign of a synthesis, including greener analytical methods. The greener reaction conditions include the following:
   - Substitute hazardous solvents with solvents that have a reduced impact on human health and the environment
   - Use solventless reaction conditions and solid-state reactions
   - Use novel processing methods
   - Eliminate energy- or material-intensive separation and purification steps
   - Improve energy efficiency, including reactions running closer to ambient conditions

A successful example of applying green reaction conditions is from Merck and Codexis. Both companies have developed a second-generation green synthesis of sitagliptin, the active ingredient in Januvia™, a treatment for type 2 diabetes. This collaboration has led to an enzymatic process that reduces waste, improves yield and safety, and eliminates the need for a metal catalyst. This work also received the 2010 Greener Reaction Conditions Award from the EPA.

3. Design of greener chemicals—Applications include the design of chemical products that are less hazardous than the products or technologies they replace. These chemical products are:
    - Less toxic than current products
    - Inherently safer with regard to accident potential
    - Recyclable or biodegradable after use
    - Safe for the atmosphere (e.g., do not deplete ozone or form smog)

An example of the design of green chemicals is spinosad. It is an environmentally safe pesticide but is not stable in water and therefore cannot be used to control mosquito larvae. Clarke, the winner of the 2010 Designing Greener Chemicals Award from the EPA, has developed a way to encapsulate spinosad in a plaster matrix, allowing it to be released slowly in water and provide effective control of mosquito larvae. This pesticide, Natular™, replaces organophosphates and other traditional, toxic pesticides and is approved for use in certified organic farming.

Conclusion

The concepts of green chemistry for producing green products and examples of green chemistry implementation in industrial processes have been discussed in this entry. It is known that many industrial processes use hazardous materials or generate wastes that are harmful to human health and the environment. Implementation of green chemistry to the design of chemical products and processes that reduce or eliminate the use and generation of hazardous substances has become essential to increase the standard of human living without the adverse effect on the environment. Using the green production system in the industry can gain the benefit of enabling chemical production that is more profitable, less wasteful, less damaging to the environment, and more socially acceptable. In this entry, the applications of green chemistry through green product production have been reviewed. The process of green products in industrial manufacturing can be applied through several green concepts, including waste minimization, waste or material recycling, solvent selection, atom utilization, intensive processing, and clean synthesis. The challenge for the industrial research and development sector is to develop products, processes, and services to improve the quality of life, the environment, and the industry competitiveness with the sustainable aspect toward the future. Recently, green chemistry has been implemented in several industrial products, including medicines, food, energy, packaging materials, household and commercial cleaning products, electronic and automotive chemicals, and a wide range of consumer goods. Biocatalysts, nanoparticles, and green solvents are the next challenge in green synthesis to produce high-quality merchandise to serve the need of consumers. Beyond green chemistry, green engineering and green technology have been focused on and proposed by several scientists and researchers worldwide. They are in the process of reinforcing steps to create cleaner production in industrial processes in the near future.

References

1. Allenby, B.R.; Richards, D. *The Greening of Industrial Ecosystems*; National Academy Press: Washington, D.C., 1994.
2. Graedel, T.E.; Allenby, B.R. *Industrial Ecology*; Prentice Hall: New York, 1995.
3. Graedel, T E.; Allenby, B.R. *Design for Environment*; Prentice Hall: New York, 1996.

4. Anastas, P.T. Benign by design chemistry. In *Benign by Design: Alternative Synthetic Design for Pollution Prevention*; Anastas, P.T., Farris, C.A., Eds.; ACS Symposium Series 577; American Chemical Society: Washington, D.C., 1994; 2–22.
5. Anastas, P.T.; Williamson, T.C. Green chemistry: An overview. In *Green Chemistry: Designing Chemistry for the Environment*; Anastas, P.T., Williamson, T.C., Eds.; ACS Symposium Series 626; American Chemical Society: Washington, D.C., 1996; 1–17.
6. Garrett, R.L. Pollution prevention, green chemistry, and the design of safer chemicals. In *Designing Safer Chemicals: Green Chemistry for Pollution Prevention*; ACS Symposium Series 640; DeVito, S.C., Garrett, R.L., Eds.; American Chemical Society: Washington, D.C., 1996; 2–15.
7. Socolow, R.; Andrews, C.; Berkhout, F.; Thomas, V. *Industrial Ecology and Global Change*; Cambridge University Press: Cambridge, U.K., 1994.
8. Frosch, R.; Gallopoulos, N. Strategies for manufacturing. Sci. Am. **1989**, *261* (3), 144–152.
9. Becker, M.; Ashford, N.A. Exploiting opportunities for pollution prevention in EPA enforcement agreements. Environ. Sci. Technol. **1995**, *29* (5), 220A–226A.
10. Breen, J.J.; Dellarco, M.J. Pollution prevention: The new environmental ethic. In *Pollution Prevention in Industrial Processes: The Role of Process Analytical Chemistry*; Breen, J.J., Dellarco, M.J., Eds.; ACS Symposium Series 508; American Chemical Society: Washington, D.C., 1992; 2–12.
11. Baccini, P.; Brunner, P.H. *Metabolism of the Anthropo sphere*; Springer-Verlag: New York, 1991.
12. Fiksel, J. Introduction. In *Design for Environment: Creating Eco-Efficient Products and Processes;* Fiksel, J., Ed.; McGraw-Hill: New York, 1996; 2–8.
13. National Science and Technology Council. *Technology for a Sustainable Future: A Framework for Action;* Office of Science and Technology: Washington, D.C., 1994.
14. Anastas, P.T.; Breen, J.J. Design for the environment and Green Chemistry: The heart and soul of industrial ecology. J. Cleaner Prod. **1997**, *5* (1–2), 97–102.
15. Lenox, M.; Jordan, B.; Ehrenfeld J. In *The Diffusion of Design for Environment: A Survey of Current Practice*, Proceedings of the IEEE International Symposium on Electronics and the Environment, Dallas, TX, May 6–8, 1996; 25–30.
16. Mason, T.J. Sonochemistry and the environment—Providing a "green" link between chemistry, physics and engineering. Ultrason. Sonochem. **2007**, *14* (4), 476–483.
17. Nameroff, T.J.; Garant, R.J.; Albert, M.B. Adoption of green chemistry: An analysis based on US patents. Res. Policy **2004**, *33* (6–7), 959–974.
18. U.S. Environmental Protection Agency, Green Chemistry, available at http://www.epa.gov/greenchemistry (accessed May 14, 2012).
19. Anastas, P.T., Heine, L.G., Williamson, T.C. Green chemical syntheses and processes: Introduction. In *Green Chemical Syntheses and Processes*; Anastas, P.T., Heine, L.G., Williamson, T.C., Eds.; ACS Symposium Series 767; American Chemical Society: Washington, D.C., 2000; 1–17.
20. Anastas, P.T.; Warner, J.C. *Green Chemistry: Theory and Practice*; Oxford University Press: New York, 1998.
21. Anastas, P.T. Meeting the challenges to sustainability through green chemistry. Green Chem. **2003**, G29–G34.
22. Clark J.H. Part 1 Green chemistry for sustainable development. In *Green Separation Processes*; Afonso, C.A.M., Crespo, J.G., Eds.; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2005; 1–31.
23. Poliakoff, M., Licence, P. Sustainable technology: Green chemistry. Nature **2007**, *450* (7171), 810–812.
24. Mestres R. A brief structured view of green chemistry issues. Green Chem. **2004**, (1), G10–G12.
25. European Commission. *Directive 91/61/EC Concerning Integrated Pollution Prevention and Control, No. L2J7, HMSO;* Official Journal of the European Communities, European Union: London, 1996.

26. European Commission, Joint Research Centre, Institute for Prospective Technological Studies. Reference document. Available at http://eippcb.jrc.es/reference/ (accessed May 14, 2012).
27. European Commission. Alphabetical index. Available at http://ec.europa.eu/atoz_en.htm (accessed March 14, 2012).
28. Demirci, U.B. How green are the chemicals used as liquid fuels in direct liquid-feed fuel cells? Environ. Int. **2009**, *35* (3), 626–631.
29. *Boehlert,* S.L. *H.R. 1215 (109th): Green Chemistry Research and Development Act of 2005;* Report 109–82; Committee on Science and Technology, U.S. House of Representatives: Washington, D.C., 2005; 1–65.
30. Markets and Business. Southeast Asia attracting green-chemistry investments. Focus Catalysts **2010**, *2010* (1), 2.
31. Manley, J.B.; Anastas, P.T.; Cue, B.W., Jr. Frontiers in green chemistry: Meeting the grand challenges for sustainability in R&D and manufacturing. J. Clean. Prod. **2008**, *16* (6), 743–750.
32. Centi, G.; Cavani, F.; Trifirò, F. *Selective Oxidation by Heterogeneous Catalysis;* Kluwer Academic Publishers/Plenum Press: New York, 2000.
33. Centi, G. Vanadyl pyrophosphate—A critical overview. Catal. Today **1993**, *16* (1), 5–26.
34. Centi, G.; Perathoner, S. Catalysis and sustainable (green) chemistry. Catal. Today **2003**, *77* (4), 287–297.
35. Sato, K.; Aoki, M.; Noyori, R. A “green” route to adipic acid: Direct oxidation of cyclohexenes with 30 percent hydrogen peroxide. Science **1998**, *281* (5383), 1646–1647.
36. Kirchhoff, M.M. Promoting sustainability through green chemistry. Resour. Conserv. Recycling **2005**, *44* (3), 237–243.
37. Patel, R.N. Microbial/enzymatic synthesis of chiral pharmaceutical intermediates. *Curr. Opin. Drug Discov. Devel.* **2003**, *6* (6), 902–920.
38. Buckland, B.C.; Robinson, D.K.; Chartrain, M. Biocatalysis for pharmaceuticals—Status and prospects for a key technology. *Metab. Eng.* **2000**, *2* (1), 42–48.
39. Pesti, J.A.; DiCosimo, R. Recent progress in enzymatic resolutions and desymmetrization of pharmaceuticals and their intermediates. *Curr. Opin. Drug Discov. Devel.* **2003**, *6* (6), 884–901.
40. Tao, J.; Xu, J.-H. Biocatalysis in development of green pharmaceutical processes. Curr. Opin. Chem. Biol. **2009**, *13* (1), 43–50.
41. Liu, S.; Xiao, J. Toward green catalytic synthesis—Transition metal-catalyzed reactions in non-conventional media. J. Mol. Catal. A Chem. **2007**, *270* (1–2), 1–43.
42. Welton, T. Room-temperature ionic liquids. Solvents for synthesis and catalysis. Chem. Rev. **1999**, *99* (8), 2071–2084.
43. Wasserscheid, P.; Keim, W. Ionic liquids—New “solutions” for transition metal catalysis. Angew. Chem. Int. Ed. Engl. **2000**, *39* (21), 3772–3789.
44. Dupont, J.; de Souza, R.F.; Suarez, P.A.Z. Ionic liquid (molten salt) phase organometallic catalysis. Chem. Rev.**2002** *102* (10), 3667–3692.
45. Xu, W.; Cooper, E.I.; Angell, C.A. Ionic liquids: Ion mobilities, glass temperatures and fragilities. J. Phys. Chem. *B* **2003**, *107* (25), 6170–6178.
46. Dupont, J.; Fonseca, G.S.; Umpierre, A.P.; Fichtner, P.F.P.; Teixeira, S.R. Transition-metal nanoparticles in imidazo- lium ionic liquids: Recyclable catalysts for biphasic hydrogenation reactions. J. Am. Chem. Soc. **2002**, *124* (16), 4228–4229.
47. Zhao, C.; Wang, H.-Z.; Yan, N.; Xiao, C.-X.; Mu, X.-D.; Dyson, P.J.; Kou, Y. Ionic-liquid-like copolymer stabilized nanocatalysts in ionic liquids: II. Rhodium-catalyzed hydrogenation of arenes. J. Catal. **2007**, *250* (1), 33–40.
48. Huang, J.; Jiang, T.; Han, B.X.; Gao, H.X.; Chang, Y.H.; Zhao, G.Y.; Wu, W.Z. Hydrogenation of olefins using ligand-stabilized palladium nanoparticles in an ionic liquid. Chem. Commun. **2003**, *14*, 1654–1655.

49. Kim, K.S.; Demberelnyamba, D.; Lee, H. Size-selective synthesis of gold and platinum nanoparticles using novel thiolfunctionalized ionic liquids. Langmuir **2004**, *20* (3), 556–560.
50. Zhu, Y.J.; Wang, W.W.; Qi, R.J.; Hu, X.L. Microwave-assisted synthesis of single-crystalline tellurium nanorods and nanowires in ionic liquids. Angew. Chem. Int. Ed. Engl. **2004**, *43* (11), 1410–1414.
51. Zhou, Y.; Antonietti, M. Synthesis of very small TiO_2 nanocrystals in a room-temperature ionic liquid and their selfassembly toward mesoporous spherical aggregates. J. Am. Chem. Soc. **2003**, *125* (49), 14960–14961.
52. Nakashima, T.; Kimizuka, N. Interfacial synthesis of hollow TiO_2 microspheres in ionic liquids. J. Am. Chem. Soc. **2003**, *125* (21), 6386–6387.
53. Cao, J.M.; Wang, J.; Fang, B.Q.; Chang, X.; Zheng, M.B.; Wang, H.Y. Microwave-assisted synthesis of flower-like ZnO nanosheet aggregates in a room-temperature ionic liquid. Chem. Lett. **2004**, 33 (10), 1332–1333.
54. Wang, J.; Cao, J.M.; Fang, B.Q.; Lu, P.; Deng, S.G.; Wang, H.Y. Synthesis and characterization of multipod, flowerlike, and shuttle-like ZnO frameworks in ionic liquids. Mater. *Lett.* **2005**, *59* (11), 1405–1408.
55. Goharshadi, E.K.; Ding, Y.; Nancarrow, P. Green synthesis of ZnO nanoparticles in a room-temperature ionic liquid 1-ethyl-3-methylimidazolium bis(trifluoromethylsulfonyl)imi de. J. Phys. Chem. Solids **2008**, *69* (8), 2057–2060.
56. Goharshadi, E.K.; Ding, Y.; Jorabchi, M.N.; Nancarrow, P. Ultrasound-assisted green synthesis of nanocrystalline ZnO in the ionic liquid (hmim)(NTf_2). Ultrason. Sonochem. **2009**, *16* (1), 120–123.
57. Wang, Y.; Yang, H. Synthesis of CoPt nanorods in ionic liquids. J. Am. Chem. Soc. **2005**, *127* (15), 5316–5317.
58. Hou, Y.; Kondoh, H.; Shimojo, M.; Sako, E.O.; Ozaki, N.; Kogure, T.; Ohta, T. Inorganic nanocrystal self-assembly via the inclusion interaction of β-cyclodextrins: Toward 3D spherical magnetite. J. Phys. Chem. *B* **2005**, *109* (11), 4845–4852.
59. Cain, J.L.; Harrison, S.R.; Nikles, J.A.; Nikles, D.E. Preparation of α-Fe particles by reduction of ferrous ion in lecithin/cyclohexane/water association colloids. J. Magn. Mater. **1996**, *155* (1–3), 67–69.
60. Mondal, K.; Lorethova, H.; Hippo, E.; Wiltowski, T.; Lalvani, S.B. Reduction of iron oxide in carbon monoxide atmosphere—Reaction controlled kinetics. Fuel Process. Technol. **2004**, *86* (1), 33–47.
61. Jian, P.; Yahui, H.; Yang, W.; Linlin, L. Preparation of polysulfone–Fe_3O_4 composite ultrafiltration membrane and its behavior in magnetic field. J. Membr. Sci. **2006**, *284* (1–2), 9–16.
62. Lu, W.; Shen, Y.; Xie, A.; Zhang, W. Green synthesis and characterization of superparamagnetic Fe_3O_4 nanoparticles. J. Magn. Magn. Mater. **2010**, *322* (13), 1828–1833.
63. Chairam, S.; Poolperm, C.; Somsook, E. Starch vermicelli template-assisted synthesis of size/shape-controlled nanoparticles. Carbohydr. Polym. **2009**, *75* (4), 694–704.
64. Kassaee, M.Z.; Akhavan, A.; Sheikh, N.; Beteshobabrud, R. γ-Ray synthesis of starch-stabilized silver nanoparticles with antibacterial activities. Radiat. Phys. Chem. **2008**, *77* (9), 1074–1078.
65. Vigneshwaran, N.; Nachane, R.P.; Balasubramanya, R.H.; Varadarajan, P.V. A novel one-pot "green" synthesis of stable silver nanoparticles using soluble starch. Carbohydr. Res. **2006**, *341* (12), 2012–2018.
66. Shankar, S.S.; Rai, A.; Ahmad, A.; Sastry, M. Rapid synthesis of Au, Ag, and bimetallic Au core–Ag shell nanoparticles using neem *(Azadirachta indica)* leaf broth. J. Colloid Interface Sci. **2004**, *275* (2), 496–502.
67. Philip, D. Biosynthesis of Au, Ag and Au–Ag nanoparticles using edible mushroom extract. Spectrochim. Acta A Mol. Biomol. Spectrosc. **2009**, *73* (2), 374–381.
68. Kasthuri, J.; Veerapandian, S.; Rajendiran, N. Biological synthesis of silver and gold nanoparticles using apiin as reducing agent. Colloids Surf. B Biointerfaces **2009**, *68* (1), 55–60.

69. Bar, H.; Bhui, D.K.; Sahoo, G.P.; Sarkar, P.; De, S.P.; Misra, A. Green synthesis of silver nanoparticles using latex of *Jatropha curcas***.** Colloids Surf. A Physicochem. Eng. Asp. **2009**, *339* (1–3), 134–139.
70. Raveendran, P.; Fu, J.; Wallen, S.L. Completely "green" synthesis and stabilization of metal nanoparticles. J. Am. Chem. Soc. **2003**, *125* (46), 13940–13941.
71. Mallikarjuna, N.N.; Varma, R.S. Microwave-assisted shape- controlled bulk synthesis of noble nanocrystals and their catalytic properties. Cryst. Growth Des. **2007**, *7* (4) 686–690.
72. Panigrahi, S.; Kundu, S.; Ghosh, S.K.; Nath, S.; Pal, T. Sugar assisted evolution of mono- and bimetallic nanoparticles. Colloids Surf. A Physicochem. Eng. Asp. **2005**, *264* (1–3), 133–138.
73. Filippo, E.; Serra, A.; Buccolieri, A.; Manno, D. Green synthesis of silver nanoparticles with sucrose and maltose: Morphological and structural characterization. J. Non. Cryst. Solids **2010**, *356* (6–8), 344–350.
74. Nersisyan, H.H.; Lee, J.H.; Son, H.T.; Won, C.W.; Maeng, D.Y. A new and effective chemical reduction method for preparation of nanosized silver powder and colloid dispersion. Mater. Res. Bull. **2003**, *38* (6), 949–956.
75. Li, C.-J. Water as a benign solvent for chemical syntheses. In *Green Chemistry: Frontiers in Benign Chemical Syntheses and Processes;* Anastas, P.T., Williamson, T.C., Eds.; Oxford University Press: New York, 1998; 234.
76. Oakes, R.S.; Heppenstall, T.J.; Shezad, N.; Clifford, A.A.; Rayner, C.M. Use of scandium tris(trifluoromethanesulfona te) as a Lewis acid catalyst in supercritical carbon dioxide: Efficient Diels–Alder reactions and pressure dependent enhancement of endo:exo stereoselectivity. Chem. Commun. **1999**, *16,* 1459–1460.
77. Heldebrant, D.J.; Jessop, P.G. Liquid poly(ethylene glycol) and supercritical carbon dioxide: A benign biphasic solvent system for use and recycling of homogeneous catalysts. J. Am. Chem. Soc. **2003**, *125* (19), 5600–5601.
78. Hou, Z.; Theyssen, N.; Brinkmann, A.; Leitner, W. Biphasic aerobic oxidation of alcohols catalyzed by poly(ethylene glycol)-stabilized palladium nanoparticles in supercritical carbon dioxide. Angew. Chem. Int. Ed. Engl. **2005**, *44* (9) 1346–1349.
79. Zalipsky, S.; Harris, J.M. Introduction to chemistry and biological applications of poly(ethylene glycol). In *Poly(Ethylene Glycol): Chemistry and Biological Applications;* Harris, J.M., Zalipsky, S., Eds.; ACS Symposium Series 680; American Chemical Society: Washington, D.C., 1997; 1–13.
80. Kajitvichyanukul, P.; Amornchat, P. Effects of diethylene glycol on TiO_2 thin film properties prepared by sol–gel process. Sci. Technol. Adv. Mater. **2005**, *6* (3–4), 344–347.
81. Code of Federal Regulations, *Food and Drug, 2001,* Title 21, vol. 3, CITE: 21CFR172.820; Food and Drug Administration (FDA): Washington, D.C., 2001.

Index